VISIONLIGHT: DYNAMIC FLOW–DRIVEN UNCERTAINTY MODELING IN 3D VISUAL REINFORCEMENT LEARNING FOR TRAFFIC SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Most RL-based traffic signal control (TSC) methods rely on features such as vehicle coordinates and waiting times, which are available in simulation but not at real intersections. We present **VisionLight**, an RL framework that operates on real-time video input through two modes: (1) end-to-end processing of raw footage and (2) image-based feature extraction compatible with existing TSC systems. To address unpredictable traffic fluctuations (uncertainty), VisionLight introduces an Entropy Attention & Multi-agent Mechanism tuned for turn-based traffic. It achieves an average 56.8% improvement over SOTA baselines across three metrics, matches feature-driven RL models, and generalizes robustly under extreme weather without retraining, making it practical for real-world deployment.

1 INTRODUCTION

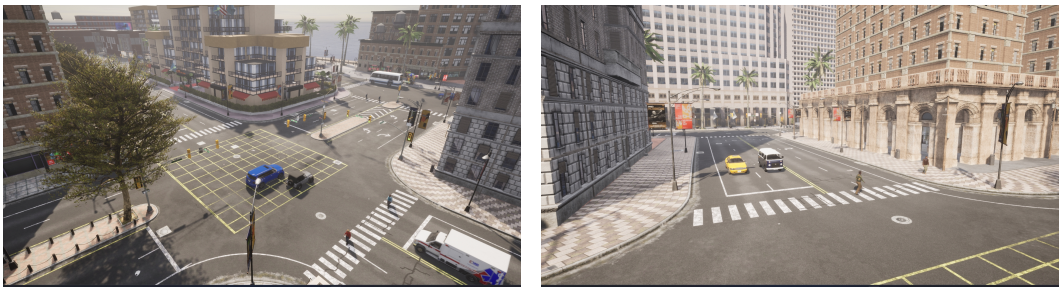


Figure 1: Illustration of video-based traffic signal control: (Left) A 3D-rendered traffic simulation environment. (Right) A camera perspective capturing real-time traffic flow for RL-based control.

Background. Traffic congestion at intersections remains a critical issue in urban areas, causing significant delays and economic losses (Fonseca & Garcia, 2021; Cheng et al., 2023). Efficient Traffic Signal Control (TSC) is essential, especially as urbanization continues to increase (Wei et al., 2019). Recent advances in Reinforcement Learning (RL) have shown great promise in improving TSC systems by dynamically adjusting signal timings in response to real-time traffic conditions, often outperforming traditional approaches in simulations (Koh et al., 2020; Li et al., 2021; Noaeen et al., 2022). As a result, RL has become a popular research direction for intersection control, with a growing body of literature exploring its potential.

GAP & PROBLEM. However, most RL-based TSC research depends on input features such as vehicle coordinates, queue lengths, or waiting times, which are easily extracted in simulations but not obtainable at real intersections (Comert & Cetin, 2021). High-precision sensors like LiDAR could provide such data, but they are costly and rarely deployed, leaving practical and scalable TSC solutions elusive.

SOLUTION. To address above challenge, our work explores both 1. *End-to-End* and 2. *Image-based Feature Extraction* approaches based on camera input. Unlike previous methods that rely on pre-processed features, our model directly operates on raw traffic video data (Li et al., 2023b; He et al., 2023; Wu et al., 2023). Benefiting from training in a highly realistic 3D-rendered simula-

tion with camera perspectives that mirror real-world conditions, we aim to bridge the gap between simulation and deployment, making the model suitable for real-world applications.

CONTRIBUTION. In summary, our contributions are as follows:

- We propose an end-to-end solution integrating surveillance camera data with RL for TSC, leveraging real-time video input instead of relying on pre-processed features.
- We introduce a feature extraction framework that enables video-based traffic signal control to synchronize with existing research approaches, helping theoretical research transition into real-world applications.
- We design a robust Entropy Attention Mechanism to assess the uncertainty of dynamic traffic flow change, enhancing turn-based traffic signal control.
- We conduct extensive simulation efforts in highly realistic 3D-rendered environments, providing a more accurate reflection of real-world scenarios.

2 Related Work

While RL methods have been effective for optimizing traffic signals (Wei et al., 2021; Xu et al., 2023), current approaches focus primarily on theoretical simulation without practical consideration of real-world sensor inputs like surveillance cameras. Our VisionLight model extends multi-agent RL concepts by incorporating real-time video data, offering more adaptive and practical traffic control solutions (Huang et al., 2021). Multi-agent RL has demonstrated great promise in improving urban traffic flow, and video input enhances decision-making capabilities further (Liu et al., 2023).

The integration of video surveillance into traffic management systems has been shown to enhance vehicle detection and traffic flow predictions through deep learning techniques applied to surveillance images (Dilshad et al., 2020; Hu et al., 2021). Research has focused on detecting vehicle density from video data for real-time signal optimization (Jamebozorg & Hami, 2024). Additionally, sensor fusion combining video cameras with LiDAR improves vehicle localization, though the high cost of LiDAR limits its scalability (Liu et al., 2023). Our research leverages the widespread deployment of traffic cameras to bridge the gap between theoretical solutions and practical implementation in signal control (Luo et al., 2018).

Real-time vehicle detection in challenging conditions such as fog or low light has benefited from models like YOLO, which is widely adopted for traffic applications (Wang et al., 2022; Meng et al., 2023). Incorporating such deep learning models into traffic systems improves detection accuracy and signal timing adjustments (Patel & Ganatra, 2023; Meng et al., 2023). Multi-stream temporal structures have further enhanced congestion detection from video, directly supporting traffic control strategies (He et al., 2023).

Our research addresses the limitations of existing RL-based traffic control systems, which often lack real-world applicability, by integrating video data to provide a scalable and intelligent solution for practical intersection management (He et al., 2023).

3 Preliminaries

3.1 Traffic Intersection Description

At a typical four-legged intersection, each incoming direction has two lanes: one for left turn exclusive and one for straight & right turns (Papageorgiou et al., 2003). These lanes are grouped into lane sets, which are activated during the same signal phase when there are no conflicting movements. The incoming and outgoing lanes are defined as:

$$L_{in} = \{l_W^l, l_W^{s/r}, l_E^l, l_E^{s/r}, l_N^l, l_N^{s/r}, l_S^l, l_S^{s/r}\}, \quad L_{out} = \{l'_W, l'_E, l'_N, l'_S\}$$

where l_W^l represents the west incoming left-turn lane and $l_W^{s/r}$ represents the west incoming straight/right-turn lane. Traffic movements are defined as $(l_i^{\text{type}}, l'_j)$, grouping non-conflicting lane sets for signal timing adjustments. Intersection setting as shown in Figure 2.

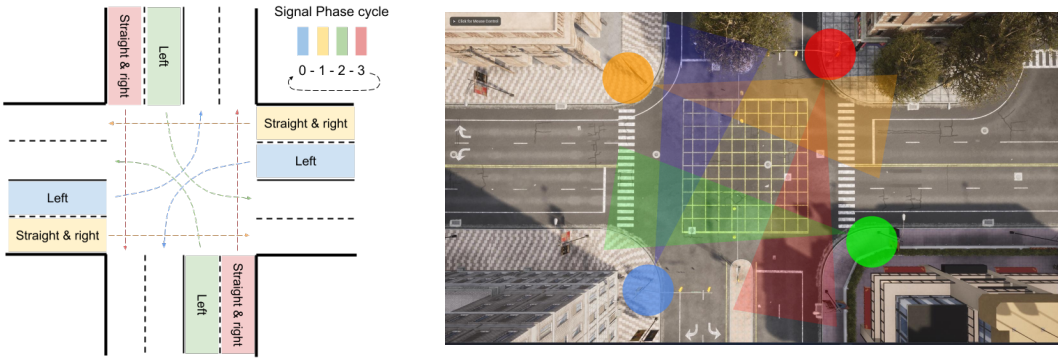


Figure 2: (a) Traffic intersection signal phases and cycling. (b) Camera placement and its perspective coverage.

3.2 Signal Phases and Action Space

The intersection operates with four distinct signal phases, each controlling traffic from different directions and movements (Chen et al., 2015):

Phase 0 :	W-E left-turn protection	Phase 1 :	W-E straight/right-turn
Phase 2 :	N-S left-turn protection	Phase 3 :	N-S straight/right-turn

In each phase, non-conflicting lane sets are activated, allowing traffic to flow from specific lanes. The phase activation is represented as:

$$p_k = \{(l_i^{\text{type}}, l'_j) \mid a(l_i^{\text{type}}, l'_j) = 1\}$$

where p_k is the active phase, and $a(l_i^{\text{type}}, l'_j) = 1$ indicates that the signal is green for the movement from incoming lane l_i^{type} to outgoing lane l'_j.

For example, during Phase 0, both l_W^l (west left-turn lane) and l_E^l (east left-turn lane) may have green lights, while opposing movements are stopped to prevent conflicts. Signal phase rotation shown in Figure 2.

In the simulation, our model makes a decision every 5 seconds, and the **ACTION SPACE** at each decision point consists of:

- *Retain*: Continue with the current phase p_k for an additional 5 seconds.
- *Switch*: Transition to the next phase p_{k+1}, with a 3-second yellow light followed by a 5-second green light.

The action space A is defined as:

$$A = \{a_{\text{retain}}, a_{\text{switch}}\}$$

where a_{retain} maintains the current phase, and a_{switch} initiates the transition to the next phase. This action definition ensures flexibility while simplifying decision-making. (Salah Bouktif, 2021).

The camera and its perspective coverage have been shown in the Figure 2. The four dots indicate the camera locations, while the triangles represent their perspective and coverage:

3.3 Markov Decision Process (MDP) Formulation

We model the traffic signal control problem as a Markov Decision Process (MDP), consistent with prior work (Puterman, 1990; Wang et al., 2023). The full MDP specification, including state, action, transition, and reward definitions, is provided in Appendix A.

4 METHODOLOGY

4.1 INPUT

The input to our solution consists of images captured at three time steps: $t-30$, $t-15$, and t_{now}, where t_{now} denotes the current time. At each step, four directional images are included, corresponding to the north (N), south (S), west (W), and east (E) views of the intersection. Thus, a total of 12 images are fed into the model as input, capturing the real-time and past traffic conditions from surveillance cameras positioned at the intersection. The input images are visualized in Figure 3.

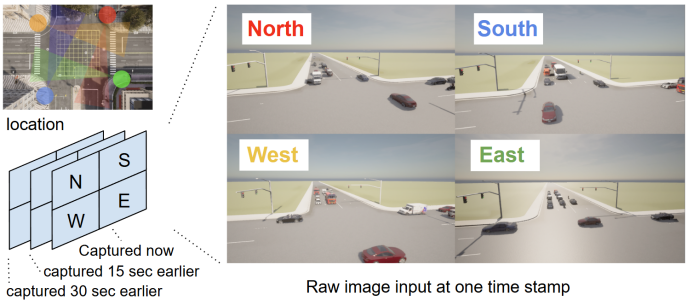


Figure 3: Input. Traffic images from four directions (N, S, W, E) captured at three time steps ($t-30$, $t-15$, t_{now}).

This temporal sequence of images provides the model with both current and past traffic states, enabling it to learn traffic dynamics over time. These inputs, combined with temporal features, are processed by the model to inform the decision-making process.

4.2 END-TO-END SOLUTION

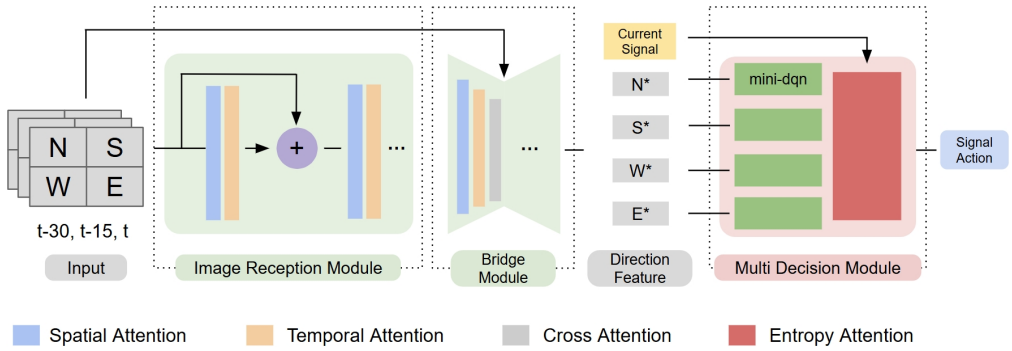


Figure 4: Overall End-to-End VisionLight Model Structure.

As shown in Figure 4, the End-to-End VisionLight model integrates three main components:

- **Image Processing Module**: This component extracts high-level semantic features from the input images, transforming them into a lower-dimensional feature space. The module is designed to be modular, allowing flexibility to replace it with various image processing techniques as required.

- **Feature Space Mapping (Bridge)**: This part of the model serves as a bridge, aligning and mapping the extracted image features with additional traffic metrics, such as vehicle density and queue lengths, into a unified decision-making space. This mapping forms a trainable latent space that aligns heterogeneous inputs, enabling reinforcement learning to operate on a unified representation.

- **Multi-Decision Agent**: Each agent focuses on one lane set (direction), processing its features through an entropy mechanism. The entropy module incorporates both phase timing (phase timing distance, i.e., the time until the next or subsequent signal phases) and state uncertainty. The agent optimizes a policy $\pi(s_t)$, balancing rewards such as minimizing queue lengths and reducing vehicle stopping times.

This modular and flexible architecture enables the system to make informed, real-time traffic control decisions based on image data and traffic metrics. The end-to-end solution algorithm is shown in Figure 5 (left).

Algorithm 1: VisionLight: End-to-End

Input: Image frames from past and current time; current signal phase
Output: Updated traffic phase

Initialize system and signal phase
while running every 5 seconds **do**
    Capture latest image
    Prepare input stack with recent frames and phase
    Send raw image stack to model for action
    if switch **then**
        Show yellow light, wait, then change to next green phase
    else
        Keep current phase
    end if
    Apply phase to simulation
    Collect current reward
    Combine with past rewards (weighted)
    Update image history
end while

Algorithm 2: VisionLight: Feature Extraction

Input: Image frames from past and current time; current signal phase
Output: Updated traffic phase

Initialize system and signal phase
while running every 5 seconds **do**
    Capture latest image
    Prepare input stack with recent frames and phase
    Extract traffic features from image stack
    Send features to FE model for action
    if switch **then**
        Show yellow light, wait, then change to next green phase
    else
        Keep current phase
    end if
    Apply phase to simulation
    Collect current reward
    Combine with past rewards (weighted)
    Update image history
end while

Figure 5: VisionLight solution algorithms.

4.3 Feature Extraction Solution

We noticed that most traffic signal control (TSC) models utilize input features such as the number of waiting vehicles, waiting times, or traffic pressure for each lane. To explore the compatibility with existing TSC reinforcement learning (RL) models, we extract and prepare these features, enabling fast and seamless deployment in real-world scenarios.

With the help of state-of-the-art object detection models such as YOLO (Jocher et al., 2023), and EfficientDet (Tan & Le, 2020), we build an image preprocessing system that detects the number of vehicles on each lane and calculates traffic pressure. These features are extracted from the time-series image input described above (i.e., images at $t - 30$, $t - 15$, and t_{now}), ensuring that short-term temporal dynamics are captured during processing. The traffic pressure extraction system is illustrated in Figure 6, which shows how traffic pressure is estimated for use in the reinforcement learning model.

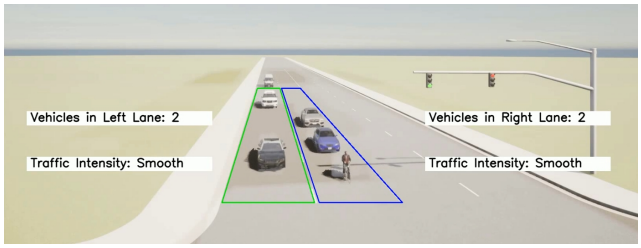


Figure 6: Traffic Pressure Detection System on predefined lane sets.

Additionally, we develop our own feature extraction model *VisionLight-FE* based on deep Q-networks (DQN) to further enhance feature representation and adapt to dynamic traffic environments. This model processes raw image data, extracts meaningful traffic-related features, and refines them for integration with RL-based decision-making systems. The feature extraction solution algorithm is shown in Figure 5 (right).

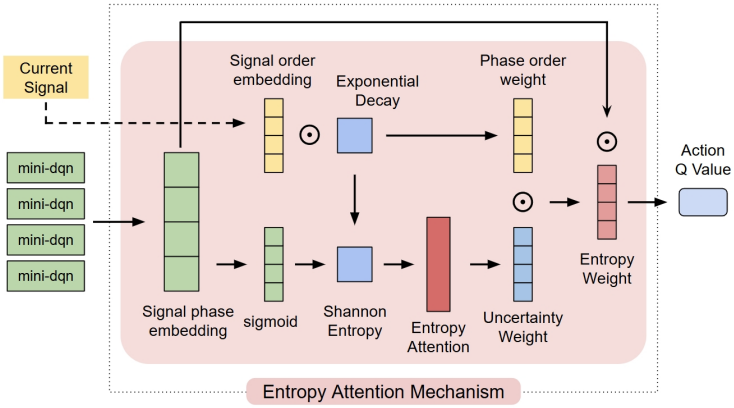


Figure 7: Entropy Attention Mechanism: Integrating phase timing order, entropy, and uncertainty weight to compute entropy-adjusted weights for each decision.

4.4 Image Reception Module

The image reception module processes traffic images captured at three timestamps: $t_{\text{-30s}}$, $t_{\text{-15s}}$, and t_{now}, across four directions (N, S, W, E). To extract useful representations from these images, we explore three neural architectures, corresponding to three VisionLight variants:

VisionLight-R uses a ResNet with spatial-temporal attention; **VisionLight-A** adopts anchor-based feature encoding inspired by YOLO; **VisionLight-T** applies a Transformer-based design for long-range traffic pattern modeling.

More details are provided in Appendix B.

4.5 Bridge Layer Module

The Bridge Layer maps high-dimensional visual features to the RL decision space, enabling communication between perception and control. It combines non-linear projection and cross-attention to capture spatial-temporal dependencies. Details are in Appendix C.

4.6 Multi-decision Module & Entropy Attention Mechanism

The Multi Mini Decision Agent Module consists of four mini-agents (Zhang et al., 2021), each responsible for one lane set (North-South left, North-South straight/right, West-East left, West-East straight/right), processing features and output a direction oriented features.

In our framework, four mini-agents generate Q-value features, each prioritizing decisions based on their focused traffic direction. However, effective decision-making should not only optimize for the current traffic state but also ensure stability over a future period. Given the rapid fluctuations in traffic flow, it is crucial to determine which agent's Q-value output is the most reliable and holds greater authenticity.

To address this, we propose a **trainable Entropy Attention Mechanism** that dynamically assesses uncertainty based on the input features from all four mini-agents. Entropy Attention Structure shown in figure 7. This mechanism assigns entropy-based weights to each agent's output, reducing the influence of decisions derived from highly volatile traffic conditions that are more likely to change. By emphasizing decisions from more stable directions, our approach ensures that the final aggregated Q-value remains both **robust for the present** and **adaptive for future traffic conditions**, leading to more reliable long-term signal control.

Each mini-agent calculates an action score $S_i = \sigma(W_d F_i + b_d)$, where σ is the sigmoid function, W_d and b_d are the weight matrix and bias, and F_i represents the feature set for a given lane. The

scores S_1, S_2, S_3, S_4 are then normalized into probabilities:

$$P_i = \frac{S_i}{\sum_{j=1}^{4} S_j}$$

Shannon entropy (Lin, 1991) $H = -\sum_{i=1}^{4} P_i \log(\max(P_i, \epsilon))$, is used to quantify decision uncertainty and adjust the decay constant $k' = k \cdot H$. The weights $w_i = e^{-k' d_i}$ are then computed based on phase timing distance d_i. These integrated weights are multiplied by the positional embedding (Vaswani et al., 2017; Li et al., 2023c) of the traffic signal phase sequence delay, ensuring that phase timing is incorporated into the decision-making process for more robustness:

$$w_i' = w_i \cdot \alpha_{\text{cr}}(S_i)$$

Finally, the weighted scores are used to compute the Q-values for two possible actions, with the action corresponding to the larger Q-value being selected:

$$Q_{\text{retain}}, Q_{\text{switch}} = f_{\text{Q}}(w_i' \cdot S_i)$$

The Entropy Attention Mechanism boosts performance, with its effectiveness proven by the ablation study results.

5 Experiment

5.1 Metrics

We evaluated the VisionLight model using two key metrics (Kim et al., 2023; Ault & Sharon, 2021): average stopping time (AST) and average queue length (AQL) (Akçelik, 1980).

AST is defined as:

$$\text{AST} = \frac{1}{T} \sum_{i=1}^{N} t_{\text{stop}}(S_i), \quad \text{AQL} = \frac{1}{T} \sum_{i=1}^{N} q(S_i)$$

where $t_{\text{stop}}(S_i)$ and $q(S_i)$ represent the stopping time and queue length for lane set S_i, respectively, and T is the flow duration in minutes. To maintain balance, the metrics were weighted equally, considering variations in vehicle spawning probabilities.

5.2 Traffic Flows

These metrics were measured over a single run across diverse traffic flow settings to ensure the model's robustness. The traffic flow settings, detailed in the Appendix F, provide a comprehensive view of traffic dynamics under light, heavy, balanced, and unbalanced conditions.

5.3 Training Strategy

The training strategy consists of optimizing the model using a dueling network, pre-training the multi-decision module on SUMO, and finally fine-tuning the entire system on Carla. For a detailed explanation, see Appendix D.

5.4 Baselines

Since no direct comparisons exist for our end-to-end video-to-signal control model, we evaluate against two categories of baselines: (1) fixed-time signal strategies and (2) reinforcement learning (RL) models that rely on structured feature inputs. All baselines were tested under diverse weather conditions (sunny, foggy, rainy, and night) to assess generalization without specific training for those scenarios (Figure 8).

Fixed-Time Methods. *Fixed-Time 30 & 40*: Pre-defined signal cycles with fixed 30s or 40s durations, independent of traffic flow.

RL Models with Feature Inputs. We further compare with established RL-based traffic signal control methods, including *MPLight*, *AttendLight*, *CoLight (SOTA)*, and *PressLight*. Each of these models is evaluated with two suffixes to denote the input source:

- **-SUMO**: The model directly consumes simulator-provided features such as vehicle positions and queue lengths and runs in SUMO environment.
- **-FE**: The model relies on image-based preprocessing, where features such as vehicle counts and traffic pressure are extracted from raw camera images.

Detailed descriptions of the models are as follows:

- *MPLight-SUMO / MPLight-FE*: Utilizes traffic pressure as both input and reward, incorporating the FRAP network to handle unbalanced traffic conditions.
- *AttendLight-SUMO / AttendLight-FE*: Integrates an attention mechanism to extract key features from observations and predict phase transitions, enhancing decision-making efficiency.
- *CoLight-SUMO / CoLight-FE (SOTA)*: A decentralized RL model using a graph attention network (GAT) to enable communication between adjacent intersections, improving coordination.
- *PressLight-SUMO / PressLight-FE*: A DRL model inspired by MaxPressure, optimizing intersection pressure by strategically managing vehicle flow.

For the model specs and latency see Appendix G.

5.5 RESULTS

Sunny **Rainy** **Foggy** **Night**

Figure 8: Camera shots from different weather conditions.

Table 1: Performance under different weather, **Best** in bold, Second-best underlined).
METRIC - AR: Average Reward (/min), AST: Average Stopping Time (seconds/min), AQL: Average Queue Length (vehicles/min).
SUFFIX - SUMO: simulation-provided features; FE: image feature extraction; R, A, T: end-to-end variant structures.

Models	Sunny			Rainy			Foggy			Night		
	AR ↑	AST ↓	AQL ↓	AR ↑	AST ↓	AQL ↓	AR ↑	AST ↓	AQL ↓	AR ↑	AST ↓	AQL ↓
Traditional Method												
Fixed-time 30	-59.03	1790.99	57.82	-63.10	**1852.76**	**62.31**	**-58.22**	**1765.80**	**55.73**	-61.41	1820.50	60.11
Fixed-time 40	-93.00	3007.91	73.84	-95.76	3150.33	78.90	-88.53	2920.50	70.12	-91.89	3055.62	75.45
Feature Input Method												
MPLight-SUMO	-15.31	259.23	29.65	-	-	-	-	-	-	-	-	-
AttendLight-SUMO	-13.43	264.28	26.49	-	-	-	-	-	-	-	-	-
CoLight-SUMO	**-12.37**	243.81	28.92	-	-	-	-	-	-	-	-	-
PressLight-SUMO	-13.96	**235.42**	30.35	-	-	-	-	-	-	-	-	-
Feature Extraction Method												
MPLight-FE	-29.36	650.79	52.80	-112.65	3230.28	88.04	-108.51	3969.53	76.45	-49.97	1867.26	**44.83**
AttendLight-FE	-28.90	725.68	59.89	-107.94	3469.61	87.23	-115.65	3797.51	82.61	-51.86	1856.13	51.88
PressLight-FE	-27.78	645.42	53.67	-108.64	3524.34	77.95	-115.16	4023.66	77.46	-54.86	1554.58	49.18
CoLight-FE(SOTA)	-27.15	655.98	51.33	-103.30	3319.79	82.71	-110.67	4083.33	80.97	-45.36	1935.80	48.46
VisionLight-FE(ours)	-28.19	633.06	56.74	-104.94	3498.96	83.20	-108.44	3917.51	81.27	-49.42	1727.92	47.51
End-to-End Method												
VisionLight-R(ours)	-13.76	267.72	27.94	-65.57	2213.90	237.05	-86.74	2776.76	262.08	**-38.97**	1450.76	117.40
VisionLight-A(ours)	-12.46	238.04	**25.72**	-70.49	2182.21	223.00	-82.97	2784.09	281.73	-40.55	**1364.12**	115.73
VisionLight-T(ours)	-15.23	316.52	34.73	**-59.36**	2319.12	256.04	-87.03	2636.48	253.94	-43.08	1405.84	105.85

From Table 1 we know:

SUNNY conditions, all three VisionLight end-to-end variants performed on par with direct feature-input methods, with VisionLight-A (ours) achieving the highest average score across three metrics. In contrast, feature-extraction methods, which rely on object detection for vehicle counts and traffic pressure estimation, introduced errors that reduced decision-making accuracy. Even with these errors, all models still outperformed traditional fixed-time control. Notably, our customized VisionLight-FE matched the best feature-input baselines, confirming its effectiveness.

EXTREME WEATHER (rain, fog, and night), all three VisionLight end-to-end variants remained competitive, performing only slightly worse than the fixed 30-second signal strategy. VisionLight-A and VisionLight-T (ours) achieved the second- and third-best average performance, even under reduced visibility. In contrast, feature-extraction methods dropped sharply, as blurred inputs impaired vehicle counts and pressure estimation, leading to incorrect decisions. Interestingly, fixed-time control outperformed RL methods in this setting, since traffic lights remain visible to drivers regardless of weather. At night, however, VisionLight's end-to-end models maintained strong performance, surpassing fixed-time control and matching the best feature-input baselines. Direct feature-input methods tested on SUMO were excluded, as SUMO does not simulate weather.

OVERALL, VisionLight scored the highest average performance across all scenarios, demonstrating strong adaptability with raw video input. It achieved results comparable to direct feature-input methods while offering a practical solution for real-world traffic signal control. For the cumulative reward over training epochs, see Figure 11.

5.6 Ablation Study: Impact of Entropy Attention Mechanism

To demonstrate the effectiveness of the Entropy Attention mechanism in capturing dynamic flow changes and stabilizing training, we evaluate its impact on all VisionLight variants, as shown in Table 2. Across all four models, adding entropy consistently improves the final average reward, with an average gain of 46.2% over their counterparts without entropy. Models with entropy also converge faster during training, indicating improved learning stability. Detailed reward progression over training epochs is provided in Appendix E.

Model	With Entropy	Without Entropy
VisionLight-FE	**-28.19**	-58.61
VisionLight-R	**-13.76**	-22.59
VisionLight-A	**-12.46**	-26.33
VisionLight-T	**-15.23**	-25.86

Table 2: **Average Reward** of VisionLight variants with and without the Entropy Attention mechanism.

6 Conclusion

To address gaps in traffic signal control (TSC), where most RL methods rely on simulation-only features such as vehicle coordinates and waiting times, we presented **VisionLight**, a video-driven RL framework for dynamic TSC management. VisionLight achieved the highest average performance across all scenarios, surpassing state-of-the-art (SOTA) baselines while remaining practical for real-world deployment. We also introduced a multi-agent architecture and an Entropy Attention Mechanism to handle dynamic flow changes and stabilize training. Ablation studies confirmed its effectiveness, showing a significant reward improvement over models without entropy.

VisionLight further demonstrated robustness under extreme weather conditions such as rain, fog, and night, maintaining strong performance and in several cases outperforming SOTA approaches without retraining.

For future work, we will extend VisionLight to multi-intersection coordination and pedestrian-aware integration during rush hours, further improving the robustness and resilience of video-based TSC systems (Figure 9).

Figure 9: Impact of rush hour, pedestrians, and multi-intersection.

REFERENCES

Rahmi Akçelik. Time-dependent expressions for delay, stop rate and queue length at traffic signals. Technical Report AIR 367-1, Australian Road Research Board, Victoria, Australia, 1980.

Ao Wang, Hui Chen, Lihao Liu, et al. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*, 2024.

James Ault and Guni Sharon. Reinforcement learning benchmarks for traffic signal control. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/f0935e4cd5920aa6c7c996a5ee53a70f-Abstract-round1.html.

Weiwei Cai and Zhanguo Wei. Remote sensing image classification based on a cross-attention mechanism and graph convolution. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2020.

Li Chen, Cynthia Chen, and Reid Ewing. Left-turn phase: Permissive, protected, or both? a quasi-experimental design in new york city. *Accident Analysis & Prevention*, 76:102–109, 2015.

Ruixiang Cheng, Zhihao Qiao, Jiarui Li, and Jiejun Huang. Traffic signal timing optimization model based on video surveillance data and snake optimization algorithm. *Sensors*, 23(11):5157, 2023.

Gurcan Comert and Mecit Cetin. Queue length estimation from connected vehicles with range measurement sensors at traffic signals. *Applied Mathematical Modelling*, 99:418–434, 2021.

Naqqash Dilshad, JaeYoung Hwang, JaeSeung Song, and NakMyoung Sung. Applications and challenges in video surveillance via drone: A brief survey. In *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 728–732. IEEE, 2020.

A. P. Fonseca and R. C. Garcia. Deep reinforcement learning model to mitigate congestion in real-time traffic light networks. *Infrastructures*, 6(10):138, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 770–778, 2016.

L. He, Q. Yu, and Y. Chen. Reinforcement learning for traffic signal control under missing data. In *Proceedings of the 2023 IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2023.

Liang Hu, L Wang, Z Zhou, Z Sheng, and Y Zhang. Deep learning-based video surveillance for predicting vehicle density in real-time scenario. *Journal of Ambient Intelligence and Humanized Computing*, 2021.

Jian Huang, Wei Wang, Lirong Wang, Hui Chen, Qingshan Deng, Heying Fan, and Yaohua Yu. Application of deep reinforcement learning in optimization of traffic signal control. In *2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, pp. 1958–1964. IEEE, 2021.

Mahdi Jamebozorg and Mohsen Hami. Traffic control using intelligent timing of traffic lights with reinforcement learning technique and real-time processing of surveillance camera images. In *6th International Conference on Traffic Management and Safety*, 2024.

Glenn Jocher et al. Yolov8: Real-time object detection and segmentation. *Ultralytics Repository*, 2023. URL https://github.com/ultralytics/ultralytics.

Jangmin Kim et al. Effects of reward functions on reinforcement learning for traffic signal control. *PLOS ONE*, 18(6):e0278473, 2023. doi: 10.1371/journal.pone.0278473. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0278473.

Songsang Koh, Bo Zhou, Hui Fang, Po Yang, Zaili Yang, Qiang Yang, Lin Guan, and Zhigang Ji. Real-time deep reinforcement learning based vehicle navigation. *Applied Soft Computing*, 96: 106694, 2020.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Wei Li et al. Weighted mean-field multi-agent reinforcement learning via reward attribution decomposition. *SpringerLink*, 34:111–134, 2023a.

X. Li, X. Du, and Z. Han. Traffic signal control using deep reinforcement learning in a connected vehicle environment. *IEEE Transactions on Vehicular Technology*, 72(5):1243–1256, 2023b.

Yue Li et al. The impact of positional encoding on length generalization in transformers. *arXiv preprint arXiv:2305.19466*, 2023c.

Zhenning Li, Jie Zhang, and Yan Liu. A deep reinforcement learning approach for traffic signal control optimization. *arXiv preprint arXiv:2107.06115*, 2021.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

C Liu, H Zhang, and G Zheng. Scalable multi-agent reinforcement learning for traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using sumo. In *The 21st IEEE International Conference on Intelligent Transportation Systems*. IEEE, 2018. URL https://elib.dlr.de/124092/.

Z Luo, PM Jodoin, SZ Su, SZ Li, and H Larochelle. Traffic analytics with low-frame-rate videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(4):878–891, 2018.

X Meng, Y Liu, L Fan, and J Fan. Yolov5s-fog: an improved model based on yolov5s for object detection in foggy weather scenarios. *Sensors*, 23(11):5321, 2023.

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. Awac: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

Mohammad Noaeen, Atharva Naik, Liana Goodman, Jared Crebo, Taimoor Abrar, Zahra Shakeri Hossein Abad, Ana LC Bazzan, and Behrouz Far. Reinforcement learning in urban network traffic signal control: A systematic literature review. *Expert Systems with Applications*, 199: 116830, 2022.

Markos Papageorgiou, Christina Diakaki, Vaya Dinopoulou, Apostolos Kotsialos, and Yibing Wang. Review of road traffic control strategies. *Proceedings of the IEEE*, 91(12):2043–2067, 2003.

P Patel and A Ganatra. Improving traffic surveillance with deep learning powered vehicle detection, identification, and recognition. *SpringerLink*, 2023.

Jan Peters, Hirotaka Hachiya, and Masashi Sugiyama. Reward-weighted regression converges to a global optimum. *arXiv preprint arXiv:1010.1362*, 2010.

Martin L Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.

Ali Ouni Salah Bouktif, Abderraouf Cheniki. Traffic signal control using hybrid action space deep reinforcement learning. *Sensors*, 21(7):2302, 2021. doi: 10.3390/s21072302.

Mingxing Tan and Quoc Le. Efficientdet: Scalable and efficient object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10781–10790, 2020. URL https://arxiv.org/abs/1911.09070.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems (NeurIPS)*, pp. 5998–6008, 2017.

Xiaoyu Wang, Scott Sanner, and Baher Abdulhai. Sequence decision transformer for adaptive traffic signal control. *MDPI Systems and Control*, 2023. URL https://www.mdpi.com/2218-6581/13/2/55.

Zhangu Wang, Jun Zhan, Chunguang Duan, Xin Guan, Pingping Lu, and Kai Yang. A review of vehicle detection techniques for intelligent vehicles. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8):3811–3831, 2022.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, pp. 1995–2003. PMLR, 2016.

H. Wei, G. Zheng, H. Yao, and Z. Li. Intellilight: A reinforcement learning approach for intelligent traffic light control. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2496–2505, 2019.

Hua Wei, Guanjie Zheng, Vikash Gayah, and Zhenhui Li. Recent advances in reinforcement learning for traffic signal control: A survey of models and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(2):12–18, 2021.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020.

Y. Wu, Z. Liu, and W. Zhang. Deep reinforcement learning for intelligent traffic signal control in complex urban networks. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

K. Xu, G. Zheng, and Y. Yu. Multi-objective deep reinforcement learning for adaptive traffic signal control. *IEEE Transactions on Vehicular Technology*, 72(6):4567–4579, 2023.

Kaixiang Zhang, Zhaoran Yang, and Tamer Basar. Multi-agent deep reinforcement learning: A survey. *IEEE transactions on knowledge and data engineering*, 34(3):1112–1131, 2021.

APPENDIX

A MARKOV DECISION PROCESS FORMULATION

The traffic signal control problem is modeled as a Markov Decision Process (MDP), defined by the tuple (S, A, P, R, γ):

State space: s_t includes vehicle density, queue length, and current signal phase:

$$s_t = \left(\{x(l_i), q(l_i)\}_{i\in\{W,E,N,S\}}, p_k\right)$$

Action space: Retaining or switching the signal phase.

Transition function: $P(s_{t+1} \mid s_t, a)$ models system dynamics.

Reward function:

$$R(s_t, a) = -\left(\alpha \sum_i t_{\text{stop}}(l_i) + \beta \sum_i q(l_i)\right)$$

where α and β weigh stopping time and queue length.

The objective is to find the optimal policy:

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right]$$

B IMAGE RECEPTION AND FEATURE EXTRACTION

The reception module extracts structured traffic features from a sequence of images taken at $t_{\text{-30s}}$, $t_{\text{-15s}}$, and t_{now} across four intersection directions. These features are consumed by the RL decision module. We implement three backbone architectures for feature extraction, each forming a VisionLight variant:

- **VisionLight-R (ResNet-based)**: Applies spatial and temporal attention mechanisms (Vaswani et al., 2017; He et al., 2016) to enhance a standard ResNet. Spatial attention emphasizes salient regions:

$$\alpha_{\text{sp}}(p) = \frac{\exp(W_{\text{sp}}^T p + b_{\text{sp}})}{\sum_{p'} \exp(W_{\text{sp}}^T p' + b_{\text{sp}})}$$

$$\alpha_{\text{tmp}}(t_i) = \frac{\exp(U_{\text{tmp}}^T h_{t_i} + c_{\text{tmp}})}{\sum_{t_j} \exp(U_{\text{tmp}}^T h_{t_j} + c_{\text{tmp}})}$$

  Residual connections $R(x) = x + \text{Attention}(x)$ maintain core information.

- **VisionLight-A (Anchor-based)**: Inspired by the YOLO object detection pipeline (Ao Wang et al., 2024), this model adapts YOLO's image encoding layers to extract structured traffic patterns, such as vehicle count and lane occupancy.

- **VisionLight-T (Transformer-based)**: Based on ViT-Base-Patch16-224 (Wu et al., 2020), this model uses self-attention to capture long-range spatial and temporal dependencies, enabling robust understanding of traffic flow dynamics and congestion levels.

These three variants allow VisionLight to flexibly accommodate varying deployment needs, from lightweight inference to high-capacity vision modeling.

C BRIDGE LAYER MODULE

The Bridge Layer Module maps the enriched feature space from the Video Reception Module into actionable inputs for the SUMO pre-trained RL agent (Lopez et al., 2018). The concatenated feature map F, combining original input and attention-enhanced features, is transformed into a lower-dimensional space for decision-making via:

$$\tilde{F} = \phi(W_b F + b_b)$$

where ϕ is a non-linear activation, W_b the weight matrix, and b_b the bias vector. The output $\tilde{F}$ captures both spatial and temporal relationships.

A cross-attention mechanism α_{cr} (Cai & Wei, 2020) integrates spatial focus and temporal changes:

$$\alpha_{\text{cr}}(F_{\text{sp},i}, F_{\text{tmp},j}) = \frac{\exp(W_c^T(F_{\text{sp},i} \oplus F_{\text{tmp},j}))}{\sum_k \exp(W_c^T(F_{\text{sp},i} \oplus F_{\text{tmp},k}))}$$

where $F_{\text{sp},i}$ and $F_{\text{tmp},j}$ represent spatial and temporal features, respectively. This mechanism ensures effective integration of spatial and temporal dependencies. The final output is sent to the Multi-Decision Module for traffic signal control.

D TRAINING STRATEGY

D.1 DUELING NETWORK

The decision-making process in the Multi Mini Decision Agent Module uses a Dueling Network architecture (Wang et al., 2016). The Q-value $Q(s, a)$ is split into the value function $V(s)$, which estimates the reward of being in a state, and the advantage function $A(s, a)$, which measures the benefit of an action:

$$Q(s, a) = V(s) + \left(A(s, a) - \frac{1}{|A|} \sum_{a'} A(s, a') \right)$$

This separation improves learning by distinguishing the value of the state from the relative advantage of each action, stabilizing the decision-making process (Konda & Tsitsiklis, 1999).

The Dueling Network, combined with entropy-based attention and the multi-agent framework, enhances the system's ability to make intelligent, real-time traffic control decisions.

D.2 MULTI MINI DECISION AGENT PRE-TRAINING ON SUMO

The Multi Mini Decision Agent Module was pre-trained in the SUMO traffic simulation environment using non-image features such as vehicle counts, queue lengths, and cumulative stopping times. The objective was to establish decision-making capability in traffic signal control before introducing complex video inputs.

The intersection settings mirrored those in later stages, with 3 timestamps ($t - 30$, $t - 15$, and t_{now}) per lane set. Each timestamp captured cumulative stopping time $t_{\text{stop}}(l_i)$ and queue length $q(l_i)$ for lane set l_i. This simplified pre-training setup ensures faster convergence and explainability before integrating video-based inputs.

D.3 CARLA FINE-TUNING

Fine-tuning in the Carla simulation environment involved using real-time image inputs to capture dynamic traffic flow. Carla offers more realistic vehicle dynamics, making it crucial for testing in real-world intersection scenarios. The setup and data collection were similar to SUMO, with images taken at $t - 30$, $t - 15$, and t_{now}.

A key component is the weighted reward function (Li et al., 2023a; Peters et al., 2010):

$$R_{\text{weighted}} = 0.4 \cdot r_1 + 0.3 \cdot r_2 + 0.2 \cdot r_3 + 0.1 \cdot r_4$$

This reflects the cumulative impact of decisions, with rewards normalized across light ($p_{\text{light}} = 0.01$) and heavy ($p_{\text{heavy}} = 0.04$) traffic flows. Stopping time weight $\alpha = 0.1$ and queue length weight $\beta = 1$ were chosen to balance both metrics based on prior training.

Fine-tuning showed that the VisionLight model can handle complex, real-world scenarios by leveraging both image inputs and weighted rewards, optimizing signal control in adaptive and intelligent ways.

D.4 Hybrid Online-Offline Training Strategy

We use a hybrid online-offline strategy to improve training efficiency (Nair et al., 2020). Let $\mathcal{E}$ be the Carla environment and $\pi_\theta(a|s)$ the policy parameterized by θ.

1. Data Collection We interact with $\mathcal{E}$ to gather experiences $\mathcal{D}_0 = \{(s_t, a_t, r_t, s_{t+1})\}$, storing them in a buffer $\mathcal{B}$: $\mathcal{B} = \mathcal{D}_0$.

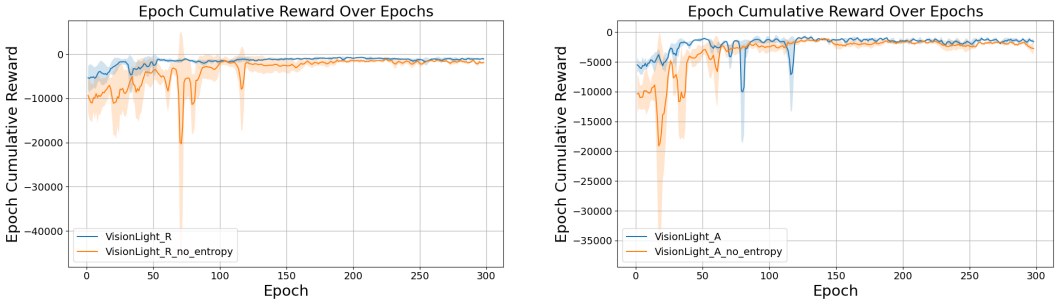
2. Offline Training In the offline phase, π_θ is updated using mini-batches of size 150 from $\mathcal{B}$. The policy parameters are adjusted as $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta)$, where $\eta = 0.0005$. We run 5 iterations per cycle, improving sample efficiency.

3. Periodic Online Updates After each offline cycle, new experiences $\mathcal{D}_n$ are collected from $\mathcal{E}$ and added to $\mathcal{B}$, replacing old data: $\mathcal{B} \leftarrow \mathcal{B} \cup \mathcal{D}_n$. Roughly 10-20% of the buffer is updated.

4. Iteration This process repeats: $\pi_\theta^{(n)} \xrightarrow{\text{Offline}} \pi_\theta^{(n+1)} \xrightarrow{\text{Online}} \mathcal{D}_{n+1} \rightarrow \mathcal{B}$.

Hyperparameters Key parameters include: buffer size = 2000, $\gamma = 0.97$, $\eta = 0.0005$, $\tau = 0.01$, batch size = 150, and ϵ annealing from 1 to 0.1 over 20,250 steps.

E Training Curves for Entropy Ablation



(a) VisionLight-R (b) VisionLight-A

Figure 10: Average reward over training epochs, with and without entropy attention.

F TRAFFIC FLOW SETTING

Time Period	North-South		East-West		South-North		West-East	
	S & R	Left	S & R	Left	S & R	Left	S & R	Left
0-6000s	600	600	600	600	600	600	600	600
250-750s	500	0	0	0	0	500	0	0
1000-1500s	0	0	500	0	0	0	0	500
1750-2250s	0	0	0	0	500	0	0	0
2500-3000s	0	0	0	0	0	0	500	0
3250-3750s	0	0	0	500	0	0	0	0
4000-4500s	0	0	0	0	0	500	0	0
4750-5250s	0	500	0	0	0	0	0	0
5500-6000s	0	0	0	0	0	0	0	0
6000-8000s	480	480	480	480	480	480	480	480
Total Vehicles	1580	1580	1580	1580	1580	1580	1580	1580
Avg Throughput (veh/hr)	360	360	360	360	360	360	360	360

Table 3: Avg. generated vehicles and throughput for each traffic flow.

G VISIONLIGHT MODEL SPECS

Model	Parameters (M)	GFLOPs	Latency (ms)
VisionLight-FE	7.5	12.1	45.6
MPLight-FE	6.8	11.5	42.3
AttendLight-FE	8.2	13.4	47.8
PressLight-FE	7.0	12.0	44.2
CoLight-FE	7.3	12.3	43.7
VisionLight-R	12.5	20.3	35.8
VisionLight-A	10.8	18.2	33.5
VisionLight-T	14.3	24.1	38.6

Table 4: Comparison of models based on parameters, FLOPs, and latency.

H TRAINING REWARD OVER EPOCHS

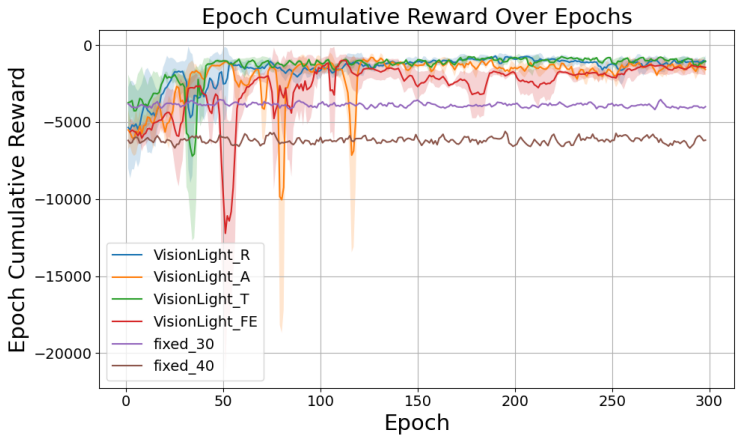


Figure 11: Cumulative Reward Comparison: VisionLights