SEATTLE, WA

CVPR
JUNE 17-21, 2024

# DMR: Decomposed Multi-Modality Representations for Frames and Events Fusion in Visual Reinforcement Learning

Haoran Xu[1,2], Peixi Peng[2,3]*, Guang Tan[1]*, Yuan Li[4], Xinhai Xu[4], Yonghong Tian[2,3,5]

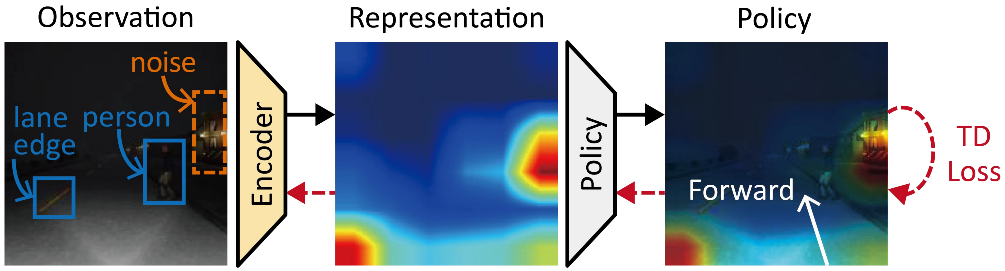
[1]School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-sen University
[2]Peng Cheng Laboratory
[3]School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University
[4]Academy of Military Sciences [5]School of Computer Science, Peking University

(* corresponding authors)

# Quick Preview

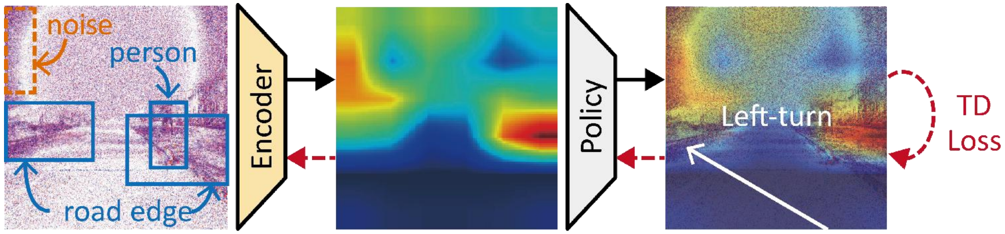
- We explore **visual reinforcement learning (RL) using two complementary visual modalities**: **frame-based RGB** camera and **event-based Dynamic Vision Sensor** (DVS).
- Existing multi-modality visual RL methods often encounter challenges in effectively extracting task-relevant information from multiple modalities while suppressing the increased noise, only using indirect reward signals instead of pixel-level supervision.
- To tackle this, we propose a Decomposed Multi-Modality Representation (DMR) framework for visual RL. It explicitly decomposes the inputs into **three distinct components**: **combined task-relevant features** (co-features), **RGB-specific noise**, and **DVS-specific noise**.
- Extensive experiments demonstrate that, by explicitly separating the different types of information, our approach achieves substantially improved policy performance compared to state-of-the-art approaches.

# Problem



(i)

**(i) RGB Frame**: insufficient ambient light causes RGB underexposure, leading to the overlooking of the front pedestrian and resulting in a forward policy aligned with the lane direction that could cause collisions.

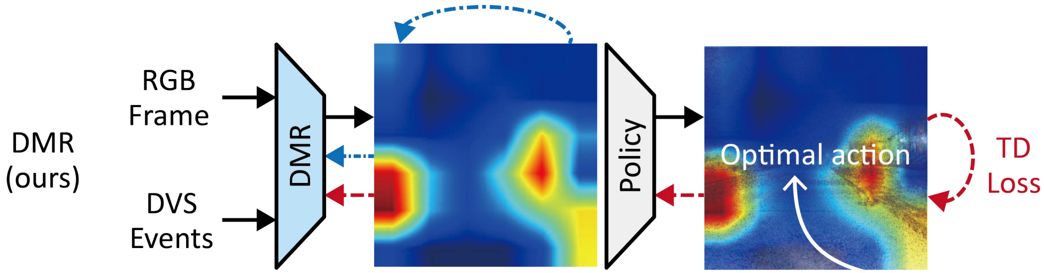


(ii)

**(ii) DVS Events**: the lack of texture in DVS causes the person and the background to blend, leading to a left-turn policy to avoid the highlighted area on the right.

The integration of frame- and event-based cameras has been explored for tasks like object detection and depth estimation. However, in vision-based RL, where entire observations are mapped to decisions **only using temporal-difference (TD) loss, without pixel-level or instance-level supervision**, simply aggregating frames and events can result in increased noise and task-irrelevant information. This phenomenon results in noise injection in the latent state space and leads to reduced RL performance.

# General solution



(i)

**(iii) Decomposed Multi-modality Representations (DMR)**: can fully take advantage of RGB and DVS to extract task-relevant information and eliminate task-irrelevant and noisy information through joint TD and DMR learning, thereby obtaining an optimal evasion policy.

We categorize the information from frames and events into **three distinct types**:

1) Combined task-relevant feature (referred to as *co-feature*);

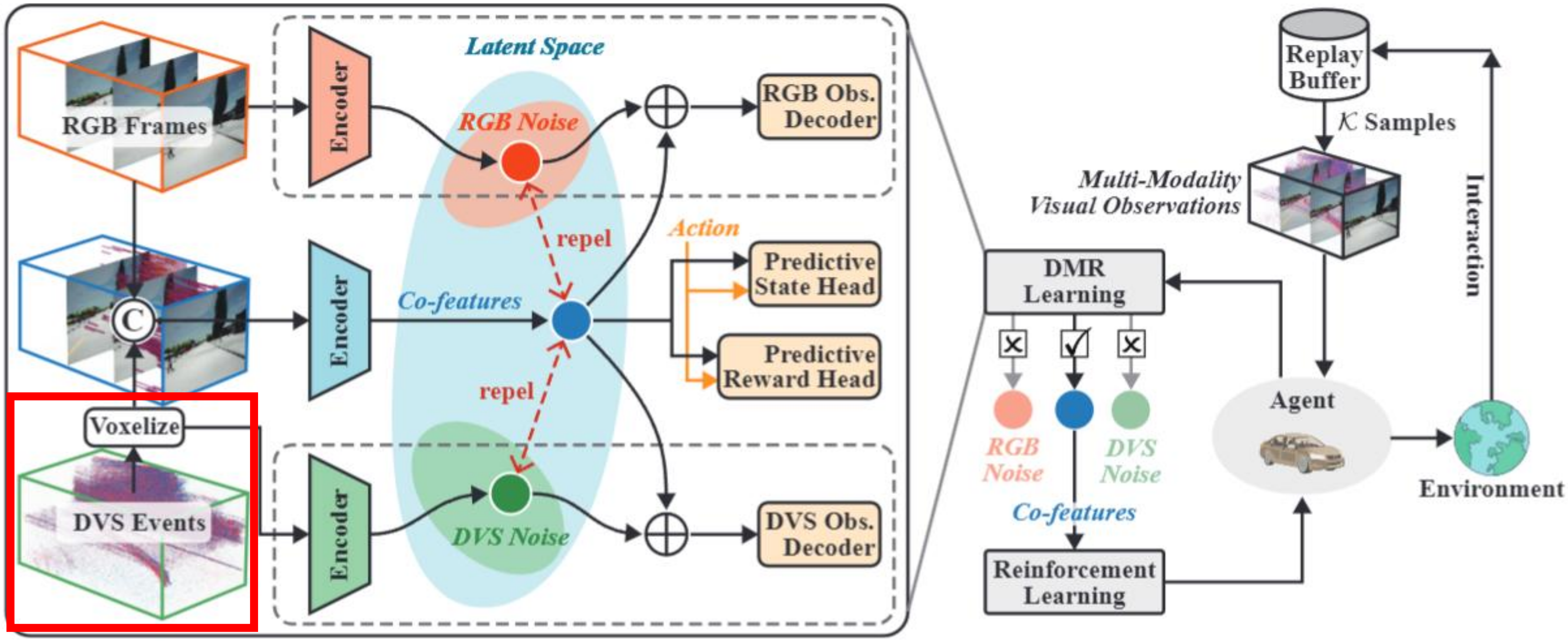
2) RGB-specific noise and task-irrelevant feature, or simply *RGB noise*;

3) DVS-specific noise and task-irrelevant feature, or *DVS noise*.

**The co-features represent the full information from both modalities that is essential for the RL task, while the noise represents unwanted information that may negatively impact the RL process.**

Combining frames and events helps to extract important regions, including the pedestrian and road edges (as shown in the above figure). These regions are difficult to identify precisely using either modality alone. It is notable that these three parts are all latent and only the rewards collected by interacting with environments are available as external guidance during learning, which is consistent with the standard RL pipeline.

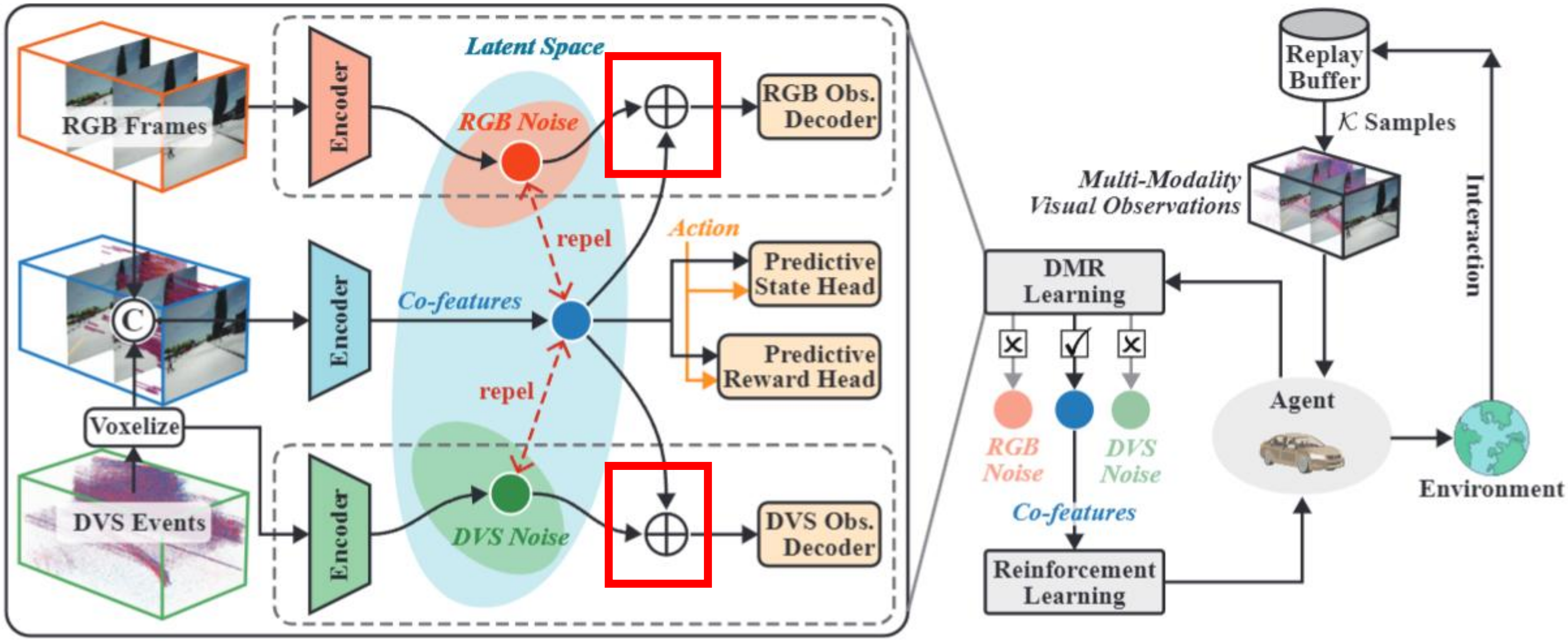
# Methodology



- We first **process asynchronous DVS events**. It is a common practice to convert events within a fixed-length temporal window into **a fixed-size tensor representation**, referred to as a voxel grid. To synchronize events with the low sampling rate of RGB frames, we partition the incoming events within the fixed time interval of RGB frames. The events occurring between the pair-wise frames are discretized into **a spatio-tempral voxel grid**. Each element in the voxel grid has three dimensions, two-dimensional location $(x_l, y_m)$, and temporal dimension $(t_n)$. Formally:

$$\mathcal{E}_t(x_l, y_m, t_n) = \sum_{x_l, y_m = x_i, y_i} p_i \max(1 - |t_n - t_i^*|)$$

# Methodology

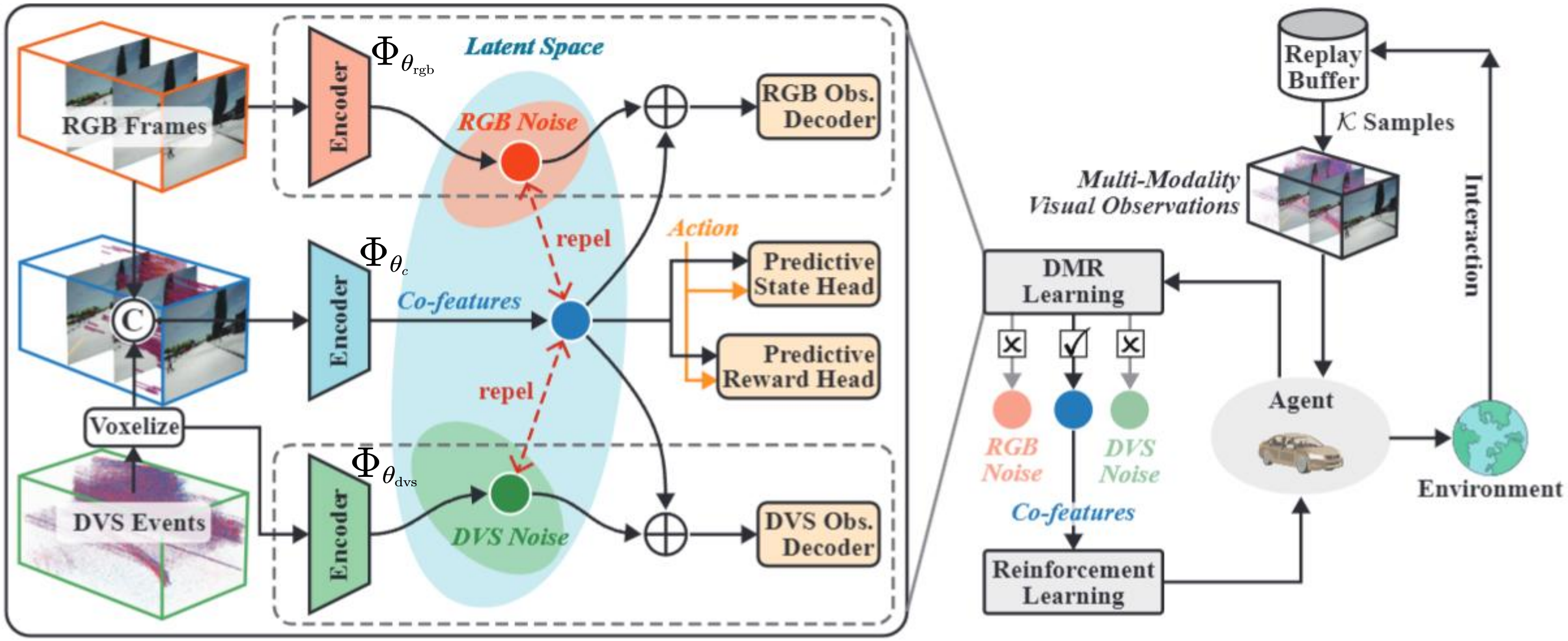


**Decomposition idea:**

- Let $z_t^i$ denote the representation for the original observation $o_t^i$ of modality $i \in \{\text{rgb}, \text{dvs}\}$.
- The representations $z_t^i$ may **differ significantly** for the two modalities even when they yield similar policies, because of the different working principles of RGB and DVS cameras.
- We **decompose** $z_t^i$ into co-features $z_t^c$ and modality-specific noises $h_t^i$ as

$$z_t^i = z_t^c \oplus h_t^i. \qquad (9)$$

# Methodology

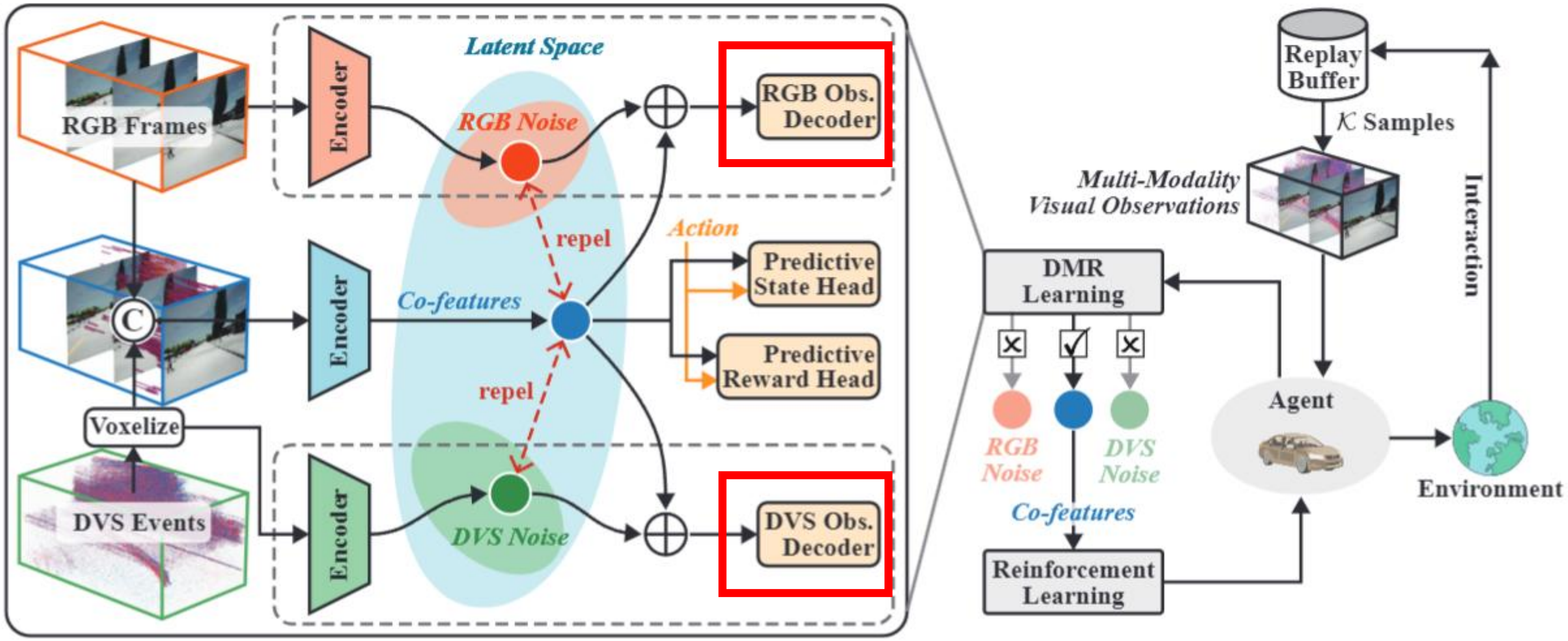


To achieve this, DMR comprises **three branches:**

- The **upper and lower branches** take RGB frames and DVS events as inputs, respectively. The data then pass through their respective encoders, denoted as $\Phi_{\theta_{\text{rgb}}}$ and $\Phi_{\theta_{\text{dvs}}}$ , to generate modality-specific noise ( $h_t^{\text{rgb}}$ , $h_t^{\text{dvs}}$).
- The **intermediate branch** takes the concatenation of RGB and DVS as input. Its output, co-features $z_t^c$, are generated by the intermediate encoder parameterized as $\Phi_{\theta_c}$ .

# Methodology

CVPR
SEATTLE, WA
JUNE 17-21, 2024

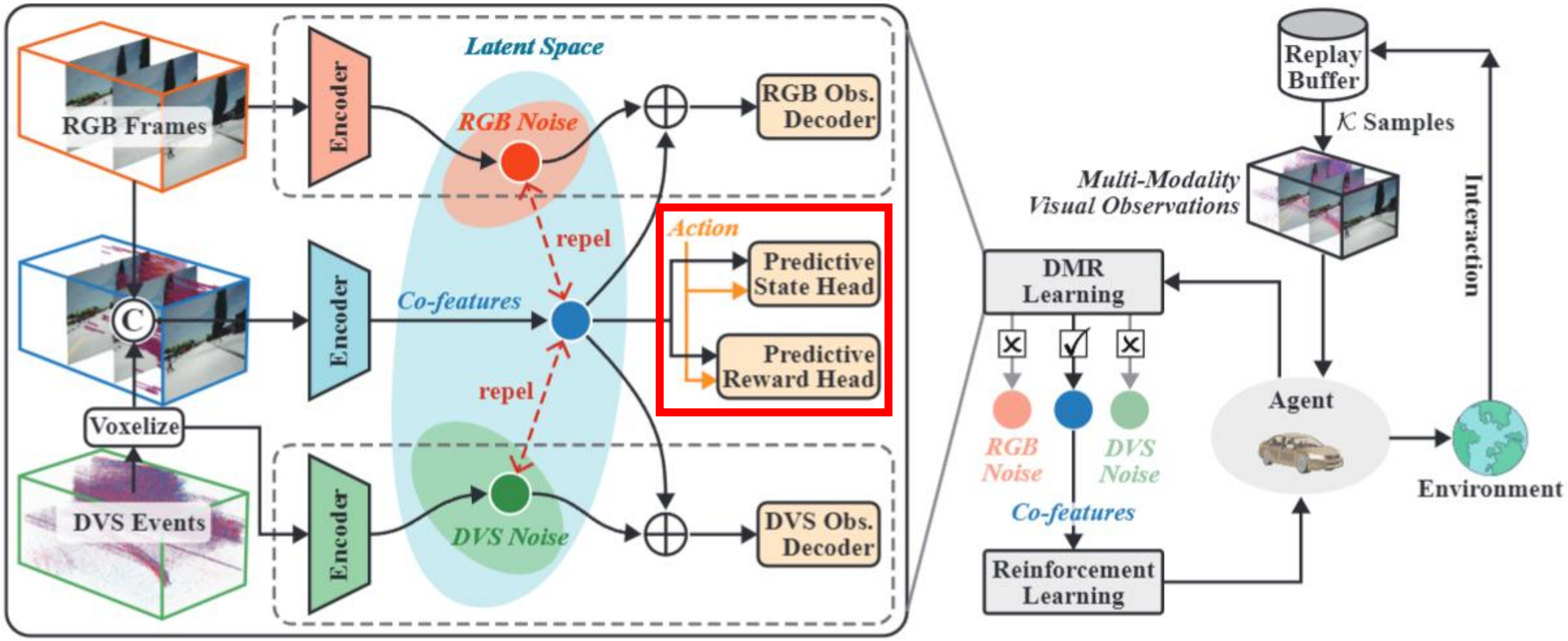


To ensure the **completeness of information**, we employ **reconstruction** decoders, denoted as $\mathcal{D}_{\theta_i}$ , to ensure that the respective original observations $o_t^i$ can be recovered:

$$\mathcal{L}_{\mathcal{D}} = \sum_{i \in \{\mathrm{rgb}, \mathrm{dvs}\}} \left\| \mathcal{D}_{\theta_i}(z_t^c + h_t^i) - o_t^i \right\|_2 , \qquad (10)$$

where $t \in \mathcal{K}$ and $\mathcal{K}$ is the set of sample indices in a training batch that are from different time steps in different MDPs.

# Methodology



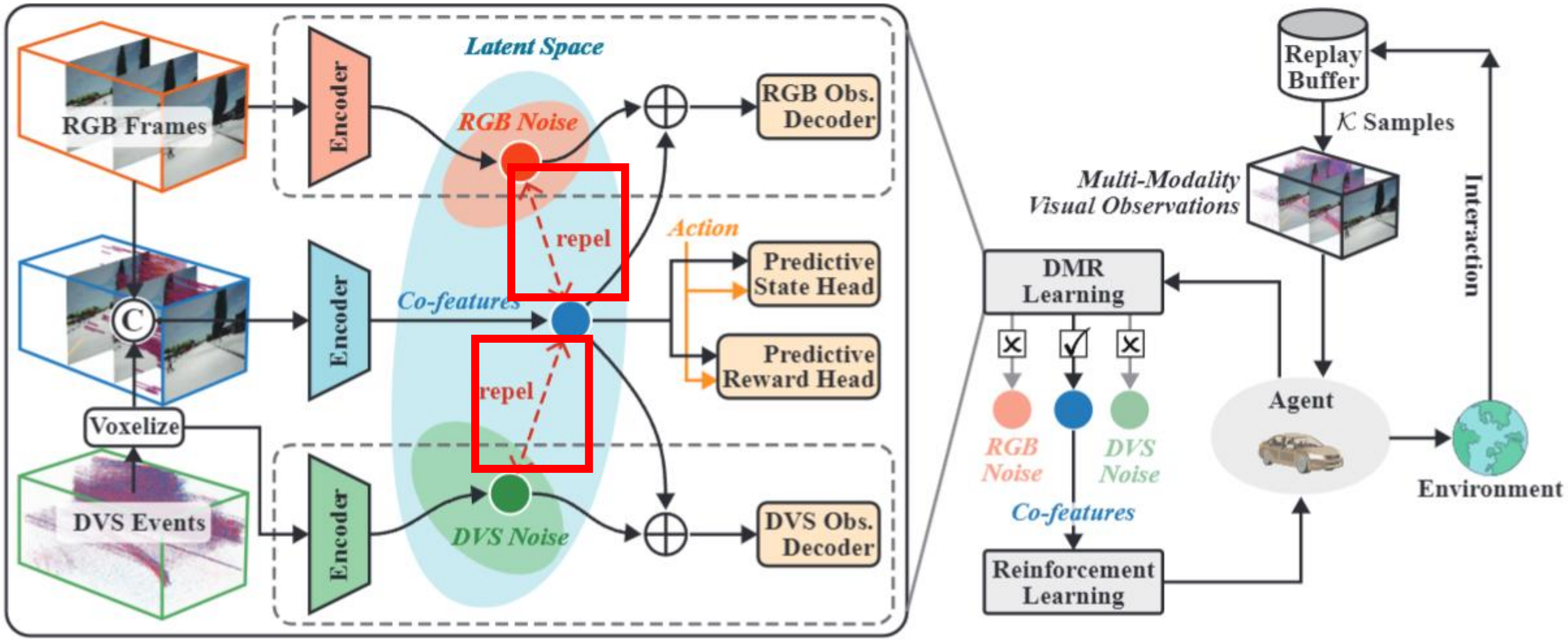
While ensuring the completeness of $z_t^i$, we utilize the task-relevant predictive heads to guide the extraction of the co-features $z_t^c$. Here, we incorporate the tractable reward and state head from DeepMDP into the predictive head.

$$\mathcal{L}_{\mathcal{P}} = \left\| \bar{\mathcal{P}}_{\theta_p}(z_t^c, a_t) - z_{t+1}^c \right\|, \tag{11}$$

$$\mathcal{L}_{\mathcal{R}} = \left\| \bar{\mathcal{R}}_{\theta_r}(z_t^c, a_t) - r_{t+1} \right\|, \tag{12}$$

where $\bar{\mathcal{P}}_{\theta_p}$ and $\bar{\mathcal{R}}_{\theta_r}$ are state and reward predictive heads, respectively. These auxiliary models share the same structure except that the output of $\bar{\mathcal{R}}_{\theta_r}$ is a one-dimension scalar.

# Methodology

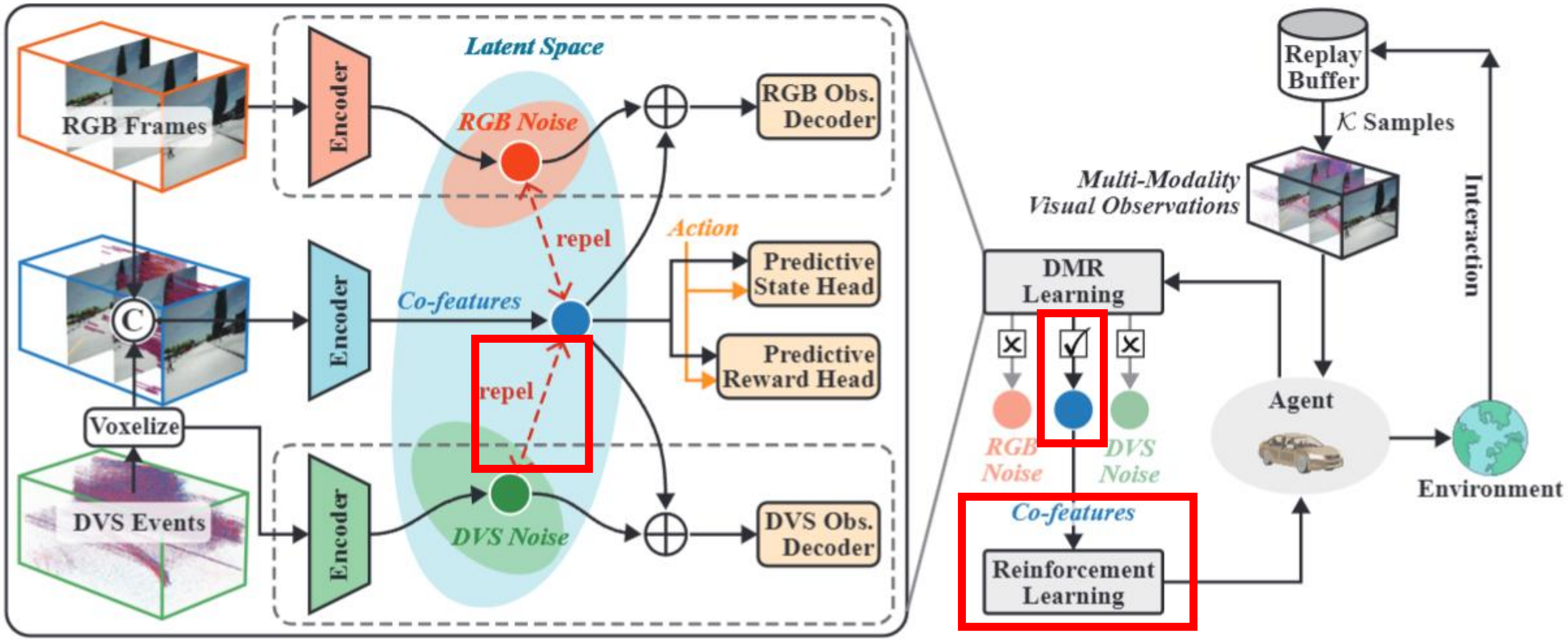


Finally, the noise should exhibit clear dissimilarity from the co-features. In other words, there should be minimal overlap between $h_t^i$ and $z_t^c$. To achieve this distinction, we design the following contrastive constraint:

$$\mathcal{L}_C = -\log \frac{f(z_t^c, \tilde{z}_t^c)}{f(z_t^c, \tilde{z}_t^c) + \sum_{i\in\{\text{rgb},\text{dvs}\}} \sum_{k\in\mathcal{K}} f(z_k^c, \tilde{h}_k^i)}, \quad (13)$$

where $\tilde{z}_t^c$ and $\tilde{h}_t^i$ indicate the moving-averaged target values of $z_t^c$ and $h_t^i$, respectively, and the function $f(a,b) = \exp(\langle a, b\rangle/\tau)$measures the similarity between $a$ and $b$ using the dot product $\langle a, b\rangle$ and the temperature parameter $\tau$.

# Methodology



With the full sensory input decomposed, we can proceed to develop policies for the downstream task using the extracted co-features. **These co-features are isolated from irrelevant information, enabling them to more effectively support the objectives of downstream control.**

In this process, we estimate the action-value and state-value by utilizing the Bellman equation and the co-features generated from DMR. Then, we can derive the policy $\pi_{\emptyset}$:

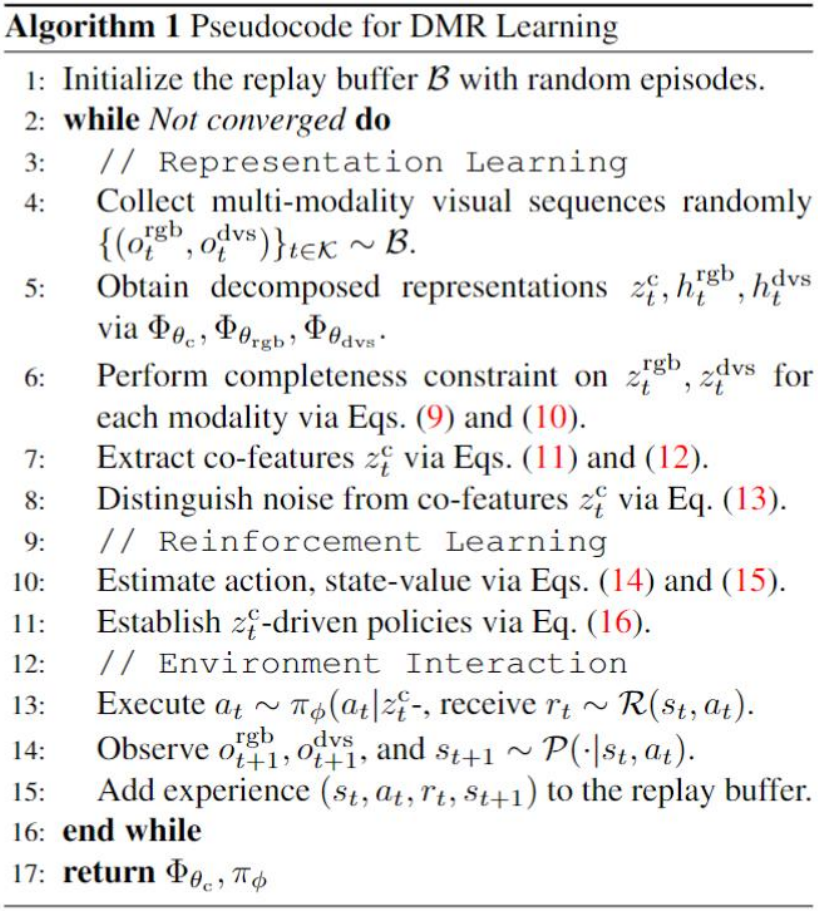
$$\mathcal{L}_Q = \mathop{\mathbb{E}}_{t\in\mathcal{K}}\left[Q(z_t^c, a_t) - (r_t + \lambda V(z_{t+1}^c))\right], \tag{14}$$

$$V(s_{t+1}) = \mathop{\mathbb{E}}_{t\in\mathcal{K}}\left[\tilde{Q}(z_{t+1}^c, a_{t+1}) - \alpha \log \pi(a_{t+1}|z_{t+1}^c)\right]. \tag{15}$$

$$\mathcal{L}_\pi = \mathop{\mathbb{E}}_{t\in\mathcal{K}}\left[\alpha \log \pi_\phi(a_t|z_t^c) - \tilde{Q}(z_t^c, a_t)\right]. \tag{16}$$

# Methodology

The full training pipeline of DMR is provided in Algorithm 1:



**Algorithm 1** Pseudocode for DMR Learning

```
Initialize the replay buffer $\mathcal{B}$ with random episodes.
while Not converged do
    // Representation Learning
    Collect multi-modality visual sequences randomly $\{(o_t^{\mathrm{rgb}}, o_t^{\mathrm{dvs}})\}_{t \in \mathcal{K}} \sim \mathcal{B}$.
    Obtain decomposed representations $z_t^{\mathrm{c}}, h_t^{\mathrm{rgb}}, h_t^{\mathrm{dvs}}$ via $\Phi_{\theta_{\mathrm{c}}}, \Phi_{\theta_{\mathrm{rgb}}}, \Phi_{\theta_{\mathrm{dvs}}}$.
    Perform completeness constraint on $z_t^{\mathrm{rgb}}, z_t^{\mathrm{dvs}}$ for each modality via Eqs. (9) and (10).
    Extract co-features $z_t^{\mathrm{c}}$ via Eqs. (11) and (12).
    Distinguish noise from co-features $z_t^{\mathrm{c}}$ via Eq. (13).
    // Reinforcement Learning
    Estimate action, state-value via Eqs. (14) and (15).
    Establish $z_t^{\mathrm{c}}$-driven policies via Eq. (16).
    // Environment Interaction
    Execute $a_t \sim \pi_\phi(a_t|z_t^{\mathrm{c}}$-, receive $r_t \sim \mathcal{R}(s_t, a_t)$.
    Observe $o_{t+1}^{\mathrm{rgb}}, o_{t+1}^{\mathrm{dvs}}$, and $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$.
    Add experience $(s_t, a_t, r_t, s_{t+1})$ to the replay buffer.
end while
return $\Phi_{\theta_{\mathrm{c}}}, \pi_\phi$
```

Representation Learning

Reinforcement Learning

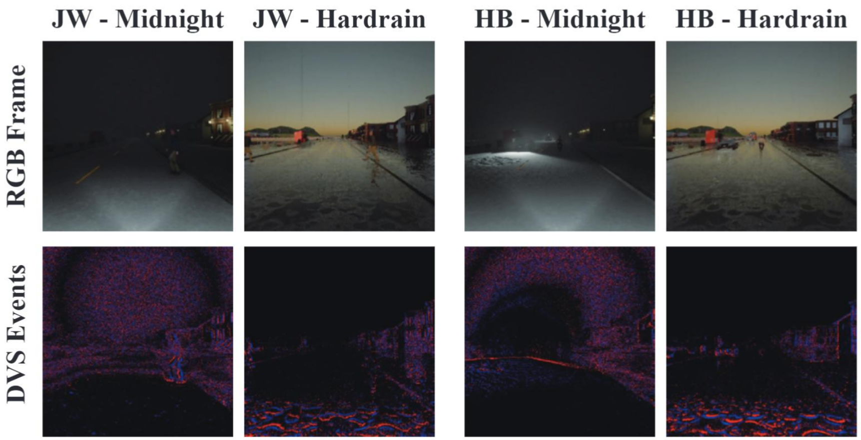
Environment Interaction

# Experimental analysis

We adopt the widely-used Carla simulator to establish our new Carla benchmark. Carla supports a rich set of scenarios with varying lighting and weather conditions.

Our Carla benchmark features **two traffic scenarios**:
(i) the **HighBeam (HB)** scenario, where an ego-vehicle experiences varying lighting conditions while encountering a cyclist, and
(ii) the **JayWalk (JW)** scenario, where the ego-vehicle encounters both stationary and moving pedestrian obstacles intermittently.

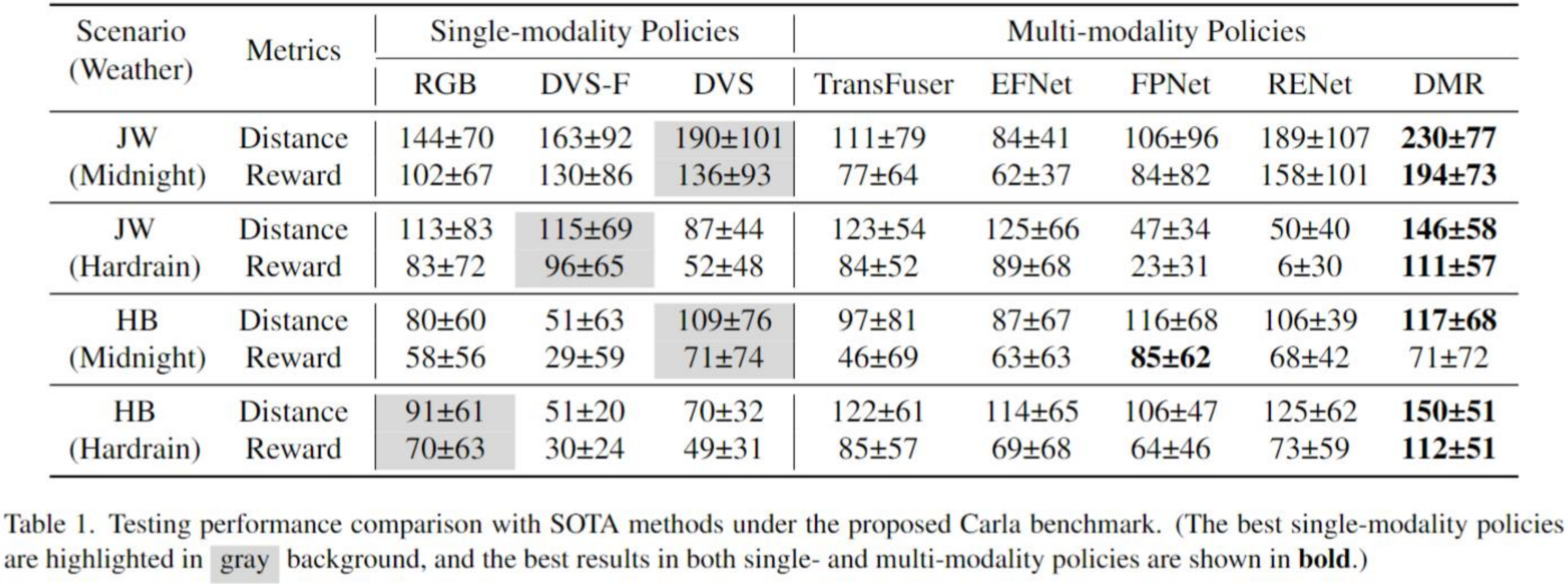
Moreover, the benchmark includes extreme weather conditions (**Midnight and Hardrain**) that can cause RGB camera failure or excessive noise with DVS cameras.



**Metrics**: All experiments are trained across 3 random seeds and 20 evaluation rollouts per seed, yielding **mean and standard deviation of the metrics of episode reward and distance.**

**DVS details**: For multi-modality observations, we focus on the fusion of RGB frames (RGB for short) and DVS voxel grids (DVS). In addition, we introduce the frame-based DVS events, termed **DVS-F**, as a type of observation to show the effectiveness of DVS voxelization.

# Experimental analysis

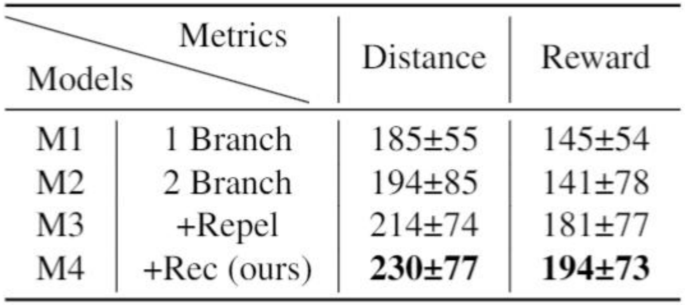


| Scenario (Weather) | Metrics | Single-modality Policies | | | Multi-modality Policies | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | RGB | DVS-F | DVS | TransFuser | EFNet | FPNet | RENet | DMR |
| JW (Midnight) | Distance | 144±70 | 163±92 | 190±101 | 111±79 | 84±41 | 106±96 | 189±107 | **230±77** |
| | Reward | 102±67 | 130±86 | 136±93 | 77±64 | 62±37 | 84±82 | 158±101 | **194±73** |
| JW (Hardrain) | Distance | 113±83 | 115±69 | 87±44 | 123±54 | 125±66 | 47±34 | 50±40 | **146±58** |
| | Reward | 83±72 | 96±65 | 52±48 | 84±52 | 89±68 | 23±31 | 6±30 | **111±57** |
| HB (Midnight) | Distance | 80±60 | 51±63 | 109±76 | 97±81 | 87±67 | 116±68 | 106±39 | **117±68** |
| | Reward | 58±56 | 29±59 | 71±74 | 46±69 | 63±63 | **85±62** | 68±42 | 71±72 |
| HB (Hardrain) | Distance | 91±61 | 51±20 | 70±32 | 122±61 | 114±65 | 106±47 | 125±62 | **150±51** |
| | Reward | 70±63 | 30±24 | 49±31 | 85±57 | 69±68 | 64±46 | 73±59 | **112±51** |

Table 1. Testing performance comparison with SOTA methods under the proposed Carla benchmark. (The best single-modality policies are highlighted in gray background, and the best results in both single- and multi-modality policies are shown in **bold**.)

- **The learned policies of SOTA multi-modality methods often fail to match the performance achieved by single-modality methods.**
  - This could be attributed to the common adoption of multi-scale and attention mechanisms in current state-of-the-art multi-modality methods. These approaches often mix task-relevant information with accumulated noise, complicating the extraction of information crucial for downstream tasks.
- **In contrast, our method offers a solution by explicitly eliminating noise and providing refined co-features for the RL. Compared to alternative multi-modality RL methods, our approach obviates the need for constructing intricate and resource-intensive fusion networks, while still attaining advantages in sample efficiency and learning performance.**

# Experimental analysis



| Models \ Metrics | | Distance | Reward |
|---|---|---|---|
| M1 | 1 Branch | 185±55 | 145±54 |
| M2 | 2 Branch | 194±85 | 141±78 |
| M3 | +Repel | 214±74 | 181±77 |
| M4 | +Rec (ours) | **230±77** | **194±73** |

Table 2. Effect of components in DMR.

- **M2** slightly improves on M1 in terms of distance while having little effect on reward. This is possibly because of the uncertainty in replay buffer sampling during RL training.
- With the introduction of three branches and contrastive constraints (**M3**), there is a significant improvement in both distance and reward.
- with the incorporation of the reconstruction decoder (**M4**), reward and distance further improve, indicating the necessity of the information completeness constraint.

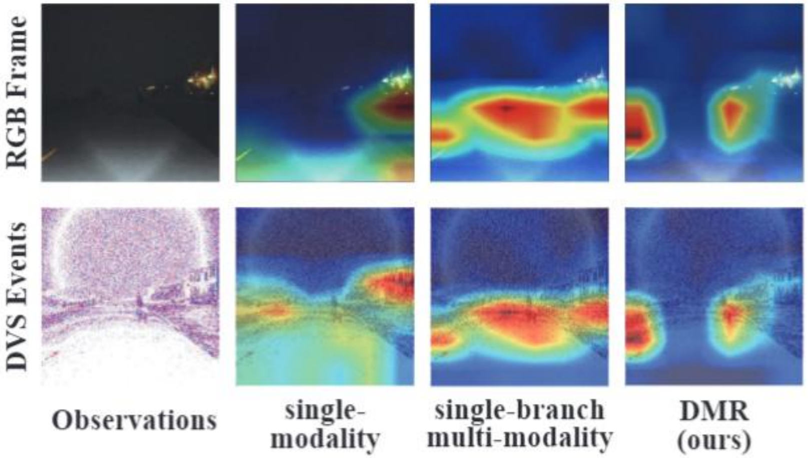
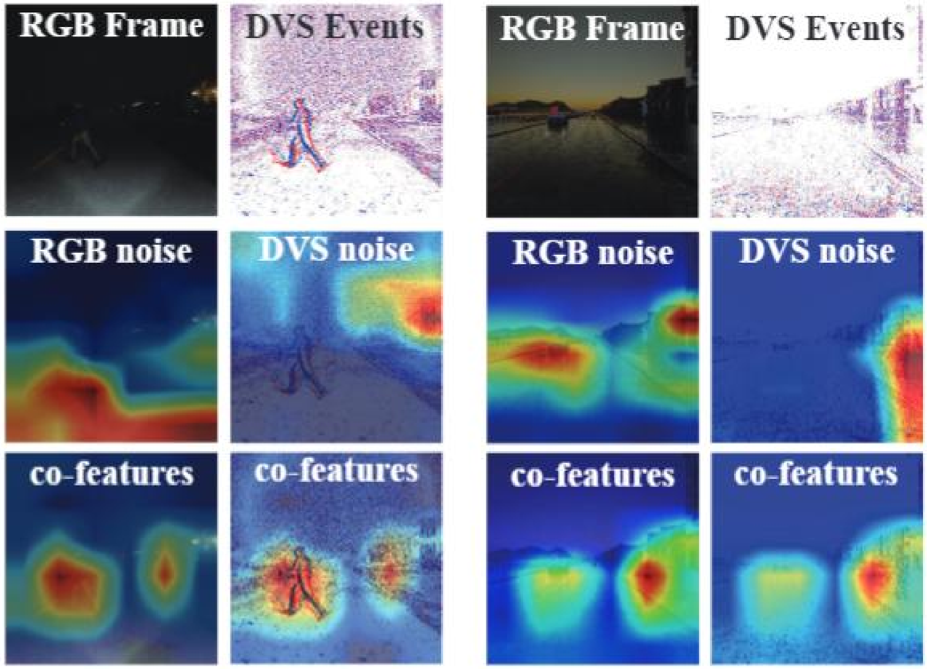


Figure 5. CAMs under different modality RL configurations in the JW-Midnight scenario.

- the CAMs for single-modality models primarily highlight the front road and the adjacent buildings, activating an unnecessarily broad space.
- The simple multi-modality model without using decomposition and contrastive constraints generates a more focused area, but still contains task-irrelevant regions.
- Our DMR effectively captures pertinent areas for RL while eliminating irrelevant regions.

# Experimental analysis



(a) JW-Midnight: an extremely low-light condition.

(b) HB-Hardrain: a rapidly-changing illumination condition.

Figure 6. CAMs of DMR under different illumination conditions.

- In the extremely low-light condition (JW-Midnight), DVS can capture the front pedestrian while RGB camera suffers from exposure failure. It can be seen that RGB noise highlights the high beam region on the road, while DVS noise is activated across a broader region, with the highest activation on the building. We can also see that the co-features attentively grasp the pedestrian and the right roadside simultaneously.

- In the rapidly-changing illumination condition (HB-Hardrain), DVS generates excessive event noise, while RGB can capture rich texture information. Notably, RGB noise mainly highlights brighter regions such as the front road and nearby vehicles and buildings, while DVS noise is prominent around puddles and splashing water. We observe that the co-features distinctly focus on the front cyclist, left vehicle, and right building, which are crucial for driving decision-making.

# Experimental analysis

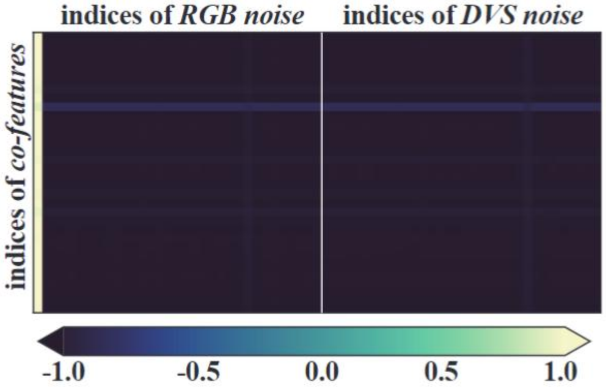


Figure 7. A similarity matrix example at the 100K'th training step.

- We present the similarity matrix between co-features and the modality-specific noises from RGB frames and DVS events, obtained from a training batch of 32 samples at the 100K'th training step.
- Each row in the similarity matrix depicts similarities between the co-feature and itself, RGB noise, and DVS noise.
- **We can see that the co-features exhibit strong coherence among themselves, while their similarity with the noises is remarkably low, illustrating a clear contrast.**

# Conclusion

- This paper explores a new decomposition perspective to address the multi-modality visual RL problem. We propose a novel three-branch multi-modality fusion framework, called DMR, designed for highly-complementary frame- and event-based visual modalities. DMR can explicitly extract task-relevant features from both modalities while mitigating the impact of irrelevant information and noise from each modality. Experimental results demonstrate the efficacy and superiority of DMR in policy performance.

- We will focus on exploring the generalization of modalities and stability in more diverse and realistic scenarios in a sim2real fashion.