

Semore: VLM-guided Enhanced Semantic Motion Representations for Visual Reinforcement Learning

Wentao Wang¹, Chunyang Liu², Kehua Sheng², Bo Zhang^{2*}, Yan Wang^{1*}

¹ Institute for AI Industry Research, Tsinghua University

² Didi Chuxing

kensorpl@aliyun.com, {liuchunyang, shengkehua, zhangbo}@didiglobal.com, wangyan@air.tsinghua.edu.cn

Abstract

The growing exploration of Large Language Models (LLM) and Vision-Language Models (VLM) has opened avenues for enhancing the effectiveness of reinforcement learning (RL). However, existing LLM-based RL methods often focus on the guidance of control policy and encounter the challenge of limited representations of the backbone networks. To tackle this problem, we introduce Enhanced Semantic Motion Representations (Semore), a new VLM-based framework for visual RL, which can simultaneously extract semantic and motion representations through a dual-path backbone from the RGB flows. Semore utilizes VLM with common-sense knowledge to retrieve key information from observations, while using the pre-trained clip to achieve the text-image alignment, thereby embedding the ground-truth representations into the backbone. To efficiently fuse semantic and motion representations for decision-making, our method adopts a separately supervised approach to simultaneously guide the extraction of semantics and motion, while allowing them to interact spontaneously. Extensive experiments demonstrate that, under the guidance of VLM at the feature level, our method exhibits efficient and adaptive ability compared to state-of-art methods. All codes are released ¹.

Introduction

Thanks to its ability to directly convert complex visual signals into actions, visual Reinforcement Learning (RL) has achieved great success in intelligent agent control in recent years (Arulkumaran et al. 2017; Ze et al. 2023; Zheng et al. 2024). It has wide applications in various domains such as autonomous driving, electronic sports, and robotic control (Liang et al. 2018; Kiran et al. 2021; Lample and Chaplot 2017; Nair et al. 2018). Due to the high dimensionality of visual signals and the inefficiency of RL interactions, the algorithm struggles to understand the environment well, making it difficult to capture optimal rewards and lacking interpretability. Efficiently extracting task-relevant representations from visual observations is crucial for breaking through the bottleneck of RL.

Previous researches leverage diverse state abstraction approaches including observation reconstruction (Vemprala,

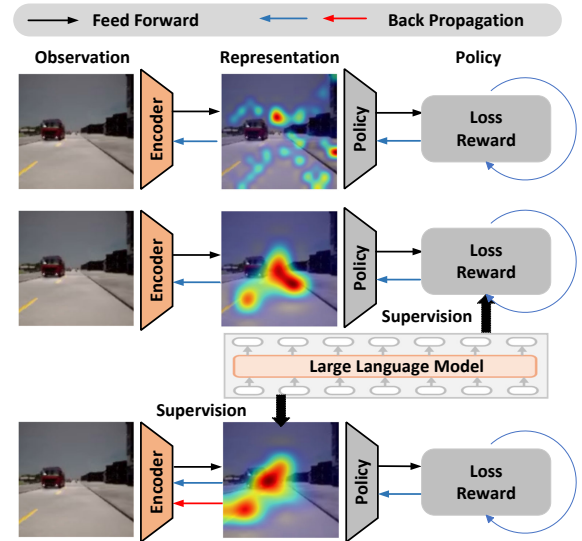


Figure 1: (a) In the first row, the overly large sampling space of RL leads to difficulty in capturing key objectives in extreme scenarios; (b) In the second row, due to the complex high-dimensional feature space and the back propagation, guidance at the policy level cannot ensure that the encoder extracts reliable features; (c) In contrast, our method can fully take advantage of the capability of VLMs to enhance the task-specific representations.

Mian, and Kapoor 2021; Yu et al. 2022), transition dynamics prediction (Gelada et al. 2019) and bisimulation (Zhang et al. 2020), resulting in the issue of high cost and data redundancy (Wang et al. 2024a). Consequently, a question arises regarding the existence of a more efficient method for explicitly extracting task-relevant representations.

Promisingly, Large Language Models (LLMs) have been actively developed in recent years, bridging human interaction and reasoning (Wang et al. 2024a; Gbagbe et al. 2024; Hu et al. 2024). Based on the advancements in LLMs, they can provide a more holistic understanding of the environment, allowing agents to respond more effectively to various scenarios with human-like logic (Han et al. 2024; Huang et al. 2024). Some works leverage LLM to guide the

*Corresponding authors.

¹ <https://github.com/StephensonRRR/Semore>.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

learning of RL at policy level (Chen et al. 2024; Ma et al. 2024; Hu et al. 2024) and indicate the enormous potential of LLMs. Due to the long forward propagation chain of the RL model, the guidance at the policy level cannot effectively enhance the extraction capability of representations, particular in complex visual input tasks (as shown in Fig. 1 (b)). This motivates the idea that the common-sense knowledge embedded in LLM can also be exploited to enhance the extraction capability of task-relevant representations at the feature level, which is shown in Fig. 1 (c).

In order to address the aforementioned limitations, we introduce a novel Enhanced Semantic Motion Representation (Semore) Learning framework for visual RL, which employs a two-stream network to separately extract semantic and action representations. This design can decouple different features, thereby fully utilizing the LLM to guide the representation learning. Specifically, the semantic stream models the environmental semantics and can identify key objects in the scenario, while the motion stream models motion clues from the residual frame of adjacent frames. To align the extracted representations and the actual surrounding environment, we introduce a VLM-based feature-level supervision module. We utilize VLM to generate task-specific feature masks, highlighting key regions in the observations.

Semantic and motion representations have strong complementarity and therefore can be enhanced through interaction with each other. Unlike previous work that used transformer-based networks to fuse two specific feature maps, we inject the feature map generated by the VLM during the training process. Specifically, we adopt feature similarity loss to align the extracted semantic features to the VLM semantics for the semantic path. Meanwhile, we adopt cross-attention between motion features and VLM semantics to enhance the motion representation of key regions. Essentially, the supervision for both pathways can let encoders to focus on key regions. Note the interaction between semantics and motion is spontaneously achieved, with the knowledge-aware features provided by the VLM serving as a mediator in this process. Both semantic and motion representations are enhanced and fused for decision-making.

In summary, the contributions of this paper are three-fold:

- We propose Semore, a novel VLM-based visual reinforcement learning framework that can enhance representations by integrating VLM-based common-sense knowledge guidance at the feature level.
- We designed a decoupled supervision module. For the semantic flow, we use explicit supervision for alignment, while for the motion flow, we use cross-attention to guide the focus areas.
- We conduct comprehensive experiments using Carla benchmarks. Experimental results demonstrate the state-of-the-art performance of our proposed method and the effectiveness of the corresponding components.

Related Works

Visual Reinforcement Learning. In vision-based RL, agents extract compact representations from low-dimensional visual observations to achieve decision-

making. In this process, representation learning is the key to improving the performance of visual RL. Existing works can be roughly divided into three main approaches: (i) data augmentation technique (Huang et al. 2023b; Zhang et al. 2020); (ii) self-supervised representations (Castro 2020; Hansen et al. 2020); (iii) modeling environment dynamics (Pan et al. 2022; Fu et al. 2021; Lee et al. 2020). However, due to the extensive exploration required in the RL process, existing methods without any prior knowledge struggle to efficiently extract representations, especially in complex environments. The emergence of LLMs and VLMs brings new opportunities for addressing this issue.

Dual-stream Network. Dual-stream networks are particularly popular for extracting diverse representations (Gao et al. 2018; Simonyan and Zisserman 2014; Wang et al. 2024c). Generally, this structure is used for encoding heterogeneous modalities such as point cloud and text for feature fusion (Xiang, Xu, and Ma 2023; Liu et al. 2023; Liang et al. 2022). Some studies have demonstrated that separately extracting different representations containing specific information from images can achieve better performance than single-stream networks in extracting diverse representations (Kim, Jones, and Hager 2021; Liu et al. 2021; Huang, Zhao, and Wu 2023; Liu et al. 2022). Simoun (Huang et al. 2023b) adopts this design in visual RL learning and constructs a structure interaction module to leverage the correlations of the dual-stream features. We adopt a two-stream structure to decouple the semantic and motion representation learning, allowing the VLM to separately supervise the feature extraction and interaction.

VLM-based Learning. VLMs have shown significant potential in learning high-quality representations for diverse downstream tasks (Du et al. 2024; Singh et al. 2022; Chen et al. 2025). Their success largely stems from training transformer architectures on large-scale datasets of image-text pairs sourced from the web, using contrastive learning techniques. Notably, CLIP (Radford et al. 2021) proposed a promising alternative that directly learns transferable visual concepts from large-scale collected image-text pairs. In this paper, we first use the VLM to retrieve semantic information from the observations, such as relevant objects, and then use a clip-based approach to generate the corresponding visual features, thereby embedding common-sense knowledge into the representation learning.

Methodology

The overall framework is illustrated in Fig. 2. We start by formalizing the task of visual RL and then discuss the details of Semore.

Problem Formulation

Visual RL can be normally formulated as a Partially Observable Markov Decision Process (POMDP), denoted as a tuple $\mathcal{M} = \langle \mathcal{O}, \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, where $\mathcal{O}$ denotes the observation space containing RGB frames o_t at different time step and $\mathcal{A}$ denotes the action space. The interaction process of the agent in a POMDP can be defined as follows: (i) the agent perceives visual observations o_t ; (ii) the agent then selects an action $a_t \in \mathcal{A}$ based on a stochastic policy $\pi(a_t|o_t)$.

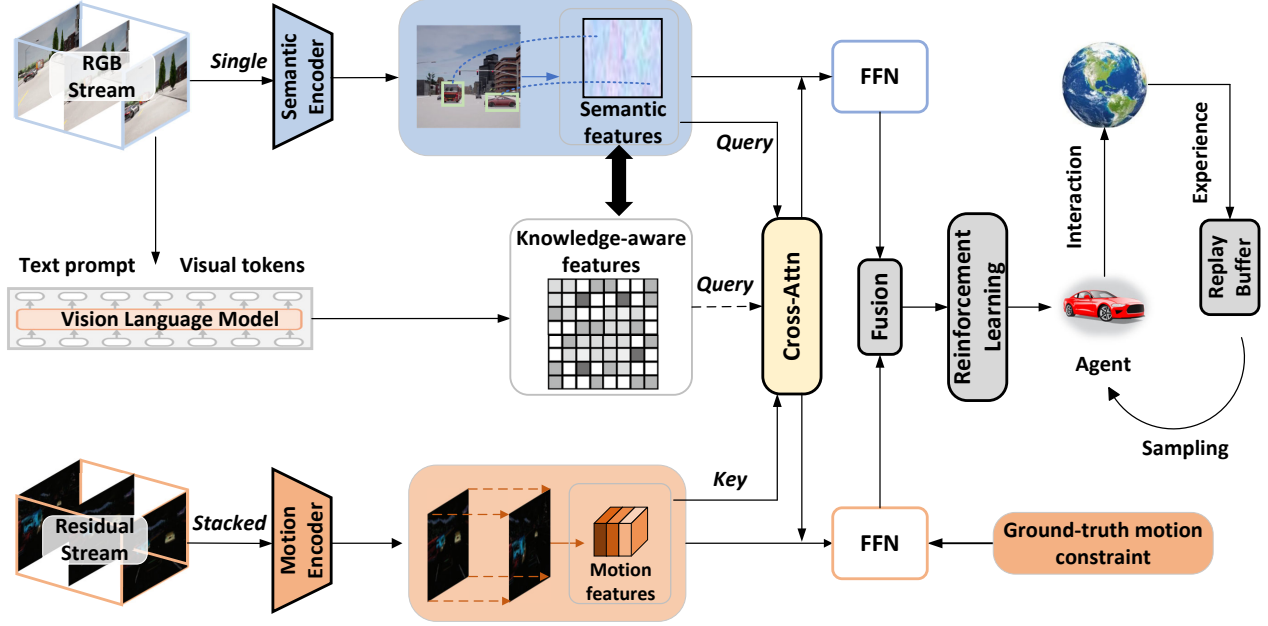


Figure 2: The overall VLM-guided learning framework. It integrates two key modules: 1) the VLM-guided semantics module employs the similarity loss to explicitly supervise the extraction of semantic representations, while the motion supervision module introduces knowledge-aware features into the motion extraction using bidirectional cross-attention.

$P(o_t, a_t)$ is the observation transition, $\mathcal{R}(o_t, a_t)$ is the reward function, and γ denotes the discount factor. The goal of this formulation is to find an optimal policy that maximizes the expected cumulative reward based on the visual observations across the entire traversal of MDPs:

$$J(\pi) = \sum_t E_{(o_t, a_t) \sim \pi} [\mathcal{R}(o_t, a_t)] \quad (1)$$

During training, π is used to interact with the environment and the generated experience is stored in a replay buffer $\mathcal{B}$.

Feature Extraction

Semantic Encoding. Semantic information is vital for visual representations as it can provide environmental understanding for the agent. Specifically speaking, the agent can identify objects or events relevant to its control based on visual representations from observations, thereby improving the decision-making capabilities. Semantic encoding aims to extract the visual semantic representations of the surrounding environment from the observed frames. Semantic encoder adopts a four-layer convolutional structure with 3×3 kernel size and ReLU non-linearity. And the output feature map of the last convolution layer is represented as $F_s^t \in R^{C \times H \times W}$. Then we use a fully connected layer with layer normalization to reduce the dimension of F_s^t and output compact feature vector f_s^t .

Motion Encoding Motion information is critical for vision-based agents to understand the dynamics of the surrounding scenarios. In other words, motion features enable the agent to have predictive capabilities, allowing it to make

more reasonable decisions. The goal of motion encoding is to extract the motion features at the pixel level (such as the movement of objects within the perception range) from multiple adjacent visual frames. Given a sequence of 3 adjacent observation frames $[o_{t-2}, o_{t-1}, o_t]$ sampled from the replay buffer, we can obtain the motion input by residual of adjacent frames $[o_{t-1} - o_{t-2}, o_t - o_{t-1}]$. Similar to semantic encoding, the motion encoder adopts a four-layer convolutional encoder with a different number of first-layer's input channels to extract feature map $F_m^t \in R^{C \times H \times W}$.

VLM-guided Semantics

In the absence of supervision on the feature level in previous visual RL models, the extracted representation of observations is not guaranteed to align with the expressive nature of the real environment. In particular, the long propagation chain and the vast sampling space of RL make it difficult to learn ground-truth representations through the mere supervision of the policy level. The difference from the true representation imposes inherent limitations on the model. To mitigate this shortfall, we introduce VLM-guided representations by using the pretrained VLM and clip-based image segmentation model CRIS (Wang et al. 2022), as illustrated in Fig. 3. The core objective is to align the extracted representations with the ground-truth as possible, thereby enhancing the agent's understanding of the environment.

We first employ the pretrained VLM to extract comprehensive semantic information from raw visual observations. The VLM takes a single frame o as input and uses a visual encoder g_V to extract the visual features, which are then

converted into language embedding tokens. Meanwhile, the prompt P_m is fed into the text encoder g_T to obtain text tokens. This can be formulated as:

$$\begin{aligned} H_V &= g_V(O) \\ H_T &= g_T(P_m) \end{aligned} \quad (2)$$

where H_V , H_T are the visual and text tokens, respectively. Then the visual tokens H_V and the text tokens H_T are fed into the VLM f for generating responses:

$$Y = f([H_V, H_T]) \quad (3)$$

where Y is the output of task-specific prompts, as the text semantics.

Given the text semantics, we need to generate a semantic feature map of the observations for representation learning. Although CLIP (Radford et al. 2021) learns powerful image-level visual concepts by aligning the textual representation with the image-level representation, this type of knowledge is suboptimal for referring image segmentation, due to the lack of more fine-grained visual concepts. Hence, we apply CRIS (Wang et al. 2022), a clip-driven image segmentation framework to accurately generate more discriminative visual representations through the alignment of text and visual features at the pixel-level. Specifically, given image o and text sequence $Y = y_1, \dots, y_n$, CRIS can compute and output a similarity map. Then we use the sigmoid function (Cybenko 1989) to segment specific objects in the image, thereby generating corresponding high-confidence masks. We add the feature masks of different objects together to obtain the complete mask map. This process can be described as:

$$\begin{aligned} h_y &= CRIS(o, y), y \in Y \\ h_{ka} &= Sigmoid(h_y) \\ H_{ka} &= \sum_{y \in Y} h_{ka} \end{aligned} \quad (4)$$

where o , Y is the input observed frame and text semantics. $Sigmoid$ denotes the Sigmoid function.

To impose effective supervision on semantics, we employ a similarity loss (Zhao et al. 2016) between extracted and knowledge-aware representations, which is denoted as:

$$L_{S-G} = \sum_{\mathcal{T}} \frac{\|F_s(t) - \hat{H}_{ka}(t)\|_1}{\|\hat{H}_{ka}(t)\|_1} \quad (5)$$

where $\hat{H}_{ka}$ is obtained by input H_{ka} into the semantic encoder, thus having the same dimension as F_s . Based on this process, VLM can leverage common-sense information to provide explicit guidance for semantics extraction. This guidance enables the visual encoder to focus on critical factors from the observations, thereby allowing it to detect objects that are key to control.

Motion Enhancement and Interaction

In previous work, popular approaches to handling different features were to use attention mechanisms for interaction, thereby outputting fused features. However, this process is a black-box operation, and supervision exists only at the

end-point, thus lacking a comprehensive understanding. In particular, when processing motion and semantic features, attention-based fusion essentially aggregates them from different spaces into a common space. However, due to the impact of sampling efficiency, the learned fusion space is prone to overfitting and has low interpretability. To address this issue, we use VLM to enhance the motion features and guide the interaction between semantics.

Since motion information is extracted using frame differences, which is sparse. This makes it difficult to apply feature alignment for supervision. Therefore, during the training phase, we employ bidirectional cross-attention to guide the motion encoder in focusing on key areas using VLM-generated features. Specifically, as shown in Fig. 4, during the training, the input of the interaction module are knowledge-aware feature map H_{ka} and a motion feature map F_m . Then an interactive attention map X can be obtained:

$$X = \sigma(\sigma(\tilde{H}_{ka}^T \tilde{F}_m) + \sigma(\tilde{F}_m \tilde{H}_{ka}^T)) \quad (6)$$

where $\tilde{H}_{ka}$ is obtained by inputting to a convolution layer for reducing the spatial complexity, $\tilde{F}_m$ is as well. σ denotes the Softmax function. Then we use the interactive feature map containing both semantic and motion information to simultaneously enhance original representations. Specifically, we separately use a fully connected layer to process the interaction and then add them to the feature maps of the semantic and motion feature maps:

$$\begin{aligned} F_s &= F_s + FN(X), \\ F_m &= F_m + FN(X) \end{aligned} \quad (7)$$

where FN denotes the fully connected layer. In this way, the semantic and motion representations complement each other, thereby enhancing understanding. Note that the semantics generated by VLM are only used to compute attention weights during training, while in reality, it is the semantic features and motion features that interact.

To ensure the motion encoder can effectively extract sufficient features and remove noise to prevent redundancy, we adopt a transition constraint via an MLP predictor $\mathcal{P}_m$. Specifically, the obtained feature vector f_t^m and action a_t at time step t is input into the motion predictor. And the predictor can predict future features, thereby enhancing the encoder's ability to extract motion information. Then the transition loss can be defined as:

$$\mathcal{L}_{trans} = \|\mathcal{P}_m(f_t^m, a_t), f_{t+1}^m\|_2^2 \quad (8)$$

where $\mathcal{P}_m$ represents the motion predictor and $\|\cdot\|_2$ is the L2-norm.

After completing the interaction, we use fully connected layers to reduce the dimensions of F_t^s and F_t^m , obtaining compact features f_t^s and f_t^m . Then we concatenate them and the final representation is $f_t = [f_t^s, f_t^m]$. To reduce noise, we use a prediction head to further purify the information related to RL rewards. Motivated by DeepMDP (Gelada et al. 2019), we utilize a reward predictive head by incorporating the tractable reward and state head from DeepMDP (Gelada et al. 2019) to predict the reward value of each observation-action pair:

$$\mathcal{L}_R = \|\mathcal{R}(f_t, a_t) - r_{t+1}\| \quad (9)$$

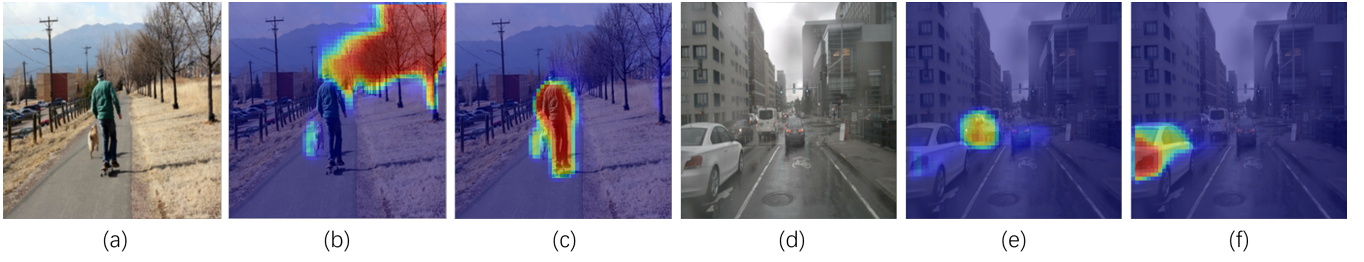


Figure 3: The illustration of VLM-generated knowledge-aware representations. (a) and (d) show the road scenarios of rural and urban areas, respectively. The text prompts are: (b): The right trees; (c): The pedestrian; (e): The white van on the left side of the black car ahead; (f): The left white car.

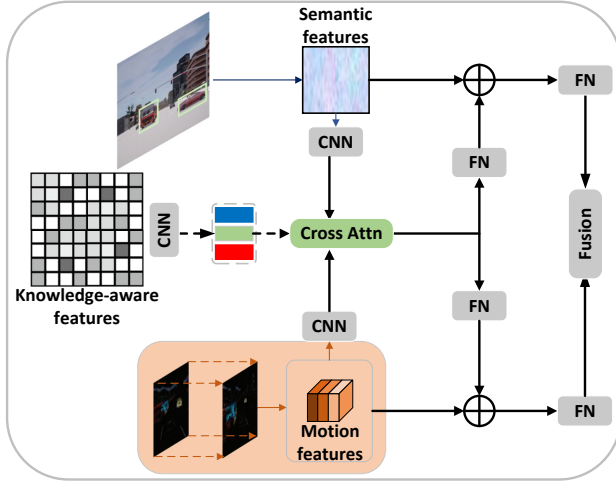


Figure 4: Interaction of motion and semantic features. (Fusion is concat and knowledge-aware features are only used in training stage.).

where r_{t+1} is the actual reward value at the next time step, which is returned from the interaction with the environment.

Reinforcement Learning based on Semore

We adopt the baseline RL algorithm SAC (Haarnoja et al. 2018) to maximize the expected cumulative reward to find the optimal policy via approximating the action-value Q and a stochastic policy π based on an α -discounted maximum entropy $\mathcal{H}(\cdot)$. The action-value function Q is learned by minimizing the soft Bellman error and the soft state value V can be estimated by sampling an action under the current policy. The above process can be formulated as:

$$\begin{aligned}
 J(\pi) &= \sum_t E_{(o_t, a_t) \sim \pi} [r(o_t, a_t) + \alpha \mathcal{H}(\pi(\cdot | o_t))], \\
 \mathcal{L}_Q &= E_{(o_t, a_t)} (Q(o_t, a_t) - (r_t + \lambda V(o_{t+1})))^2, \\
 V(o_{t+1}) &= E_{a_{t+1} \sim \pi} [\tilde{Q}(o_{t+1}, a_{t+1}) - \alpha \log \pi(a_{t+1} | o_{t+1})],
 \end{aligned} \tag{10}$$

where $\tilde{Q}$ denotes the exponential moving average of the parameters of Q . And the policy is optimized by decreasing

the difference between the exponential of the soft-Q function and the policy:

$$\mathcal{L}_\pi = E_{a_t \sim \pi} [\alpha \log \pi(a_t | o_t) - Q(o_t, a_t)]. \tag{11}$$

Training Details

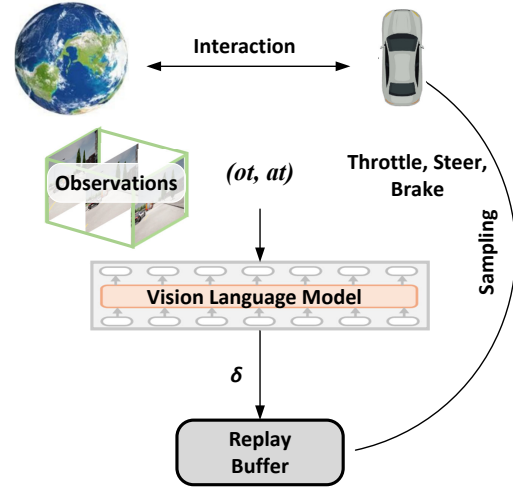


Figure 5: VLM-based selective replay buffer.

Selective Replay Buffer RL agents are very likely to perform ineffective exploration in the initial stage due to the lack of prior knowledge. To alleviate the low exploration efficiency for large continuous action space in visual RL that often prohibits the use of challenging tasks, we design the selective replay buffer to provide better exploration. The intuition is that if the agent is provided with some positive training data such as expert supervision at the beginning, it can acquire a certain level of initial execution capability, thus avoiding the high cost of excessive random sampling. Specifically, our RL framework will be first warmed up by learning knowledge via the observation-action pairs that are deemed qualified by the LLM to initialize the action exploration in a reasonable space.

LLM is not able to make precise decision signals, but can offer macro-level guidance, such as braking, turning left, or

turning right. Therefore, it can be fully leveraged to evaluate the reasonableness of observation-action pairs. As illustrated in Fig. 5, the generated pair (o_t, a_t) is fed into the image LLM and we can obtain the output that is reasonable or unreasonable. We additionally adopted a decay factor δ to represent the probability of adding the observation-action pair to the replay buffer. Its initial value is set to 1, and it decreases along the training process. When its value reaches 0.5, it indicates that no further selective additions will be made, and instead, all interactions with the environment will be added to the replay buffer.

Training Objective. Based on the fused semantic motion representations, Semore learns from visual to control signals in an end-to-end manner via optimizing the following equation:

$$\mathcal{L} = \underbrace{\mathcal{L}_{trans}}_{\text{motion}} + \underbrace{\mathcal{L}_{S-G}}_{\text{semantic}} + \underbrace{\mathcal{L}_R}_{\text{state}} + \underbrace{\mathcal{L}_\pi + \mathcal{L}_Q}_{\text{RL}} \quad (12)$$

where the objective jointly considers the semantic and motion representations, as well as the purification of reward-related information for RL learning.

Experiments

Experimental Setup

To evaluate our approach under realistic and challenging vision-based environments, we employ the CARLA simulator (Dosovitskiy et al. 2017), which is a widely used open-source simulator for autonomous driving (Liang et al. 2018; Xu et al. 2024; Huang et al. 2023a). CARLA provides a rich and realistic urban environment to evaluate autonomous driving agents in various traffic scenarios. As shown in Fig. 6, we evaluate our method in three traffic scenarios: the HighBeam (HB) scenario, where the ego-vehicle encounters a cyclist, JayWalk (JW) scenario, where the ego-vehicle encounters both stationary and moving pedestrians intermediately and Highway (HW) scenario, where the ego-vehicle is driving on an eight-lane highway with numerous vehicles traveling in the same direction. Similar to (Zhang et al. 2020; Fan et al. 2021), the reward function can encourage the agent to avoid crashes with other moving and static objects and travel as long as possible. We set the single camera on the ego-vehicle’s roof with a view of 60-degree.

Our method is implemented based on SAC (Haarnoja et al. 2018) and DeepMDP (Gelada et al. 2019). The same encoder network architecture and training hyperparameters are adopted for all comparative methods. The spatial resolution of the input RGB images is $128 \times 128 \times 3$. All methods are trained for 110k frames using 5 random seeds to report the mean and standard deviation of the rewards. And more details can be found in the attached supplementary material.

We adopt Qwen2-VL-7B-Instruct (Wang et al. 2024b) in the experiments and Fig. 7 shows the input prompts for the VLM.

Methods Compared: We consider the following baseline methods for comparison: 1) SAC (Haarnoja et al. 2018), a widely-used RL algorithm based on α -discounted maximum entropy; 2) Flare (Shang et al. 2021), a multi-frame visual RL method that utilizes temporal information through latent



Figure 6: Visualization of the CARLA scenarios, where the left column is JW, the middle column is HB, and the right column is HW.

vector differences. 3) CURL (Laskin, Srinivas, and Abbeel 2020), which integrates contrastive learning with model-free RL with minimal changes to the architecture and training pipeline. 4) DrQ (Yarats, Kostrikov, and Fergus 2021), built upon the SAC by adding a convolutional encoder and data augmentation in the form of random shifts. 5) DeepMDP (Gelada et al. 2019), a latent model of an MDP and has been trained to minimize two tractable losses: predicting the rewards and predicting the distribution of the next latent states. 6) Simoun (Huang et al. 2023b), a dual-stream visual RL method that simultaneously extracts appearance and motion information, and enhances representations through interaction and intrinsic rewards.

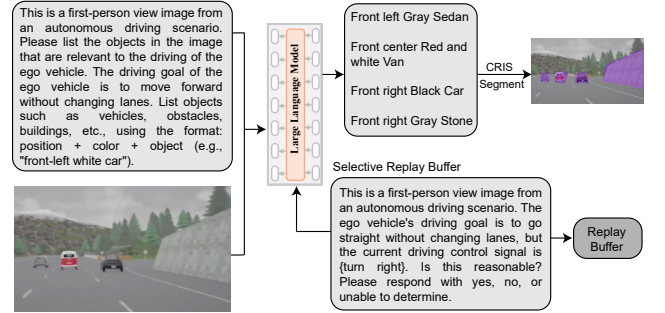


Figure 7: The input prompts for the VLM.

Overall Performance

We compare Semore with benchmarks and the results are shown in Tab. 1. It can be observed that our method outperforms all other methods in terms of the episode reward. And the average driving distance is farther than other methods and the average crash intensity is lower. In particular, the observed improvements over Simoun emphasize the effectiveness of VLM in guiding representation learning. Note that our method did not achieve the best driving smoothness with a higher average brake and steer value. Combining the driving distance and crash intensity, this is likely because the comparative methods did not make appropriate obstacle avoidance decisions based on specific objectives. This is particularly evident in scenarios with higher traffic density. When the number of objects on the road increases, the ego-vehicle must take action to alter its current state for obstacle avoidance. Fig. 8 visualizes the computed interaction feature masks using the Equation. 7 of the HW scenario.

Scen.	Metrics	SAC	CURL	Flare	DrQ	DeepMDP	Simoun	Ours
JW	Episode reward $\uparrow$	103 $\pm$ 72	112 $\pm$ 81	136 $\pm$ 52	97 $\pm$ 59	146 $\pm$ 48	168 $\pm$ 79	201 $\pm$ 54
	Distance (m) $\uparrow$	143 $\pm$ 58	128 $\pm$ 61	123 $\pm$ 68	109 $\pm$ 39	132 $\pm$ 53	208 $\pm$ 74	233 $\pm$ 66
	Crash intensity $\downarrow$	4633 $\pm$ 184	3829 $\pm$ 153	2974 $\pm$ 171	3013 $\pm$ 187	2627 $\pm$ 98	2382 $\pm$ 103	2043 $\pm$ 98
	Average steer (%) $\downarrow$	15.40	14.79	12.53	13.29	10.88	13.69	13.91
	Average brake (%) $\downarrow$	2.06	2.83	2.27	2.02	1.95	2.80	2.86
HB	Episode reward $\uparrow$	73 $\pm$ 62	86 $\pm$ 67	82 $\pm$ 54	93 $\pm$ 71	101 $\pm$ 45	104 $\pm$ 66	123 $\pm$ 60
	Distance (m) $\uparrow$	86 $\pm$ 43	96 $\pm$ 57	92 $\pm$ 56	91 $\pm$ 64	112 $\pm$ 59	128 $\pm$ 61	163 $\pm$ 52
	Crash intensity $\downarrow$	4850 $\pm$ 153	4145 $\pm$ 193	3692 $\pm$ 133	3511 $\pm$ 176	2921 $\pm$ 102	2604 $\pm$ 109	2418 $\pm$ 103
	Average steer (%) $\downarrow$	15.53	14.20	12.13	12.97	10.14	13.56	13.03
	Average brake (%) $\downarrow$	2.43	3.11	2.94	2.81	2.57	3.06	3.08
HW	Episode reward $\uparrow$	126 $\pm$ 24	134 $\pm$ 19	138 $\pm$ 29	167 $\pm$ 23	182 $\pm$ 35	268 $\pm$ 30	316 $\pm$ 33
	Distance (m) $\uparrow$	102 $\pm$ 17	137 $\pm$ 35	114 $\pm$ 22	108 $\pm$ 29	129 $\pm$ 22	202 $\pm$ 22	263 $\pm$ 17
	Crash intensity $\downarrow$	3870 $\pm$ 98	3122 $\pm$ 104	2548 $\pm$ 91	2487 $\pm$ 98	2153 $\pm$ 76	1816 $\pm$ 63	1671 $\pm$ 56
	Average steer (%) $\downarrow$	17.03	15.89	12.28	15.95	10.56	15.10	14.24
	Average brake (%) $\downarrow$	1.84	2.56	1.82	1.63	1.54	2.14	2.13

Table 1: Quantitative results of different models for driving policies. ($\uparrow$ indicates that larger is better while $\downarrow$ means opposite.)

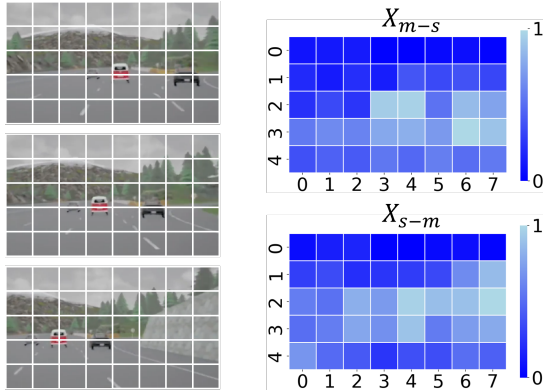


Figure 8: Visualization of the feature masks.

mance improvements. However, due to its inability to effectively integrate motion information, and considering the highly dynamic nature of the scenes, its obstacle avoidance capability improves only marginally compared to M2.

Model		M1	M2	M3	M4
Metrics	Distance(m)	127	163	174	233
	Intensity	3420	2582	2471	2043
HB	Distance(m)	76	119	137	163
	Intensity	3846	2673	2549	2418
HW	Distance(m)	146	203	238	263
	Intensity	3073	1940	1886	1671

Table 2: Effect of components in Semore.

Ablation Study

Effectiveness of Semore Components. To validate the contribution of each component, we incrementally incorporate individual components of the framework and obtain a series of models labeled M1 to M4. Specifically, M1 utilizes solely the semantic-stream branch for decision-making without the supervision of explicit similarity loss; In M2, the semantic and motion branches are employed, and the features from both streams are directly concatenated to feed into the policy learning. Both M3 and M4 utilize VLM-generated features as explicit supervision of semantic representation. M3 leverages semantic supervision to align to the knowledge-aware representations in terms of Eq. 5. M4 builds upon M3 by incorporating motion supervision and interaction.

Tab. 2 shows the performance of each model. It is clear that M1 degrades to a conventional multi-frame input visual RL, while M2, by decoupling semantic and motion information, enhances the representation extraction capability. This demonstrates the effectiveness of dual-stream design. With the guidance of the VLM, M3 achieves significant perfor-

Conclusion

In this paper, we propose Semore, a novel framework aimed at addressing the issue of limited representation learning capability in visual RL. Semore can leverage knowledge-aware supervision in both semantic and motion representation learning under the guidance of VLM. Based on the decoupled two-stream network architecture, semantic extraction can be enhanced through feature alignment under explicit supervision. Simultaneously, a bidirectional cross-attention mechanism is used to enhance motion extraction while achieving semantic-motion interaction. Thus, the knowledge of the VLM is distilled into our encoders, thereby enhancing the representations. Extensive experiments in different challenging scenarios demonstrate the efficacy and superiority.

Acknowledgments

This work is supported by Xiongan AI Institute, Didi Chuxing and Wuxi Research Institute of Applied Technologies, Tsinghua University under Grant 20242001120.

References

- Arulkumaran, K.; Deisenroth, M. P.; Brundage, M.; and Bharath, A. A. 2017. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6): 26–38.
- Castro, P. S. 2020. Scalable methods for computing state similarity in deterministic markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10069–10076.
- Chen, S.; Jiang, B.; Gao, H.; Liao, B.; Xu, Q.; Zhang, Q.; Huang, C.; Liu, W.; and Wang, X. 2024. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv preprint arXiv:2402.13243*.
- Chen, Y.; Ding, Z.-h.; Wang, Z.; Wang, Y.; Zhang, L.; and Liu, S. 2025. Asynchronous Large Language Model Enhanced Planner for Autonomous Driving. In *European Conference on Computer Vision*, 22–38. Springer.
- Cybenko, G. 1989. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4): 303–314.
- Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*, 1–16. PMLR.
- Du, P.; Wang, Y.; Sun, Y.; Wang, L.; Liao, Y.; Zhang, G.; Ding, E.; Wang, Y.; Wang, J.; and Liu, S. 2024. LaMI-DETR: Open-Vocabulary Detection with Language Model Instruction. *arXiv preprint arXiv:2407.11335*.
- Fan, L.; Wang, G.; Huang, D.-A.; Yu, Z.; Fei-Fei, L.; Zhu, Y.; and Anandkumar, A. 2021. Secant: Self-expert cloning for zero-shot generalization of visual policies. *arXiv preprint arXiv:2106.09678*.
- Fu, X.; Yang, G.; Agrawal, P.; and Jaakkola, T. 2021. Learning task informed abstractions. In *International Conference on Machine Learning*, 3480–3491. PMLR.
- Gao, J.; Ge, R.; Chen, K.; and Nevatia, R. 2018. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6576–6585.
- Gbagbe, K. F.; Cabrera, M. A.; Alabbas, A.; Alyunes, O.; Lykov, A.; and Tsetserukou, D. 2024. Bi-VLA: Vision-Language-Action Model-Based System for Bimanual Robotic Dexterous Manipulations. *arXiv preprint arXiv:2405.06039*.
- Gelada, C.; Kumar, S.; Buckman, J.; Nachum, O.; and Bellemare, M. G. 2019. Deepmdp: Learning continuous latent space models for representation learning. In *International conference on machine learning*, 2170–2179. PMLR.
- Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, 1861–1870. Pmlr.
- Han, W.; Guo, D.; Xu, C.-Z.; and Shen, J. 2024. Dme-driver: Integrating human decision logic and 3d scene perception in autonomous driving. *arXiv preprint arXiv:2401.03641*.
- Hansen, N.; Jangir, R.; Sun, Y.; Alenyà, G.; Abbeel, P.; Efros, A. A.; Pinto, L.; and Wang, X. 2020. Self-supervised policy adaptation during deployment. *arXiv preprint arXiv:2007.04309*.
- Hu, S.; Fang, Z.; Fang, Z.; Deng, Y.; Chen, X.; Fang, Y.; and Kwong, S. 2024. AgentsCoMerge: Large Language Model Empowered Collaborative Decision Making for Ramp Merging. *arXiv preprint arXiv:2408.03624*.
- Huang, X.; Liu, W.; Chen, X.; Wang, X.; Wang, H.; Lian, D.; Wang, Y.; Tang, R.; and Chen, E. 2024. Understanding the planning of LLM agents: A survey. *arXiv preprint arXiv:2402.02716*.
- Huang, X.; Zhao, C.; and Wu, Z. 2023. A video anomaly detection framework based on appearance-motion semantics representation consistency. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.
- Huang, Y.; Peng, P.; Zhao, Y.; Xu, H.; Geng, M.; and Tian, Y. 2023a. Hierarchical adaptive value estimation for multi-modal visual reinforcement learning. *Advances in Neural Information Processing Systems*, 36: 46724–46736.
- Huang, Y.; Peng, P.; Zhao, Y.; Zhai, Y.; Xu, H.; and Tian, Y. 2023b. Simoun: Synergizing Interactive Motion-appearance Understanding for Vision-based Reinforcement Learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 176–185. IEEE.
- Kim, T. S.; Jones, J.; and Hager, G. D. 2021. Motion guided attention fusion to recognize interactions from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13076–13086.
- Kiran, B. R.; Sobh, I.; Talpaert, V.; Mannion, P.; Al Sal-lab, A. A.; Yogamani, S.; and Pérez, P. 2021. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6): 4909–4926.
- Lample, G.; and Chaplot, D. S. 2017. Playing FPS games with deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.
- Laskin, M.; Srinivas, A.; and Abbeel, P. 2020. Curl: Contrastive unsupervised representations for reinforcement learning. In *International conference on machine learning*, 5639–5650. PMLR.
- Lee, K.; Seo, Y.; Lee, S.; Lee, H.; and Shin, J. 2020. Context-aware dynamics model for generalization in model-based reinforcement learning. In *International Conference on Machine Learning*, 5757–5766. PMLR.
- Liang, T.; Xie, H.; Yu, K.; Xia, Z.; Lin, Z.; Wang, Y.; Tang, T.; Wang, B.; and Tang, Z. 2022. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35: 10421–10434.
- Liang, X.; Wang, T.; Yang, L.; and Xing, E. 2018. Cirl: Controllable imitative reinforcement learning for vision-based self-driving. In *Proceedings of the European conference on computer vision (ECCV)*, 584–599.
- Liu, R.; Wu, Z.; Yu, S.; and Lin, S. 2021. The emergence of objectness: Learning zero-shot segmentation from videos. *Advances in neural information processing systems*, 34: 13137–13152.
- Liu, Y.; Liu, J.; Zhao, M.; Yang, D.; Zhu, X.; and Song, L. 2022. Learning appearance-motion normality for video

- anomaly detection. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.
- Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D. L.; and Han, S. 2023. Bevfusion: Multi-task multi-sensor fusion with unified bird’s-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, 2774–2781. IEEE.
- Ma, R.; Luijckx, J.; Ajanovic, Z.; and Kober, J. 2024. ExploRLLM: Guiding Exploration in Reinforcement Learning with Large Language Models. *arXiv preprint arXiv:2403.09583*.
- Nair, A. V.; Pong, V.; Dalal, M.; Bahl, S.; Lin, S.; and Levine, S. 2018. Visual reinforcement learning with imagined goals. *Advances in neural information processing systems*, 31.
- Pan, M.; Zhu, X.; Wang, Y.; and Yang, X. 2022. Iso-dream: Isolating and leveraging noncontrollable visual dynamics in world models. *Advances in neural information processing systems*, 35: 23178–23191.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Shang, W.; Wang, X.; Srinivas, A.; Rajeswaran, A.; Gao, Y.; Abbeel, P.; and Laskin, M. 2021. Reinforcement learning with latent flow. *Advances in Neural Information Processing Systems*, 34: 22171–22183.
- Simonyan, K.; and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27.
- Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; and Kiela, D. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15638–15650.
- Vemprala, S.; Mian, S.; and Kapoor, A. 2021. Representation learning for event-based visuomotor policies. *Advances in Neural Information Processing Systems*, 34: 4712–4724.
- Wang, B.; Qu, Y.; Jiang, Y.; Shao, J.; Liu, C.; Yang, W.; and Ji, X. 2024a. LLM-Empowered State Representation for Reinforcement Learning. *arXiv preprint arXiv:2407.13237*.
- Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024b. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Wang, Z.; Fan, S.; Huo, X.; Xu, T.; Wang, Y.; Liu, J.; Chen, Y.; and Zhang, Y.-Q. 2024c. Emiff: Enhanced multi-scale image feature fusion for vehicle-infrastructure cooperative 3d object detection. *arXiv preprint arXiv:2402.15272*.
- Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; and Liu, T. 2022. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11686–11695.
- Xiang, H.; Xu, R.; and Ma, J. 2023. HM-ViT: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 284–295.
- Xu, H.; Peng, P.; Tan, G.; Li, Y.; Xu, X.; and Tian, Y. 2024. DMR: Decomposed Multi-Modality Representations for Frames and Events Fusion in Visual Reinforcement Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26508–26518.
- Yarats, D.; Kostrikov, I.; and Fergus, R. 2021. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International conference on learning representations*.
- Yu, T.; Zhang, Z.; Lan, C.; Lu, Y.; and Chen, Z. 2022. Mask-based latent reconstruction for reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 25117–25131.
- Ze, Y.; Hansen, N.; Chen, Y.; Jain, M.; and Wang, X. 2023. Visual reinforcement learning with self-supervised 3d representations. *IEEE Robotics and Automation Letters*, 8(5): 2890–2897.
- Zhang, A.; McAllister, R.; Calandra, R.; Gal, Y.; and Levine, S. 2020. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*.
- Zhao, H.; Gallo, O.; Frosio, I.; and Kautz, J. 2016. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1): 47–57.
- Zheng, R.; Wang, X.; Sun, Y.; Ma, S.; Zhao, J.; Xu, H.; Daumé III, H.; and Huang, F. 2024. Temporal Latent Action-Driven Contrastive Loss for Visual Reinforcement Learning. *Advances in Neural Information Processing Systems*, 36.