Active Vision Reinforcement Learning under Limited Visual Observability

단국대학교

이동규

운동·감각 정책 공동 학습하여 고정 시각 한계 극복 적응적 시각을 도입

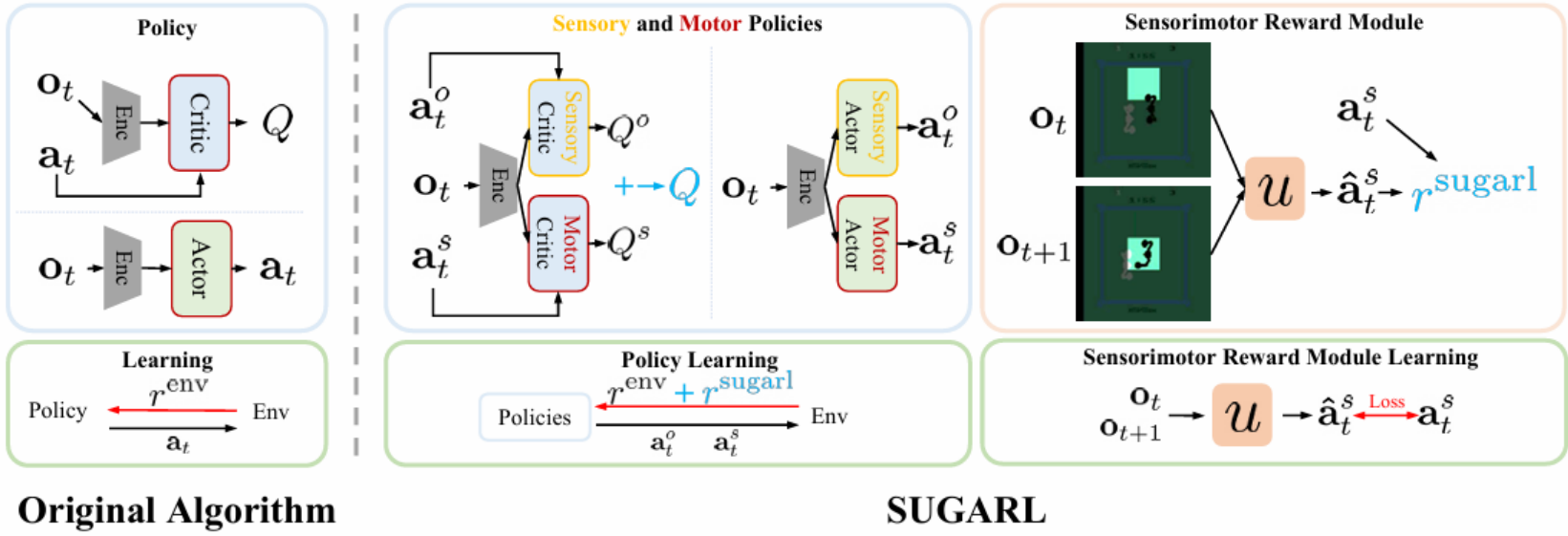
- 기존 Visual RL은 사전 정의된 최적 뷰를 obs space로 제공
- agent가 스스로 적응하지 못함. (예: table-top manipulator는 overhead 고정 카메라 사용)

- Full vision 환경에서는 효과적이지만, 관측 범위가 제한되면 문제 발생.
- Active RL은 embodied agent가 스스로 시각 정보를 선택·조정하며 새로운 관찰을 얻음.

- 인간이 눈과 손을 분리된 시스템으로 가지고도 학습 과정에서 함께 조율
- 모델도 분리 설계 + 공동 학습을 통해 최적 조정 전략을 학습.

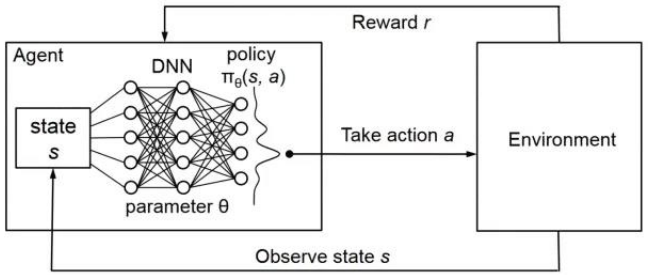
- task return 극대화 + 관측 제어를 동시에 달성하기 위해 motor policy + sensory policy를 분리해 학습.
- motor policy는 깔끔한 시각 입력을 원함
- sensory policy는 motor action을 고려해 최적의 관찰을 선택해야 함.

DQN, SAC의 변형
Sensor, Motor 분리
동시 학습의 장점 활용



- Motor Policy, Sensory Policy 분리 (π^s, π^o)
- Sensorimotor Reward 도입 r^{sugarl}
- DQN : 2종류의 Q-value 출력 Q^s, Q^o
- SAC : 확장된 Head
- Joint Learning of Motor and Sensory Polices

DQN 출력 확장
Q-value 결합 (단순 합)
Reward 추가



DQN

$$\mathcal{L}_i^Q(\theta_i) = \mathbb{E}_{(\mathbf{o}_t, \mathbf{a}_t^s, \mathbf{a}_t^o)\sim\mathcal{D}}\left[\left(y_i - \left(Q_{\theta_i}^s(\mathbf{o}_t, \mathbf{a}_t^s) + Q_{\theta_i}^o(\mathbf{o}_t, \mathbf{a}_t^o)\right)\right)^2\right]$$

$$y_i = \mathbb{E}_{\mathbf{o}_{t+1}}\left[r_t^{\text{env}} + \beta r_t^{\text{sugarl}} + \gamma\left(\max_{\mathbf{a}_{t+1}^s} Q_{\theta_{i-1}}^s(\mathbf{o}_{t+1}, \mathbf{a}_{t+1}^s) + \max_{\mathbf{a}_{t+1}^o} Q_{\theta_{i-1}}^o(\mathbf{o}_{t+1}, \mathbf{a}_{t+1}^o)\right)\right],$$

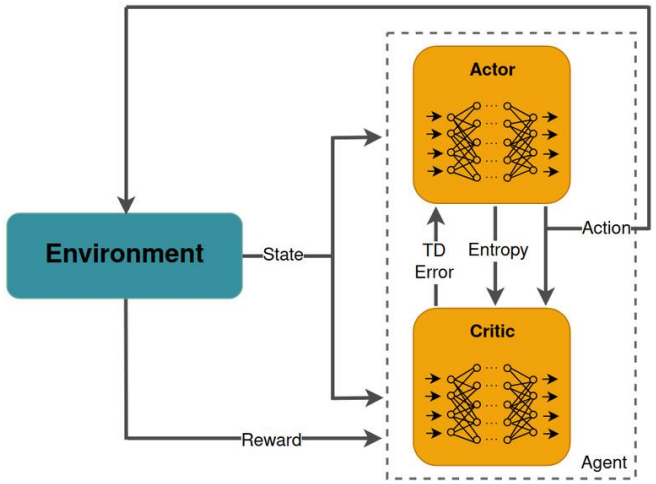
- Network를 이용해 Q 함수를 근사
- 이 결과 값을 가지고 action을 선택
- Network update는 오차제곱합을 최소화하도록
- 학습 시 환경에서의 reward와 sensorimotor reward(가중치 곱)를 단순 합
- Q 단순 합하여 통합 학습 및 기존의 손실함수 형태를 따르도록 함

SAC : Background

- SAC는 Return을 극대화하면서도 policy의 entropy를 최대화하는 모델
- Entropy $\mathcal{H}(X) = \mathbb{E}\left[-\log p(X)\right] = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$
- State Value $V = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^{\pi}(s,a) - \log\pi(a|s)\right] = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^{\pi}(s,a)\right] + \mathcal{H}(\pi(\cdot|s))$
- Objective fun (v 학습) $J_V(\psi) = \mathbb{E}_{s_t\sim\mathcal{D}}\left[\frac{1}{2}\left(V_\psi(s_t) - \mathbb{E}_{a_t\sim\pi_\phi}\left[Q_\theta(s_t,a_t) - \log\pi_\phi(a_t|s_t)\right]\right)^2\right]$
- Objective fun (Q 학습) $J_Q(\theta) = \mathbb{E}_{(s_t,a_t)\sim\mathcal{D}}\left[\frac{1}{2}\left(Q_\theta(s_t,a_t) - \hat{Q}_\theta(s_t,a_t)\right)^2\right]$ where $\hat{Q}_\theta(s_t,a_t) = r(s_t,a_t) + \gamma\,\mathbb{E}_{s_{t+1}\sim p(\cdot|s_t,a_t)}\left[V_{\bar{\psi}}(s_{t+1})\right]$
- Objective fun (policy 학습) $J_\pi(\phi) = \mathbb{E}_{s_t\sim\mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi_\phi(\cdot\mid s_t)\,\middle\|\,\frac{\exp(Q_\theta(s_t,\cdot))}{Z_\theta(s_t)}\right)\right]$
- KL Divergence의 정의를 사용하여 근사적으로 표현

$$\mathrm{D}_{\mathrm{KL}}(p\,\|\,q) = \mathbb{E}_{X\sim p(x)}\left[\log\frac{p(X)}{q(X)}\right] \qquad J_\pi(\phi) \approx \mathbb{E}_{s_t\sim\mathcal{D}}\left[\mathbb{E}_{a_t\sim\pi_\phi(\cdot|s_t)}\left[\log\pi_\phi(a_t|s_t) - Q_\theta(s_t,a_t)\right]\right]$$

SAC head 확장
Q-value 결합 (단순 합)
Reward 가중치 곱



SAC

$$\mathcal{L}^V(\psi) = \mathbb{E}_{\mathbf{o}_t \sim \mathcal{D}} \Big[\frac{1}{2} \big(V_\psi^s(o_t) + V_\psi^o(o_t) - \mathbb{E}_{\mathbf{a}_t^s \sim \pi_\phi^s, \mathbf{a}_t^o \sim \pi_\phi^o} \big[Q_\psi^s(\mathbf{o}_t, \mathbf{a}_t^s) + Q_\psi^o(\mathbf{o}_t, \mathbf{a}_t^o) - \log \pi_\phi^s(\mathbf{a}_t^s | o_t) - \log \pi_\phi^o(\mathbf{a}_t^o | o_t) \big] \big)^2 \Big],$$

and the actor loss $\mathcal{L}^\pi$ is

$$\mathcal{L}^\pi(\phi) = \mathbb{E}_{\mathbf{o}_t \sim \mathcal{D}} \big[\log \pi_\phi^s(\mathbf{a}_t^s | \mathbf{o}_t) + \log \pi_\phi^o(\mathbf{a}_t^o | \mathbf{o}_t) - Q_\psi^s(\mathbf{o}_t, \mathbf{a}_t^s) - Q_\psi^o(\mathbf{o}_t, \mathbf{a}_t^o) \big],$$

- Sensor와 Motor 각각에 대한 head 존재
- 총 4개의 head이나, 학습은 2개씩 한번에 진행
- Network update 시 기존 SAC의 학습 방식을 활용
- 학습 시 환경에서의 reward와 sensorimotor reward(가중치 곱)를 단순 합
- Q 단순 합하여 기존의 손실함수 형태를 따르도록 함

Module에서 관측 평가 Reward 생성 Module은 단독 학습

Sensorimotor Reward

- Action에 대해서는 환경의 Reward가 평가
- 감각(시각인식)에 대한 적절한 피드백을 위해 Sensorimotor reward 사용
- Sensorimotor reward module에서 할당되는 reward

Sensorimotor Reward Module

- Transition이 주어졌을 때 obs만 가지고 action을 추론하는 Module (o_t, a_t^s, o_{t+1})
- Module이 잘 학습되었다면, 높은 Prediction error는 잘못된 시각 조정을 뜻함

$$r_t^{sugarl} = -(1 - p(a_t^s | o_t, o_{t+1}; u_\xi))$$

$$\mathcal{L}^u(\xi) = \mathbb{E}_{o_t, o_{t+1}, a_t^s \sim \mathcal{D}} \left[\text{Error}(a_t^s - u_\xi(o_t, o_{t+1})) \right]$$

가중치는 Env Reward를 참조하여 구성됨 영향력 조정

Sensorimotor Reward Weight

- Reward combine 시 sensorimotor reward에 가중치가 곱해짐

- 환경에서 나오는 reward와 유사한 scale을 가지도록
- 예) Env Reward: 500, Sensorimotor Reward: 5

- 가중치는 Env Reward의 산술 평균으로 사전에 정의된 값
- 모든 관측 가능한 환경에서 baseline agent 활용 or 환경 정보

$$r_t = r_t^{env} + \beta r_t^{sugarl}$$

$$\beta = E_\tau[\sum_{t=1}^{T} r_t^{env}/T]$$

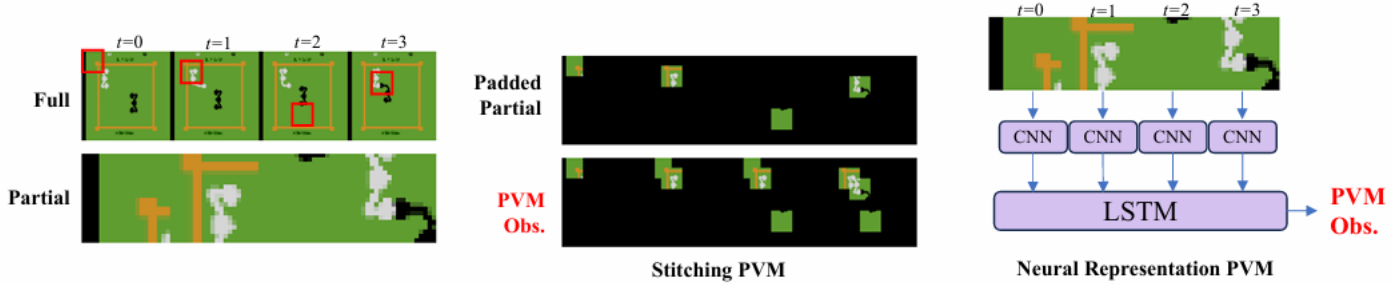
PVM을 사용하여 부분 관측 영역 확장 인간 눈의 잔상 효과 반영

Persistence-of-Vision Memory

- 인간 눈의 잔상에서 영감을 받음
- 효과적으로 관측 가능한 영역을 확장함

$$PVM(o_t) = f(o_{t-B+1}, ..., o_t)$$

- 이전 B step만큼 buffer에 저장
- 이것들을 하나의 PVM obs로 통합하여 기존의 obs를대체함
- Stitching PVM에서는 부분 관찰들(partial observations)을 퍼즐처럼 결합
- Neural Representation PVM 등 을 결합함수로 사용할 수 있음

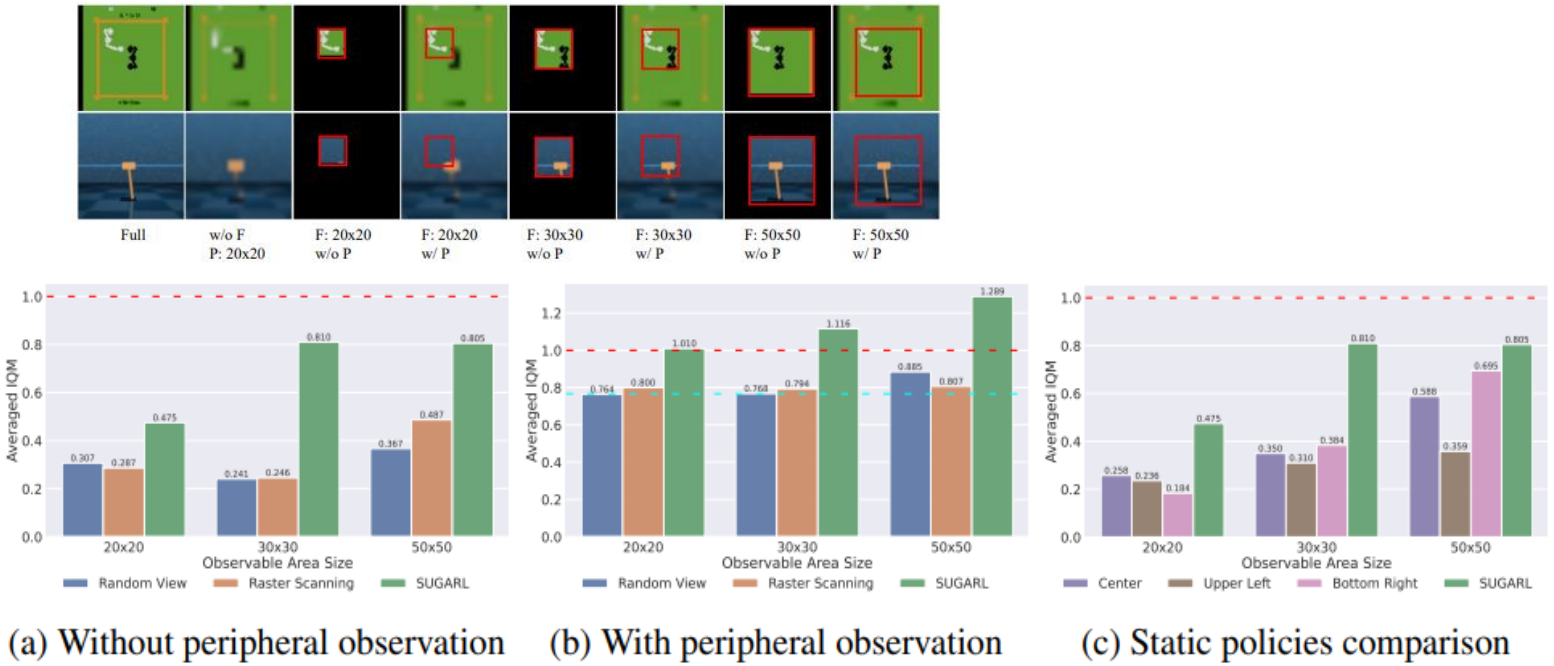


능동적 시각 선택 중요성
동시 학습의 중요성
PVM 종류에 따른 차이

Approach	Wipe	Door	NutAssemblySquare	Lift	Stack
SUGARL-DrQ (Stacking PVM)	56.0	274.8	78.0	79.2	12.7
SUGARL-DrQ (LSTM PVM)	58.5	266.9	**108.6**	88.8	31.5
SUGARL-DrQ (3D Transformation+LSTM PVM)	**74.1**	**291.0**	65.2	87.5	32.4
SUGARL-DrQ w/o Joint Learning	43.6	175.4	58.0	107.2	12.0
SUGARL-DrQ w/o PVM	52.8	243.3	37.9	55.6	7.7
Single Policy	12.4	22.8	8.42	10.7	0.53
DrQ w/ Object Detection (DETR)	15.2	43.1	54.8	15.4	7.5
DrQ w/ End-to-End Attention	14.2	141.4	28.5	33.0	13.6
Eye-in-hand View (hand-coded, moving camera)	16.1	114.6	102.9	**233.9**	**73.0**
Front View (hand-coded, fixed camera)	49.4	240.6	39.6	69.0	13.8
Agent View (hand-coded, fixed camera)	12.7	190.3	49.9	122.6	14.7
Side View (hand-coded, fixed camera)	25.9	136.2	34.5	56.6	12.8

- 5가지 과제: Lift, Stack, NutAssemblySquare, Door, Wipe.
- SUGARL이 항상 베이스라인보다 우수함
- 특히 난이도 높은 과제(Wipe, Door, NutAssemblySquare)에서 최고의 성능을 달성.
- Hand-coded views 보다도 대부분 상황에서 더 나은 결과.
- Persistence-of-Vision Memory (PVM) 디자인 비교
  - 3D Transformation + LSTM > LSTM > Stacking 순으로 성능이 좋음.
  - 공간 정렬 + 시간 결합을 모두 고려하는 방식이 가장 효과적임.

능동적 시각 선택 중요성
주변시 여부에 따른 차이
제한된 시야에서 우수성



(a) Without peripheral observation (b) With peripheral observation (c) Static policies comparison

- Peripheral observation 없음 vs 있음.
- foveal resolution: 20×20, 30×30, 50×50.
- 2D 환경에서 능동적 시각 선택과 기타 선택 사이의 비교.
- Peripheral observation이 있을 때는 full observation baseline보다 우수
- Static policies와 비교

큰 관측 영역(50×50)에서는 격차가 작으나, 작은 영역에서는 SUGARL이 큰 폭으로 우수.
능동적 정책이 제한된 시야에서 특히 유리함.

Action 분리 필요성 SAC 성능 평가 2D 연속 행동공간에서 우수함 증명

Table 2: Evaluation results on different conditions and algorithm backbones

(a) Action modeling

Model	20	30	50
SUGARL (abs)	**0.475**	**0.810**	0.805
SUGARL (rel)	0.367	0.745	**0.945**
Single Policy	0.132	0.222	0.171

(b) Train more steps

Steps	Model	20	30	50
1M	SUGARL	**0.475**	**0.810**	**0.805**
	Single Policy	0.132	0.222	0.171
5M	SUGARL	**1.170**	**1.121**	**1.553**
	Single Policy	0.332	0.640	1.145

(c) SUGARL with SAC

Model	20	30	50
SUGARL	**0.424**	**0.730**	**0.785**
SUGARL w/o r^{sugarl}	0.300	0.307	0.504
SAC-raster scanning	0.117	0.195	0.136
SAC-random view	0.155	0.104	0.134

(d) Different PVMs

Model	20	30	50
Stitching PVM	**0.475**	**0.815**	**0.810**
LSTM PVM	0.397	0.448	0.470

Table 3: SUGARL on DMC

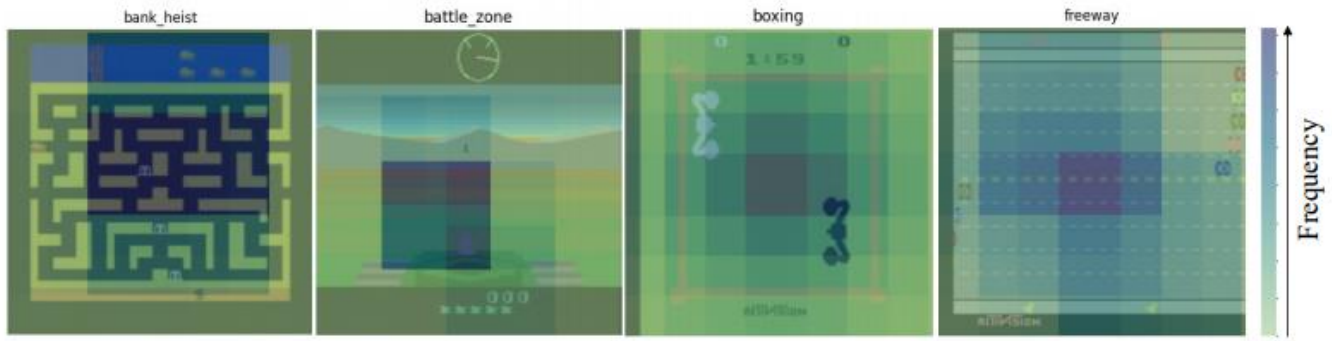
Model	20	30	50
SUGARL-DrQ	**0.686**	**0.717**	**1.052**
DrQ-Single Policy	0.540	0.570	0.776
DrQ-Raster Scanning	0.609	0.566	0.913
DrQ-Random View	0.569	0.591	0.768
SUGARL-DrQ w/o PVM	0.672	0.620	0.930
SUGARL w/o Joint Learning	0.377	0.446	0.355

Table 4: Varing α of SUGARL-SAC.

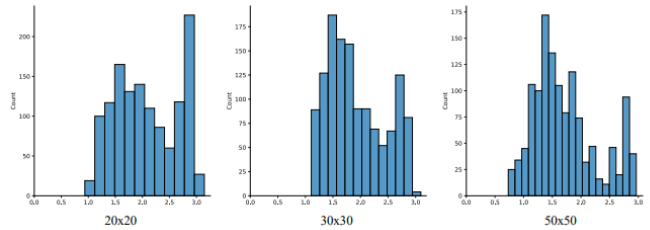
Model	20	30	50
autotune α	0.271	0.358	0.444
fixed-$\alpha = 0.2$	**0.424**	**0.730**	**0.785**

- 단일 정책은 액션 공간이 너무 커져 학습 실패.
- 작은 관측 영역에서는 절대 제어(abs)가 빠른 탐색에 유리.
- 큰 영역에서는 상대 제어(rel)가 더 안정적이라 성능 우수.
- 1M → 5M step으로 늘리면 SUGARL은 성능이 계속 향상.
- 단, max-entropy 기반 SAC에서는 sensory policy가 균일하게 분포되지 않아, entropy weight를 줄여야 안정적임.
- exploration 특성 제어
- SUGARL-DrQ가 Single Policy, Raster Scanning, Random View보다 성능 우수. (2D)

Sensory Policy
환경별 적응
특정 부분에 집중하는 경향성



Model	r^{sugarl}	Joint Learning	PVM	Reward Balance (β)	IQM
Random View			✓		0.367
SUGARL = Base RL algorithm					-
+Naive positive intrinsic reward	positive				0.281
+Joint learning	positive	✓			0.322
Positive → negative r^{sugarl}	negative	✓			0.360
+PVM	negative	✓	✓		0.423
+Reward Balance	negative	✓	✓	✓	**0.805**
SUGARL w/o r^{sugarl}		✓	✓		0.421
SUGARL w/o r^{sugarl} and w/o PVM		✓			0.231



- 학습된 Sensory Policy가 Task에 따라 고정된 영역을 더 많이 보고 동적인 환경은 전반적으로 보는 경향이 있음
- KL-divergence 분석: 학습된 정책은 항상 uniform distribution과 차이를 보이며 특정 영역을 선호.
- 관측 크기가 커질수록 분포는 균등에 가까워지지만, 여전히 집중 영역 유지.
- Ablation Study 결과 PVM, Negative Reward, Reward Balance, Joint Learning 유효성 입증