Information-Theoretic State Space Model for Multi-View Reinforcement Learning

HyeongJoo Hwang, Seokin Seo, Youngsoo Jang, Sungyoon Kim

Geon-Hyeong Kim, Seunghoon Hong, Kee-Eung Kim

KAIST

Multi-View Reinforcement Learning (MVRL)

- **Goal**: Learning an optimal policy from multi-view observations.
  - Practical problem (e.g. autonomous cars).

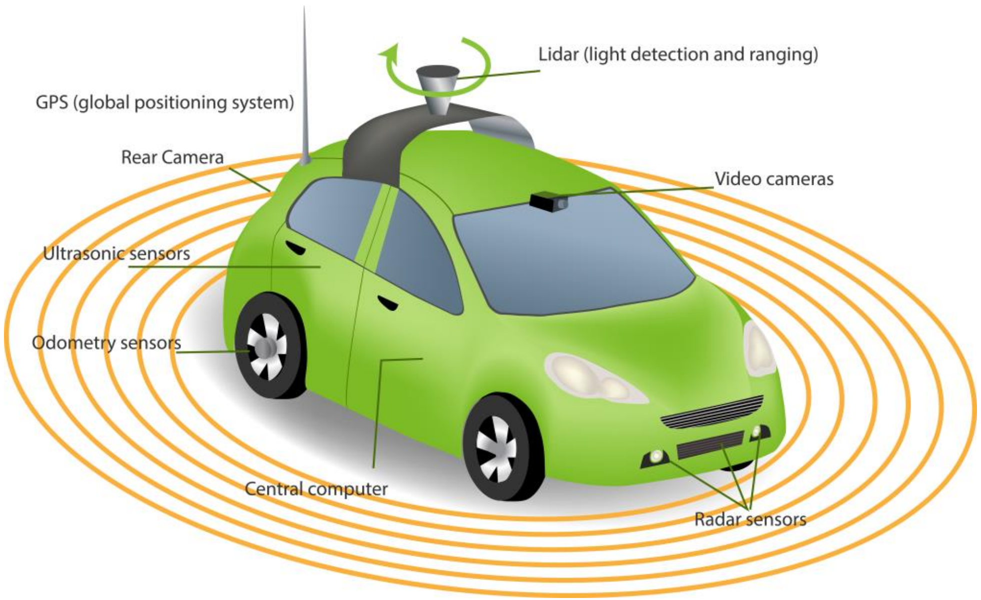


Image from Adobe Stock

Multi-View Reinforcement Learning (MVRL)

- Goal: Learning an optimal policy from multi-view observations.
- **Challenge**: Some views can be randomly missing s.t. $\tilde{o}_t \subseteq \vec{o}_t = \{o_t^v\}_{v=1}^V$.

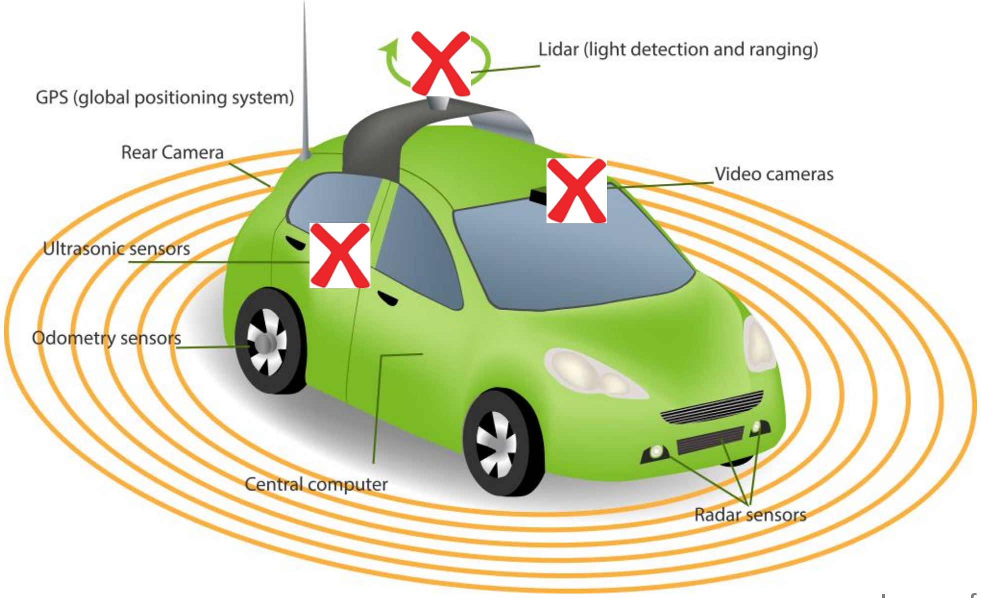


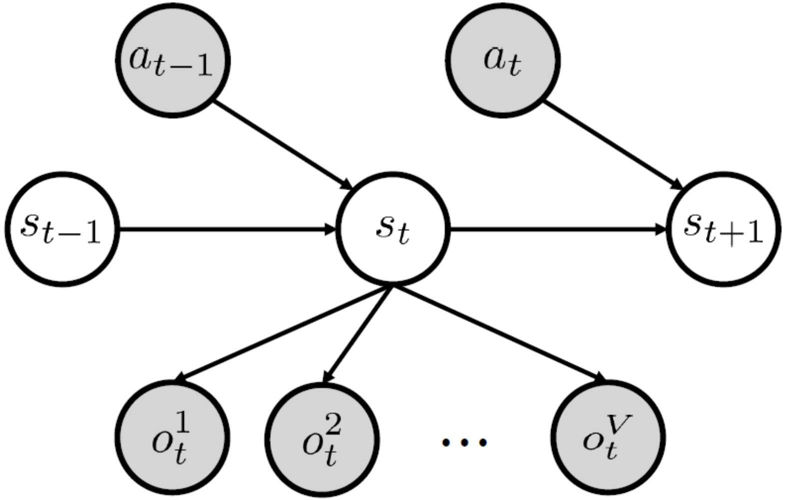
Image from Adobe Stock

Multi-View Reinforcement Learning (MVRL)

- Goal: Learning an optimal policy from multi-view observations.
- Challenge: Some views can be randomly missing s.t. $\tilde{o}_t \subseteq \vec{o}_t = \{o_t^v\}_{v=1}^V$.
- **Solution**: Learn the latent state robust to missing views!

Desiderata of representation for MVRL

1. **Informative for optimal control**
   - Any ideal representation should be informative for the optimal control as much as “state”.
   - Capturing the underlying transition dynamics is the key to learn the informative latent state.

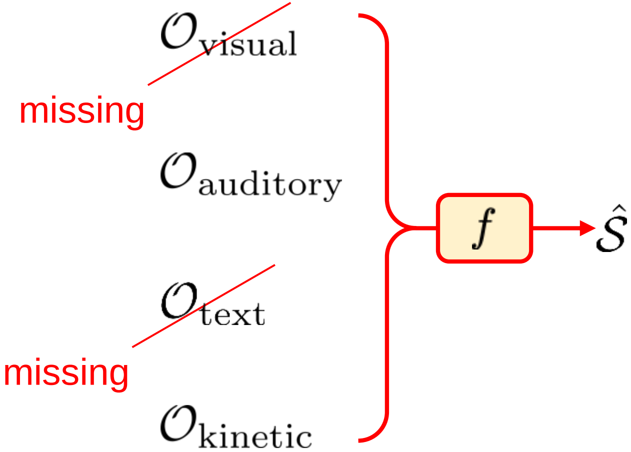


Desiderata of representation for MVRL

2. Robust to missing views

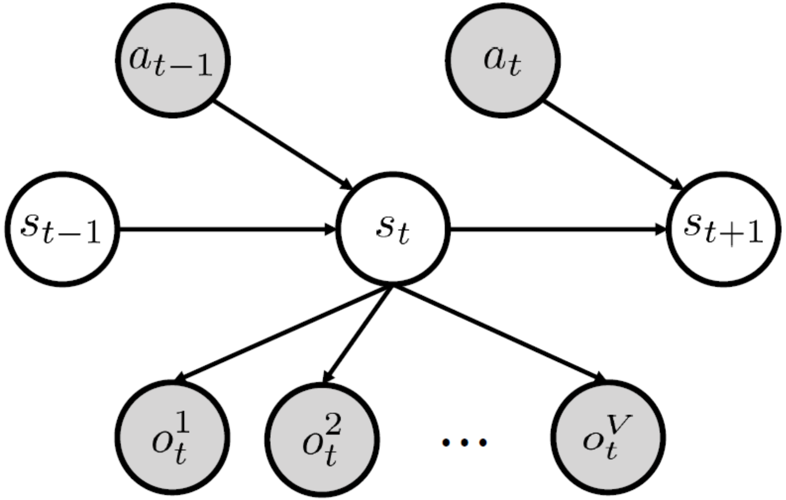
- The representation should be able to handle missing views in test time.
  - Handling missing views in train time would be even more practical.

- This could be very important in a modularized sensor systems



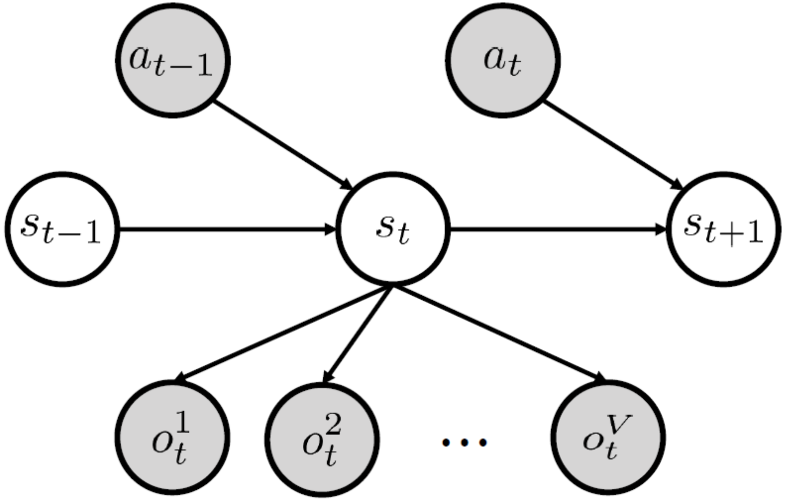
Relating states, actions, and multi-view observations

- To learn the underlying dynamics, we need to note:
  - $\langle S_{t-1}, A_{t-1} \rangle$ generates S_t.
  - S_t generates $O_t^1, O_t^2, \ldots, O_t^V$.



Relating states, actions, and multi-view observations

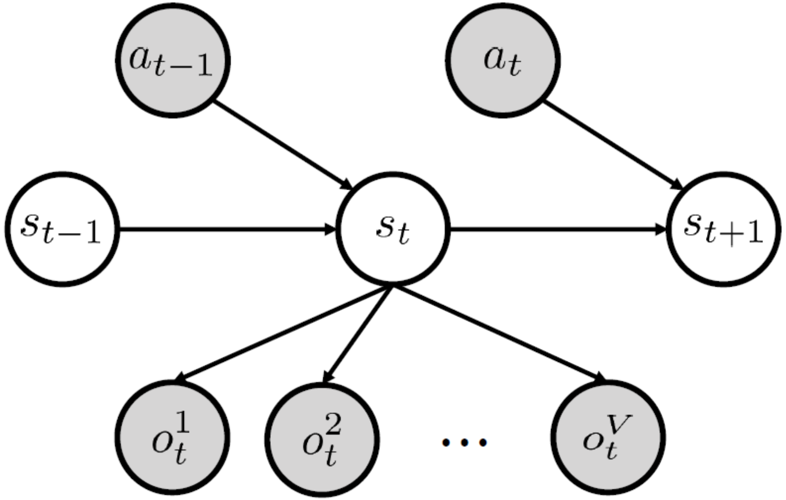
- To learn the underlying dynamics, we need to note:
  - $\langle S_{t-1}, A_{t-1} \rangle$ generates S_t.
  - S_t generates $O_t^1, O_t^2, \dots, O_t^V$.



- Thus, following two properties hold:
  1) There exists strong dependence among $\langle S_{t-1}, A_{t-1} \rangle, O_t^1, \dots, O_t^V$.

Relating states, actions, and multi-view observations

- To learn the underlying dynamics, we need to note:
  - $\langle S_{t-1}, A_{t-1} \rangle$ generates S_t.
  - S_t generates $O_t^1, O_t^2, \dots, O_t^V$.



- Thus, following two properties hold:
  1) There exists strong dependence among $\langle S_{t-1}, A_{t-1} \rangle, O_t^1, \dots, O_t^V$.
  2) However, $\langle S_{t-1}, A_{t-1} \rangle, O_t^1, \dots, O_t^V$ are conditionally independent given S_t.

Information theoretic approach

- **Total Correlation (TC)** measures dependence among multiple RVs.

$$TC\left(A,B,C\right) \triangleq D_{KL}\left[p\left(A,B,C\right) \| p(A)p(B)p(C)\right]$$

$$TC\left(A,B,C \mid Z\right) \triangleq \mathbb{E}_{p(Z)}\left[D_{KL}\left[p\left(A,B,C \mid Z\right) \| p(A \mid Z)p(B \mid Z)p(C \mid Z)\right]\right]$$

- **Our objective** we train $p_\theta(\hat{s}_t \mid \vec{o}, \hat{s}_{t\text{-}1}, a_{t\text{-}1})$ by maximizing

$$TC_\theta(\langle \hat{S}_{t\text{-}1}, A_{t\text{-}1}\rangle, \vec{O}_t; \hat{S}_t) \triangleq \underbrace{TC_\theta(\langle \hat{S}_{t\text{-}1}, A_{t\text{-}1}\rangle, \vec{O}_t)}_{\text{1) Strong \textbf{dependence} among } \langle S_{t-1}, A_{t-1}\rangle, O_t^1, \dots, O_t^V.} - \underbrace{TC_\theta(\langle \hat{S}_{t\text{-}1}, A_{t\text{-}1}\rangle, \vec{O}_t \mid \hat{S}_t)}_{\text{2) Conditional \textbf{independence} given } S_t.}, \quad (1)$$

where $\vec{o}_t = \{o_t^v\}_{v=1}^V$ denote complete-view observations from V different views.

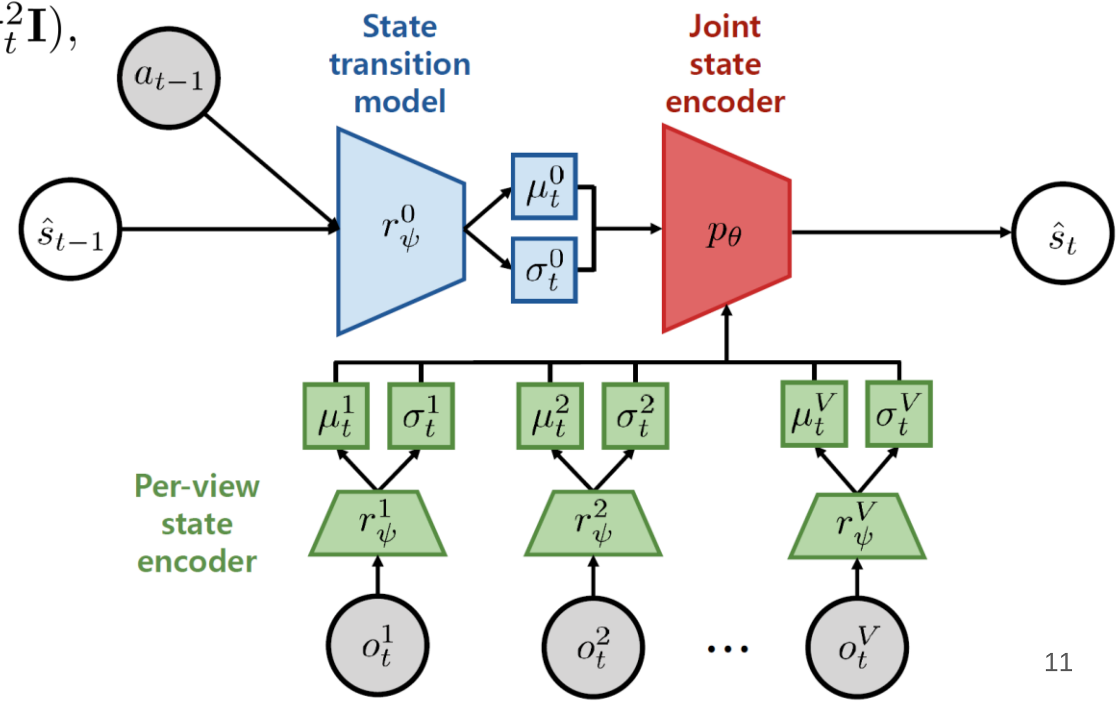
Inverse-Variance Weighted Average

- F2C combines per-view latent states with ($\theta = \{\psi^v\}_{v=0}^V$)

$$p_\theta(\hat{s}_t \mid \vec{o}_t, \hat{s}_{t-1}, a_{t-1}) \triangleq N(\mu_t, \sigma_t^2 \mathbf{I}),$$

$$\text{where } \mu_t \triangleq \frac{\sum_{v=0}^V \mu_t^v / (\sigma_t^v)^2}{\sum_{v=0}^V 1/(\sigma_t^v)^2}$$

$$\text{and } \sigma_t^2 \triangleq \frac{1}{\sum_{v=0}^V 1/(\sigma_t^v)^2}.$$



- Missing views
  can be ignored by $\sigma_t^v = \infty$.

  However…

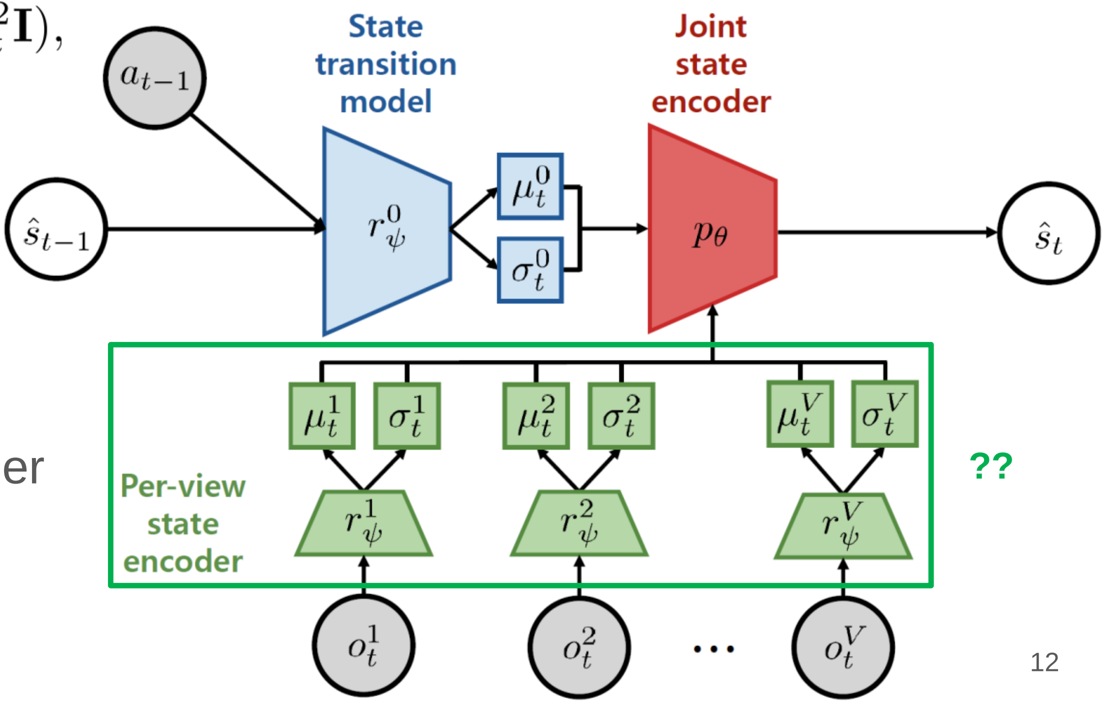
Inverse-Variance Weighted Average

- F2C combines per-view latent states with ($\theta = \{\psi^v\}_{v=0}^V$)

$$p_\theta(\hat{s}_t \mid \vec{o}_t, \hat{s}_{t-1}, a_{t-1}) \triangleq N(\mu_t, \sigma_t^2 \mathbf{I}),$$

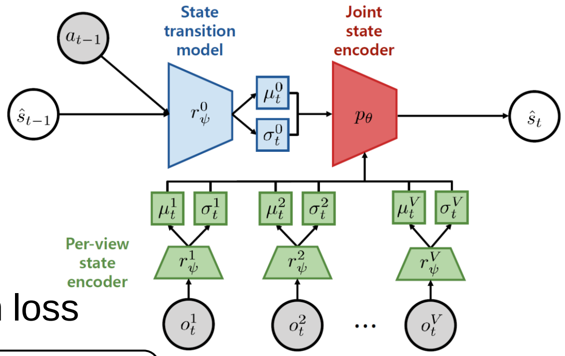
$$\text{where } \mu_t \triangleq \frac{\sum_{v=0}^V \mu_t^v/(\sigma_t^v)^2}{\sum_{v=0}^V 1/(\sigma_t^v)^2}$$

$$\text{and } \sigma_t^2 \triangleq \frac{1}{\sum_{v=0}^V 1/(\sigma_t^v)^2}.$$



- 2 issues:
  1) would each per-view encoder extract meaningful info?
  2) balanced dependence?

Connection to State Space Models



- Lower bound with single CVIB:

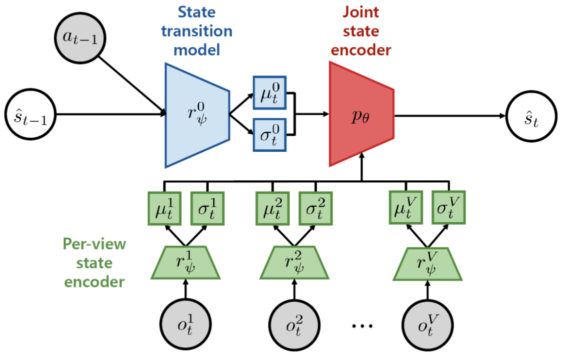
$$TC_\theta(\langle \hat{S}_{t\text{-}1}, A_{t\text{-}1}\rangle, \vec{O}_t; \hat{S}_t) \geq \sum_{v=1}^{V} \Big[H(O_t^v) + \overbrace{\mathbb{E}_{p_\theta(o_t^v, \hat{s}_t)}\left[\ln q_\phi\left(o_t^v | \hat{s}_t\right)\right]}^{\text{Reconstruction loss}}\Big] \qquad (2)$$

$$- \underbrace{\mathbb{E}\left[D_{KL}\left[p_\theta\left(\hat{s}_t | \vec{o}_t, \hat{s}_{t\text{-}1}, a_{t\text{-}1}\right) \| r_\psi^0\left(\hat{s}_t | \hat{s}_{t\text{-}1}, a_{t\text{-}1}\right)\right]\right]}_{\text{Conditional Variational Information Bottleneck (CVIB) of transition model}} .$$

- Reduces to the **ELBO** objective for (R)SSMs [2,3,4,5,6] when V=1.

- Limit: No explicit optimization of per-view encoders.
  $\longrightarrow$ Unbalanced dependence: Joint encoder may ignore some views.

Fuse2Control (F2C)



- Lower bound with **multiple CVIBs**:

$$TC_\theta(\langle \hat{S}_{t-1}, A_{t-1}\rangle, \vec{O}_t; \hat{S}_t) \geq \frac{V}{V+1} I_{\text{NCE}}(t;\theta)$$

$$+ \frac{V}{V+1} \sum_{v=1}^{V} \Big[H\left(O_t^v\right) + \overbrace{\mathbb{E}_{p_\theta(o_t^v, \hat{s}_t)}\left[\ln q_\phi\left(o_t^v | \hat{s}_t\right)\right]}^{\text{Reconstruction loss}} \Big] \qquad (3)$$

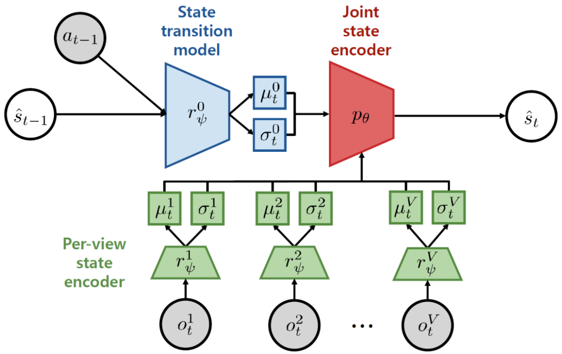
$$- \frac{1}{V+1} \mathbb{E}\left[D_{KL}\left[p_\theta\left(\hat{s}_t | \vec{o}_t, \hat{s}_{t\text{-}1}, a_{t\text{-}1}\right) \| r_\psi^0\left(\hat{s}_t | \hat{s}_{t\text{-}1}, a_{t\text{-}1}\right)\right]\right]$$

$$- \frac{1}{V+1} \sum_{v=1}^{V} \underbrace{\mathbb{E}\left[D_{KL}\left[p_\theta\left(\hat{s}_t | \vec{o}_t, \hat{s}_{t\text{-}1}, a_{t\text{-}1}\right) \| r_\psi^v\left(\hat{s}_t | o_t^v\right)\right]\right]}_{\substack{\text{Conditional Variational Information Bottleneck} \\ \text{(CVIB) of each view}}} .$$

- 2 advantages:

  1) optimize $r_\psi^v\left(\hat{s}_t | o_t^v\right)$ to extract **meaningful info**.

Fuse2Control (F2C)



- Lower bound with **multiple CVIBs**:

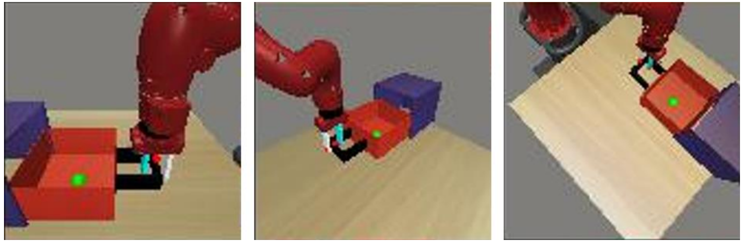
$$
\begin{aligned}
TC_\theta(\langle \hat{S}_{t-1}, A_{t-1}\rangle, \vec{O}_t; \hat{S}_t) \geq \frac{V}{V+1} I_{\mathrm{NCE}}(t;\theta) \\
+ \frac{V}{V+1}\sum_{v=1}^{V}\Big[H\left(O_t^v\right) + \overbrace{\mathbb{E}_{p_\theta(o_t^v, \hat{s}_t)}\left[\ln q_\phi\left(o_t^v|\hat{s}_t\right)\right]}^{\text{Reconstruction loss}}\Big] \quad (3) \\
- \frac{1}{V+1}\mathbb{E}\left[D_{KL}\left[p_\theta\left(\hat{s}_t|\vec{o}_t, \hat{s}_{t\text{-}1}, a_{t\text{-}1}\right) \| r_\psi^0\left(\hat{s}_t|\hat{s}_{t\text{-}1}, a_{t\text{-}1}\right)\right]\right] \\
- \frac{1}{V+1}\sum_{v=1}^{V}\underbrace{\mathbb{E}\left[D_{KL}\left[p_\theta\left(\hat{s}_t|\vec{o}_t, \hat{s}_{t\text{-}1}, a_{t\text{-}1}\right) \| r_\psi^v\left(\hat{s}_t|o_t^v\right)\right]\right]}_{\text{Conditional Variational Information Bottleneck (CVIB) of each view}}.
\end{aligned}
$$

- 2 advantages:
  1) optimize $r_\psi^v\left(\hat{s}_t|o_t^v\right)$ to extract meaningful info.
  2) regularize $p_\theta(\hat{s}_t|\vec{o}, \hat{s}_{t\text{-}1}, a_{t\text{-}1})$ to have **balanced dependence**.

Complex manipulation tasks with 3 camera views

Can F2C be jointly trained with policy under missing-view scenarios?

- Env: Metaworld
  - 3 camera views
    - each view is randomly missing
      with probability 0.5.



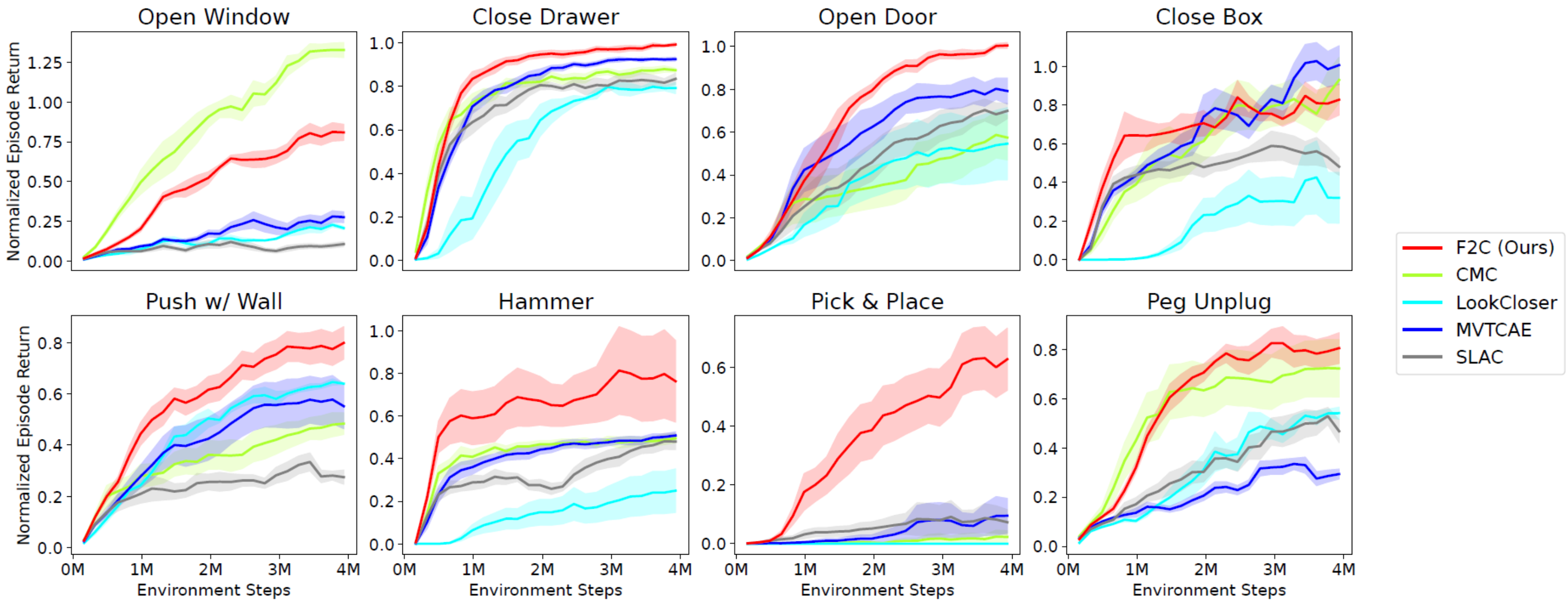
  - Goal: bringing the object close to the goal position in each task.

  - Evaluation protocol
    - jointly train representation and PPO directly from missing-view observation.

Complex manipulation tasks with 3 camera views

Can F2C be jointly trained with policy under missing-view scenarios?

- Env: Metaworld



Thank you!

Summary

(1) Principled extension of MVL to MVRL.

(2) Showed close relationship between TC objective and existing (R)SSMs.

(3) Reformulated TC objective to learn the latent state robust to missing views.