MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Carnegie Mellon University

Causal Representation Learning for Video Understanding

Guangyi Chen

https://chengy12.github.io/

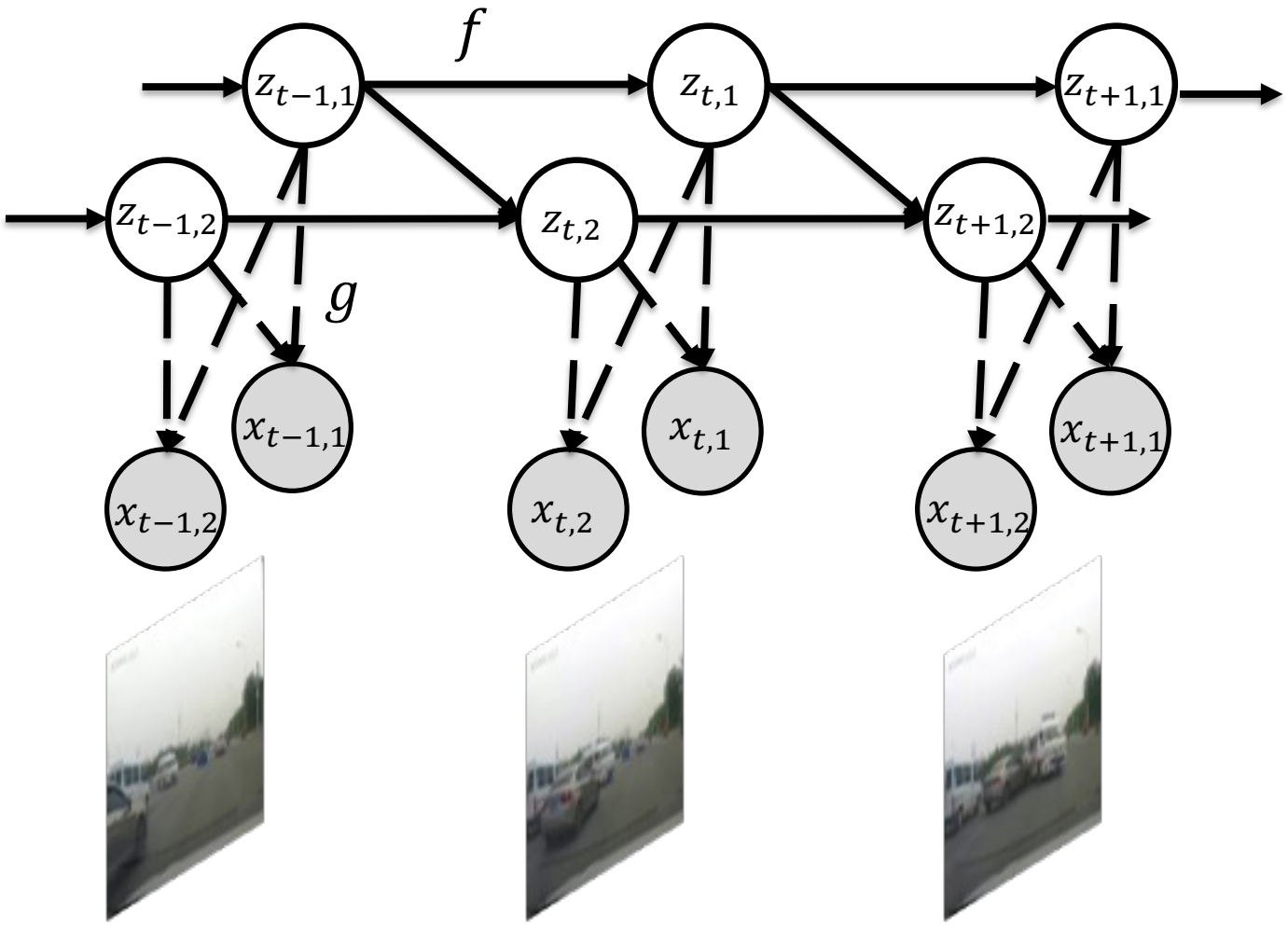
Carnegie Mellon University, Pittsburgh PA, USA

Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

Pacific Causal Inference Conference, 6th July

Causal Representation Learning

- Given the video data, causal representation learning (CRL) aims to recover the data generation process from the observation to obtain the disentangled latent representation.



- Data generation process:

$$x_t = g(z_t)$$

- Latent causal process:

$$z_{it} = f_i(\mathbf{Pa}(z_{it}), \varepsilon_{it})$$

- Representation learning:

$$\hat{z}_t = \hat{g}^{-1}(x_t)$$

- Component-wise identifiability

$$p_{\hat{g},\hat{f},\hat{p}_\epsilon}(x_t) = p_{g,f,p_\epsilon}(x_t)$$
$$\Longrightarrow \quad \hat{g} = g \circ T \circ \pi$$

- Explainable, independently controllable, better transferable

CRL for Video Understanding

- When the learned representation can be identifiable?
- How to learn the causal representation from the video data?
- What can we do for video understanding if causal representation is learned?

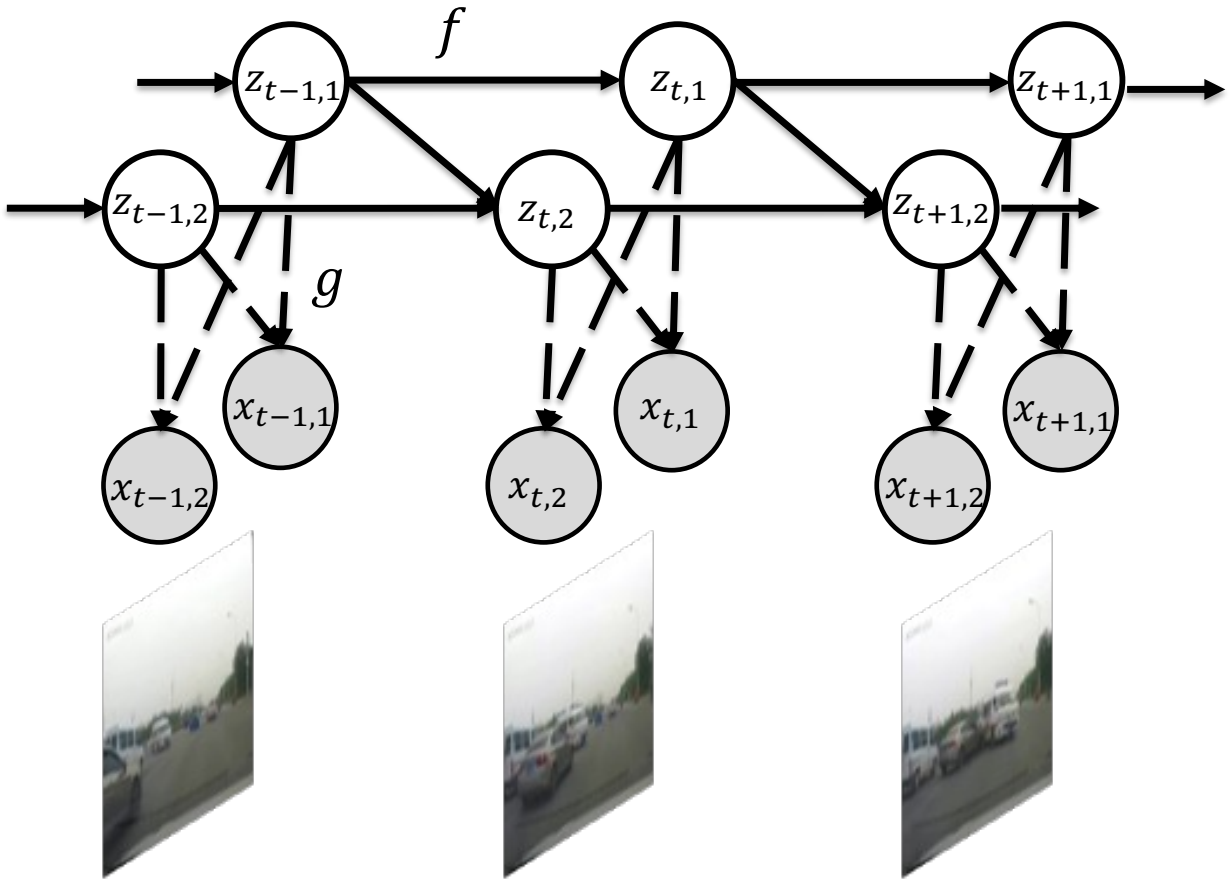
Identification Theorem

↓

Practical Algorithm

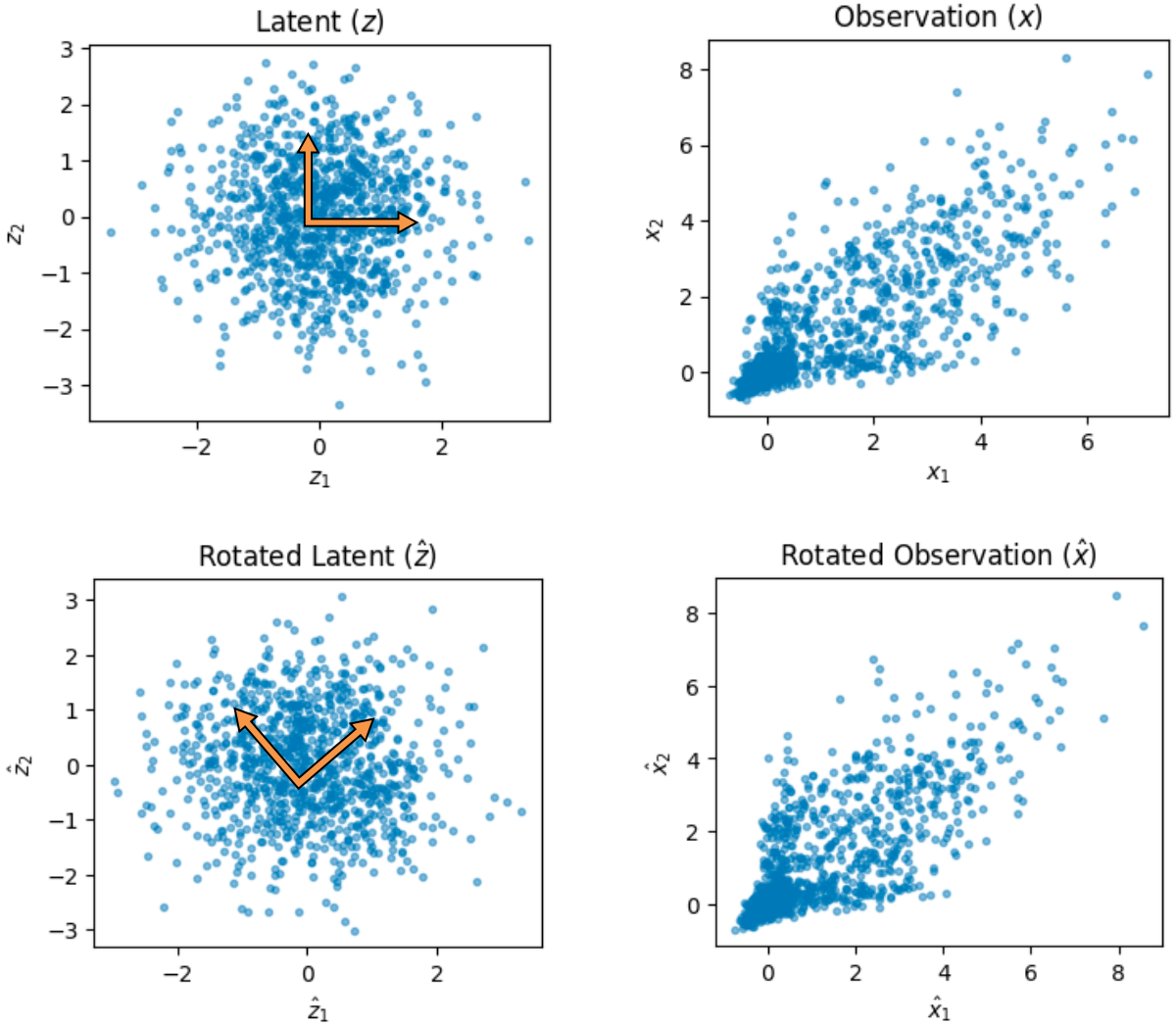
↓

Real-world Application



Causal Representation Learning is Hard

- For the linear Gaussian case, when we add a rotation on the latent variable, the generated observation distribution is unchanged.



- Original generation process

$$z \sim \mathcal{N}(\mathbf{0}, I) \qquad x = g(z)$$

- Adding a rotation on the latent variable doesn't change the observation

$$\hat{z} = \sigma(z) \qquad \hat{x} = g(\hat{z})$$

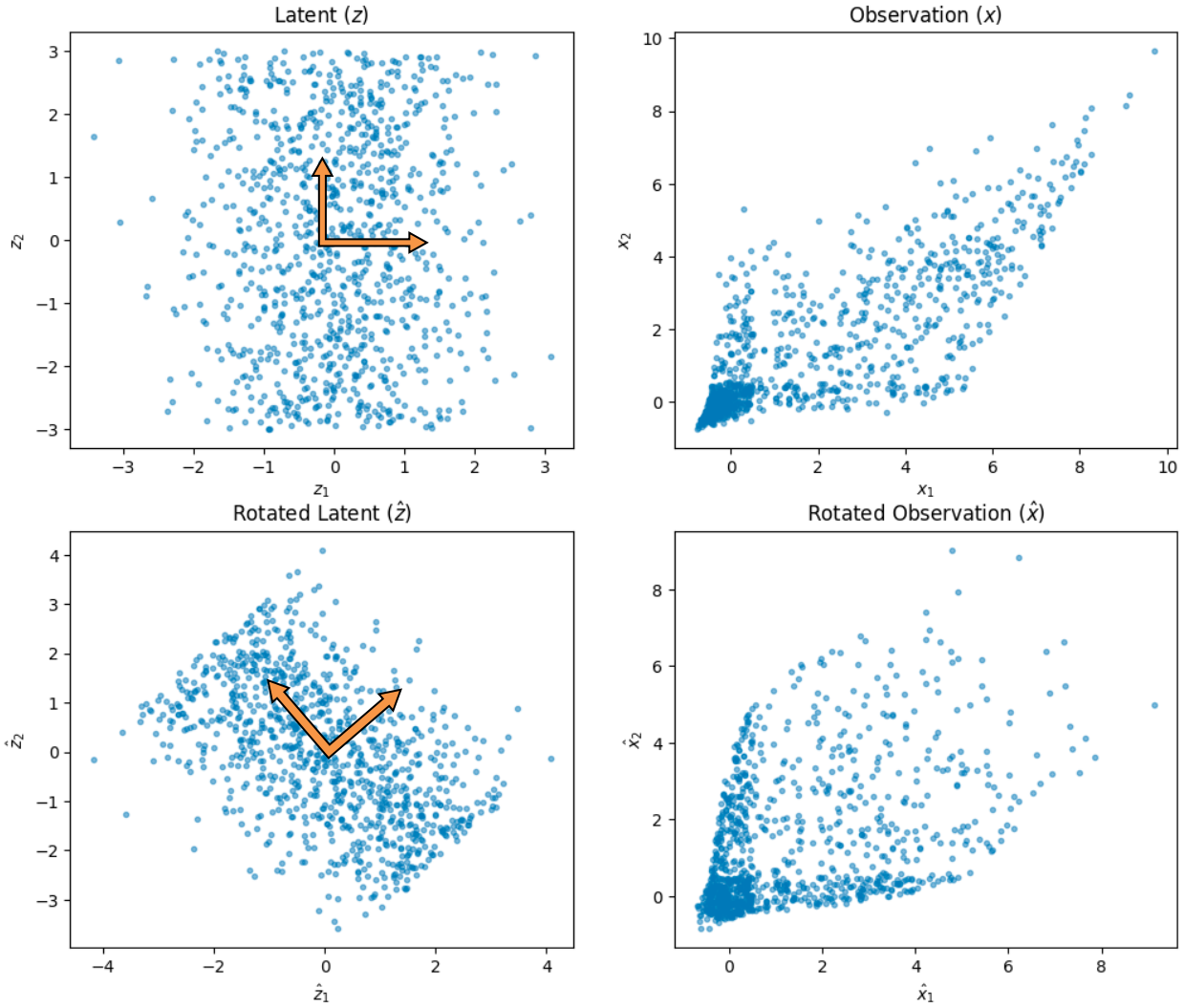
$$p(\hat{z}) = p(\sigma(z)) = p(z)$$

- We may learn the representations with the entanglement

$$\hat{g}^{-1} = g^{-1} \qquad \hat{g}^{-1} = \sigma \circ g^{-1}$$

Causal Representation Learning Is Hard

- For the linear non-Gaussian case, we can recover the latent variables.
- For the non-linear case, we can leverage the auxiliary variables such as the domain index.



- The Non-Gaussian property provides the "changeability" (sufficient change)
- For the more complex non-linear case, we need more "changeability".
- Auxiliary variables, such as labels and domain index, can provide such side information.

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Identification Theorem

Carnegie Mellon University

Generation Process contains the stationary latent temporal dynamic transition and the invertible mixing function:

$$\mathbf{x}_t = \mathbf{g}(\mathbf{z}_t), \quad z_{it} = f_i\left(\mathbf{z}_H, \epsilon_{it}\right).$$

Theorem (Identifiability under Stationary Process). *For a series of observations $\mathbf{x}_t$ and estimated latent variables $\hat{\mathbf{z}}_t$, suppose there exists function $\hat{\mathbf{g}}$ which is subject to observational equivalence,*

$$\mathbf{x}_t = \hat{\mathbf{g}}(\hat{\mathbf{z}}_t).$$

If assumptions

- *(Smooth and Positive Density) the probability density of latent variables is third-order differentiable and positive,*
- *(conditional independence) the components of $\hat{\mathbf{z}}_t$ are mutually independent conditional on $\hat{\mathbf{z}}_H$,*
- *(sufficiency) let $\eta_{kt} \triangleq \log p(z_{kt}|\mathbf{z}_H)$, and*

$$\mathbf{v}_{lt} \triangleq \left(\frac{\partial^2 \eta_{1t}}{\partial z_{1t}\partial z_{l,H}}, \frac{\partial^2 \eta_{nt}}{\partial z_{nt}\partial z_{l,H}}, \frac{\partial^3 \eta_{1t}}{\partial z_{1t}^2\partial z_{l,H}}, \frac{\partial^3 \eta_{nt}}{\partial z_{nt}^2\partial z_{l,H}}\right)^\intercal,$$

for $l = 1, 2, \cdots, n$. For each value of $\mathbf{z}_t$, there exists $2n$ different values of $z_{l,H}$ such that the $2n$ vector functions $\mathbf{v}_{lt} \in \mathbf{R}^{2n}$ are linearly independent,

are satisfied, then $\mathbf{z}_t$ must be a component-wise transformation of a permuted version of $\hat{\mathbf{z}}_t$.

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Identification Theorem

Carnegie Mellon University

- Stationary: The temporal transition function is fixed for the whole sequence.
  Known non-stationary is an opportunity more than a challenge.
  Unknown non-stationary cases cannot be identifiable since it is difficult to distinguish the domain change or variable change.

- Invertibility: The observation contains all information about the latent process.
  If we cannot recover the missing information, we cannot achieve the identification.

- Conditional independence: there are no instantaneous relations among the latent variables.
  If two variables always change jointly, we cannot say what is the unique effect of each.

- Sufficiency: It means that the conditional independent change of the latent variables has sufficient influence and these changes can be captured from the observation.
  Image: requiring domain index or labels
  Video: using the historical state as the side information

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

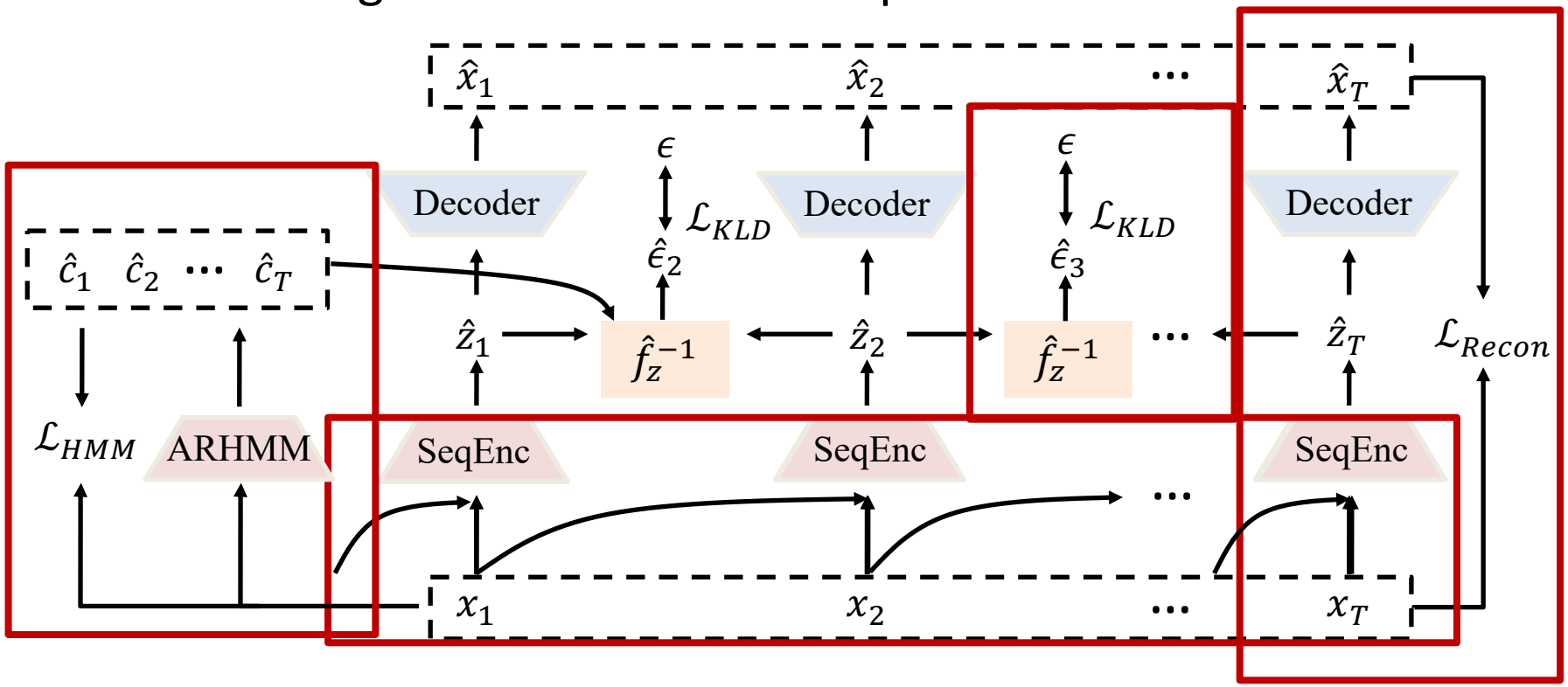
Identification Under More Challenges

Carnegie Mellon University

Challenges	Scenarios	Extra Assumption	Reference
Unknown non-stationary	Bioinformatics, Speech, Volleyball Game	1) Domain index follows the Markov process, 2) Mechanism Sparsity	NCTRL[3] IDEA[4]
Non-invertible mixing function	Occlusion, Video, Motion Blue	There exists context to complete the missing info.	CaRiNG[2]
Instantaneous relations	Skeleton, Stock	Sparse relations	IDOL[5]

Practical Algorithm

- **Auto-Encoder Model**: Estimates the mixing function $x_t = g(z_t)$ and de-mixing function $\hat{z}_t = \hat{g}^{-1}(x_t)$. Invertibility is enforced by reconstruction loss.
- **Prior Network**: Estimates the prior distribution $p(z_t|z_H)$ by learning the inverse dynamics f_z^{-1}, by change of variable $p(z_t|z_H) = p_\varepsilon\left(f_z^{-1}(z_t, z_H)\right)\left|\frac{\partial f_z^{-1}}{\partial z_t}\right|$. By constraining this prior, we encourage the conditional independence.



- **Sequential encoder** to leverage the context to recover the lost information
- **Autoregressive hidden Markov module** to estimate transition matrix of the unknown non-stationary.

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Synthetic Experiments

Carnegie Mellon University

- Apply a 2-layer MLP as the transition function and a random three-layer MLP for mixing.
- Consider both the stationary scenario and non-stationary scenarios. For non-stationary scenarios, there may be only casual dynamic changes or both casual dynamic changes and global observations vary.
- To add changes to the causal dynamic, we vary the values of the first layer of the MLP.
- The global change component is sampled from i.i.d Gaussian distribution whose mean and variance are modulated by domain index.
- Compared with baseline betaVAE, i-VAE/ TCL with independent factors, and LEAP which only models nonstationary noise.
- Mean Correlation Coefficient (MCC) is used to evaluate the learned representation.

Experiment Settings	Method								
	TDRL	LEAP	SlowVAE	PCL	i-VAE	TCL	betaVAE	KVAE	DVBF
Fixed	**0.954 ±0.009**	–	0.411 ±0.022	0.516 ±0.043	–	–	0.353 ±0.001	0.832 ±0.038	0.778 ±0.045
Changing	**0.958 ±0.017**	0.726 ±0.187	0.511 ±0.062	0.599 ±0.041	0.581 ±0.083	0.399 ±0.021	0.523 ±0.009	0.711 ±0.062	0.648 ±0.071
Modular	**0.993 ±0.001**	0.657 ±0.108	0.406 ±0.045	0.564 ±0.049	0.557 ±0.005	0.297 ±0.078	0.433 ±0.045	0.632 ±0.048	0.678 ±0.074

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Application: Video Reasoning

Carnegie Mellon University

- Video reasoning aims to answer the neural language reasoning questions based on the video content, whose challenge lies in understanding the latent causal process.

Counterfactual Inference

Q: Would the accident still happen if there were fewer vehicles on the road?

- ✓ Yes, the road is not congested at the first place, and the accident is not related to the density of the vehicles on the road.
- ✗ No, fewer vehicles would have provided enough space to safely avoid the accident.
- ✗ No, there is no accident.

Introspection

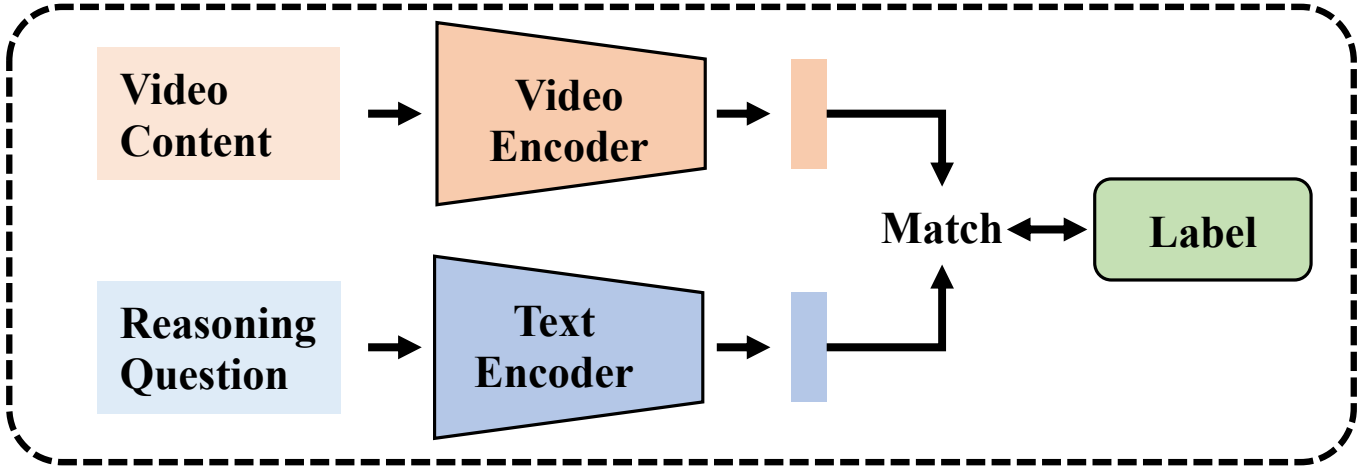
Q: What could have been done to prevent this accident from happening?

- ✗ The accident could have been avoided if the white sedan had slowed down.
- ✗ The accident could have been avoided if the black sedan had changed the lane.
- ✗ The accident could have been prevented if the road is marked clearly.
- ✓ The accident could have been avoided if the white sedan had stayed on its lane.

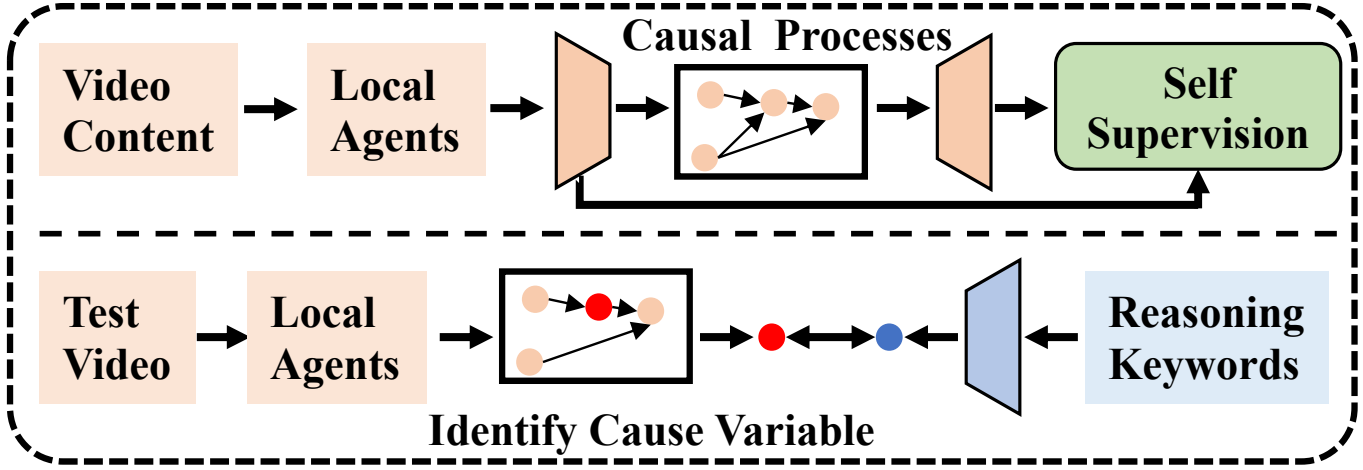
Credited to the SUTD-TrafficQA dataset (Xu. et.al. 2021)

Compared with Existing VideoQA methods

- Once we identify the causal dynamics, we can efficiently conduct video reasoning (such as attribution and counterfactual questions) as a causal inference process.



- Formulate VideoQA as cross-modality matching.
- Rely on question-answer (QA) pairs, Learn relations

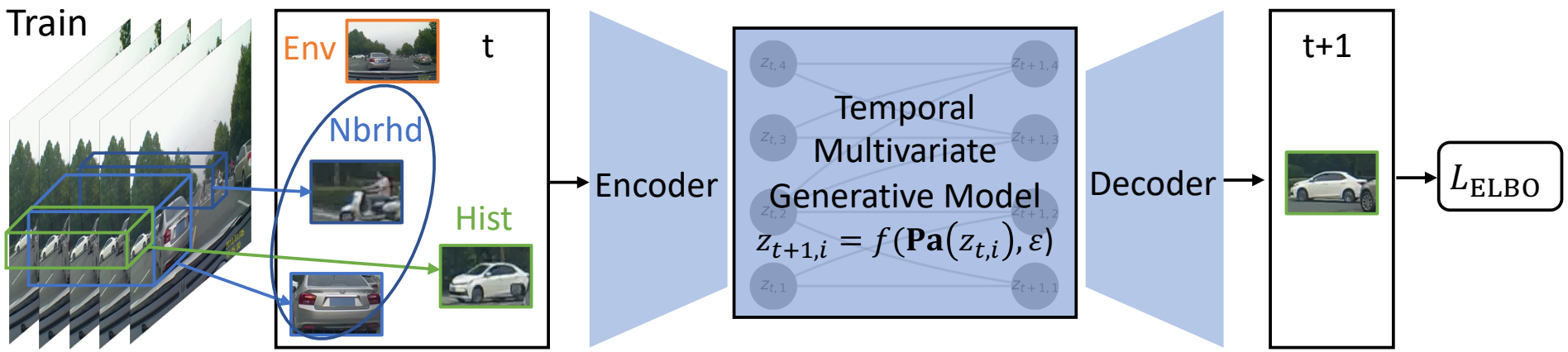


- Don't require the QA pairs.
- Can efficiently answer the attribution and counterfactual questions

Chen et al. LLCP: Learning Latent Causal Processes for Reasoning-based Video Question Answer. ICLR, 2024

Overall Framework

- In the training, we learn the latent causal process from normal videos by self-supervision
- During inference, we first identify the root cause and use it to select answer



Algorithm Details

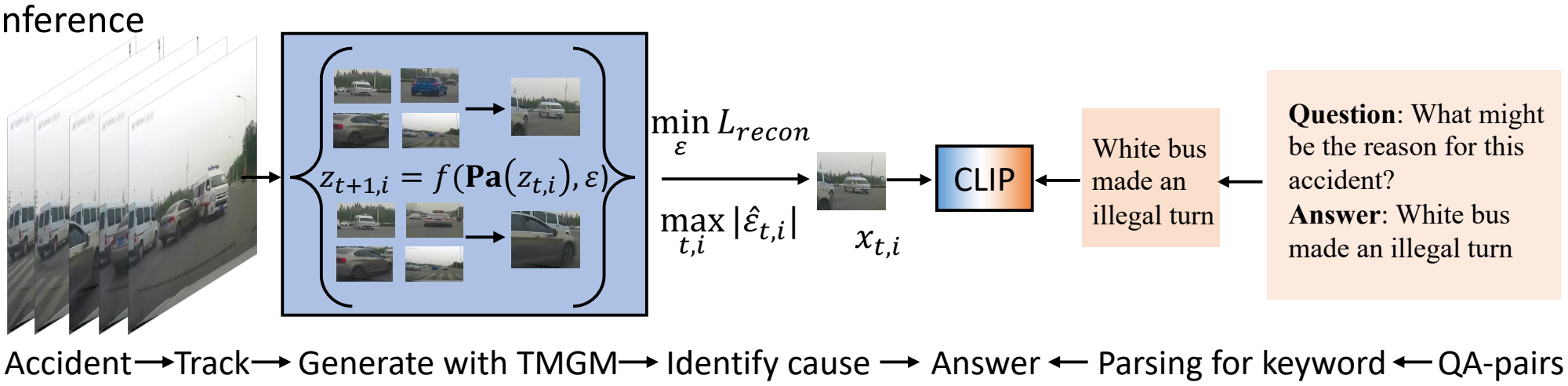
- **Identify root cause**: we find the root cause of abnormal videos by comparing them with learned causal structures. Specifically, we find the variable with a notable shift in the learned local causal processes.

$$\mathbf{x'_C} = \arg\max_{t,i} \left\| \hat{\mathbf{x}}'_{t,i} - \mathbf{x}'_{t,i} \right\|,$$

Or identify the variable that needs outlier noise for reconstruction:

$$\mathbf{x'_C} = \arg\max_{t,i} \hat{\epsilon}_{t,i} \quad , \qquad \hat{\epsilon}_{t,i} = \min_{\epsilon_k} \left\| \hat{\mathbf{x}}'_{t,i}(\epsilon_k) - \mathbf{x}'_{t,i} \right\|$$

- **Counterfactual prediction**: 1) Estimate unique characters of the query video; 2) Change the state into the counterfactor; 3) Predict with estimated characters and counterfactual state.

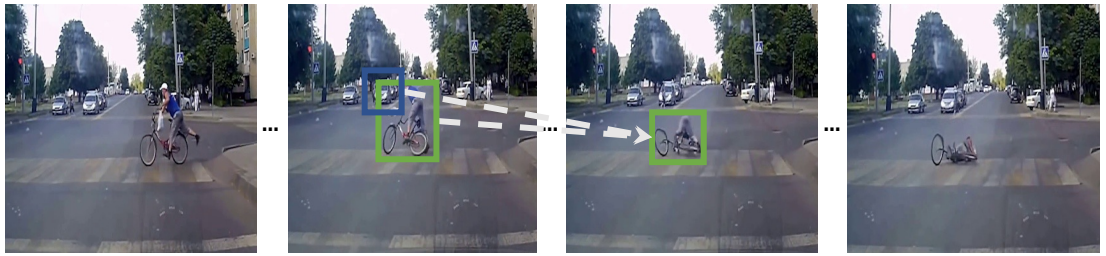


MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Showcase of Experimental Results

Carnegie Mellon University

➢ We can identify the causal relations and thus find the root cause.



Q: What types of vehicles that if get removed from the videos, there won't be an accident?

☑ A: Bicycle or tricycle or non-motor vehicles .

(a) Counterfactual (success)

Q: Could the accident be prevented if the involved vehicles change lane or turn properly?

☑ A: Yes.

(b) Introspection (success)

Q: What might be the reason which led to this accident?

☑ A: The white sedan did an illegal lane changing.

(c) Attribution (success)

Q: Could any involved vehicles stop in time to prevent the accident?

☒ A: Yes, there was enough time to react.

(d) Introspection (fail)

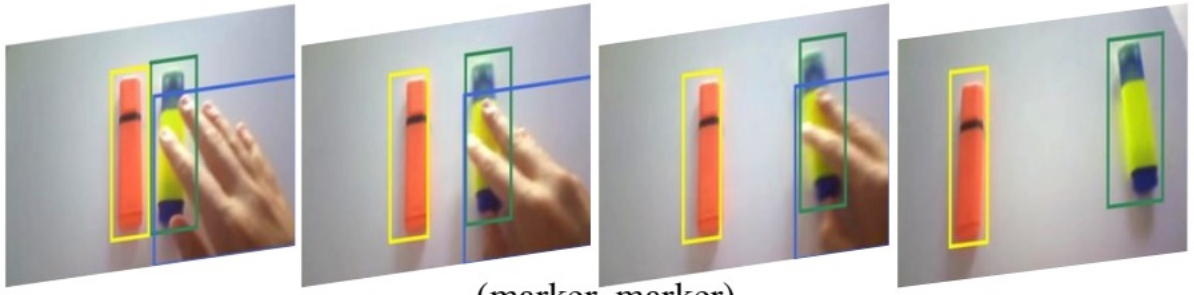
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Application: Few-shot Action Recognition

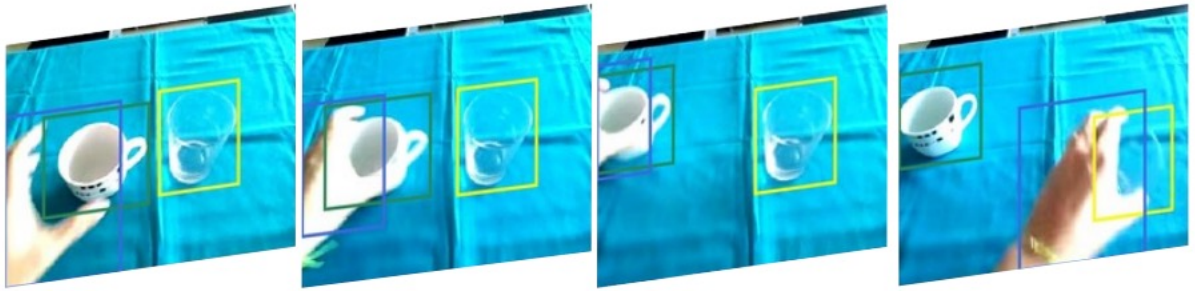
Carnegie Mellon University

- Few-shot action recognition targets to efficiently transfer the learned action recognition knowledge into the new domain with few-shot examples.

Moving [something] and [something] away from each other

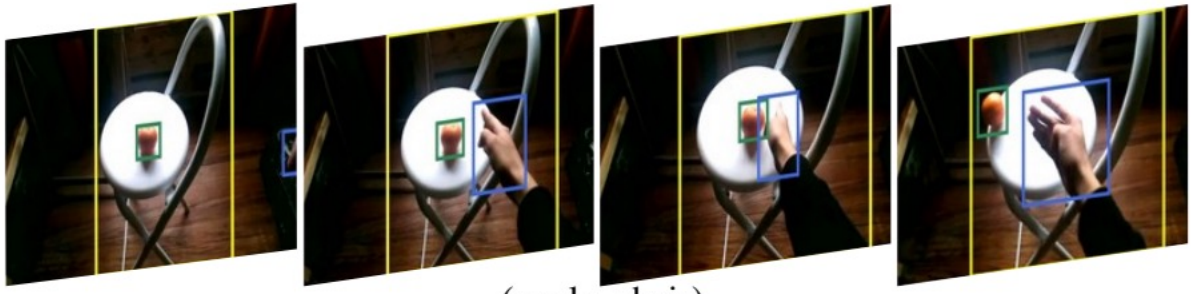


(marker, marker)

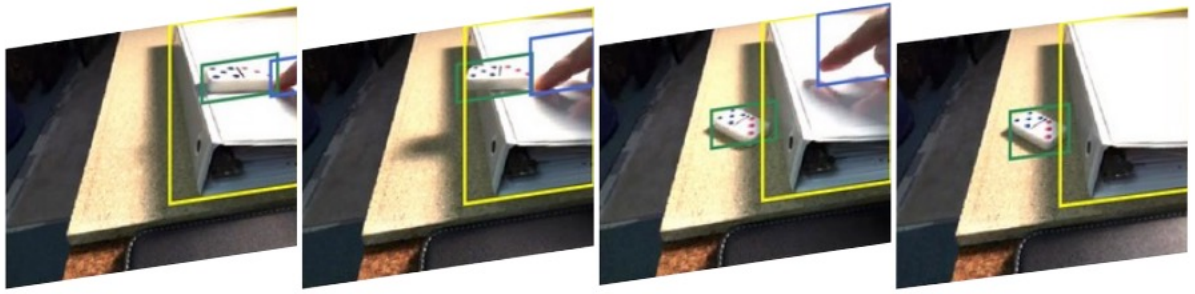


(cup, glass)

Pushing [something] off of [something]



(apple, chair)

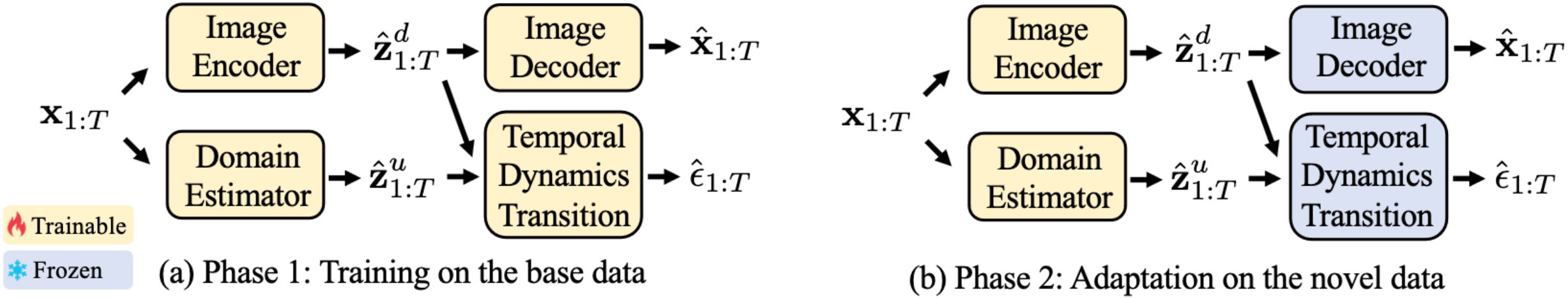


(domino, folder)

Credited to the Something-Else dataset (Materzynska. 2020)

Application: Few-shot Action Recognition

- Key insight: Despite different representations of actions, the underlying physical laws are invariant across different actions.
- Once the causal dynamic model is identified, we fix the temporal dynamic transition function (in each domain) as a constraint to adapt to novel data.



Li et al. Learning Causal Domain-Invariant Temporal Dynamics for Few-Shot Action Recognition. ICML, 2024

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Quantitative Experimental Results

Carnegie Mellon University

➢ Significantly improve the few-shot action recognition performance over previous SOTAs

	SSv2				HMDB-51				UCF-101			
	2-shot	4-shot	8-shot	16-shot	2-shot	4-shot	8-shot	16-shot	2-shot	4-shot	8-shot	16-shot
XCLIP(Ni et al., 2022)	3.9	4.5	6.8	10.0	53.0	57.3	62.8	64.0	70.6	71.5	73.0	91.4
ActionCLIP(Wang et al., 2021b)	4.1	5.8	8.4	11.1	47.5	57.9	57.3	59.1	70.6	71.5	73.0	91.4
VicTR(Kahatapitiya et al., 2023)	4.2	6.1	7.9	10.4	60.0	63.2	66.6	70.7	87.7	92.3	93.6	95.8
VideoPrompt(Ju et al., 2022)	4.4	5.1	6.1	9.7	39.7	50.7	56.0	62.4	71.4	79.9	85.7	89.9
ViFi-CLIP(Rasheed et al., 2023)	6.2	7.4	8.5	12.4	57.2	62.7	64.5	66.8	80.7	85.1	90.0	92.7
VL Prompting(Rasheed et al., 2023)	6.7	7.9	10.2	13.5	63.0	65.1	69.6	72.0	**91.0**	93.7	95.0	96.4
VideoMAE (Tong et al., 2022)	8.2	10.0	15.1	18.2	63.7	69.4	70.9	75.3	**91.0**	94.1	94.8	97.7
CDTD$_{NCE}$ (ours)	**9.5**	**11.6**	**14.8**	**19.5**	**65.8**	**70.2**	**72.5**	**77.9**	90.6	**94.7**	**96.2**	**98.5**

➢ Also performs well in compositional action recognition task

	Loss	Sth-Else	
		5-shot	10-shot
ORViT (Herzig et al., 2022)	CE	33.3	40.2
SViT (Ben Avraham et al., 2022)		34.4	42.6
CDTD$_{CE}$ (ours)		**37.6**	**44.0**
ViFi-CLIP (Rasheed et al., 2023)	NCE	44.5	54.0
VL Prompting (Rasheed et al., 2023)		44.9	58.2
CDTD$_{NCE}$ (ours)		**48.5**	**63.9**

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Takeaway Points

Carnegie Mellon University

- Video data provides more “changeability” than static data, which can help the identification of the latent representation.
- The identification results can be extended to more challenging scenarios, such as unknown non-stationary, non-invertibility, and instantaneous relations.
- In the algorithm, we encourage conditional independence by adding the constraint between the learned posterior and conditional independent prior.
- Learning causal representation can benefit lots of applications such as video reasoning, few-shot action recognition, and so on.

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

References

Carnegie Mellon University

[1] Yao, W., Chen, G., and Zhang, K., “Temporally disentangled representation learning.” NeurIPS, 2022.
[2] Chen, G., Shen, Y., Chen, Z., Song, X., Sun, Y., Yao, W., Liu, X., and Zhang, K., "CaRiNG: Learning Temporal Causal Representation under Non-Invertible Generation Process." ICML, 2024.
[3] Song, X., Yao, W., Fan, Y., Dong, X., Chen, G., Niebles, J.C., Xing, E., and Zhang, K., "Temporally disentangled representation learning under unknown nonstationarity." NeurIPS, 2023.
[4] Li, Z., Cai, R., Yang, Z., Huang, H., Shen, Y., Chen, Z., Song, X., Hao, Z., Chen, G., and Zhang, K., “ When and How: Learning Identifiable Latent States for Nonstationary Time Series Forecasting. ” Preprint, 2024.
[5] Li, Z., Shen, Y., Zheng, K., Cai, R., Song, X., Gong, M., Hao, Z., Zhu, Z., Chen, G., and Zhang, K., “On the Identification of Temporally Causal Representation with Instantaneous Dependence.” Preprint, 2024.
[6] Chen, G., Li, Y., Liu, X., Li, Z., Al Suradi, E., Wei, D., and Zhang, K., LLCP: Learning Latent Causal Processes for Reasoning-based Video Question Answer. ICLR, 2024.
[7] Li, Y., Chen, G., Abramowitz, B., Anzellotti, S., and Wei, D., “Learning Causal Domain-Invariant Temporal Dynamics for Few-Shot Action Recognition.” ICML, 2024.

MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

Carnegie
Mellon
University

Thanks for your listening