CCL: Causal-aware In-context Learning for Out-of-Distribution Generalization

Hoyoon Byun, Gyeongdeok Seo, Joonseong Kang, Taero Kim, Jihee Kim, Kyungwoo Song

Out-of-Distribution in In-context learning (ICL)

Demonstration distribution

Task: Math problem, Env.: Equipment inventory

x_1: A server has 10 GPUs, and 4 are currently in use. How many GPUs are available?

y_1 : 10-4 = 6

c_1 : **M-N (Math problem)**

Task: Sentiment analysis, Env.: Grocery industry

x_2 : **Tom** has already bought this **banana 3** times in the past **7** days. How do you think he feels?

y_2 : Positive

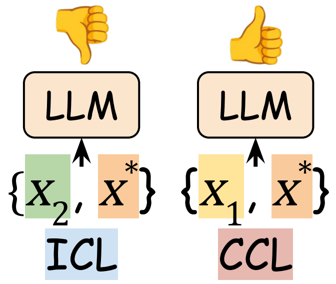
c_2 : Positive (Sentiment analysis)

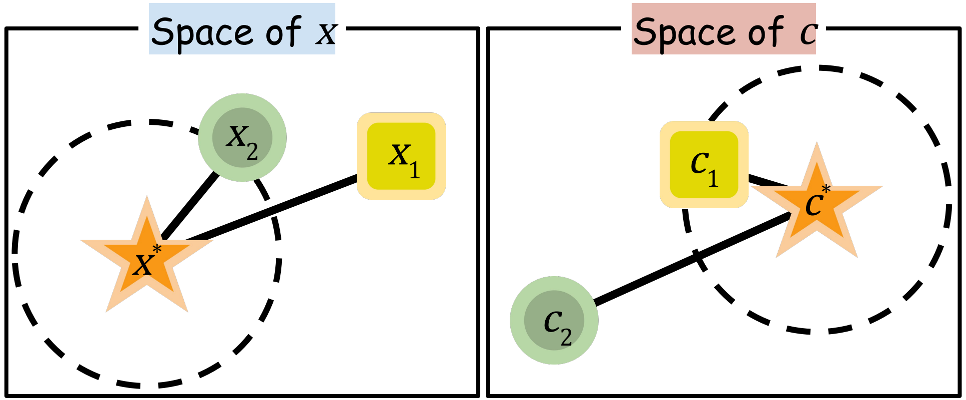
Out-of-distribution

Task: Math problem, Env.: Daily life

x^* : **Tom** ate **3** out of **7 bananas**. How many **bananas** are left?

c^* : **M-N (Math problem)**

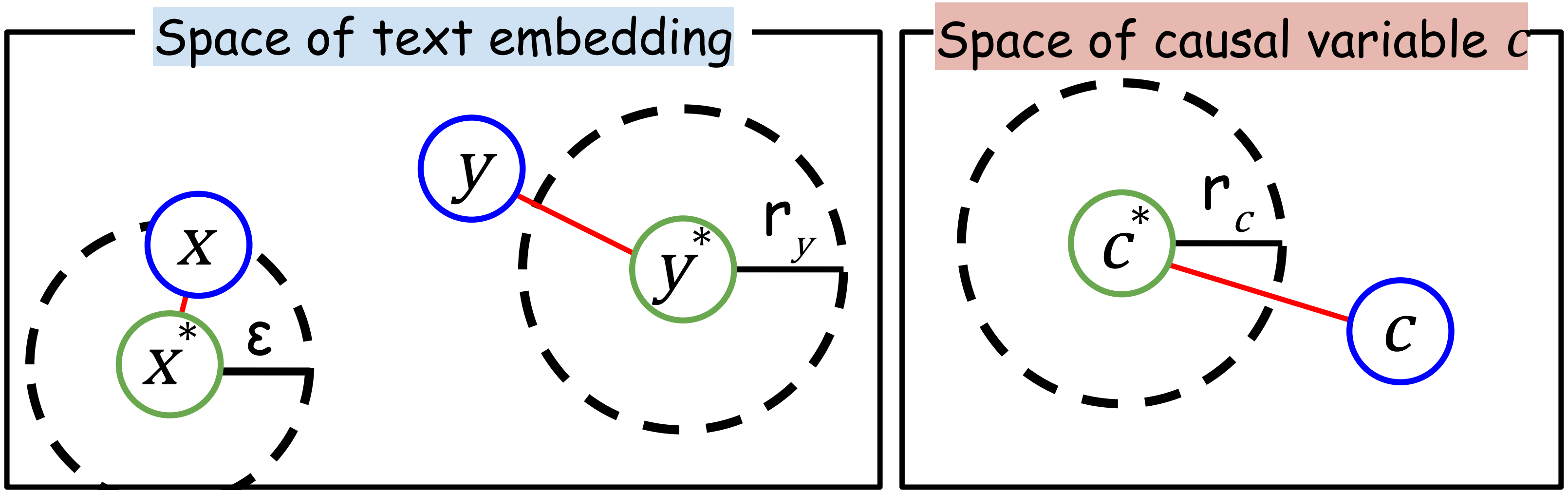




To ensure strong ICL, it is essential to **choose examples that are semantically close to the task-relevant meaning inherent in the query input**, especially when the target and demonstration distributions differ.

Motivation of causal-aware ICL: beyond x-space retrieval

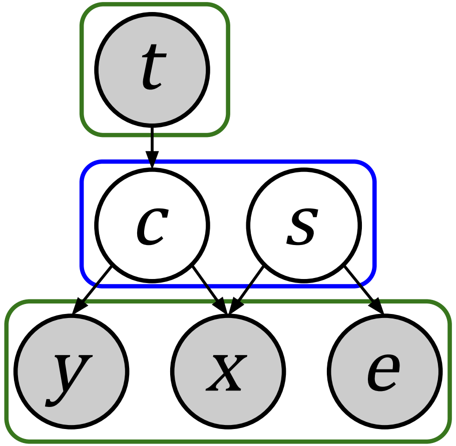
Theorem 3.3: "*Close at the x level can still be distant at the c level.*"



CCL captures task-relevant causal features as latent c, enabling example selection based on the task-relevant causal factors rather than surface x-similarity.

Data-generating process (DGP) for NLP

< Generative model >



We assume that the **domain shift** in the observed data is induced by **changes in s**, while **c remains invariant.**

Observable variables

t : **task variable (Descr. of task)**

Ex. "*Sentiment analysis is a natural language processing (NLP) task that involves determining the emotional tone or sentiment expressed in a piece of text.*"

e : **environment variable (Descr. of data source (or domain))**

Ex. "*This dataset contains reviews of 29 different categories of products collected from the Amazon website, one of the largest e-commerce platforms globally. ……*"

x : **input query variable**

Ex. "*Worked for about 4 months. DVD player will not eject or accept disks. Do not buy.*"

y : **the (ground truth) answer (or response) variable**

Ex. "*Negative*"

Latent causal variables

c : **domain-invariant variables**

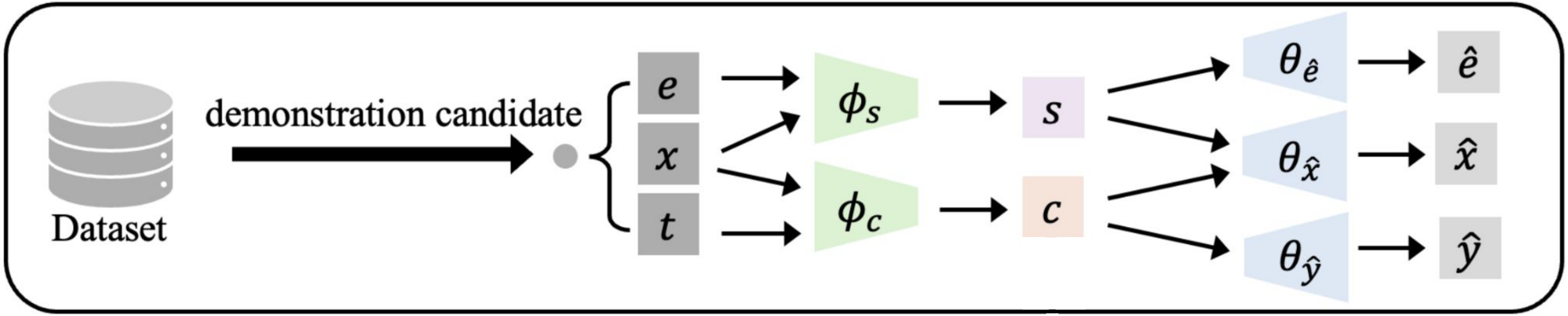
The latent variable that cause query x and answer y represents the underlying task intention

s : **domain-variant variables**

The latent variable represents the domain-specific information.

We define a data-generating process (or causal graph) with two latent variables c and s representing domain-invariant and domain-variant information.

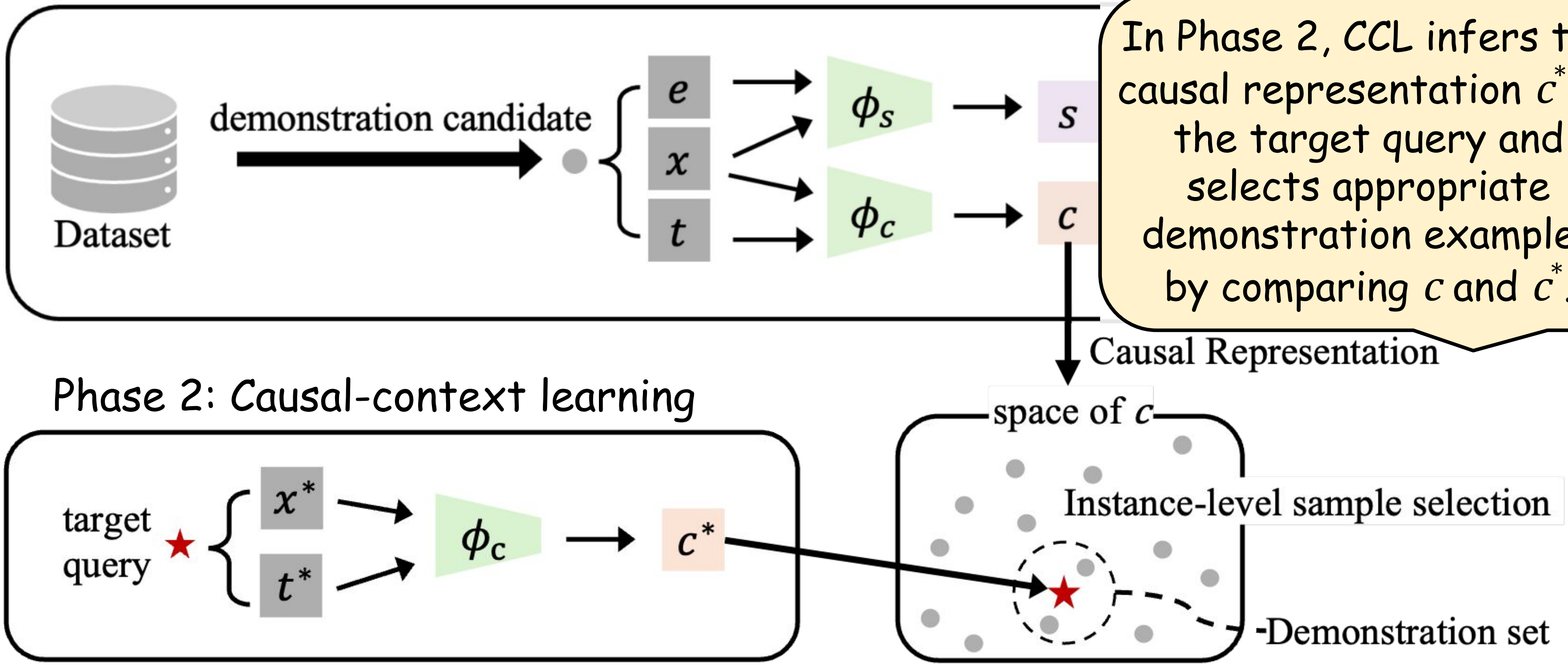
Phase 1: Causal representation learning with In-pool (In-distribution) dataset



We optimize a VAE-based framework to learn causal representations and store the resulting latent causal variables c for the in-distribution dataset.

CCL Frameworks

Phase 1: Causal representation learning with In-pool (In-distribution) dataset



In Phase 2, CCL infers the causal representation c^* of the target query and selects appropriate demonstration examples by comparing c and c^*.

Phase 2: Causal-context learning

Experimental results

MGSM

CCL's causal embedding c achieves better cross-lingual problem retrieval than raw x-embeddings.

Metric	x embedding	c embedding
Total Accuracy	81.03	**85.84**
ID Accuracy	97.05	**99.74**
OOD Accuracy	53.00	**61.52**
Total NDCG	86.00	**88.73**
ID NDCG	99.12	**99.89**
OOD NDCG	63.03	**69.21**

Method	Total	ID	OOD
ZS	87.71	89.43	84.70
ICL (Fix.)	91.20	91.26	91.10
ICL (KNN)	94.07	95.83	91.00
CCL	**94.55**	**96.11**	**91.80**

source : Liang Wang et al.,. Learning to retrieve in-context examples for large language models. EACL 2024

Experimental results

MGSM

CCL's causal embedding c achieves better cross-lingual problem retrieval than raw x-embeddings.

Metric	x embedding	c embedding
Total Accuracy	81.03	**85.84**
ID Accuracy	97.05	**99.74**
OOD Accuracy	53.00	**61.52**
Total NDCG	86.00	**88.73**
ID NDCG	99.12	**99.89**
OOD NDCG	63.03	**69.21**

Method	Total	ID	OOD
ZS	87.71	89.43	84.70
ICL (Fix.)	91.20	91.26	91.10
ICL (KNN)	94.07	95.83	91.00
CCL	**94.55**	**96.11**	**91.80**

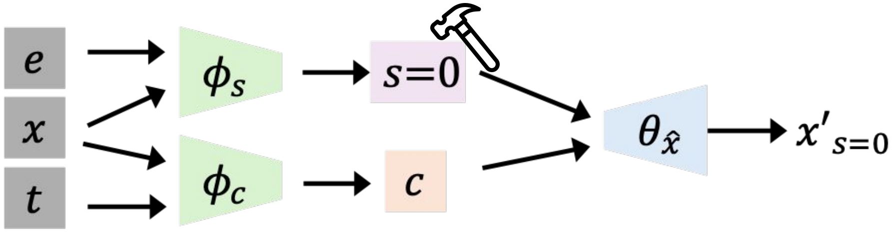
Language model	Retrieval method	QNLI	PIQA	WSC273	YELP	Avg.
Llama-3.2-3B-IT	ZS	43.36	**71.33**	55.31	*88.98*	64.75
	LLM-R	29.93	69.91	61.17	79.48	60.12
	ICL (K-means)	68.13	69.04	49.82	75.81	*65.70*
	CCL	**75.18**	*70.46*	**61.91**	**95.44**	**75.74**
Phi-4-mini-IT	ZS	**86.34**	**76.01**	64.10	95.76	80.55
	LLM-R	*85.21*	74.10	65.93	**96.37**	80.40
	ICL (K-means)	83.18	74.81	*71.06*	96.25	*81.33*
	CCL	82.26	*75.73*	**71.43**	*96.33*	**81.44**
GPT-4o	ZS	**91.30**	*94.07*	90.84	97.47	93.42
	LLM-R	90.32	**94.23**	*92.67*	*98.27*	93.87
	ICL (K-means)	88.28	93.04	87.55	98.17	91.76
	CCL	*90.77*	93.15	**93.77**	**98.36**	**94.01**

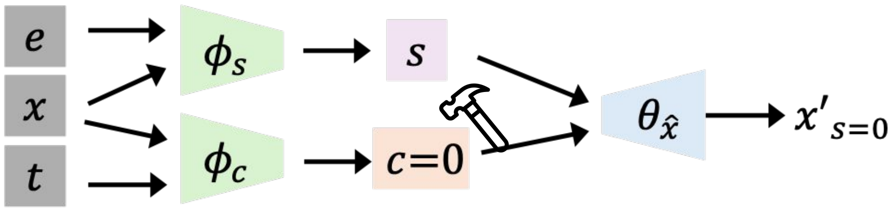
OOD NLP

CCL shows consistently superior performance across various LLMs in OOD NLP experiments.

source : Liang Wang et al.,. Learning to retrieve in-context examples for large language models. EACL 2024

Quantitative analysis





"*the red velvet **pancakes** were horrible and brown, and **potatos** were over cooked and bland.. would not recommend*"

x	$x'_{s=0}$	$x'_{c=0}$
horribleappetizers	unappetizing	review
pancakes	flavorless	reviewers
potatos	horribleappetizers	critiques
hadhorrible	inedible	soggy
bad	trashed	reviews

"Worked for about 4 months. **DVD** player will not eject or accept **disks**. Do not buy."

x	$x'_{s=0}$	$x'_{c=0}$
dvd	unusable	reverb
eject	expired	throw
disks	cancelled	film
unusable	crappy	review
purchased	trashed	trip

We verify that latent variables c and s learn task-relevant and domain-relevant features by zeroing out each latent features in turn and examining the neighboring words of the reconstructed embeddings.

[Paper link] [Code link]

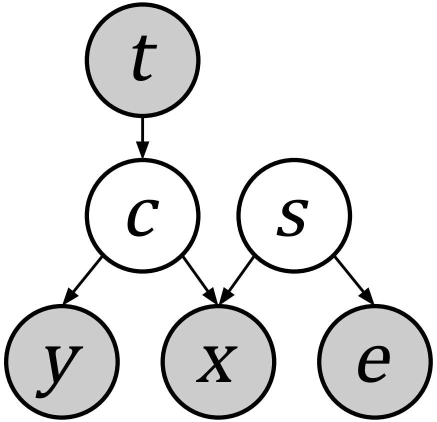
Thanks for Watching

Appendix

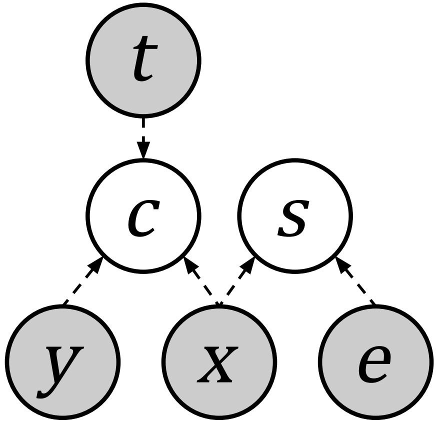
Appendix - VAE-based Causal representation learning (CRL)

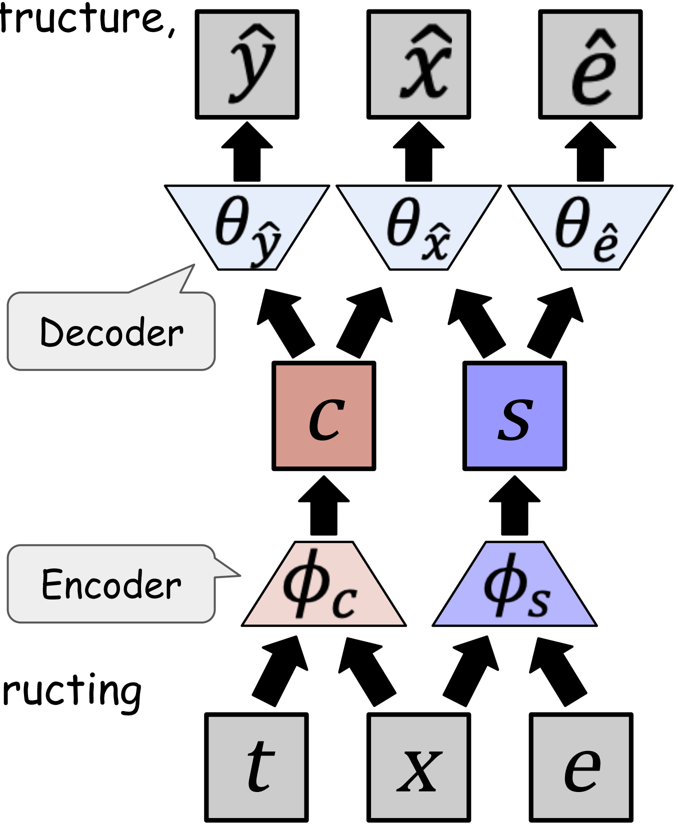
CRL aims to learn latent variables that capture the causal structure, enabling the discovery of causal patterns in observed data.

‹ Generative model ›



‹ Inference model ›





CCL leverages CRL for OOD generalization in ICL by constructing causal representations using a VAE-based model.

source : Liu et al., Learning Causal Semantic Representation for Out-of-Distribution Prediction, NeurIPS 2021,
Lu et al., Invariant Causal Representation Learning for Out-of-Distribution Generalization, ICLR 2022

Appendix - Learning causal representations via variational inference

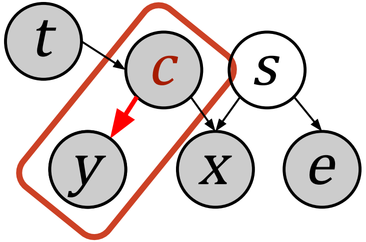
$$\log p_\theta(x,y,t,e) = \log \int p_\theta(x,y,t,e,c,s)\,dc\,ds = \log \mathbb{E}_{q_\phi(c,s|x,y,t,e)}\left[\frac{p_\theta(x,y,t,e,c,s)}{q_\phi(c,s\mid x,y,t,e)}\right]$$

$$\geq \mathbb{E}_{q_\phi(c,s|x,y,t,e)}\left[\log \frac{p_\theta(x,y,t,e,c,s)}{q_\phi(c,s\mid x,y,t,e)}\right] := L_{\text{ELBO}}$$

At test time, **y is always unobserved**, as it is the target variable we aim to infer.

To modify the variational inference objective without conditioning on y, we factorize the inference model by leveraging the conditional independence ($y \perp (x,t,e,s) \mid c$) structure implied by the DGP.

Causal mechanism



$$q_\phi(c,s,y\mid x,t,e) = q_\phi(c,s\mid x,t,e)\boxed{p_\theta(y\mid c)}$$

source : Liu et al., Learning Causal Semantic Representation for Out-of-Distribution Prediction, NeurIPS 2021

Appendix - Reformulating variational inference for unobserved y

$$\log p_\theta(x,y,t,e) = \log \int p_\theta(x,y,t,e,c,s)\,dc\,ds = \log \mathbb{E}_{q_\phi(c,s|x,y,t,e)}\left[\frac{p_\theta(x,y,t,e,c,s)}{q_\phi(c,s \mid x,y,t,e)}\right]$$

$$\geq \mathbb{E}_{q_\phi(c,s|x,y,t,e)}\left[\log \frac{p_\theta(x,y,t,e,c,s)}{q_\phi(c,s \mid x,y,t,e)}\right] := L_{\mathrm{ELBO}}$$

Reformulating ELBO with causal mechanism

$$\max_{\theta,\phi} \mathbb{E}_{p_D(x,y,t,e)}[L_{\mathrm{ELBO}}] = \mathbb{E}_{p_D(x,y,t,e)}\Bigg[\log \Phi_{y|x,t,e} + \frac{1}{\Phi_{y|x,t,e}} \mathbb{E}_{q_\phi(c,s|x,t,e)}\left[\boxed{p_\theta(y|c)} \times \log \frac{p_\theta(x,t,e,c,s)}{q_\phi(c,s|x,t,e)}\right]\Bigg]$$

$$\Phi_{y|x,t,e} = \mathbb{E}_{q_\phi(c,s|x,t,e)}\boxed{[p_\theta(y|c)]}$$

CCL infers latent variables without using y, removing the need for an auxiliary model for y.
By incorporating causal relations into the decoding process, it ensures that c captures task-relevant information.

source : Liu et al., Learning Causal Semantic Representation for Out-of-Distribution Prediction, NeurIPS 2021