NEURAL INFORMATION PROCESSING SYSTEMS

Rethinking Misalignment in Vision-Language Model Adaptation from a Causal Perspective

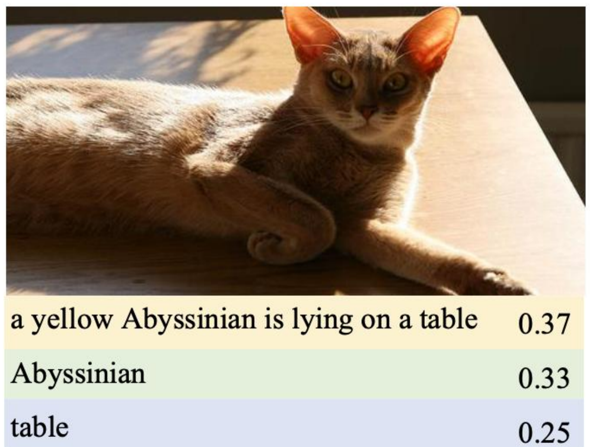
Yanan Zhang*, Jiangmeng Li*,
Lixiang Liu, Wenwen Qiang

中国科学院大学
University of Chinese Academy of Sciences

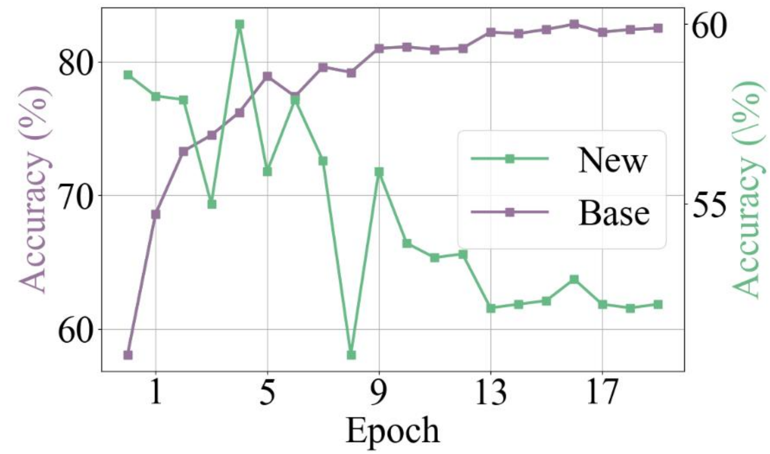
ISCAS

Motivation

There exists a two-level misalignment between CLIP and downstream tasks, which hinders its adaptation to these tasks.



(a) A motivating example of task misalignment.



(b) A motivating experiment on data misalignment

Two-level misalignment:

(a) task misalignment: caused by the discrepancy between the pre-training objectives of CLIP and the objectives of downstream tasks.

(b) data misalignment: inconsistency exists between the training and testing data.

Problem Analysis

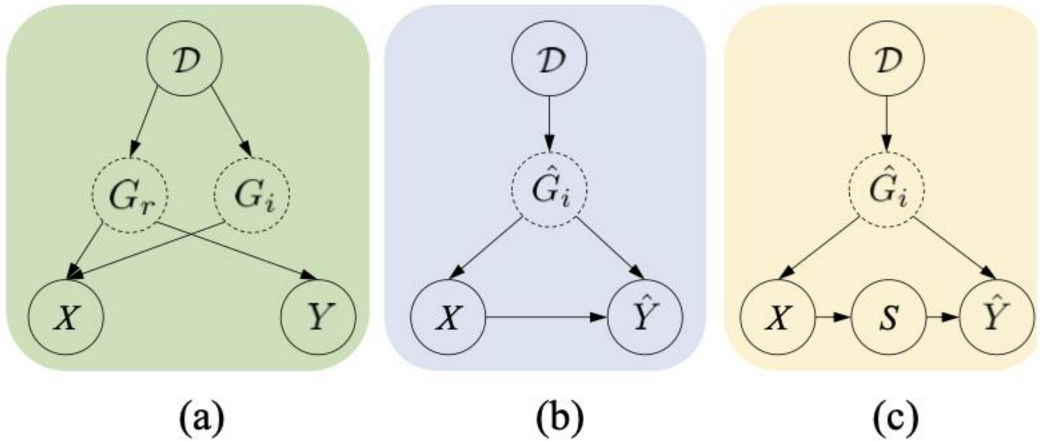


Figure. Structural Causal Model.

Confounder: the set of task-irrelevant generative factors that are incorrectly retained

Front-Door Adjustment:

$$P(\hat{Y}=y|do(x)) = \sum_{s} \underbrace{P(s|x)}_{\text{first term}} \underbrace{\sum_{x'} P(y|x',s)P(x')}_{\text{second term}}.$$

Causality-Guided Semantic Decoupling and Classification

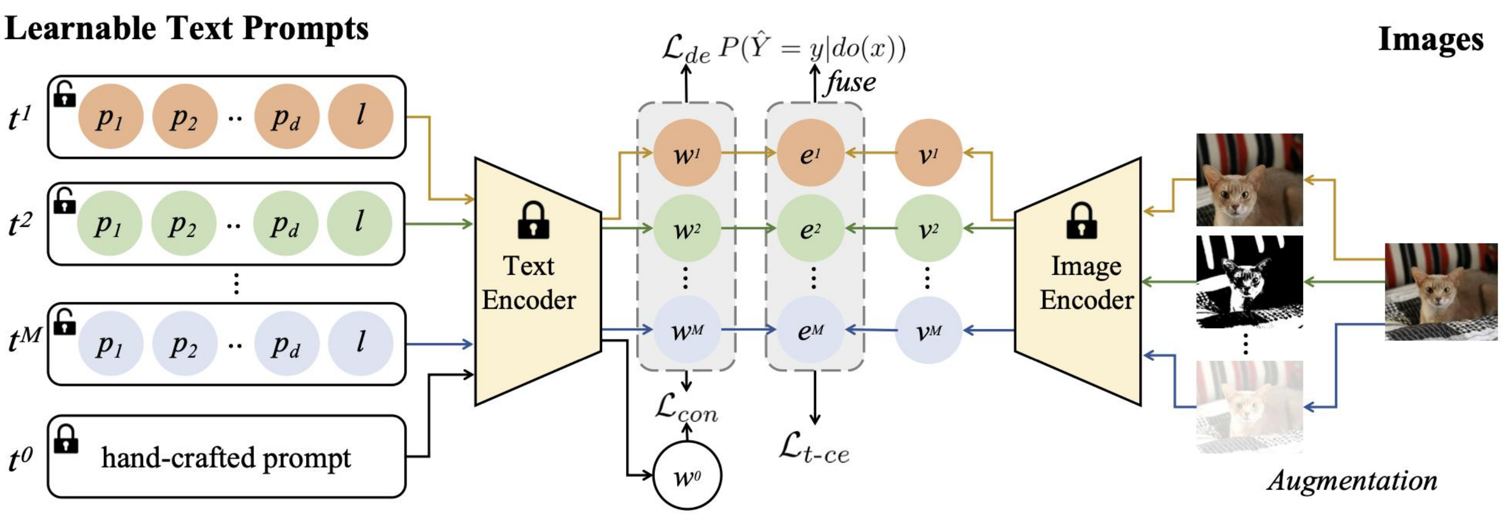


Figure Framework of CDC.

Visual-Language Dual Semantic Decoupling

- Visual: data augmentations
- Language:

$$\mathcal{L}_{de} = \frac{1}{M-1}\frac{1}{C}\sum_{m'=1,m'\neq m}^{M}\sum_{c=1}^{C}\sum_{\bar{c}=1}^{C} P(\bar{c}|w_c^m, w^{m'})\log P(\bar{c}|w_c^m, w^{m'})$$

$$\mathcal{L}_{con} = -\frac{1}{C}\sum_{m=1}^{M}\sum_{c=1}^{C}\log P(c|w_c^m, w^0)$$

Causality-Guided Semantic Decoupling and Classification

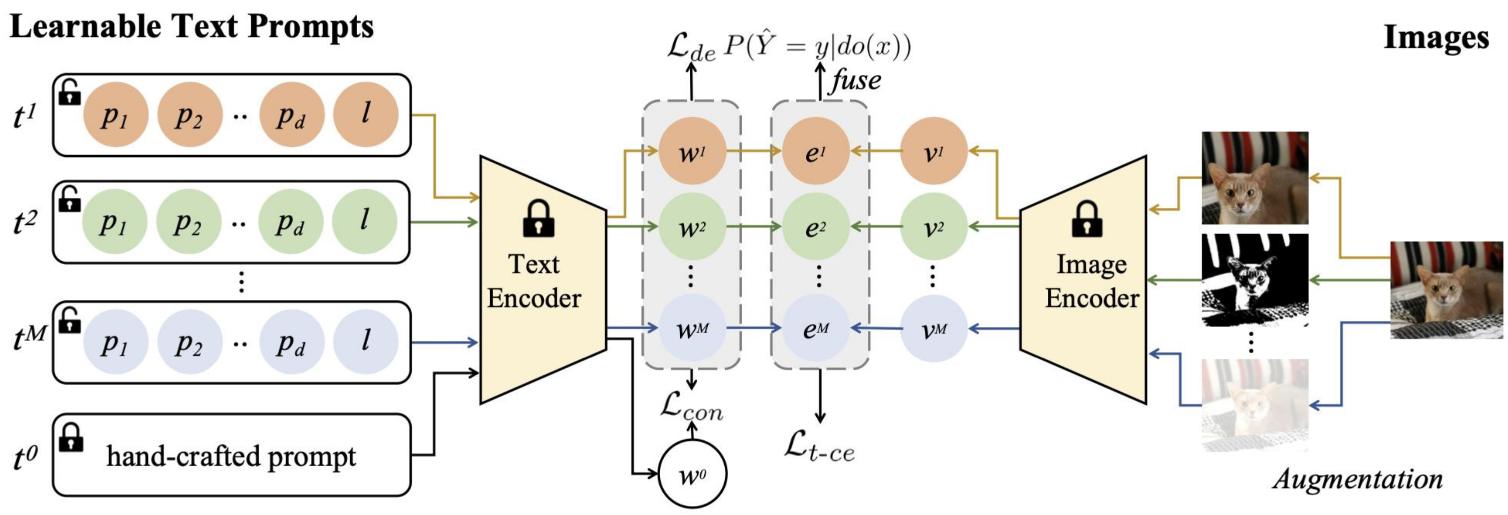


Figure Framework of CDC.

Decoupled Semantic Trusted Classification

$$B_c^m = \begin{cases} b_c^1, & \text{if } m = 1 \\ \frac{1}{1-C}(B_c^{m-1} b_c^m + B_c^{m-1} u^m + b_c^m U^{m-1}), & \text{if } 1 < m \le M \end{cases},$$

$$U^m = \begin{cases} u^1, & \text{if } m = 1 \\ \frac{1}{1-C} U^{m-1} u^m, & \text{if } 1 < m \le M \end{cases}.$$

Evaluation

- Compared to the baseline method MaPLe with base-to-new setting

Table 1: The comparison with baseline methods on base-to-novel generalization setting.

Dataset	CoOp [4]			CoCoOp [5]			MaPLe [6]			CDC			
	Base	New	HM	Base	New	HM	Base	New	HM	Base	New	HM	Δ
Avg	82.69	63.22	71.66	80.47	71.69	75.83	82.28	75.14	78.55	**83.34**	**77.38**	**80.25**	+1.70
ImageNet	76.47	67.88	71.92	75.98	70.43	73.10	76.66	70.54	73.47	**77.50**	**71.73**	**74.51**	+1.04
Caltech	98.00	89.91	93.73	97.96	93.81	95.84	97.74	94.36	96.02	**98.20**	**94.37**	**96.25**	+0.23
Pets	93.67	95.29	94.47	95.20	97.69	96.43	95.43	97.76	96.58	**96.07**	**98.00**	**97.02**	+0.44
Cars	**78.12**	60.40	68.13	70.49	73.59	72.01	72.94	**74.00**	73.47	73.80	73.97	**73.88**	+0.41
Flowers	**97.60**	59.67	74.06	94.87	71.75	81.71	95.92	72.46	82.56	96.93	**75.07**	**84.61**	+2.05
Food	88.33	82.26	85.19	90.70	91.29	90.99	90.71	92.05	91.38	**90.87**	**92.33**	**91.59**	+0.21
Aircraft	**40.44**	22.30	28.75	33.41	23.71	27.74	37.44	35.61	36.50	37.47	**37.50**	**37.48**	+0.98
SUN	80.60	65.89	72.51	79.74	76.86	78.27	80.82	78.70	79.75	**82.37**	**80.03**	**81.18**	+1.43
DTD	79.44	41.18	54.24	77.01	56.00	64.85	80.36	59.18	68.16	**82.70**	**64.10**	**72.22**	+4.06
SAT	92.19	54.74	68.90	87.49	60.04	71.21	94.07	73.23	82.35	**95.10**	**82.33**	**88.26**	+5.91
UCF	84.69	56.05	67.46	82.33	73.45	77.64	83.00	78.66	80.77	**85.70**	**81.73**	**83.67**	+2.90

- The cross-dataset generalization

Table 2: Comparison of CDC with recent approaches on cross-dataset evaluation.

	Source	**Target**										
	ImageNet	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN	DTD	SAT	UCF	Avg
CoOp	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
Co-CoOp	71.02	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
MaPLe	70.72	93.53	90.49	65.57	72.23	86.20	**24.74**	67.01	46.49	48.06	**68.69**	66.30
CDC	**71.76**	**94.47**	**90.77**	**66.27**	**72.67**	**86.27**	24.50	**68.07**	**46.60**	**49.13**	68.60	**66.73**

Conclusion

- We identify the two-level misalignment in adaptation of CLIP.
- We develop a Structural Causal Model and discover the confounder which hinders the estimation of true causal relationships between new samples and their categories.
- Empirically, CDC outperforms state-of-the-art methods on multiple experimental settings.

Thanks！