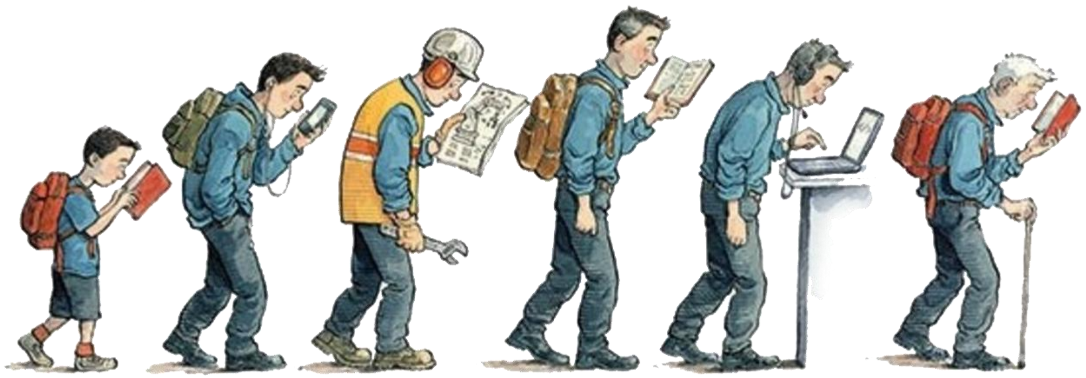
Stabilizing Zero-Shot Prediction: A Novel Antidote to Forgetting in Continual Vision-Language Tasks

Zijian Gao[1,2]†, Xingxing Zhang[3]†, Kele Xu[1,2]*, Xinjun Mao[1,2], Huaimin Wang[1,2]
[1]School of Computer, National University of Defense Technology, Changsha, 410000, China.
[2]State Key Laboratory of Complex & Critical Software Environment, Changsha, 410000, China.
[3]School of Computer Science, Tsinghua University, Beijing, 100049, China.
{gaozijian19, xukelele, xjmao, hmwang}@nudt.edu.cn, xxzhang1993@gmail.com

Continual Learning

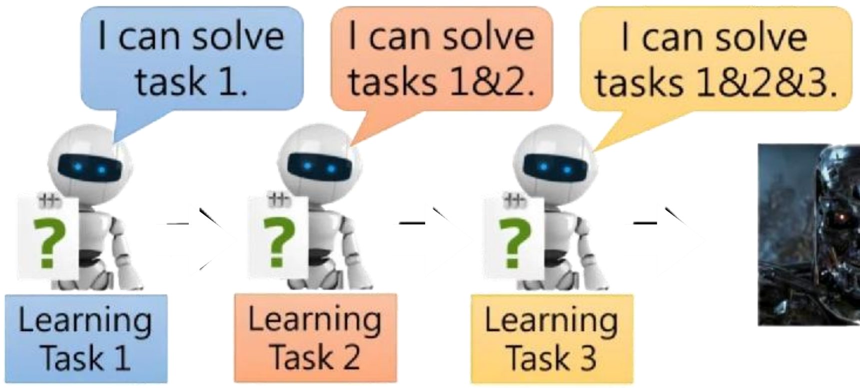
Biological intelligence

"Live and learn"



Machine intelligence

"Catastrophic forgetting"



Incremental learning,
Lifelong learning,
Never ending learning,
Continual learning

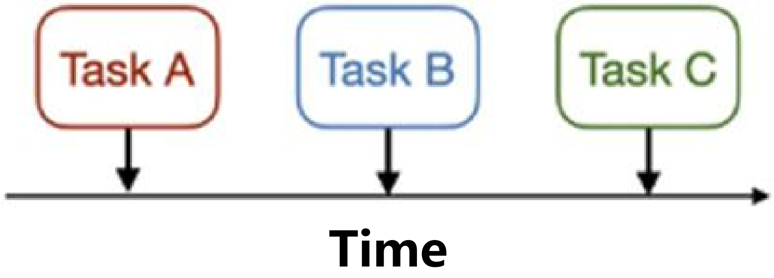
"Continual learning is the constant development of increasingly complex behaviors; the process of building more complicated skills on top of those already developed."

Ring (1997). CHILD: A First Step Towards Continual Learning, Machine Learning.
Sutton (2024). Loss of plasticity in deep continual learning, Nature.

Core Challenge

Catastrophic Forgetting (McCloskey, 1989)

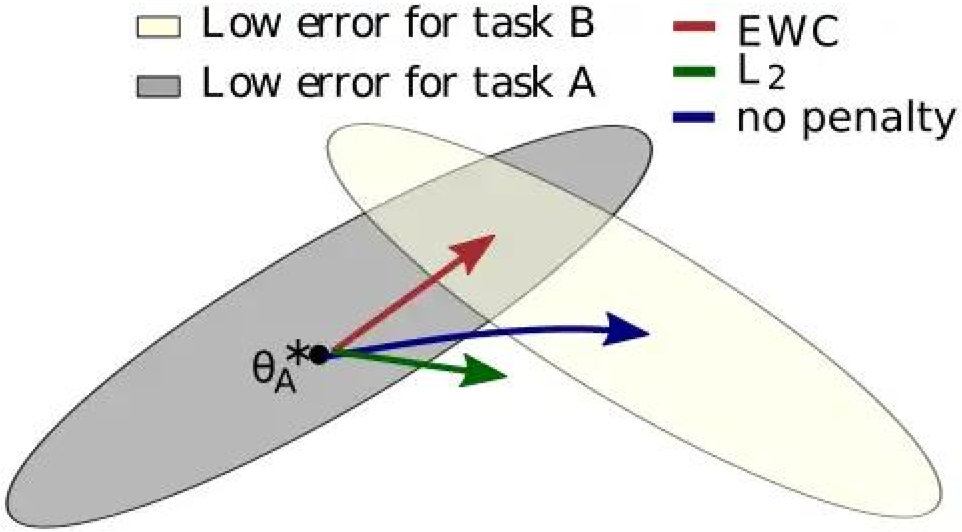
"... the process of learning a **new** set of patterns suddenly and completely **erased** a network's knowledge of **what it had already learned.**" (French, 1999)



Stability vs. Plasticity Dilemma (Carpenter, 1987)

Plasticity ⇔ ability to **adapt** to a new task **(Learning)**

Stability ⇔ ability to **retain** the learned skills on the old tasks **(Anti-forgetting)**

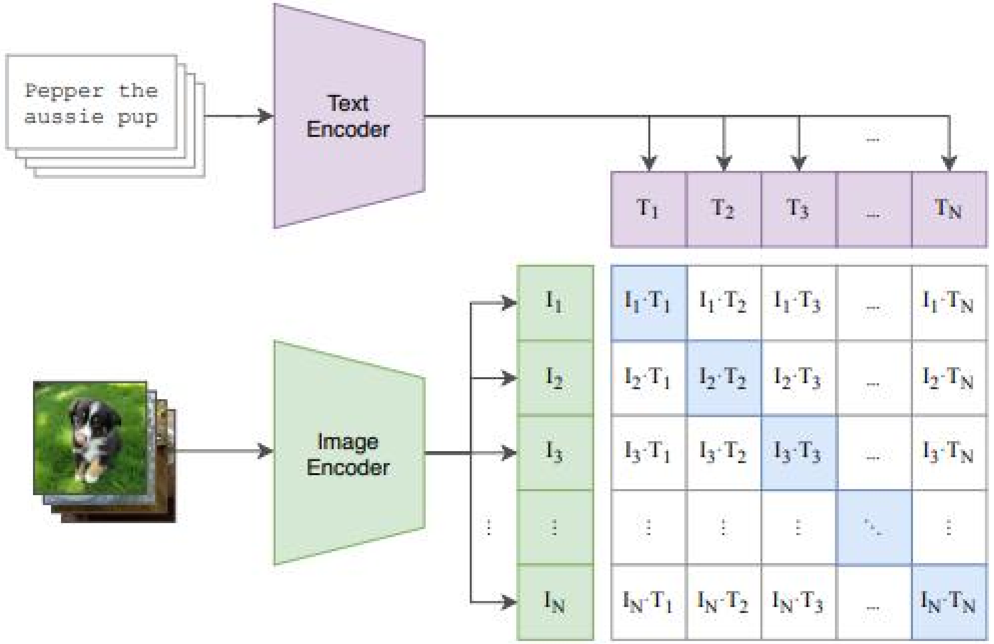


Carpenter (1987). ART 2: Self-organization of stable category recognition codes for analog input patterns, Applied optics.
McCloskey (1989). Catastrophic interference in connectionist networks: the sequential learning problem, Psychol. Learn. Motiv.
French (1999). Catastrophic forgetting in connectionist networks, Trends Cogn. Sci.

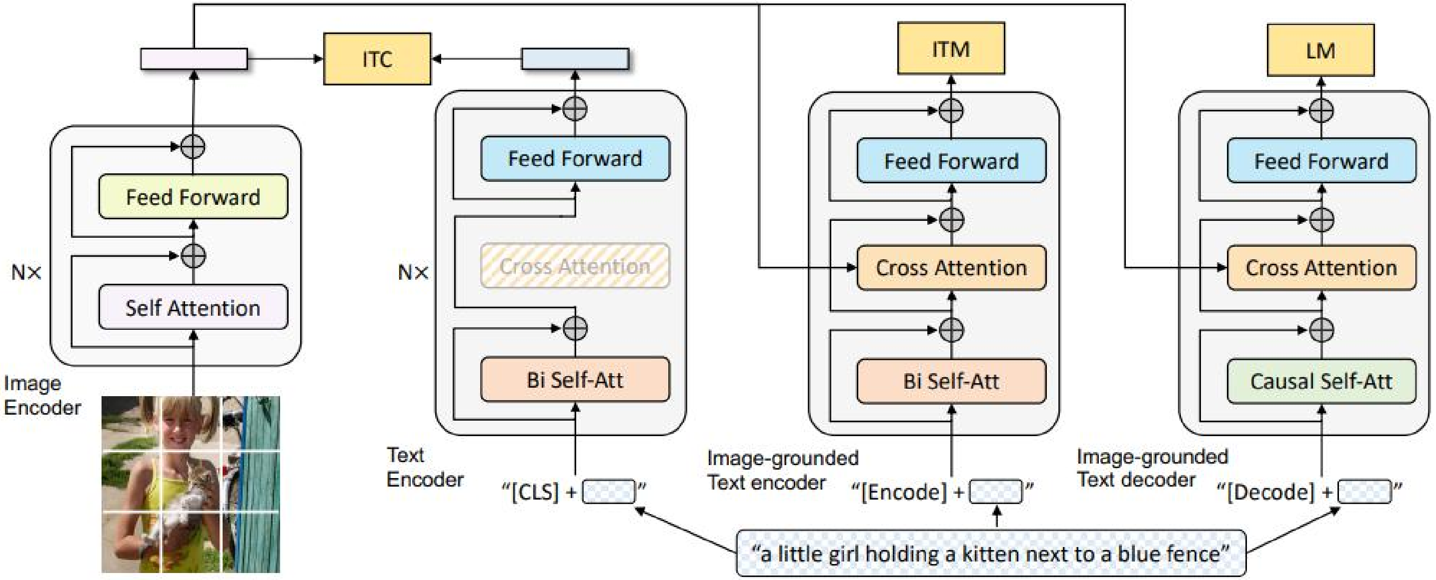
Continual Concept NVLR

Examples of Structured VL Concept Reasoning Task

Multi-modal Foundation Model



Contrastive Language-Image Pre-training (CLIP) Model



Bootstrapping Language-Image Pre-training (BLIP) Model

A. Radford (2021). Learning transferable visual models from natural language supervision, ICML.
Junnan Li (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, ICML.

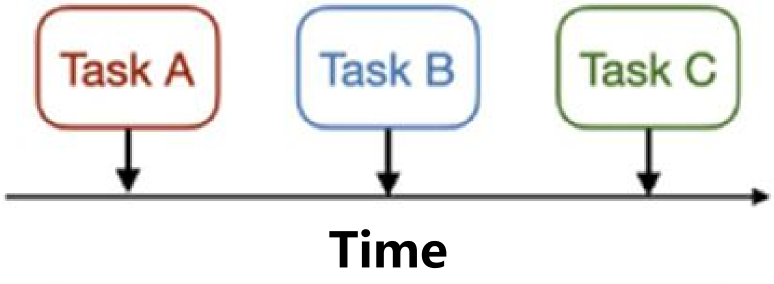
In the Era of Foundation Model

Stability vs. Plasticity Dilemma (Carpenter, 1987)

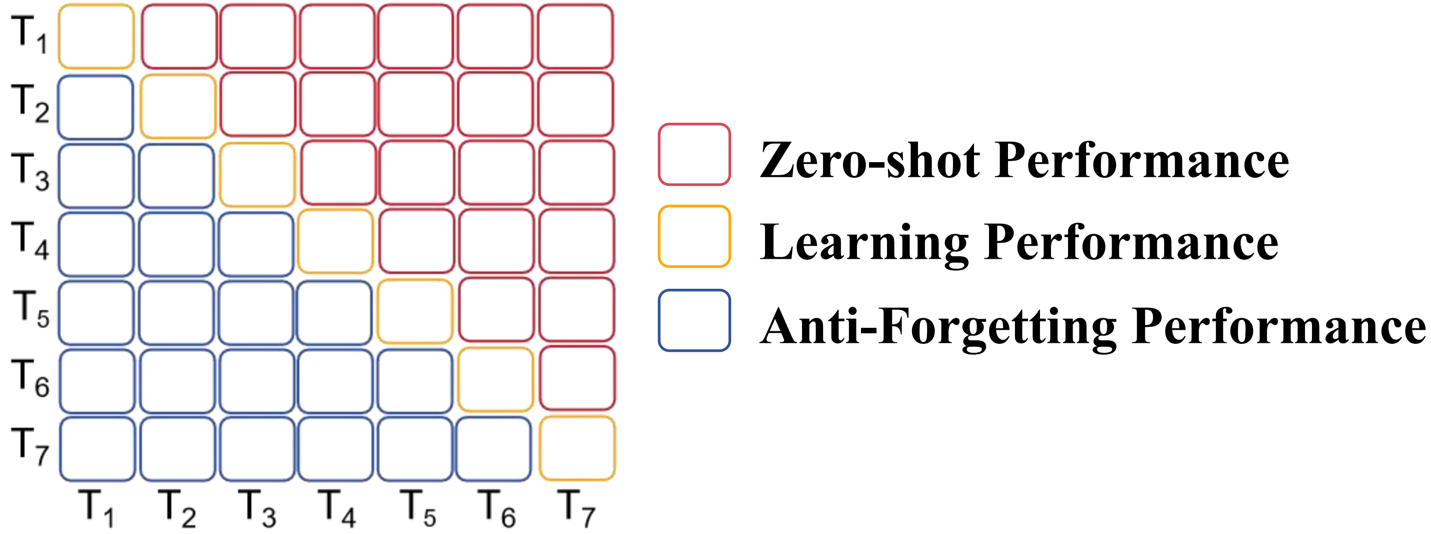
Plasticity ⇔ ability to adapt to a new task (Learning)

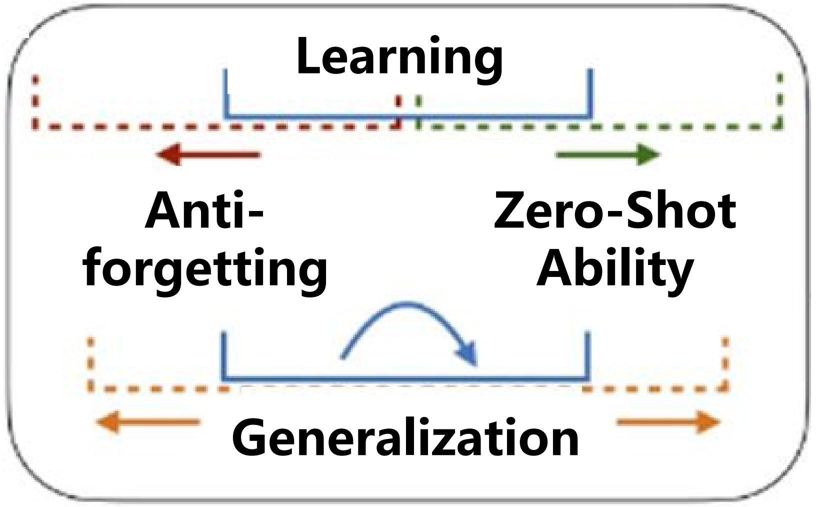
Stability ⇔ ability to retain the learned skills on the old tasks (Anti-forgetting)

Transferability ⇔ ability to transfer the learned skills on the future tasks (Zero-shot Ability)

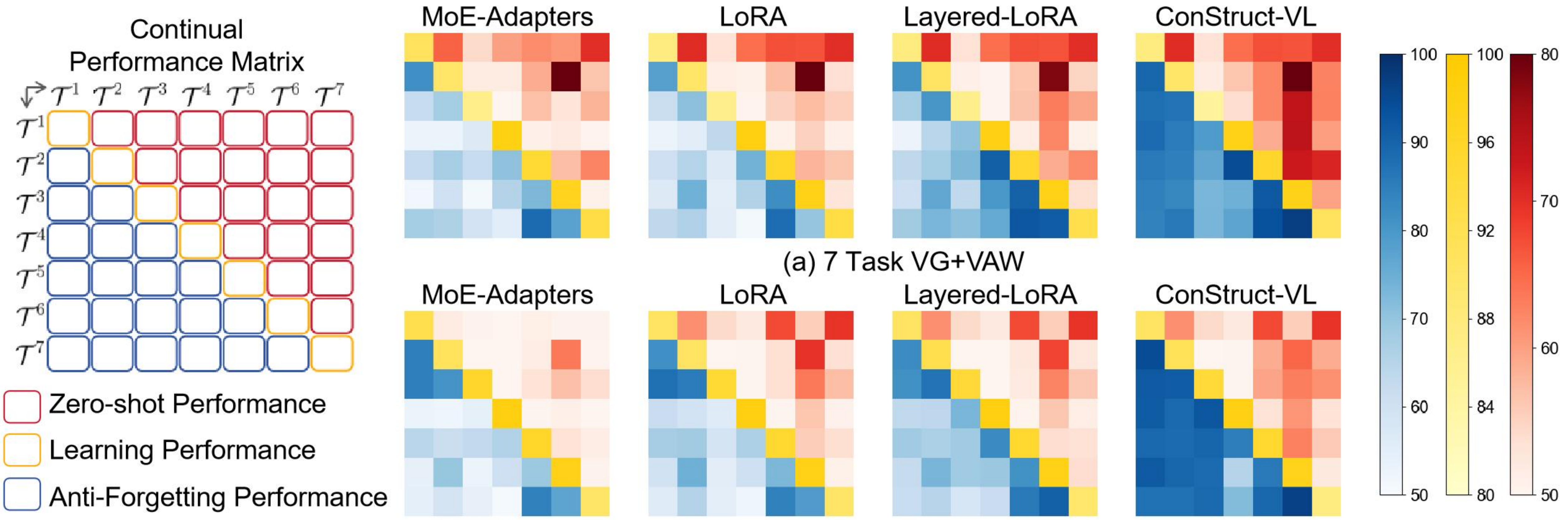


Continual Learning Performance Matrix





Empirical Study



(a) 7 Task VG+VAW

(b) 7 Task VG

A model's stability in zero-shot predictions can reflect its anti-forgetting capabilities.

Theoretical Study

Proposition 1 *For continual learning with pre-trained VL models, let $\mathcal{M}^t$ denote a solution of the continually learned tasks $\mathcal{T}^1, \cdots, \mathcal{T}^t$. In particular, $\mathcal{M}^t = \arg\min_{||\mathcal{M}-\mathcal{M}^{t-1}||_2 \leq \Delta} \hat{\mathcal{E}}_t(\mathcal{M})$ where $||\mathcal{M} - \mathcal{M}^{t-1}||_2 \leq \Delta$ represents the weight vectors for continual tasks are only minor variations. For any $\delta \in (0,1)$ with probability at least $1-\delta$:*

$$\boxed{\forall s \in \{1, \cdots, t-1\}} \quad \mathcal{E}_s(\mathcal{M}^t) \leq \hat{\mathcal{E}}_{1:t}(\mathcal{M}^t) + \frac{1}{2t}\sum_{i=1}^{t} \mathrm{Div}(\mathcal{T}_i, \mathcal{T}_s) + \sqrt{\frac{d[\ln(\bar{N}/d)] + \ln(1/\delta)}{2\bar{N}}}, \quad (2)$$

$$\boxed{\forall k \in \{t+1, \cdots, n\}} \quad \mathcal{E}_k(\mathcal{M}^t) \leq \hat{\mathcal{E}}_{1:t}(\mathcal{M}^t) + \frac{1}{2t}\sum_{i=1}^{t} \mathrm{Div}(\mathcal{T}_i, \mathcal{T}_k) + \sqrt{\frac{d[\ln(\bar{N}/d)] + \ln(1/\delta)}{2\bar{N}}}, \quad (3)$$

empirical error of continual tasks *discrepancy between task distributions*

complexity of the parameter space

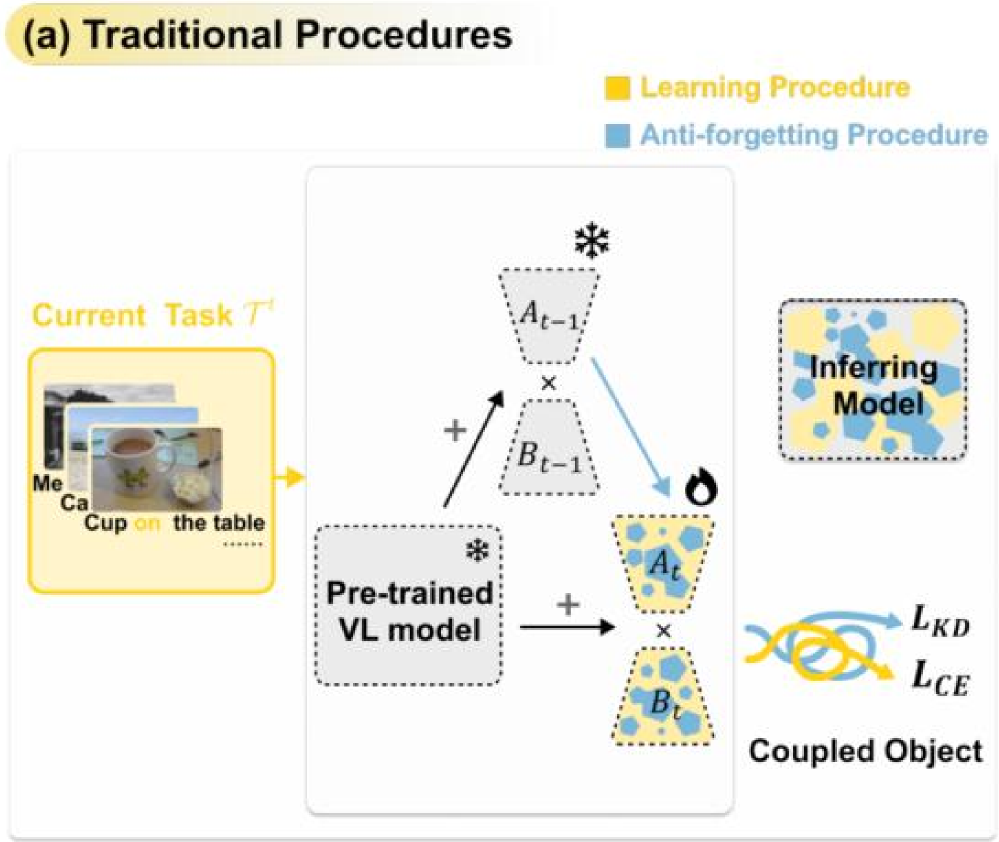
The model $\mathcal{M}^t$ has consistent upper bounds on the generalization errors for both previously learned and future tasks.

Rethinking the Dilemma

Given task $\mathcal{T}^t$, old model $\mathcal{M}^{t-1}$ and current model $\mathcal{M}^t$:

Stability vs. Plasticity Balance (Carpenter, 1987)

Dilemma: Learning $\Longleftrightarrow$ Anti-Forgetting



Optimization Goal:

$$\min \mathcal{L}_{\mathrm{CE}}(P^t(\mathcal{T}^t), \overline{P}(\mathcal{T}^t)) + \mathcal{L}_{\mathrm{KD}}(P^t(\mathcal{T}^t), P^{t-1}(\mathcal{T}^t))$$

Anti-Forgetting

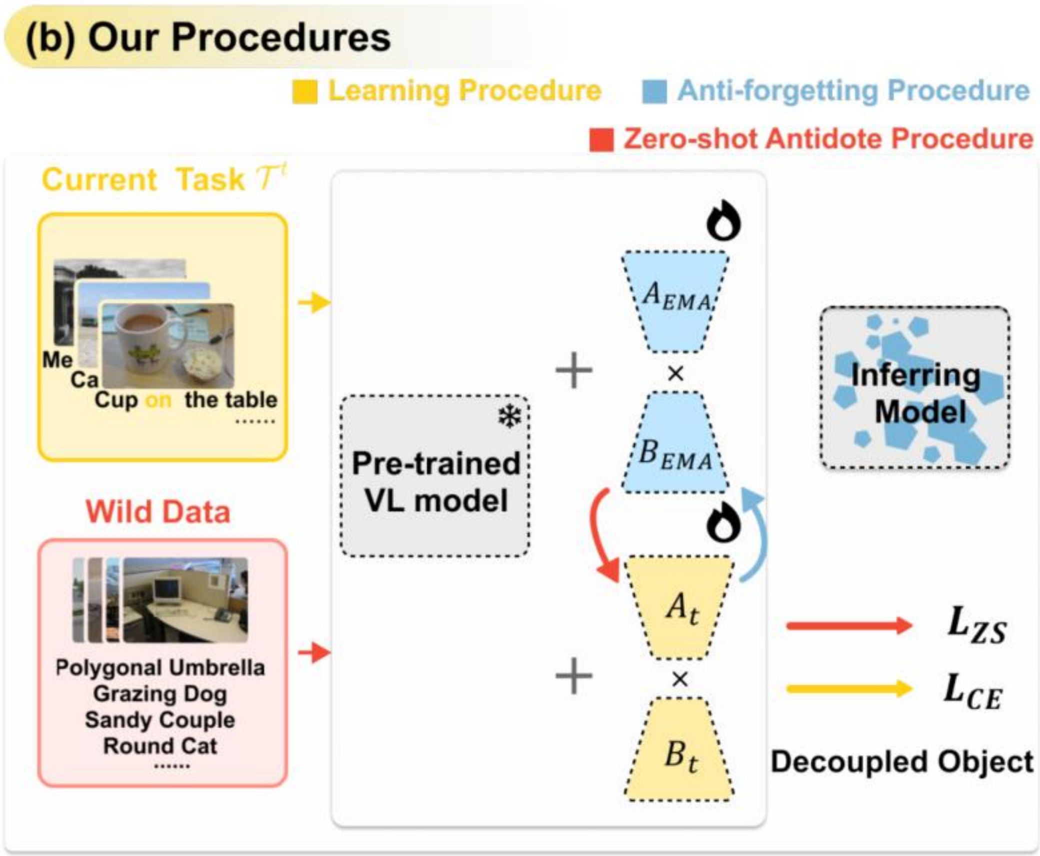
Zero-Shot Stability

Our Paradigm

Given task $\mathcal{T}^t$, old model $\mathcal{M}^{t-1}$, current model $\mathcal{M}^t$ and wild data $\mathcal{D}_{\text{wild}}$:

1. **Wild data**
2. **Zero-shot Stability Supervision**

Win-Win: Learning $\Longleftrightarrow$ Zero-Shot Stability $\Longleftrightarrow$ Anti-Forgetting



Optimization Goal:

$$\min \mathcal{L}_{\text{CE}}(P^t(\mathcal{T}^t), \overline{P}(\mathcal{T}^t)) + \mathcal{L}_{\text{KD}}(P^t(\mathcal{T}^t), P^{t-1}(\mathcal{T}^t))$$

$$\min \mathcal{L}_{\text{CE}}(P^t(\mathcal{T}^t), \overline{P}(\mathcal{T}^t)) + \mathcal{L}_{\text{ZS}}(P^t(\mathcal{D}_{\text{wild}}), \widehat{P}^t(\mathcal{D}_{\text{wild}}))$$

Exponential Moving Average (EMA):

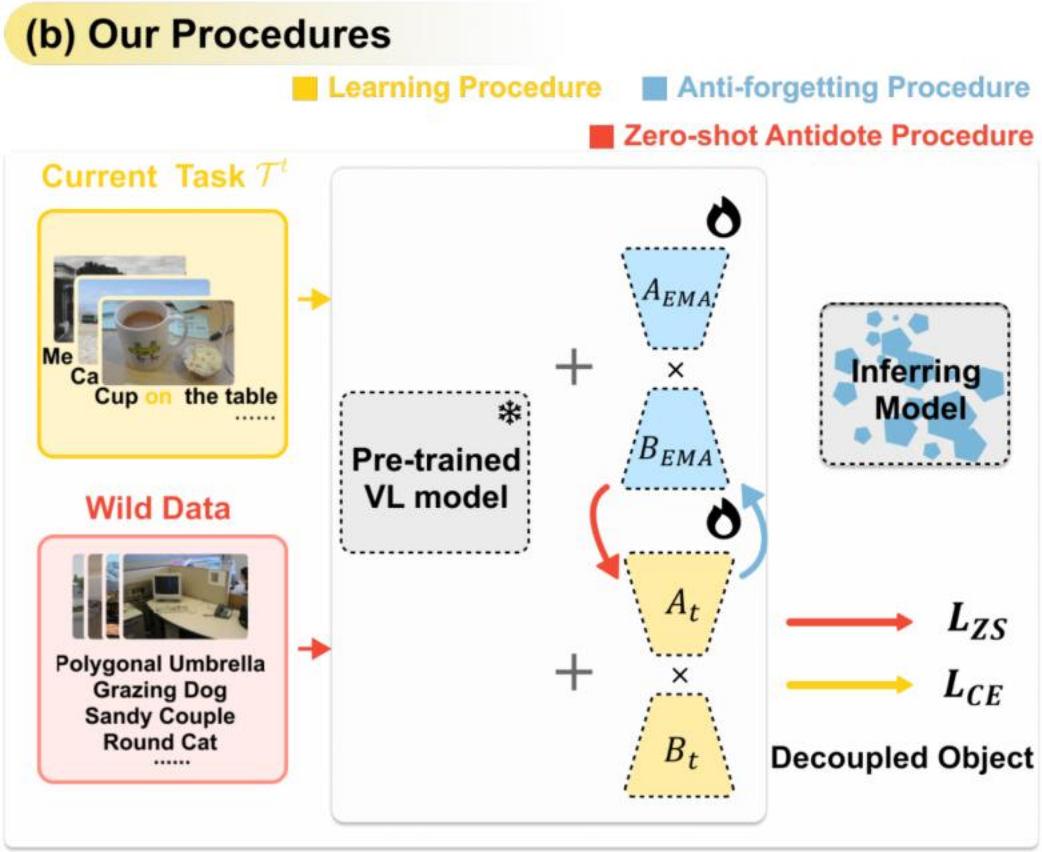
$$\widehat{\mathcal{W}} \leftarrow \alpha \widehat{\mathcal{W}} + (1-\alpha)\mathcal{A} \cdot \mathcal{B}$$

Our Paradigm

Given task $\mathcal{T}^t$, old model $\mathcal{M}^{t-1}$, current model $\mathcal{M}^t$ and wild data $\mathcal{D}_{\text{wild}}$:

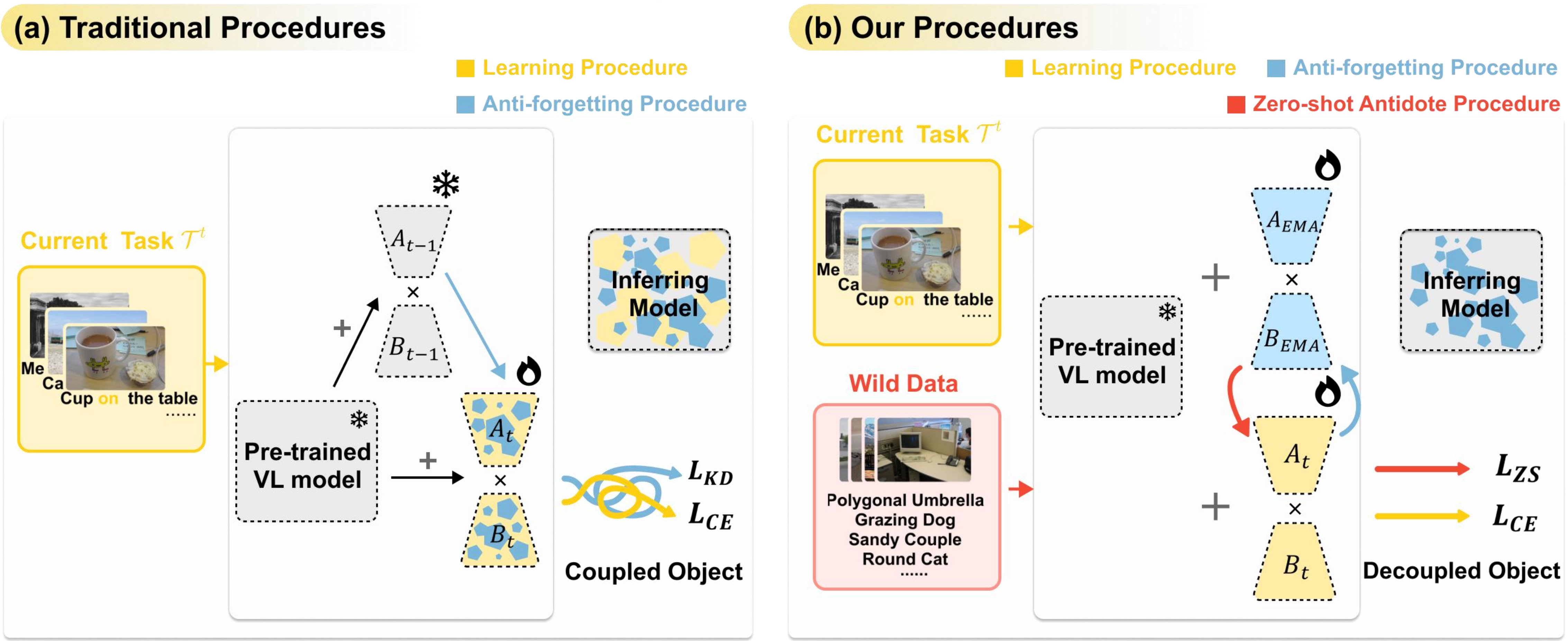
1. Wild data
2. Zero-shot Stability Supervision

Win-Win: Learning ⟺ Zero-Shot Stability ⟺ Anti-Forgetting



By systematically stabilizing zero-shot predictions during continual learning, we can significantly enhance the model's ability to retain historical knowledge without compromising the acquisition of new information.

Comparison



Comparison of training and inference procedures between traditional and our CL paradigm.

Existing Paradigm: $\min \mathcal{L}_{\text{CE}}(P^t(\mathcal{T}^t), \overline{P}(\mathcal{T}^t)) + \mathcal{L}_{\text{KD}}(P^t(\mathcal{T}^t), P^{t-1}(\mathcal{T}^t))$

Our Paradigm: $\min \mathcal{L}_{\text{CE}}(P^t(\mathcal{T}^t), \overline{P}(\mathcal{T}^t)) + \mathcal{L}_{\text{ZS}}(P^t(\mathcal{D}_{\text{wild}}), \widehat{P}^t(\mathcal{D}_{\text{wild}}))$

Wild Data (Unpaired & Unlabeld)

Results

Table 1: Overall performance (%) of CL methods across three benchmarks under various VL models.

VL models	Method	7 Task VG+VAW			7 Task VG			5 Task VAW		
		FAA (↑)	CAA (↑)	FFM (↓)	**FAA** (↑)	CAA (↑)	FFM (↓)	**FAA** (↑)	CAA (↑)	FFM (↓)
BLIP	Joint Learning	91.90	-	-	95.27	-	-	92.60	-	-
	Continual-FT [7]	65.21	73.98	30.32	63.91	73.97	31.34	67.07	78.35	28.14
	LoRA [10]	75.39	76.59	20.73	69.16	75.89	28.20	71.54	79.07	22.48
	Layered-LoRA [33]	76.68	78.51	18.96	70.13	79.66	28.08	83.77	83.47	9.20
	LwF [21]	70.93	73.62	26.26	69.62	77.05	29.05	80.07	84.32	14.93
	ZSCL [49]	66.87	66.00	19.08	67.32	75.65	27.45	66.53	75.05	25.13
	MoE-Adapters [45]	69.90	74.47	27.11	64.50	77.18	34.98	80.09	83.02	14.36
	ConStruct-VL [33]	87.27	86.98	6.14	89.01	91.87	5.80	83.73	86.34	6.47
	ZAF (Ours)	**90.05**	**89.45**	**3.32**	**92.49**	**92.39**	**1.97**	**89.13**	**90.03**	**3.93**
	Improvement	***2.78***	***2.47***	***2.82***	***3.48***	***0.52***	***3.83***	***5.40***	***3.69***	***2.54***
BLIP w/ CapFilt-L	Joint Learning	93.72	-	-	95.31	-	-	92.90	-	-
	Continual-FT [7]	67.20	74.85	28.02	70.05	75.17	23.99	71.95	79.31	22.18
	LoRA [10]	71.97	76.07	25.27	69.97	77.52	28.49	79.66	82.36	13.78
	Layered-LoRA [33]	76.66	76.27	19.20	70.43	78.00	27.16	81.89	82.66	11.18
	LwF [21]	73.39	75.42	23.81	70.02	77.62	28.47	79.83	84.21	15.63
	ZSCL [49]	62.90	64.29	22.06	67.12	76.21	27.14	68.13	77.15	24.67
	MoE-Adapters [45]	69.76	73.29	27.34	63.99	76.19	35.34	80.01	84.10	14.43
	ConStruct-VL [33]	85.16	87.61	8.75	88.95	90.69	5.22	83.33	85.57	6.28
	ZAF (Ours)	**89.61**	**89.65**	**4.18**	**92.53**	**92.20**	**1.72**	**89.43**	**90.20**	**3.02**
	Improvement	***4.45***	***2.04***	***4.57***	***3.58***	***1.51***	***3.50***	***6.10***	***4.63***	***3.26***
BLIP w/ NLVR	Joint Learning	93.37	-	-	95.07	-	-	92.36	-	-
	Continual-FT [7]	67.23	73.60	27.96	73.40	78.60	20.55	73.19	80.58	20.69
	LoRA [10]	69.55	75.03	27.25	68.73	78.03	29.62	75.63	81.87	19.37
	Layered-LoRA [33]	80.62	79.89	13.92	73.03	81.12	24.99	83.73	84.26	9.29
	LwF [21]	73.00	77.26	23.12	71.11	79.39	27.09	82.10	84.69	11.24
	ZSCL [49]	60.27	67.94	28.48	65.82	78.06	27.68	62.03	74.33	31.20
	MoE-Adapters [45]	72.50	74.81	23.74	67.09	76.54	31.83	79.05	84.21	15.58
	ConStruct-VL [33]	85.97	87.00	6.94	86.96	90.47	7.91	84.36	85.93	5.36
	ZAF (Ours)	**89.67**	**89.30**	**3.38**	**91.78**	**91.74**	**2.02**	**88.74**	**89.03**	**2.67**
	Improvement	***3.70***	***2.30***	***3.56***	***4.82***	***1.27***	***5.89***	***4.38***	***3.10***	***2.69***

BLIP	**7 Task VG+VAW**	**7 Task VG**	**5 Task VAW**
Zero-shot Accuracy	50.74	50.83	50.42
Final Forgetting w/o L_{ZS}	20.11	32.63	12.69
Final Forgetting w/ L_{ZS}	3.32	1.97	3.93

BLIP w/ CapFilt-L	**7 Task VG+VAW**	**7 Task VG**	**5 Task VAW**
Zero-shot Accuracy	49.60	50.88	49.23
Final Forgetting w/o L_{ZS}	20.66	23.54	14.08
Final Forgetting w/ L_{ZS}	4.18	1.72	3.02

BLIP w/ NLVR	**7 Task VG+VAW**	**7 Task VG**	**5 Task VAW**
Zero-shot Accuracy	67.89	68.82	70.39
Final Forgetting w/o L_{ZS}	17.30	21.55	10.18
Final Forgetting w/ L_{ZS}	3.38	2.02	2.67

Results

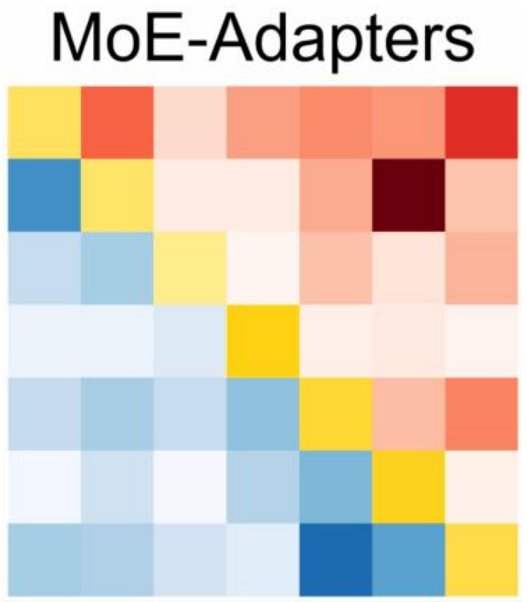
Table 2: Comparison of training complexity among various CL methods across three benchmarks.

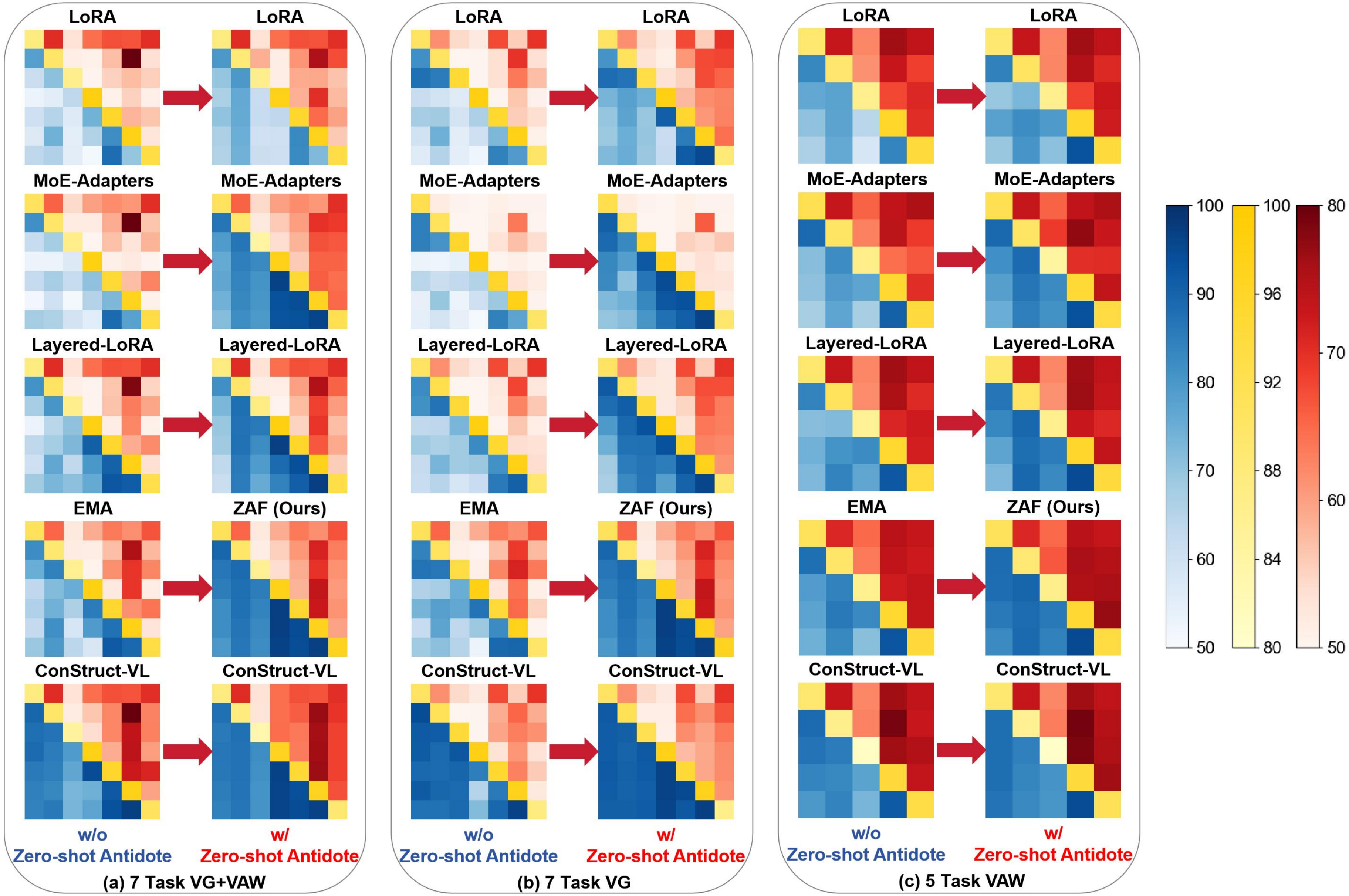
Method	Model Size (M)	Train Params (M)	Train Times (h)		
			7 Task VG+VAW	7 Task VG	5 Task VAW
Continual-FT [7]	223.94	223.94	5.57	2.25	2.90
LoRA [10]	230.13	6.19	4.64	1.86	2.82
Layred-LoRA [33]	230.13 ∼ 267.29	6.19	11.10	4.04	5.60
LwF [21]	230.13 ∼ 236.33	6.19	8.13	5.19	6.34
MoE-Adapters [45]	251.04	27.10	6.01	2.82	4.19
ZSCL [49]	223.94	223.94	11.83	6.44	7.83
ConStruct-VL [33]	230.13 ∼ 267.29	6.19	247.78	102.59	73.24
ZAF (Ours)	236.33	6.19	8.35	5.81	6.67
Training Speed Acceleration			**247.78/8.35 ≈ 29.67**	**102.59/5.81 ≈ 17.66**	**73.24/6.67 ≈ 10.98**

Table 3: Comparison of plugin performance for various CL methods across three benchmarks.

Method	7 Task VG+VAW			7 Task VG			5 Task VAW		
	FAA (↑)	CAA (↑)	FFM (↓)	**FAA** (↑)	CAA (↑)	FFM (↓)	**FAA** (↑)	CAA (↑)	FFM (↓)
Joint Learning	93.37	-	-	95.07	-	-	92.36	-	-
LoRA [10]	69.55	75.03	27.25	68.73	78.03	29.62	75.63	81.87	19.37
w/ Zero-shot Antidote	**72.47**	**77.78**	**23.15**	**79.12**	**83.47**	**16.85**	**81.55**	**84.01**	**12.39**
Layered-LoRA [33]	80.62	79.89	13.92	73.03	81.12	24.99	83.73	84.26	9.29
w/ Zero-shot Antidote	**83.81**	**85.11**	**10.37**	**84.10**	**88.75**	**11.26**	**86.66**	**87.02**	**4.98**
MoE-Adapters [45]	72.50	74.80	23.74	67.09	76.54	31.83	79.05	84.21	15.58
w/ Zero-shot Antidote	**86.78**	**86.78**	**6.29**	**83.62**	**87.89**	**12.27**	**85.55**	**87.47**	**6.67**
ConStruct-VL [33]	85.97	87.00	6.94	86.96	90.47	7.91	84.36	85.93	5.36
w/ Zero-shot Antidote	**89.60**	**88.13**	**1.26**	**92.05**	**92.06**	**0.88**	**86.94**	**86.72**	**1.22**
EMA-LoRA	77.78	80.96	17.30	75.02	82.22	21.55	83.08	86.21	10.18
w/ Zero-shot Antidote (ZAF)	**89.67**	**89.30**	**3.38**	**91.78**	**91.74**	**2.02**	**88.74**	**89.03**	**2.67**
Average Improvement	*7.18*	*5.88*	*8.94*	*11.96*	*7.10*	*14.52*	*4.72*	*2.35*	*6.37*

Visual Analysis





Summary

Empirical Finding:

A model's stability in zero-shot predictions can reflect its anti-forgetting capabilities.

Theoretical Study:

The model $\mathcal{M}^t$ has consistent upper bounds on the generalization errors for both previously learned and future tasks.

New CL Paradigm:

Stability vs. Plasticity $\xrightarrow{\textit{Wild data}}$ Zero-Shot Stability & Plasticity

CL Algorithm-independent
Network Architecture-independent
Foundation Model-independent
Task-independent
Task Boundary-independent

Thank you !

Email: gaozijian19@nudt.edu.cn

Project Page: https://github.com/Zi-Jian-Gao/Stabilizing-Zero-Shot-Prediction-ZAF