NEURAL INFORMATION PROCESSING SYSTEMS

Align Your Prompts: Test-Time Prompting using Distribution Alignment for Zero-Shot Generalization

Jameel Hassan[1], Hanan Gani[1], Noor Hussein[1], Uzair Khattak[1], Muzammal Naseer[1], Fahad Shahbaz Khan[1,2], Salman Khan[1,3]

[1]Mohamed Bin Zayed University of Artificial Intelligence [2]Linköping University [3]Autralian National University

جـــامـــعـــة مـحـمـد بـــن زايـــد للـــذكــاء الاصـطـنـــاعـــي
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

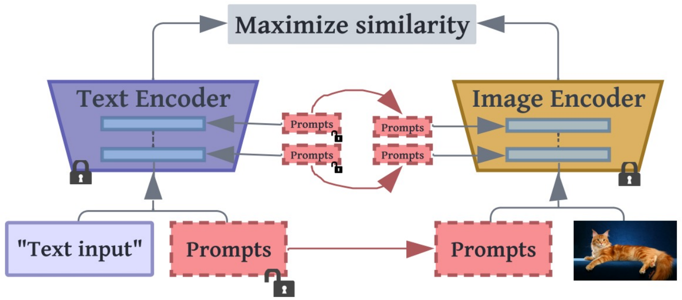
LINKÖPING UNIVERSITY

Australian National University

Background

- Foundational Vision-Language (VL) models
  - Pre-trained models on large scale image-text pairs
  - Good generalization to unseen data

- Multi-modal prompt learning
  - Prompt learning Preserves model generalization, instead of overfitting

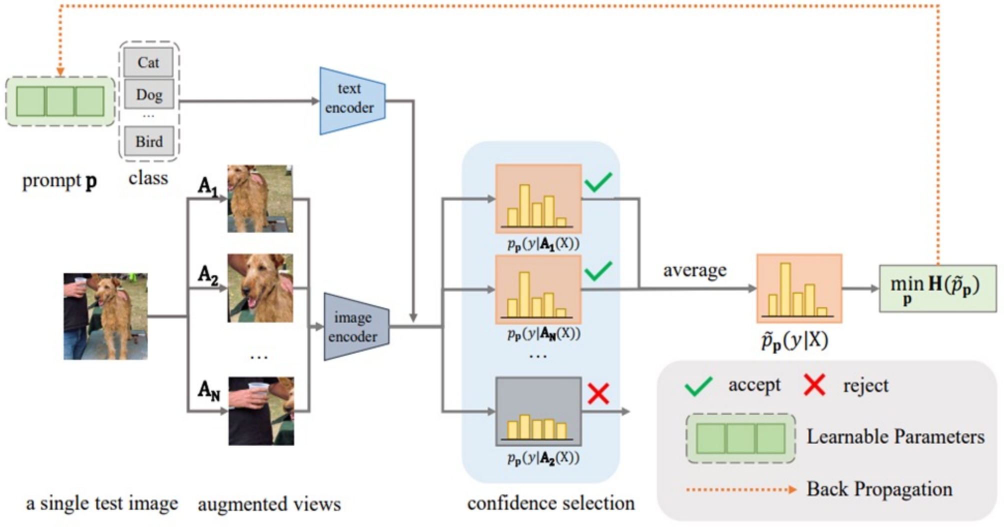


- Test-Time Prompt Tuning for efficient adaptation
  - An effective lightweight adaptation mechanism at test time for foundation models

Khattak et al. "Maple: Multi-modal prompt learning.

Problem Statement

- Adapting large scale Vision-Language models like CLIP at test time.
  - Light weight adaptation
  - Using a single sample for adaptation



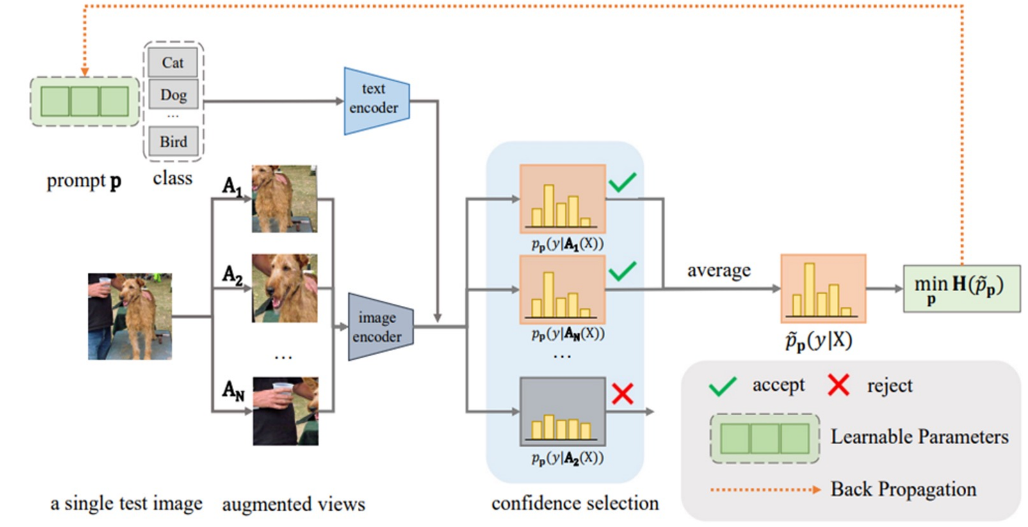
Test-time prompt tuning (Shu et al. 2022)

Existing solution

- Prompt update using entropy minimization across augmented views
  - Fails to handle the distribution shift in test data

Problem Statement

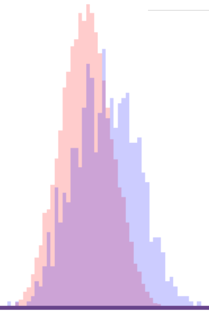
- Adapting large scale Vision-Language models like CLIP at test time.
  - Light weight adaptation
  - Using a single sample for adaptation



Test-time prompt tuning (Shu et al. 2022)

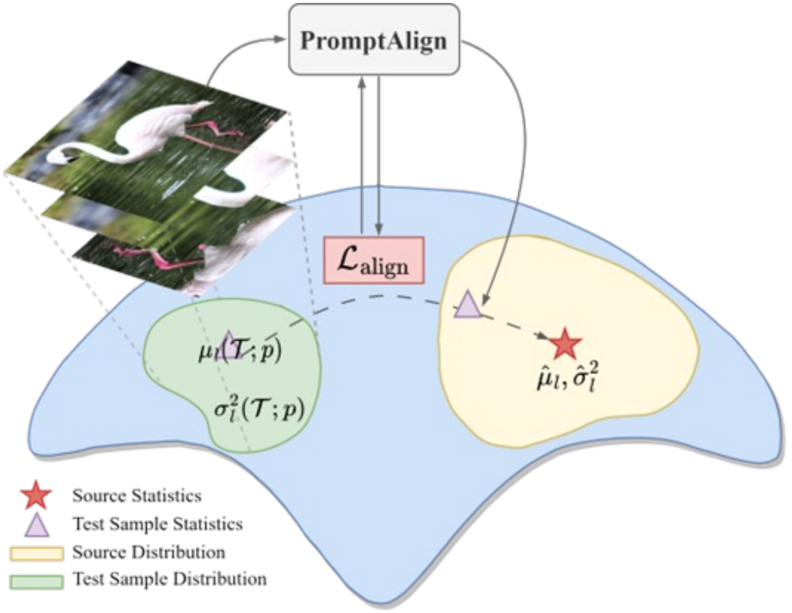
Existing solution

- Prompt update using entropy minimization across augmented views
  - Fails to handle the distribution shift in test data



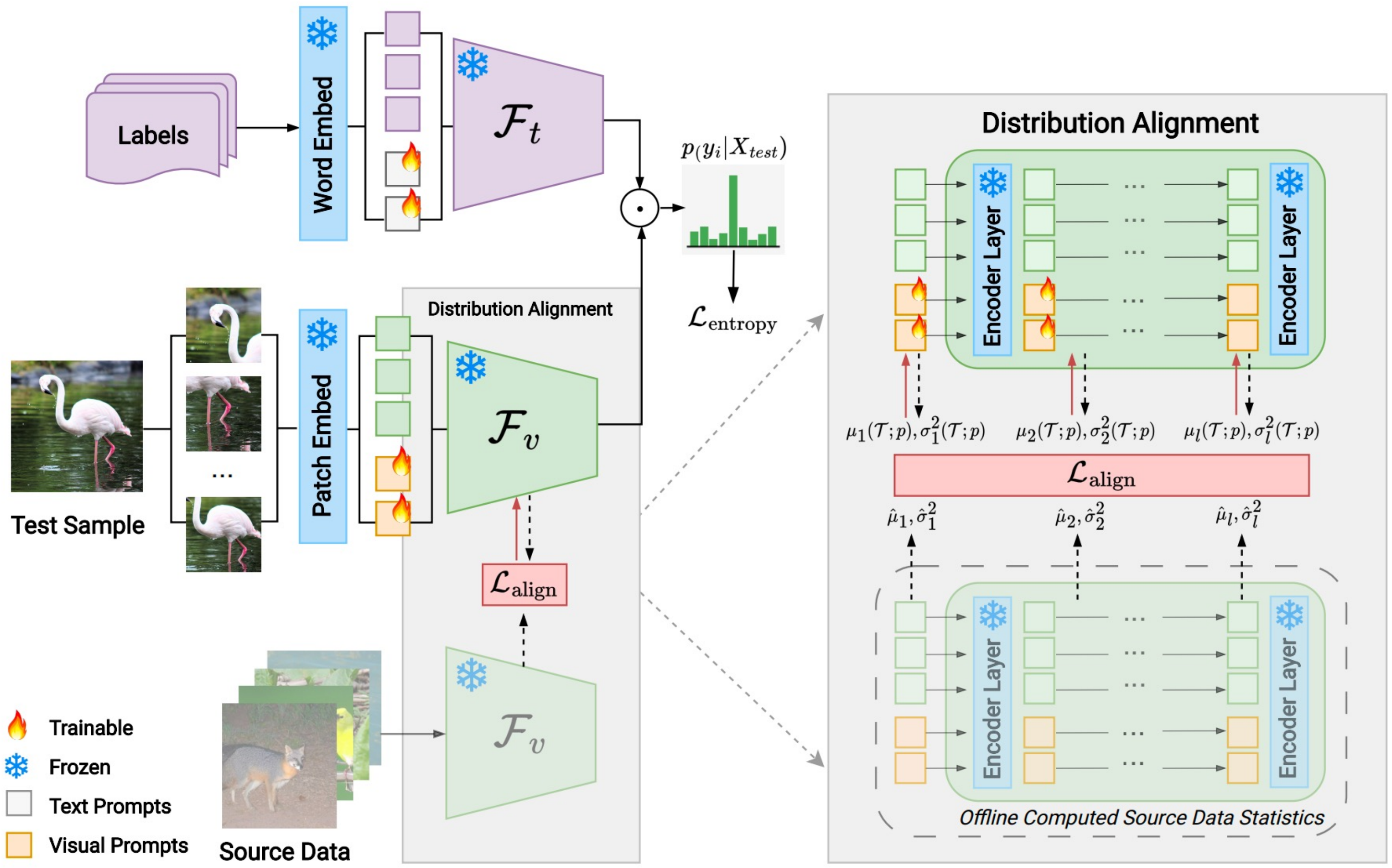
Prompt Align

- We explicitly handle the distibution shift in test data
  - The distribution aware prompts helps narrowing the distribution gap in the test domain
- We formulate a distribution alignment loss utilizing offline computed source data statistics
  - The test sample token distributions are aligned with the source data token distributions
- We study and validate the use of ImageNet as a proxy dataset for CLIP pre-training



(a) Our proposed PromptAlign method

Prompt Align Design



Token distribution statistics:

$$\boldsymbol{\mu}_l(\mathcal{T};\boldsymbol{p}) = \frac{1}{N_k}\sum_{\mathrm{x}\in\mathcal{H}(X)} \tilde{\boldsymbol{X}}^{\boldsymbol{p}}_{l,\mathrm{x}} \quad ,$$

$$\boldsymbol{\sigma}_l^2(\mathcal{T};\boldsymbol{p}) = \frac{1}{N_k}\sum_{\mathrm{x}\in\mathcal{H}(X)}\left(\tilde{\boldsymbol{X}}^{\boldsymbol{p}}_{l,\mathrm{x}} - \boldsymbol{\mu}_l(\mathcal{T};\boldsymbol{p})\right)^2,$$

$$\hat{\boldsymbol{\mu}}_l = \boldsymbol{\mu}_l(\mathcal{D},\theta_v) \quad \text{and} \quad \hat{\boldsymbol{\sigma}}_l^2 = \boldsymbol{\sigma}_l^2(\mathcal{D},\theta_v)$$

Alignment loss:

$$\mathcal{L}_{\text{align}} = \frac{1}{L}\sum_{l=1}^{L}\Big(\|\boldsymbol{\mu}_l(\mathcal{T};\boldsymbol{p}) - \hat{\boldsymbol{\mu}}_l\|_1 + \|\boldsymbol{\sigma}_l^2(\mathcal{T};\boldsymbol{p}) - \hat{\boldsymbol{\sigma}}_l^2\|_1\Big).$$

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{entropy}} + \beta\mathcal{L}_{\text{align}}$$

Experiments

We conduct experiments on two generalization tasks

- Domain Generalization
    - Trained on ImageNet dataset
    - Evaluated on 4 Out of Distribution variants of ImageNet and PUG ImageNet variant
- Cross-dataset evaluation
    - Trained on ImageNet and tested on the 11 cross datasets

Experiments: Domain Generalization

Table 1: **Effect of token distribution alignment strategy for domain generalization.** The base model MaPLe is trained on ImageNet and evaluated on datasets with domain shifts.

	Imagenet V2	Imagenet Sketch	Imagenet A	Imagenet R	OOD Avg.
MaPLe [18]	64.07	49.15	50.90	76.98	60.28
MaPLe+TPT	64.87	48.16	58.08	78.12	62.31
PromptAlign	**65.29**	**50.23**	**59.37**	**79.33**	**63.55**

	Imagenet V2	Imagenet Sketch	Imagenet A	Imagenet R	OOD Avg.
CLIP [28]	60.86	46.09	47.87	73.98	57.20
CLIP+TPT [32]	64.35	47.94	54.77	77.06	60.81
CoOp [46]	64.20	47.99	49.71	75.21	59.28
CoOp+TPT [32]	**66.83**	49.29	57.95	77.27	62.84
Co-CoOp [45]	64.07	48.75	50.63	76.18	59.91
Co-CoOp+TPT [32]	64.85	48.27	58.47	78.65	62.61
PromptAlign	65.29	**50.23**	**59.37**	**79.33**	**63.55**

Experiments: Domain Generalization

Evaluation on the recent Photorealistic Unreal Graphics (PUG) dataset

Table 3: **Effect of token distribution alignment strategy for domain generalization.** The base model MaPLe is trained on ImageNet and evaluated on PUG-ImageNet.

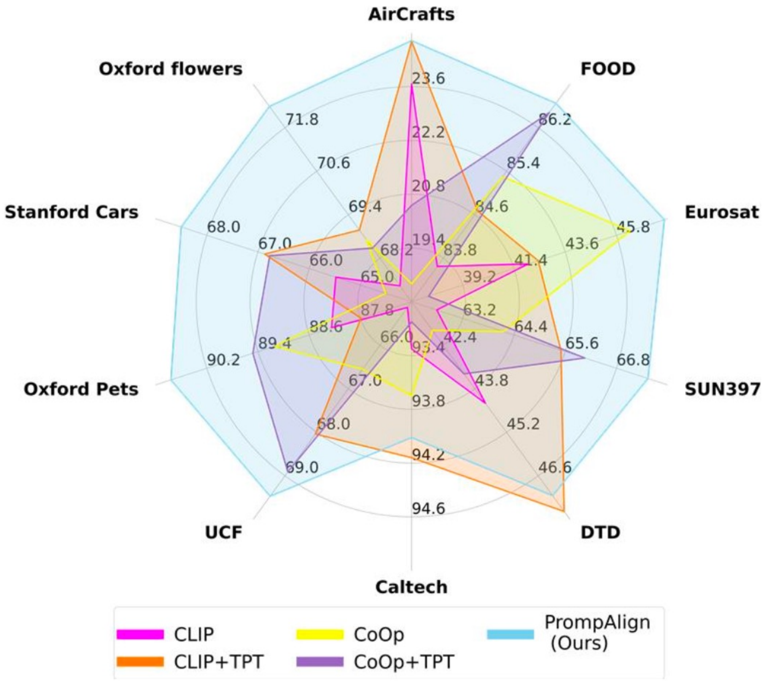
	Camera (Yaw/ Pitch/ Roll)	Pose (Yaw/ Pitch/ Roll)	Scale	Texture	Lighting	Worlds
MaPLe [18]	48.73/ 39.93/ 32.13	48.10/ 28.40/ 27.80	46.90	37.90	15.50	32.13
MaPLe+TPT	57.04/ 45.99/ 39.23	56.26/ 35.64/ 33.26	54.87	43.73	22.52	42.00
PromptAlign	**58.14/ 46.93/ 40.45**	**57.43/ 36.31/ 34.32**	**56.18**	**44.97**	**23.06**	**43.24**

Bordes et al. "Pug: Photorealistic and semantically controllable synthetic data for representation learning."

Experiments: Cross Dataset

Table 4: **Comparison of PromptAlign in cross-dataset evaluation.** Prompt learning methods are trained on ImageNet and evaluated on cross-datasets.

	Caltech	Pets	Cars	Flowers	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
CLIP [28]	93.35	88.25	65.48	67.44	83.65	23.67	62.59	44.27	42.01	65.13	63.58
CLIP+TPT [32]	**94.16**	87.79	66.87	68.98	84.67	24.78	65.50	**47.75**	42.44	68.04	65.10
CoOp [46]	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
CoCoOp [45]	93.79	90.46	64.90	70.85	83.97	22.29	66.89	45.45	39.23	68.44	64.63
ProDA [44]	86.70	88.20	60.10	77.50	80.80	22.20	-	50.90	58.50	-	65.62
MaPLe	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	**48.06**	68.69	66.30
MaPLe+TPT	93.59	90.72	66.50	72.37	86.64	24.70	**67.54**	45.87	47.80	69.19	66.50
PromptAlign	94.01	**90.76**	**68.50**	**72.39**	**86.65**	**24.80**	**67.54**	47.24	47.86	**69.47**	**66.92**



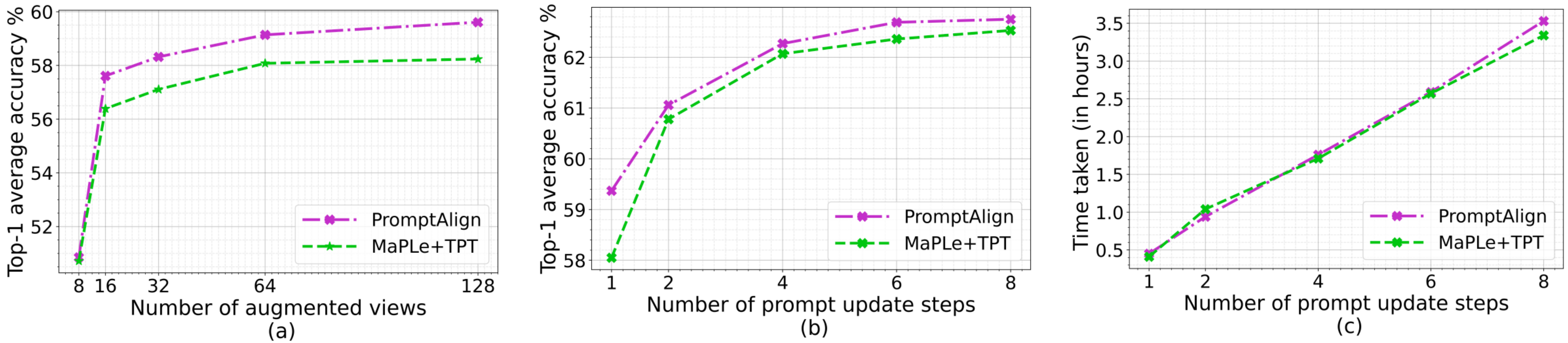
Analysis of Distribution Alignment

Effect of the distribution alignment loss

Method	Entropy loss	Distribution alignment	Top-1 Acc.
MaPLe [18]	✗	✗	50.90
MaPLe+TPT	✓	✗	58.08
PromptAlign†	✗	✓	50.85
PromptAlign	✓	✓	**59.37**

Analysis of Distribution Alignment

Effect of distribution alignment with number of augmented views and prompt update steps



Conclusion

- We introduce a distribution alignment loss to enhance test-time adaptation of Vision-Language models for zero-shot generalization.
- The proposed method bridges the gap between the test sample and source distributions explicitly, facilitated by multi-modal prompts.
- Validates ImageNet as a valid proxy source dataset for the distribution alignment loss
- Extensive experiments show improvement in domain generalization and cross dataset evaluation.