DPA: Dual Prototypes Alignment for Unsupervised Adaptation of Vision-Language Models

[1, 2]Eman Ali [1]Sathira Silva [1]Muhammad Haris Khan

[1]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

[2]Alexandria University, Egypt

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

Background & Motivation

- CLIP excels in zero-shot classification but struggles with adaptation.
- Domain gaps between pre-training data and task-specific images hinder performance.
- Unsupervised learning is key for costly or sensitive domains like security and healthcare.

Background & Motivation

- CLIP excels in zero-shot classification but struggles with adaptation.
- Domain gaps between pre-training data and task-specific images hinder performance.
- Unsupervised learning is key for costly or sensitive domains like security and healthcare.

Background & Motivation

- CLIP excels in zero-shot classification but struggles with adaptation.
- Domain gaps between pre-training data and task-specific images hinder performance.
- Unsupervised learning is key for costly or sensitive domains like security and healthcare.

Background & Motivation

- CLIP excels in zero-shot classification but struggles with adaptation.
- Domain gaps between pre-training data and task-specific images hinder performance.
- Unsupervised learning is key for costly or sensitive domains like security and healthcare.

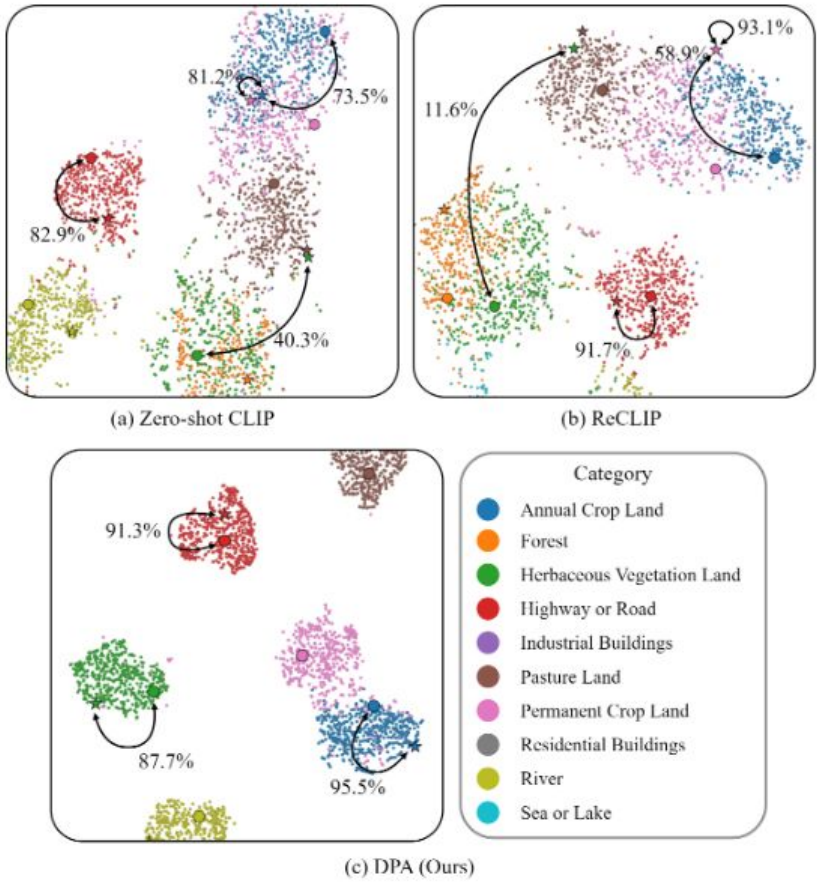
Problem Statement

How can we adapt CLIP to target domains using unlabeled data while mitigating challenges like noisy pseudo-labels and domain gap for enhanced classification?

Contributions

- Proposed a novel **dual-modality prototypic alignment** framework for unsupervised adaptation of VLMs.
- Introduced ranking pseudo-labels to mitigate noise.
- Achieved significant performance enhancements across 13 downstream tasks.

Contributions



★ textual prototypes
● image prototypes

- Proposed a novel **dual-modality prototypic alignment** framework for unsupervised adaptation of VLMs.
- Introduced ranking pseudo-labels to mitigate noise.
- Achieved significant performance enhancements across 13 downstream tasks.

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

Limitations of Existing Approaches

- **Unsupervised Adaptation Techniques:**
  - Pseudo-labeling methods suffer from noise and modality gaps.
  - Projection spaces and label propagation reduce gaps but are costly and inefficient.
- **Key Gaps in Current Methods:**
  - Limited in inductive settings.
  - Struggle with scaling to large-class datasets.

Limitations of Existing Approaches

- **Unsupervised Adaptation Techniques:**
  - Pseudo-labeling helps but suffers from noise and modality gaps.
  - Projection spaces and label propagation reduce gaps but are costly and inefficient.
- **Key Gaps in Current Methods:**
  - Limited in inductive settings.
  - Struggle with scaling to large-class datasets.

Limitations of Existing Approaches

- **Unsupervised Adaptation Techniques:**
  - Pseudo-labeling helps but suffers from noise and modality gaps.
  - Projection spaces and label propagation reduce gaps but are costly and inefficient.
- **Key Gaps in Current Methods:**
  - Limited in inductive settings.
  - Struggle with scaling to large-class datasets.

Our Solution: DPA

- **Core Components:**
  - Dual Prototypes
    - Convex Combination for Pseudo-Labels
    - Alignment of Visual* and Textual Prototypes
    - Fine-tuning LayerNorm layers

Our Solution: DPA

- **Dual Prototypes:**
  - Image prototypes serve as non-parametric classifiers, that are tolerant to noise.
  - Textual prototypes initialized using class-name embeddings, are fine-tuned.

Our Solution: DPA

- **Dual Prototypes:**
  - Image prototypes serve as non-parametric classifiers, that are tolerant to noise.
  - Textual prototypes initialized using class-name embeddings, are fine-tuned.

Our Solution: DPA

- **Convex Combination for Pseudo-Labels:**
  - Combines image and textual prototypes for accurate pseudo-labeling.
- **Alignment of Visual* and Textual Prototypes:**
  - Bridges modality gaps by aligning textual prototypes with image prototypes.
- **Fine-tuning LayerNorm layers:**
  - Parameter-efficient fine-tuning of CLIP visual backbone.

Our Solution: DPA

- **Convex Combination for Pseudo-Labels:**
  - Combines outputs of image and textual prototypes for accurate pseudo-labeling.
- **Alignment of Visual* and Textual Prototypes:**
  - Aligns textual prototypes with image prototypes to close the modality gap.
- **Fine-tuning LayerNorm layers:**
  - Parameter-efficient fine-tuning of CLIP visual backbone.

* we use visual prototypes and image prototypes interchangeably.

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

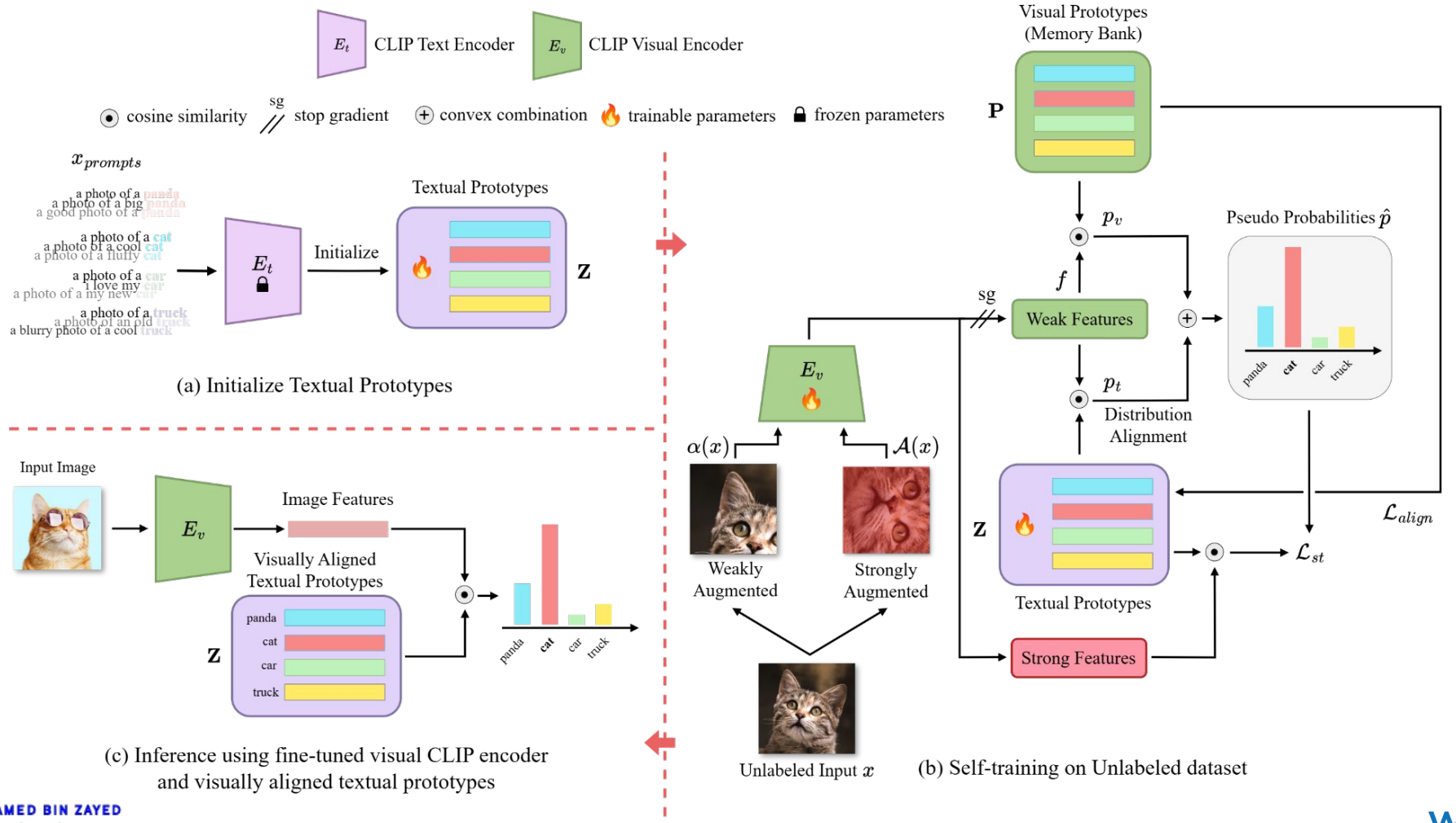
Our Solution: DPA

- **Convex Combination for Pseudo-Labels:**
  - Combines outputs of image and textual prototypes for accurate pseudo-labeling.
- **Alignment of Visual* and Textual Prototypes:**
  - Aligns textual prototypes with image prototypes to close the modality gap.
- **Fine-tuning LayerNorm layers:**
  - Parameter-efficient fine-tuning of CLIP visual backbone.

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

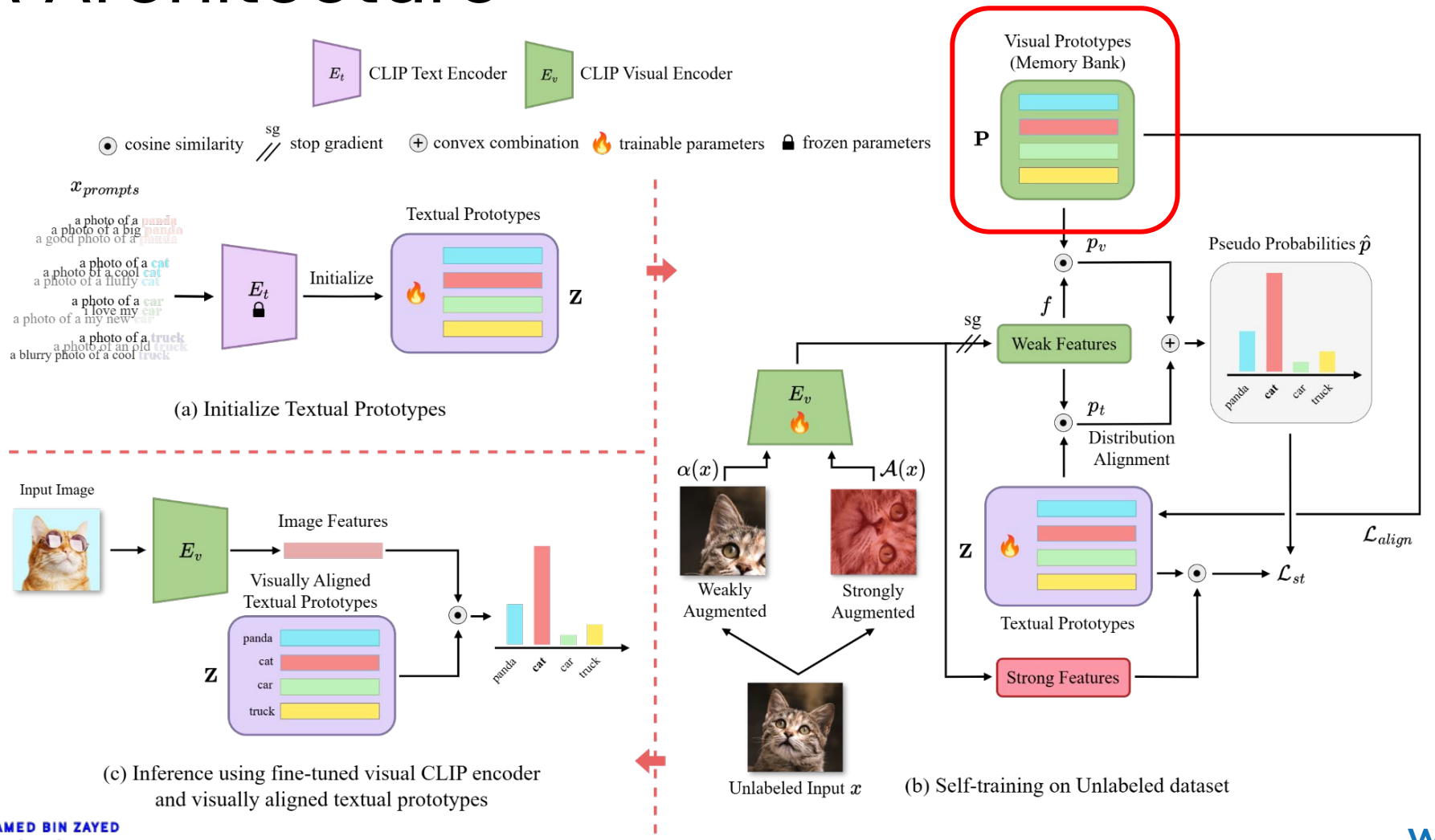
DPA Architecture



MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

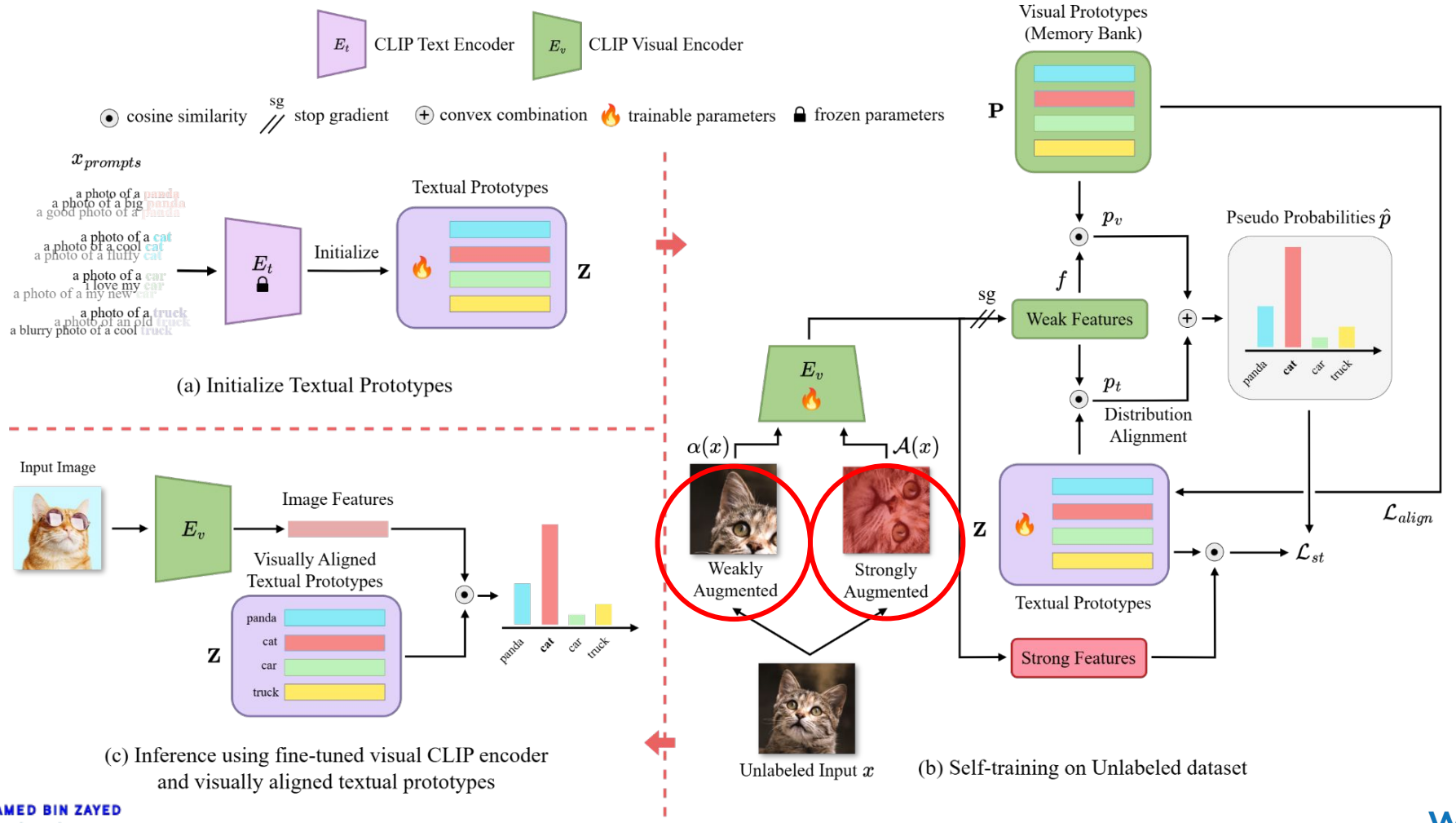
DPA Architecture



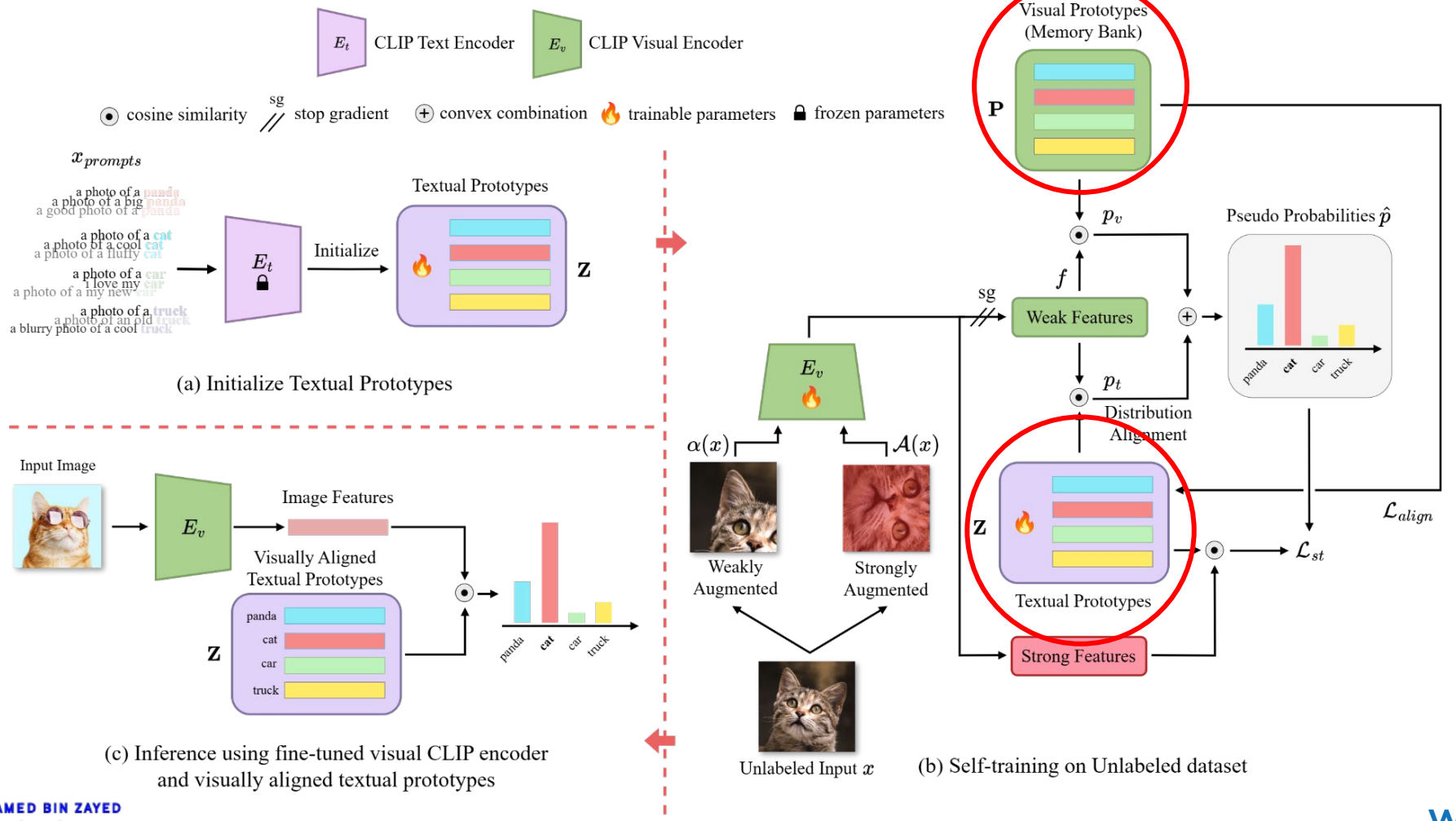
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

DPA Architecture



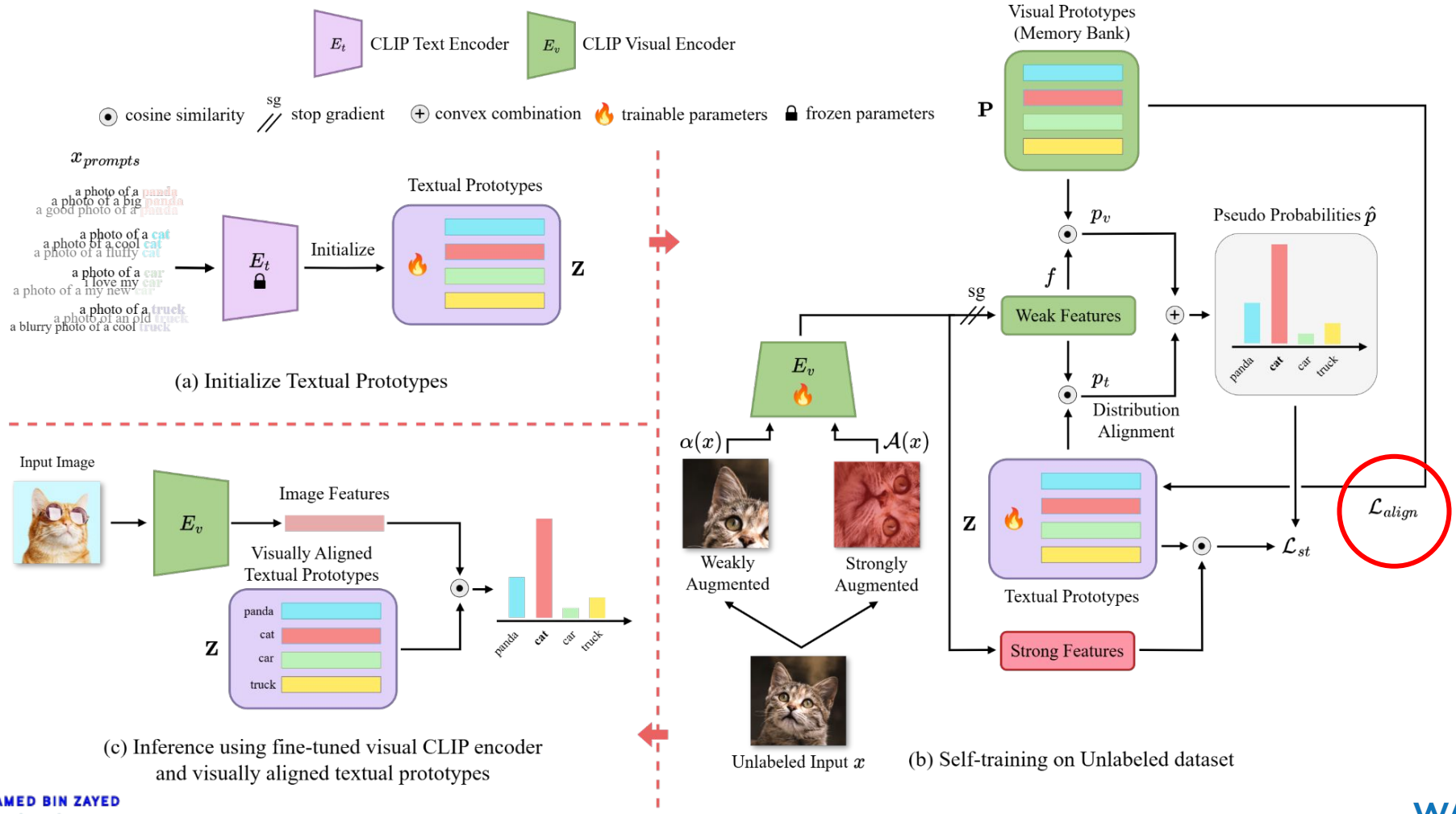
DPA Architecture



MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

DPA Architecture



MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

Datasets and State-of-the-art Comparison

- **Datasets:**
  - Evaluated on 13 diverse datasets spanning:
    - **General classification** (Ex: ImageNet).
    - **Specialized domains** (Ex: EuroSAT).
    - **Fine-grained tasks** (Ex: OxfordPets).
    - **Scene recognition** (Ex: UCF101).
- Compared against 4 SOTA methods: **CLIP**, **UPL**, **POUF**, **LaFTer**.

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

Experiments: Comparative Results

Method	ImgNet	Caltech	DTD	ESAT	FGVCA	Food	Flower	OxPets	SUN	StCars	CIFAR10	CIFAR100	UCF	Avg
Zero-shot CLIP [38]	63.30	90.69	44.42	43.84	19.50	82.40	66.46	87.50	61.99	58.74	89.80	65.10	64.20	64.46
UPL [18]	58.22	92.36	45.37	51.88	17.07	84.25	67.40	83.84	62.12	49.41	91.26	67.41	62.04	64.05
POUF [45]	52.20	94.10	46.10	62.90	18.20	82.10	67.80	87.80	60.00	57.70	90.50	62.00	61.20	64.82
LaFTer [33]	61.63	94.39	50.32	69.96	19.86	82.45	72.43	84.93	65.87	57.44	94.57	69.79	65.08	68.36
DPA	**64.64**	**96.06**	**55.69**	**80.04**	**20.67**	**84.76**	**75.56**	**90.71**	**68.13**	**62.62**	**95.97**	**76.47**	**68.49**	**72.29**

Table 1. Comparison of state-of-the-art unsupervised adaptation methods using the ViT-B/32 backbone.

+4-8% overall improvements to top-1 accuracy compared to SOTA

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

Experiments: Comparative Results

Method	ImgNet	Caltech	DTD	ESAT	FGVCA	Food	Flower	OxPets	SUN	StCars	CIFAR10	CIFAR100	UCF	Avg
Zero-shot CLIP [38]	63.30	90.69	44.42	43.84	19.50	82.40	66.46	87.50	61.99	58.74	89.80	65.10	64.20	64.46
UPL [18]	58.22	92.36	45.37	51.88	17.07	84.25	67.40	83.84	62.12	49.41	91.26	67.41	62.04	64.05
POUF [45]	52.20	94.10	46.10	62.90	18.20	82.10	67.80	87.80	60.00	57.70	90.50	62.00	61.20	64.82
LaFTer [33]	61.63	94.39	50.32	69.96	19.86	82.45	72.43	84.93	65.87	57.44	94.57	69.79	65.08	68.36
DPA	**64.64**	**96.06**	**55.69**	**80.04**	**20.67**	**84.76**	**75.56**	**90.71**	**68.13**	**62.62**	**95.97**	**76.47**	**68.49**	**72.29**

Table 1. Comparison of state-of-the-art unsupervised adaptation methods using the ViT-B/32 backbone.

+8.24%

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

Experiments: Comparative Results

Method	ImgNet	Caltech	DTD	ESAT	FGVCA	Food	Flower	OxPets	SUN	StCars	CIFAR10	CIFAR100	UCF	Avg
Zero-shot CLIP [38]	63.30	90.69	44.42	43.84	19.50	82.40	66.46	87.50	61.99	58.74	89.80	65.10	64.20	64.46
UPL [18]	58.22	92.36	45.37	51.88	17.07	84.25	67.40	83.84	62.12	49.41	91.26	67.41	62.04	64.05
POUF [45]	52.20	94.10	46.10	62.90	18.20	82.10	67.80	87.80	60.00	57.70	90.50	62.00	61.20	64.82
LaFTer [33]	61.63	94.39	50.32	69.96	19.86	82.45	72.43	84.93	65.87	57.44	94.57	69.79	65.08	68.36
DPA	**64.64**	**96.06**	**55.69**	**80.04**	**20.67**	**84.76**	**75.56**	**90.71**	**68.13**	**62.62**	**95.97**	**76.47**	**68.49**	**72.29**

Table 1. Comparison of state-of-the-art unsupervised adaptation methods using the ViT-B/32 backbone.

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

Experiments: Comparative Results

Method	ImgNet	Caltech	DTD	ESAT	FGVCA	Food	Flower	OxPets	SUN	StCars	CIFAR10	CIFAR100	UCF	Avg
Zero-shot CLIP [38]	63.30	90.69	44.42	43.84	19.50	82.40	66.46	87.50	61.99	58.74	89.80	65.10	64.20	64.46
UPL [18]	58.22	92.36	45.37	51.88	17.07	84.25	67.40	83.84	62.12	49.41	91.26	67.41	62.04	64.05
POUF [45]	52.20	94.10	46.10	62.90	18.20	82.10	67.80	87.80	60.00	57.70	90.50	62.00	61.20	64.82
LaFter [33]	61.63	94.39	50.32	69.96	19.86	82.45	72.43	84.93	65.87	57.44	94.57	69.79	65.08	68.36
DPA	**64.64**	**96.06**	**55.69**	**80.04**	**20.67**	**84.76**	**75.56**	**90.71**	**68.13**	**62.62**	**95.97**	**76.47**	**68.49**	**72.29**

Table 1. Comparison of state-of-the-art unsupervised adaptation methods using the ViT-B/32 backbone.

+7.47%

Experiments: Comparative Results

Method	ImgNet	Caltech	DTD	ESAT	FGVCA	Food	Flower	OxPets	SUN	StCars	CIFAR10	CIFAR100	UCF	Avg
Zero-shot CLIP [38]	63.30	90.69	44.42	43.84	19.50	82.40	66.46	87.50	61.99	58.74	89.80	65.10	64.20	64.46
UPL [18]	58.22	92.36	45.37	51.88	17.07	84.25	67.40	83.84	62.12	49.41	91.26	67.41	62.04	64.05
POUF [45]	52.20	94.10	46.10	62.90	18.20	82.10	67.80	87.80	60.00	57.70	90.50	62.00	61.20	64.82
LaFTer [33]	61.63	94.39	50.32	69.96	19.86	82.45	72.43	84.93	65.87	57.44	94.57	69.79	65.08	68.36
DPA	**64.64**	**96.06**	**55.69**	**80.04**	**20.67**	**84.76**	**75.56**	**90.71**	**68.13**	**62.62**	**95.97**	**76.47**	**68.49**	**72.29**

Table 1. Comparison of state-of-the-art unsupervised adaptation methods using the ViT-B/32 backbone.

+3.93%

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

Experiments: Comparative Results

Method	Caltech	DTD	ESAT	FGVCA	Food	Flower	OxPets	StCars	CIFAR10	CIFAR100	UCF	Avg
Zero-shot CLIP	90.69	44.42	43.84	19.50	82.40	66.46	87.50	58.74	89.80	65.10	64.20	64.79
Base	93.57	48.99	61.94	19.20	84.10	68.45	90.24	59.25	95.95	73.55	65.45	69.15
Center	95.44	55.53	70.56	19.80	84.65	75.27	**90.71**	61.53	95.96	76.01	67.30	72.07
Center+w	95.46	54.54	**80.06**	19.56	84.63	75.44	90.49	61.19	**95.97**	75.92	67.51	72.80
Center+w+Align (DPA)	**96.06**	**55.69**	80.04	**20.67**	**84.76**	**75.56**	**90.71**	**62.62**	**95.97**	**76.47**	**68.49**	**73.37**

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

Experiments: Efficiency comparison

- DPA outperforms baselines with higher efficiency, fewer parameters, and lower computational complexity.

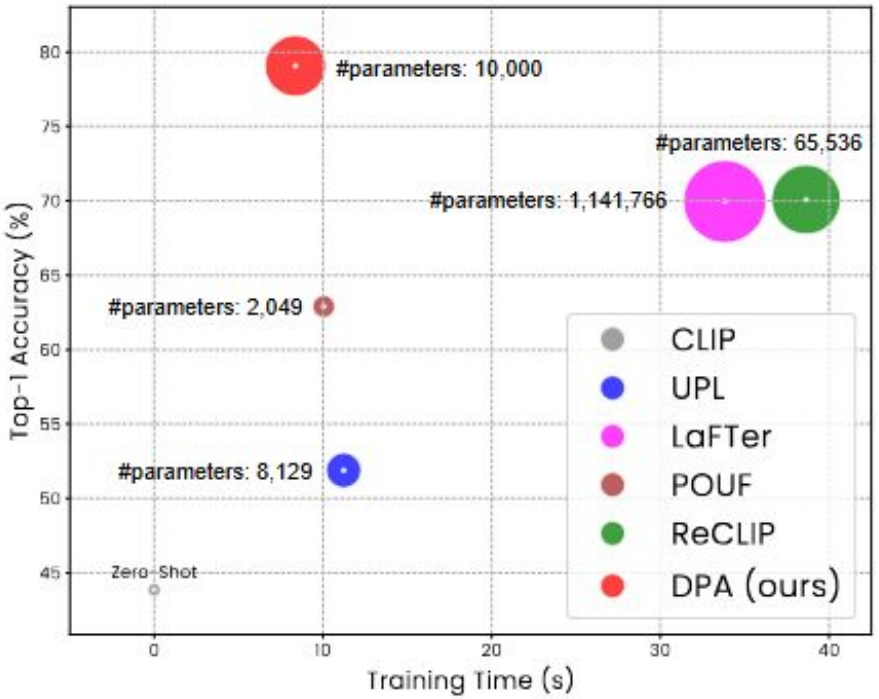
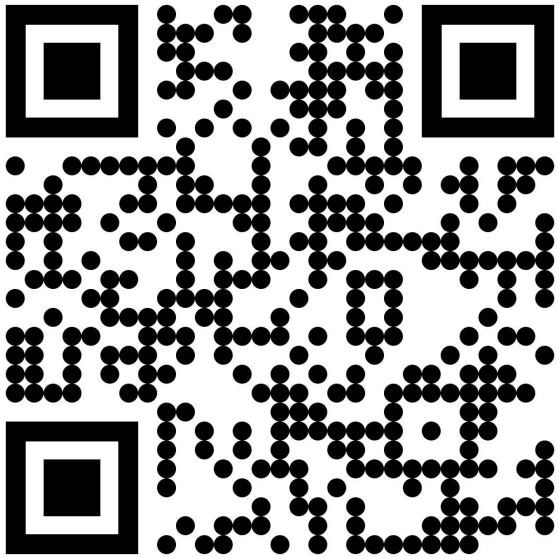


Figure 6. Efficiency comparison of DPA with baselines. The radius of each circle represents trainable parameters in each method.

Conclusion

- **DPA: Bridging the Domain Gap in VLMs**
  - **Dual Prototypes:**
    - Novel fusion of two distinct classifiers via a convex combination.
    - Improves pseudo-label robustness during self-training.
  - **Robust Self-Training:**
    - Effective pseudo-label ranking stabilizes early training.
  - **Enhanced Alignment:**
    - Strengthens visual-textual prototype adaptation for better domain alignment.
- **Performance Highlights:**
  - Significant improvements in pseudo-labeling accuracy.
  - Outperforms zero-shot CLIP and state-of-the-art methods on **13 diverse vision tasks.**

Thank You!



⬆️ Link to the paper

Contact: Eman Ali <eman.ali@mbzuai.ac.ae>
Sathira Silva <sathira.silva@mbzuai.ac.ae>
Haris Khan <muhammad.haris@mbzuai.ac.ae>

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4