Vision Language Model-based Test-Time Adaptation

2025. 4. 17

Data Mining & Quality Analytics Lab.

황 순 혁

발표자 소개

❖ 황 순 혁 (Sunhyeok Hwang)

- 고려대학교 일반대학원 산업경영공학과 재학
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- 석 · 박사 통합과정 9학기차 (2022. 03 ~)

❖ Research Interest

- Test-Time Adaptation
- Vision Language Model
- Domain Adaptation/Generalization

❖ Contact

- shhwang1@korea.ac.kr

Contents

1. Introduction

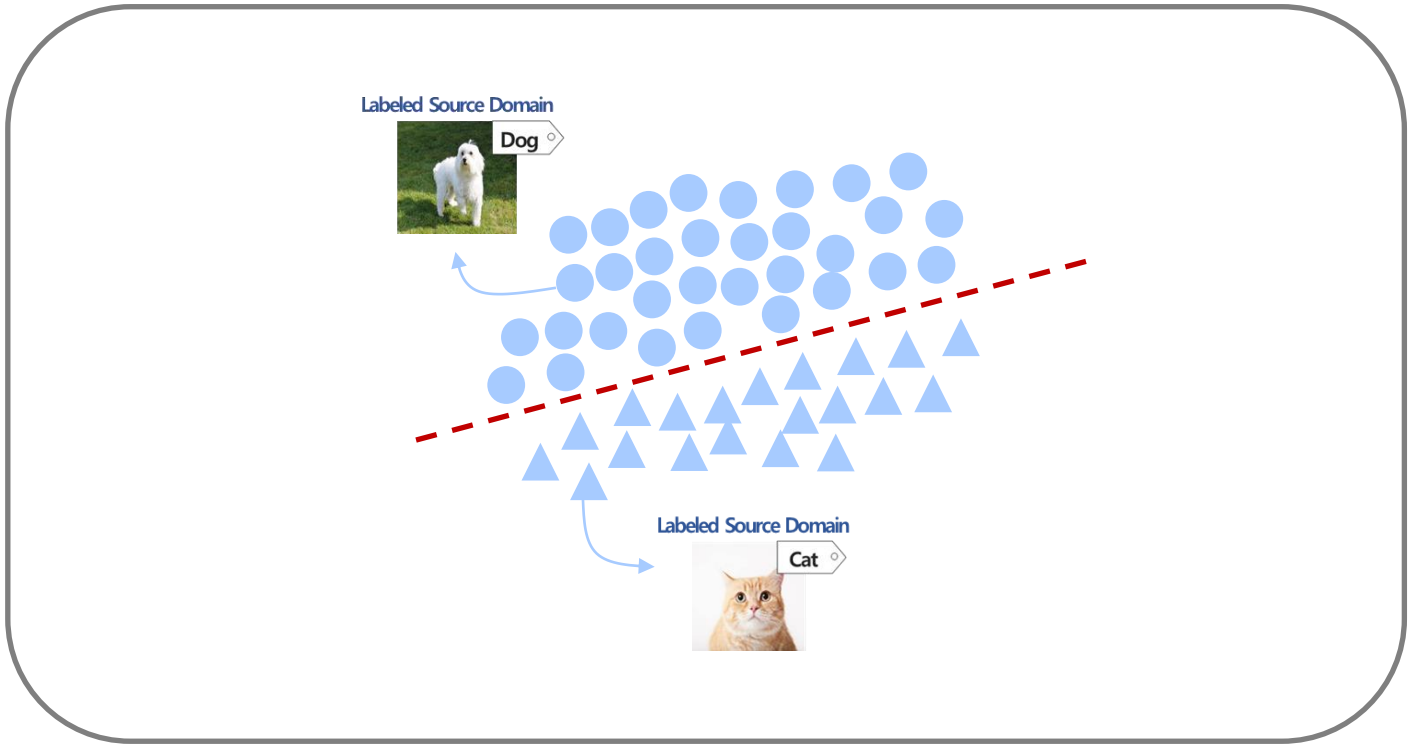
Introduction

Background

도메인 적응 방법론 (Domain Adaptation)

→ **서로 다른 도메인** 데이터셋간 **일반화 성능**을 높이는 방법론

특징 : 유사하지만 분포가 서로 약간 다름



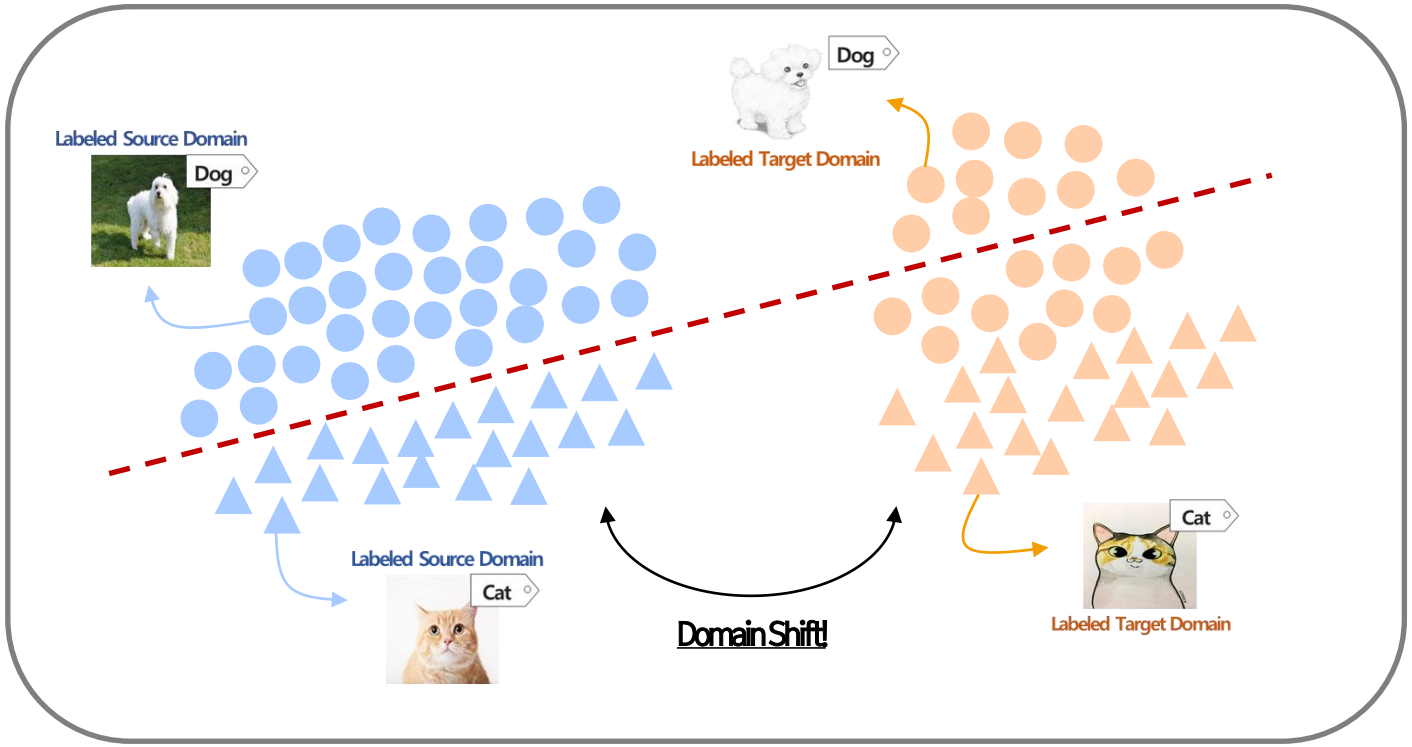
Introduction

Background

도메인 적응 방법론 (Domain Adaptation)

→ **서로 다른 도메인** 데이터셋간 **일반화 성능**을 높이는 방법론

특징 : 유사하지만 분포가 서로 약간 다름



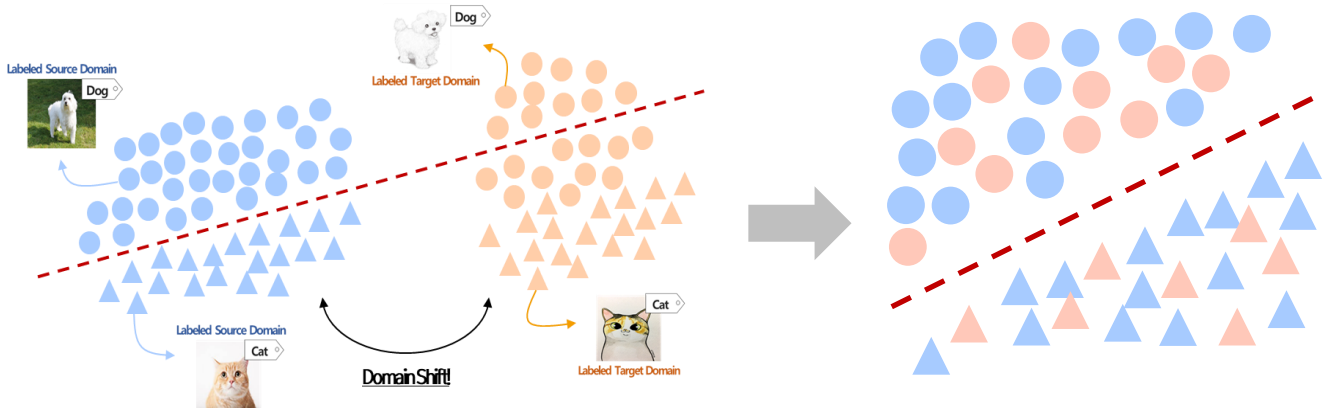
Introduction

Background

도메인 적응 방법론 (Domain Adaptation)

→ **서로 다른 도메인** 데이터셋간 **일반화 성능**을 높이는 방법론

특징 : 유사하지만 분포가 서로 약간 다름



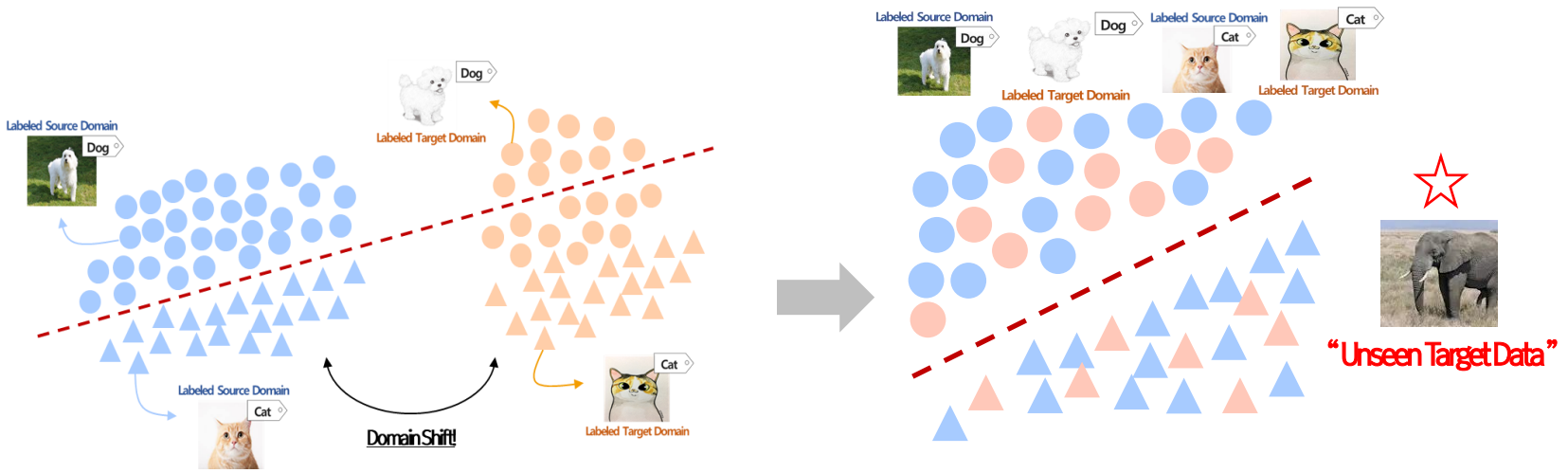
✓ 두 도메인을 잘 Alignment해서 모두 잘 분류되도록 만들자!

Introduction

Background

테스트 시점 적응 방법론 (Test-Time Adaptation)

→ Source Domain의 도움 없이 테스트 시점의 새로운 Target Domain 데이터를 잘 일반화



✓ 테스트 시점에서 학습된 모델에 학습 단계에서 보지 못한 분포의 데이터가 투입된다면?

Introduction

Background

테스트 시점 적응 방법론 (Test-Time Adaptation)

→ Source Domain의 도움 없이 테스트 시점의 새로운 Target Domain 데이터를 잘 일반화

ex. 제품 Recipe 정보(Categorical Variables)

Sample No.	Step 1	Step 2	Step 3	· · ·	Step 100	Y
1	A	K	AA	· · ·	Z	0.59
2	B	L	AB	· · ·	Z	0.66
3	B	K	AA	· · ·	Y	0.58
4	A	K	AC	· · ·	Y	0.61
· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·
998	A	K	AC	· · ·	M	?
999	C	L	AB	· · ·	Z	?
1,000	B	O	AD	· · ·	M	?

Train (Source Domain)

Test (Target Domain)

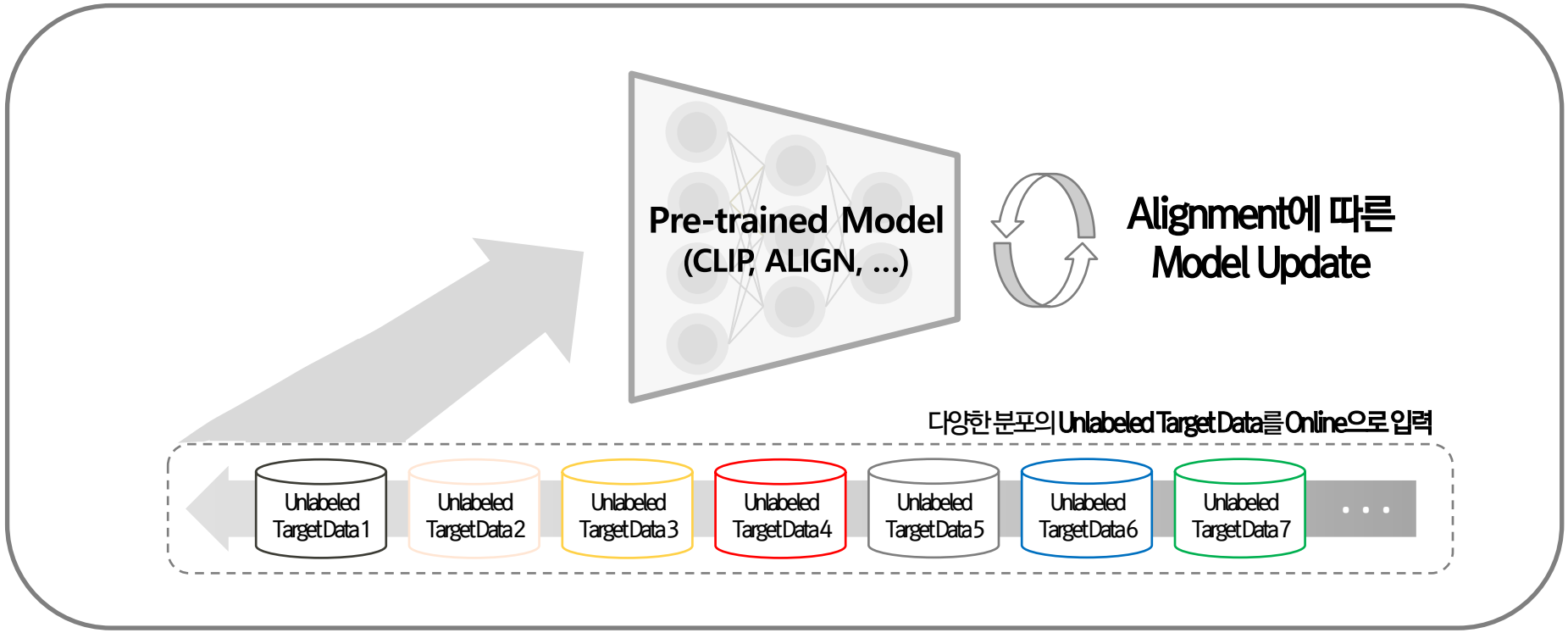
✓ 테스트 시점에서 학습된 모델에 학습 단계에서 보지 못한 분포의 데이터가 투입된다면?

Introduction

Background

테스트 시점 적응 방법론 (Test-Time Adaptation)

→ Source Domain의 도움 없이 테스트 시점의 새로운 Target Domain 데이터를 잘 일반화

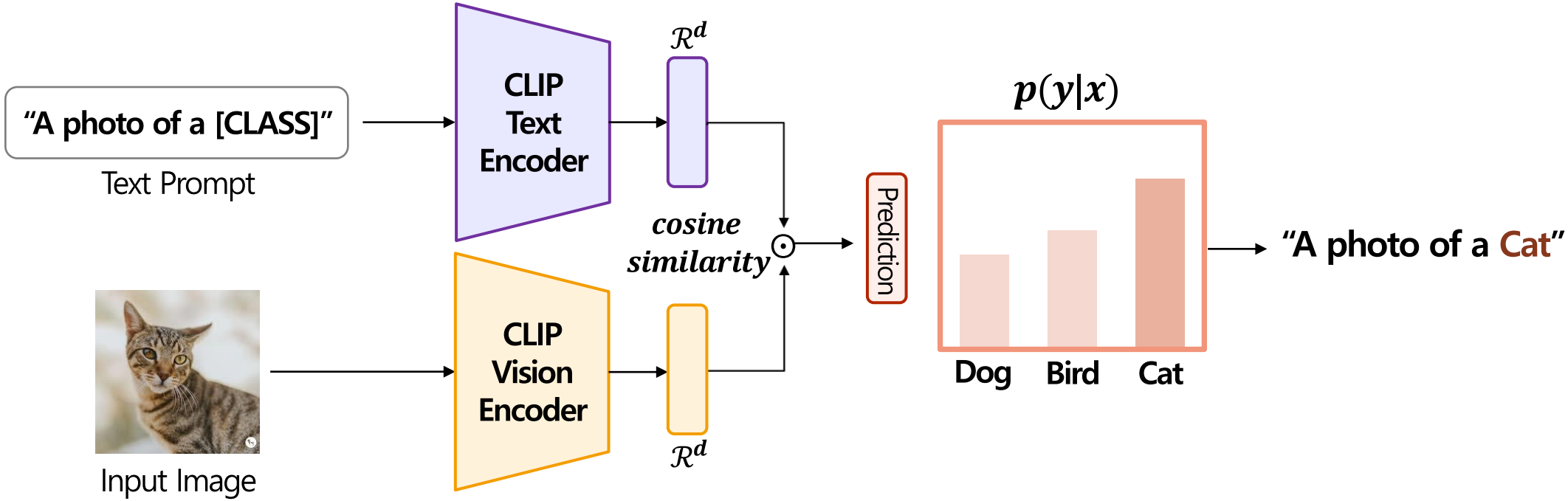


Introduction

Background

비전-언어 모델 (Vision-Language Model, VLM)

→ Text의 도움을 받아 이미지의 의미를 이해하고 분류하는 **멀티모달 학습 모델**



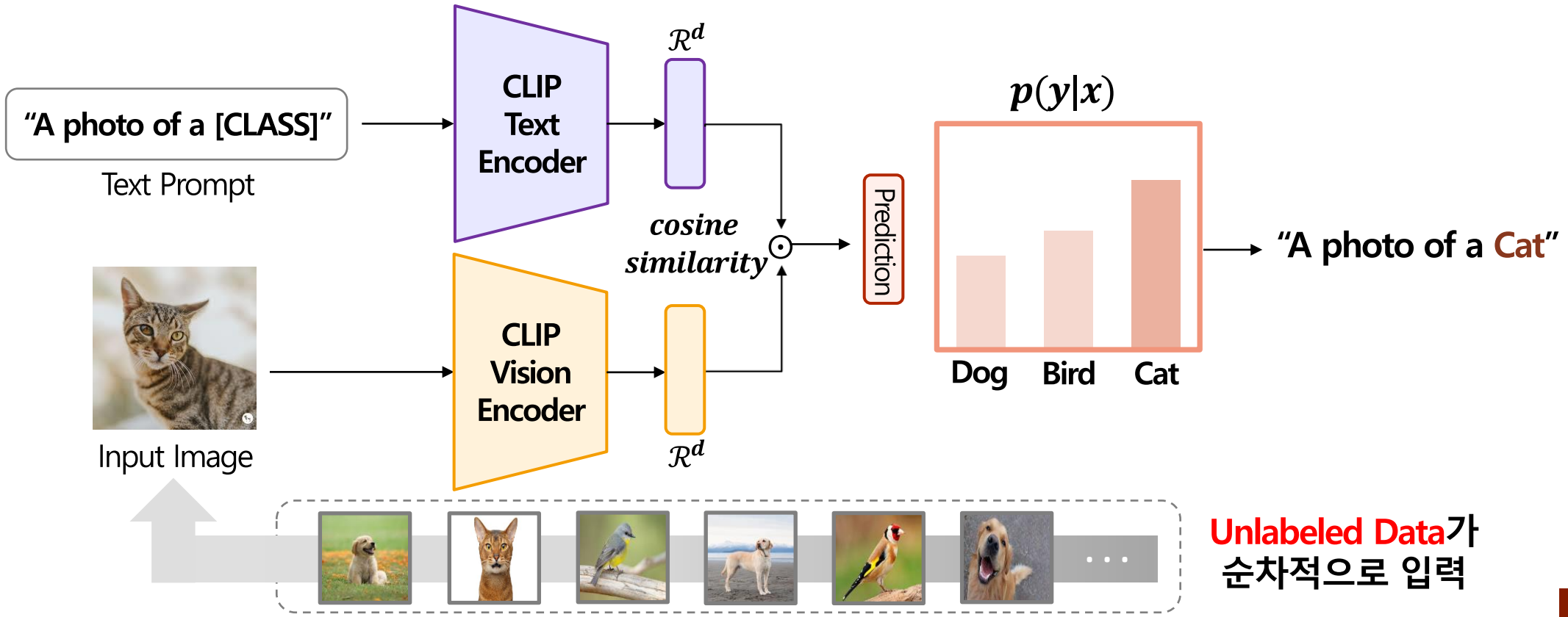
CLIP Algorithm

Introduction

Background

VLM-based Test-time Adaptation

→ Test-time Adaptation 문제를 VLM을 활용하여 해결하는 방법론

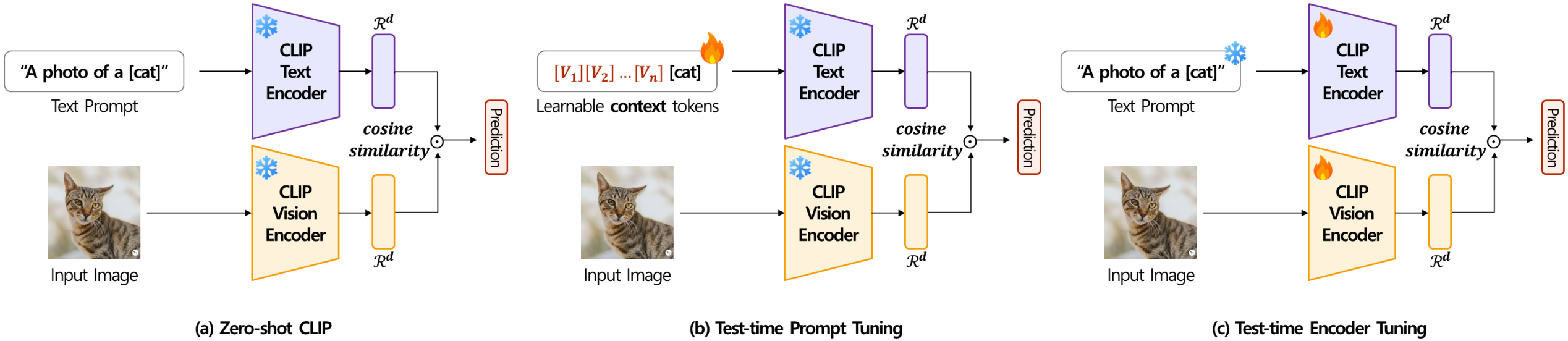


Introduction

Background

VLM-based Test-time Adaptation

→ Test-time Adaptation 문제를 VLM을 활용하여 해결하는 방법론



Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model (2024 ICLR)

Methods

Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

❖ Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model [1]

- 2024년에 제안된 Label propagation 기반 training-free test-time adaptation 방법론
- 유사도 보정과 그래프 기반 propagation을 활용한 고품질 pseudo-label 생성 및 안정적인 adaptation 방법론 제안

Published as a conference paper at ICLR 2025

EFFICIENT AND CONTEXT-AWARE LABEL PROPAGATION FOR ZERO-/FEW-SHOT TRAINING-FREE ADAPTATION OF VISION-LANGUAGE MODEL

Yushu Li[1,2,4*], **Yongyi Su**[1,2*], **Adam Goodge**[2], **Kui Jia**[3], **Xun Xu**[2†]
[1] South China University of Technology
[2] Institute for Infocomm Research (I²R), A*STAR
[3] School of Data Science, The Chinese University of Hong Kong, Shenzhen
[4] Shanghai AI Laboratory
eeyushuli@mail.scut.edu.cn, eesuyongyi@mail.scut.edu.cn,
goodge_adam_david@i2r.a-star.edu.sg, kuijia@cuhk.edu.cn,
xu_xun@i2r.a-star.edu.sg

ABSTRACT

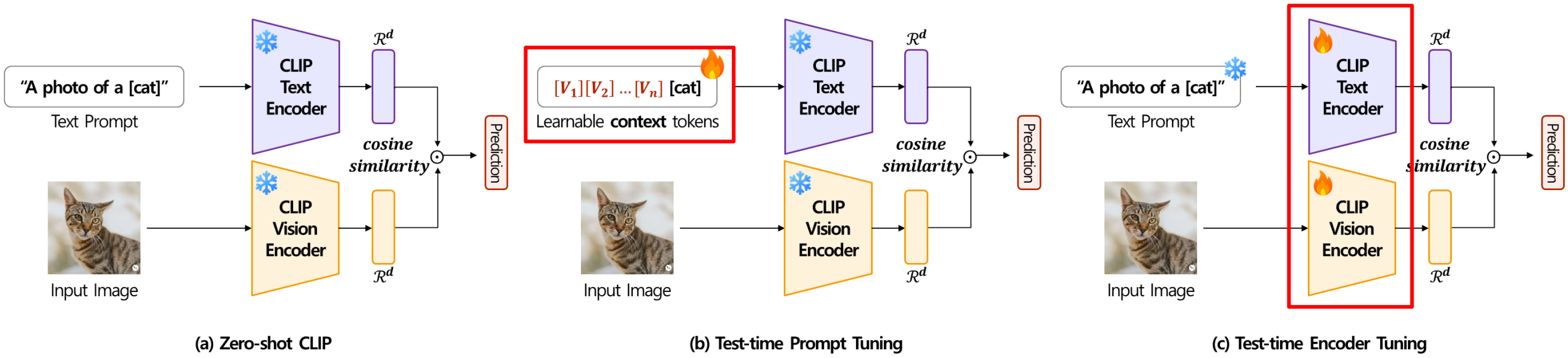
Vision-language models (VLMs) have revolutionized machine learning by leveraging large pre-trained models to tackle various downstream tasks. Although label, training, and data efficiency have improved, many state-of-the-art VLMs still

Methods

Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

❖ Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

- 기존 VLM 기반 TTA 방법은 prompt/encoding의 학습이 필요하여 높은 계산 비용과 자원 소모 발생
- Test data의 구조를 충분히 활용하지 못하고 cosine similarity에 의존해 pseudo-label 품질이 제한

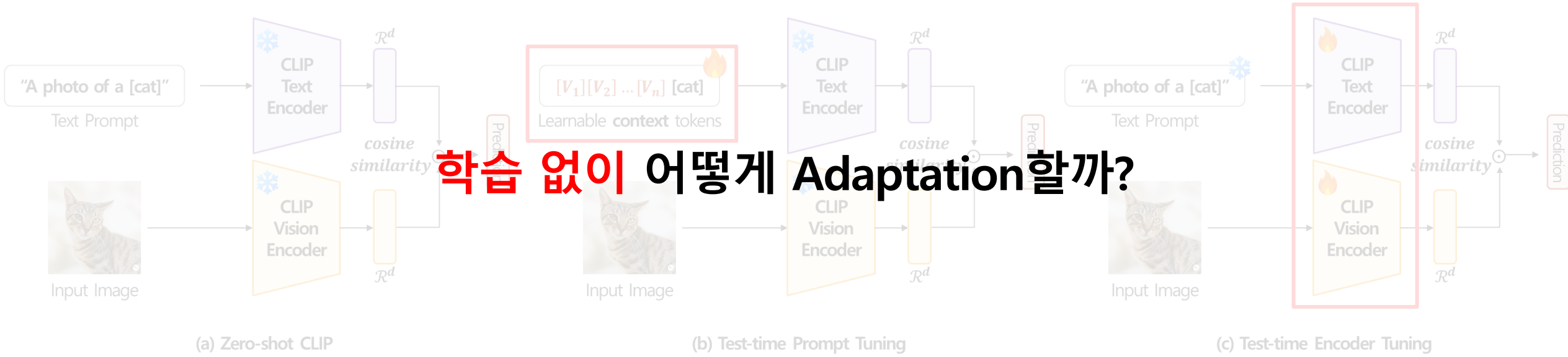


Methods

Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

❖ Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

- 기존 VLM 기반 TTA 방법은 prompt/encoding의 학습이 필요하여 높은 계산 비용과 자원 소모 발생
- Test data의 구조를 충분히 활용하지 못하고 cosine similarity에 의존해 pseudo-label 품질이 제한



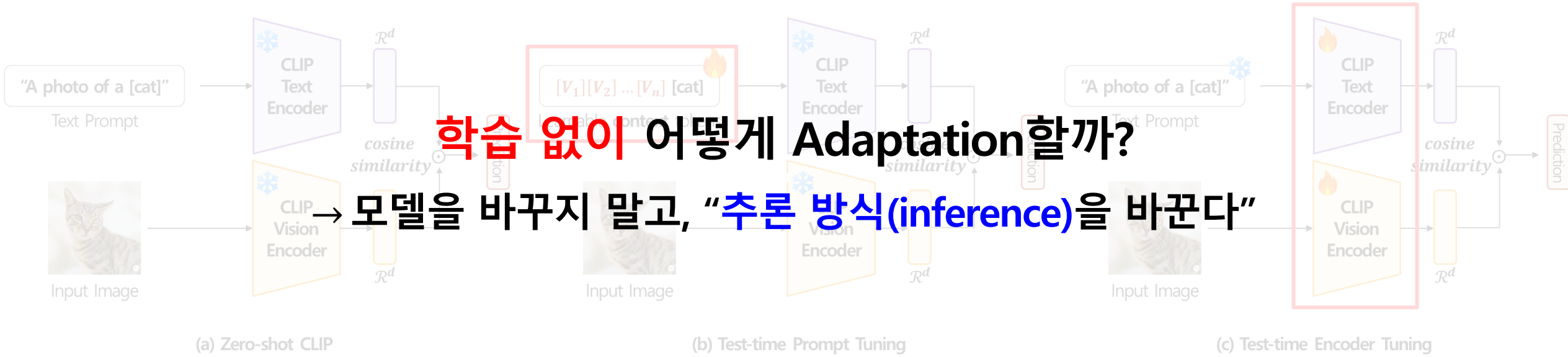
학습 없이 어떻게 Adaptation할까?

Methods

Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

❖ Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

- 기존 VLM 기반 TTA 방법은 prompt/encoding의 학습이 필요하여 높은 계산 비용과 자원 소모 발생
- Test data의 구조를 충분히 활용하지 못하고 cosine similarity에 의존해 pseudo-label 품질이 제한



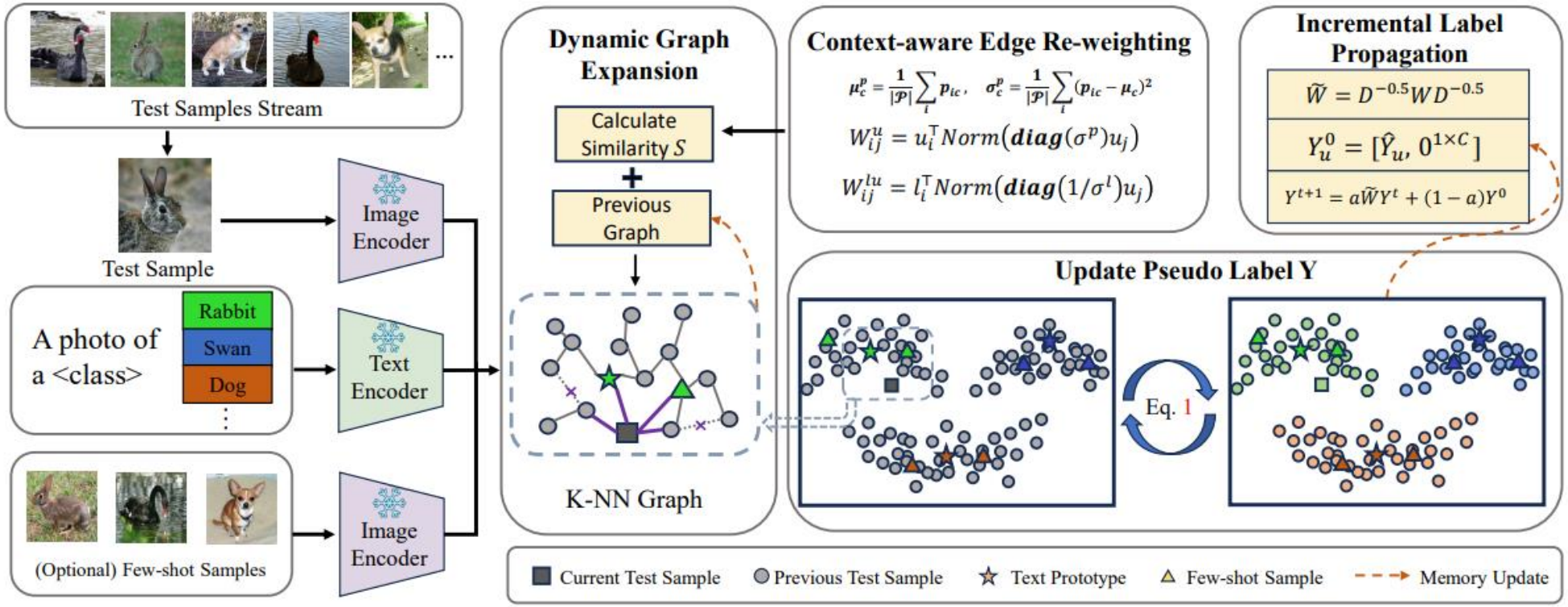
학습 없이 어떻게 Adaptation할까?

→ 모델을 바꾸지 말고, "추론 방식(inference)을 바꾼다"

Methods

Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

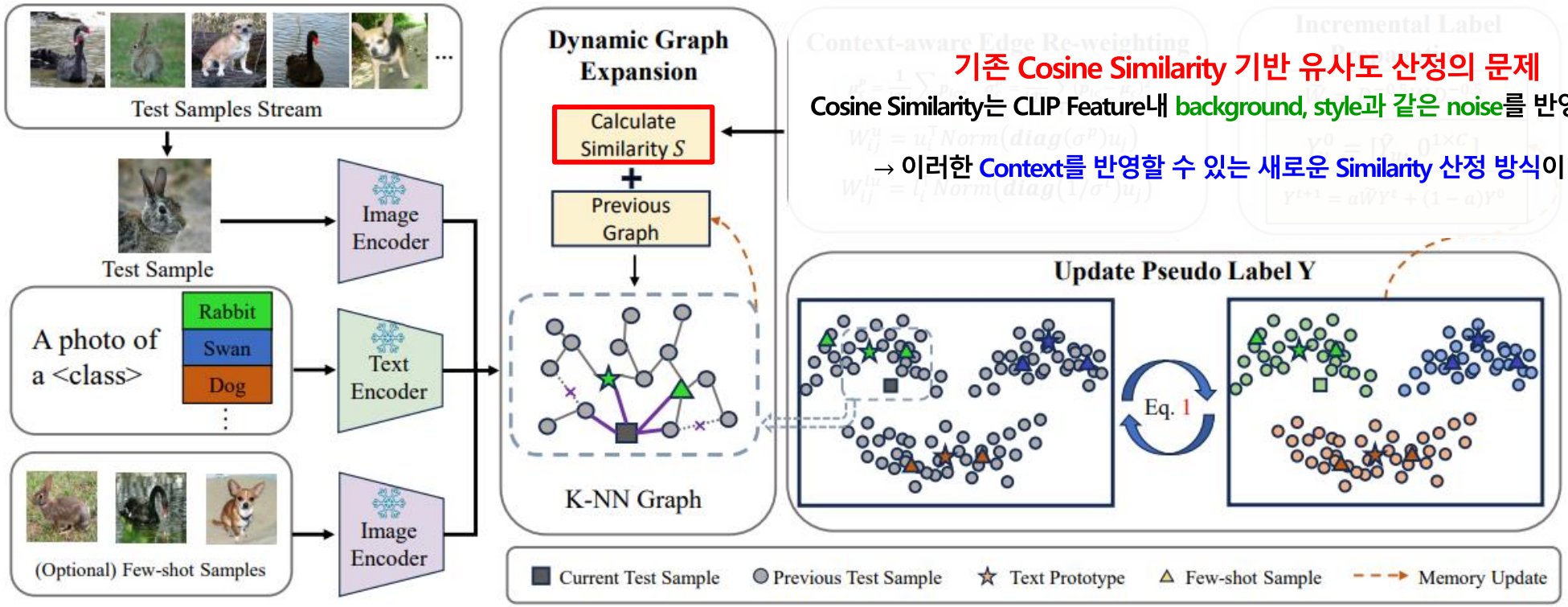
❖ Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model



Methods

Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

❖ Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model



기존 Cosine Similarity 기반 유사도 산정의 문제

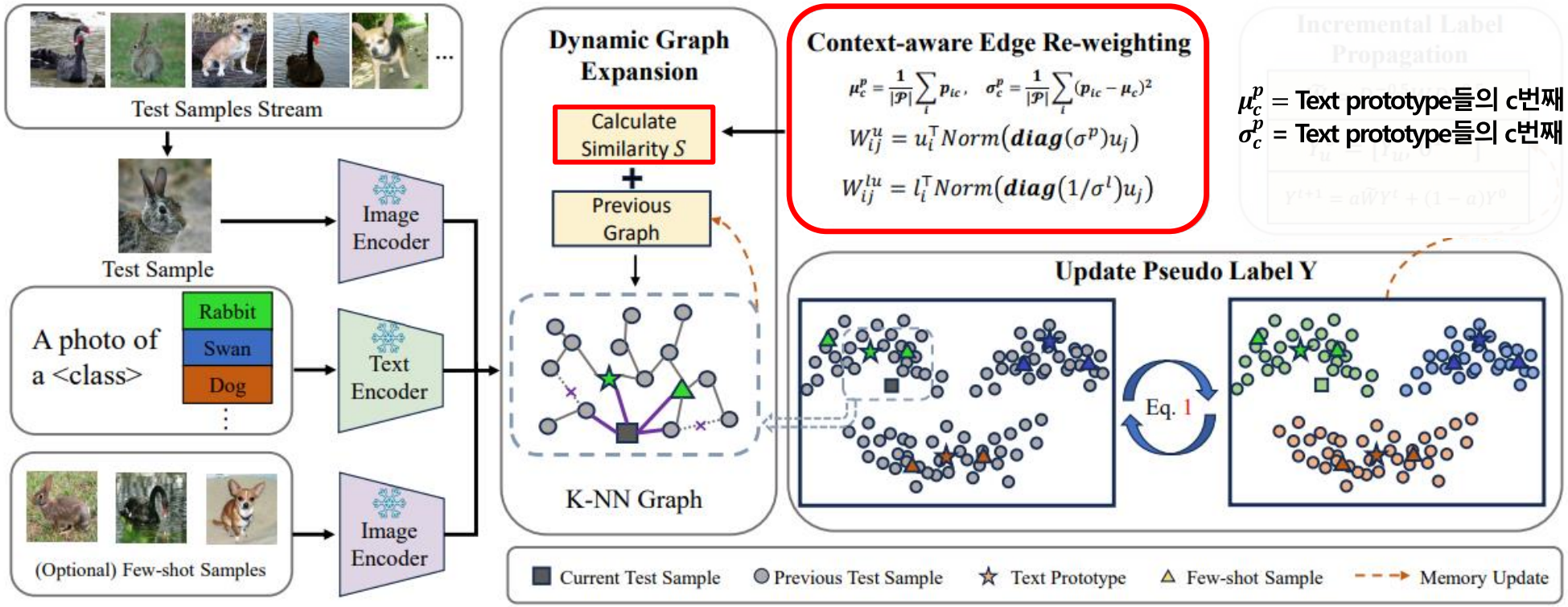
Cosine Similarity는 CLIP Feature내 background, style과 같은 noise를 반영하지 못함

→ 이러한 Context를 반영할 수 있는 새로운 Similarity 산정 방식이 필요

Methods

Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

❖ Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

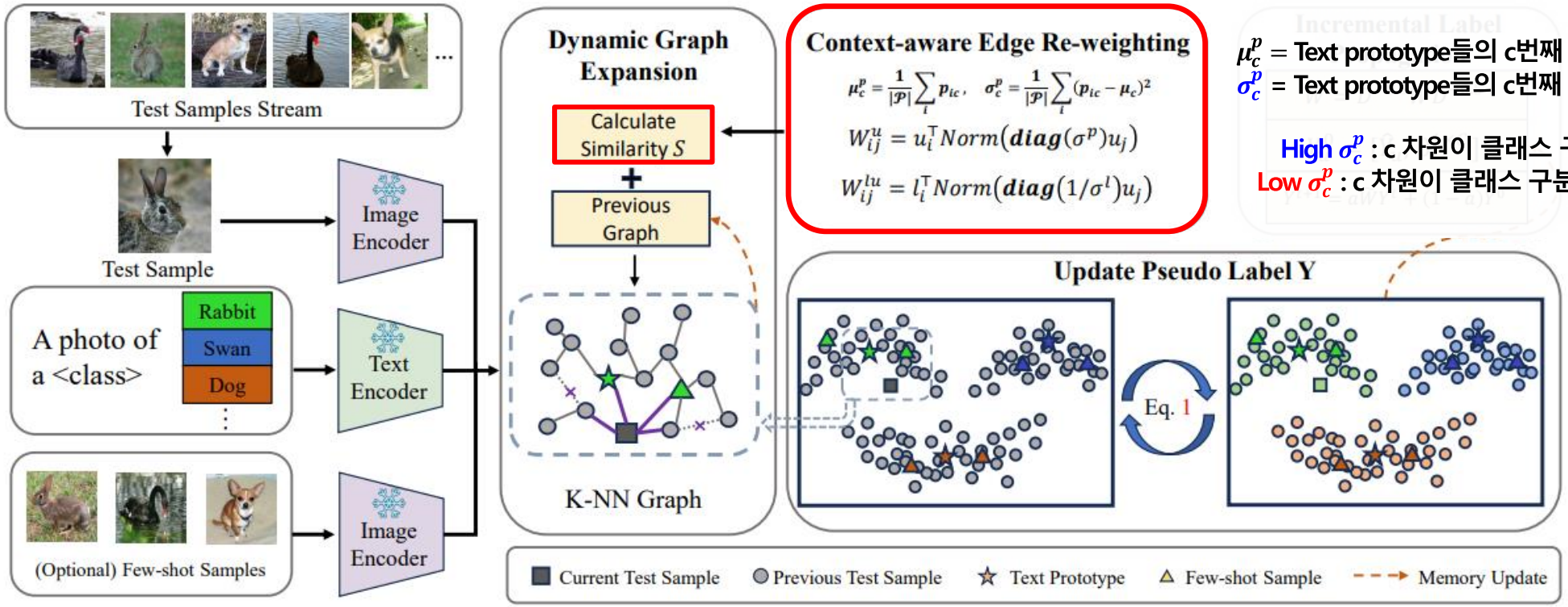


μ_c^p = Text prototype들의 c번째 feature 차원의 평균
σ_c^p = Text prototype들의 c번째 feature 차원의 분산

Methods

Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

❖ Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model



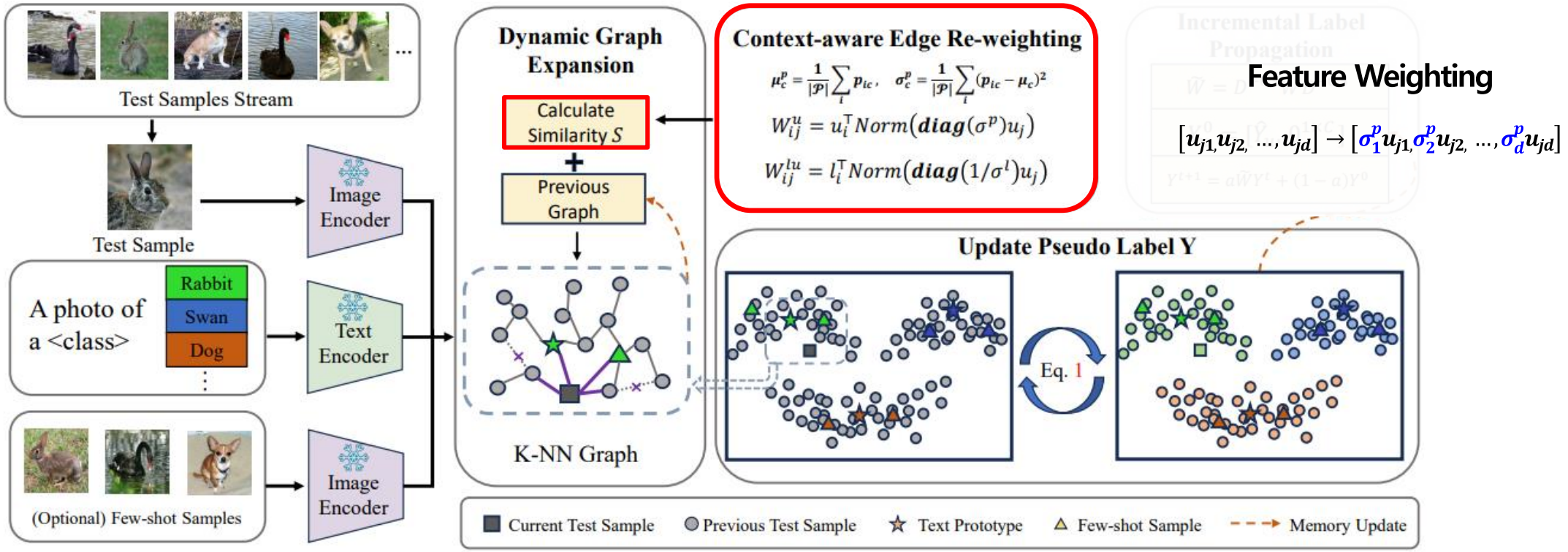
μ_c^p = Text prototype들의 c번째 feature 차원의 평균
σ_c^p = Text prototype들의 c번째 feature 차원의 분산

High σ_c^p : c 차원이 클래스 구분에 중요하다
Low σ_c^p : c 차원이 클래스 구분에 중요하지 않다

Methods

Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

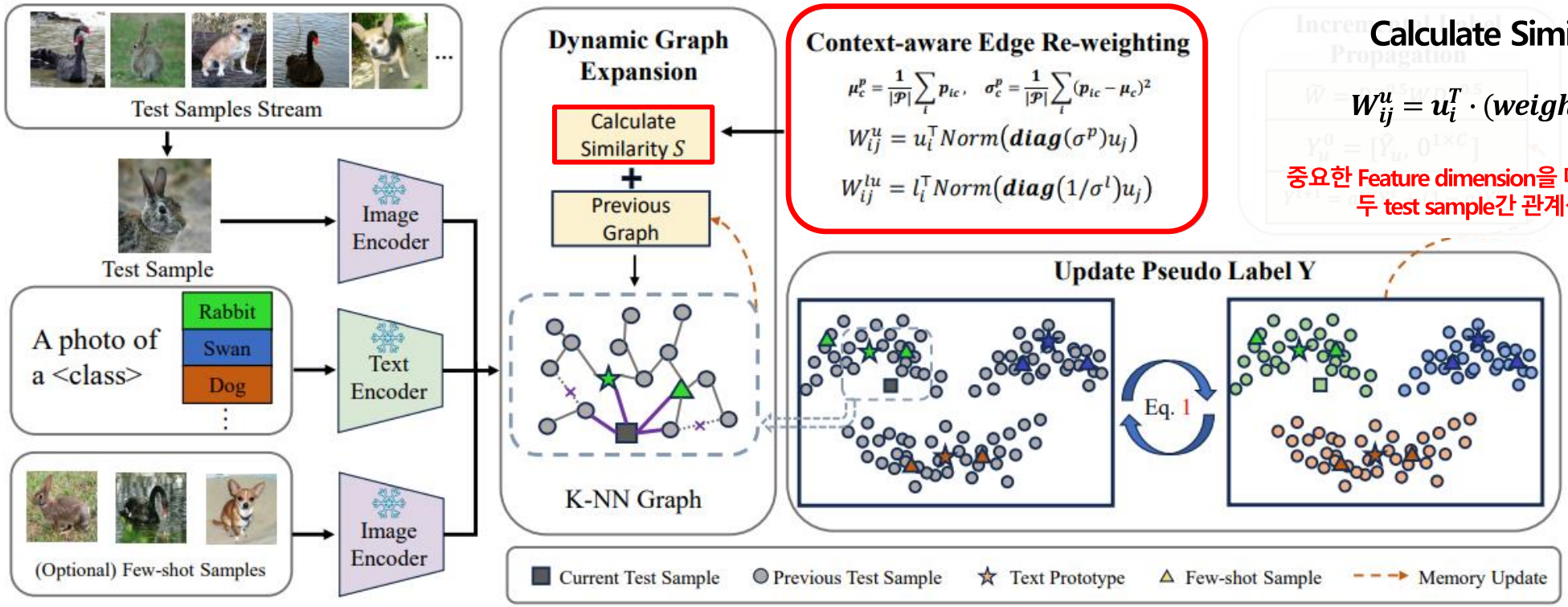
❖ Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model



Methods

Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

❖ Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model



Calculate Similarity

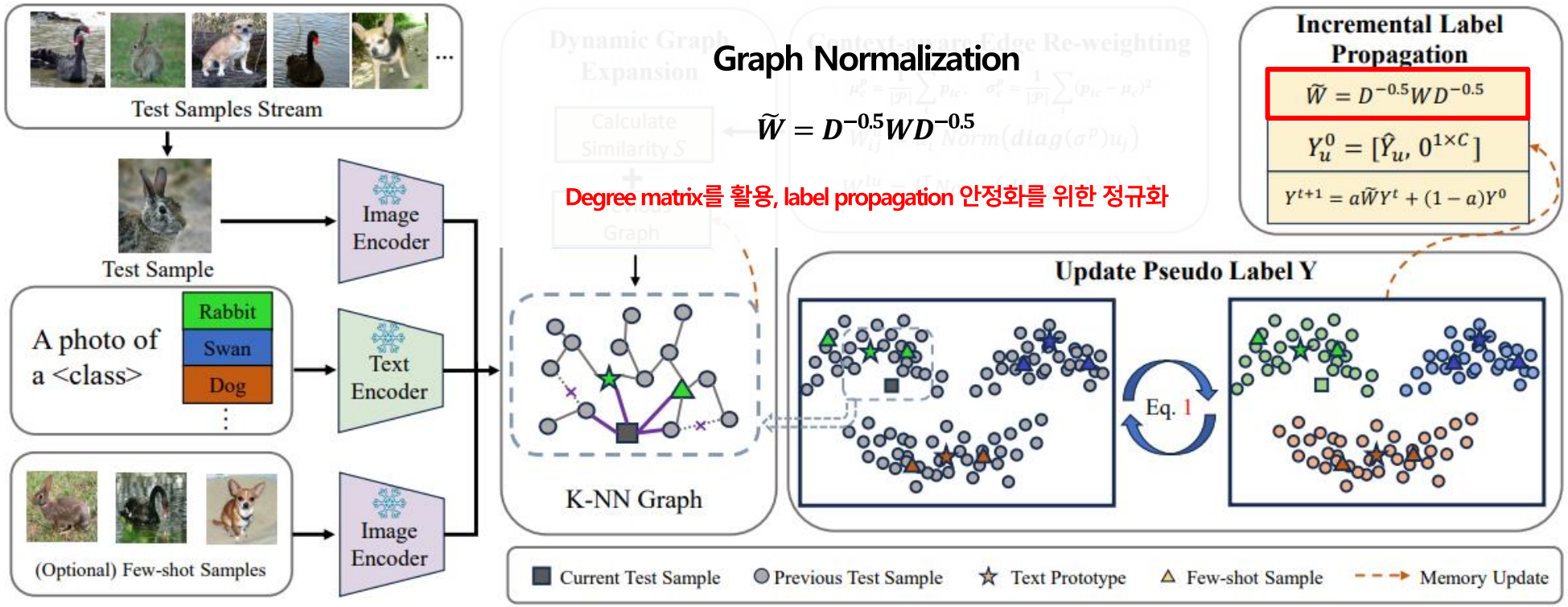
$$W_{ij}^u = u_i^T \cdot (\boldsymbol{weighted}\, u_j)$$

중요한 Feature dimension을 더 크게 반영해서
두 test sample간 관계성을 계산

Methods

Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

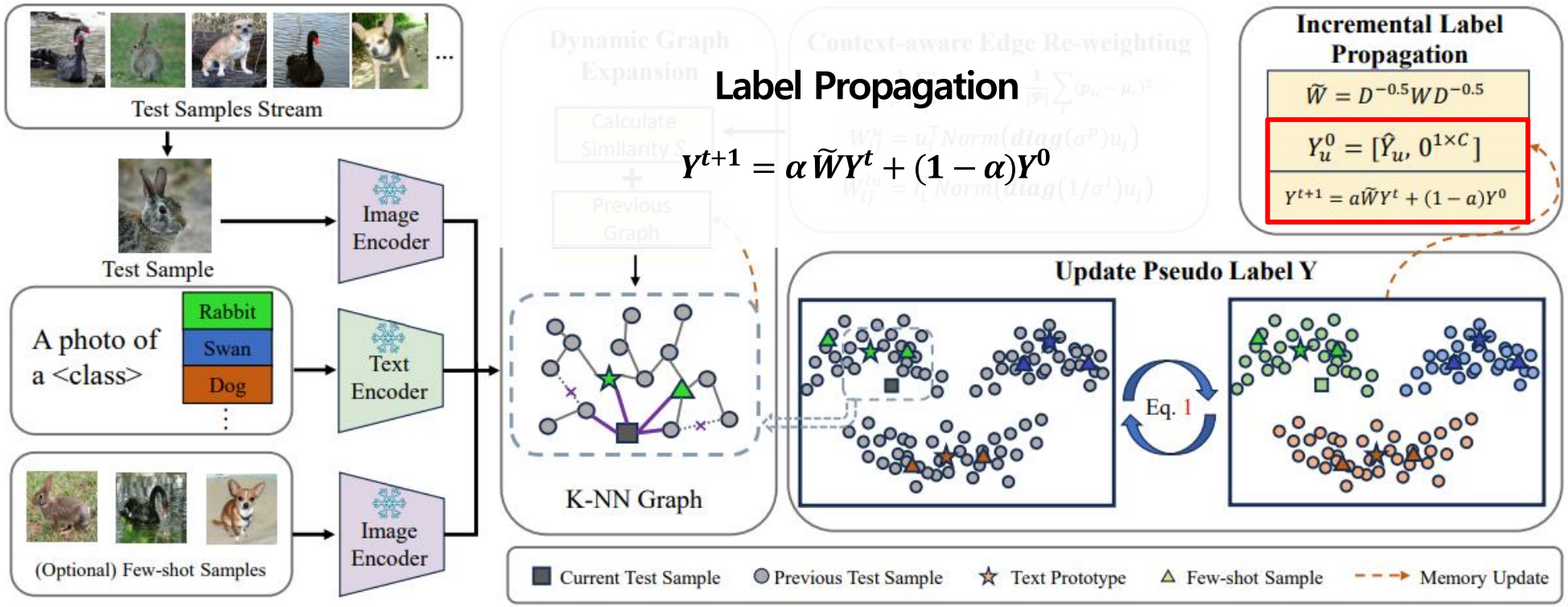
❖ Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model



Methods

Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

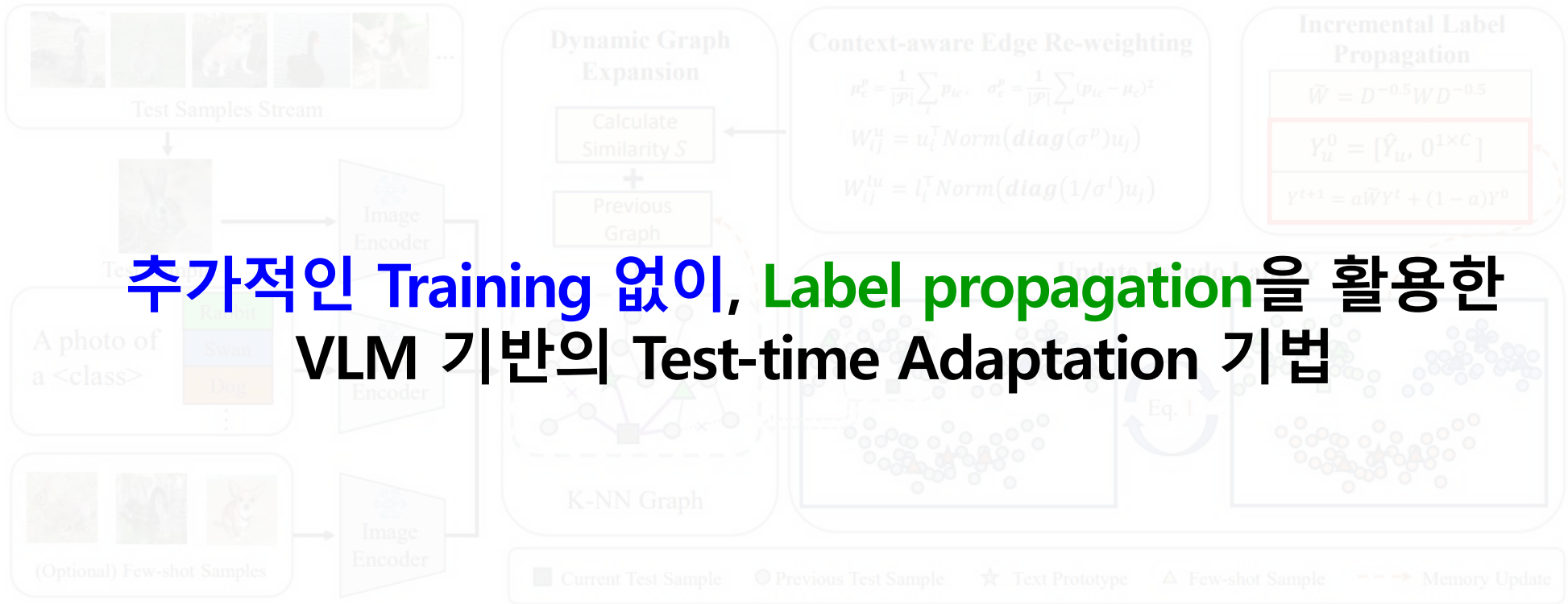
❖ Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model



Methods

Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

❖ Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model



추가적인 Training 없이, Label propagation을 활용한 VLM 기반의 Test-time Adaptation 기법

Experiments

Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

❖ Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

Method	ImageNet	Flower	DTD	Pets	Cars	UCF	Caltech	Food	SUN	Aircraft	EuroSAT	Mean
CLIP-RN50 (Radford et al., 2021)	58.16	61.75	40.37	83.57	55.70	58.84	85.88	73.97	58.80	15.66	23.69	56.04
DN (Zhou et al., 2023)	60.16	63.32	41.21	81.92	56.55	55.60	87.25	74.64	59.11	17.43	28.31	56.86
TPT (Shu et al., 2022)	60.74	62.69	40.84	84.49	58.46	60.82	87.02	74.88	61.46	17.58	28.33	57.94
DiffTPT (Feng et al., 2023)	60.80	63.53	40.72	83.40	**60.71**	62.67	86.89	**79.21**	62.72	17.60	41.04	59.94
VisDesc (Menon & Vondrick, 2023)	59.68	65.37	41.96	82.39	54.76	58.47	88.11	76.80	59.84	16.26	37.60	58.29
Ensemble (Zhang et al., 2022)	60.32	66.10	40.07	85.83	55.71	61.33	83.94	77.32	58.53	17.10	37.54	58.53
CALIP (Guo et al., 2023)	60.57	66.38	42.39	86.21	56.27	61.72	87.71	77.42	58.59	17.76	38.90	59.45
CuPL (Pratt et al., 2023)	61.45	65.44	48.64	84.84	57.28	58.97	89.29	76.94	62.55	19.59	38.38	60.31
SuS-X (Udandarao et al., 2023)	61.84	67.72	50.59	85.34	57.27	61.54	89.53	77.58	62.95	19.47	45.57	61.76
DMN (Zhang et al., 2024b)	62.02	68.33	50.53	86.29	58.36	64.02	89.09	74.69	63.70	20.22	44.94	62.02
DMN* (Zhang et al., 2024b)	63.87	67.93	50.41	86.78	60.02	65.34	90.14	76.70	64.39	22.77	48.72	63.37
TDA (Karmanov et al., 2024)	61.35	68.74	43.74	86.18	57.78	64.18	89.70	77.75	62.53	17.61	42.11	61.06
ZLaP† (Kalantidis et al., 2024)	62.20	69.27	42.79	80.32	56.42	62.81	86.90	77.87	61.83	17.37	31.85	59.06
ECALP (Ours)	**62.64**	**69.39**	**54.49**	**88.20**	60.56	**66.67**	**89.94**	76.97	**64.97**	**21.12**	**49.09**	**64.00**
CLIP-ViTB/16 (Radford et al., 2021)	66.73	64.44	44.27	88.25	65.48	65.13	93.35	83.65	62.59	23.67	42.01	63.87
Ensemble (Zhang et al., 2022)	68.34	66.99	45.04	86.92	66.11	65.16	93.55	82.86	65.63	23.22	50.42	64.93
TPT (Shu et al., 2022)	68.98	68.98	47.75	87.79	66.87	68.04	94.16	84.67	65.50	24.78	42.44	65.45
DiffTPT (Feng et al., 2023)	70.30	70.10	47.00	88.20	67.01	68.22	92.49	**87.23**	65.74	25.60	43.13	65.90
DMN (Zhang et al., 2024b)	70.51	75.32	54.85	91.22	67.01	71.95	93.63	84.05	69.14	28.29	56.22	69.29
DMN*(Zhang et al., 2024b)	72.25	74.49	55.85	92.04	67.96	72.51	95.38	85.08	70.18	30.03	59.43	70.47
TDA (Karmanov et al., 2024)	69.51	71.42	47.40	88.63	67.28	70.66	94.24	86.14	67.62	23.91	**58.00**	67.71
ZLaP† (Kalantidis et al., 2024)	70.17	73.49	48.58	87.14	65.63	71.45	93.06	86.92	67.44	25.44	55.62	67.72
ECALP (Ours)	**71.26**	**75.96**	**56.32**	**92.31**	**68.20**	**75.44**	**94.40**	85.72	**70.35**	**29.49**	56.53	**70.54**

Fine-grained Categorization 실험 성능

Experiments

Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

❖ Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

Method	ImageNet-A	ImageNet-V2	ImageNet-R	ImageNet-S	Mean
CLIP-RN50 (Radford et al., 2021)	21.83	51.41	56.15	33.37	40.69
CoOp† (Zhou et al., 2022b)	23.06	55.40	56.60	34.67	42.43
CoCoOp† (Zhou et al., 2022a)	23.32	55.72	57.74	34.48	42.82
TPT (Shu et al., 2022)	26.67	54.70	59.11	35.09	43.89
DiffTPT (Feng et al., 2023)	**31.06**	55.80	58.80	37.10	45.69
CALIP (Guo et al., 2023)	23.96	53.70	60.81	35.61	43.52
TDA (Karmanov et al., 2024)	30.29	55.54	62.58	38.12	46.63
DMN (Zhang et al., 2024b)	28.57	56.12	61.44	39.84	46.49
ECALP (Ours)	28.80	**56.92**	**63.68**	**41.51**	**47.73**
CLIP-ViT-B/16 (Radford et al., 2021)	47.87	60.86	73.98	46.09	57.20
CoOp† (Zhou et al., 2022b)	49.71	64.20	75.21	47.99	59.14
CoCoOp† (Zhou et al., 2022a)	50.63	64.07	76.18	48.75	59.91
MaPLe† (Khattak et al., 2023)	50.90	64.07	76.98	49.15	60.28
TPT (Shu et al., 2022)	54.77	63.45	77.06	47.94	60.81
DiffTPT (Feng et al., 2023)	55.68	65.10	75.00	46.80	60.65
TDA (Karmanov et al., 2024)	**60.11**	64.67	80.24	50.54	63.89
DMN (Zhang et al., 2024b)	58.28	65.17	78.55	53.20	63.80
ECALP (Ours)	58.52	**65.72**	**80.77**	**54.66**	**64.92**

Style-transfer 실험 성능

Experiments

Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

❖ Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

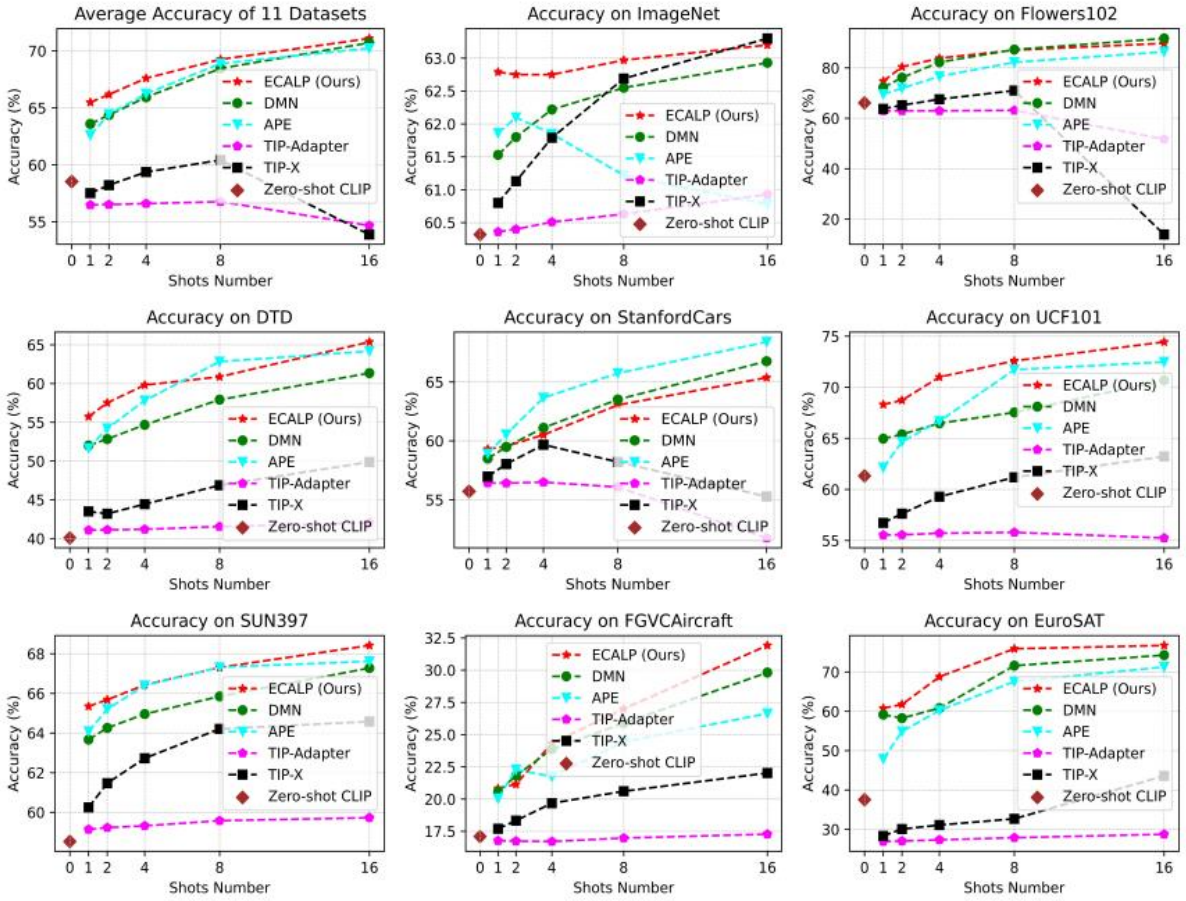
Method	Gauss.	Shot	Impu.	Defo.	Glas.	Moti.	Zoom	Snow	Fros.	Fog	Brig.	Cont.	Elas.	Pix.	JPEG	Average
CLIP-RN50 (Radford et al., 2021)	1.63	2.18	1.64	10.06	3.42	7.85	12.83	12.58	15.67	21.95	40.27	6.28	4.75	11.12	13.03	11.02
TDA (Karmanov et al., 2024)	2.26	3.10	2.31	11.30	5.12	9.26	15.43	15.47	19.11	26.45	45.30	8.34	7.30	13.01	15.83	13.31
ZLaP† (Kalantidis et al., 2024)	1.77	2.33	1.65	10.30	3.54	7.99	13.47	13.66	17.15	23.20	44.67	6.55	5.15	11.61	14.23	11.82
DMN (Zhang et al., 2024b)	2.14	2.78	2.30	10.91	4.48	8.59	14.31	14.14	17.92	24.16	44.57	7.88	6.11	12.40	14.93	12.51
ECALP (Ours)	**2.71**	**3.30**	**2.82**	**12.29**	**5.49**	**10.56**	**16.82**	**16.66**	**20.60**	**27.83**	**47.02**	**9.08**	**7.72**	**14.46**	**16.88**	**14.28**
CLIP-ViTB/16 (Radford et al., 2021)	11.34	12.31	11.85	23.78	15.12	24.05	22.72	32.70	30.43	36.69	54.57	16.84	12.77	31.22	33.00	24.63
TDA (Karmanov et al., 2024)	15.42	16.46	16.03	26.53	17.91	27.35	25.90	36.50	34.84	40.53	58.57	20.16	16.62	35.65	36.69	28.34
ZLaP† (Kalantidis et al., 2024)	12.83	14.03	13.27	24.88	16.13	25.77	24.36	34.43	32.63	38.56	58.42	17.53	14.21	33.72	35.52	26.42
DMN (Zhang et al., 2024b)	14.33	15.33	14.69	26.06	17.19	26.61	25.23	34.81	33.48	38.93	58.70	19.38	15.40	35.32	36.49	27.46
ECALP (Ours)	**15.92**	**16.84**	**16.32**	**27.85**	**18.78**	**28.59**	**27.62**	**37.82**	**36.01**	**41.65**	**60.57**	**21.26**	**17.77**	**37.39**	**38.11**	**29.50**

Out-of-Distribution 실험 성능

Experiments

Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

❖ Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model



Zero/Few-shot 개수에 따른 실험 성능

R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning (CVPR 2025)

Methods

R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning

❖ R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning [2]

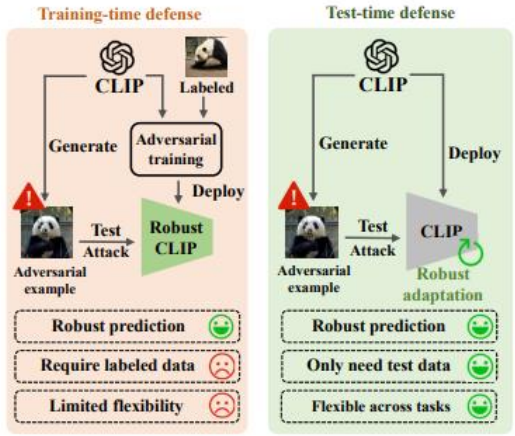
- 2025년에 제안된 adversarial-aware test-time prompt tuning 기반 방법론
- pointwise entropy 최적화와 reliability-weighted ensembling을 활용한 강건한 예측 및 안정적인 adaptation 방법론 제안

R-TPT: Improving Adversarial Robustness of Vision-Language Models through Test-Time Prompt Tuning

Lijun Sheng[1,2], Jian Liang[2,3*], Zilei Wang[1], Ran He[2,3]
[1] University of Science and Technology of China
[2] NLPR & MAIS, Institute of Automation, Chinese Academy of Sciences
[3] University of Chinese Academy of Sciences
slj0728@mail.ustc.edu.cn, liangjian92@gmail.com

Abstract

Vision-language models (VLMs), such as CLIP, have gained significant popularity as foundation models, with numerous fine-tuning methods developed to enhance performance on downstream tasks. However, due to their inherent vulnerability and the common practice of selecting from a limited set of open-source models, VLMs suffer from a higher risk of adversarial attacks than traditional vision models. Existing defense techniques typically rely on adversarial fine-tuning during training, which requires labeled data and lacks of flexibility for downstream tasks. To address these limitations, we propose robust test-time prompt tuning (R-TPT), which mitigates the impact of adversarial attacks during the inference stage. We first reformulate the classic marginal entropy objective by eliminating the term that introduces conflicts under adversarial conditions,

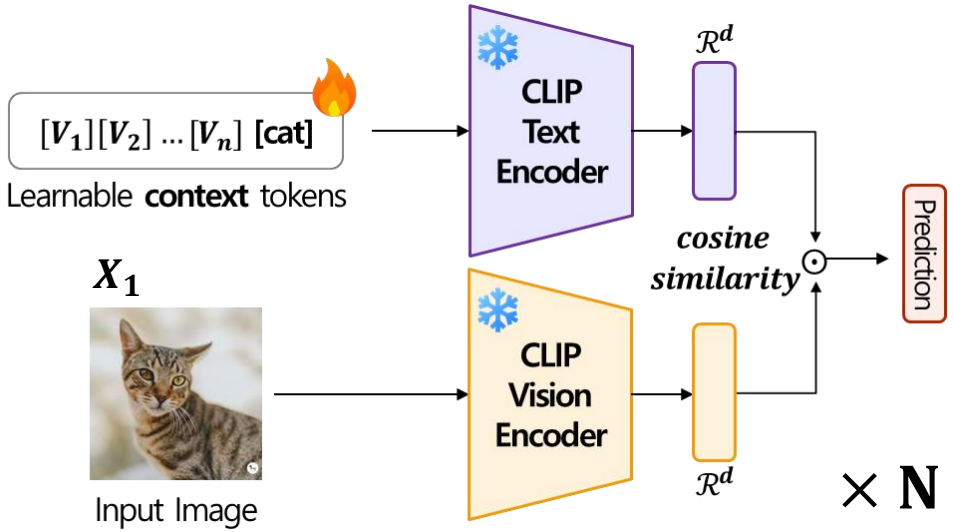


Methods

R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning

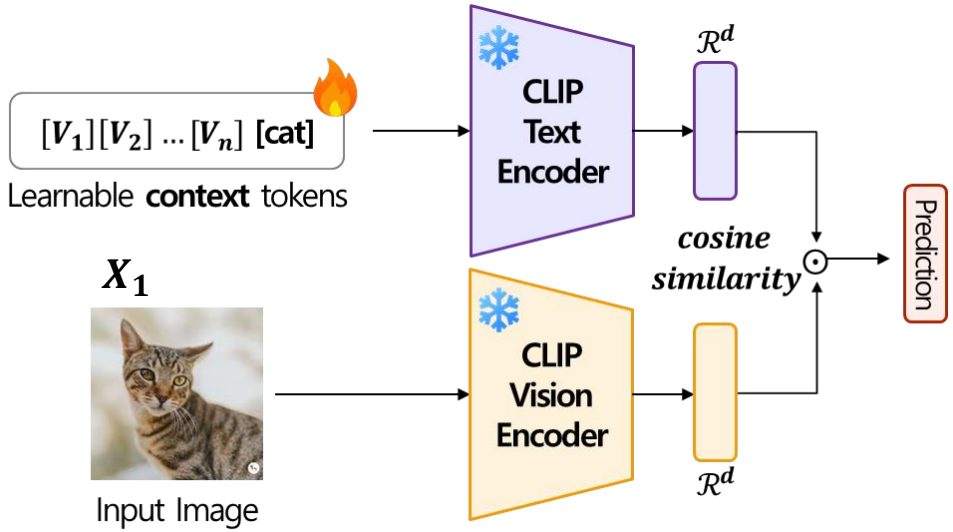
❖ VLM-based Test-time Prompt Tuning

- VLM 기반의 TPT(Test-time Prompt Tuning)은 두 가지 형식을 가짐
  1) 매 샘플마다 Learnable context parameters를 초기화하는 방식 (Episodic TPT)
  2) 매 샘플의 정보를 Learnable context parameters에 누적하여 반영하는 방식 (Online TPT)



X_1 샘플로 특정 **N Step간 Prompt optimization** 후 Prediction

(a) Episodic TPT



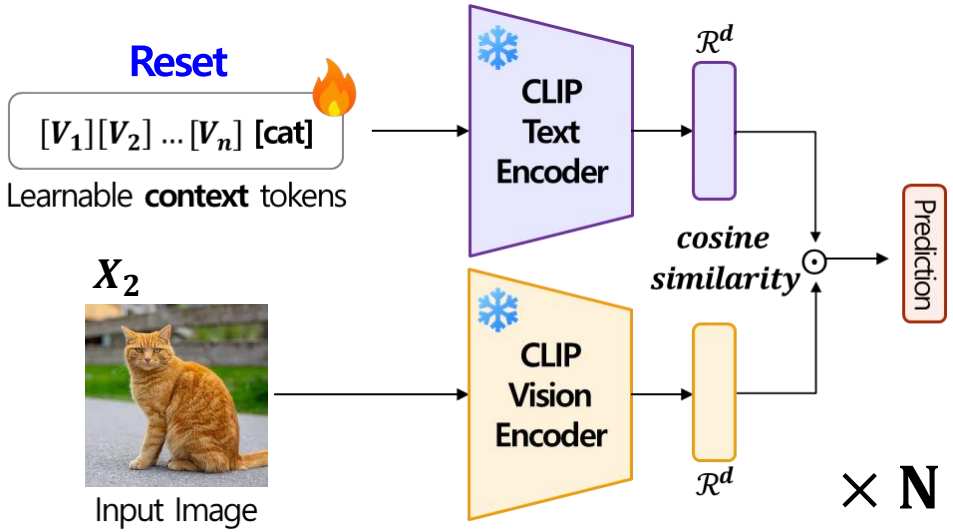
(b) Online TPT

Methods

R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning

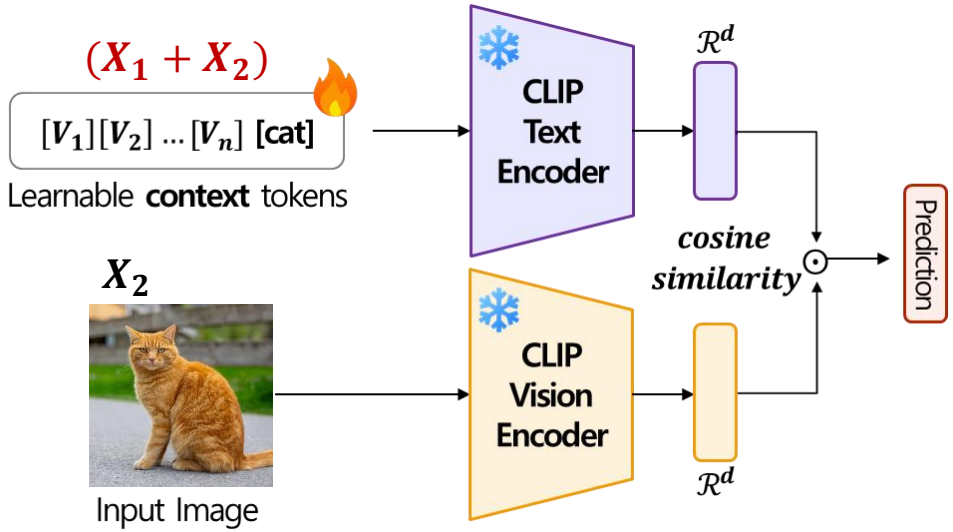
❖ VLM-based Test-time Prompt Tuning

- VLM 기반의 TPT(Test-time Prompt Tuning)은 두 가지 형식을 가짐
  1) 매 샘플마다 Learnable context parameters를 초기화하는 방식 (Episodic TPT)
  2) 매 샘플의 정보를 Learnable context parameters에 누적하여 반영하는 방식 (Online TPT)



X_1 샘플로 특정 **N Step간 Prompt optimization** 후 Prediction

(a) Episodic TPT



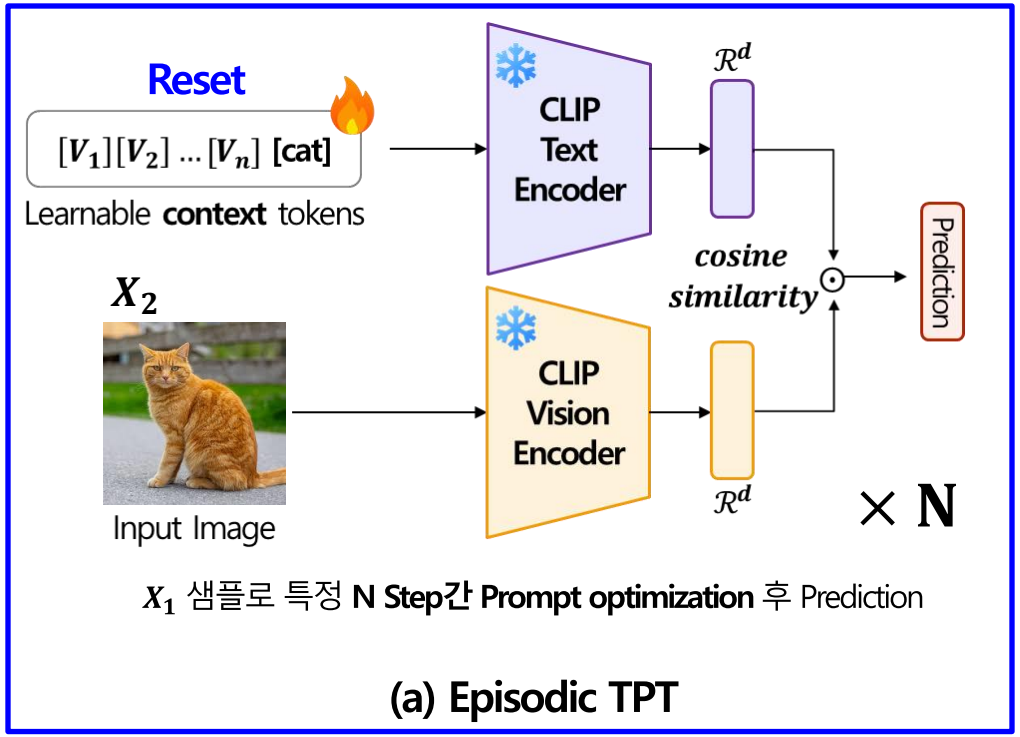
(b) Online TPT

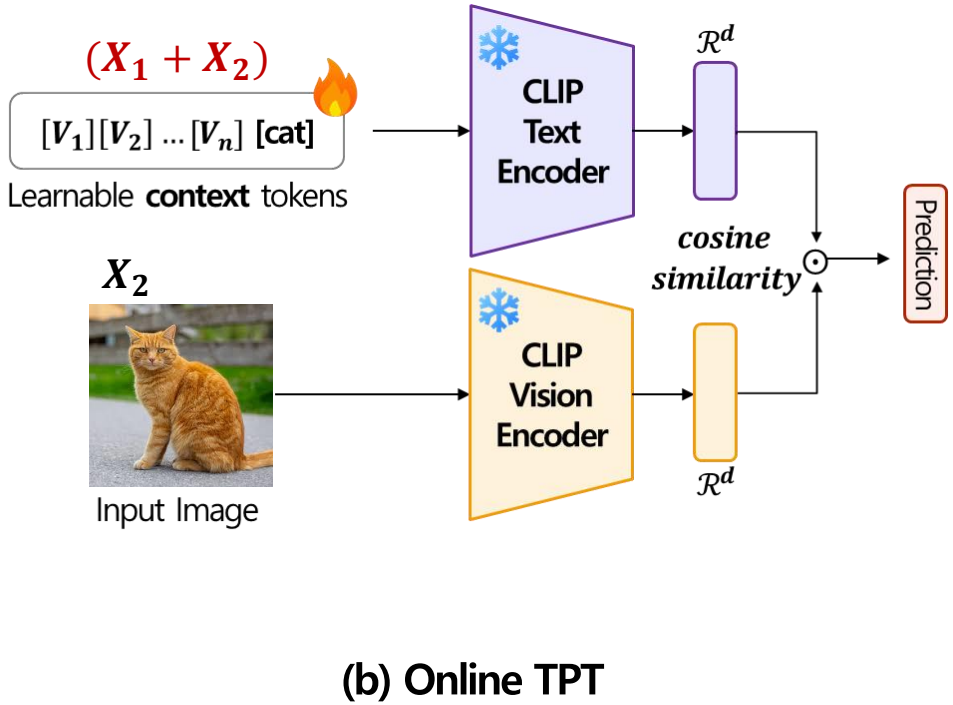
Methods

R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning

❖ VLM-based Test-time Prompt Tuning

- VLM 기반의 TPT(Test-time Prompt Tuning)은 두 가지 형식을 가짐
  1) 매 샘플마다 Learnable context parameters를 초기화하는 방식 (Episodic TPT)
  2) 매 샘플의 정보를 Learnable context parameters에 누적하여 반영하는 방식 (Online TPT)

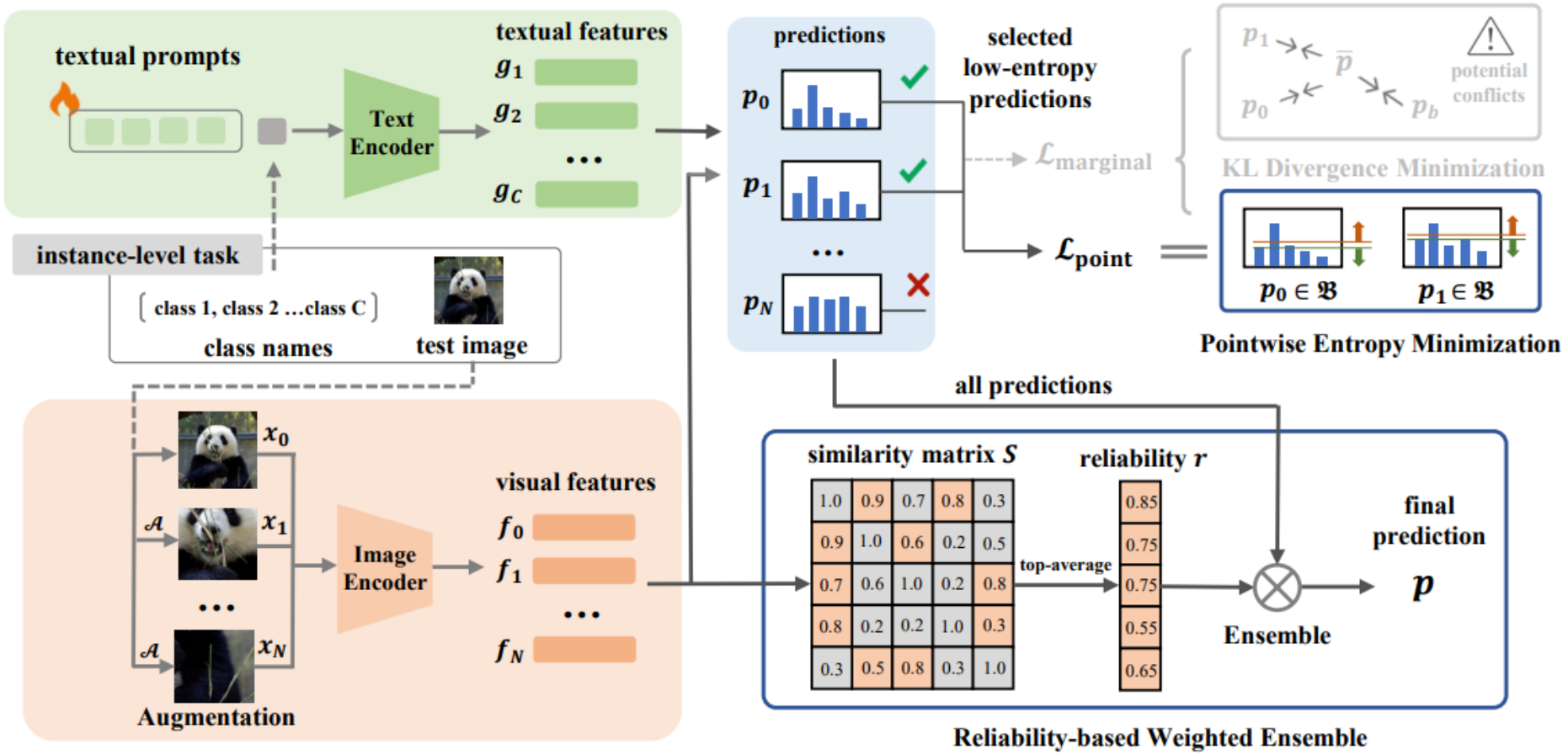




Methods

R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning

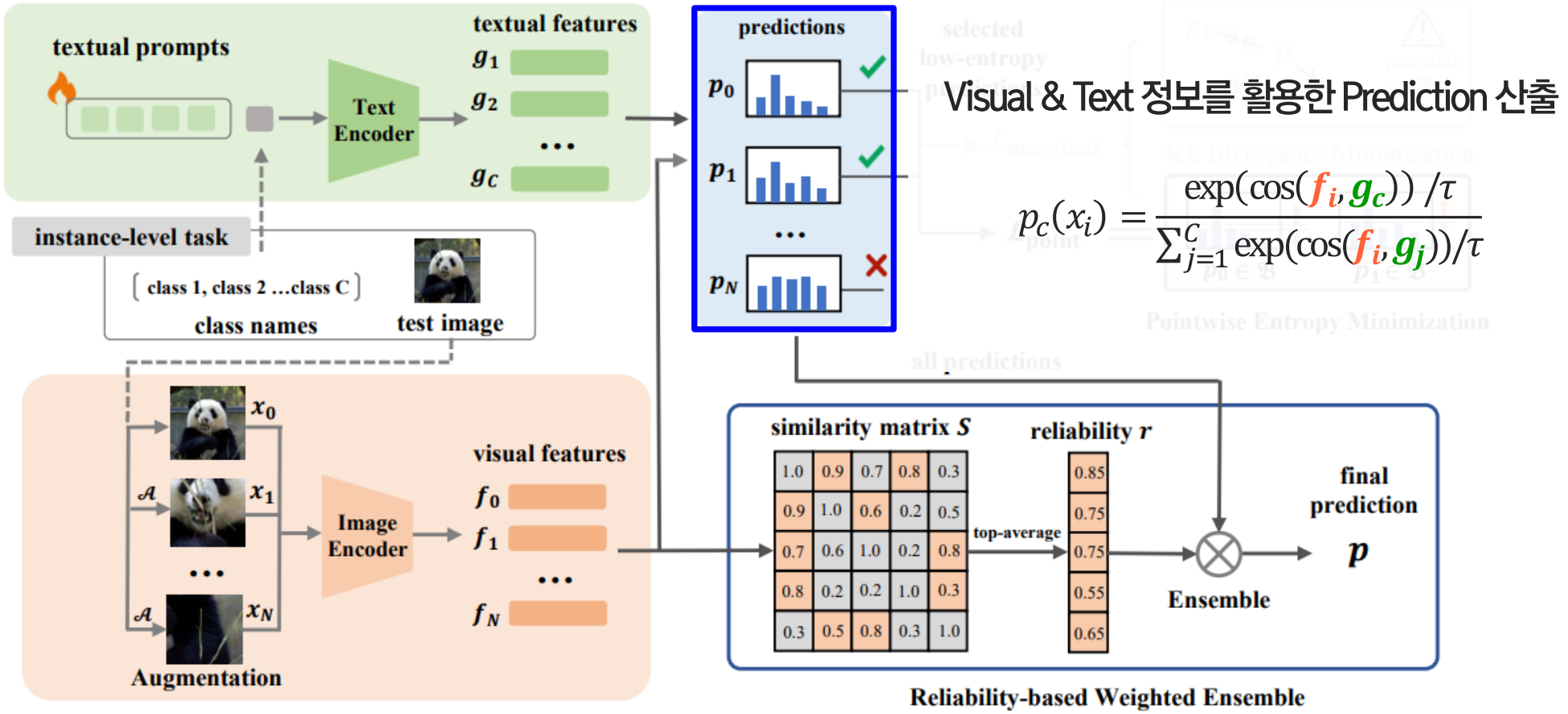
❖ R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning



Methods

R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning

❖ R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning



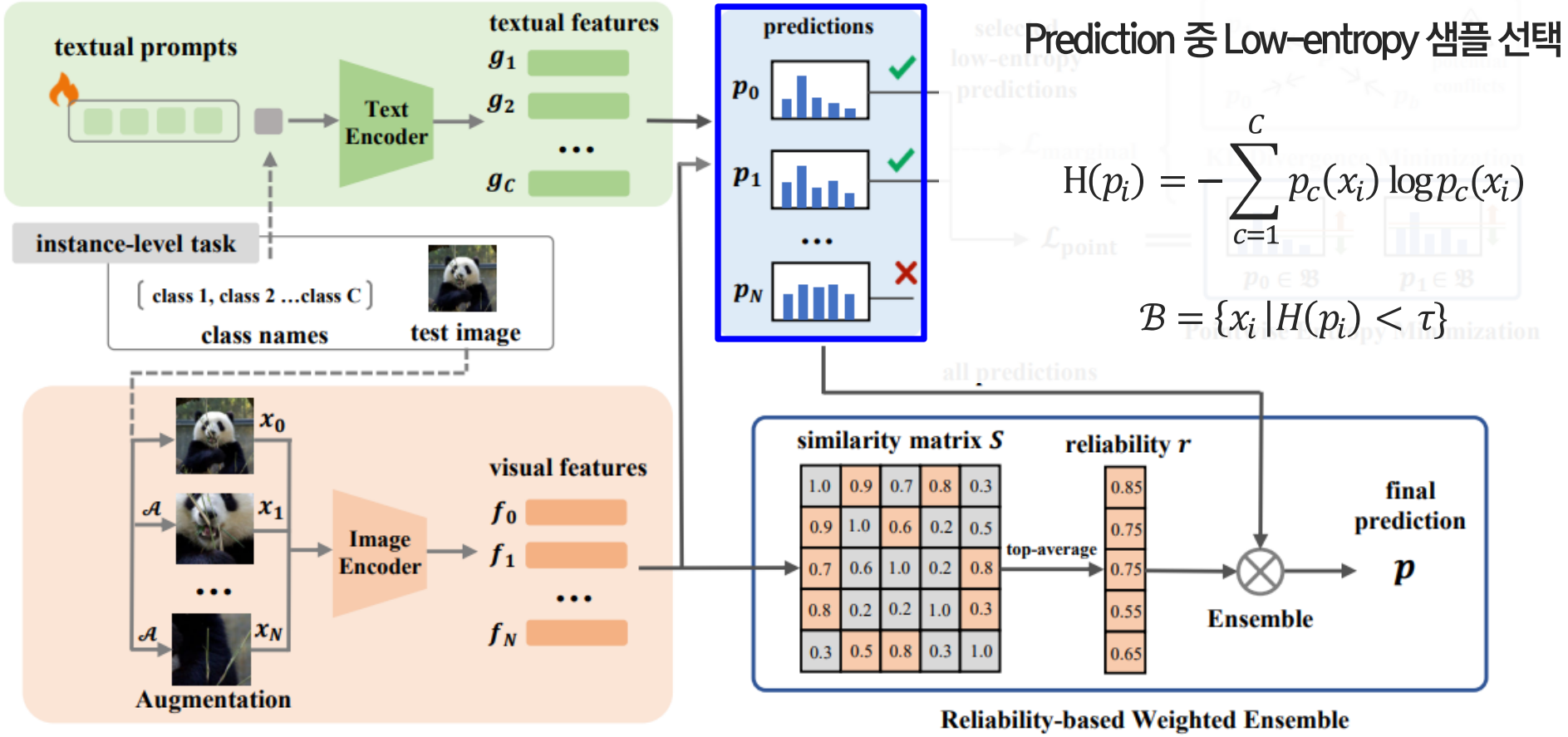
Visual & Text 정보를 활용한 Prediction 산출

$$p_c(x_i) = \frac{\exp(\cos(f_i, g_c))\ /\tau}{\sum_{j=1}^{C} \exp(\cos(f_i, g_j))/\tau}$$

Methods

R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning

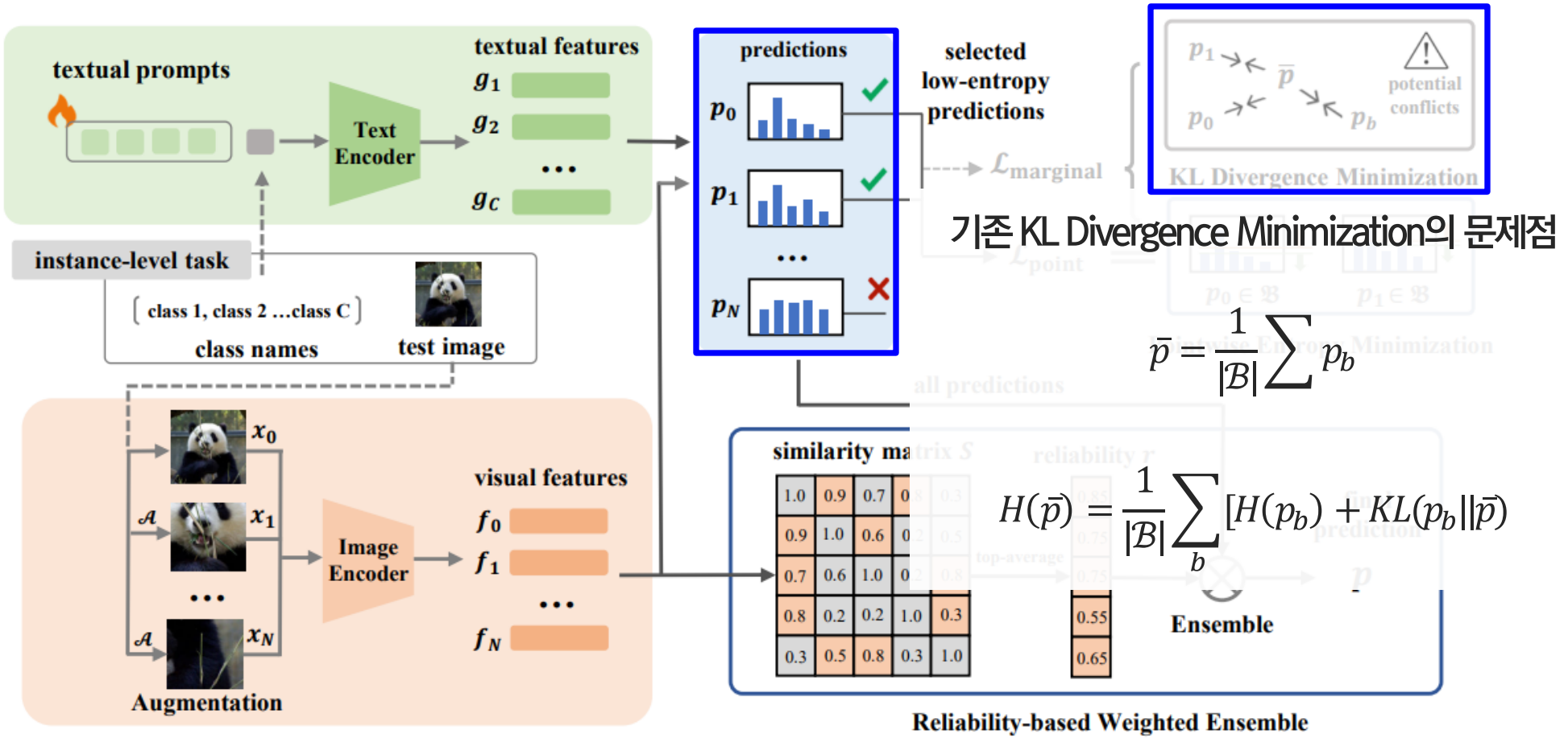
❖ R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning



Methods

R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning

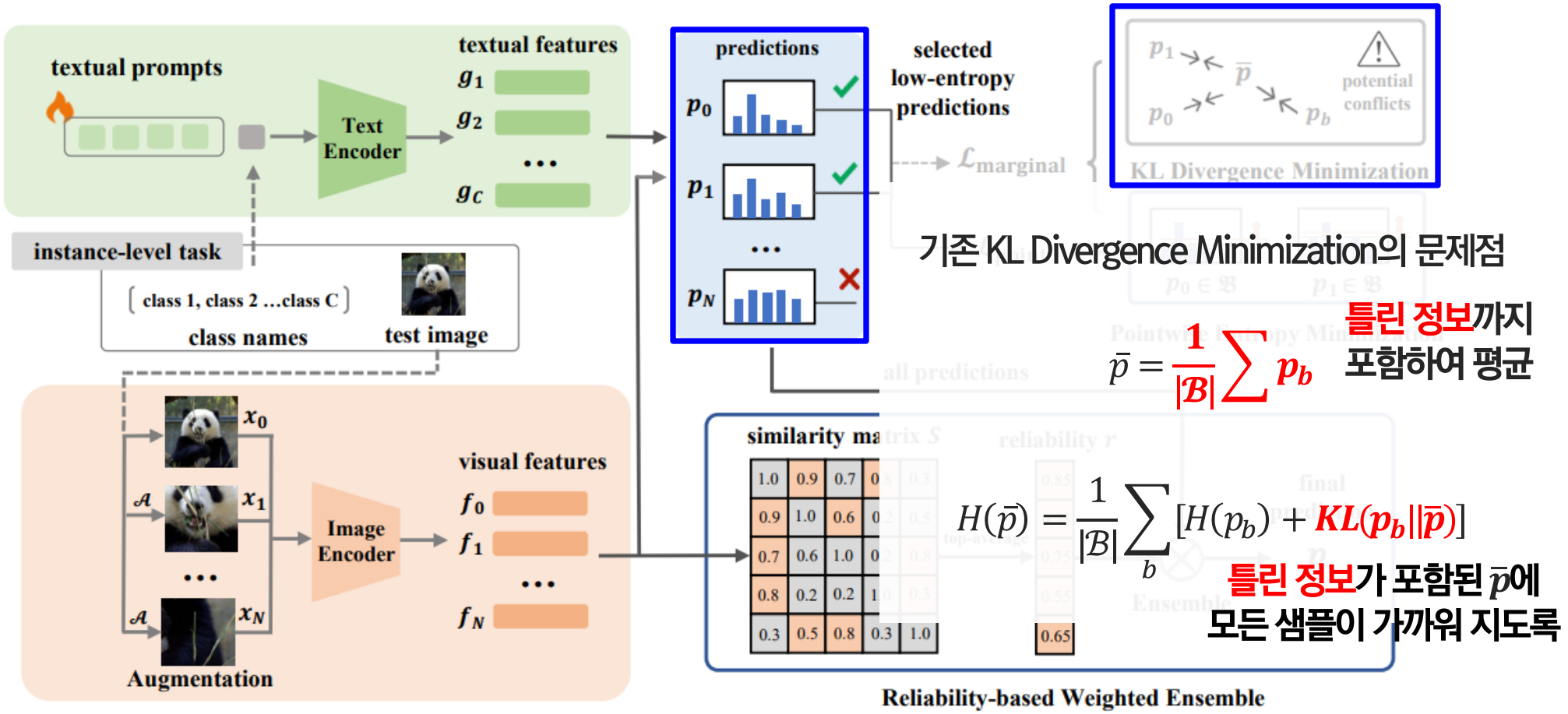
❖ R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning



Methods

R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning

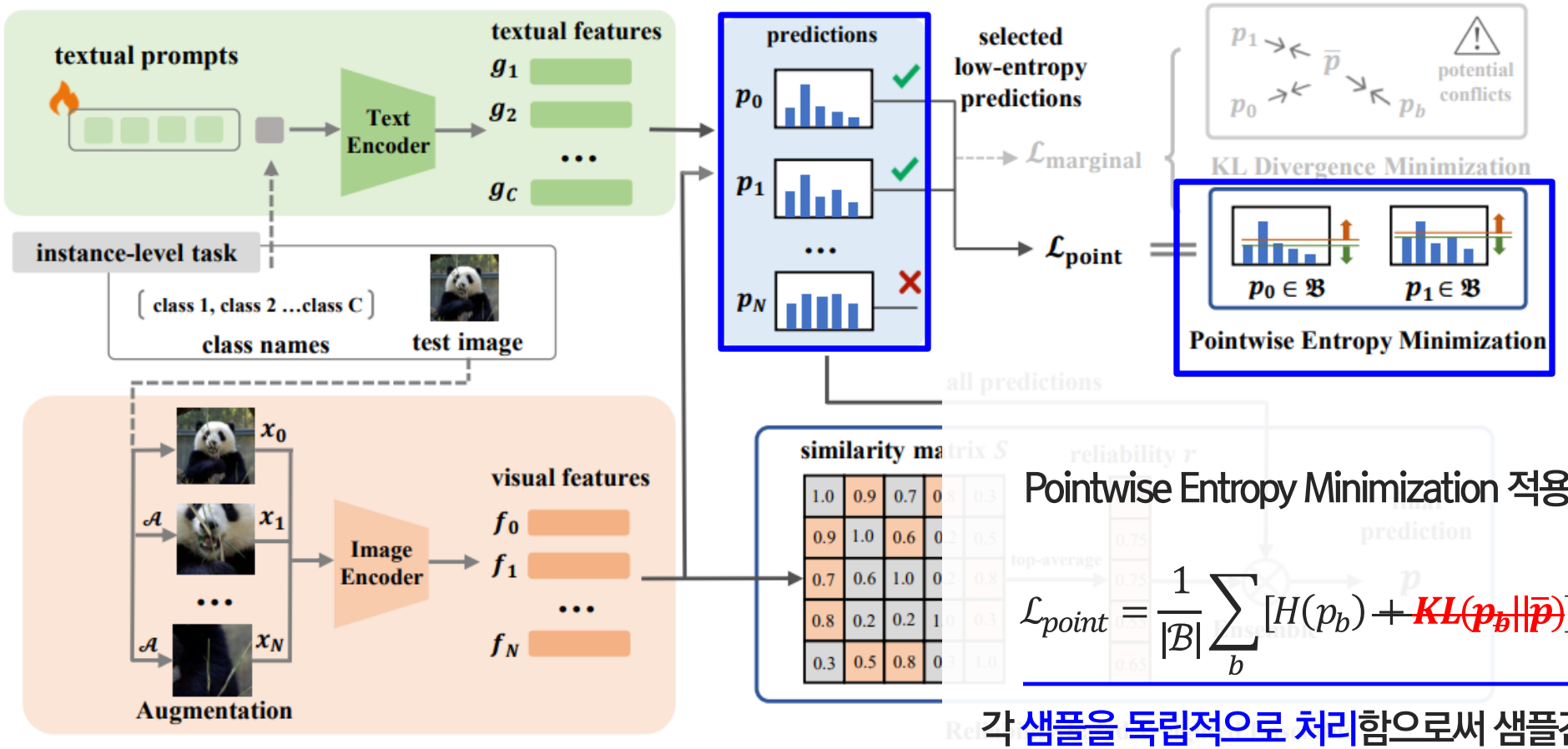
❖ R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning



Methods

R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning

❖ R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning



Pointwise Entropy Minimization 적용

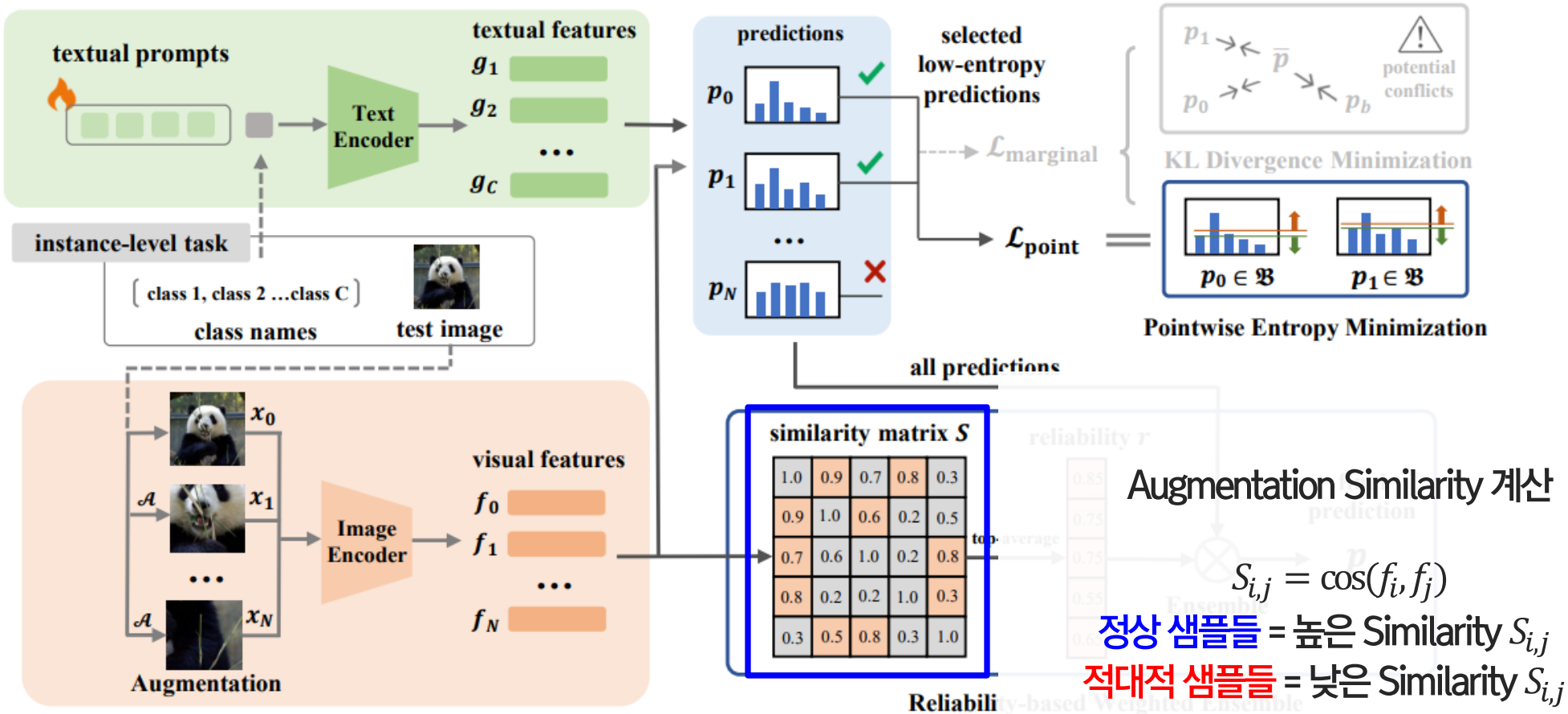
$$\mathcal{L}_{point} = \frac{1}{|\mathcal{B}|}\sum_{b}[H(p_b) + \cancel{KL(p_b||\bar{p})}]$$

각 샘플을 독립적으로 처리함으로써 샘플간 상호작용으로 인한 오류 전파를 차단함

Methods

R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning

❖ R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning



Augmentation Similarity 계산

$$S_{i,j} = \cos(f_i, f_j)$$

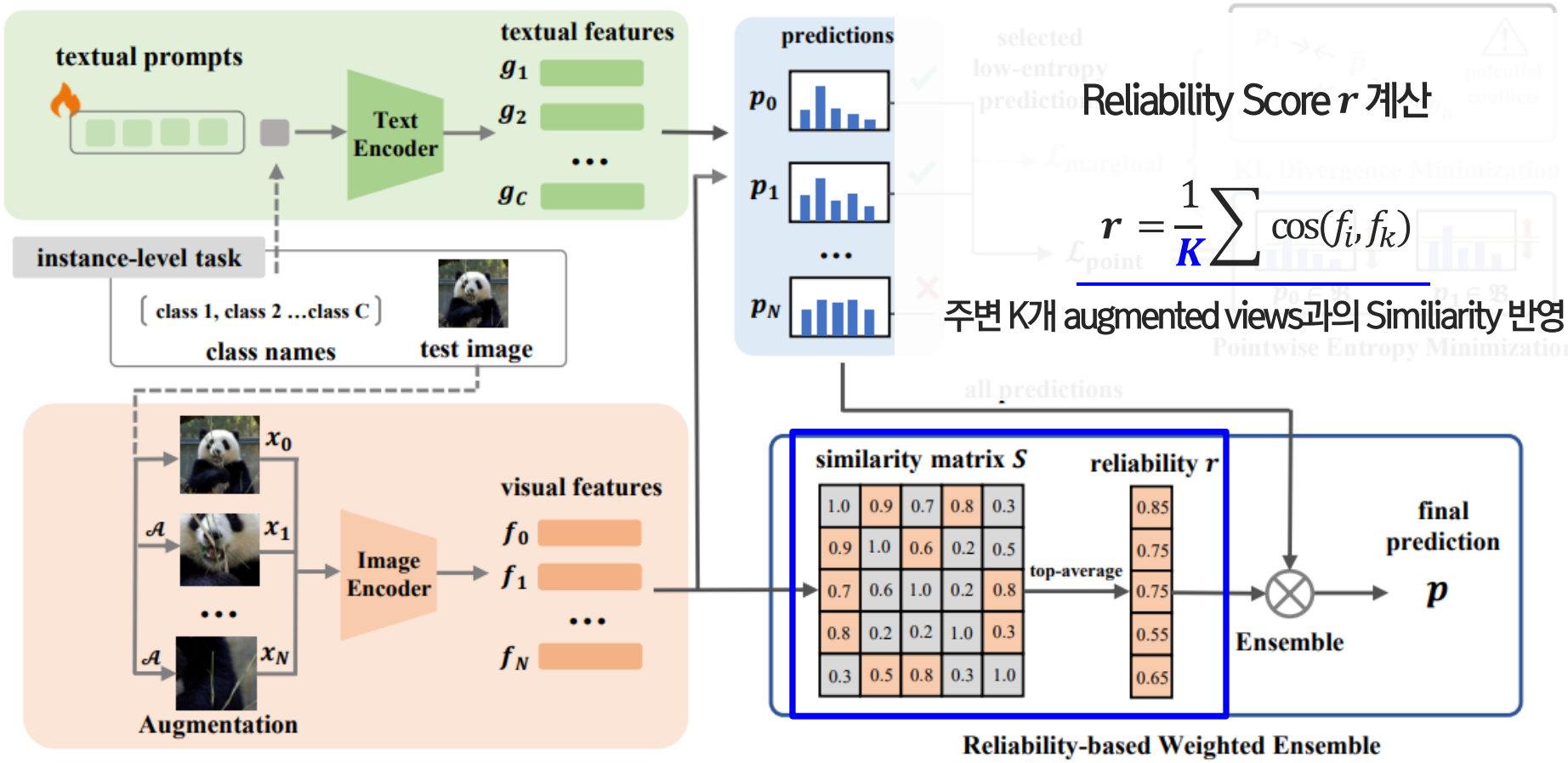
정상 샘플들 = 높은 Similarity $S_{i,j}$

적대적 샘플들 = 낮은 Similarity $S_{i,j}$

Methods

R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning

❖ R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning



Reliability Score r 계산

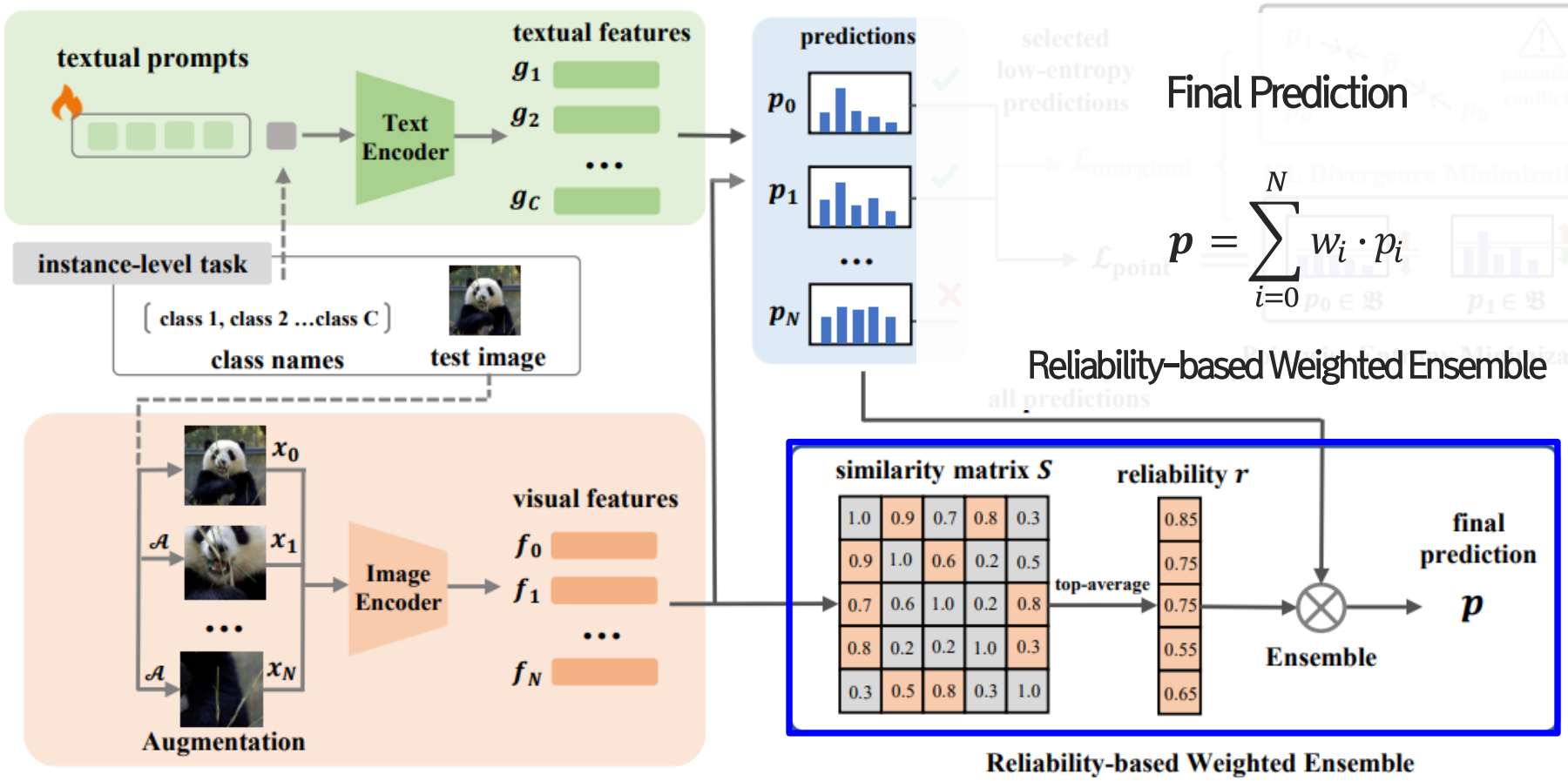
$$r = \frac{1}{K}\sum \cos(f_i, f_k)$$

주변 K개 augmented views과의 Similiarity 반영

Methods

R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning

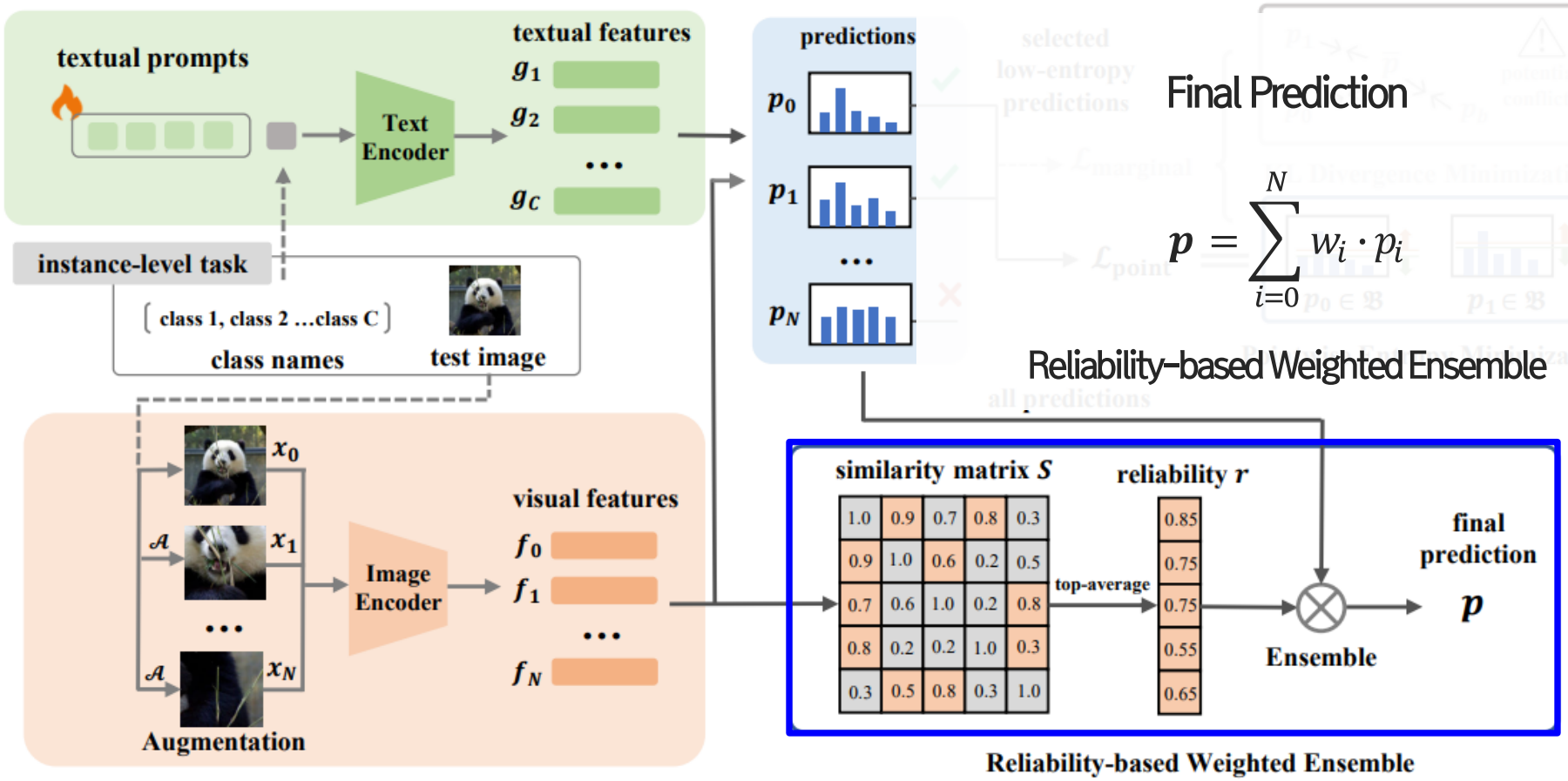
❖ R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning



Methods

R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning

❖ R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning



Experiments

R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning

❖ R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning

Method	Caltech101		Pets		Cars		Flower102		Aircraft		DTD		EuroSAT		UCF101		Avg.	
	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.
CLIP [32]	94.0	0.0	**88.3**	0.0	65.5	0.0	67.4	0.0	23.9	23.9	44.4	0.0	42.2	0.0	65.2	0.0	61.4	3.0
Ensemble	91.9	74.7	86.2	51.2	65.7	26.0	65.9	36.3	23.4	23.4	43.2	25.1	28.2	2.2	63.0	30.6	58.4	33.7
TPT [37]	94.1	0.0	87.4	0.0	66.5	0.0	69.1	0.0	23.4	23.4	**46.9**	0.0	**42.6**	0.0	**67.9**	0.0	62.2	2.9
C-TPT [50]	93.9	0.0	88.2	0.0	65.8	0.0	**69.6**	0.0	23.9	23.9	45.9	0.0	42.3	0.0	65.5	0.0	61.9	3.0
MTA [53]	**94.3**	72.1	88.0	51.8	**67.7**	18.5	67.4	27.9	**25.0**	25.0	46.5	16.2	42.5	1.2	67.5	27.5	**62.3**	30.0
R-TPT	93.7	82.0	87.2	60.2	67.0	34.7	68.7	44.6	23.9	23.9	46.4	32.8	34.7	8.5	67.2	43.2	61.1	41.2

Fine-grained Classification Task 실험 성능

Clean accuracy (Acc.)는 열세하지만, Adversarial accuracy (Rob.)에서 크게 우세함

Experiments

R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning

❖ R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning

Method	ImageNet		ImageNet-A		ImageNet-V2		ImageNet-R		ImageNet-S		Avg.	
	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.
CLIP [32]	58.2	0.1	21.8	0.0	51.5	0.1	56.1	0.8	33.3	0.5	44.2	0.3
Ensemble	58.0	40.1	22.6	10.1	52.0	37.2	51.3	39.3	29.5	20.7	42.7	29.5
TPT [37]	60.7	0.3	26.5	0.0	54.8	0.3	**58.9**	1.8	35.0	1.4	**47.2**	0.7
C-TPT [50]	60.4	0.1	24.1	0.0	54.3	0.1	57.7	1.0	34.7	0.9	46.2	0.4
MTA [53]	60.4	30.0	27.5	5.6	54.2	24.6	58.4	29.8	**35.2**	11.3	47.1	20.3
R-TPT	**60.9**	47.7	**28.4**	14.4	**54.9**	41.6	57.6	46.9	34.0	26.2	47.1	35.4

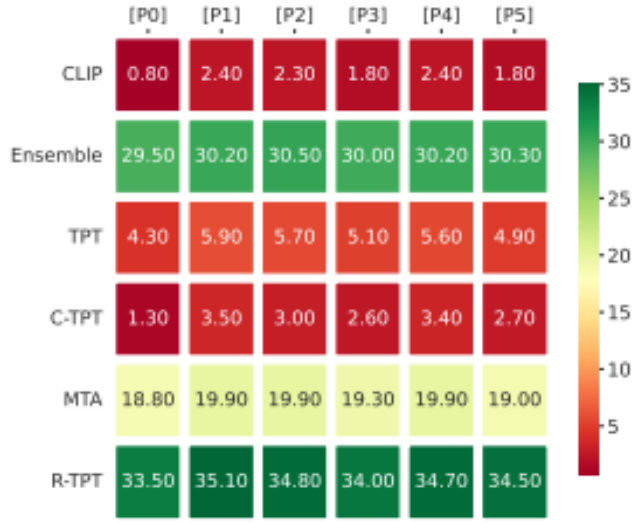
ImageNet-OOD Classification Task 실험 성능

마찬가지로, Clean accuracy (Acc.)는 열세하지만, Adversarial accuracy (Rob.)에서 크게 우세함

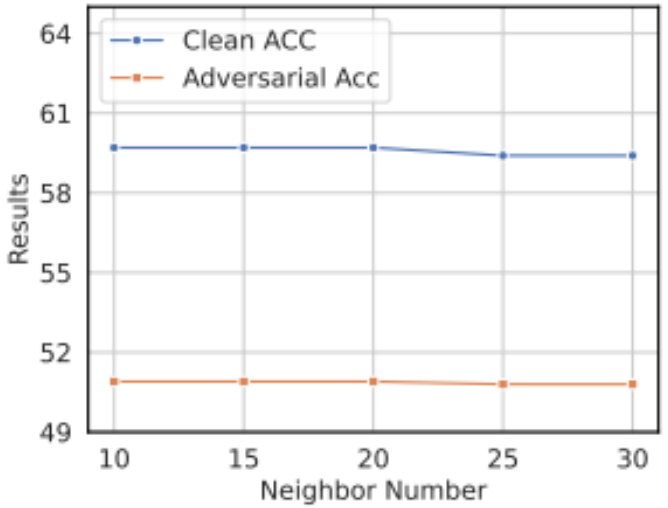
Experiments

R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning

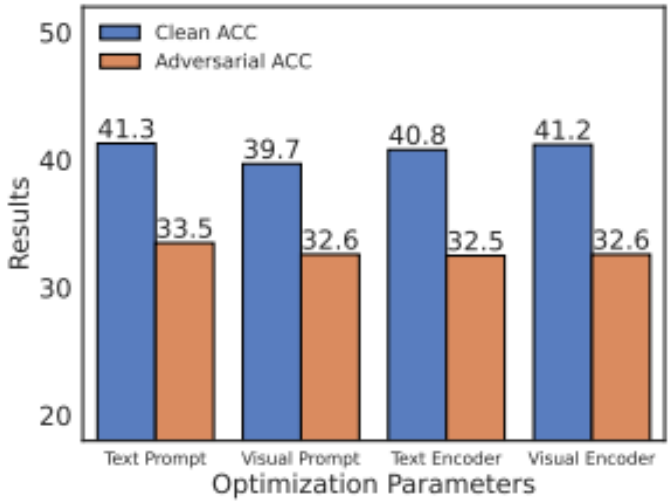
❖ R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning



(a) Different prompt templates.



(b) Sensitivity of neighbor number.



(c) Sensitivity of optimization parameters.

(a) Prompt templates에 따른 성능 변화

사용하는 Prompt template에 관계없이,
제안 방법론이 높은 Rob. 성능을 보임

(b) Neighbor 수에 따른 성능 변화

K-NN의 Neighbor 수에 관계없이,
일관된 Acc., Rob. 성능을 보임

(c) Update 파라미터 종류에 따른 성능 변화

어느 파라미터를 Update하는지에 관계없이,
제안 방법론이 높은 Acc., Rob. 성능을 보임

여러가지 조건에 관계없이 Robust한 성능을 보이는 Test-time Prompt Tuning 알고리즘 R-TPT

Batclip: Bimodal online test-time adaptation for clip (ICCV 2025)

Methods

Batclip: Bimodal online test-time adaptation for clip

❖ Batclip: Bimodal online test-time adaptation for clip [3]

- 2025년에 제안된 Bimodal online test-time adaptation 방법론
- Text, Visual 두 모달리티를 모두 고려하면서 일부 layer만 업데이트하는 효율적인 test-time adaptation 방법론 제안

BATCLIP: Bimodal Online Test-Time Adaptation for CLIP

Sarthak Kumar Maharana[1], Baoming Zhang[1], Leonid Karlinsky[2], Rogerio Feris[2], Yunhui Guo[1]
[1]The University of Texas at Dallas [2]MIT-IBM Watson AI Lab
{sarthak.maharana, yunhui.guo}@utdallas.edu

Abstract

Although open-vocabulary classification models like Contrastive Language Image Pretraining (CLIP) have demonstrated strong zero-shot learning capabilities, their robustness to common image corruptions remains poorly understood. Through extensive experiments, we show that zero-shot CLIP lacks robustness to common image corruptions during test-time, necessitating the adaptation of CLIP to unlabeled corrupted images using test-time adaptation (TTA). However, we found that existing TTA methods have severe limitations in adapting CLIP due to their unimodal nature. To address these limitations, we propose BATCLIP,

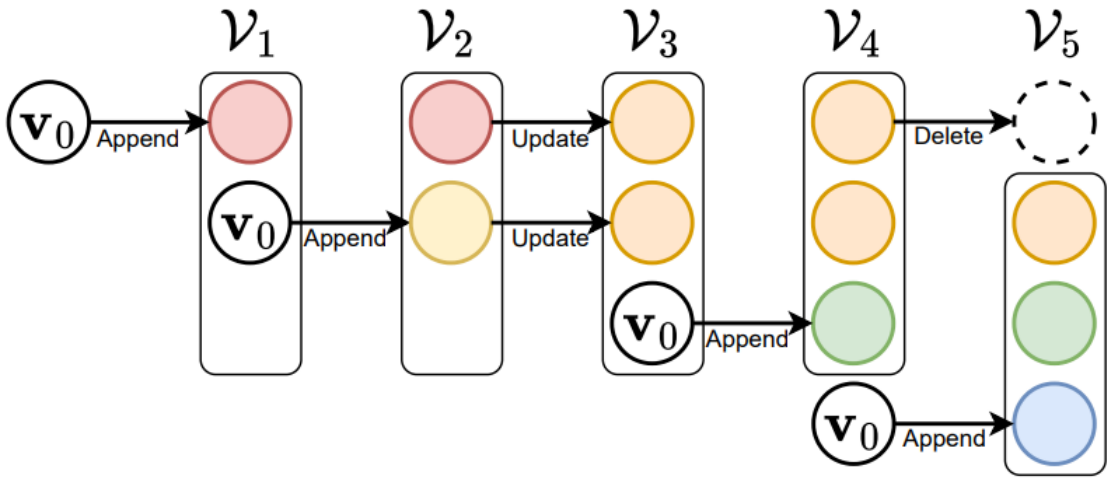
achieve. For example, models deployed for safety-critical applications like autonomous driving [1], could face rapid distributional shifts of blurriness, pixel changes, snowy nights, or other weather conditions [38]. In particular, our findings on the zero-shot performance of CLIP with a ResNet-101 [17] vision backbone reveals that the accuracy on the test set of CIFAR100 [21] with *Gaussian* noise of severity level 5, plummets to 10.79% from 49% on the clean set. Similar trends are observed with ViT-B/16, -B/32, and -L/14 [11] as backbones. The performance degradation caused by image corruption can have significant consequences in real-world scenarios, particularly in safety-critical applications like self-driving cars.

Methods

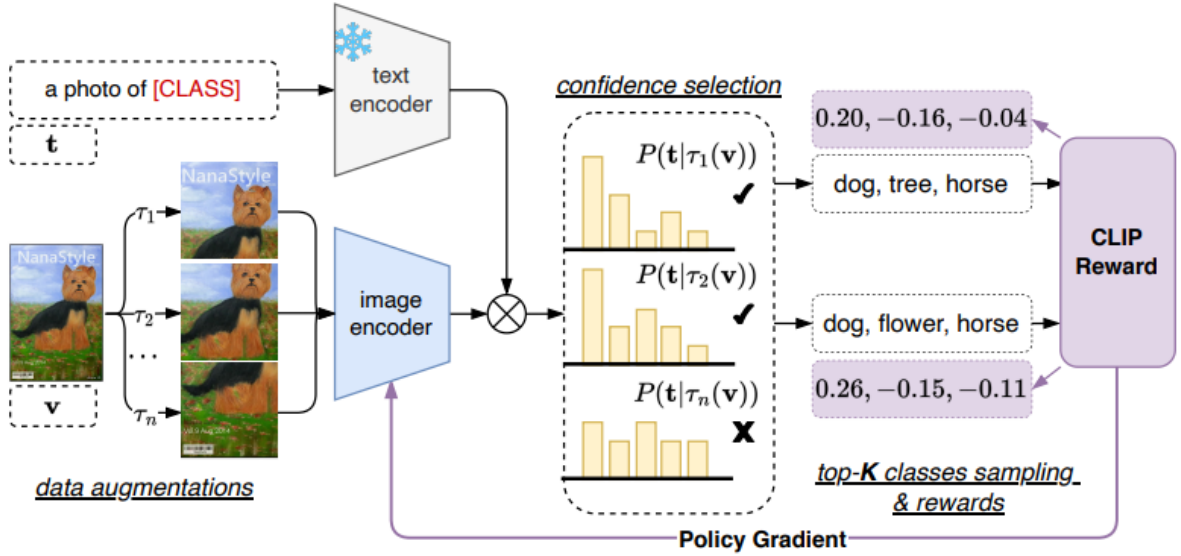
Batclip: Bimodal online test-time adaptation for clip

❖ Batclip: Bimodal online test-time adaptation for clip [3]

- 대부분의 VLM-based TTA 연구는 특정 모달리티만 고려하는 경우가 다수



Text modality만 고려하는 Dynaprompt 알고리즘 [4]



Visual modality만 고려하는 RLCF 알고리즘 [5]

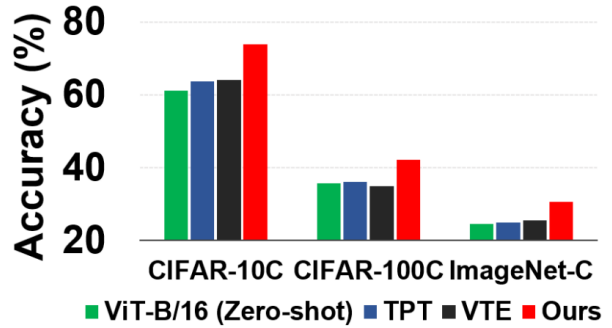
Methods

Batclip: Bimodal online test-time adaptation for clip

❖ Batclip: Bimodal online test-time adaptation for clip [3]

- 대부분의 VLM-based TTA 연구는 특정 모달리티만 고려하는 경우가 다수

a multimodal model like CLIP to fully leverage its multi-modal nature for adaptation. Specifically, it prevents the encoders from jointly adjusting their features, resulting in suboptimal alignment between the visual and text modalities after adaptation. For instance, when a test image in-



BATCLIP;
특정 모달리티(Text or Visual)만 고려하는 것은
Suboptimal alignment 문제를 야기할 수 있음

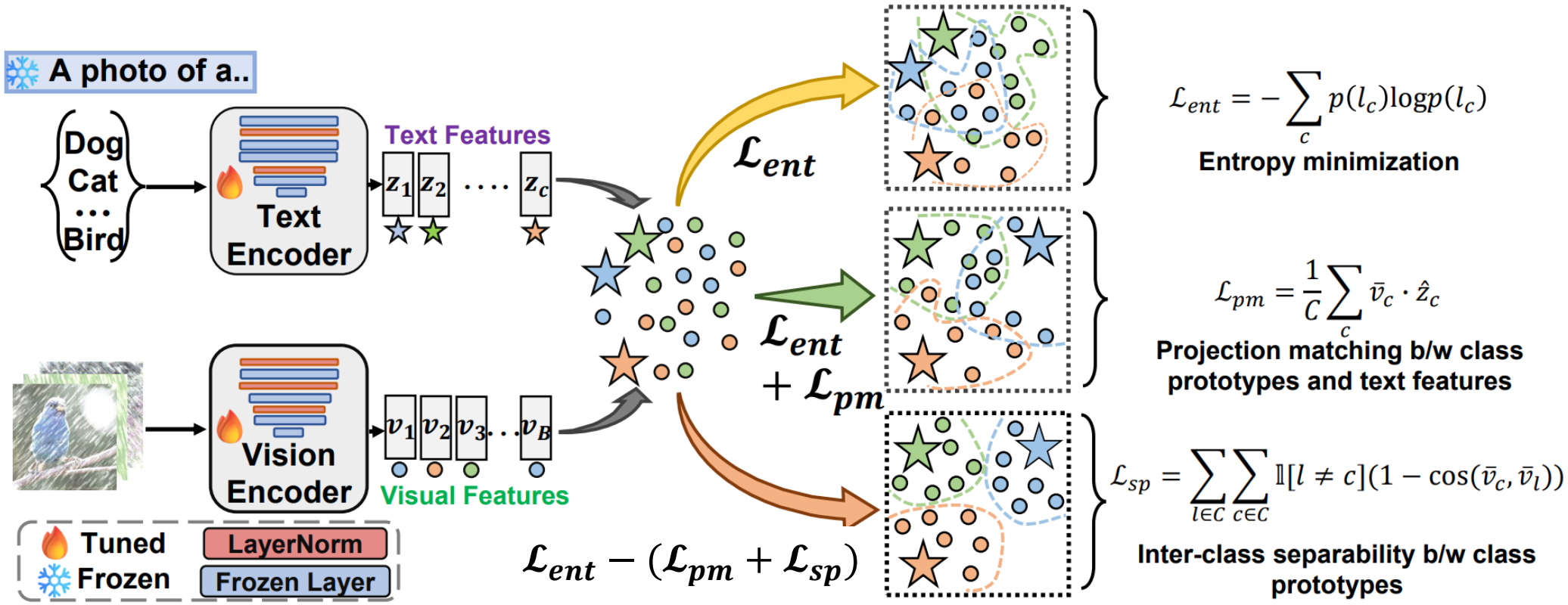
Text modality만 고려하는 Dynaprompt 알고리즘 [2]

Visual modality만 고려하는 RLCF 알고리즘 [3]

Methods

Batclip: Bimodal online test-time adaptation for clip

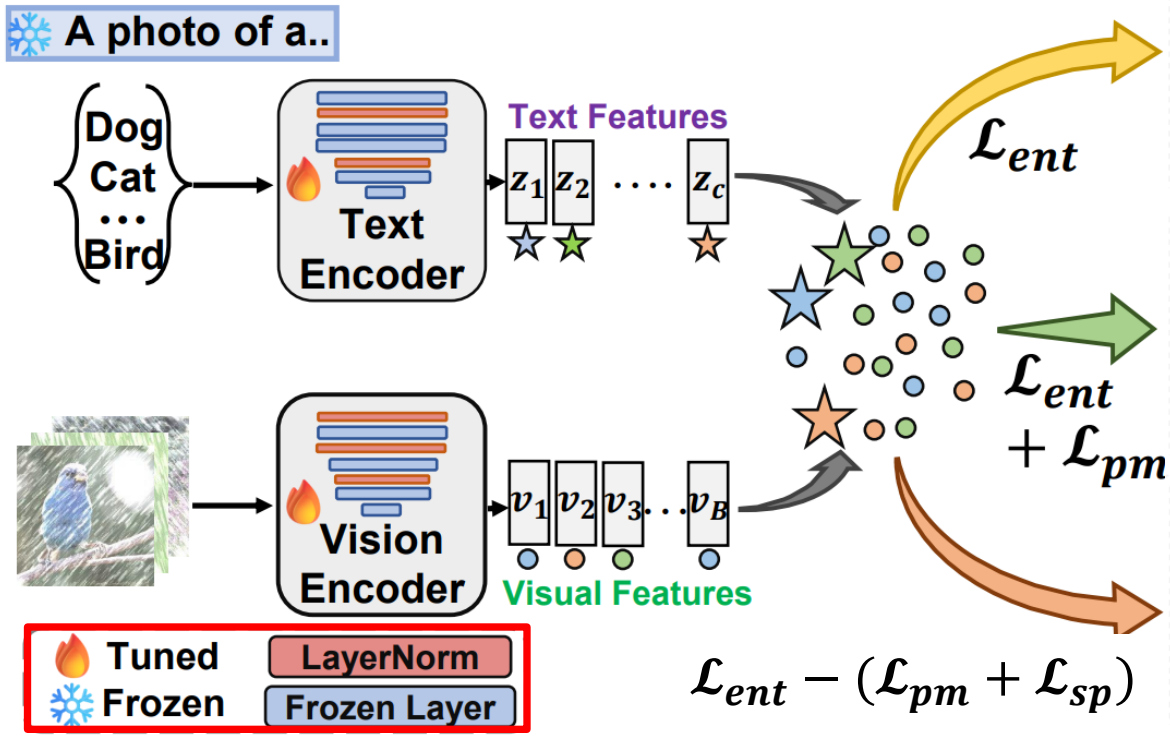
❖ Batclip: Bimodal online test-time adaptation for clip [3]



Methods

Batclip: Bimodal online test-time adaptation for clip

❖ Batclip: Bimodal online test-time adaptation for clip [3]



LayerNorm 부분을 업데이트하는 이유

1) **Distribution shift**를 다루는 유일한 Layer

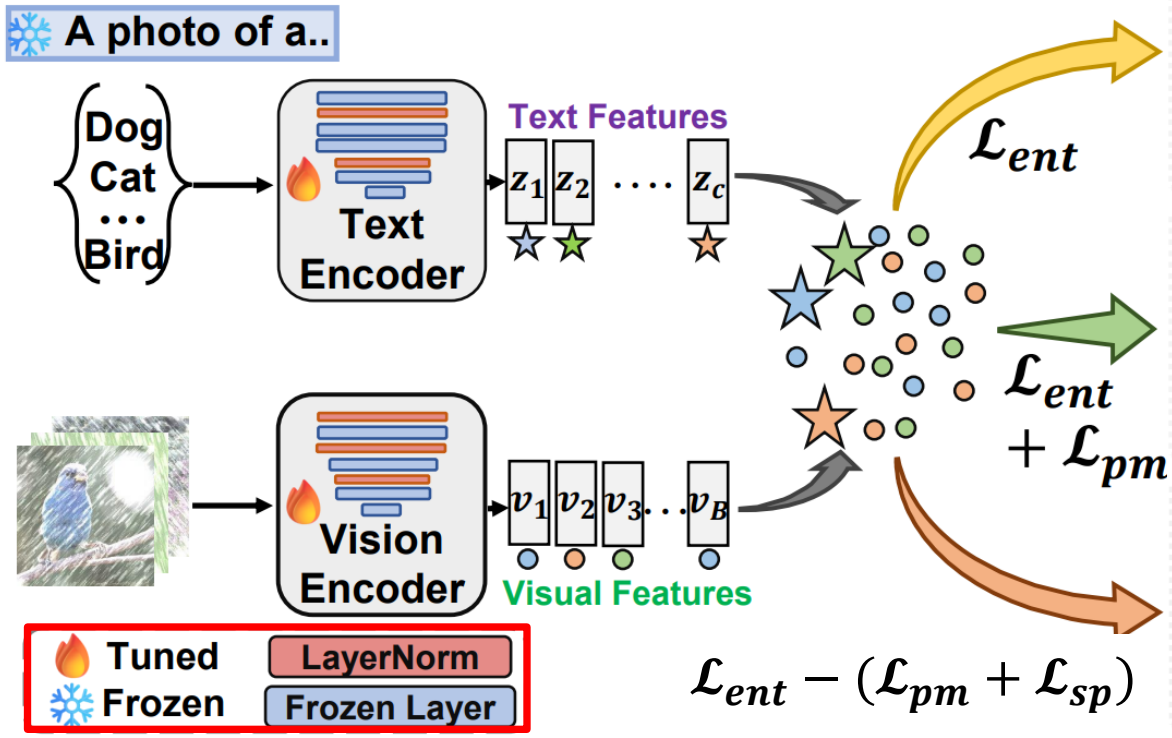
$$\hat{x} = \frac{x - \mu}{\sigma}$$

→ Feature distribution 자체를 재정렬하는 역할

Methods

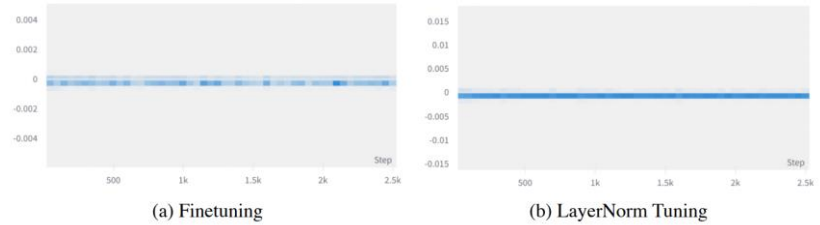
Batclip: Bimodal online test-time adaptation for clip

❖ Batclip: Bimodal online test-time adaptation for clip [3]



LayerNorm 부분을 업데이트하는 이유

2) 안정적인 Gradient 흐름



(a) Finetuning (b) LayerNorm Tuning

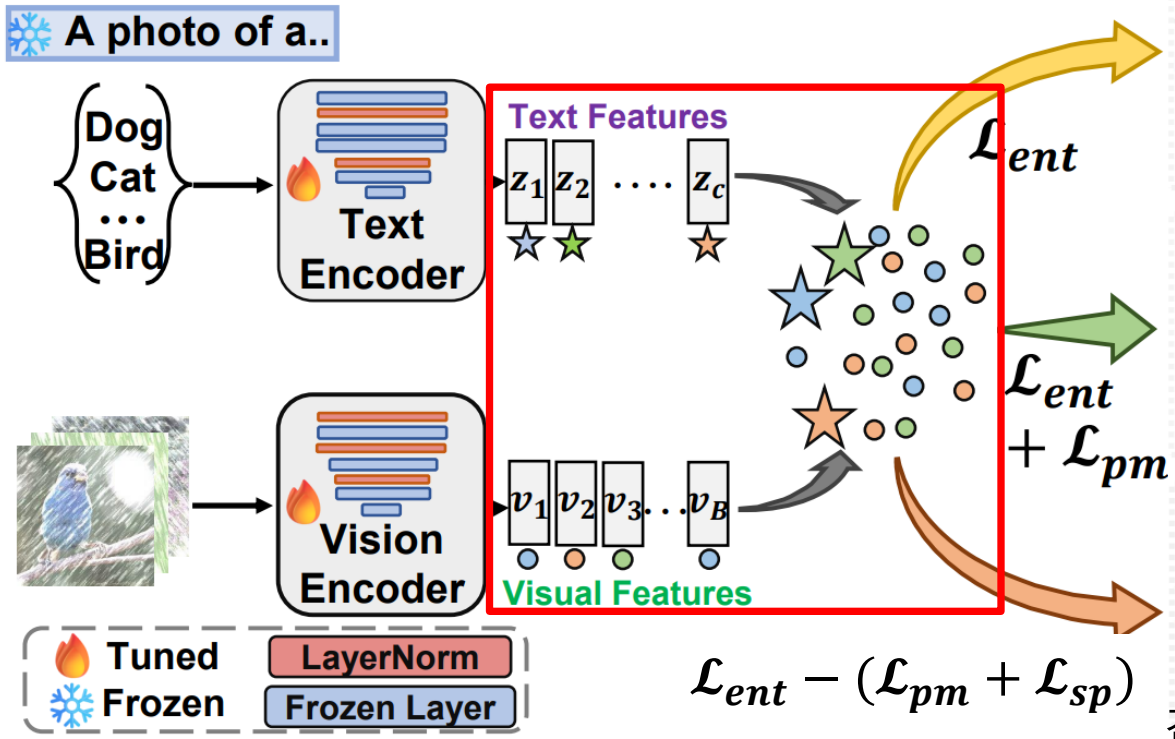
LayerNorm은 full parameter fine-tuning의 경우보다 Gradient variance가 낮음

→ 업데이트가 흔들리지 않고 안정적

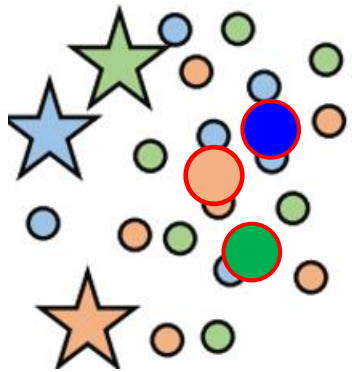
Methods

Batclip: Bimodal online test-time adaptation for clip

❖ Batclip: Bimodal online test-time adaptation for clip [3]



Visual Feature의 Feature 정보 v_k 를 활용한
Class Prototype $\bar{v}_c$구성



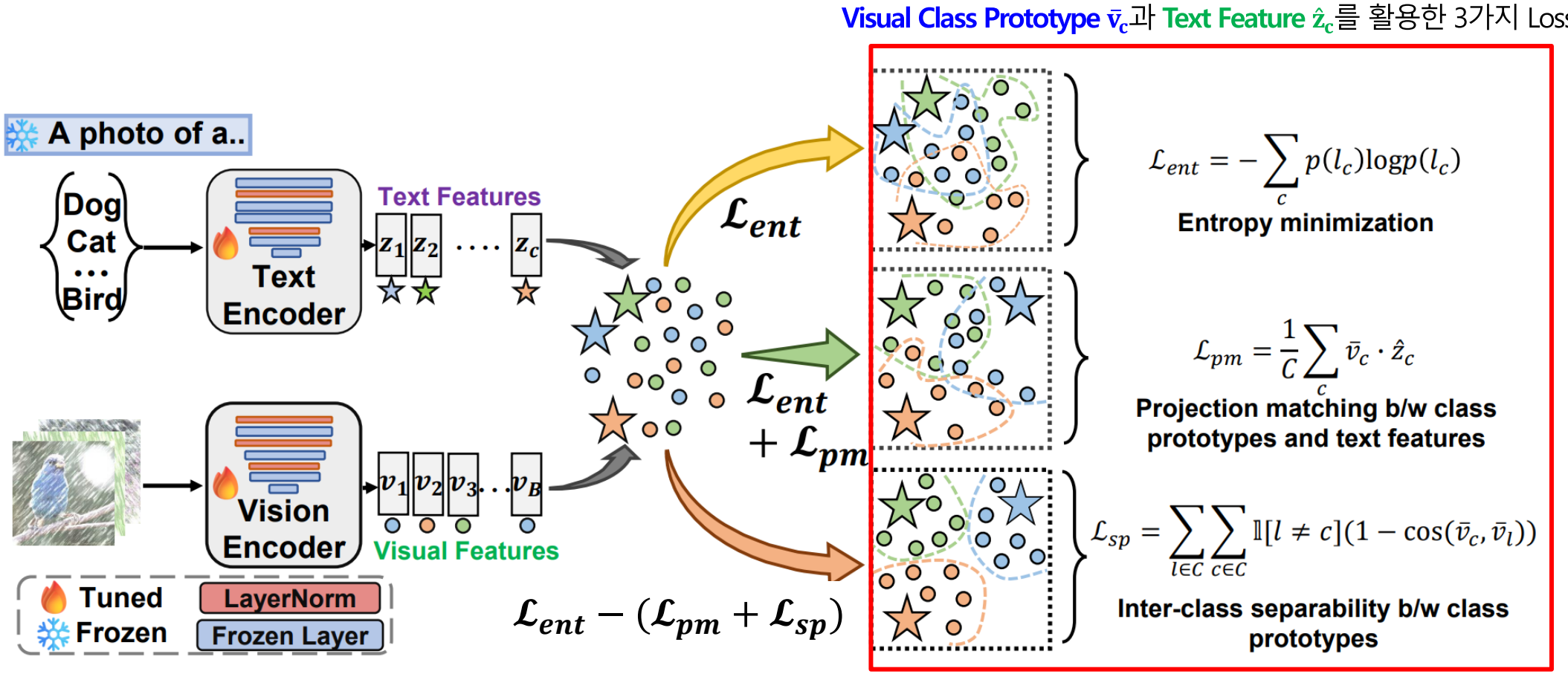
$$\bar{v}_c = \frac{1}{\sum_{k=1}^{B} \mathbb{I}[\hat{y}_k = c]} \sum_{k=1}^{B} \mathbb{I}[\hat{y}_k = c] \cdot v_k$$

같은 pseudo-label class를 갖는 **visual feature들의** **평균**

Methods

Batclip: Bimodal online test-time adaptation for clip

❖ Batclip: Bimodal online test-time adaptation for clip [3]

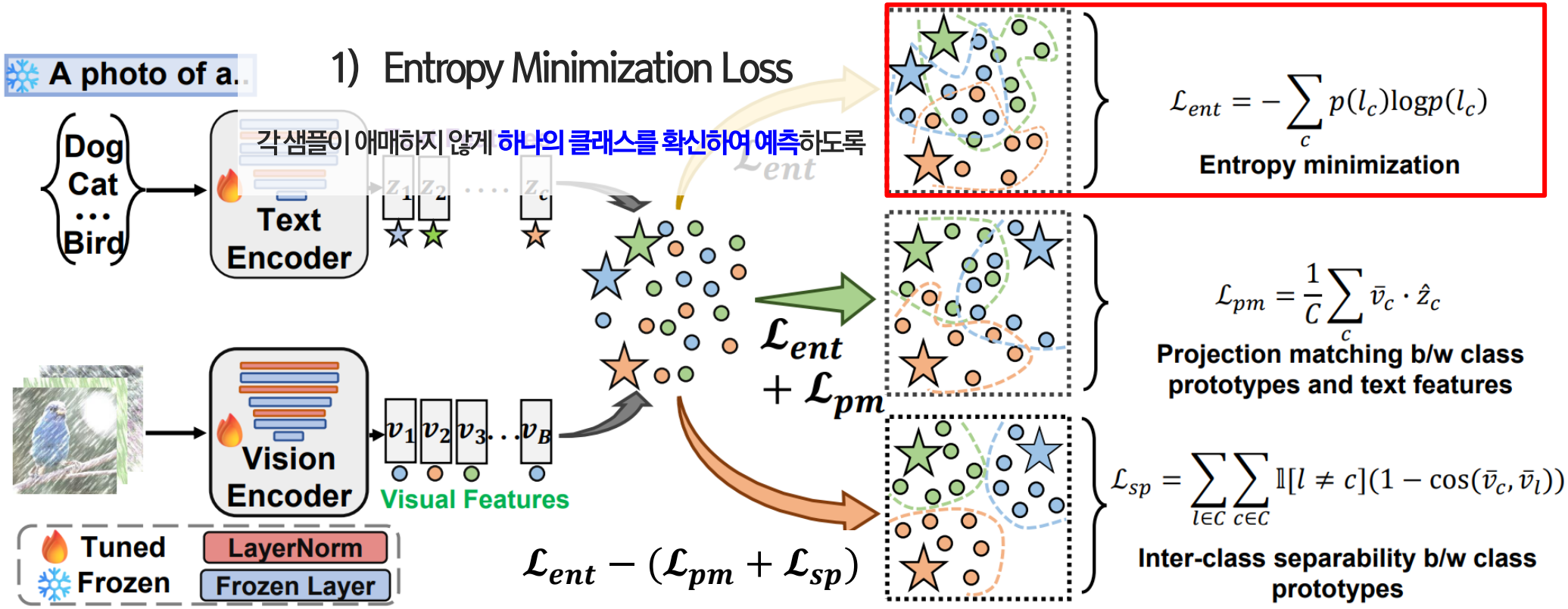


Methods

Batclip: Bimodal online test-time adaptation for clip

❖ Batclip: Bimodal online test-time adaptation for clip [3]

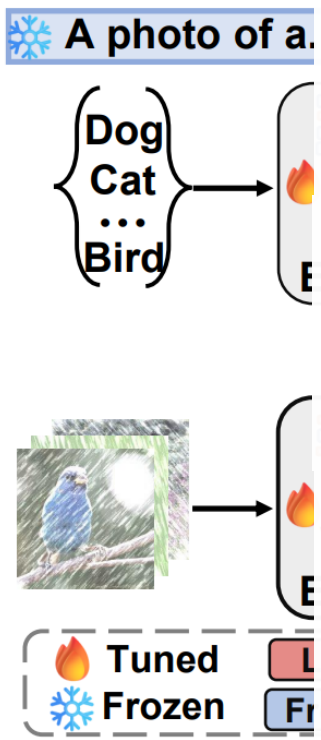
Visual Class Prototype $\bar{v}_c$과 Text Feature $\hat{z}_c$를 활용한 3가지 Loss



Methods

Batclip: Bimodal online test-time adaptation for clip

❖ Batclip: Bimodal online test-time adaptation for clip [3]



1) Entropy Minimization Loss

각 샘플이 애매하지 않게 하나의 클래스를 확신하여 예측하도록

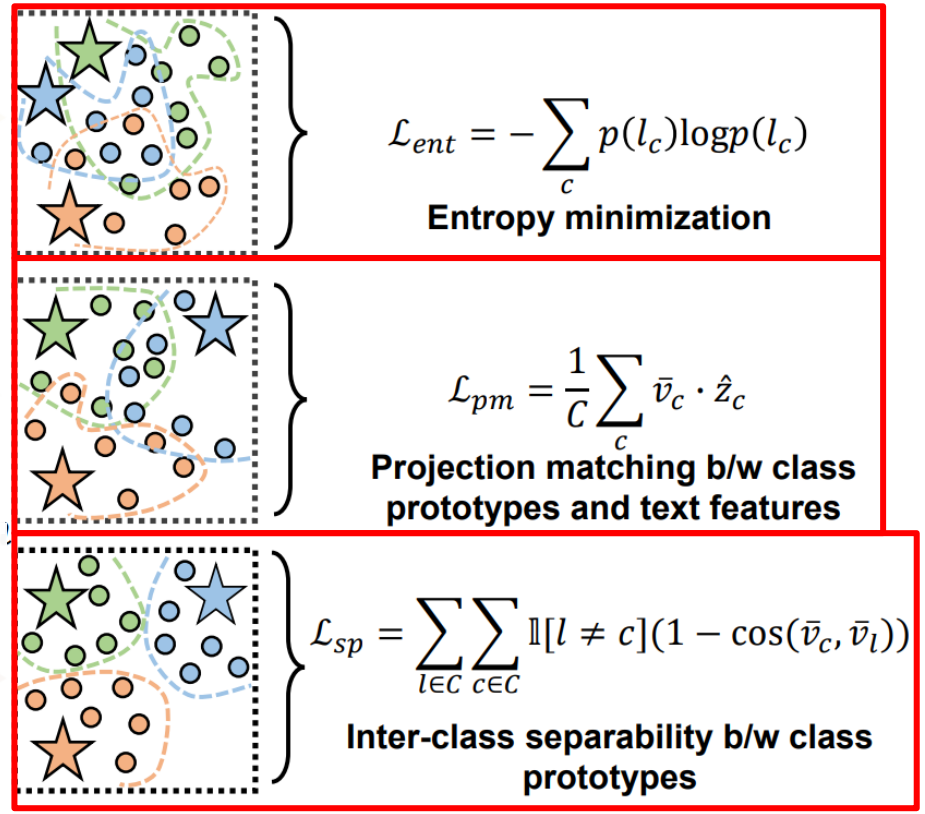
2) Projection Matching Loss

동일 클래스의 이미지와 텍스트를 서로 가깝게 만들도록

3) Inter-class Separability Loss

서로 다른 클래스의 Prototype들끼리 멀어지도록

Visual Class Prototype $\bar{v}_c$과 Text Feature $\hat{z}_c$를 활용한 3가지 Loss

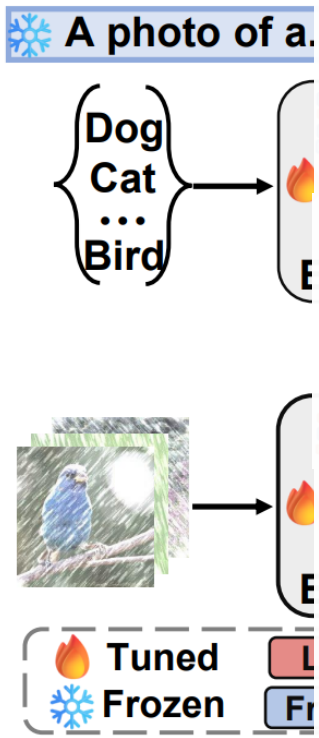


Methods

Batclip: Bimodal online test-time adaptation for clip

❖ Batclip: Bimodal online test-time adaptation for clip [3]

Visual Class Prototype $\bar{v}_c$과 Text Feature $\hat{z}_c$를 활용한 3가지 Loss



1) Entropy Minimization Loss

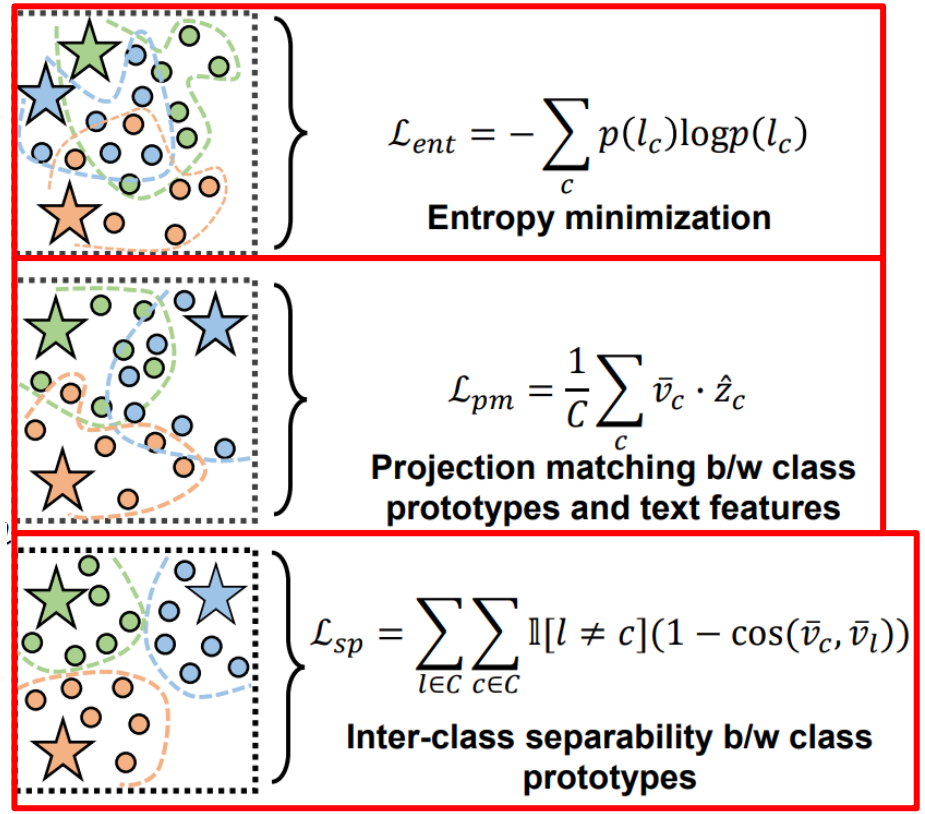
각 샘플이 애매하지 않게 하나의 클래스를 확신하여 예측하도록

2) Projection Matching Loss

동일 클래스의 이미지와 텍스트를 서로 가깝게 만들도록

3) Inter-class Separability Loss

서로 다른 클래스의 Prototype들끼리 멀어지도록



$$\mathcal{L}_{total} = \mathcal{L}_{ent} - (\mathcal{L}_{pm} + \mathcal{L}_{sp})$$

Experiments

Batclip: Bimodal online test-time adaptation for clip

❖ Batclip: Bimodal online test-time adaptation for clip

	Method	Venue	*Gaussian*	*Shot*	*Impulse*	*Defocus*	*Glass*	*Motion*	*Zoom*	*Snow*	*Frost*	*Fog*	*Brightness*	*Contrast*	*Elastic*	*Pixelate*	*JPEG*	Mean
CIFAR-10C	ZS	ICLR'21	37.92	41.7	54.42	71.75	40.89	67.93	73.62	73.89	77.35	70.22	84.45	62.36	53.81	47.65	59.43	61.16
	TENT	ICLR'21	15.49	18.28	38.12	81.59	21.73	76.32	82.35	**84.62**	82.19	80.60	**91.83**	80.55	63.52	58.57	54.71	62.03
	RoTTA	CVPR'23	39.17	42.90	55.41	72.18	41.32	68.02	74.01	74.38	78.01	70.80	84.80	63.19	54.62	49.32	60.15	61.89
	RPL	arXiv	15.47	17.43	40.73	**81.76**	20.08	69.89	**82.93**	84.43	83.19	81.84	91.80	79.42	64.89	54.07	54.90	61.52
	SAR	ICML'22	47.98	53.60	60.56	74.30	47.56	73.15	76.43	77.91	79.88	75.66	86.79	71.62	58.34	62.03	64.71	67.37
	TPT	NeurIPS'22	37.74	42.24	60.57	72.88	44.80	69.69	75.37	75.96	78.84	72.12	85.68	62.04	58.90	55.14	62.64	63.64
	VTE	ECCV-W'24	42.42	46.26	64.23	71.10	45.58	68.50	73.66	76.75	78.27	71.02	85.28	57.24	59.54	60.59	61.85	64.15
	WATT-P*	NeurIPS'24	44.70	49.66	60.48	74.35	46.33	72.21	76.28	78.18	80.97	74.90	87.66	68.63	58.35	56.13	63.47	66.15
	WATT-S*	NeurIPS'24	57.22	61.86	**67.63**	78.61	53.07	77.85	80.14	81.84	83.46	80.38	89.67	78.45	65.76	**68.63**	**67.67**	72.81
	Ours		**61.13**	**64.09**	65.76	80.51	**54.96**	**80.65**	81.94	83.04	**84.19**	**80.84**	88.95	**82.15**	**69.16**	62.68	66.64	**73.85**
CIFAR-100C	ZS	ICLR'21	19.64	21.40	25.26	42.54	20.03	43.17	47.95	48.35	49.74	41.57	57.02	34.58	29.15	23.96	32.43	35.79
	TENT	ICLR'21	7.60	8.21	8.33	51.81	7.95	**52.45**	55.34	54.16	36.17	50.92	65.63	**54.51**	36.52	**43.99**	35.81	37.96
	RoTTA	CVPR'23	20.65	22.22	26.17	42.48	20.26	42.90	47.88	48.75	49.92	41.86	57.00	34.52	29.27	25.08	32.88	36.12
	RPL	arXiv	6.44	7.09	7.09	**52.16**	11.81	52.33	**55.50**	**54.20**	38.83	**51.99**	**66.07**	54.45	**36.86**	42.83	**39.45**	38.47
	SAR	ICML'22	**25.30**	27.19	32.78	47.12	23.42	47.16	51.70	51.94	**52.48**	48.77	61.54	44.50	32.26	33.67	38.06	41.19
	TPT	NeurIPS'22	17.95	19.51	27.13	43.53	20.08	42.65	48.63	49.11	49.48	42.14	57.35	33.26	31.13	27.59	32.75	36.15
	VTE	ECCV-W'24	17.96	18.72	28.17	40.38	19.60	39.50	45.33	48.24	46.87	40.73	55.31	30.04	32.47	30.35	31.45	35.01
	WATT-P*	NeurIPS'24	20.53	22.22	27.3	43.14	17.51	42.37	48.17	47.31	49.34	41.49	57.07	35.29	27.75	25.83	31.89	35.81
	WATT-S*	NeurIPS'24	21.20	23.11	28.23	44.16	18.45	43.44	49.14	48.16	50.05	42.35	57.82	36.43	28.36	26.85	32.93	36.71
	Ours		24.91	**27.73**	**33.66**	50.11	**26.27**	48.49	54.85	52.35	51.62	48.38	63.27	45.21	34.74	32.38	37.31	**42.09**
ImageNet-C	ZS	ICLR'21	11.18	12.54	12.04	23.36	15.18	24.50	22.58	32.32	29.88	35.88	54.18	17.20	12.72	30.96	33.26	24.51
	TENT	ICLR'21	5.14	5.70	7.44	25.22	19.34	26.80	24.16	33.56	30.42	37.74	54.24	22.50	13.90	35.02	36.08	25.15
	RoTTA	CVPR'23	11.34	12.96	12.32	23.38	15.50	24.66	22.90	32.56	30.02	35.98	54.32	17.20	12.80	31.06	33.46	24.78
	RPL	arXiv	9.04	10.04	10.96	24.40	17.40	26.28	23.76	32.70	30.62	36.64	54.04	19.38	13.24	33.14	34.60	25.08
	SAR	ICML'22	17.96	20.46	20.68	25.72	23.04	29.52	26.04	34.92	32.74	39.00	55.00	27.14	19.64	36.66	37.50	29.73
	TPT	NeurIPS'22	8.48	9.46	10.20	23.98	15.16	25.10	24.00	33.94	32.12	37.08	55.64	16.54	13.68	34.06	33.58	24.87
	VTE	ECCV-W'24	9.18	10.76	10.78	24.72	14.30	24.36	25.24	35.38	32.46	38.16	55.56	16.14	14.26	38.72	33.98	25.60
	StatA (γ=0.1)	CVPR'25	10.56	11.22	10.86	22.11	14.12	22.39	20.85	31.39	29.90	34.63	53.32	16.00	12.46	29.55	32.79	23.47
	StatA (γ=-1)	CVPR'25	11.83	13.02	12.41	23.71	15.02	23.64	21.79	31.80	30.22	36.52	54.15	17.56	13.00	32.01	33.37	24.67
	Ours		**19.32**	**21.38**	**19.60**	**26.58**	**21.94**	**30.88**	**29.02**	**36.48**	32.00	**40.98**	**56.72**	**26.14**	**23.74**	37.67	**38.34**	**30.72**

Experiments

Batclip: Bimodal online test-time adaptation for clip

❖ Batclip: Bimodal online test-time adaptation for clip

Dataset	Domain	ZS	TENT	WATT-P*	WATT-S*	Ours
OfficeHome	Art	78.29	79.4	80.35	**80.43**	79.86
	Clipart	64.03	63.18	66.85	**66.9**	66.44
	Product	87.11	88.42	87.5	87.54	**88.51**
	Real World	88.96	89.6	**90.04**	89.99	89.67
PACS	Art	97.22	**98.05**	97.75	97.75	97.56
	Cartoon	97.4	**97.65**	97.53	97.53	97.48
	Photo	99.58	99.58	99.59	99.52	**99.72**
	Sketch	86.23	**88.75**	88.52	88.65	87.76
VLCS	Caltech101	99.43	99.43	99.36	99.36	**99.51**
	LabelMe	68.15	68.14	66.92	68.49	**68.94**
	SUN09	73.4	73.4	74.53	74.68	**75.23**
	VOC2007	84.75	84.75	84.0	84.03	**85.6**
TerraInc	L38	20.3	20.3	27.79	29.08	**37.33**
	L43	31.52	31.52	33.98	**34.13**	32.71
	L46	28.98	28.98	27.07	28.13	**31.05**
	L100	52.35	52.35	43.59	42.32	**55.22**

Domain Generalization Task에 대한 실험 성능

→ VLCS, TerraInc 데이터셋에서 우수한 성능을 보임

Conclusions

1) Zero/Few-shot based TTA – Efficient and Context-Aware Label Propagation for Zero-/Few-Shot Training-Free Adaptation of Vision-Language Model

- 추가 학습 없이 테스트 데이터 간 유사도를 활용한 label propagation 기반 pseudo-label 생성
- context-aware edge re-weighting을 통해 노이즈를 완화하고 안정적인 adaptation 수행

2) Test-time Prompt Tuning – R-TPT: Improving adversarial robustness of vision-language models through test-time prompt tuning

- 테스트 시점에서 텍스트 프롬프트를 동적으로 업데이트하여 데이터 분포 변화에 대응
- Augmentation 기반 consistency와 robust optimization을 통해 adversarial 상황에서도 안정적인 성능 확보

3) Encoder Tuning – BATCLIP: Bimodal online test-time adaptation for CLIP

- Visual encoder와 text encoder를 동시에 업데이트하는 bimodal adaptation 구조 적용
- Prototype 기반 alignment와 feature space 재구성을 통해 test-time distribution shift에 효과적으로 대응

Reference

References

[1] Li, Y., Su, Y., Goodge, A., Jia, K., & Xu, X. (2024). Efficient and context-aware label propagation for zero-/few-shot training-free adaptation of vision-language model. In ICLR.

[2] Sheng, L., Liang, J., Wang, Z., & He, R. (2025). R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning. In CVPR.

[3] Maharana, S., Zhang, B., Karlinsky, L., Feris, R., & Guo, Y. (2025). Batclip: Bimodal online test-time adaptation for clip. In ICCV.

[4] Xiao, Z., Yan, S., Hong, J., Cai, J., Jiang, X., Hu, Y., ... & Snoek, C. G. (2025). Dynaprompt: Dynamic test-time prompt tuning. In ICLR.

[5] Zhao, Shuai, et al. "Test-time adaptation with clip reward for zero-shot generalization in vision-language models." In ICLR.

Thank You