NeurIPS 2026

Bridging CLIP with DINO

Cross-Modal Information Maximization for Online Test-Time Adaptation

Taehoon Kim[1], Wonjun Hwang[1]*, Hyunsouk Cho[2]*

[1]Korea University [2]Ajou University *Corresponding authors

Modality bias feeds confirmation bias

Modality bias Under corruption, CLIP projects distorted visual features onto wrong text anchors, with high confidence.

Confirmation bias Entropy minimization sharpens the class already predicted, so a confident error is reinforced instead of corrected.

Label collapse The predicted class distribution concentrates on a small subset of classes.

Confident errors under CLIP + TENT

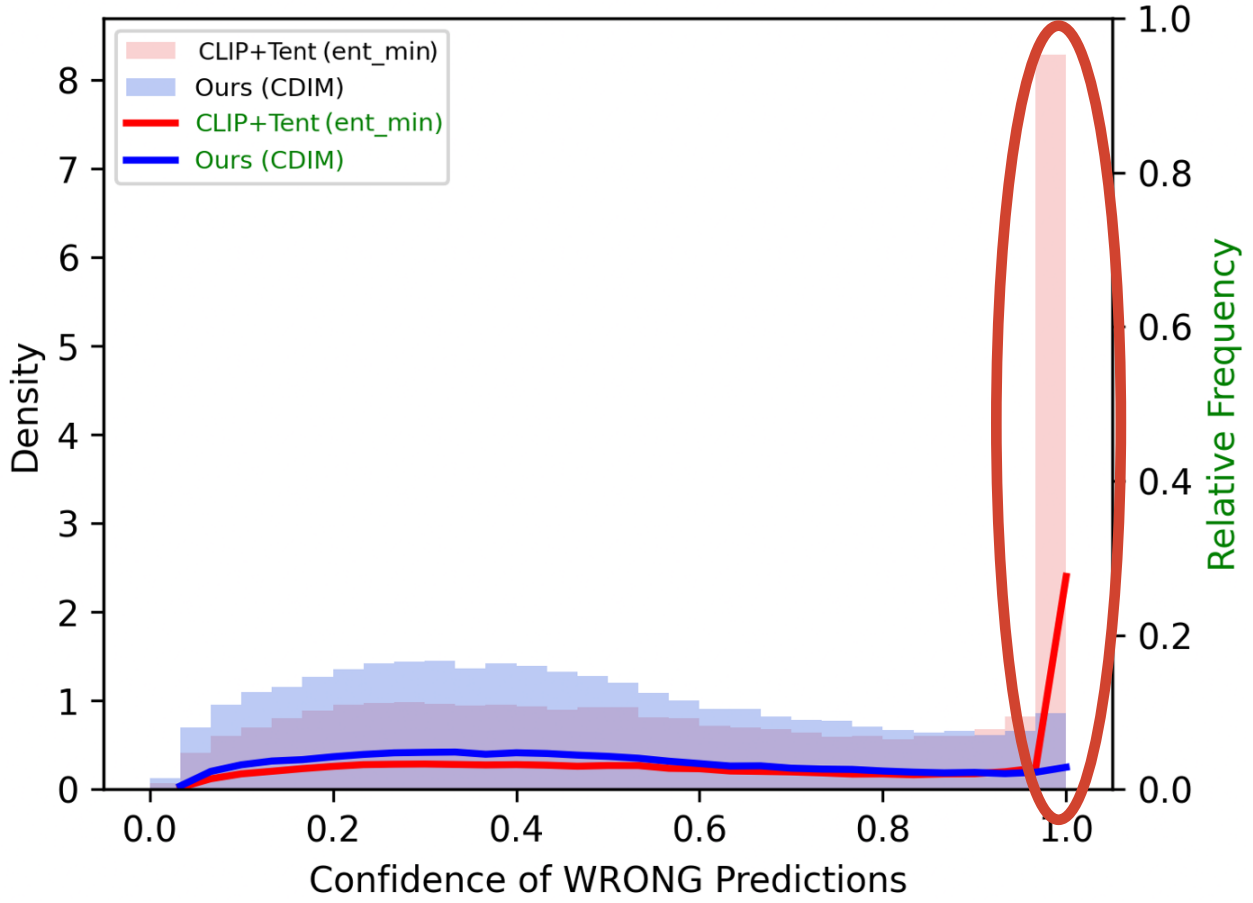


Figure 1: confidence of wrong predictions, ImageNet-C (all corruptions, severity 5)

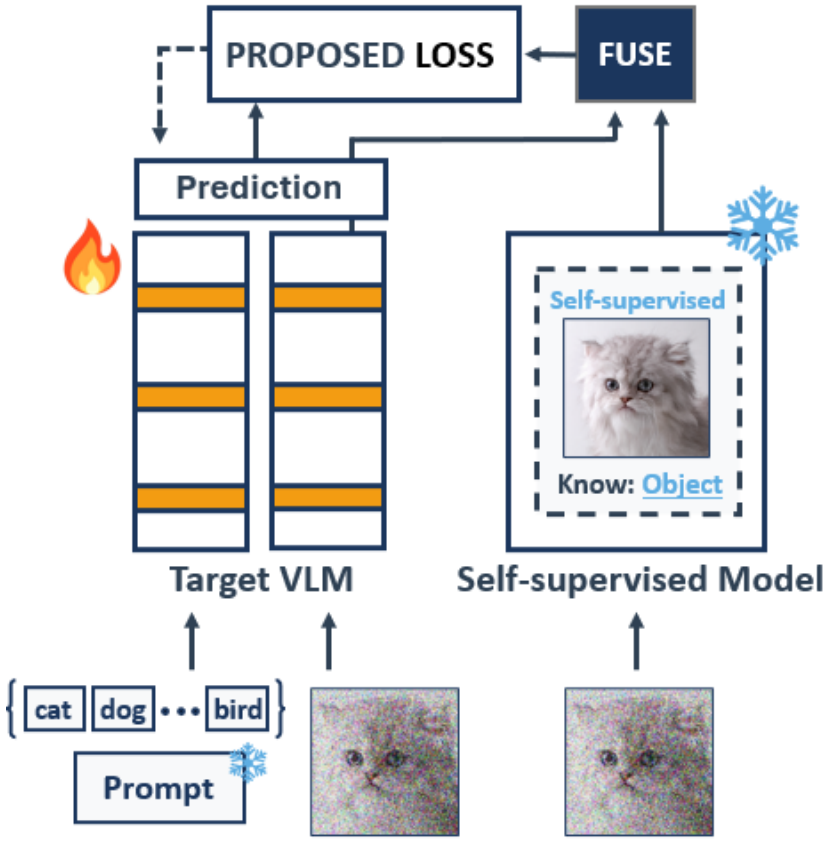
CLIP + TENT

Errors concentrate above 0.95 confidence. Mean accuracy: 25.2%

CDIM

Wrong predictions move to moderate confidence (0.1 to 0.4)

CDIM: break the loop at three levels



1 **Representation**

Cross-modal logit ensemble with a frozen DINOv2

2 **Optimization**

Gated test-time information maximization (GT-IM)

3 **Prediction**

Online logit adjustment (OLA)

**Only normalization-layer affine parameters of CLIP are updated.
**DINOv2 is never updated.

1 Cross-modal logit ensemble

DINO-space text prototypes (Eq. 1)

$$\bar{M} = \frac{1}{B} C^\top D, \quad \tilde{T}^D = \mathrm{norm}(\bar{M}^\top T^\top)$$

Per-backbone logits (Eq. 2)

$$Z^C = \tau_C^{-1} C T^\top, \quad Z^D = \tau_D^{-1} D \tilde{T}^D$$

Ensemble (Eq. 3)

$$Z^{\mathrm{ens}} = \frac{1}{2}(Z^C + Z^D)$$

C, D: L2-normalized CLIP and DINOv2 image features of the batch. **T**: L2-normalized text features of the K class prompts.

Inverse temperatures: 20 for the DINO branch.

Projection: from text to DINO space

CLIP text prototypes T



cross-covariance of the batch

DINO feature space

DINO image features D stay in their native space and are never projected toward text.

ECE (Fig. 2b)

CLIP + TENT

0.400

↓

+ DINO ensemble

0.210

- The ensemble alone lowers the calibration error of CLIP + TENT.
- The construction is parameter-free and batch-local: no learned mapping between the two backbones.

2 Gated information maximization

Uncertainty score (Eq. 4)

$$E_i = -\log \sum_c \exp(z_{ic}^{\text{ens}})$$

Gate: drop the k least confident samples (Eq. 5)

$$G_i = \mathbf{1}\{E_i < E_{(B-k+1)}\}, \quad k = \max(1, \lfloor \rho B \rfloor)$$

GT-IM objective (Eq. 6)

$$\mathcal{L}_{\text{GT-IM}} = \frac{\sum_i G_i \, H(p_i)}{\sum_i G_i} - H(\bar{p})$$

Sharpening uses the retained samples (ρ = 0.05).

Diversity uses the full batch. Both terms are weighted equally.

3 Online logit adjustment

EMA of the model-induced class prior (Eq. 9)

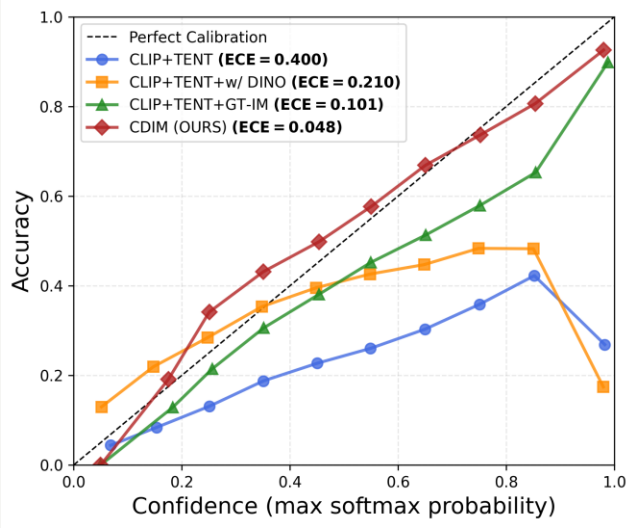
$$\bar{\pi}_c^{(t)} = \alpha\, \bar{\pi}_c^{(t-1)} + (1-\alpha)\, \hat{\pi}_c^{(t)}$$

Adjusted logits and prediction (Eq. 8)

$$z_{ic}^{\text{adj}} = z_{ic}^{\text{ens}} - \log \bar{\pi}_c^{(t)}$$

- Momentum α = 0.9. The prior starts uniform, 1/K for each class.
- No learnable parameters, no gradient, one K-dimensional state vector.
- Assumes that the target class distribution is close to uniform.

Why it works



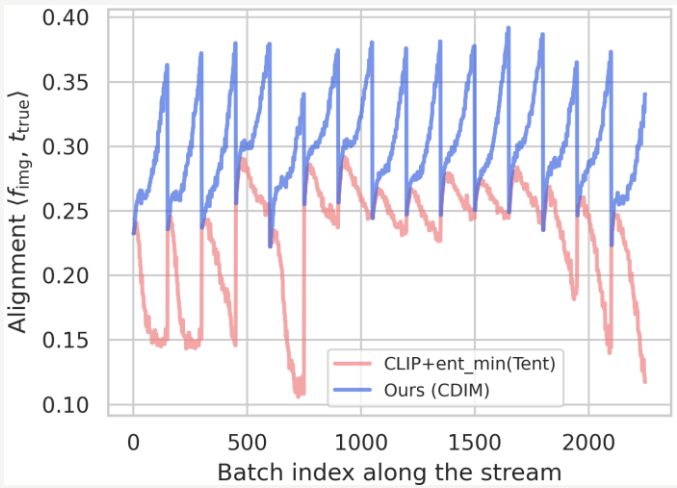
Calibration (Fig. 2b)

ECE: CLIP + TENT 0.400

+ DINO ensemble 0.210

+ GT-IM 0.101

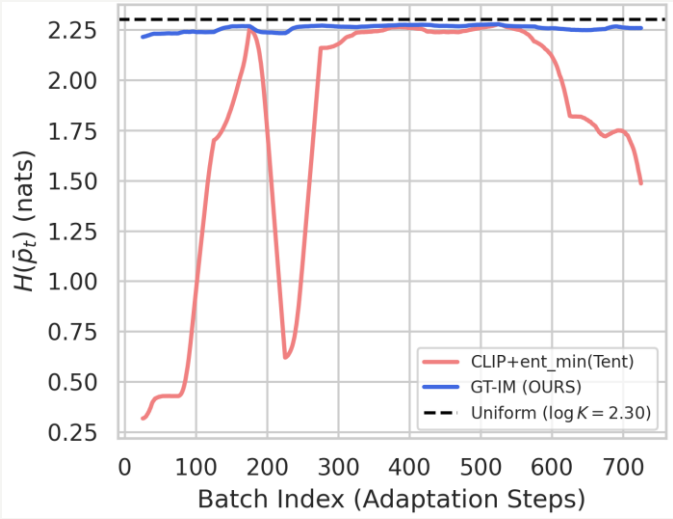
Full CDIM 0.048



Alignment (Fig. 3a)

CDIM stays higher and recovers

after each corruption change



Marginal entropy (Fig. 3c)

CDIM stays near log K = 2.30

CLIP + TENT drops sharply

Figures 3a and 3c: CIFAR10-C, CLIP + TENT vs CDIM. In Figure 2b, adding DINO ensemble or GT-IM alone gives 0.210 and 0.101.

Online TTA results

Method	ImageNet-C	CIFAR10-C	CIFAR100-C
Zero-shot CLIP	24.5	61.2	35.8
TENT	25.2	61.9	37.0
Mint	29.2	70.2	42.4
BATCLIP	30.1	73.8	42.2
SubTTA	32.1	76.3	44.5
CDIM (ours)	**36.9**	**79.8**	**48.2**

+4.8
ImageNet-C

+3.5
CIFAR10-C

+3.7
CIFAR100-C

Accuracy (%), CLIP ViT-B/16, severity 5, mean of 15 corruptions. Gains are over SubTTA, the strongest compared baseline.

Continual and mixed-domain: competitive

Continual (CLIP ViT-B/16)

Method	C10-C	C100-C	IN-C
TENT	61.4	31.7	23.5
EATA	63.5	45.8	34.1
BATCLIP	57.1	27.0	18.4
Mint	67.4	34.5	25.3
CDIM	**80.4**	**46.3**	**34.6**

Mixed-domain (backbones follow Mint)

Method	C10-C	C100-C	IN-C
CLIP	59.0	35.8	39.6
CLIPArTT	56.9	38.7	40.5
WATT-S	63.6	39.0	43.9
Mint	65.9	39.8	45.2
CDIM	**71.2**	**40.1**	**46.9**

Continual: margin over the strongest baseline is +13.0 / +0.5 / +0.5 (C10-C / C100-C / IN-C).

Mixed: margin over Mint is +5.3 / +0.3 / +1.7. CDIM is not designed for continual TTA.

Selected baselines are shown. See Table 2 of the paper for all.

Each component contributes

ENS	GT-IM	OLA	ImageNet-C	CIFAR10-C	CIFAR100-C
			25.2	61.9	37.0
✓			28.7	63.1	40.0
	✓		33.8	76.9	45.2
	✓	✓	35.0	78.3	46.2
✓	✓		36.3	79.2	47.5
✓	✓	✓	**36.9**	**79.8**	**48.2**

The first row is the CLIP + TENT baseline. GT-IM alone gives the largest single gain (+15.0 on CIFAR10-C).

Cost and limitations

Method (CIFAR100-C, V100)	Acc. (%)	Peak mem. (MB)	Latency (ms)
TENT, ViT-B/16	37.0	13146	555
BATCLIP, ViT-B/16	42.2	14505	665
CDIM, ViT-B/16 + DINOv2-base	**48.2**	**16643**	**1347**
CDIM, TinyCLIP + DINOv2-small	**45.5**	**5335**	**1139**

Limitations

The discard ratio ρ is fixed. The marginal term and OLA assume a near-uniform class distribution.

Summary

Problem Modality bias feeds confirmation bias in CLIP-based TTA.

Method One frozen DINOv2 anchor, gated information maximization, and online logit adjustment.

Result +4.8 / +3.5 / +3.7 over SubTTA on ImageNet-C / CIFAR10-C / CIFAR100-C.

Code: **github.com/B1ueC1oud/CDIM**