Question: How to effectively detect noisy samples to mitigate their negative impacts in TTA? 🤔

Noisy Test-Time Adaptation in Vision-Language Models

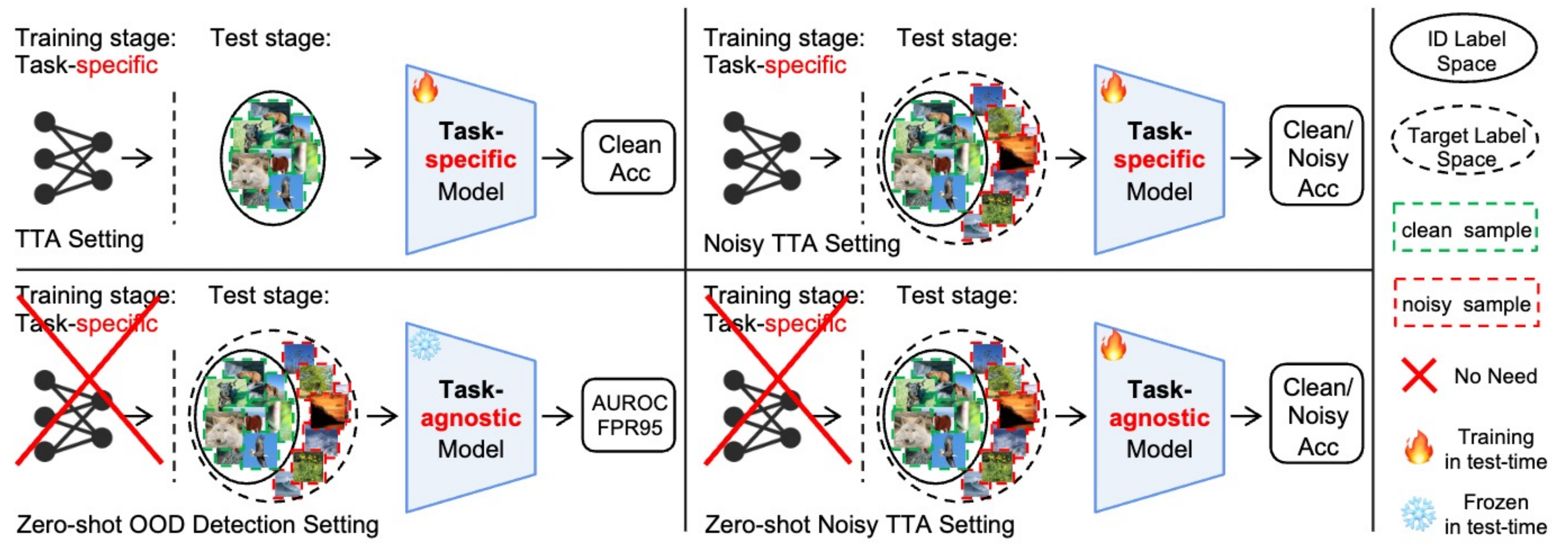
Chentao Cao, Zhun Zhong, Zhanke Zhou, Tongliang Liu, Yang Liu, Kun Zhang, Bo Han

香港浸會大學 HONG KONG BAPTIST UNIVERSITY
THE UNIVERSITY OF SYDNEY
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Paper Code

Problem: Zero-shot Noisy TTA

Comparison between TTA, noisy TTA, zero-shot OOD detection, and the proposed zero-shot noisy TTA.



Test set: $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}_{\text{id}} \cup \mathcal{Y}_{\text{noisy}}\}$

The ID classes are defined based on the classification task of interest rather than the classes used in pre-training.
Noisy samples refer to data that lie outside the ID label space, whereas clean samples stay within it.

Simple baseline (ZS-CLIP):

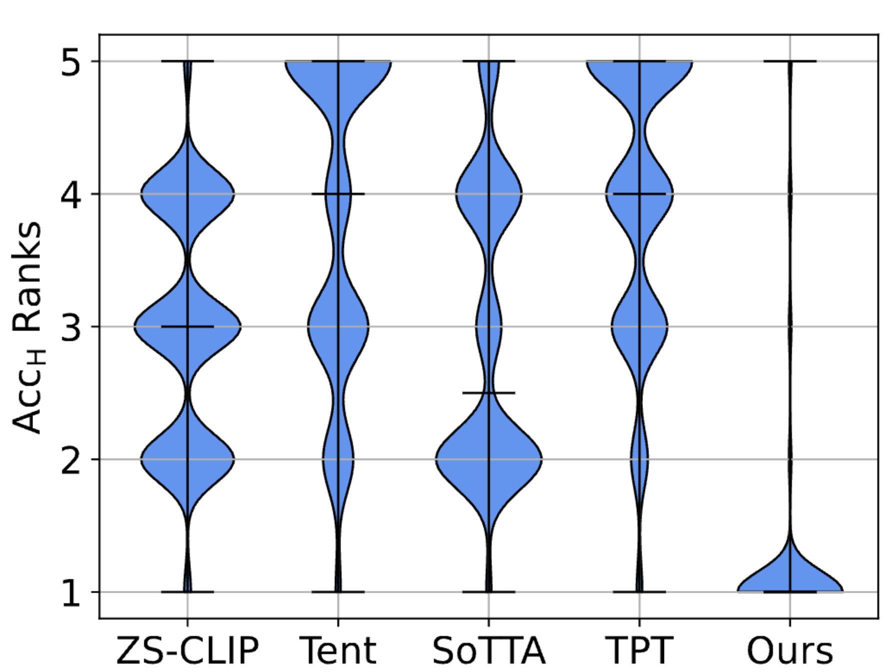
$$G_\lambda(x_i) = \begin{cases} \text{Clean} & S(x_i) \geq \lambda \\ \text{Noise} & S(x_i) < \lambda \end{cases}, \quad \text{where} \quad S(x_i) = \max_k \frac{e^{s_k(x_i)/\tau}}{\sum_{j=1}^K e^{s_j(x_i)/\tau}},$$

$S(\cdot)$ denotes the MCM score and $s_k(x_i)$ is the cosine similarity between the image and text features

How to detect noisy sample online (credit to OWTTT):

$$\min_\lambda \frac{1}{N_{\text{id}}} \sum_i [S(x_i) - \frac{1}{N_{\text{id}}} \sum_j \mathbb{1}(S(x_j) > \lambda) S(x_j)]^2 + \frac{1}{N_{\text{ood}}} \sum_i [S(x_i) - \frac{1}{N_{\text{ood}}} \sum_j \mathbb{1}(S(x_j) \leq \lambda) S(x_j)]^2,$$

Failure Case Study



Performance ranking distribution of five TTA methods across 44 ID-OOD dataset pairs.

Existing TTA methods often underperform the frozen model under ZS-NTTA setting.

Comprehensive Analysis

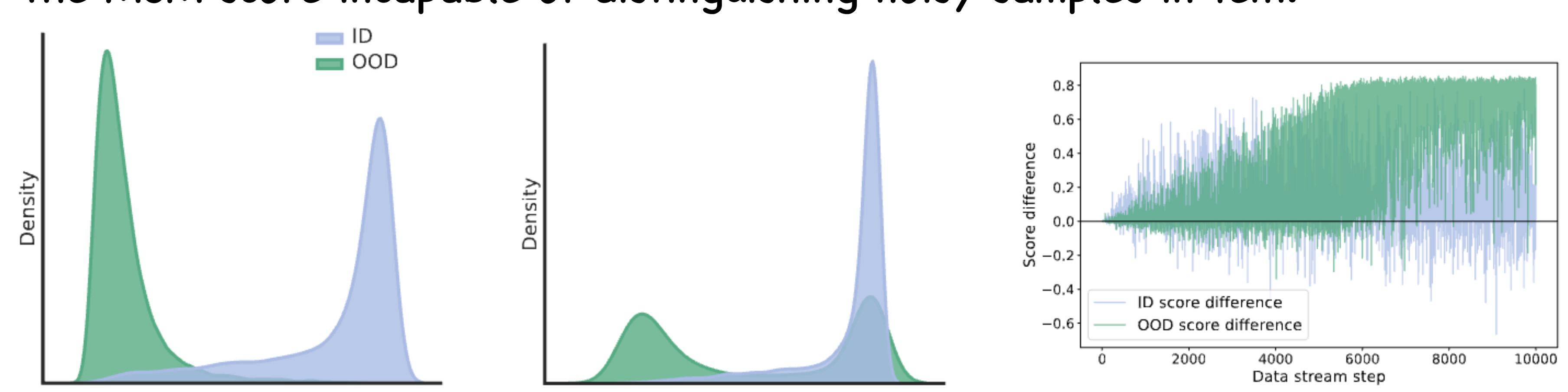
We analyze the failure case, i.e., ZS-CLIP outperforms most tuning-based methods on most ID datasets, highlighting three key observations.

Observation 1. Noisy samples have a significant negative impact on model adaptation during TTA.

Table 1: Failure case study of existing TTA methods with CIFAR-10 as the ID dataset. Green indicate an improvement over ZS-CLIP while red indicates the opposite.

Method	SVHN			LSUN			Texture			Places			Avg		
	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H
ZS-CLIP	83.55	98.39	90.36	83.11	97.82	89.87	82.18	91.82	86.73	81.73	76.26	78.90	82.64	91.07	86.47
Tent (GT)	90.77	96.99	93.78	90.40	93.55	91.95	90.07	90.22	90.14	89.87	74.50	81.47	90.28	88.81	89.34 (+2.87%)
Tent (Normal)	87.18	52.90	65.85	89.03	73.96	80.80	89.78	88.48	89.13	88.78	65.44	75.34	88.69	70.19	77.78 (-8.69%)
Tent (All-update)	81.74	43.13	56.47	80.17	55.59	65.65	89.28	84.64	86.90	87.86	56.27	68.60	84.76	59.91	69.41 (-17.06%)
SoTTA (GT)	90.45	97.47	93.83	90.03	94.88	92.39	89.68	91.39	90.53	89.30	75.96	82.09	89.87	89.92	89.71 (+3.25%)
SoTTA (Normal)	90.21	81.71	85.75	90.13	91.06	90.59	89.56	90.96	90.25	89.04	74.17	80.93	89.73	84.47	86.88 (+0.42%)
SoTTA (All-update)	89.69	73.13	80.57	89.88	90.76	90.32	89.47	90.54	90.00	89.05	74.50	81.13	89.52	82.23	85.50 (-0.96%)
TPT (GT)	85.86	98.46	91.73	85.86	98.00	91.53	85.19	92.30	88.60	84.88	77.33	80.93	85.45	91.52	88.20 (+1.73%)
TPT (Normal)	81.76	98.85	89.50	81.53	97.93	88.98	80.43	92.11	85.87	79.88	77.18	78.51	80.90	91.52	85.72 (-0.75%)
TPT (All-update)	85.18	96.98	90.70	84.84	91.15	87.88	83.92	75.36	79.41	83.59	54.11	65.69	84.38	79.40	80.92 (-5.55%)

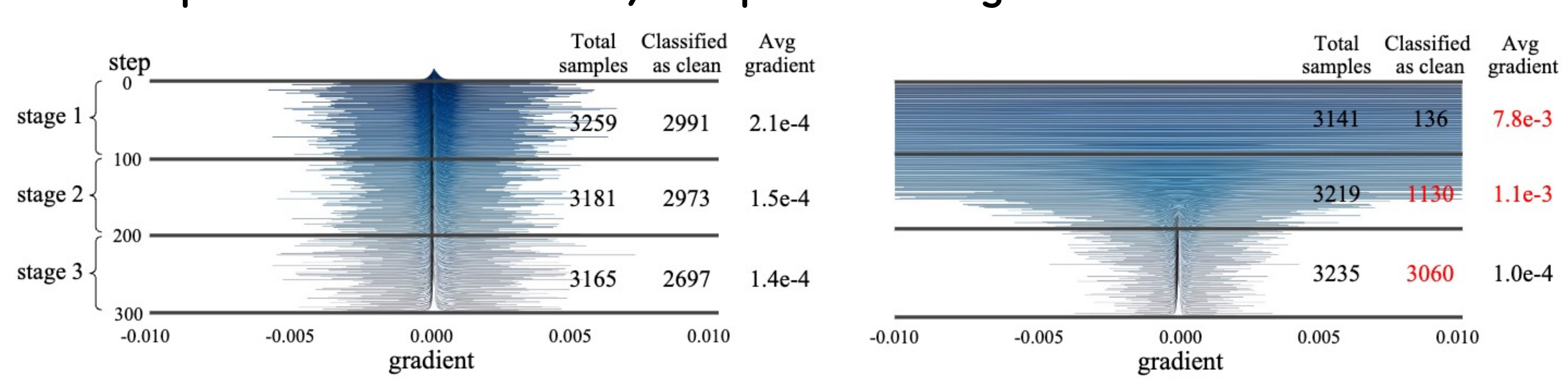
Observation 2. Noisy samples' score gradually increase, ultimately rendering the MCM score incapable of distinguishing noisy samples in Tent.



(a) ZS-CLIP (b) Tent (c) Score difference

Observation 3. Few inaccuracies during the early TTA stages can gradually lead the model to overfit to noisy samples.

The impact of clean and noisy samples on the gradients:

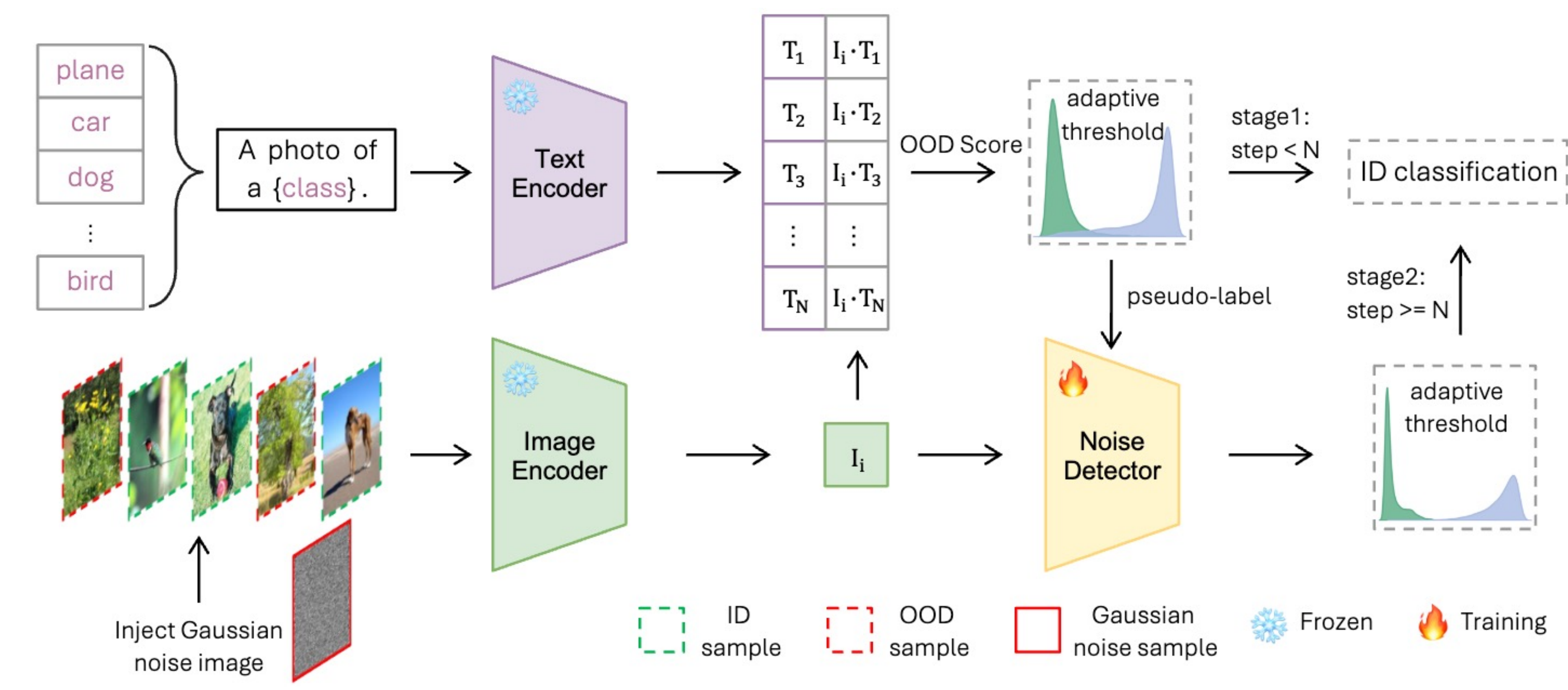


(a) Clean sample (b) Noisy sample

Motivation: We naturally consider whether decoupling the classifier and detector might be a superior strategy for the ZS-NTTA task.

Method: Adaptive Noise Detector

We use the detection results from ZS-CLIP as pseudo-labels to train the Adaptive Noise Detector.
To further handle the clean data stream case, we intentionally inject Gaussian noise as additional noisy samples to avoid wrongly assigning too many clean samples as noisy ones.



Experiments

On ImageNet, AdaND enhances the average performance by 8.32% in terms of ACC_H for ZS-NTTA.

Table 2: Zero-shot noisy TTA results for ImageNet as the ID dataset.

ID	Method	iNaturalist			SUN			Texture			Places			Avg		
		Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H	Acc_S	Acc_N	Acc_H
ImageNet	ZS-CLIP	54.01	86.53	66.51	53.43	83.96	65.30	52.71	78.52	63.08	53.35	80.50	64.17	53.38	82.38	64.77
	Tent	48.56	35.74	41.18	55.44	75.54	63.95	54.94	70.93	61.92	55.76	73.98	63.59	53.67	64.05	57.66
	SoTTA	53.15	62.68	57.52	53.16	68.76	59.96	53.64	68.05	59.99	53.60	69.16	60.39	53.39	67.16	59.47
	TPT	52.58	88.93	66.09	51.91	86.09	64.77	51.11	80.01	62.38	51.80	82.89	63.76	51.85	84.48	64.25
	AdaND (Ours)	**63.26**	**96.87**	**76.54**	**61.34**	**89.44**	**72.77**	**62.45**	**83.54**	**71.47**	**61.92**	**84.82**	**71.58**	**62.24**	**88.67**	**73.09**

AdaND is computationally efficient and comparable to ZS-CLIP.

Table 3: Runtime and GPU memory with varying batch sizes on ImageNet for a sample.

Resource	ZS-CLIP ($bs = 1$)	SoTTA ($bs = 1$)	TPT ($bs = 1$)	Ours ($bs = 1$)	ZS-CLIP ($bs = 128$)	Tent ($bs = 128$)	Ours ($bs = 128$)
Time (s)↓	0.1125	0.1193	0.3219	0.1272	0.0015	0.0037	0.0017
Memory (GiB)↓	3.80	9.13	21.23	3.83	4.54	14.99	4.57

On ImageNet, AdaND enhances the average performance by 9.40% in terms of FPR95 for zero-shot OOD detection.

Table 4: Zero-shot OOD detection results for ImageNet as the ID dataset.

Method	iNaturalist		SUN		Texture		Places		Avg	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Max-Logit	89.31	61.66	87.43	64.39	71.68	86.61	85.95	63.67	83.59	69.08
Energy	85.09	81.08	84.24	79.02	65.56	93.65	83.38	75.08	79.57	82.21
MCM	94.61	30.91	92.57	37.59	86.11	57.77	89.77	44.69	90.77	42.74
CLIPN	95.27	23.94	93.93	26.17	90.93	40.83	92.28	33.45	93.10	31.10
NegLabel	**99.49**	**1.91**	95.49	20.53	90.22	43.56	91.64	35.59	94.21	25.40
AdaND (Ours)	98.91	4.19	**95.86**	**17.08**	**93.01**	**21.76**	**94.55**	**20.95**	**95.58**	**16.00**