Bilateral Information-aware Test-time Adaptation for Vision-Language Models

Jingwei Sun
Hong Kong Baptist University

with Jianing Zhu, Jiangchao Yao, Gang Niu, Masashi Sugiyama, Bo Han

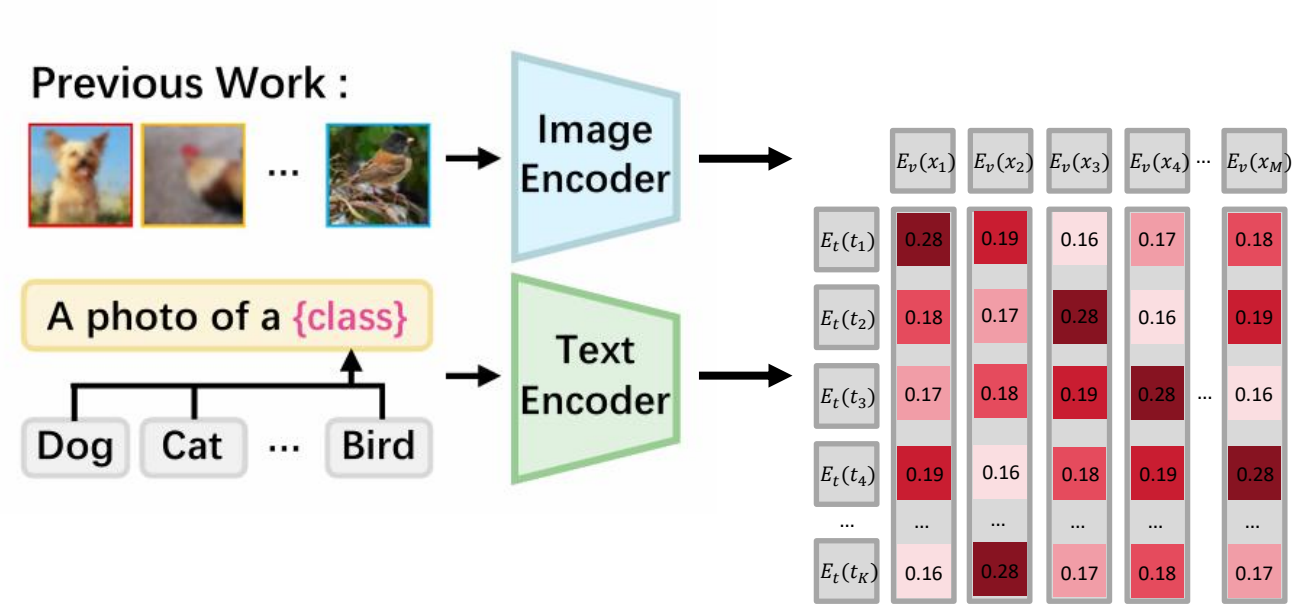
ICLR 2026 Paper: https://openreview.net/forum?id=vv8EcCoBfr *Code: https://github.com/tmlr-group/BITTA*

Outline

- Background
- Method
- Experiment
- Summary

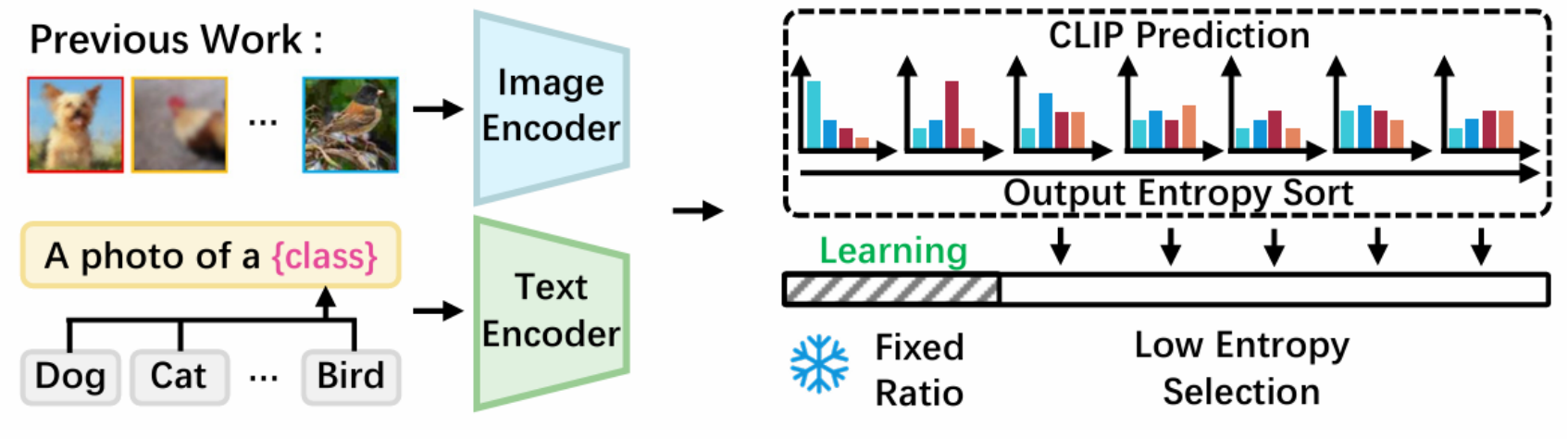
Background | Zero-shot prediction

Pre-trained VLMs (e.g., CLIP) enable Zero-shot prediction by computing the similarity between the extracted image feature and textual representation.



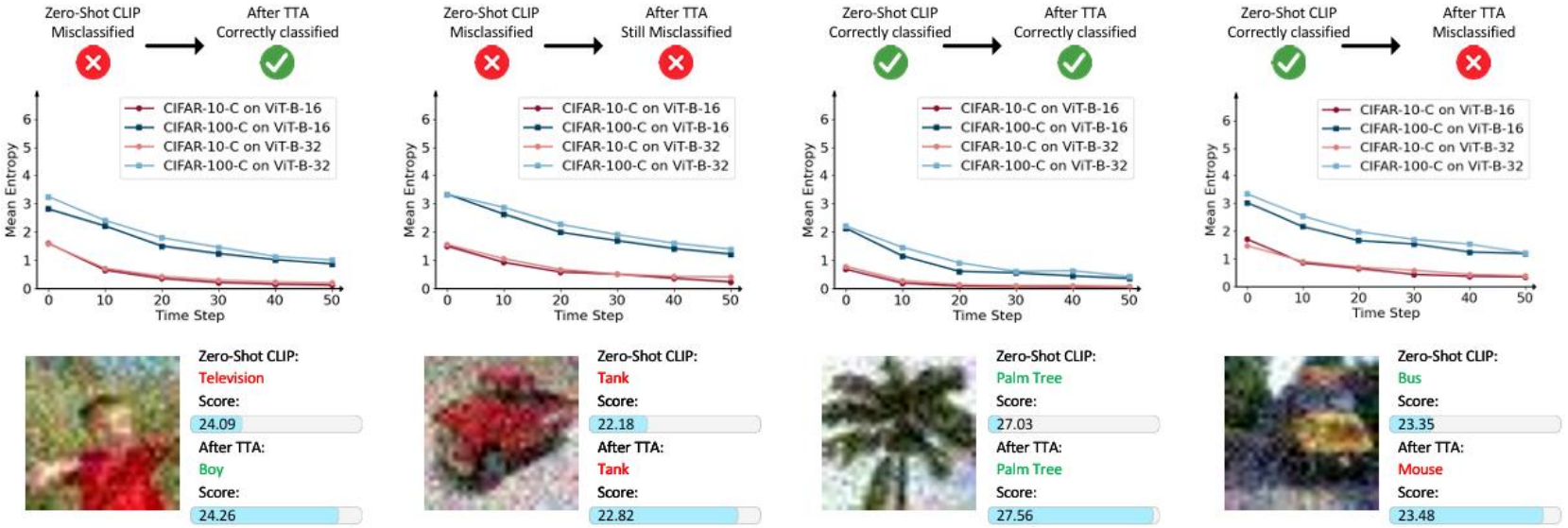
Background | Test-time adaptation

Test-time adaptation (TTA) allows models to be fine-tuned based on new data encountered during inference.



Method | Motivation

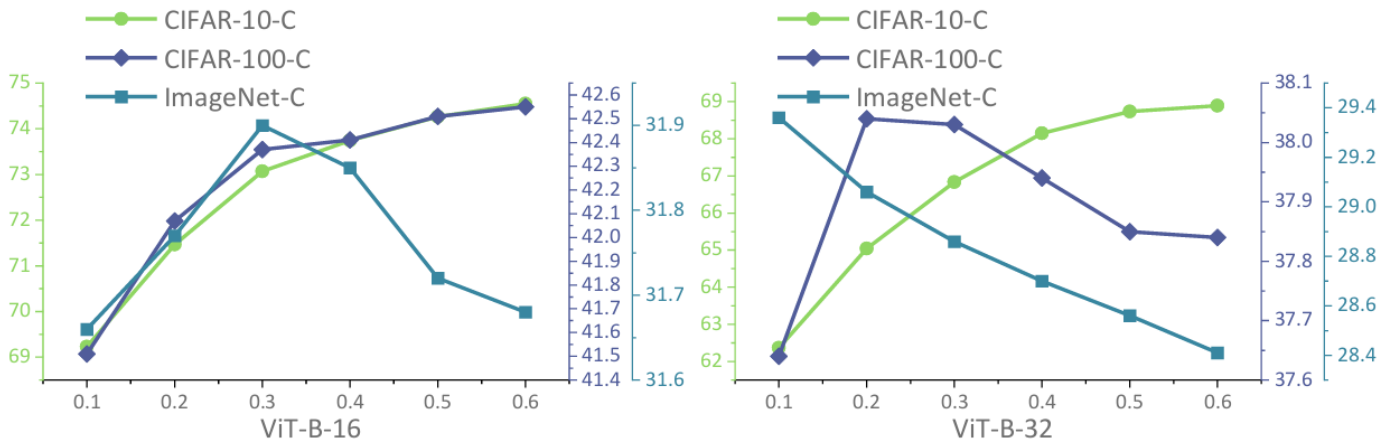
Does the fixed low-entropy selection beneficially affect adaptation on all data?



Overfitting on atypical feature: Only learning on confident samples can induce unexpected overfitting (as decreasing entropy values) on misclassified samples.

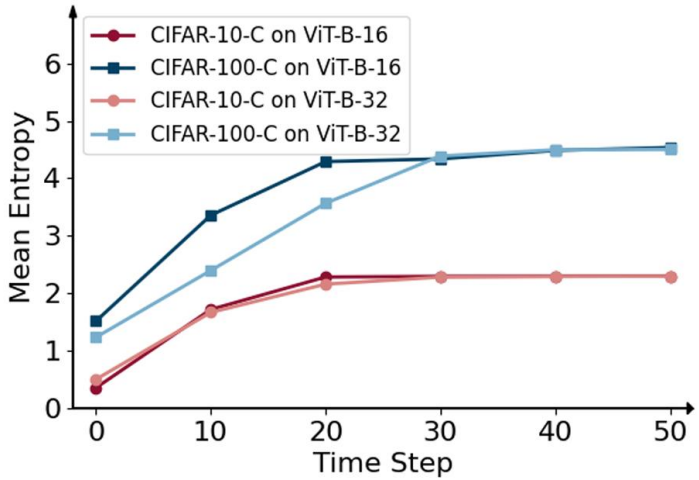
Method | Motivation

Does the fixed low-entropy selection beneficially affect adaptation on all data?



Incompatibility of fixed selection ratio: Optimal performance is achieved at different low-entropy selection ratios for each dataset.

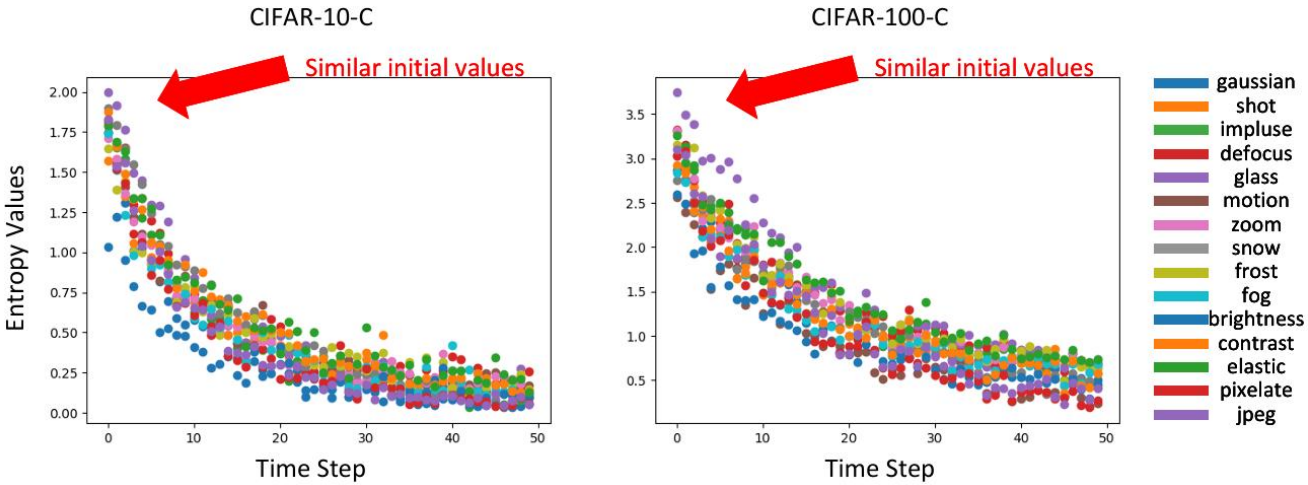
Method | Regularization



Atypical feature: Both the indiscriminability of high-entropy samples and the misclassification of some low-entropy samples stem from noise information

Entropy change during TTA: The increase in entropy of high-entropy samples also triggers the increase in entropy of low-entropy misclassified samples.

Method | Standardization

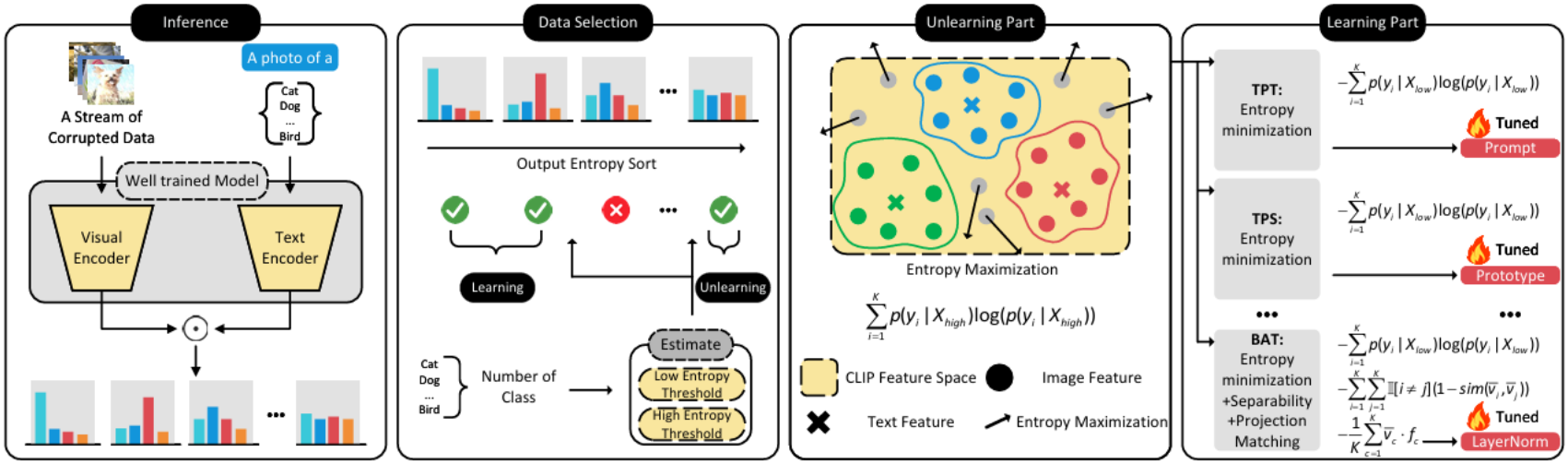


Within the same dataset, samples exhibit similar entropy across these different optimal ratios.

A linear relationship between the initial entropy $\boldsymbol{\tau_l^n}$ and the number of categories K:

$$\frac{\boldsymbol{\tau_l^n}}{\boldsymbol{Max(\mathcal{H}(x_i))}} = \boldsymbol{-0.00038K + 0.83}$$

Method | Framework



Appropriate sample selection ratio is estimated according to the output entropy and number of categories. High-entropy samples will be used for unlearning to avoid overfitting. Low-entropy samples will be used for learning core representations.

Optimization Object: $\mathcal{L}_{\text{learning}} + \lambda \mathcal{L}_{\text{unlearning}}$

Experiment | Main Results

Table 1: **Comparison of TTA mean accuracy (%) across different corruption types, using two different CLIP models**. Δ highlighted the improvement in green over the top-performing baseline.

	Method	Gaussian	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Brightness	Contrast	Elastic	Pixelate	JPEG	Mean
CIFAR-10-C	ViT-B-16	37.99	41.73	54.40	71.70	40.91	67.92	73.61	73.89	77.41	70.26	84.47	62.29	53.82	47.60	59.39	61.16
	TPT	39.91	44.91	58.76	72.25	43.53	70.03	74.77	75.48	78.51	72.71	85.06	70.84	57.23	52.29	61.27	63.84
	TDA	42.94	46.04	63.27	72.27	46.88	69.76	74.65	77.46	79.35	71.57	85.86	62.33	60.48	63.72	62.94	65.30
	BCA	31.51	34.53	54.17	66.50	35.64	64.93	70.59	73.90	76.27	68.45	84.02	55.42	51.52	55.80	55.95	58.61
	DMN-ZS	44.97	49.50	59.77	73.60	48.75	70.85	74.80	77.05	77.45	70.05	86.20	67.55	62.75	55.30	62.05	65.38
	DPE	34.78	38.66	57.81	73.14	44.30	68.59	75.89	74.26	78.89	71.17	86.90	64.43	57.49	52.37	58.54	62.48
	BAT	60.47	65.48	63.40	80.09	52.34	80.40	82.00	83.05	83.55	81.25	89.62	82.39	67.54	60.52	68.06	73.34
	BAT+BITTA	**62.59**	**67.42**	**64.97**	**80.63**	**56.25**	**80.74**	**82.29**	**83.52**	**84.29**	**81.92**	**89.90**	**83.25**	**69.13**	**65.50**	**69.26**	**74.78**
	Δ	**+2.12**	**+1.94**	**+1.57**	**+0.54**	**+3.91**	**+0.34**	**+0.29**	**+0.47**	**+0.74**	**+0.67**	**+0.28**	**+0.86**	**+1.59**	**+1.78**	**+1.20**	**+1.44**
	ViT-B-32	35.58	40.07	43.16	69.98	41.50	64.51	70.19	70.80	72.34	66.66	81.38	64.51	59.66	48.16	56.58	59.01
	TPT	43.02	57.23	46.77	71.24	46.46	68.01	72.67	73.68	75.75	68.94	83.86	73.51	62.53	50.39	57.80	63.46
	TDA	47.40	51.38	52.15	72.68	54.42	69.19	75.08	76.77	78.01	72.27	84.70	63.33	66.55	56.54	60.55	65.40
	BCA	47.88	51.13	51.54	71.65	52.97	68.64	73.75	76.14	77.55	70.76	84.43	63.07	65.76	56.86	60.39	64.83
	DMN-ZS	45.90	48.60	43.85	70.10	48.80	68.00	72.20	72.10	73.05	67.95	81.40	67.70	63.10	53.20	59.05	62.33
	DPE	38.83	41.34	43.76	72.49	42.00	65.56	73.05	73.99	75.19	67.46	83.45	65.56	62.58	54.25	57.80	61.15
	BAT	49.40	54.79	**52.40**	76.16	53.08	75.06	76.29	76.81	78.32	74.83	86.33	78.83	66.58	53.06	60.39	67.49
	BAT+BITTA	**52.34**	**57.61**	52.28	**76.82**	**55.99**	**75.79**	**76.83**	**77.64**	**78.68**	**76.79**	**86.50**	**80.41**	**67.44**	**58.45**	**61.69**	**69.02**
	Δ	**+2.94**	**+0.38**	**-0.12**	**+0.67**	**+1.57**	**+0.73**	**+0.54**	**+0.83**	**+0.36**	**+1.96**	**+0.17**	**+1.58**	**+0.86**	**+1.59**	**+1.14**	**+1.53**

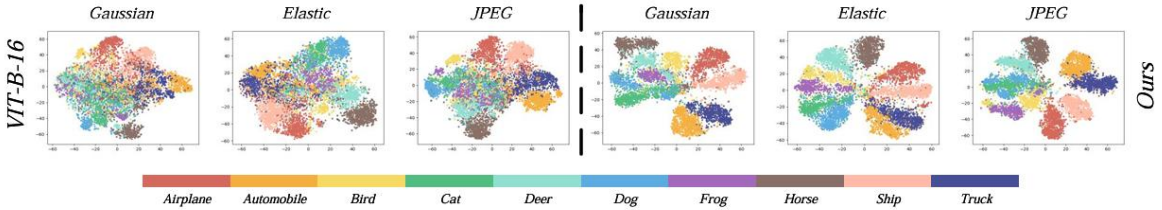


Figure 5: **t-SNE plots of image features from BITTA and zero-shot ViT-B-16.** The results indicate that our method produces more discriminative features that are more tightly clustered.

Experiment | Main Results

Table 2: **Comparison of BITTA combined with different TTA methods over CIFAR-10-C using ViT-B-16**. Δ highlighted the improvement in green over the original method without BITTA.

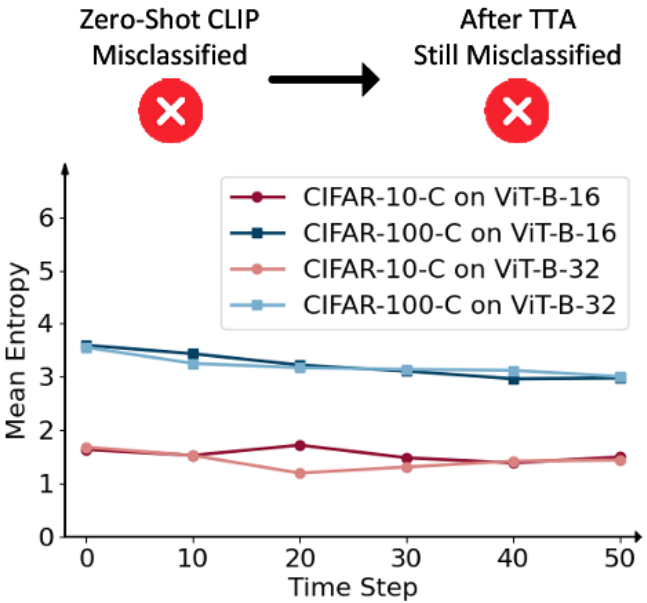
Method	Gaussian	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Brightness	Contrast	Elastic	Pixelate	JPEG	Mean
TPT	39.91	44.91	58.76	72.25	43.53	70.03	74.77	75.48	78.51	72.71	85.06	70.84	57.23	52.29	61.27	63.84
TPT + BITTA	**42.37**	**47.13**	**60.89**	**73.62**	**45.92**	**71.67**	**75.72**	**76.45**	**79.70**	**73.71**	**86.25**	**71.03**	**59.10**	**54.94**	**62.06**	**65.37**
Δ	**+2.46**	**+2.22**	**+2.13**	**+1.37**	**+2.39**	**+1.64**	**+0.95**	**+0.97**	**+1.19**	**+1.00**	**+1.19**	**+0.19**	**+1.87**	**+2.65**	**+0.79**	**+1.53**
DiffTPT	39.25	45.65	59.40	71.80	43.75	71.10	75.50	75.95	78.85	73.80	84.90	**72.55**	57.20	51.05	62.85	64.24
DiffTPT + BITTA	**41.91**	**48.50**	**62.40**	**73.00**	**46.35**	**73.05**	**76.75**	**76.85**	**80.10**	**74.90**	**85.95**	71.90	**60.15**	**54.85**	**63.75**	**66.03**
Δ	**+2.66**	**+2.85**	**+3.00**	**+1.20**	**+2.60**	**+1.95**	**+1.25**	**+0.90**	**+1.25**	**+1.10**	**+1.05**	**-0.65**	**+2.95**	**+3.80**	**+0.90**	**+1.79**
CTPT	38.18	42.74	56.73	71.66	41.58	68.91	74.18	74.41	77.05	70.6	84.55	63.95	54.98	49.33	58.79	61.84
CTPT + BITTA	**39.40**	**44.10**	**58.31**	**72.16**	**43.04**	**69.46**	**74.67**	**74.77**	**78.05**	**71.08**	**85.15**	**64.05**	**55.66**	**50.15**	**59.12**	**62.61**
Δ	**+1.22**	**+1.36**	**+1.58**	**+0.50**	**+1.46**	**+0.55**	**+0.49**	**+0.36**	**+1.00**	**+0.48**	**+0.60**	**+0.10**	**+0.68**	**+0.82**	**+0.33**	**+0.77**
TPS	40.11	44.48	59.73	71.72	43.05	70.25	74.32	74.69	77.67	**73.36**	83.78	**77.30**	58.35	52.85	**62.59**	64.28
TPS + BITTA	**42.62**	**46.92**	**61.32**	**72.27**	**43.87**	**71.32**	**74.82**	**74.97**	**77.82**	72.87	**84.39**	77.20	**59.82**	**57.26**	61.06	**65.24**
Δ	**+2.51**	**+2.44**	**+1.59**	**+0.55**	**+0.82**	**+1.07**	**+0.50**	**+0.28**	**+0.15**	**-0.49**	**+0.61**	**-0.10**	**+1.47**	**+4.41**	**-1.53**	**+0.96**

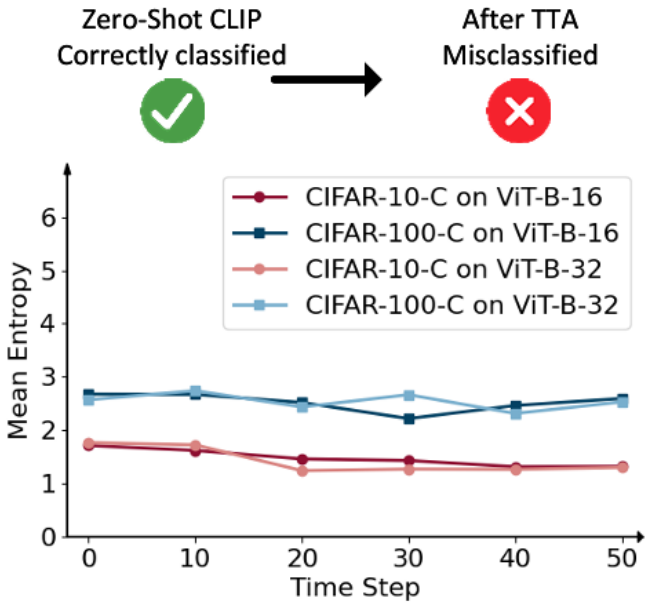
Experiment | More Analysis

Table 7: **The comparison of expected calibration error (ECE ↓).** Δ highlighted the decline ratio in green over the original method without BITTA.

	Method	Gaussian	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Brightness	Contrast	Elastic	Pixelate	JPEG	Mean
CIFAR-10-C	BAT (ViT-B-16)	24.04	20.11	24.17	10.99	30.22	9.86	9.43	8.60	8.15	9.42	5.11	7.96	17.59	24.35	18.71	15.25
	BAT+BITTA	**20.73**	**17.59**	**20.55**	**9.67**	**24.48**	**8.71**	**8.84**	**7.13**	**6.85**	**8.64**	**3.94**	**6.52**	**15.43**	**18.30**	**16.54**	**12.93**
	Δ	**13.77%**	**12.53%**	**14.98%**	**12.01%**	**18.99%**	**11.66%**	**6.26%**	**17.09%**	**15.95%**	**8.28%**	**22.89%**	**18.09%**	**12.28%**	**24.85%**	**11.59%**	**15.21%**
	BAT (ViT-B-32)	50.04	34.31	31.14	12.93	28.51	13.83	12.78	11.58	10.48	12.66	6.16	9.50	19.03	25.45	23.41	20.12
	BAT+BITTA	**23.47**	**21.51**	**26.90**	**11.20**	**23.90**	**11.58**	**10.57**	**9.53**	**8.66**	**10.52**	**4.79**	**7.66**	**16.09**	**21.02**	**19.38**	**15.12**
	Δ	**53.09%**	**37.31%**	**13.62%**	**13.38%**	**16.17%**	**16.27%**	**17.29%**	**17.70%**	**17.37%**	**16.90%**	**22.24%**	**19.37%**	**15.45%**	**17.41%**	**17.21%**	**24.86%**
CIFAR-100-C	BAT (ViT-B-16)	50.81	47.82	32.37	13.31	25.07	13.78	12.06	12.97	15.35	17.15	8.85	16.93	21.31	20.30	37.43	23.03
	BAT+BITTA	**21.18**	**20.53**	**13.59**	**11.36**	**21.56**	**11.54**	**9.75**	**10.13**	**10.93**	**13.33**	**6.79**	**12.15**	**18.34**	**17.43**	**15.49**	**14.27**
	Δ	**58.32%**	**57.07%**	**58.02%**	**14.65%**	**14.00%**	**16.26%**	**19.15%**	**21.89%**	**28.79%**	**22.27%**	**23.28%**	**28.23%**	**13.94%**	**14.14%**	**58.62%**	**38.03%**
	BAT (ViT-B-32)	40.08	24.86	22.12	13.89	23.76	15.50	12.45	13.12	15.05	13.09	9.17	16.97	20.21	20.82	18.52	18.64
	BAT+BITTA	**18.42**	**17.56**	**14.67**	**10.41**	**20.05**	**11.76**	**9.72**	**10.28**	**10.74**	**10.18**	**6.27**	**11.48**	**16.51**	**16.12**	**15.11**	**13.28**
	Δ	**54.04%**	**29.36%**	**33.68%**	**25.05%**	**15.61%**	**24.13%**	**21.93%**	**21.65%**	**28.64%**	**22.23%**	**31.62%**	**32.35%**	**18.31%**	**22.57%**	**18.41%**	**28.73%**
ImageNet-C	BAT (ViT-B-16)	27.58	27.27	27.62	24.10	28.36	29.78	29.11	19.08	24.59	18.41	12.86	22.06	28.03	19.83	22.41	24.07
	BAT+BITTA	**20.66**	**20.82**	**19.28**	**16.85**	**19.67**	**16.95**	**18.84**	**16.21**	**18.29**	**14.61**	**9.15**	**14.66**	**22.23**	**15.48**	**18.47**	**17.47**
	Δ	**25.09%**	**23.65%**	**30.20%**	**30.08%**	**30.64%**	**43.08%**	**35.28%**	**15.04%**	**25.62%**	**20.64%**	**28.85%**	**33.54%**	**20.69%**	**21.94%**	**17.58%**	**27.39%**
	BAT (ViT-B-32)	26.99	25.19	26.00	21.62	23.81	19.96	24.63	20.22	24.02	19.56	15.33	27.79	23.05	18.08	18.49	22.32
	BAT+BITTA	**17.94**	**17.50**	**17.41**	**15.42**	**17.17**	**14.76**	**19.57**	**15.24**	**18.37**	**14.48**	**10.16**	**16.41**	**18.26**	**14.48**	**14.97**	**16.14**
	Δ	**33.53%**	**30.53%**	**33.04%**	**28.68%**	**27.89%**	**26.05%**	**20.54%**	**24.63%**	**23.52%**	**25.97%**	**33.72%**	**40.95%**	**20.78%**	**19.91%**	**19.04%**	**27.66%**

Experiment | More Analysis





Summary

- ✓ Conceptually, we reveal two critical phenomenons in TTA with VLMs from new perspective, i.e., the fixed previous low-entropy selection criteria causes unexpected overfitting on atypical features, and cannot adapt to different data distribution.
- ✓ Empirically, we recognize high-entropy samples can function as auxiliary set to effectively offset the atypical features memorized during the adaptation process and the entropy values can serve as standardized signals to indicate optimal selection ratio.
- ✓ Technically, we propose BITTA, which innovatively selects bilateral information with dynamic ratio and simultaneously implement learning on low-entropy samples to fit core representations and unlearning on high-entropy samples to avoid overfitting.
- ✓ Experimentally, we conduct extensive explorations to verify the effectiveness of BITTA under different scenarios, including the significant improvement on various datasets, compatibility with diverse methods and model architectures, etc.