PROTOTTA: PROTOTYPE-GUIDED TEST-TIME ADAPTATION

Mohammad Mahdi Abootorabi[1,3], Parvin Mousavi[2,3], Purang Abolmaesumi[1], Evan Shelhamer[1,3]

[1] University of British Columbia [2] Queen's University [3] Vector Institute

ABSTRACT

Deep networks that rely on prototypes—interpretable representations that can be related to the model input—have gained significant attention for balancing high accuracy with inherent interpretability, which makes them suitable for critical domains such as healthcare. However, these models are limited by their reliance on training data, which hampers their robustness to distribution shifts. While test-time adaptation (TTA) improves the robustness of deep networks by updating parameters and statistics, the prototypes of interpretable models have not been explored for this purpose. We introduce ***ProtoTTA***, a general framework for prototypical models that leverages intermediate prototype signals rather than relying solely on model outputs. *ProtoTTA* minimizes the entropy of the prototype-similarity distribution to encourage more confident and prototype-specific activations on shifted data. To maintain stability, we employ geometric filtering to restrict updates to samples with reliable prototype activations, regularized by prototype-importance weights and model-confidence scores. Experiments across four prototypical backbones on four diverse benchmarks spanning fine-grained vision, histopathology, and NLP demonstrate that ProtoTTA improves robustness over standard output entropy minimization while restoring correct semantic focus in prototype activations. We also introduce novel interpretability metrics and a vision-language model (VLM) evaluation framework to explain TTA dynamics, confirming ProtoTTA restores human-aligned semantic focus and correlates reliably with VLM-rated reasoning quality. Code is available at: `https://github.com/DeepRCL/ProtoTTA`.

1 INTRODUCTION AND RELATED WORK

Prototype-based neural networks have emerged as a compelling solution for deep learning in critical domains, balancing high accuracy with inherent interpretability. Unlike standard black-box models, these architectures provide internal explainability by classifying inputs through similarity matching with learned prototypes, enabling intuitive "this looks like that" reasoning with direct visual evidence for predictions (Chen et al., 2019). This transparency has driven growing adoption in high-stakes applications, particularly healthcare (Wei et al., 2024; Vaseli et al., 2023; Sethi et al., 2025), where understanding model decisions is as important as their accuracy. The foundational model, ProtoPNet (Chen et al., 2019), introduced the concept of classifying images by comparing input feature patches to learned prototypes. However, ProtoPNet relies on spatially rigid prototypes, limiting its ability to capture geometric variations. To address this, Deformable ProtoPNet (Donnelly et al., 2022) replaced single rigid prototypes with flexible sub-prototypes that adapt their positions to better match input features. Subsequently, ProtoViT (Ma et al., 2024) extended this paradigm to Vision Transformer (ViT) (Dosovitskiy, 2020) backbones with coherence-aligned sub-prototypes for capturing complex geometric variations. The field continues to evolve with innovations such as Hyperbolic Hierarchical Part Prototypes (Vaseli et al., 2025), demonstrating sustained community interest in advancing prototype-based architectures.

Despite these architectural advances, prototype-based models remain vulnerable to distribution shifts, where the discrepancy between training and test distributions degrades the semantic validity of selected prototypes. To address distribution shifts in general deep learning, Test-Time Adaptation (TTA) (Liang et al., 2024) has emerged as a paradigm to dynamically adjust models at inference using unlabeled target-domain data. Foundational methods like Tent (Wang et al., 2020) update normalization parameters via entropy minimization, while subsequent approaches have improved stability and efficiency through active sample selection (EATA) (Niu et al., 2022), sharpness-aware

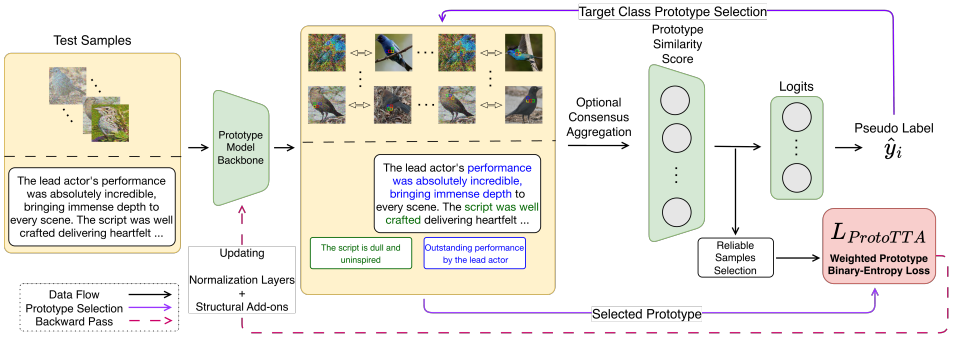

Figure 1: Overview of the *ProtoTTA* pipeline. The loss minimizes the binary entropy of prototype similarities for decisive activations. Geometric Filtering masks uncertain samples while Consensus Aggregation refines scores. Finally, updates to normalizations and structural add-ons (e.g., attention biases, 1×1 convolutions) restore semantic focus under domain shifts.

minimization (SAR) (Niu et al., 2023), or enforcing consistency across augmented views (MEMO) (Zhang et al., 2022). Beyond standard discriminative tasks, TTA has expanded to generative models (Prabhudesai et al., 2023) and even LLM agents that adapt to unseen environments by adjusting to local observation formats (Chen et al., 2025). However, the intersection of TTA and model interpretability remains largely unexplored. Existing TTA methods treat underlying networks as black boxes, operating only on output logits or global feature statistics. While recent work such as TIDE (Agarwal et al., 2025) shows that training with local concept supervision can enable test-time correction, it requires specialized training pipelines and externally generated concept annotations. In the context of prototype-based models, current approaches overlook the rich interpretable intermediate signals, such as prototype activation patterns, spatial similarity maps, and semantic feature associations, that distinguish these architectures from conventional networks. This raises a fundamental question: Can we leverage a model's inherent interpretability signals to enable more semantic and reliable test-time adaptation? To this end, we introduce *ProtoTTA*, the first TTA framework designed specifically for prototype-based architectures. Beyond improving robustness, our framework provides diagnostic transparency: we can trace why adaptation succeeds or fails by observing whether the model reactivates domain-invariant prototypes suppressed by corruption. To validate this, we introduce interpretability-aware metrics that assess semantic consistency, prototype alignment, and prediction stability, complementing standard accuracy measures with insights into the model's reasoning process. We further introduce a vision-language model (VLM)-based evaluation framework that explains TTA dynamics through language and prototype evidence quantitatively and qualitatively (Figure 2), and show that our metrics correlate strongly with VLM-rated reasoning quality.

2 METHOD

Prototype-based architectures provide three forms of granular interpretability absent in black-box models: (i) similarity scores quantifying input-prototype matching (prototype activations), (ii) prototype-to-class weights in the final classification head, and (iii) spatial localization binding prototypes to specific image regions. We leverage these intrinsic properties to design a specialized TTA method. Figure 1 illustrates the ProtoTTA pipeline. The core intuition behind our approach is that distribution shifts can corrupt the prototype selection mechanism, causing the model to activate semantically irrelevant prototypes and suppress correct ones, degrading both accuracy and interpretability. Our adaptation strategy must therefore encourage reactivation of relevant prototypes while suppressing spurious activations. We achieve this by minimizing the entropy of prototype activations, compelling the model toward confident and unambiguous prototype matching.

To prevent adaptation on ambiguous or corrupted samples that could destabilize the model, we update only on a reliable set $\mathcal{R}$ using geometric filtering: we select samples whose maximum prototype similarity (after sub-prototype aggregation) exceeds a threshold τ, optionally combined with a low-entropy constraint on the prediction distribution to ensure model confidence and sample utility. For each test sample $x_i \in \mathcal{R}$, the forward pass yields a pseudo-label $\hat{y}_i$, and we focus our adaptation specifically on the set of target prototypes $\mathcal{P}_t$ associated with this pseudo-label. Unlike standard TTA methods that minimize Shannon entropy on output logits, a distinct challenge here is that prototype activations are similarity scores (e.g., cosine similarity $\in [-1, 1]$) rather than probability distributions. Moreover, unlike logits, where only one class should dominate, prototype activa-

Table 1: Test-time adaptation performance on CUB-200-C (Bird Classification) across diverse corruption types. We report accuracy (%) for ProtoViT (Transformer-based). The final column shows the mean ± std across all corruptions. Best results in **bold**, second-best underlined.

Method	Noise					Blur			Weather				Digital					Total
	Gauss	Shot	Impul	Speck	**Avg**	Defoc	Gauss	**Avg**	Brit	Fog	Frost	**Avg**	Contr	Elast	Jpeg	Pixel	**Avg**	
Backbone: ProtoViT (LN) — Dataset: CUB-200-C																		
Unadapted	37.3	36.5	39.6	48.8	40.5	42.8	39.0	40.9	69.2	61.6	43.4	58.1	49.6	71.0	67.7	68.1	64.1	51.9 ± 13.0
Memo	37.2	36.5	38.6	48.3	40.2	43.3	39.2	41.2	69.8	62.9	43.5	58.7	53.3	71.7	68.2	69.8	65.8	52.5 ± 13.5
SAR	40.0	38.4	41.3	49.3	42.2	43.0	38.5	40.7	70.1	62.7	45.2	59.3	47.0	71.4	67.7	68.3	63.6	52.5 ± 12.8
Tent	45.8	41.2	40.2	45.5	43.2	41.5	39.2	40.4	72.2	63.0	49.9	61.7	52.9	73.8	67.0	70.3	66.0	54.0 ± 12.8
EATA	50.1	51.4	51.7	61.4	53.7	45.1	41.5	43.3	73.4	69.4	52.7	65.2	55.3	74.8	68.0	70.8	67.2	58.9 ± 10.8
***ProtoTTA* (Ours)**	**52.0**	**52.3**	**53.6**	**65.0**	**55.7**	**46.7**	**43.3**	**45.0**	**73.9**	**70.1**	**53.1**	**65.7**	**56.1**	**75.0**	**68.3**	**72.3**	**67.9**	**60.1** ± 10.6

Table 2: Test-time adaptation on Amazon-C review classification using ProtoLens trained on Yelp. We report accuracy (%) across 5 textual corruption types at 4 severity levels ($s \in \{20, 40, 60, 80\}$). The final column shows the mean ± std across all 20 scenarios. Best in **bold**, second-best underlined.

Method	Qwerty					Swap					Remove					Mixed					Aggressive					Total
	20	40	60	80	**Avg**	20	40	60	80	**Avg**	20	40	60	80	**Avg**	20	40	60	80	**Avg**	20	40	60	80	**Avg**	
Unadapted	89.4	85.4	80.0	71.5	81.6	89.9	84.4	76.2	70.6	80.3	**90.5**	86.8	82.5	**76.8**	84.2	84.1	84.1	84.1	80.3	83.2	87.5	79.6	71.2	61.3	74.9	80.81 ± 7.43
SAR	81.3	85.7	81.3	66.5	78.7	88.7	82.8	69.8	63.9	76.3	81.3	87.1	80.3	70.8	79.8	81.3	82.8	81.3	77.4	80.7	81.3	76.6	**78.1**	55.2	72.8	77.65 ± 8.19
Tent	87.5	85.2	**90.6**	68.5	82.9	**90.2**	84.1	75.4	68.9	79.6	87.5	86.8	81.4	75.1	82.7	81.3	83.6	81.3	79.0	81.3	81.3	77.9	**78.1**	58.4	73.9	80.08 ± 7.80
EATA	87.5	85.4	**90.6**	71.6	**83.8**	89.8	84.2	76.3	**70.7**	80.2	87.5	86.8	82.5	76.8	83.4	81.3	84.1	81.3	80.2	81.7	81.3	79.5	**78.1**	61.6	75.1	80.84 ± 6.91
***ProtoTTA* (Ours)**	**89.5**	**86.2**	81.1	**72.2**	82.2	90.0	**85.0**	**76.4**	69.9	**80.3**	90.5	**87.7**	**83.2**	76.1	**84.4**	**85.0**	**85.0**	**85.0**	**81.3**	**84.0**	**88.1**	**80.6**	71.9	**62.3**	**75.7**	**81.33** ± 7.45

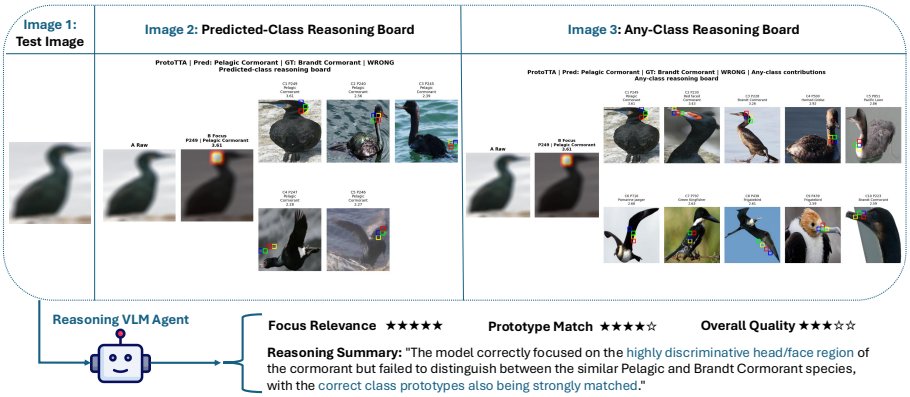

Figure 2: The VLM agent scores prototype-based reasoning boards, providing language-grounded evaluation of TTA dynamics unavailable for black-box methods.

tions represent independent matching signals where each prototype can legitimately exhibit strong or weak similarity. To address this, we map the raw similarity scores s_{ip} into a probability space $\bar{s}_{ip} \in [0, 1]$ using linear scaling (for pre-normalized similarities) or other normalization schemes as needed, such as log-scaling. We then minimize the binary entropy of each mapped activation independently, which encourages each prototype to produce decisive similarities (near 0 or 1) rather than ambiguous mid-range values. In this mapped space, a value of 0.5 corresponds to maximum entropy (equivalent to cosine similarity of 0), indicating high uncertainty or noise that should be suppressed. The adaptation objective is:

$$\mathcal{L}_{\text{ProtoTTA}} = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} c_i \cdot \sum_{p \in \mathcal{P}_t} w_p \cdot H(\bar{s}_{ip}) \tag{1}$$

where $H(\bar{s}_{ip}) = -\bar{s}_{ip} \log(\bar{s}_{ip}) - (1 - \bar{s}_{ip}) \log(1 - \bar{s}_{ip})$ is the binary entropy of the mapped prototype activation. c_i is the confidence score of sample x_i; w_p represents the prototype weights from the final classification head for that specific class; and $\mathcal{P}_t$ is the set of target prototypes derived from pseudo-label $\hat{y}_i$. Minimizing this objective suppresses low-activation noise and reinforces strong prototype associations. Standard backbones with sub-prototypes typically derive the representative similarity score via maximum pooling or global averaging. However, maximum pooling is sensitive to outliers, while averaging weakens decisive signals by mixing them with low-activation noise. To mitigate this, we use a Top-K Mean strategy that aggregates the k most relevant scores to enforce robust consensus. Finally, we update normalization parameters to correct global shifts and fine-tune structural add-ons, specifically attention biases (for Transformers) or 1×1 convolutions (for CNNs). These targeted updates recalibrate the model's semantic focus under domain shift while preserving the structural knowledge learned during training (Algorithm Pseudocode in Appendix §E).

Table 3: Efficiency and Interpretability Analysis across distinct backbones/datasets.

Method	Interpretability Metrics			Efficiency Metrics	
	Semantic Consis. (PAC) ↑	Proto. Alignment (PCA-W) ↑	Prediction Stability ↑	Selection Rate ↓	Rel. Speed ↑
Backbone: ProtoViT (LN + Attention Biases) — Dataset: CUB-200-C					
Unadapted	88.2 ± 3.6	70.8 ± 11.2	$54.1 \pm 14.6\%$	0.0%	99.8%
Memo	88.3 ± 3.6	71.1 ± 11.2	$54.4 \pm 14.6\%$	100.0%	2.3%
SAR	88.9 ± 3.5	74.3 ± 10.3	$57.8 \pm 14.0\%$	98.9%	48.7%
Tent	88.3 ± 5.8	59.1 ± 32.3	$46.5 \pm 30.2\%$	100.0%	91.7%
EATA	91.3 ± 3.1	81.1 ± 7.7	$66.5 \pm 11.8\%$	68.1%	94.9%
***ProtoTTA* (Ours)**	$\mathbf{91.9 \pm 2.9}$	$\mathbf{82.6 \pm 7.1}$	$\mathbf{68.7 \pm 11.5\%}$	**58.0%**	**95.7%**
Backbone: ProtoLens (LN + Attention Biases) — Dataset: Amazon-C					
Unadapted	17.9 ± 0.7	64.4 ± 3.9	$50.7 \pm 0.5\%$	0.0%	100.0%
SAR	$\mathbf{20.8 \pm 6.2}$	63.8 ± 3.8	$\mathbf{51.6 \pm 4.7\%}$	7.8%	62.8%
Tent	18.1 ± 2.8	64.1 ± 4.1	$50.1 \pm 1.9\%$	100.0%	98.0%
EATA	18.6 ± 2.4	64.5 ± 3.6	$50.2 \pm 1.9\%$	**1.4%**	100.0%
***ProtoTTA* (Ours)**	18.2 ± 0.7	$\mathbf{64.8 \pm 3.9}$	$50.8 \pm 0.4\%$	28.2%	**100.0%**

3 Experiments and Results

Dataset and Experimental Settings. We evaluate ProtoTTA on vision (CUB-200-C (Wah et al., 2011)) and NLP (Amazon-C) benchmarks using ProtoViT (Ma et al., 2024) and ProtoLens (Wei & Zhu, 2025) as primary backbones; results on ProtoPNet (Chen et al., 2019)/SICAPv2-C (Silva-Rodríguez, 2020) and ProtoPFormer (Xue et al., 2024)/Stanford Dogs-C (Khosla et al., 2011) are in Appendix §B. Beyond classification accuracy, we report three interpretability-oriented metrics: *(i) Prototype Activation Consistency (PAC)*, measuring cosine similarity between clean and adapted activations to quantify semantic stability; *(ii) Weighted Prototype Alignment (PCA-W)*, evaluating if highly activated prototypes align with the ground truth, weighted by activation strength and classification-layer importance; and *(iii) Prediction Stability*, measuring agreement between clean and adapted predictions. For efficiency, we additionally report the *Selection Rate* (fraction of test samples triggering adaptation) and *Relative Speed* compared to the unadapted model. More details about experiments, datasets, and metrics can be found in Appendix §A.

Results. *ProtoTTA* consistently outperforms all baselines across vision and NLP benchmarks (Tables 1, 2), demonstrating that prototype-specific signals enable more effective adaptation than treating interpretable models as black boxes. Critically, *ProtoTTA* is entirely source-free, whereas EATA, the closest competitor, requires $\approx$2000 source samples. ProtoViT's sub-prototype structure provides ample capacity for semantic refocusing, yielding substantial gains. We note that blur corruptions are uniquely challenging over all vision backbones, suggesting patch-prototype matching relies heavily on vulnerable high-frequency local features. Cross-domain NLP results confirm generalization beyond vision despite the added challenge of shared non-class-specific prototypes in ProtoLens. Generalization to CNN and additional Transformer backbones as well as integration with existing TTA methods is demonstrated in Appendix §B. Beyond accuracy, Table 3 reveals a crucial takeaway: *ProtoTTA* achieves superior performance while adapting on fewer samples, maintaining both efficiency and interpretability. High PAC and PCA-W scores confirm that adaptation restores the model's original reasoning process, reactivating correct-class prototypes suppressed by corruption, rather than merely improving output statistics (qualitative analysis in Appendix §G). Ablation studies (Appendix §F) validate our design choices, and continual adaptation experiments confirm resistance to catastrophic forgetting (Appendix §H).

VLM-Based Interpretability Evaluation. Prototype-based TTA is inherently explainable: reasoning boards expose which prototypes drive each prediction, enabling automated language-grounded evaluation unavailable for black-box methods. We pass reasoning boards from a representative CUB-200-C subset to a VLM agent that scores focus relevance, prototype match, and overall adaptation quality (protocol in Appendix §C). Table 6 shows that *ProtoTTA* achieves the highest scores across all dimensions, with the largest gains in focus relevance ($+0.12$ over EATA) and prototype match ($+0.20$ over unadapted), confirming that adaptation improves semantic reasoning quality beyond accuracy. Furthermore, sample-level PCA-W correlates positively with VLM-rated overall quality across all methods (Pearson r=0.53, Spearman ρ=0.59), validating that it captures genuine interpretability signals aligned with human-proxy evaluation. Notably, this correlation strengthens significantly under *ProtoTTA* (r=0.68), as our adaptation explicitly suppresses the noise-induced semantic hallucinations that otherwise decouple mathematical activations from visual reality (analysis in Appendix §C.2). Ultimately, this framework enables a direct comparison between unadapted and adapted predictions, providing a language-grounded analysis of exactly how adaptation alters semantic focus and showcasing the unique power of explainable TTA (details in Appendix §D).

REFERENCES

Aishwarya Agarwal, Srikrishna Karanam, and Vineet Gandhi. Tide: Training locally interpretable domain generalization models enables test-time correction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30210–30220, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Arthur Chen, Zuxin Liu, Jianguo Zhang, Akshara Prabhakar, Zhiwei Liu, Shelby Heinecke, Silvio Savarese, Victor Zhong, and Caiming Xiong. Grounded test-time adaptation for llm agents. *arXiv preprint arXiv:2511.04847*, 2025.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

Jon Donnelly, Alina Jade Barnett, and Chaofan Chen. Deformable protopnet: An interpretable image classifier using deformable prototypes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10265–10275, 2022.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2, 2011.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, July 2024. ISSN 1573-1405. doi: 10.1007/s11263-024-02181-w. URL http://dx.doi.org/10.1007/s11263-024-02181-w.

Chiyu Ma, Jon Donnelly, Wenjun Liu, Soroush Vosoughi, Cynthia Rudin, and Chaofan Chen. Interpretable image classification with adaptive prototype-based vision transformers. *Advances in Neural Information Processing Systems*, 37:41447–41493, 2024.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *The Internetional Conference on Machine Learning*, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *Internetional Conference on Learning Representations*, 2023.

Mihir Prabhudesai, Tsung-Wei Ke, Alex Li, Deepak Pathak, and Katerina Fragkiadaki. Diffusiontta: Test-time adaptation of discriminative models via generative feedback. *Advances in Neural Information Processing Systems*, 36:17567–17583, 2023.

Barbara Rychalska, Dominika Basaj, Alicja Gosiewska, and Przemysław Biecek. Models in the wild: On corruption robustness of neural nlp systems. In *International Conference on Neural Information Processing*, pp. 235–247. Springer, 2019.

Sahil Sethi, David Chen, Thomas Statchen, Michael C Burkhart, Nipun Bhandari, Bashar Ramadan, and Brett Beaulieu-Jones. Protoecgnet: Case-based interpretable deep learning for multi-label ecg classification with contrastive learning. *arXiv preprint arXiv:2504.08713*, 2025.

Julio Silva-Rodríguez. SICAPv2 - prostate whole slide images with gleason grades annotations, 2020. URL https://data.mendeley.com/datasets/9xxm58dvs3/1.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867, 2020.

Hooman Vaseli, Ang Nan Gu, S Neda Ahmadi Amiri, Michael Y Tsang, Andrea Fung, Nima Kondori, Armin Saadat, Purang Abolmaesumi, and Teresa SM Tsang. Protoasnet: Dynamic prototypes for inherently interpretable and uncertainty-aware aortic stenosis classification in echocardiography. In *International conference on medical image computing and computer-assisted intervention*, pp. 368–378. Springer, 2023.

Hooman Vaseli, Victoria Wu, Nima Kondori, Nguyen Nhat Minh To, Andrea Fung, Ang Nan Gu, and Purang Abolmaesumi. Happi: Hyperbolic hierarchical part prototypes for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 685–694, 2025.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Jul 2011.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Bowen Wei and Ziwei Zhu. ProtoLens: Advancing prototype learning for fine-grained interpretability in text classification. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4503–4523, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.226. URL https://aclanthology.org/2025.acl-long.226/.

Yuanyuan Wei, Roger Tam, and Xiaoying Tang. Mprotonet: A case-based interpretable model for brain tumor classification with 3d multi-parametric magnetic resonance imaging. In *Medical Imaging with Deep Learning*, pp. 1798–1812. PMLR, 2024.

Mengqi Xue, Qihan Huang, Haofei Zhang, Jingwen Hu, Jie Song, Mingli Song, and Canghong Jin. Protopformer: concentrating on prototypical parts in vision transformers for interpretable image recognition. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 1516–1524, 2024.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022.

A Dataset and Experimental Details

Vision Benchmarks and Models. We evaluate ProtoTTA on two challenging vision datasets that require fine-grained discrimination. For *CUB-200* (Wah et al., 2011), a fine-grained bird classification dataset (200 classes), we use the ProtoViT (Ma et al., 2024) model with a DeiT-S/16 backbone. This model contains 2,000 prototypes (10 per class, with 4 sub-prototypes each) and achieves 85.4% accuracy on clean test data. We construct the *CUB-200-C* dataset using 13 of the corruption types with severity 5, following ImageNet-C (Hendrycks & Dietterich, 2019), categorized as follows:

- **Noise:** Gaussian, Shot, Impulse, Speckle
- **Blur:** Gaussian, Defocus
- **Weather:** Fog, Frost, Brightness
- **Digital:** JPEG Compression, Contrast, Pixelate, Elastic Transform

For *SICAPv2* (Silva-Rodríguez, 2020), a challenging histopathology dataset for prostate cancer grading that requires distinguishing subtle morphological differences between cancer grades, we employ the legacy ProtoPNet architecture (Chen et al., 2019) with a VGG19-BN (Simonyan & Zisserman, 2014) backbone. This model is configured with 50 prototypes (10 per class) and achieves 63.4% accuracy on clean test data. To bridge the gap between ProtoPNet's distance-based metric and the probabilistic requirements of our adaptation method, we apply a specific transformation (a form of Log-Inverse Distance Kernel) to the prototype activations. ProtoPNet natively outputs minimum squared Euclidean distances, d_{min}. We map these to a probability distribution by first computing a raw similarity score s_{raw} via logarithmic activation:

$$s_{raw} = \log\left(\frac{d_{min} + 1.0}{d_{min} + 10^{-4}}\right) \tag{2}$$

which is subsequently normalized to the $[0, 1]$ range. We construct *SICAPv2-C* following the same corruption protocol as CUB-200-C.

For *Stanford Dogs* (Khosla et al., 2011), a fine-grained dog breed classification dataset (120 classes), we use the ProtoPFormer (Xue et al., 2024) model with a DeiT-S/16 backbone. ProtoPFormer extends prototype learning to Vision Transformers through a *token reservation* mechanism: at the last transformer block, only the top-k most-attended patch tokens are retained for prototype comparison, concentrating the model's attention on discriminative image regions. The model employs a dual-branch architecture with 1,800 prototypes in total, 1,200 local patch-level prototypes (10 per class), and 600 global CLS-token prototypes (5 per class), each of dimension 384, with predictions formed as an equal-weight combination of both branches. It achieves 90.75% accuracy on clean test data. We construct *Stanford Dogs-C* using the same 13 corruption types at severity 5 following ImageNet-C (Hendrycks & Dietterich, 2019), with the same category groupings as CUB-200-C. Results are reported in Appendix §B.

NLP Benchmark and Model. We evaluate ProtoLens (Wei & Zhu, 2025) for review sentiment classification. We use a ProtoLens model with an *all-mpnet-base-v2* (Song et al., 2020) backbone, configured with 50 shared prototypes. Each prototype represents a semantic concept and is associated with some representative sub-sentences extracted from the training set via a 5-word sliding window for interpretability. The model was trained on the Yelp dataset (94.0% accuracy) and achieves 91.97% accuracy on the clean Amazon test dataset. We create the *Amazon-C* benchmark following WildNLP (Rychalska et al., 2019) with 5 corruption types applied across 4 severity levels (20%, 40%, 60%, and 80% word-level corruption), resulting in 20 total experimental settings. The corruptions are categorized as follows:

- **Keyboard:** QWERTY
- **Character:** Swap, Remove Char
- **Combined:** Mixed, Aggressive

During test-time adaptation, we apply a temperature-scaled sigmoid function to the prototype similarities to map them into the $[0, 1]$ range with an appropriate semantic spread:

$$p_i = \sigma(\tau \cdot s_i) = \frac{1}{1 + e^{-\tau \cdot s_i}} \tag{3}$$

where s_i is the cosine similarity for prototype i, and τ is a temperature hyperparameter (typically set to 5.0) that controls the sharpness of the resulting probability distribution.

Optimization. Optimization is performed using Adam (Kingma & Ba, 2017) with a learning rate of 10^{-3} and momentum $\beta_1 = 0.9$. We use a test batch size of 128 and perform a single gradient update per batch, though our method remains effective across varying batch sizes. All of our experiments are conducted in non-episodic settings. We observed the same trend in episodic settings as well, and we plan to incorporate these findings into future designs.

A.1 Evaluation Metrics

To rigorously assess both performance and interpretability preservation, we report the following metrics alongside standard classification accuracy.

Prototype Activation Consistency (PAC). PAC quantifies the semantic stability of the model by measuring the cosine similarity between the prototype activation vectors of the clean ($\mathbf{a}_i^{\text{clean}}$) and adapted ($\mathbf{a}_i^{\text{adapted}}$) inputs for a given sample i:

$$\text{PAC} = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbf{a}_i^{\text{clean}} \cdot \mathbf{a}_i^{\text{adapted}}}{\|\mathbf{a}_i^{\text{clean}}\|_2 \|\mathbf{a}_i^{\text{adapted}}\|_2} \tag{4}$$

where N is the total number of test samples. A higher PAC score indicates that the adaptation process preserves the original semantic representation of the input.

Weighted Prototype Alignment (PCA-W). To verify that the model attends to semantically correct features, we introduce PCA-W. For a sample i with ground truth class y_i, we analyze the set of top-k activated prototypes $\mathcal{T}_i$. We weight each prototype $p \in \mathcal{T}_i$ by its contribution to the true class, defined as $c_{i,p} = a_{i,p} \cdot |W_{y_i,p}|$, where $a_{i,p}$ is the activation strength and $W_{y_i,p}$ is the weight connecting prototype p to class y_i:

$$\text{PCA-W} = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{p \in \mathcal{T}_i} c_{i,p} \cdot \mathbb{1}[\text{class}(p) = y_i]}{\sum_{p \in \mathcal{T}_i} c_{i,p}} \tag{5}$$

Unlike simple accuracy, PCA-W confirms that the model is right for the right reasons—specifically, that the high-activation prototypes actually belong to the correct semantic category.

Prediction Stability. We measure the stability of the decision boundary by calculating the prediction agreement between the clean and adapted models:

$$\text{Prediction Stability} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\hat{y}_i^{\text{adapted}} = \hat{y}_i^{\text{clean}}] \tag{6}$$

High Prediction Stability indicates that the adaptation improves robustness without arbitrarily flipping predictions for samples that were already handled correctly by the clean model.

Efficiency. We report *Selection Rate*, defined as the percentage of test samples that trigger the adaptation update ($N_{\text{adapted}}/N_{\text{total}}$), and *Relative Speed*, defined as the ratio of the inference throughput (samples/sec) of the adapted model to that of the unadapted baseline.

B CNN and Additional Transformer Backbone Results

We evaluate ProtoTTA on two additional backbones to demonstrate generalization across architectures. For the CNN-based ProtoPNet on SICAPv2-C, the architecture presents distinct challenges, including a lack of sub-prototypes, limited depth, and lower separation of prototypes in the representation space. These limitations reduce the available adaptation headroom. To address this, we introduce *ProtoTTA+*, a hybrid approach that leverages logit distributions. By combining our prototype-aware loss (W=0.7) with standard entropy minimization on logits (W=0.3), *ProtoTTA+*

bridges this gap and achieves best-in-class performance. This demonstrates that our prototype-guided approach is complementary and can be seamlessly integrated with existing TTA methods rather than simply replacing them.

Furthermore, for ProtoPFormer on Stanford Dogs-C—which extends prototype learning to Vision Transformers via token reservation—*ProtoTTA+* achieves the best overall accuracy (57.7%). This confirms that our hybrid approach generalizes effectively across diverse prototype architectures and complex fine-grained recognition tasks. These results demonstrate that our approach is complementary and can be effectively integrated with existing TTA methods such as SAR or Tent, rather than simply replacing them.

Table 4: Test-time adaptation performance on SICAPv2-C (histopathology prostate cancer grading) using ProtoPNet (CNN-based), and on Stanford Dogs-C (fine-grained dog breed classification) using ProtoPFormer (Transformer-based). The final column shows the mean ± std across all corruptions. Best results in **bold**, second-best underlined.

	Noise					Blur			Weather				Digital					
Method	Gauss	Shot	Impul	Speck	**Avg**	Defoc	Gauss	**Avg**	Brit	Fog	Frost	**Avg**	Contr	Elast	Jpeg	Pixel	**Avg**	**Total**
Backbone: ProtoPNet (BN + 1×1 Convs) — Dataset: SICAPv2-C																		
Unadapted	16.4	11.1	13.1	11.5	13.0	36.9	34.4	35.7	31.9	30.3	31.3	31.2	30.9	57.6	49.4	55.5	48.4	31.6 ± 15.2
Memo	58.5	**58.7**	57.5	57.1	**58.0**	51.1	54.1	52.6	**59.7**	50.1	59.5	56.4	**44.7**	56.8	51.7	60.2	53.4	55.4 ± 4.5
SAR	58.2	58.3	57.9	55.8	57.6	**54.2**	**56.1**	**55.2**	57.6	**51.6**	60.1	56.4	43.5	57.5	51.5	61.6	53.5	55.7 ± 4.5
Tent	56.6	54.6	**58.8**	54.9	56.2	53.0	55.7	54.4	52.1	45.0	58.4	51.8	39.8	60.1	50.9	60.2	52.8	53.8 ± 5.7
EATA	58.3	58.3	58.1	55.7	57.6	53.4	55.0	54.2	58.4	50.7	**61.3**	**56.8**	43.3	58.2	52.2	60.3	53.5	55.6 ± 4.7
***ProtoTTA* (Ours)**	**59.4**	58.1	57.9	55.1	57.6	51.6	54.4	53.0	56.5	47.9	60.6	55.0	43.8	58.7	52.2	61.7	54.1	55.2 ± 5.0
***ProtoTTA+* (Ours)**	58.3	57.6	58.6	**57.3**	**58.0**	51.9	**56.1**	54.0	59.2	49.2	59.2	55.9	43.7	**60.3**	**52.3**	**61.9**	**54.6**	**55.8 ± 4.9**
Backbone: ProtoPFormer (LN + Attention Biases) — Dataset: Stanford Dogs-C																		
Unadapted	46.6	46.2	48.1	55.6	49.1	40.4	37.4	38.9	**69.6**	51.2	48.0	56.3	46.5	57.8	**68.5**	**70.8**	60.9	52.8 ± 10.6
Tent	52.7	51.8	**54.7**	60.7	55.0	45.8	42.3	44.1	68.3	60.4	53.8	60.8	49.3	64.9	67.1	70.6	63.0	57.1 ± 8.6
SAR	36.7	40.9	41.8	48.1	41.9	31.5	30.0	30.8	59.6	34.4	30.6	41.5	31.8	50.3	59.5	63.1	51.2	42.9 ± 11.5
EATA	52.0	53.3	53.4	60.5	54.8	**46.7**	43.7	**45.2**	68.2	59.7	52.9	60.3	**49.5**	65.3	67.2	70.7	**63.2**	57.2 ± 8.4
***ProtoTTA+* (Ours)**	**53.8**	**54.9**	53.8	**60.9**	**55.9**	45.2	**45.2**	**45.2**	68.1	**61.8**	**54.2**	**61.4**	48.0	**65.6**	67.9	70.0	62.9	**57.7 ± 8.4**

C VLM-Based Interpretability Evaluation

C.1 Scoring Protocol

We construct a fixed 100-sample subset of CUB-200-C by sampling uniformly across corruption types and severity levels. For each sample and each TTA method, we assemble a reasoning board comprising: (i) the corrupted test image, (ii) the predicted-class reasoning board showing the top-matched prototypes with their spatial focus regions highlighted, and (iii) the any-class contribution board showing the highest-contributing prototypes across all classes alongside their contribution scores. These boards are passed to a VLM agent (Qwen/Qwen3-VL-32B-Thinking (Bai et al., 2025)) that returns three scores on a 1–5 integer scale:

- **Focus Relevance:** whether the highlighted image region corresponds to a semantically meaningful and species-discriminative part rather than background or noise.
- **Prototype Match:** whether the retrieved prototype patches are visually similar to the highlighted region in the test image.
- **Overall Adaptation Quality:** the overall semantic convincingness of the model's prototype-based reasoning for the predicted class.

Table 6 reports the complete results, including per-metric standard deviations. Furthermore, we measured the overall quality scores exclusively on samples where the models predicted correctly (mean$=4.62$ for ProtoTTA). These filtered scores consistently mirrored the general performance trends, demonstrating that the positive correlation between predictive accuracy and VLM-judged reasoning quality persists across adaptation methods.

C.2 PCA-W and VLM Correlation

To validate that our proposed PCA-W metric captures genuine interpretability signals, we compute sample-level PCA-W from the saved prototype contribution scores and correlate them with the

Table 5: Efficiency and interpretability analysis for ProtoPNet on SICAPv2-C and ProtoPFormer on Stanford Dogs-C. Best in **bold**, second-best underlined.

Method	Interpretability Metrics			Efficiency Metrics	
	Semantic Consis. (PAC) ↑	Proto. Alignment (PCA-W) ↑	Prediction Stability ↑	Selection Rate ↓	Rel. Speed ↑
Backbone: ProtoPNet (BN + 1×1 Convs) — Dataset: SICAPv2-C					
Unadapted	**88.8 ± 2.6**	51.3 ± 5.9	20.8 ± 13.0%	0.0%	99.9%
Memo	86.1 ± 0.4	60.7 ± 1.5	**31.1 ± 1.3%**	100.0%	45.4%
SAR	85.9 ± 0.3	**62.0 ± 1.9**	28.3 ± 1.8%	100.0%	22.6%
Tent	84.5 ± 0.5	60.9 ± 2.6	22.7 ± 1.2%	100.0%	45.3%
EATA	86.2 ± 0.4	61.9 ± 2.0	26.6 ± 1.2%	3.0%	**89.5%**
***ProtoTTA* (Ours)**	87.7 ± 1.1	60.4 ± 2.4	27.9 ± 2.1%	**1.2%**	75.3%
***ProtoTTA*+ (Ours)**	87.4 ± 0.6	59.8 ± 1.4	27.2 ± 2.2%	76.3%	45.1%
Backbone: ProtoPFormer (LN + Attention Biases) — Dataset: Stanford Dogs-C					
Unadapted	92.1 ± 0.4	53.5 ± 10.2	55.5%	0.0%	100.0%
Tent	91.2 ± 0.2	64.8 ± 7.3	61.8%	100.0%	74.6%
EATA	91.3 ± 0.3	**65.1** ± 7.1	**63.4%**	69.7%	72.1%
SAR	**93.8** ± 0.2	35.3 ± 11.3	58.0%	**0.7%**	**94.5%**
***ProtoTTA*+ (Ours)**	91.2 ± 0.2	64.4 ± 7.4	58.6%	71.4%	82.5%

Table 6: Full VLM evaluation results on the 100-sample CUB-200-C subset. All scores on a 1–5 scale. Best in **bold**, second-best underlined.

Method	Focus Relevance ↑	Prototype Match ↑	Overall Quality ↑
Unadapted	4.14 ± 0.98	3.66 ± 0.87	3.53 ± 1.02
Tent	4.03 ± 0.92	3.64 ± 0.92	3.59 ± 1.02
EATA	4.18 ± 0.90	3.79 ± 0.82	3.75 ± 0.94
ProtoTTA (Ours)	**4.30 ± 0.90**	**3.86 ± 0.81**	**3.78 ± 0.97**

VLM's overall adaptation quality scores. Sample-level PCA-W is defined as:

$$\text{PCA-W}(x) = \frac{\sum_{p \in \mathcal{P}_{GT}} c_p}{\sum_{p \in \mathcal{P}_{top}} c_p} \tag{7}$$

where $\mathcal{P}_{top}$ are the top contributing prototypes for sample x, $\mathcal{P}_{GT} \subseteq \mathcal{P}_{top}$ are those belonging to the ground-truth class, and c_p is the contribution score. A value of 1.0 means all top prototype mass comes from ground-truth class prototypes; 0.0 means none does.

Table 7 reports correlations pooled across all scored samples and within ProtoTTA alone. The moderate positive pooled correlation (Pearson r=0.53) confirms that PCA-W aligns with VLM-rated interpretability across diverse adaptation methods, establishing it as a valid, method-agnostic proxy for semantic reasoning.

Crucially, the correlation strengthens significantly when evaluated exclusively on ProtoTTA samples (r=0.68). This amplification highlights a fundamental mechanism of our approach: realigning mathematical activations with visual reality. Under distribution shifts, unadapted models and standard TTA baselines often suffer from "semantic hallucinations"—mathematically assigning high activation weights to correct-class prototypes based on spurious matching with random background noise. This noise artificially inflates or deflates mathematical metrics, decoupling them from the actual visual evidence (which a VLM accurately scores as poor). By explicitly minimizing prototype activation entropy and employing geometric filtering, ProtoTTA suppresses these noise-induced spurious matches. Consequently, it ensures that when the mathematical score (PCA-W) is high, it is driven by genuinely discriminative, clean visual features, thus tightening the correlation with human-aligned VLM evaluations.

D Explainable Test-Time Adaptation via Prototype Reasoning

A key advantage of prototype-based architectures is that test-time adaptation is inherently explainable. Unlike black-box TTA methods that operate on output logits, prototype models expose the full evidence chain: which image regions are attended to, which training prototypes are retrieved, and

Table 7: Pearson (r) and Spearman (ρ) correlations between sample-level PCA-W and VLM scores.

Subset	Overall Quality		Proto Match		Focus Rel.	
	r	ρ	r	ρ	r	ρ
All (N=397)	0.53	0.59	0.49	0.56	0.32	0.39
ProtoTTA (N=97)	0.68	0.73	0.55	0.62	0.32	0.31

how their contributions change under adaptation. This allows adaptation dynamics to be narrated in semantically meaningful terms.

Figure 3 illustrates a representative example. The corrupted test image (left) shows a heavily noise-corrupted bird. The unadapted model (right, red) misclassifies it as a Crested Auklet by fixating on noisy beak/crown artifacts — the reasoning board shows wrong-class Crested Auklet prototypes dominating both the predicted-class and any-class boards. After ProtoTTA adaptation (middle, green), the model corrects its prediction to the ground-truth Bronzed Cowbird. The reasoning boards reveal *why*: the model shifts attention to the species-discriminative red eye region and retrieves multiple correct-class Bronzed Cowbird prototypes, while suppressing the spurious wrong-class evidence visible in the unadapted boards.

The VLM agent automatically generates this comparative analysis from the reasoning boards, producing structured output covering: unadapted failure analysis, adaptation success analysis, change impact, and a comparative checklist (attention location, wrong-class suppression, prototype match, semantic part focus). This framework enables language-grounded diagnosis of TTA behaviour that is uniquely possible with prototype-based models.

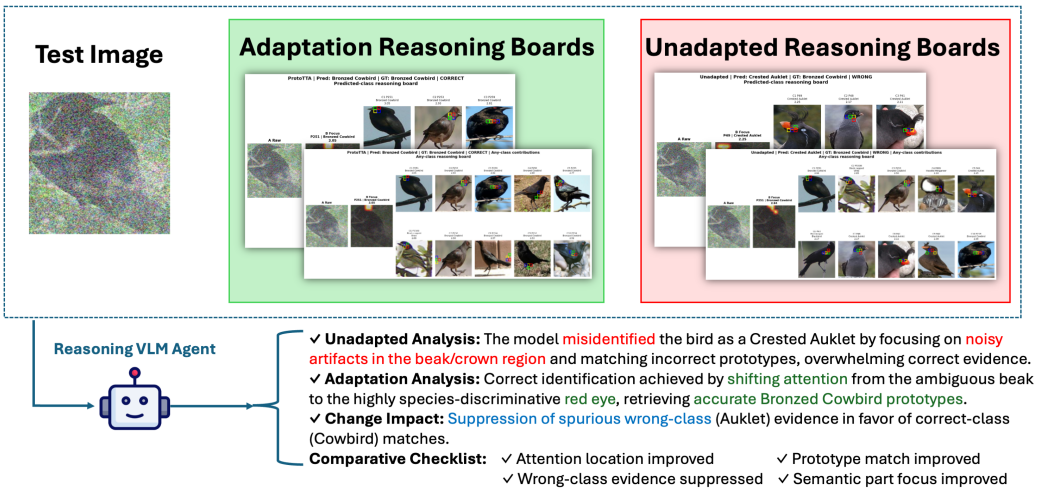

Figure 3: Explainable TTA via prototype reasoning boards. The unadapted model (right) misclassifies the corrupted bird by matching wrong-class prototypes in the noisy beak/crown region. ProtoTTA (middle) corrects the prediction by shifting to the species-discriminative red eye and retrieving correct-class Bronzed Cowbird prototypes, suppressing spurious evidence. The VLM agent automatically narrates this before/after analysis from the reasoning boards.

E PROTOTTA PSEUDOCODE

In this section, we provide the detailed pseudocode for our proposed method, ProtoTTA. Algorithm 1 outlines the complete adaptation procedure, including the geometric filtering mechanism and the prototype-guided entropy minimization steps.

Algorithm 1 ProtoTTA: Test-Time Prototype Updates

Require: Pre-trained prototype model f_θ, test stream $\mathcal{D}_{test}$, similarity threshold τ, learning rate η
Ensure: Adapted parameters Θ (normalization layers + structural add-ons)

Initialize $\Theta \leftarrow \Theta_{init}$ ▷ LayerNorm/BatchNorm + Attention Biases/1×1 Convs
for each batch $\{\mathbf{x}_i\}_{i=1}^{B} \in \mathcal{D}_{test}$ **do**
  // Forward pass and pseudo-labeling
  Compute prototype similarities s_{ip} for all prototypes p
  Map similarities to probability space: $\bar{s}_{ip} \in [0, 1]$ ▷ Linear scaling or sigmoid or log-scale
  Obtain pseudo-labels: $\hat{y}_i \leftarrow \arg\max_c f_\theta(\mathbf{x}_i)$
  // Geometric filtering for reliable set
  $\mathcal{R} \leftarrow \{i \mid \max_p(\bar{s}_{ip}) > \tau\}$ ▷ Optional: add entropy constraint
  if $|\mathcal{R}| > 0$ **then**
    // Consensus aggregation (Top-K Mean for sub-prototypes)
    Aggregate sub-prototype scores via Top-K Mean
    // Compute weighted binary entropy loss
    $\mathcal{L}_{ProtoTTA} \leftarrow \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} c_i \cdot \sum_{p \in \mathcal{P}_{\hat{y}_i}} w_p \cdot H(\bar{s}_{ip})$
      where $H(\bar{s}_{ip}) = -\bar{s}_{ip} \log(\bar{s}_{ip}) - (1 - \bar{s}_{ip}) \log(1 - \bar{s}_{ip})$
      c_i: confidence score, w_p: prototype importance weight, $\mathcal{P}_{\hat{y}_i}$: target prototypes
    // Update only normalization + structural parameters
    $\Theta \leftarrow \Theta - \eta \nabla_\Theta \mathcal{L}_{ProtoTTA}$
  end if
end for
return Adapted model f_Θ

F Ablation Studies

In this section, we provide a detailed analysis of the architectural and algorithmic choices governing *ProtoTTA* for ProtoViT backbone. We validate the necessity of our geometric filtering, the selection of update parameters, and our consensus strategies.

F.1 Impact of Geometric Filtering and Update Parameters

Table 8 illustrates the impact of our two most critical components: the Geometric Filter and the choice of trainable parameters.

Geometric Filtering. The inclusion of the Geometric Filter is the most decisive factor in our framework. As shown in the first section of Table 8, removing the filter (adapting on all samples regardless of reliability) causes a drastic performance drop of nearly 4% (60.12% $\rightarrow$ 56.33%). This confirms our hypothesis that adapting to ambiguous or low-confidence samples introduces noise that degrades the model, whereas restricting updates to the Reliable Set $\mathcal{R}$ ensures positive transfer.

Adaptation Parameters. We investigate which parts of the model should be updated during test-time. Updating all learnable parameters ("All Adaptive") leads to model collapse (51.86%), likely due to catastrophic forgetting on the small batch size. Restricting updates to LayerNorm parameters stabilizes the model (59.41%), a finding consistent with standard TTA literature. However, our proposed strategy of updating Attention Biases alongside LayerNorms yields the best performance (60.12%). This suggests that adjusting the attention mechanism is crucial for prototype-based models to "re-focus" on relevant semantic features under corruption.

F.2 Design Refinements: Aggregation and Weighting

Table 9 examines the finer design choices regarding prototype aggregation, target selection, and loss weighting.

Consensus Strategy. Standard prototype models typically use the `max` operator to aggregate sub-prototype scores. While `max` performs well (60.04%), it is sensitive to outliers. Our Top-K Mean

Table 8: Comparison of Geometric Filtering Strategies and Adaptation Modes on ProtoViT. The "Geometric Filter" is essential for stability, preventing the model from learning on noisy data. For parameters, adding Attention Bias to LayerNorm updates yields the best results.

Configuration	Mean Acc	Std Dev	Min	Max
Geometric Filtering				
No Filter	56.33%	13.74	34.90	74.61
With Filter (Ours)	**60.12%**	**10.55**	**43.29**	**74.96**
Adaptation Parameters				
All Adaptive	51.86%	13.03	36.49	71.06
LayerNorm Only	59.41%	10.89	42.04	75.06
LN + Attn Bias (Ours)	**60.12%**	**10.55**	**43.29**	**74.96**

strategy marginally outperforms it (60.12%) and offers lower variance (10.55 Std Dev), validating that a consensus-based approach is more robust to the noise induced by corruptions.

Weighting and Target Selection. We observe that weighting samples by both confidence (c_i) and prototype importance (w_p) yields the highest accuracy, though the gain over unweighted adaptation is incremental. Similarly, we compared adapting on "Target Prototypes Only" (derived from pseudo-labels) versus "All Prototypes". The results are nearly identical, justifying our design choice to use "Target Only" for its computational efficiency without sacrificing accuracy.

Table 9: Ablation of Consensus strategies, Prototype Targeting, and Loss Weighting on ProtoViT. Our choices (Top-K Mean, Target Only, and Combined Weighting) consistently provide the most robust performance.

Configuration	Mean Acc	Std Dev	Min	Max
Consensus Strategy				
Mean	59.75%	10.77	42.68	75.22
Max (Standard)	60.04%	10.59	43.22	74.89
Top-K Mean (Ours)	**60.12%**	**10.55**	**43.29**	**74.96**
Target vs. All Prototypes				
All Prototypes	60.11%	10.57	42.94	74.99
Target Only (Ours)	**60.12%**	**10.55**	**43.29**	**74.96**
Weighting Strategy				
No Weighting	60.03%	10.62	42.91	74.91
Importance Only	60.04%	10.61	42.91	74.92
Confidence Only	60.07%	10.65	42.79	75.03
Both (Ours)	**60.12%**	**10.55**	**43.29**	**74.96**

G Qualitative Analysis

In this section, we leverage the inherent interpretability of prototype-based models to visually diagnose failure modes under distribution shifts and compare the effectiveness of different adaptation strategies. Unlike black-box models, this architecture allows us to trace the reasoning process by examining which prototypes are activated and how strongly they contribute to the final prediction.

Figure 4 compares the activation intensity of the top-k prototypes across different scenarios on ProtoViT. The clean baseline (Top Left) displays the ideal global activation landscape, where the model attends to a mixture of relevant prototypes. Under corruption, the unadapted model (Top Right) exhibits a distinct failure mode: activations for the ground-truth class are severely suppressed. In contrast, prototypes belonging to incorrect classes are spuriously sharpened, causing the model to "hallucinate" and misclassify. While EATA (Bottom Left) attempts to adapt, it struggles to suppress these hallucinations or fully restore the signal strength of the correct class. In contrast, *ProtoTTA*

(Bottom Right) successfully recovers the original semantic focus. It restores the activation profile to closely mirror the clean baseline, thereby suppressing noise-induced hallucinations and enabling the model to make accurate predictions with high confidence.

Figure 5 provides a deeper insight into why baseline methods fail. By visualizing the top-activated prototypes for the *incorrect* class predicted by EATA, we observe that the model assigns high activation scores to irrelevant semantic features. This confirms that EATA fails to suppress noise-induced artifacts effectively. Instead, the noise deceives the model into "hallucinating" strong matches with incorrect prototypes. This phenomenon, amplified activations for the wrong class (Figure 5) combined with suppressed activations for the correct class (Figure 4), fully explains the drop in robustness.

To further illustrate the contrast in adaptation dynamics, Figure 7 provides a direct comparison of the top-activated prototypes for the predicted class between EATA and ProtoTTA on a corrupted sample (Ground Truth Class: 20). While EATA incorrectly predicts Class 31 by hallucinating strong semantic matches with irrelevant features, ProtoTTA successfully suppresses these noise-induced artifacts. By explicitly minimizing prototype activation entropy , ProtoTTA recovers the correct semantic associations, attending to the appropriate visual features and accurately predicting the ground-truth class.

Finally, Figure 6 visualizes the spatial distribution of prototype activations. We observe that *ProtoTTA* acts as a high-fidelity approximation of the uncorrupted model, triggering the same semantic regions of interest as the clean baseline. In contrast, both the Unadapted model and EATA (the strongest baseline) fail to attend to these discriminative features. Instead, they drift towards irrelevant background noise or spurious artifacts, likely "easier" high-frequency paths for the corrupted model to latch onto, which directly leads to the hallucinations and misclassifications discussed earlier.

H Forgetting Analysis

In our primary experimental setup, the model is reset between different corruption types. To investigate the stability of *ProtoTTA* during prolonged adaptation on ProtoViT, we analyze the performance trend over a single long-sequence corruption (Gaussian Noise, Severity 5, approx. 5,500 samples) without intermediate resets. Figure 8 illustrates the per-batch accuracy of *ProtoTTA* compared to EATA and the unadapted baseline. We observe that *ProtoTTA* maintains a consistent performance advantage over the unadapted model throughout the entire sequence. Crucially, there is no downward trend in accuracy across later batches, indicating that our method does not suffer from *catastrophic forgetting* or error accumulation. The performance fluctuations align with the intrinsic difficulty of specific batches (mirrored by the unadapted baseline and EATA), confirming that our geometric filtering and reliability-weighted updates effectively prevent semantic drift.

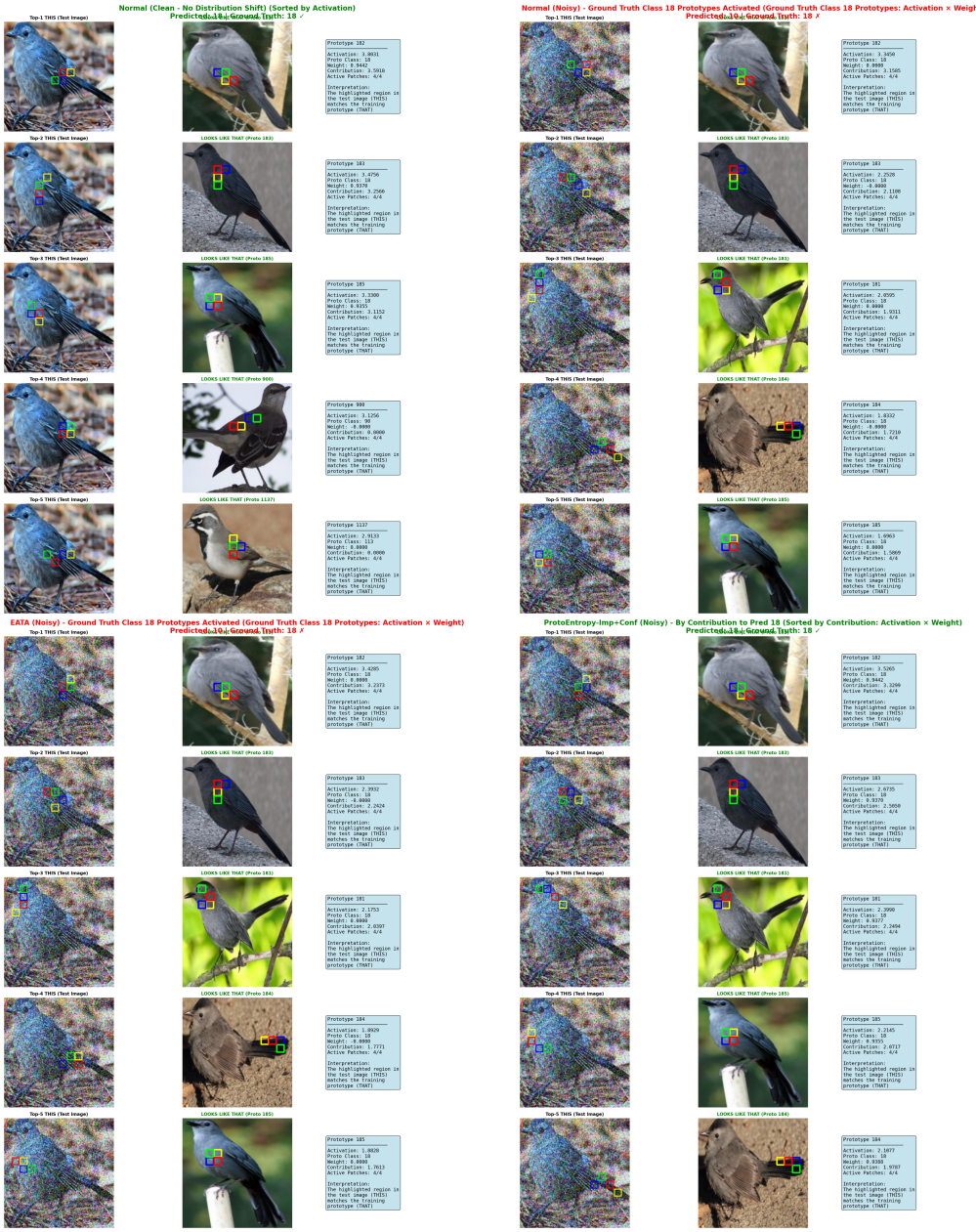

Figure 4: Visualization of Prototype Contributions. We compare the activation strengths of ground-truth class prototypes across three scenarios: (Top-Left) Normal model on clean data prototype activations (global), serving as the "Gold Standard"; (Top-Right) Unadapted model on noisy data, showing suppressed activations; (Bottom-Left) EATA adaptation, which yields incomplete recovery; and (Bottom-Right) *ProtoTTA*, which successfully restores the activation profile to match the clean baseline.

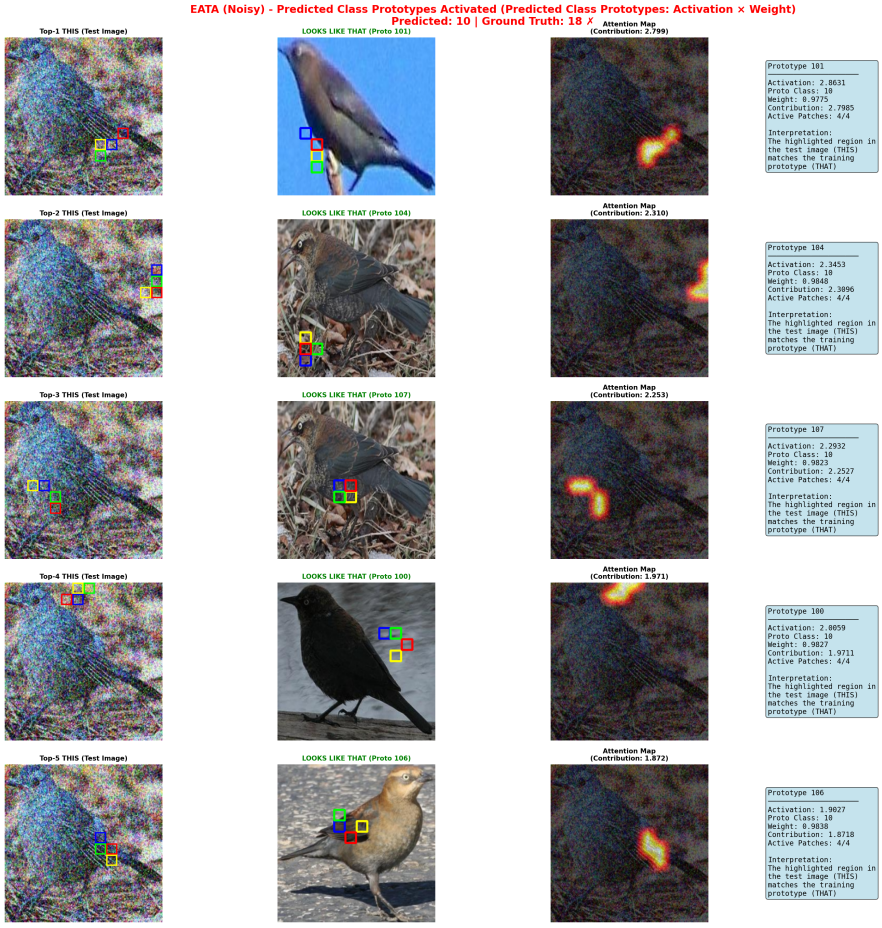

Figure 5: Analysis of EATA Misclassification. We visualize the top-5 activated prototypes for the incorrect class predicted by EATA. Despite the mismatch, the model registers high activation scores (hallucinations), indicating that the adaptation failed to filter out noise-induced features. This contrasts with Figure 4, where ground-truth activations were suppressed.

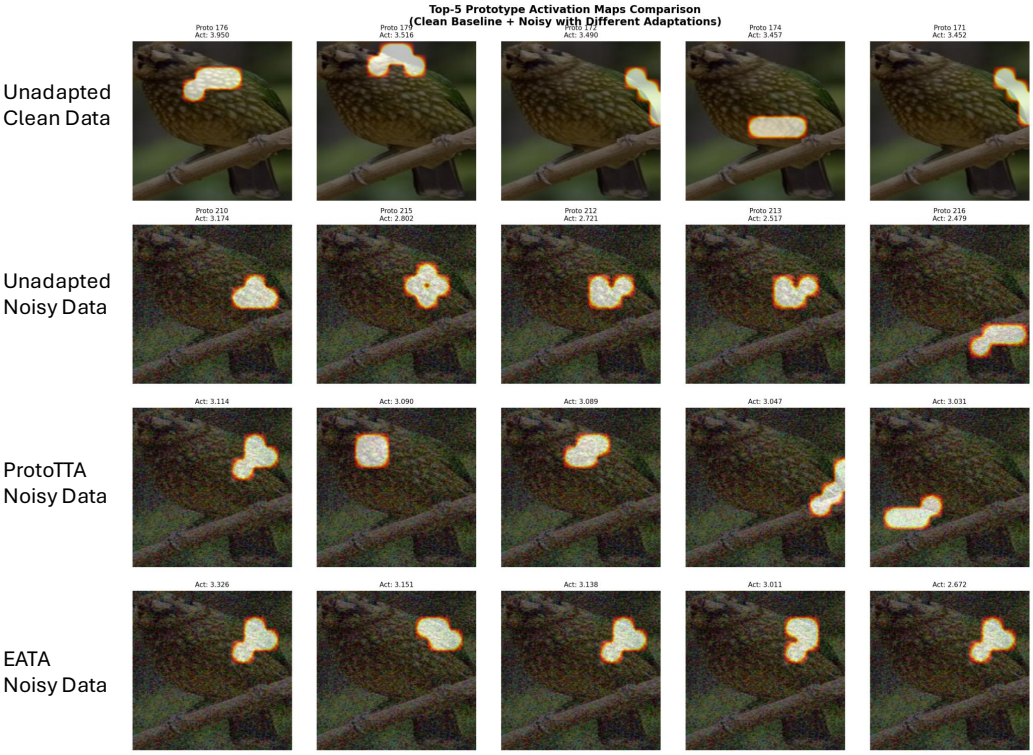

Figure 6: Comparison of Prototype Attention Maps. We visualize the spatial regions of input samples triggering the top prototype activations across different settings. *ProtoTTA* (Third Row) successfully realigns the model's attention, closely matching the focus of the Clean Baseline (Top Row) on the object of interest. Conversely, the Unadapted (Second Row) and EATA (Bottom Row) models diverge significantly, consistently attending to non-informative background regions or noise artifacts.

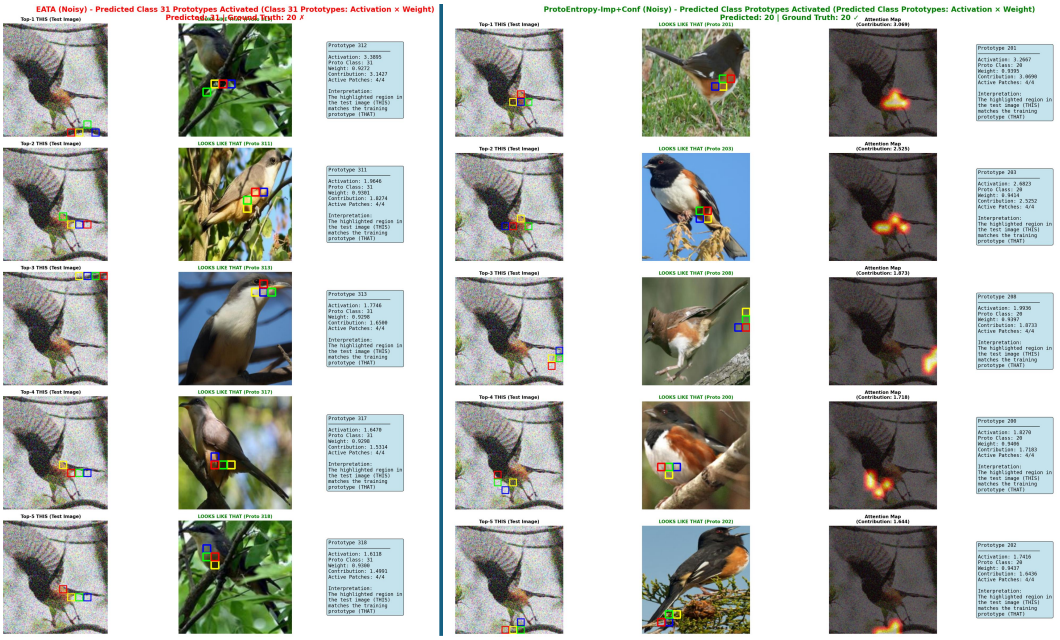

Figure 7: Comparison of Prototype Contributions (Activation × Weight) for the predicted class between EATA (left) and ProtoTTA (right) under noisy conditions. EATA suffers from semantic hallucination, confidently activating irrelevant prototypes for an incorrect class (Predicted: 31, Ground Truth: 20) based on spurious background artifacts. In contrast, ProtoTTA accurately maps spatial features to the ground-truth class prototypes (Predicted: 20, Ground Truth: 20), effectively suppressing noise-induced hallucinations and restoring correct semantic focus.

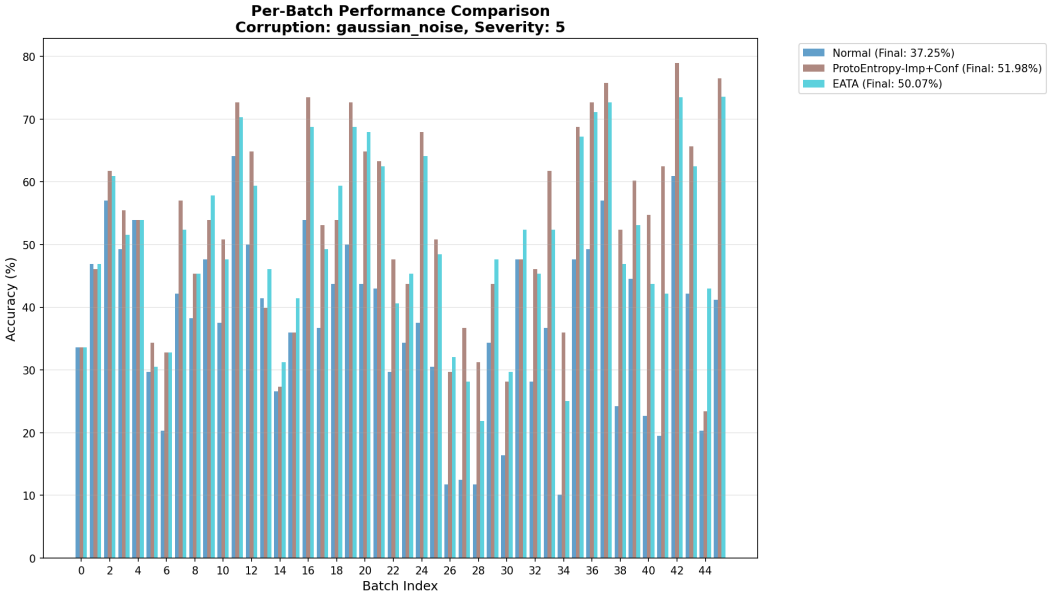

Figure 8: Per-Batch Performance Stability. We track classification accuracy across consecutive batches for Gaussian Noise (Severity 5). *ProtoTTA* (Brown bars) consistently outperforms the Unadapted baseline (Blue) and remains competitive with EATA (Cyan) throughout the sequence. The absence of performance degradation in later batches confirms that *ProtoTTA* avoids catastrophic forgetting, successfully maintaining stability by restricting updates to the reliable set.