A. Detailed Dataset Information

We construct the Openset dataset using the same method as in the AdaND[1] work, where we mix OOD (Out-of-Distribution) data into the ID (In-Distribution) data. We used 9 ID datasets and 4 OOD datasets, ensuring no semantic overlap between the ID and OOD data. The mixing methodology is detailed in Table 1 and Figure 1.

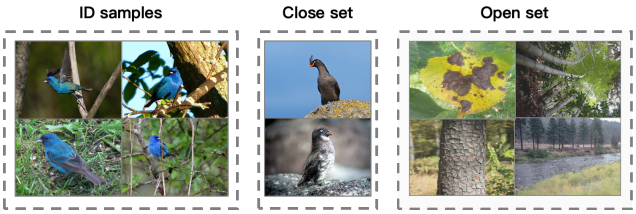

Figure 1. Openset Benchmark. Visual examples including (Left) in-distribution (ID) instances of "Indigo Bunting," (Middle) close-set samples misclassified as "Indigo Bunting," and (Right) out-of-distribution (OOD) samples from an auxiliary dataset.

Table 1. ID and OOD datasets used in our experiments.

ID Dataset	OOD Dataset
Fine-grained Datasets	
CUB-200-2011	iNaturalist, SUN, Texture, Places
Stanford Cars	iNaturalist, SUN, Texture, Places
Food-101	iNaturalist, SUN, Texture, Places
Oxford-IIIT Pet	iNaturalist, SUN, Texture, Places
ImageNet Variants	
ImageNet	iNaturalist, SUN, Texture, Places
ImageNet-K	iNaturalist, SUN, Texture, Places
ImageNet-A	iNaturalist, SUN, Texture, Places
ImageNet-V2	iNaturalist, SUN, Texture, Places
ImageNet-R	iNaturalist, SUN, Texture, Places

In Table 2, we provide comprehensive statistics for each ID dataset utilized in our experiments, detailing the number of classes and the samples. In Table 3, we provide the number of classes in OOD datasets.

B. Text Templates for Each ID Dataset

We adopt the prompt templates listed in Table 4 as the textual templates. Building upon these, we further integrate the fine-grained semantic descriptions generated by GPT-3 [4] to enhance text–image alignment accuracy substantially.

C. Implementation Details

C.1. Adaptation Process

As illustrated in Figure 2, the adaptation process of COSTA can be broadly divided into three phases: a warm-up period, a prototype learning period, and a stable inference period. The warm-up period is configured to last for the initial 128 inference steps. During this stage, we initialize a cache queue with a length of 128. This cache is

Table 2. The number of classes and samples for the ID and OOD datasets used in our experiments.

Dataset	Classes	Samples
Fine-grained Datasets		
CUB-200-2011	200	11788
STANFORD-CARS	196	8041
Food-101	101	25250
Oxford-IIIT Pet	37	7349
ImageNet Variants		
ImageNet	1000	49781
ImageNet-V2	1000	10000
ImageNet-A	200	7500
ImageNet-R	200	30000
ImageNet-K	1000	50899

Table 3. Number of classes in OOD datasets.

Dataset	iNaturalist	SUN	Texture	Places
Classes	110	50	47	50

Table 4. Prompt templates used in different datasets. ImageNet (-*) refers to ImageNet and variants.

Dataset	Prompt template(s)
CUB-200-2011	`a photo of a {}, a type of bird.`
STANFORD-CARS	`a photo of a {}, a type of car.`
Food-101	`a photo of {}, a type of food.`
Oxford-IIIT Pet	`a photo of a {}, a type of pet.`

ImageNet(-*)	`a bad photo of the {}.`
	`a {} in a video game.`
	`an origami {}.`
	`a photo of the small {}.`
	`art of the {}.`
	`a photo of the large {}.`
	`itap of a {}.`

Figure 2. Adaptation Process of COSTA.

used to calculate the adaptive threshold for differentiating between ID and OOD data using the OWTTT algorithm [3]. Following this, the prototype learning period spans approximately 30% of the total inference time for the dataset. In this phase, the model undergoes continuous adaptation to form stable prototypes, and the OOD detector's performance also trends toward convergence. A detector cache queue with a length of 512 is utilized during this period. The final stage is the stable inference period, which begins after the first 30% of inference steps and continues until the end of the dataset. Through experiments on 36 datasets, we have empirically observed that COSTA's adaptation generally converges around this 30% mark. Although there can be fluctuations due to variations in data difficulty, the overall convergence speed remains consistent. For comparison, the AdaND method does not integrate any Test-Time Adaptation (TTA) techniques. As a result, while AdaND converges quickly, its performance does not show further improvement thereafter.

C.2. Discussion on Hyperparameters

COSTA involves two sets of learnable parameters: the first consists of the visual and text prototype residual vectors, and the second is the autoregressive classifier within the detector. For the first set of parameters, the learning rate is uniformly set to 1×10^{-4}. We employ the AdamW optimizer with an epsilon of 1×10^{-3} and a weight decay coefficient of 1×10^{-1}. For the second set of parameters, we optimize the autoregressive classifier with Adam [2], using a learning rate of 0.0005 and no weight decay.

In the COSTA framework, the DCPE module employs an affine-transformed prototype inference strategy to obtain test-time–adapted prediction probabilities. As defined in the main text, this inference incorporates an affine transformation parameterized by α and β. The values of these hyperparameters, shown in Table 5, are selected through a search-based procedure and remain fixed for all evaluations on a given dataset.

C.3. Loss Functions

In the test-time adaptation stage, we introduce a lightweight residual-learning objective that refines the text and visual prototypes on-the-fly for each incoming test sample. Concretely, we optimize learnable residuals $\hat{v}_c, \hat{t}_c$ by minimizing the composite loss

Table 5. Weighted logits parameters (α, β) for different ID datasets. All experiments use the same set of OOD datasets (iNaturalist, SUN, Texture, Places).

ID Dataset	α	β
CUB-200-2011	1.0	1.0
STANFORD-CARS	1.0	7.0
Food-101	1.0	1.0
Oxford-IIIT Pet	2.0	7.0
ImageNet	2.0	5.0

$$\mathcal{L} = \mathcal{L}_{\text{aug}} + \lambda\, \mathcal{L}_{\text{align}}, \tag{1}$$

where the self-entropy term

$$\mathcal{L}_{\text{aug}} = -\sum_{c=1}^{C} P(y = y_c|\mathbf{x}_{\text{test}}) \log P(y = y_c|\mathbf{x}_{\text{test}}) \tag{2}$$

encourages consistent predictions across augmented views, and the contrastive alignment term

$$\mathcal{L}_{\text{align}} = \frac{1}{C}\sum_{c=1}^{C}\Big[-\log \frac{\exp(\mathbf{t}_c^\top \mathbf{v}_c)}{\sum_{c'} \exp(\mathbf{t}_c^\top \mathbf{v}_{c'})} - \log \frac{\exp(\mathbf{t}_c^\top \mathbf{v}_c)}{\sum_{c'} \exp(\mathbf{t}_{c'}^\top \mathbf{v}_c)}\Big] \tag{3}$$

draws matched text–visual prototypes together. A single gradient step on this objective yields updated prototypes that better capture the target-domain semantics, improving zero-shot generalization without revisiting source data.

D. Additional Experiments

D.1. Experiment on Time Cost

We evaluated the computational efficiency of COSTA, a critical factor for real-world application. The wall-clock time for various test-time adaptation (TTA) strategies was benchmarked on a single RTX-3090 GPU, with results shown in Table 6. The data shows that COSTA's runtime is higher than that of TDA in AdaND on CUB-200-2011 with the OOD dataset iNaturalist. This is an expected trade-off, as the additional overhead is directly attributable to the core components enabling its advanced performance: (1) multi-prototype adaptation passes, (2) precise metric calculations, and (3) the generation of diverse augmented views.

Most importantly, this computational investment is justified by the outcome. The mechanisms that increase runtime are the very ones that allow COSTA to deliver a new state-of-the-art in performance, achieving results on Acc_S, Acc_N, and Acc_H that are far superior to those of existing methods.

Table 6. Wall-clock time comparison on CUB-200-2011 with OOD dataset iNaturalist.

Method	Time (min)
COSTA (31 augmented views)	171
COSTA (1 augmented view)	87
TDA in AdaND	51

D.2. Experiment on Cache Size

To study the effect of cache size on our method, we conducted an ablation study. On the CUB-200-2011 with iNaturalist pair, we varied the size of one cache while keeping the size of the other cache constant. As shown in Table 7 and 8, we found that continuously increasing the cache size is not necessarily a good choice for our COSTA method. The performance plateaus when the Entropy Cache size is 10 and the ProtoCache size is 60 (Acc_H reaches 72.04%, and further increasing the cache size does not yield significant performance improvement). Based on the analysis of the experiments, we empirically chose to use an Entropy Cache size of 10 and a ProtoCache size of 60 in the COSTA framework. These observations further suggest that our COSTA method is relatively insensitive to the precise values of the cache parameters within a practical range.

Table 7. The effect of different Entropy Cache sizes on Acc_H. Experiments are conducted with CUB-200-2011 as ID and iNaturalist as OOD.

Size	Acc_S (%)	Acc_N (%)	Acc_H (%)
5	57.03	97.34	71.92
10	57.19	97.33	72.04
20	57.17	97.33	72.03

Table 8. The effect of different Proto Cache sizes on Acc_H. Experiments are conducted with CUB-200-2011 as ID and iNaturalist as OOD.

Size	Acc_S (%)	Acc_N (%)	Acc_H (%)
40	57.03	97.39	71.94
60	57.19	97.33	72.04
80	57.18	97.32	72.04

D.3. Experiment on Augmented Views

To investigate the impact of the number of augmented views on the performance of COSTA, we conduct experiments on CUB-200-2011 as the in-distribution (ID) dataset and iNaturalist as the out-of-distribution (OOD) dataset. We vary the number of augmented views per sample from 8 to 64 while keeping all other hyperparameters fixed. Each view is generated via random resized cropping, and the original image is included among the augmented views.

As shown in Table 9, increasing the number of views consistently improves both Acc_S and Acc_H, indicating that multiple augmentations enhance the robustness of cached prototypes and yield more stable confidence estimates. However, the improvement exhibits diminishing returns: Acc_H rises from 70.94% at 8 views to 72.21% at 64 views, but most of the gain is achieved when increasing from 8 to 32 views. Beyond 32 views, further increases contribute marginally ($< 0.2\%$), while computational cost grows linearly with the number of forward passes.

Table 9. Effect of the number of augmented views on COSTA performance. Experiments are conducted with CUB-200-2011 as ID and iNaturalist as OOD.

Views	Acc_S (%)	Acc_N (%)	Acc_H (%)
8	55.68	97.72	70.94
16	56.82	97.33	71.75
32	57.19	97.33	72.04
64	57.39	97.34	72.21

D.4. Ablation on Cache Admission Constraint

Experimental Content: To further validate the necessity of the manifold consistency constraint, we tracked the ID-to-OOD misclassification rate during the test-time adaptation process on the CUB-200-2011 benchmark. We compared the standard COSTA framework against a baseline that ablates Eq. 5, thereby relying solely on the entropy-based confidence score for cache admission

Analysis: In the high-dimensional embedding space of vision-language models like CLIP, out-of-distribution (OOD) samples often exhibit low entropy (i.e., high prediction confidence) despite being geometrically distant from the true in-distribution (ID) manifold. Simple entropy-based filtering is blind to this geometric discrepancy and fails to reject these "confident" but incorrect OOD samples. As shown in Figure 3, without Eq. 5, the cache rapidly accumulates OOD noise, leading to severe prototype drift. By explicitly enforcing geometric compactness through our distance-to-prototype constraint, COSTA effectively establishes a spatial boundary that complements the entropy boundary. Empirically, removing this constraint causes a 1.17% drop in the Acc_H metric, demonstrating that manifold consistency is not merely a heuristic, but a critical

safeguard for maintaining prototype fidelity and driving the collaborative feedback loop.

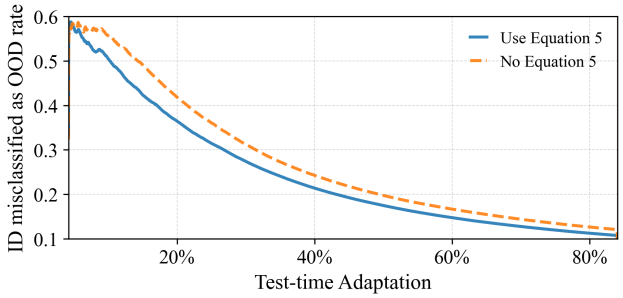

Figure 3. On CUB-200-2011, ID→OOD misclassification rate with vs. without the cache admission constraint during test-time adaptation.

D.5. Robustness under Extreme OOD Ratios

Experimental Content: Standard open-set test-time adaptation benchmarks typically assume a moderate OOD contamination ratio. To evaluate the robustness of our method under extreme noise conditions, we conducted experiments on the Oxford-IIIT Pet dataset by varying the OOD ratio from the standard 25% to a severe 75%. The results are compared against the state-of-the-art open-loop baseline, AdaND, in Table 10.

Analysis: Traditional TTA and open-loop detection methods experience a precipitous drop in performance when OOD samples dominate the data stream. This occurs because an overwhelming influx of noise easily bypasses static or post-hoc filters, hijacking the prototype update process. However, COSTA demonstrates remarkable resilience, maintaining a significant performance margin. Notably, at the 75% extreme noise level, COSTA achieves a +3.40% gain in Acc_H over AdaND (94.76% vs. 91.36%). This robust behavior highlights the advantage of COSTA's dynamic, closed-loop design. Because the Maximum Prototype Similarity (MPS) detector continuously uses dynamically refined, high-fidelity prototypes to filter the incoming stream, it prevents the majority-class OOD samples from pulling the prototype centers away from the true ID manifold.

Table 10. **Robustness under Extreme OOD Ratios (Oxford-Pet).**

Ratio	Method	Acc_S (%)	Acc_N (%)	Acc_H (%)
25%	AdaND	85.39	96.94	90.80
	COSTA	**90.08**$_{+4.69}$	**97.25**$_{+0.31}$	**93.53**$_{+2.73}$
75%	AdaND	85.89	97.59	91.36
	COSTA	**90.66**$_{+4.77}$	**99.26**$_{+1.67}$	**94.76**$_{+3.40}$

D.6. Failure Analysis on Extreme Semantic Shifts

Experimental Content: While COSTA significantly improves robustness across the vast majority of dataset pairs, we observed a narrower performance margin in the specific scenario of ImageNet-V2 (ID) shifting to the Places (OOD) dataset. To understand the boundary of our method's capabilities, we conducted a failure analysis on this extreme semantic shift.

Analysis: The underlying cause of this bottleneck is **Foreground-Background Entanglement**. The CLIP model's pre-training aligns visual features highly with distinct objects (object-centric). ImageNet-V2 consists of prominent foreground objects, while the Places dataset is entirely scene-centric, composed largely of background features. When the distribution shift represents a transition from "Object" to "Scene," the semantic shift is orthogonal to the model's primary feature representation. Consequently, the visual features of ID backgrounds and OOD scenes heavily entangle in CLIP's embedding space. Under such extreme semantic overlap, the geometric separation between ID and OOD prototypes becomes blurred, rendering distance-based and prototype-guided metrics less discriminative. This structural limitation indicates that purely prototype-based OOD detection may struggle when structural or spatial awareness (e.g., distinguishing an object from a background scene) is required, highlighting a valuable direction for future research in open-set test-time adaptation.

E. Theoretical Analysis

E.1. Problem Setting in Open-Set Environments

Unlike standard closed-set TTA, our setting explicitly handles open-set test streams. We formalize the test data stream as a mixture distribution:

$$p_{test}(x) = (1-\gamma)p_{id}(x) + \gamma p_{ood}(x) \tag{4}$$

where $\gamma \in (0,1)$ is the OOD contamination ratio, $p_{id}(x)$ is the in-distribution target distribution, and $p_{ood}(x)$ is the distribution of unknown semantic classes.

The core vulnerability of cache-based ZSTTA is prototype corruption. If an OOD sample is erroneously admitted into the cache, it shifts the estimated prototype away from the true ID manifold, inducing an estimation bias Δ_{ood} bounded by the mass of admitted OOD samples. Thus, the expected classification risk $\mathcal{E}_{open}(f)$ is heavily penalized by the OOD admission rate.

E.2. Theoretical Justification of OOD Rejection

Let A denote the event that a sample satisfies the low-entropy (high-confidence) criterion. Let $\mathcal{R} = \{x : \|g(x) - \mu_c\| \leq d_0\}$ denote the spatial alignment region enforced by

our ProtoCache (and implicitly guided by the MPS OOD detector).

Vulnerability of Entropy-Only Cache. Vision-Language Models like CLIP are known to be overconfident on OOD data. Thus, the probability of an OOD sample exhibiting low entropy, $\mathbb{P}_{ood}(A)$, is non-trivial. In an entropy-only cache of size K, the expected number of admitted OOD samples is:

$$K_{ood} = N\gamma\mathbb{P}_{ood}(A) \tag{5}$$

When $K_{ood} \gg 0$, the cached prototype becomes a convex combination of the ID center and the OOD center (as analyzed in Section 3 of the main paper). The error bound in Proposition 1 is no longer valid because the samples are drawn from the contaminated mixture p_{test} rather than the pure p_{id}, leading to a catastrophic accumulation of the excess risk: $\mathcal{E}(f_{his}) \to \infty$ as adaptation progresses.

Robustness of the Collaborative Dual-Cache. By introducing the alignment region $\mathcal{R}$ around the class prototype μ_c, a sample is admitted if and only if it satisfies both A and $\mathcal{R}$. Because OOD samples are semantically disjoint from the known ID classes, their feature embeddings $g(x)$ are geometrically distant from the ID prototype μ_c with high probability. We formalize this semantic separation as:

$$\mathbb{P}_{ood}(\mathcal{R}) \leq \epsilon_{ood} \to 0 \tag{6}$$

where ϵ_{ood} is a negligibly small constant. The number of OOD samples admitted into the ProtoCache is bounded by:

$$K'_{ood} = N\gamma\mathbb{P}_{ood}(A \cap \mathcal{R}) \leq N\gamma\mathbb{P}_{ood}(\mathcal{R}) \leq N\gamma\epsilon_{ood} \approx 0 \tag{7}$$

Conclusion on Open-Set Error Bound. Since $K'_{ood} \approx 0$, the ProtoCache acts as a strict geometric filter that effectively purifies the mixture distribution $p_{test}(x)$ back to the ID distribution $p_{id}(x)$ inside the cache. Consequently, the prototype estimation bias Δ_{ood} is eliminated. The Dual-Cache framework not only satisfies the tighter risk bound established in Proposition 2 (via the Strong Density Condition of ID samples, $c_a > c_t$), but also safeguards the adaptation process against the open-set risk explosion. This mathematically validates why our collaborative mechanism between prototype learning and semantic alignment (MPS) is strictly necessary for robust Open-Set ZSTTA.

References

[1] Chentao Cao, Zhun Zhong, Zhanke Zhou, Tongliang Liu, Yang Liu, Kun Zhang, and Bo Han. Noisy test-time adaptation in vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[3] Yushu Li, Xun Xu, Yongyi Su, and Kui Jia. On the robustness of open-world test-time training: Self-training with dynamic prototype expansion. In *ICCV*, 2023. 2

[4] Sarah M Pratt, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification, 2023. 1