Appendix

A. Additional Related Works

Difference with Existing Noise Learners: The learnable noise in BadCLIP and *TNT* serves different purposes: the former uses it as a universal trigger for injecting backdoor attack in the model in a supervised few-shot prompt learning setup, whereas *TNT* adapts noise specific to each test sample under a TTA (un-supervised) setup to improve zero-shot generalization.

Visual Prompting *vs* Noise Learning: Noise optimization can be considered analogous to a Visual Prompting; however, there are key distinctions: **(a)** While VPT learns a generic visual prompt from few-shot training data, TNT learns noise specific to each individual test sample, highlighting their application in very different settings. **(b)** VPT operates in the latent space by appending learnable tokens to the input layer of transformer (Eq. 4 of VPT), whereas TNT uses learnable noise tensor (of shape $224 \times 224 \times 3$) directly in the pixel space (Eq. 2 of our work). VPT requires access to inside of the model whereas we do not need such access, making our approach easier to implement.

Furthermore, when applied in the TTA setting, VPT requires a significantly higher number of learnable parameters compared to our approach (TNT), as illustrated in Figure 3 Visual Prompting of our work.

Comparison with Recent Methods: A recent TTA method, TDA [16], differs from our TNT in three distinctive way:

- **Memory-based modules**: TDA employs memory-based modules and avoids prompt tuning, whereas *TNT* does not rely on external memory.
- **Stream-based inference**: TDA requires a continuous stream or batch of test samples for computing dataset statistics, while TNT adapts to a single test sample without dependence on previous samples, depicting *sample-wise adaptation*.
- **Dataset-specific configurations**: TDA necessitates extensive hyperparameter tuning for each dataset. In contrast, *TNT* applies a uniform configuration across all datasets, ensuring robustness and efficiency.

Also, MTA [39] is not entirely training-free as it requires Gaussian kernel optimization for each test sample (Eq. 1 in MTA). While MTA is computationally efficient, TNT achieves better performance with around 3% better generalization on OOD samples as shown in Table 1. Additionally, CoTTA [35], which enforces cross-entropy consistency between student and teacher models, TNT instead minimizes the average distance between **image embeddings** across augmented views, eliminating the need for external models. These comparisons distinguish the TNT components as innovative elements utilized in the zero-shot generalization of VLMs.

Table 4. **Top-1 accuracy % of different methods on ResNet50.** TNT* and TNT outperform baselines across different backbones.

RN50	CLIP	CoOp	TPT	RCLF	TNT*	TNT
ImageNet	58.23	63.35	60.93	61.10	62.73	**65.12**
ImageNet-A	21.47	23.42	26.74	25.98	32.44	**34.26**

Table 5. **Impact of K on ImageNet-A Top-1 accuracy %.** Here K refers to selection of top-K views during inference as in Eq. 7.

K	1	6	12	24	32	64
TNT*	60.23	**61.87**	61.54	60.24	59.02	58.93
TNT	62.77	**63.93**	62.49	60.93	59.48	59.12

Table 6. **Performance of proposed TNT* and TNT on 2 corrupted datasets.** Here, we consistently used $Blur$ corruption with $severity = 5$ consistently.

Corruption Dataset	CLIP	CoOp	TPT	RCLF	TNT*	TNT
CIFAR10-C	82.91	83.22	82.60	82.66	83.25	**83.54**
ImageNet-C	55.74	60.09	56.24	55.86	56.51	**60.29**

Table 7. Impact of Temperature (τ) on TNT* and TNT.

Temperature τ	$9e^{-1}$	$9e^{-2}$	$9e^{-3}$	$9e^{-4}$	$9e^{-5}$
TNT*	59.57	60.14	61.25	**61.87**	61.88
TNT	60.82	61.45	63.02	**63.93**	63.89

B. Implementation Details

Number of Augmentations: All baselines, including TNT, default to using 63 augmentations. We conducted an ablation study (Figure 5(b)) showing that TNT consistently outperforms baselines across different augmentation settings.
Noise Consistency: The same single trainable noise is added across all augmentations for consistency in adaptation. Thus, it is essentially optimizing a single learnable noise during backpropagation.

C. Further Ablations

Evaluation on ResNet50 Backbone: Our proposed TNT* and TNT outperform baselines on ResNet50, as shown in Table 4.
Impact of Hyperparameters (α, β): Across datasets, we set $\alpha = 0.1$ and $\beta = 0.1$, with performance remaining stable for variations in these values $(0.2, ..., 0.5)$.
Impact of K on ImageNet-A: Table 5 shows how performance varies with different values of K.
Evaluation on CIFAR-10-C and ImageNet-C: Table 6 presents our results on corruption datasets.
Impact of Temperature (τ) on TNT* and TNT: Table 7 depicts TNT's performance across different temperatures.