Frustratingly Easy Test-Time Adaptation of Vision-Language Models

Matteo Farina[1,*]   Gianni Franchi[2]   Giovanni Iacca[1]
Massimiliano Mancini[1]   Elisa Ricci[1,3]

[1]University of Trento
[2]U2IS, ENSTA Paris, Institut Polytechnique de Paris
[3]Fondazione Bruno Kessler (FBK)

Abstract

Vision-Language Models seamlessly discriminate among arbitrary semantic categories, yet they still suffer from poor generalization when presented with challenging examples. For this reason, Episodic Test-Time Adaptation (TTA) strategies have recently emerged as powerful techniques to adapt VLMs in the presence of a single unlabeled image. The recent literature on TTA is dominated by the paradigm of prompt tuning by Marginal Entropy Minimization, which, relying on online backpropagation, inevitably slows down inference while increasing memory. In this work, we theoretically investigate the properties of this approach and unveil that a surprisingly strong TTA method lies dormant and hidden within it. We term this approach ZERO (*TTA with "zero" temperature*), whose design is both incredibly effective and frustratingly simple: augment N times, predict, retain the most confident predictions, and marginalize after setting the Softmax temperature to zero. Remarkably, ZERO requires a *single* batched forward pass through the vision encoder only and *no* backward passes. We thoroughly evaluate our approach following the experimental protocol established in the literature and show that ZERO largely surpasses or compares favorably *w.r.t.* the state-of-the-art while being almost $10\times$ faster and $13\times$ more memory friendly than standard Test-Time Prompt Tuning. Thanks to its simplicity and comparatively negligible computation, ZERO can serve as a strong baseline for future work in this field. The code is available.

1   Introduction

Groundbreaking achievements in Vision-Language pretraining [31, 14, 33, 39, 52] have increased the interest in crafting Vision-Language Models (VLMs) that can understand visual content alongside natural language, enabling a new definition of zero-shot classification. Despite huge pretraining databases [34, 37], VLMs still face limitations, suffering from performance degradation in case of large train-test dissimilarity [24] and requiring the design of highly generalizing textual templates [56].

Test-Time Adaptation (TTA) can effectively improve the robustness of VLMs by adapting a given model to online inputs. Among the various TTA setups (such as "fully" [45], "continual" [47] or "practical" TTA [49]), Episodic TTA [53] is particularly appealing, as it focuses on *one-sample* learning problems and requires no assumptions on the distribution of the test data. When presented with a test image $\mathbf{x}$, the parameters θ of a model f are optimized through a TTA objective $\mathcal{L}$ before inferring the final prediction, and reset afterward.

---

[*]Correspondence to: `m.farina@unitn.it`. Code at `https://github.com/FarinaMatteo/zero`.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

The choice of $\mathcal{L}$ is, ultimately, what characterizes TTA methods the most, with the recent literature being dominated by the objective of Marginal Entropy Minimization (MEM) [53]. Given a collection $\mathcal{A}$ of $N \in \mathbb{N}$ data augmentation functions, a test image $\mathbf{x}$ is first augmented N times to obtain a set of different views $X = \{\mathcal{A}_i(\mathbf{x})\}_{i=1}^N = \{\mathbf{x}_i\}_{i=1}^N$. The *marginal probability distribution $\overline{p}$ w.r.t.* sample $\mathbf{x}$ is then defined as the empirical expectation of Softmax-normalized model outputs over X, *i.e.*:

$$\overline{p}(\cdot|\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N p(\cdot|\mathbf{x}_i). \tag{1}$$

Under this framework, the Shannon Entropy of $\overline{p}$ is a bona fide measure of how *inconsistently* and *uncertainly* the model predicts over X, making it a tantalizing candidate to minimize, *i.e.*:

$$\mathcal{L}_{ent} = H(\overline{p}(\cdot|\mathbf{x})) = -\sum_{c=1}^C \overline{p}(y=c|\mathbf{x}) \log(\overline{p}(y=c|\mathbf{x})), \tag{2}$$

where C is the number of semantic categories. Once $\mathcal{L}_{ent}$ is computed, (some of) the parameters of f are typically updated for a few steps of Gradient Descent before inferring the final prediction over the source input $\mathbf{x}$ with updated parameters. Owing to its simplicity and effectiveness, MEM has become a *de facto* standard in modern TTA [53, 38, 33, 19, 42, 27].

In this work, we take the opposite direction and challenge this paradigm. By conducting an in-depth theoretical and empirical investigation, we find that: ① while effective in improving model robustness, MEM has *little effect* on the prediction of $\overline{p}$; ② no matter the dataset, the label space, or the parameter initialization, VLMs become much better classifiers when $\overline{p}$ replaces the standard inference protocol. Building on these insights, we show that a surprisingly strong and optimization-free TTA baseline is subtly hidden within the MEM framework. We term this baseline ZERO, which is short for TTA with "**zero**" temperature. Instead of tuning any parameters, setting to zero the Softmax temperature before marginalizing over views makes $\overline{p}$ already stronger than the model after MEM. Notably, ZERO only requires a single forward pass through the vision encoder and no backward passes.

Wrapping up, the contributions of this paper are the following:

1. We theoretically show *when* the prediction obtained through $\overline{p}$ (*i.e.*, $\arg\max \overline{p}$) is *invariant* to MEM, and empirically verify that MEM has largely *no effect* on $\arg\max \overline{p}$;

2. We theoretically and empirically demonstrate that the error rate of $\overline{p}$ is a lower bound to the base error of a VLM in the setup of TTA. Additionally, we identify augmentations-induced *overconfidence* as the primal factor undermining the reliability of $\overline{p}$;

3. Motivated by these theoretical insights, we introduce ZERO, a frustratingly simple TTA approach that recovers the reliability of $\overline{p}$ by tweaking a single parameter of the model: the temperature;

4. We thoroughly evaluate ZERO following the established experimental setup with a variety of model initializations. Our results show that ZERO surpasses or compares favorably to state-of-the-art TTA methods while being much faster and more memory efficient (*e.g.*, $10\times$ faster and $13\times$ more memory efficient than the established Test-Time Prompt Tuning [38]).

2 Understanding Marginal Entropy Minimization

In this Section, we take a step towards both theoretically and empirically understanding the paradigm of MEM. In particular, this section is devoted to answering the following research questions:

1. *How does MEM affect the marginal probability distribution?* And, in turn

2. *How does the marginal probability distribution relate to the standard inference protocol?*

First, we introduce MEM for VLMs by reviewing the established Test-Time Prompt Tuning (TPT) method [38] and its notation. Then, in Sections 2.2 and 2.3 we answer the research questions above.

2.1 Preliminaries

Zero-Shot Classification with VLMs employs a predefined template (*e.g.*, "`a photo of a`") from which a set of context vectors $\mathbf{t}_{\text{ctx}}$ is obtained by looking up a token embedding table. Expanding

2

the template with the class names (*e.g.*, "`a photo of a laptop.`" for the class "`laptop`") makes up the entire set of input vectors $[\mathbf{t}_{ctx}, \mathbf{t}_1], \ldots, [\mathbf{t}_{ctx}, \mathbf{t}_C]$, with $\mathbf{t}_i$ being the embeddings derived from the i-th class name. A text encoder $\mathbf{E}_{txt}$ transforms these class descriptions into independent and normalized text embeddings $\mathbf{z}_1^{txt}, \ldots, \mathbf{z}_C^{txt}$ and an image encoder $\mathbf{E}_{img}$ encodes an input image $\mathbf{x}$ into a normalized latent vector $\mathbf{z}^{img}$. Lastly, classification is carried out by picking the class c corresponding to the text embedding $\mathbf{z}_c^{txt}$ holding the maximum cosine similarity with $\mathbf{z}^{img}$.

MEM for VLMs. Pioneered by MEMO [53] in the scope of unimodal neural networks, MEM was re-purposed for TTA with VLMs by Test-Time Prompt Tuning [38]. In [38], a VLM such as CLIP [31] is adapted at test time by minimizing the same objective of Eq. (2). In contrast to optimizing all model parameters, TPT relies on the effectiveness of prompt tuning [56, 55, 15], optimizing only the context vectors derived from the token embeddings of the standard CLIP template "`a photo of a`". By explicitly enunciating the dependency on the context vectors $\mathbf{t}_{ctx}$ and re-using the notation of Sec. 1, one can re-write the MEM objective of [38] as:

$$\mathcal{L}_{ent} = H(\overline{p}(\cdot|\mathbf{x}, \mathbf{t}_{ctx})) = -\sum_c^C \overline{p}(y = c|\mathbf{x}, \mathbf{t}_{ctx}, \tau) \log\left(\overline{p}(y = c|\mathbf{x}, \mathbf{t}_{ctx}, \tau)\right)$$

$$\text{where } \overline{p}(y = c|\mathbf{x}, \mathbf{t}_{ctx}, \tau) = \frac{1}{N}\sum_i^N \frac{\exp\left(\mathbf{z}_i^{img} \cdot \mathbf{z}_c^{txt}(\mathbf{t}_{ctx})/\tau\right)}{\sum_k^C \exp\left(\mathbf{z}_i^{img} \cdot \mathbf{z}_k^{txt}(\mathbf{t}_{ctx})/\tau\right)}. \tag{3}$$

Here, τ is the *temperature* of the Softmax operator. In the rest of this section, we omit the dependency on τ for simplicity, writing $p(\cdot|\mathbf{x}, \mathbf{t}_{ctx})$. Similarly to [53], the objective of Eq. (3) is minimized for a single step of Gradient Descent to update the set of context vectors. The updated context vectors, denoted as $\mathbf{t}_{ctx}^*$, are then used to prompt the VLM and obtain the final prediction for $\mathbf{x}$. For any class c this is simply $\mathbf{z}_{\mathbf{x}}^{img} \cdot \mathbf{z}_c^{txt}(\mathbf{t}_{ctx}^*)$, which is easily transformed into $p(y = c|\mathbf{x}, \mathbf{t}_{ctx}^*)$ via Softmax.

2.2 How does MEM affect the marginal probability distribution?

The recent literature on TTA shows that minimizing $\mathcal{L}_{ent}$ significantly enhances the robustness of model outputs. However, the impact of this process on the marginal probability distribution $\overline{p}$ remains unclear. We start with a straightforward hypothesis: due to its nature, minimizing $\mathcal{L}_{ent}$ tends to increase the probability of the most probable class of $\overline{p}(\cdot|\mathbf{x}, \mathbf{t}_{ctx})$. More formally, denoting with $\hat{c}$ the prediction of $\overline{p}$ (*i.e.*, $\hat{c} = \arg\max \overline{p}(\cdot|\mathbf{x}, \mathbf{t}_{ctx})$), we hypothesize that $\overline{p}(y = \hat{c}|\mathbf{x}, \mathbf{t}_{ctx}^*) > \overline{p}(y = \hat{c}|\mathbf{x}, \mathbf{t}_{ctx})$. If this hypothesis is realized, it comes as a natural consequence that minimizing $\mathcal{L}_{ent}$ is unlikely to alter the prevailing class of $\overline{p}$, thus resulting in a consistent prediction pre- and post-TTA where $\arg\max \overline{p}(\cdot|\mathbf{x}, \mathbf{t}_{ctx}) = \arg\max \overline{p}(\cdot|\mathbf{x}, \mathbf{t}_{ctx}^*)$.

Hence, the first contribution of this work is to show that the prediction of the marginal probability distribution $\overline{p}$ is *invariant* to Entropy Minimization under loose constraints on confidence and gradients. To lighten the notation of the proposition, let us first define the following function g:

$$g(c, \mathbf{z}^{img}, \mathbf{z}_1^{txt}, \ldots, \mathbf{z}_C^{txt}) = \frac{\exp\left(\mathbf{z}^{img} \cdot \mathbf{z}_c^{txt}/\tau\right)}{\sum_k^C \exp\left(\mathbf{z}^{img} \cdot \mathbf{z}_k^{txt}/\tau\right)} \tag{4}$$

i.e., the probability assigned to class c given a latent image representation $\mathbf{z}^{img}$ and class-wise text embeddings $\mathbf{z}_1^{txt}, \ldots, \mathbf{z}_C^{txt}$. Additionally, let $\delta g(c, \mathbf{z}^{img})$ be the negative variation incurred to the function g when the context vectors $\mathbf{t}_{ctx}$ are updated through Entropy Minimization:

$$\delta g(c, \mathbf{z}^{img}) = g(c, \mathbf{z}^{img}, \mathbf{z}_1^{txt}(\mathbf{t}_{ctx}), \ldots, \mathbf{z}_C^{txt}(\mathbf{t}_{ctx})) - g(c, \mathbf{z}^{img}, \mathbf{z}_1^{txt}(\mathbf{t}_{ctx}^*), \ldots, \mathbf{z}_C^{txt}(\mathbf{t}_{ctx}^*)) \tag{5}$$

where, for clarity, the dependency of the text embeddings $\mathbf{z}_1^{txt}, \ldots, \mathbf{z}_C^{txt}$ on the context vectors (either $\mathbf{t}_{ctx}$ or $\mathbf{t}_{ctx}^*$) is explicit. Using this notation, we can formalize the following proposition:

Proposition 2.1. *Let $\mathbf{z}_1^{img}, \ldots, \mathbf{z}_N^{img}$ be the latent image representations resulting from the N views and $\hat{c} = \arg\max \overline{p}(\cdot|\mathbf{x}, \mathbf{t}_{ctx})$ be the initial prediction of the marginal probability distribution. If the entropy of $\overline{p}$ is minimized and $\overline{p}(y = \hat{c}|\mathbf{x}, \mathbf{t}_{ctx}) > \frac{1}{N}\sum_{i=1}^N \delta g(\hat{c}, \mathbf{z}_i^{img})$ then the prevalent class of $\overline{p}$ is invariant to MEM, i.e., $\arg\max \overline{p}(\cdot|\mathbf{x}, \mathbf{t}_{ctx}) = \arg\max \overline{p}(\cdot|\mathbf{x}, \mathbf{t}_{ctx}^*)$.*

In Appendix A, we provide a detailed proof of this proposition, highlighting that $\delta(\hat{c}, \mathbf{z}^{img})$ is directly linked to the gradient *w.r.t.* the context vectors $\mathbf{t}_{ctx}$. This relationship emerges when writing any post-update text embedding $\mathbf{z}_c^{txt}(\mathbf{t}_{ctx}^*)$ as a function of its pre-update counterpart $\mathbf{z}_c^{txt}(\mathbf{t}_{ctx})$. Specifically,

3

we can write $\mathbf{z}_c^{\text{txt}}(\mathbf{t}_{\text{ctx}}^*) = \mathbf{E}_{\text{txt}}([\mathbf{t}_{\text{ctx}} - \lambda\nabla_{\mathbf{t}_{\text{ctx}}}(\mathcal{L}_{ent}), \mathbf{t}_c])$, which is equivalent to $\mathbf{E}_{\text{txt}}([\mathbf{t}_{\text{ctx}}, \mathbf{t}_c]) - \lambda\nabla_{\mathbf{t}_{\text{ctx}}}\mathbf{E}_{\text{txt}}([\mathbf{t}_{\text{ctx}}^*, \mathbf{t}_c])^t\nabla_{\mathbf{t}_{\text{ctx}}}(\mathcal{L}_{ent})$ after a first-order Taylor Expansion around $\mathbf{t}_{\text{ctx}}^*$. Consequently, the proposition holds by a condition relating confidence (through $\overline{p}(y = c|\mathbf{x}, \mathbf{t}_{\text{ctx}})$) and gradients (through $\delta(\hat{c}, \mathbf{z}^{\text{img}})$). Alongside the proof, Appendix A presents evidence supporting this proposition for CLIP [31] on the ImageNet-1k validation set [4], as well as across various datasets for natural distribution shifts: ImageNet-A [13], ImageNet-R [12], ImageNet-v2 [32], and ImageNet-Sketch [46].

2.3 How does $\overline{p}$ relate to the standard inference protocol?

From prior work on Test-Time Augmentations (TTAug) with unimodal neural networks [40, 35], empirical evidence suggests that $\overline{p}(\cdot|\mathbf{x})$ is more robust than $p(\cdot|\mathbf{x})$. This observation leads to the hypothesis that the expected risk of predicting with $\overline{p}$ is lower than that of doing so with p. However, the literature lacks guarantees for this hypothesis, except for the peculiar case in which the risk function is the squared error, *i.e.*, $\ell(a, b) = (a - b)^2$ [16].[2]

As the second contribution of this study, we show that the error rate of $\overline{p}(\cdot|\mathbf{x})$ does indeed lower-bound the error rate of $p(\cdot|\mathbf{x})$. We do so by revisiting the theory of model ensembling, and showing that analogous ideas can emerge for TTA.

Preliminaries on model ensembling. From the theory of classifier ensembling [18], we know that if $f_1, \ldots, f_N$ are $N \in \mathbb{N}$ independent classifiers with error rate ϵ and $\mathbf{x}$ is an example whose label is $y \in \{0, 1\}$, then the probability that any group of k classifiers picks the same *wrong* label $f_i(\mathbf{x}) = \hat{y} \neq y$ can be expressed with a Binomial distribution wrapping N Bernoulli processes:

$$P_{\hat{y}\neq y}(k) = \binom{N}{k}\epsilon^k(1 - \epsilon)^{(N-k)} \tag{6}$$

Revisiting model ensembling for TTA. Eq. (6) holds as long as all events modeled as Bernoulli processes are independent. Thus, we have an equivalent error estimate for the setup in which only a single classifier f is present and $X_y = \{\mathbf{x}_i\}_{i=1}^N$ is a set of independent examples with the same underlying label y. Within this framework, any group of k examples in X_y to which the classifier has assigned the same label $\hat{y}$ is also a set of independent Bernoulli processes, whose error probability is still quantified via Eq. (6). Note that this resembles the TTA setup in the presence of N views of the source sample $\mathbf{x}$, as long as augmentations do not change their underlying labels. We refer the reader to Appendix H for a discussion about the independence assumption among different views.

$\overline{p}$ **is better than** p **(if** f **is calibrated).** The final step can be taken through the lens of *model calibration* [8], a property requiring that the confidence of a classifier matches its accuracy. For example, a calibrated classifier f whose confidence is 0.7 is expected to be correct 70% of the times. In the previous discussion, if we denote with $k(y)$ the number of examples correctly labeled as y, then the accuracy of the classifier is exactly $k(y)/N$. It follows that there is a positive correlation between accuracy and confidence if f exhibits good calibration, i.e., $\uparrow k(y)/N \implies \uparrow \overline{p}(y)$. Thus, the probability of picking the wrong class with this marginal probability is approximated by Eq. (6). Given this relationship, we have that $\overline{p}(y) = \max \overline{p}(\cdot)$ if $k(y)$ matches or exceeds the majority within N. Thus, the probability of picking the wrong class with $\overline{p}$ is approximated by marginalizing out all values of k that satisfy this criterion, which entails that the error or $\overline{p}$ can be expressed with the cumulative distribution of (6):

$$P_{\hat{y}\neq y}(\overline{p}) = \sum_{k=\lfloor N/2+1 \rfloor}^{N} \binom{N}{k}\epsilon^k(1 - \epsilon)^{(N-k)} \tag{7}$$

From the Condorcet Jury Theorem [36], we know that Eq. (7) is a *monotonically decreasing function* if the error ϵ is better than random guessing, which is likely to be the case for VLMs pretrained on a massive amount of web data such as CLIP. Hence, we conclude that the error of $\overline{p}$ is a realistic lower bound for the base model error ϵ *over a set of independent data points sharing the same label*.

Does this lower bound empirically realize? We evaluate if the error of $\overline{p}$ consistently lower bounds the error of p also in practical use cases, where model calibration is unknown and the label space is large. For this, we use `CLIP-ViT-B-16` [5], the ImageNet validation set, and four datasets reflecting

---

[2]In Appendix D, we show that this bound generalizes to any function ℓ satisfying the triangular inequality.

4

(a) Error of $\overline{p}$ vs $\epsilon(y)$.   (b) Reliability diagrams for IN-1k and its augmented version.

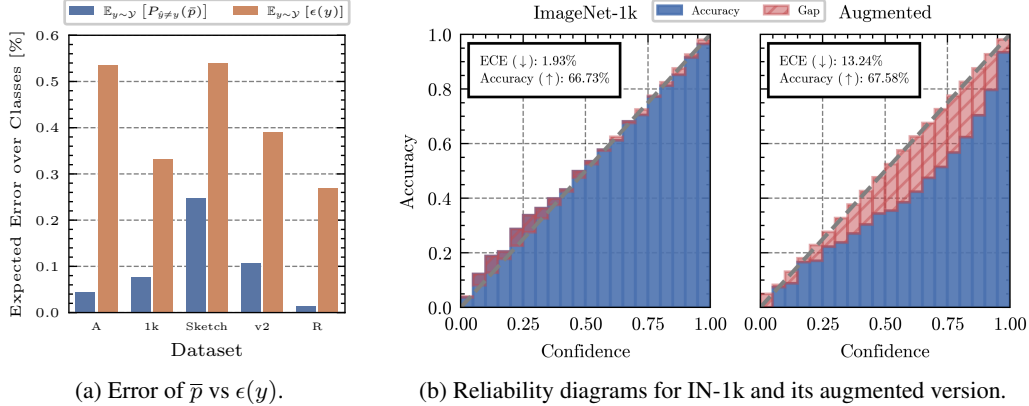



Figure 1: Motivating findings. (a) Comparison between the expected error of `CLIP-ViT-B-16`, denoted as $\epsilon(y)$, and the error of the marginal probability distribution obtained by marginalizing over examples with the same label, $P_{\hat{y} \neq y}(\overline{p})$; (b) Reliability diagrams of `CLIP-ViT-B-16` on the ImageNet validation set (left), and its augmented version (right), showing that augmentations largely un-calibrate CLIP exclusively due to overconfidence while leading to slightly better overall accuracy.

Natural Distribution Shifts [12, 13, 32, 46]. For all classes in each dataset, we first draw all images sharing the same label (X_y). Then, we compute the expected error $\epsilon(y)$ of the model on this subset, together with the error of $\overline{p}$ (ideally, Eq. (6)). Lastly, we average these errors over the entire label space $\mathcal{Y}$. We do <u>not</u> restrict to the cases where y is supported by the majority and we do <u>not</u> re-organize predictions in a *one-versus-all* scheme. Fig. 1(a) clearly shows that the error of $\overline{p}$ is a lower bound to the base error of the model also in practical use cases where the label space is large and guarantees on model calibration are possibly missing. Importantly, this phenomenon persists *no matter the dataset*.

3   Simple and surprisingly strong TTA (for free)

The main point of Section 2.2 is that MEM generally does not affect the predominant class of the marginal probability distribution $\overline{p}$. On the other hand, from Section 2.3 one can conclude that through $\overline{p}$ the model becomes a much stronger classifier. Summarizing:

$$\text{From Section 2.2: } \arg\max\left(\overline{p}(\cdot|\mathbf{x}, \mathbf{t}_{\text{ctx}})\right) = \arg\max\left(\overline{p}(\cdot|\mathbf{x}, \mathbf{t}_{\text{ctx}}^*)\right)$$

$$\text{From Section 2.3: } \begin{cases} P_{\hat{y} \neq y}(\overline{p}(\cdot|\mathbf{x}, \mathbf{t}_{\text{ctx}})) \leq P_{\hat{y} \neq y}(p(\cdot|\mathbf{x}, \mathbf{t}_{\text{ctx}})) \\ \text{and, equivalently} \\ P_{\hat{y} \neq y}(\overline{p}(\cdot|\mathbf{x}, \mathbf{t}_{\text{ctx}}^*)) \leq P_{\hat{y} \neq y}(p(\cdot|\mathbf{x}, \mathbf{t}_{\text{ctx}}^*)) \end{cases} \tag{8}$$

Chaining observations together, it emerges that:

$$P_{\hat{y} \neq y}(p(\cdot|\mathbf{x}, \mathbf{t}_{\text{ctx}}^*)) \geq P_{\hat{y} \neq y}(\overline{p}(\cdot|\mathbf{x}, \mathbf{t}_{\text{ctx}}^*)) = P_{\hat{y} \neq y}(\overline{p}(\cdot|\mathbf{x}, \mathbf{t}_{\text{ctx}})) \tag{9}$$

i.e., if all assumptions are met, the error of MEM $\geq$ error of $\overline{p}$ after MEM = error of $\overline{p}$ *without* MEM. All in all, this TTA framework is hiding a surprisingly strong and optimization-free baseline: $\overline{p}$! Next, we highlight the detrimental impact of data augmentations on this marginal probability distribution and introduce a simple trick to recover its reliability: *zeroing-out* the Softmax temperature.

3.1   Augmentations undermine the reliability of $\overline{p}$

While augmentations are essential in TTA to obtain multiple views of the test instance, noisy views may constitute Out-of-Distribution (OOD) data, thus having the undesired effect of un-calibrating the model. To sidestep this issue, one can attempt to discriminate between in-distribution (*w.r.t.* to the pretraining data) and OOD views. Given that low confidence is a common trait in OOD data, a viable way to discriminate is confidence-based filtering, such as in TPT [38]. Formally, a smaller set of confident views are obtained following $X_{filt} = \{\mathbf{x}_i \in X | H(p(\cdot|\mathbf{x}_i, \mathbf{t}_{\text{ctx}})) < \rho\}$, where ρ is a

5

threshold retaining the views whose entropy is in the bottom-10% percentile (lowest entropy). Despite its effectiveness, this filter cannot help when the reliability of $\bar{p}$ is undermined by *overconfidence*.

Augmentations lead to poor calibration. We demonstrate the impact of augmentation-induced overconfidence using the same model and datasets of Section 2.3. For each dataset, we generate an augmented counterpart following the augmentation and filtering setup of TPT [38], *i.e.*: we augment an input $N = 64$ times using simple random resized crops and horizontal flips. Then, we only retain 10% of the N views according to confidence-based filtering, resulting in 6 views per sample. Consequently, each augmented dataset contains $6\times$ more data points than its plain counterpart. The Expected Calibration Error (ECE) [8] reported in Appendix C conveys that ① zero-shot CLIP is well-calibrated (ECE < 0.1 for all datasets), strongly supporting the theory of Section 2.3 and ② *the augmented visual space greatly increases the calibration error*.

Poor calibration is frequently linked to overconfidence. We investigate the reason for the increase in ECE by presenting reliability diagrams for the ImageNet validation set in Fig. 1(b). In a reliability diagram, every bar below the identity line $y = x$ signals overconfidence (*i.e.*, the confidence on the x-axis prevails over the accuracy on the y-axis), while the opposite signals under-confidence. Notably, in the scope of our experiments, overconfidence is the primal factor leading to an increase in the ECE. The error rate, in contrast, decreases slightly. In Appendix C, we also experiment across all datasets for Natural Distribution Shifts and different CLIP models pretrained on the 2B subset of LAION [2, 34]. Importantly, this phenomenon further persists within this extended experimental suite.

3.2   ZERO: Test-Time Adaptation with "zero" temperature

Since its reliability is severely undermined by augmentations-induced overconfidence, directly predicting through $\bar{p}$ is not an enticing baseline for TTA. Concurrently, we also know that the error rate does not decrease when predicting over the augmented visual space. Hence, we are interested in finding an efficient way to capitalize on these observations: relying on the predictions over the views, while ignoring potentially misleading confidence information. The key is to note that both desiderata are obtained by explicitly tweaking a single parameter of the model: *the temperature*. Specifically, setting the temperature to (the limit of) zero corresponds to converting probability distributions into one-hot encodings, hence exclusively relying on their $\arg\max$ when marginalizing. Inspired by this idea we propose ZERO, Test-Time Adaptation with "**zero**" temperature.

Procedure. ZERO follows these simple steps: ① augment, ② predict, ③ retain the most confident predictions, ④ set the Softmax temperature to zero and ⑤ marginalize. The final prediction is the $\arg\max$ of the marginal probability distribution computed after "zeroing-out" the temperature, *i.e.*:

$$\text{ZERO}(\mathbf{x}, \mathbf{t}_{\text{ctx}}, C) = \underset{c \in [1,\dots,C]}{\arg\max} \left(\sum_{i=1}^{N} \mathbb{1}(\mathbf{x}_i \in X_{filt}) \lim_{\tau \to 0^+} p(y = c | \mathbf{x}_i, \mathbf{t}_{\text{ctx}}, \tau) \right), \qquad (10)$$

where $\mathbb{1}$ is an indicator function, whose output is 1 if $\mathbf{x}_i \in X_{filt}$ and 0 otherwise, and X_{filt} is the set of confident views *before* tweaking the temperature, *i.e.*, $\mathbf{x}_i \in X_{filt}$ if $H(p(\cdot | \mathbf{x}_i, \mathbf{t}_{\text{ctx}}, \tau) < \rho$.

Efficient Implementation. In all its simplicity, ZERO is computationally lightweight. In closed set assumptions where the class descriptions (and thus their embeddings) are fixed, ZERO only requires a single batched forward pass through the vision encoder, just as much as needed to forward the N views. Additionally, since the temperature is explicitly tweaked, ZERO needs *no backpropagation at all* and can be implemented in a few lines of code. For reference, a PyTorch-like implementation [30] is reported in Algorithm 1.

Algorithm 1 PyTorch-style code for ZERO

```
# z_txt = pre-computed text embeddings (C,hdim)
# temp = model's original temp
# augment = takes (C,H,W) and returns (N,C,H,W)
# gamma = filtering percentile (e.g., 0.1)
def zero(image, z_txt, N, gamma, temp):
  # step1: augment
  views = augment(image, num_views=N)
  # step2: predict (unscaled logits)
  l = model.image_encoder(views) @ z_txt.t()
  # step3: retain most confident preds
  l_filt = confidence_filter(l, temp, top=gamma)
  # step4: zero temperature
  zero_temp = torch.finfo(l_filt.dtype).eps
  # step5: marginalize
  p_bar = (l_filt / zero_temp).softmax(1).sum(0)
  return p_bar.argmax()
```

Equivalent perspective and final remark. We bring to attention a simple scheme which corresponds to ZERO: *voting* over (confident) augmentations. Drawing from the theory of ensembling, note that the error rate of the voting paradigm is exactly described by Eq. (6). Essentially,

6

Table 1: Natural Distribution Shifts. TTA methods are grouped according to the baseline model and top-1 accuracy is reported. **Bold text** is the best method within each group.

Method	ImageNet	A	V2	R	Sketch	Mean
CLIP-ViT-B-16						
Zero-Shot	66.73	47.87	60.86	73.98	46.09	59.11
Ensemble	68.34	49.89	61.88	77.65	48.24	61.20
TPT	68.98	54.77	63.45	77.06	47.94	62.44
ZERO	69.31±0.13	59.61±0.19	64.16±0.03	77.22±0.05	48.40±0.07	63.74
ZERO+Ensemble	**71.17**±0.06	**62.75**±0.14	**65.23**±0.08	**80.75**±0.02	**50.59**±0.08	**66.10**
MaPLe						
Zero-Shot	-	50.90	64.07	76.98	49.15	60.28
TPT	-	58.08	64.87	78.12	48.16	62.31
PromptAlign	-	59.37	65.29	79.33	50.23	63.55
ZERO	-	**63.32**±0.26	**66.81**±0.43	**79.74**±0.32	**51.07**±0.47	**65.23**
CLIP-ViT-B-16 + CLIP-ViT-L-14						
Zero-Shot	73.44	68.82	67.80	85.40	57.84	70.66
RLCF $_{t=3}^{\mathbf{t}_{ctx}}$	73.23	65.45	69.77	83.35	54.74	69.31
RLCF $_{t=3}^{\Theta_v}$	74.85	73.71	69.77	86.19	57.10	72.32
ZERO	**75.52**±0.03	**75.15**±0.26	**70.37**±0.05	**87.21**±0.09	**59.61**±0.04	**73.57**

this means that ZERO capitalizes on the theoretical insights while circumventing practical issues stemming from augmentations. We also highlight that ZERO is subtly hidden within any TTA framework relying exclusively on MEM, since computing $\bar{p}$ is inevitable therein. For this reason, we refer to ZERO as a *baseline* for TTA. Our goal diverges from introducing a "novel" state-of-the-art method for TTA. In contrast, we advocate the importance of evaluating simple baselines.

4 Experiments

In this section, we present a comprehensive experimental evaluation of ZERO. Similarly to [38, 33, 54], we always work in the setup of *single test point* adaptation. Our results show that ZERO, alongside its simplicity, is an effective and efficient approach for TTA.

4.1 Experimental Protocol

Baselines. We compare ZERO to three strategies for TTA with VLMs: ① TPT [38], ② PromptAlign [33], and ③ Reinforcement Learning from CLIP Feedback (RLCF) [54]. As introduced in Section 2, TPT works by minimizing the entropy of $\bar{p}$. In contrast, PromptAlign relies on a pretrained MaPLe initialization [15] and pairs the MEM objective with a distribution alignment loss between layer-wise statistics encountered online and pretraining statistics computed offline. Finally, RLCF does not include MEM in its framework; Zhao et al. [54] shows that, if rewarded with feedback from a stronger teacher such as CLIP-ViT-L-14, the smaller CLIP-ViT-B-16 can surpass the teacher itself.

Models. As different approaches consider different backbones in the original papers, we construct different comparison groups to ensure fair comparisons with all TTA baselines [38, 33, 54].

Group 1: When comparing to TPT, we always use CLIP-ViT-B-16. Shu et al. [38] also reports CLIP-Ensemble, *i.e.*, CLIP enriched with an ensemble of hand-crafted prompts. While the design of TPT does not allow leveraging text ensembles (as also pointed out by concurrent work [42]), ZERO seamlessly integrates with CLIP-Ensemble. We denote this variant with ZERO+Ensemble.

Group 2: When comparing to PromptAlign, we follow Samadh et al. [33] and start from a MaPLe initialization for a fair comparison. MaPLe prompts are learned on ImageNet, following [33]. Within this group, we also report TPT on top of MaPLe, as in [33].

Group 3: When comparing to RLCF, we use both CLIP-ViT-B-16 and CLIP-ViT-L-14 as in [54]. Specifically, confidence-based filtering acts on top of the output of the first model, and the selected inputs are passed to the second model for the final output. Both forward passes are inevitable in RLCF, so this scheme corresponds to "early-exiting" the pipeline, exactly as per MEM. RLCF can vary according to (i) the parameter group being optimized and (ii) the number of adaptation steps. We denote with Θ_v the full image encoder tuning, with $\mathbf{t}_{ctx}$ prompt tuning, and with t the number

7

of adaptation steps. For example, RLCF $_{t=3}^{t_{ctx}}$ indicates RLCF with prompt tuning for 3 TTA steps. Note that, since all methods need to forward more than one image to the teacher model, the zero-shot baseline of this group is exactly zero-shot classification with `CLIP-ViT-L-14`.

Pretrainings. This Section deals with models officially released by OpenAI [28]. Appendix B further reports experiments with LAION-pretrained CLIP models [2], as well as the soft prompt initialization with supervised Context Optimization (`CoOp`) from [56].

Benchmarks. We follow the established experimental setup of [38, 33], evaluating ZERO on Natural Distribution Shifts and Fine-grained Classification (also referred to as "Cross-Datasets Generalization" in previous works). For the former, we consider the ImageNet validation set and the four datasets for Natural Distribution Shifts already presented in Section 2, commonly considered Out-of-Distribution (OOD) datasets for CLIP. For fine-grained classification, we evaluate all TTA methods on 10 datasets. Specifically, we experiment with Oxford-Flowers (FWLR) [25], Describable Textures (DTD) [3], Oxford-Pets (PETS) [29], Stanford Cars (CARS) [17], UCF101 (UCF) [41], Caltech101 (CAL)[6], Food101 (FOOD) [1], SUN397 (SUN)[48], FGVC-Aircraft (AIR) [23] and EuroSAT (ESAT) [11]. For all of these datasets, we refer to the test split in Zhou et al. [56] as per the common protocol.

Textual prompts. When `+Ensemble` is specified, we do *not* use dataset-specific templates. In contrast, we use the set of 7 generic templates highlighted in the official CLIP repository [28] across all datasets. When adapting `MaPLe`, we stick to the ImageNet-learned prompts released by [15] and evaluate them cross-datasets as in [33].

Implementation Details. The augmentation pool $\mathcal{A}$ only contains random resized crops and random horizontal flips. The only hyperparameter of ZERO is the percentile for confidence-based filtering, which is set to 0.3 after validation on ImageNet (following standard practice [51]) and kept fixed *for all datasets*. We inherit the setup of TPT with $N = 64$, crafting 63 augmentations to collate with the source image. To ensure hardware differences do not play any role in comparisons, we execute all TTA methods under the same hardware setup by running the source code of each repository with no modifications. We always use 1 NVIDIA A100 GPU and FP16 Automatic Mixed Precision. Results are averaged over 3 different seeds. Unless otherwise specified, all tables report top-1 accuracy.

4.2 Results

Natural Distribution Shifts. Results for Natural Distribution Shifts are reported in Table 1.

Group 1 (TPT): ZERO *surpasses TPT consistently on all datasets.* Among OOD datasets, peak difference is reached with ImageNet-A, where ZERO outperforms TPT by $+4.84\%$. Enriching ZERO with hand-crafted prompts improves results further, with an average margin of $+3.66\%$ *w.r.t.* TPT.

Group 2 (PromptAlign): Within the second comparison group, ZERO *outperforms PromptAlign on all datasets*, with $+1.68\%$ being the gap in average performance. ZERO consistently outperforms TPT also when the baseline initialization is `MaPLe` (by an average of $+2.92\%$). Please note that we omit evaluation on ImageNet for this group, since PromptAlign adopts token-level statistics from this dataset when adapting to test points, which would render the comparison unfair. For completeness, we report that zero-shot `MaPLe` achieves an accuracy of 70.72% on ImageNet, which is improved to 72.99% by adapting with ZERO ($+2.27\%$).

Group 3 (RLCF): We follow [54] and report RLCF variants with $t = 3$ steps. In this group, ZERO outperforms RLCF in 5 out of 5 datasets, with a gap in the average performance of $+1.25\%$. Importantly, RLCF is only close to ZERO with image encoder tuning; only prompt tuning is insufficient.

Fine-grained Classification. Results for fine-grained classification are shown in Table 2. To foster readability, the standard deviations of ZERO are separately reported in Table 11 (Appendix).

Group 1 (TPT): Default ZERO improves over the zero-shot baseline `CLIP-ViT-B-16`, but is outperformed by TPT with an average margin of -0.57%. However, extending ZERO with hand-crafted prompts (something that TPT cannot do *by design*) is sufficient to outperform TPT on 7 out of 10 datasets, and obtain an average improvement of $+0.74\%$.

Group 2 (PromptAlign): On average, PromptAlign has an improvement of $+0.5\%$ over ZERO. However, note that this is mostly influenced by the performance on one dataset only (EuroSAT) and that, in contrast, ZERO *surpasses PromptAlign in 7 out of 10 datasets.* In line with the previous benchmark, ZERO better adapts `MaPLe` than TPT, again outperforming it in 7 out of 10 datasets.

8

Table 2: Fine-grained classification. TTA methods are grouped according to the reference baseline, top-1 accuracy is reported and **bold text** indicates the best performer of each group.

Method	FLWR	DTD	PETS	CARS	UCF	CAL	FOOD	SUN	AIR	ESAT	Mean	Median
CLIP-ViT-B-16												
Zero-Shot	67.44	44.27	88.25	65.48	65.13	93.35	83.65	62.59	23.67	42.01	63.58	65.31
Ensemble	67.07	45.09	**88.28**	66.16	67.51	93.91	84.04	66.26	23.22	**50.42**	65.20	66.66
TPT	**68.75**	**47.04**	87.23	66.68	68.16	93.93	84.67	65.39	23.13	42.86	64.78	67.42
ZERO	67.68	46.12	87.75	68.04	67.77	93.66	86.53	65.03	**25.21**	34.33	64.21	67.72
ZERO$_{+Ensemble}$	67.17	45.86	87.83	**68.97**	**69.18**	**94.41**	**86.77**	**67.63**	**25.21**	42.17	**65.52**	**68.30**
MaPLe												
Zero-Shot	72.23	46.49	90.49	65.57	68.69	93.53	86.20	67.01	24.74	**48.06**	66.30	67.85
TPT	72.37	45.87	90.72	66.50	69.19	93.59	86.64	67.54	24.70	47.80	66.49	68.36
PromptAlign	**72.39**	47.24	**90.76**	68.50	69.47	94.01	86.65	67.54	24.80	47.86	**66.92**	68.99
ZERO	71.62	**47.89**	90.60	**68.58**	**69.87**	**94.48**	**87.20**	**68.20**	**26.25**	39.47	66.42	**69.23**
CLIP-ViT-B-16 + CLIP-ViT-L-14												
Zero-Shot	75.76	51.83	92.86	76.16	73.70	94.04	88.03	66.96	30.54	**54.38**	70.43	74.73
RLCF $^{\mathbf{t}_{ctx}}_{t=1}$	71.58	50.34	89.01	69.76	69.84	94.09	85.90	67.33	23.71	46.87	66.84	69.80
RLCF $^{\mathbf{t}_{ctx}}_{t=3}$	72.49	51.93	89.55	72.91	72.31	95.00	86.84	69.04	25.40	45.96	68.14	72.40
RLCF $^{\Theta_v}_{t=1}$	72.56	52.21	89.51	63.12	71.49	94.65	86.90	68.50	24.06	47.74	67.07	70.00
RLCF $^{\Theta_v}_{t=3}$	71.74	53.27	91.15	70.93	73.24	94.73	87.28	69.38	28.54	47.41	68.77	71.34
ZERO	**76.41**	**53.63**	**94.08**	**78.39**	**74.68**	**95.21**	**90.66**	**69.61**	**33.62**	44.21	**71.05**	**75.55**

Table 3: Computational requirements of different TTA methods.

Metric	CLIP-ViT-B-16		CLIP-ViT-B-16 + CLIP-ViT-L-14		
	TPT	ZERO	RLCF $^{\mathbf{t}_{ctx}}_{t=3}$	RLCF $^{\Theta_v}_{t=3}$	ZERO
Time [s]	0.57±0.01	**0.06**±0.01	1.20±0.02	0.18±0.01	**0.08**±0.02
Mem [GB]	17.66	**1.40**	18.64	9.04	**2.58**

Group 3 (RLCF): As Zhao et al. [54] do not report results on fine-grained classification, we use their code to evaluate four RLCF variants: Θ_v and $\mathbf{t}_{ctx}$ tuning, with $t = 1$ and $t = 3$ adaptation steps. We find that ZERO largely outperforms RLCF regardless of the configuration. Even with respect to the strongest RLCF $^{\Theta_v}_{t=3}$ variant, ZERO obtains an average improvement of $+2.28\%$.

Computational Requirements. The complexity of ZERO does not scale linearly with the size of the label space, as it does for prompt-tuning strategies. To quantify the computational gain of ZERO *w.r.t.* other TTA methods, we report the runtime per image and peak GPU memory in Table 3 under the same hardware (*i.e.*, 1 NVIDIA RTX 4090. We compare the computational requirements of ZERO to TPT and the RLCF pipeline in a worst-case scenario where the label space is large (ImageNet). We omit PromptAlign from our analysis since it has slightly worse computational performance than TPT.

ZERO is $9.5\times$ faster than TPT taking $12.61\times$ less memory, corresponding to an order of magnitude of computational savings in both time and space. Concerning the slowest RLCF variant (prompt tuning), ZERO is $15\times$ faster and takes $7.22\times$ less memory. In the faster RLCF $^{\Theta_v}$, text classifiers are also cached; nevertheless, ZERO is $2.25\times$ faster and $3.5\times$ more memory friendly.

5   Related Work

Closest to our work is a recent and very active research thread focusing on Episodic TTA with VLMs [38, 33, 54, 42]. As discussed in the manuscript, these methods mostly rely on prompt learning, a parameter-efficient strategy that only trains over a small set of input context vectors [20]. Narrowing down to VLMs, notable examples of prompt learning approaches include CoOp [56], CoCoOp [55], and MaPLe [15]. Episodic TTA has also been explored with traditional unimodal networks, such as ResNets [10], where MEM is still a core component [53]. In this context, MEM has recently been enriched with sharpness- [27] or shape-aware filtering [19]. Due to its nature, Episodic TTA is completely agnostic to the temporal dimension and is powerful when no reliable assumptions on the test data can be taken. Some other works relax these constraints and integrate additional assumptions such as *batches* of test data being available instead of single test points [45]. When test data are

9

assumed to belong to the same domain, one can rely on various forms of knowledge retention as a powerful mechanism to gradually incorporate domain knowledge [21, 22] or avoid forgetting [26]. The synergy between TTA and retrieval is also emerging as a powerful paradigm when provided with access to huge external databases [9, 50]. We particularly believe this can be a promising direction.

Closely related to our work are also Test-Time Training (TTT) and TTAug. In TTT the same *one sample* learning problem of Episodic TTA is tackled with auxiliary visual self-supervised tasks, such as rotation prediction [43] or masked image modeling [7], which require specialized architecture heads and are not directly applicable to VLMs. TTAug has recently been theoretically studied [16]. It boils down to producing a large pool of augmentations to exploit at test time [35], or to learn from [44]. In all its simplicity, ZERO can be seen as a strong TTAug baseline for VLMs, which, differently from concurrent work [51], does not involve any form of optimization.

6 Limitations

ZERO can seamlessly adapt a wide range of VLMs on arbitrary datasets without requiring extensive computational resources and is backed by theoretical justifications. However, we delineate four major limitations to our method which we report here.

Preliminary observations. The first limitation concerns the preliminary observations which led to ZERO, such as augmentation-induced overconfidence or a comparable error rate between source and augmented datasets. These observations may not persist if VLMs or benchmarks change significantly in the future, potentially leading to poor adaptation. For example, we have observed a consistent failure case for TTA with EuroSAT [11], with ZERO incurring large performance drops *w.r.t.* simple zero-shot classification. In Appendix F we unravel this worst-case further.

Theoretical assumptions. The second limitation stems from theoretical assumptions, the core one being the invariance of the marginal probability distribution to marginal entropy minimization. While our proposition guarantees invariance if entropy is globally minimized and the negative variation to the probability of the most probable class is less than the initial probability itself, these theoretical assumptions may not hold all the time. In this work, we supported our assumptions with empirical verification but, as per the first limitation, these may not extend to the space of all models and datasets. We refer the interested readers to Appendix A for a more in-depth discussion about the invariance of the prediction of $\bar{p}$ to MEM.

Independence among views. A third worthy-of-note limitation relates to the independence assumption among the views from which the marginal probability distribution is obtained. As we discussed in Section 2.3, the views themselves do not have any *direct* dependency, but they are still partially related through the source image from which they stem. Related to this limitation, we hypothesize that extending ZERO in a Retrieval-Augmented TTA setup (or a cache-based one) could improve the results. The discussion on this topic is extended in Appendix H.

Linear complexity with respect to augmented views. Finally, despite being much lighter than the current state-of-the-art TTA strategies, ZERO's computational requirements in the visual branch scale linearly with the number of views, since all of them need to be independently forwarded. On this, we believe that exploring how to augment directly in the latent visual space to also circumvent the forward pass of the vision encoder is an intriguing direction.

7 Conclusions

We theoretically investigated Marginal Entropy Minimization, the core paradigm of the current research on Test-Time Adaptation with VLMs. Building on our theoretical insights, we introduced ZERO: a frustratingly simple yet strong baseline for TTA, which only relies on a single batched forward pass of the vision encoder. ZERO works by setting the temperature of the Softmax operator to "zero" before marginalizing across confident views, which is equivalent, in terms of output, to the widely known paradigm of majority voting. Our experimental results on Natural Distribution Shifts and Fine-grained Classification unveil that ZERO favorably compares to state-of-the-art TTA methods while requiring relatively negligible computation. We hope our findings will inspire researchers to push the boundaries of TTA further.

10

Acknowledgements. The authors acknowledge the CINECA award under the ISCRA initiative for the availability of high-performance computing resources and support. Matteo Farina is supported by the PRIN project "LEGO-AI" (Prot.2020TA3K9N) and the PAT project "AI@TN". This work was supported by the projects EU Horizon ELIAS (No. 101120237), AI4TRUST (No.101070190), FAIR - Future AI Research (PE00000013), funded by NextGeneration EU, and carried out in the Vision and Learning joint laboratory of Fondazione Bruno Kessler and the University of Trento, Italy.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, 2014.

[2] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[3] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2020.

[6] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*. IEEE, 2004.

[7] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. Test-time training with masked autoencoders. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[8] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

[9] Moritz Hardt and Yu Sun. Test-time training on nearest neighbors for large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[11] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7), 2019.

[12] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[13] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[15] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

11

[16] Masanari Kimura. Understanding test-time augmentation. In *International Conference on Neural Information Processing (ICONIP)*. Springer, 2021.

[17] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCV-W)*, 2013.

[18] Ludmila I Kuncheva. *Combining pattern classifiers: methods and algorithms*. 2014.

[19] Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *International Conference on Learning Representations (ICLR)*, 2024.

[20] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021.

[21] Zichen Liu, Hongbo Sun, Yuxin Peng, and Jiahuan Zhou. Dart: Dual-modal adaptive online prompting and knowledge retention for test-time adaptation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

[22] Xiaosong Ma, Jie Zhang, Song Guo, and Wenchao Xu. Swapprompt: Test-time prompt adaptation for vision-language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[23] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[24] Prasanna Mayilvahanan, Thaddäus Wiedemer, Evgenia Rusak, Matthias Bethge, and Wieland Brendel. Does clip's generalization performance mainly stem from high train-test similarity? In *International Conference on Learning Representations (ICLR)*, 2023.

[25] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian conference on computer vision, graphics & image processing*. IEEE, 2008.

[26] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning (ICML)*, 2022.

[27] 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 *International Conference on Learning Representations (ICLR)*, 2023.

[28] OpenAI. Clip. URL `https://github.com/openai/CLIP`.

[29] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[32] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning (ICML)*, 2019.

[33] Jameel Hassan Abdul Samadh, Hanan Gani, Noor Hazim Hussein, Muhammad Uzair Khattak, Muzammal Naseer, Fahad Khan, and Salman Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

12

[34] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[35] Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[36] Lloyd Shapley and Bernard Grofman. Optimizing group judgmental accuracy in the presence of interdependencies. *Public Choice*, 1984.

[37] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.

[38] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[39] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[40] Jongwook Son and Seokho Kang. Efficient improvement of classification accuracy via selective test-time augmentation. *Information Sciences*, 2023.

[41] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[42] Elaine Sui, Xiaohan Wang, and Serena Yeung-Levy. Just shift it: Test-time prototype shifting for zero-shot generalization with vision-language models. *arXiv preprint arXiv:2403.12952*, 2024.

[43] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning (ICML)*, 2020.

[44] Devavrat Tomar, Guillaume Vray, Behzad Bozorgtabar, and Jean-Philippe Thiran. Tesla: Test-time self-learning with automatic adversarial augmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[45] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

[46] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[47] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[48] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[49] Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[50] Luca Zancato, Alessandro Achille, Tian Yu Liu, Matthew Trager, Pramuditha Perera, and Stefano Soatto. Train/test-time adaptation with retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[51] Maxime Zanella and Ismail Ben Ayed. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? *arXiv preprint arXiv:2405.02266*, 2024.

[52] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

13

[53] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[54] Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. Test-time adaptation with CLIP reward for zero-shot generalization in vision-language models. In *International Conference on Learning Representations (ICLR)*, 2024.

[55] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[56] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022.

14

A Marginal Entropy Minimization does not influence $\arg\max \overline{p}$

A.1 Proof of Proposition 2.1.

Proof. Let us denote the pre-TTA $p^{\text{init}}(c) = \overline{p}(y = c|\mathbf{x}, \mathbf{t}_{\text{ctx}})$ and the post-TTA $p^{\text{end}}(c) = \overline{p}(y = c|\mathbf{x}, \mathbf{t}^*_{\text{ctx}})$, *i.e.*, the marginal probabilities before and after optimizing $\mathbf{t}_{\text{ctx}}$. Let c^{init} and c^{end} denote the predictions before and after TTA, *i.e.*, $c^{\text{init}} = \arg\max p^{\text{init}}$ and $c^{\text{end}} = \arg\max p^{\text{end}}$.

To simplify the notation, let us use $\mathbf{z}^{*\text{txt}}$ to write any post-TTA text embedding $\mathbf{z}^{\text{txt}}(\mathbf{t}^*_{\text{ctx}})$.

Under the assumption that entropy is minimized (the optimal scenario for MEM), we have $p^{\text{end}}(c^{\text{end}}) = 1$, and $p^{\text{end}}(c) = 0 \ \forall \ c \neq c^{\text{end}}$.

Let us rewrite the final distribution p^{end} using the function g introduced in Sec.2.2. Specifically, for any class c, we have:

$$p^{\text{end}}(c) = \frac{1}{N} \sum_i^N \frac{\exp\left(\mathbf{z}_i^{\text{img}} \cdot \mathbf{z}_c^{\text{txt}}(\mathbf{t}^*_{\text{ctx}})/\tau\right)}{\sum_k^C \exp\left(\mathbf{z}_i^{\text{img}} \cdot \mathbf{z}_k^{\text{txt}}(\mathbf{t}^*_{\text{ctx}})/\tau\right)} = \frac{1}{N} \sum_i g(c, \mathbf{z}_i^{\text{img}}, \mathbf{z}_1^{*\text{txt}}, \ldots, \mathbf{z}_C^{*\text{txt}}). \quad (11)$$

Performing a first-order Taylor expansion on g, we have:

$$\begin{aligned} g(c, \mathbf{z}_i^{\text{img}}, \mathbf{z}_1^{*\text{txt}}, \ldots, \mathbf{z}_C^{*\text{txt}}) &= g(c, \mathbf{z}_i^{\text{img}}, \mathbf{z}_1^{\text{txt}}, \ldots, \mathbf{z}_C^{\text{txt}}) + \\ (\nabla_{[\mathbf{z}_1^{\text{txt}}, \ldots, \mathbf{z}_C^{\text{txt}}]} g)^t &([\mathbf{z}_1^{*\text{txt}}, \ldots, \mathbf{z}_C^{*\text{txt}}] - [\mathbf{z}_1^{\text{txt}}, \ldots, \mathbf{z}_C^{\text{txt}}]). \end{aligned} \quad (12)$$

We can also write any post-TTA text embedding $\mathbf{z}_c^{\text{txt}}(\mathbf{t}^*_{\text{ctx}})$ as a function of the text encoder $\mathbf{E}_{\text{txt}}$ prompted with optimized context vectors:

$$\mathbf{z}_c^{\text{txt}}(\mathbf{t}^*_{\text{ctx}}) = \mathbf{E}_{\text{txt}}([\mathbf{t}^*_{\text{ctx}}, \mathbf{t}_c]) = \mathbf{E}_{\text{txt}}([\mathbf{t}_{\text{ctx}} - \lambda \nabla_{\mathbf{t}_{\text{ctx}}} H, \mathbf{t}_c]). \quad (13)$$

Through another first-order Taylor expansion (this time on $\mathbf{z}_c^{\text{txt}}(\mathbf{t}^*_{\text{ctx}})$), we have:

$$\begin{aligned} \mathbf{z}_c^{\text{txt}}(\mathbf{t}^*_{\text{ctx}}) &= \mathbf{E}_{\text{txt}}([\mathbf{t}_{\text{ctx}}, \mathbf{t}_c]) + (\nabla_{\mathbf{t}_{\text{ctx}}} \mathbf{E}_{\text{txt}})^t (\mathbf{z}_c^{\text{txt}}(\mathbf{t}^*_{\text{ctx}}) - \mathbf{z}_c^{\text{txt}}(\mathbf{t}_{\text{ctx}})) = \\ \mathbf{E}_{\text{txt}}([\mathbf{t}_{\text{ctx}}, \mathbf{t}_c]) &- \lambda(\nabla_{\mathbf{t}_{\text{ctx}}} \mathbf{E}_{\text{txt}})^t \nabla_{\mathbf{t}_{\text{ctx}}}(H), \end{aligned} \quad (14)$$

leading to an equivalent re-writing:

$$\mathbf{z}_c^{\text{txt}}(\mathbf{t}^*_{\text{ctx}}) = \mathbf{z}_c^{\text{txt}}(\mathbf{t}_{\text{ctx}}) - \lambda(\nabla_{\mathbf{t}_{\text{ctx}}} \mathbf{E}_{\text{txt}})^t \nabla_{\mathbf{t}_{\text{ctx}}}(H) \quad (15)$$

Substituting (15) into (12), we can express g as follows:

$$\begin{aligned} g(c, \mathbf{z}_i^{\text{img}}, \mathbf{z}_1^{*\text{txt}}, \ldots, \mathbf{z}_C^{*\text{txt}}) &= g(c, \mathbf{z}_i^{\text{img}}, \mathbf{z}_1^{\text{txt}}, \ldots, \mathbf{z}_C^{\text{txt}}) - \lambda(\nabla_{[\mathbf{z}_1^{\text{txt}}, \ldots, \mathbf{z}_C^{\text{txt}}]} g)^t \mathbf{d} \\ \text{where } \mathbf{d} \in \mathbb{R}^C \text{ s.t. } \mathbf{d}_k &= (\nabla_{\mathbf{t}_{\text{ctx}}} \mathbf{E}_{\text{txt}}([\mathbf{t}_{\text{ctx}}, \mathbf{t}_k]))^t \nabla_{\mathbf{t}_{\text{ctx}}}(H)([\mathbf{t}_{\text{ctx}}, \mathbf{t}_k])) \ \ \forall k \in \{1, \ldots, C\}, \end{aligned} \quad (16)$$

with $\mathbf{d}_k$ denoting the k-th entry of the C dimensional vector $\mathbf{d}$. From (12) and (16) the *negative variation* $\delta g(c, \mathbf{z}^{\text{img}})$ to g before and after MEM can be expressed as:

$$\delta g(c, \mathbf{z}^{\text{img}}) = \lambda(\nabla_{[\mathbf{z}_1^{\text{txt}}, \ldots, \mathbf{z}_C^{\text{txt}}]} g)^t \mathbf{d} \quad (17)$$

Finally, for any class, we can rewrite its final probability p^{end} as a function of its initial probability p^{init} and the variation of g before and after TTA for the same class:

$$p^{\text{end}}(c) = p^{\text{init}}(c) - \frac{\lambda}{N} \sum_i \delta g(c, \mathbf{z}_i^{\text{img}}) \quad (18)$$

From Eq.(18) we have that if $p^{\text{init}}(c^{\text{init}}) > \frac{\lambda}{N} \sum_i \delta g(c^{\text{init}}, \mathbf{z}_i^{\text{img}})$, then the final probability $p^{\text{end}}(c^{\text{init}}) > 0$. In the optimal case for MEM the entropy of p^{end} is minimized, which entails that *only one class* can have a probability strictly greater than 0. Hence, $c^{\text{init}} = c^{\text{end}}$.

$\square$

15

Table 4: Empirical evidence supporting Proposition 2.1.

Proposition	IN-1k	IN-A	IN-v2	IN-R	IN-Sketch
$\arg\max p^{\text{init}} = \arg\max p^{\text{end}}$ [%]	95.73±0.05	95.55±0.12	94.86±0.17	96.78±0.08	91.23±0.09

A.2 Experimental verification

We support the previous proposition with empirical evidence, by manually counting how often the prediction of $\overline{p}$ is invariant to Test-Time Prompt Tuning by MEM. This experiment is easy to reproduce and consists of the following: augment N times, filter by confidence, compute p^{init}, optimize by MEM, compute p^{end} and check if $\arg\max p^{\text{init}} = \arg\max p^{\text{end}}$. We report the proportion of samples for which the proposition holds for all Natural Distribution Shifts datasets in Table 4, averaged over 3 runs with different seeds (the same used in Sec. 4 of the main body). Although the proposition only accounts for the cases where entropy is globally minimized, the table shows that the marginal probability distribution is largely invariant to MEM. In the best case (ImageNet-Sketch) MEM alters the prediction of $\overline{p}$ only 8.77% of the times. In the worst case (ImageNet-R), the prediction is unaltered for 96.78% samples.

A.3 Can invariance be anticipated?

In the proof of Proposition 2.1, we express the post-MEM embeddings as a function of the pre-MEM embeddings through a Taylor expansion. For this relationship to hold, the variation needs to be small. If the initial entropy is high, the gradients from MEM (and, thus, the variation between pre- and post-MEM embeddings) can be larger than what a Taylor expansion can accurately approximate. In such cases, Prop. 2.1 cannot be guaranteed. We execute a simple experiment using the validation set of ImageNet-1k, whose recipe is described below, to visualize this relationship.

We compute pre- and post-MEM marginal probability distributions. We sort the pre-MEM distributions in order of descending entropy (most to least uncertain) and quantize them into 10 bins. Bins shall be interpreted as follows: the leftmost bin contains the top 10% of samples with the highest entropy; the second bin contains samples outside the top-10% percentile but within the top-20%, and so on; the rightmost bin contains the bottom 10% of samples with the lowest entropy. For each bin we compute the invariance ratio, measuring how often the $\arg\max$ of the pre-MEM $\overline{p}$ does *not* change after MEM. Finally, we display a histogram with this data in Figure 2.

A trend appears: as the entropy decreases (left to right), invariance holds more and more often. Hence, intuitively, the most likely cases where invariance to MEM does not hold are those of high uncertainty in the initial marginal probability distribution. However, this may still be rare: even within the top 10% of most uncertain samples, invariance holds more than 82% of the time (leftmost bin).

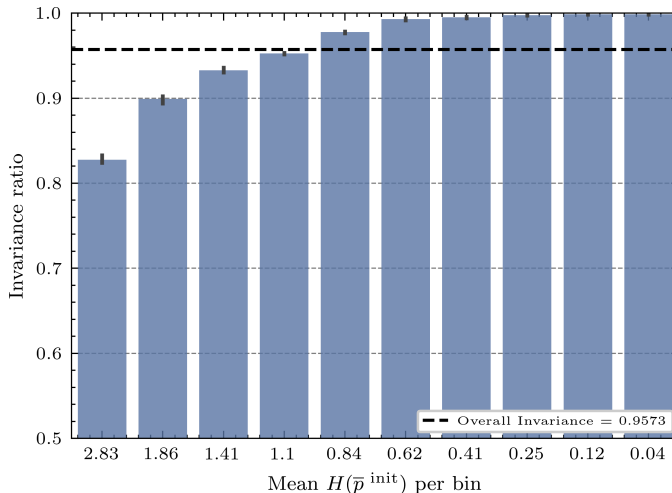



Figure 2: Entropy of the pre-TTA marginal probability distribution vs the invariance ratio.

16

Table 5: Results on Natural Distribution Shifts when adapting `CLIP-ViT-B-16` pretrained on the 2B English subset of LAION-5B. Top-1 accuracy is reported, and **bold text** indicates the best performer.

Method	ImageNet	A	V2	R	Sketch	Mean
CLIP-ViT-B-16 (LAION2B)						
Zero-Shot	69.27	37.08	61.27	78.83	54.85	60.26
Ensemble	70.43	38.32	62.28	80.41	55.54	61.40
TPT	70.61	41.94	62.96	80.40	55.48	62.28
ZERO	71.39	48.71	63.53	80.59	55.82	64.01
ZERO+Ensemble	**72.14**	**49.02**	**64.32**	**82.42**	**56.53**	**64.89**

Table 6: Results on Natural Distribution Shifts when adapting OpenAI's `CLIP-ViT-B-16`, with `CoOp`-learned prompts. Top-1 accuracy is reported, and **bold text** indicates the best performer.

Method	ImageNet	A	V2	R	Sketch	Mean
CoOp						
Zero-Shot	71.51	49.71	64.20	75.21	47.99	61.72
TPT	73.64	57.77	66.72	78.03	49.56	65.14
ZERO	**74.12**	**61.57**	**67.15**	**78.43**	**49.77**	**66.21**

Table 7: Fine-grained Classification with `CLIP-ViT-B-16` pretrained on the 2B English subset of LAION-5B. Top-1 accuracy is reported, and **bold text** indicates the best performer.

Method	FLWR	DTD	PETS	CARS	UCF	CAL	FOOD	SUN	AIR	ESAT	Mean	Median
CLIP-ViT-B-16 (LAION2B)												
Zero-Shot	69.71	54.43	89.37	89.94	64.02	95.82	81.38	70.60	26.04	47.05	68.84	70.16
Ensemble	68.70	54.55	87.76	89.98	67.64	96.51	81.64	70.62	25.68	**49.64**	69.27	69.66
TPT	69.47	54.53	89.00	90.72	66.68	96.16	81.76	**71.34**	26.73	48.81	69.52	70.41
ZERO	**70.82**	55.20	**89.77**	**91.95**	67.23	96.13	**83.65**	71.21	**28.25**	45.01	69.92	**71.02**
ZERO+Ensemble	68.01	**55.95**	87.67	91.87	**69.11**	**96.54**	83.83	71.09	28.10	47.10	**69.93**	70.10

B   Additional Experiments: LAION-2B Pretraining, Context Optimization and Hyperparameter Inheritance

This Appendix deals with enriching the experiments of Section 4, which focused on models officially released by OpenAI [28]. Here we focus on the comparison with TPT [38] and extend the analysis to: ① `CLIP-ViT-B-16` pretrained on the 2B English Subset of LAION-5B [34]; ② OpenAI's CLIP, transferred after supervised Context Optimization (`CoOp`) [56].

Implementation Details. For the experiments with LAION Pretraining, we use the `open_clip` repository, *i.e.*, the official code for [2]. The pretrained keyword for this model is `laion2b_s34b_b88k`. For `CoOp` we use the context vectors learned on ImageNet-1k officially released by [56]. The experimental setup is analogous to Section 4 in all details. We do not tune any hyperparameters for these different initializations, but inherit them from the experiments with OpenAI models.

B.1   LAION-2B Pretraining

Table 5 reports experiments on Natural Distribution Shifts, from which we observe no differences *w.r.t.* OpenAI models: ZERO largely outperforms TPT, and peak difference is reached with ImageNet-A [13]. Results on Fine-grained Classification are given in Table 7. We observe that ZERO improves the zero-shot baseline better with this pretraining, and overcomes TPT with an average margin of $+0.4\%$. In contrast, ensembling textual prompts appears less effective. We speculate this is because the 7 templates were explicitly tuned and selected for OpenAI models. The worst-case scenario is confirmed with satellite imagery [11]; please refer to Appendix F for a deeper investigation.

17

Table 8: Natural Distribution Shifts (percentile = 0.1). TTA methods are grouped according to the baseline model and top-1 accuracy is reported. **Bold text** is the best method within each group.

Method	ImageNet	A	V2	R	Sketch	Average
CLIP-ViT-B-16						
Zero-Shot	66.73	47.87	60.86	73.98	46.09	59.11
Ensemble	68.34	49.89	61.88	77.65	48.24	61.20
TPT	68.98	54.77	63.45	77.06	47.94	62.44
ZERO	69.06±0.04	61.35±0.26	64.13±0.17	77.28±0.08	48.29±0.04	64.02
ZERO$_{+Ensemble}$	**70.93**±0.02	**64.06**±0.09	**65.16**±0.21	**80.75**±0.08	**50.32**±0.09	**66.24**
MaPLe						
Zero-Shot	-	50.90	64.07	76.98	49.15	60.28
TPT	-	58.08	64.87	78.12	48.16	62.31
PromptAlign	-	59.37	65.29	79.33	50.23	63.55
ZERO	-	**64.65**±0.24	**66.63**±0.32	**79.75**±0.41	**50.73**±0.62	**65.44**
CLIP-ViT-B-16 + CLIP-ViT-L-14						
ZeroShot	73.44	68.82	67.80	85.40	57.84	70.66
RLCF $^{\mathbf{t}_{ctx}}_{t=3}$	73.23	65.45	69.77	83.35	54.74	69.31
RLCF $^{\Theta_v}_{t=3}$	**74.85**	73.71	**69.77**	86.19	57.10	72.32
ZERO	74.48±0.12	**77.07**±0.35	69.53±0.12	**86.87**±0.05	**58.59**±0.08	**73.31**

Table 9: Finegrained classification (percentile = 0.1). Formatting follows other tables.

Method	FLWR	DTD	PETS	CARS	UCF	CAL	FOOD	SUN	AIR	ESAT	Mean	Median
CLIP-ViT-B-16												
Zero-Shot	67.44	44.27	88.25	65.48	65.13	93.35	83.65	62.59	23.67	42.01	63.58	65.31
Ensemble	67.07	45.09	88.28	66.16	67.51	93.91	84.04	66.26	23.22	50.42	65.20	66.66
TPT	**68.75**	**47.04**	87.23	66.68	68.16	93.93	**84.67**	65.39	23.13	42.86	64.78	67.42
ZERO	67.07	45.80	86.74	67.54	67.64	93.51	84.36	64.49	24.40	39.60	64.11	67.31
ZERO$_{+Ensemble}$	66.82	45.86	87.20	**68.48**	**68.57**	**94.14**	84.58	**66.90**	**24.42**	43.77	65.07	**67.69**
MaPLe												
Zero-Shot	72.23	46.49	90.49	65.57	68.69	93.53	86.20	67.01	24.74	48.06	66.30	67.85
TPT	72.37	45.87	90.72	66.50	69.19	93.59	86.64	67.54	24.70	47.80	66.49	68.37
PromptAlign	**72.39**	47.24	**90.76**	**68.50**	69.47	94.01	86.65	67.54	24.80	47.86	**66.92**	**68.98**
ZERO	71.20	**47.70**	90.17	67.91	**69.49**	**94.12**	**86.78**	**67.55**	**25.57**	41.05	66.15	68.70
CLIP-ViT-B-16 + CLIP-ViT-L-14												
ZeroShot	**75.76**	51.83	92.86	76.16	73.70	94.04	**88.03**	66.96	30.54	**54.38**	**70.43**	74.73
RLCF $^{\mathbf{t}_{ctx}}_{t=1}$	71.58	50.34	89.01	69.76	69.84	94.09	85.90	67.33	23.71	46.87	66.84	69.80
RLCF $^{\mathbf{t}_{ctx}}_{t=3}$	72.49	51.93	89.55	72.91	72.31	**95.00**	86.84	69.04	25.40	45.96	68.14	72.40
RLCF $^{\Theta_v}_{t=1}$	72.56	52.21	89.51	63.12	71.49	94.65	86.90	68.50	24.06	47.74	67.07	70.00
RLCF $^{\Theta_v}_{t=3}$	71.74	53.27	91.15	70.93	73.24	94.73	87.28	**69.38**	28.54	47.41	68.77	71.34
ZERO	75.34	**54.22**	**92.90**	**77.33**	**74.26**	94.52	87.57	68.05	**32.11**	42.74	69.90	**74.80**

B.2 Context Optimization (CoOp)

For this comparison, we follow [38] and report `CoOp` on Natural Distribution Shifts only, presenting results in Table 6. We further observe patterns consistent with OpenAI models, with ZERO providing large improvements over TPT. Also here, the best-case scenario persists with ImageNet-A.

B.3 Hyperparameter Inheritance

In all experiments so far, including Section 4 as well as Tables 5, 6 and 7, we employed a percentile for confidence-based filtering set to 0.3. This value was obtained after validation on ImageNet-1k with OpenAI's `CLIP-ViT-B-16` and kept fixed for all models and datasets. Here, we show that ZERO obtains favorable performance even if the percentile for confidence-based filtering is not tuned in any way, but set to 0.1 by "inheriting" the value used in TPT [38]. These results are given in Tables 8 and 9. Surprisingly, some datasets within the Natural Distribution Shifts benchmark benefit from this more restrictive filtering (ImageNet-A above all), while we observe that Finegrained classification tends to improve when more views are retained. The core findings, however, are entirely unchanged: the best

18

case remains ImageNet-A, the worst-case remains EuroSAT, and ZERO outperforms competitor in most datasets, no matter the experimental setup.

C   Calibration and Overconfidence of CLIP on augmented Natural Distribution Shifts

In Section 3.1 of the manuscript, the validation set of ImageNet-1k is shown to convey that overconfidence emerges as a critical issue when predicting over augmented views. In this appendix, we expand the analysis to the 4 datasets for robustness to Natural Distribution Shifts (NDS) [13, 12, 32, 46]. For all datasets, we follow the augmentation setup of Sec.3.1, and generate augmented counterparts with $6\times$ more examples.

Figure 3: Expected Calibration Error (ECE) [8] of `CLIP-ViT-B-16` across 5 datasets for robustness to natural distribution shifts. Blue is the ECE of zero-shot `CLIP`, and orange is the ECE of zero-shot `CLIP` on an augmented version of the dataset after confidence-based thresholding.

First, let us define the calibration of DNNs. Calibrating DNNs is crucial for developing reliable and robust AI systems, especially in safety-critical applications. A DNN is perfectly calibrated if the probability that its prediction is correct ($\hat{y} = y$) given a confidence score random variable S is equal to its confidence score. The confidence score is commonly taken as the maximum of the output probability vector of the model, *i.e.*, $s = \max p(\cdot)$:

$$P(\hat{y} = y | S = s) = s \tag{19}$$

To evaluate the expected calibration error (ECE), we typically split the dataset into M bins B_m based on their confidence scores. We then calculate the accuracy of each bin, denoted as $\mathrm{acc}(B_m)$, and the average confidence, denoted as $\mathrm{conf}(B_m)$. The ECE is defined by the following formula:

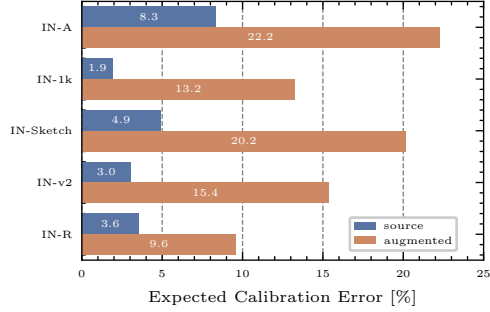
$$\mathrm{ECE} = \frac{1}{M} \sum_{m}^{M} \|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\| \tag{20}$$

Then, we show how the ECE of `CLIP-ViT-B-16` varies between "source" and augmented versions of all datasets (ImageNet-1k included) in Figure 3. From this experiment, we observe a large increase in the ECE across all datasets. In no cases, the ECE remains comparable to its default value when no augmentations are present. As we discussed in 3.1, the calibration error increases when the model is either more accurate than confident (signaling *underconfidence*) or the opposite, signaling *overconfidence*. Reliability diagrams are a standard tool to understand which is the case, hence we show them for all 4 NDS Datasets in Fig.4. These results are entirely consistent with Sec.3.1: the calibration error increases exclusively due to overconfidence, no matter the dataset. In parallel, the error rate of `CLIP-ViT-B-16` can either remain close to its default value (*e.g.*, ImageNet-Sketch), slightly decrease (*e.g.*, ImageNet-R and -v2) or largely decrease (ImageNet-A). We observe an identical pattern for CLIP models pretrained on LAION. For reference, see Figure 5.

D   On the expected risk of $\overline{p}$ and p.

The *expected* risk of a classifier f is commonly defined as the expectation of the risk function ℓ over the joint distribution of data and labels.

$$\mathcal{R}(f) = \mathbb{E}_{(x,y) \sim P_{\mathcal{X}\mathcal{Y}}} \Big[\ell(y, f(\mathbf{x})) \Big]. \tag{21}$$

In [16], the expected risk of a classifier $\overline{f}(\mathbf{x}) = \overline{p}(\cdot | \mathbf{x})$, which predicts by marginalizing over several augmented views, is theoretically shown to lower-bound the empirical risk of a standard classifier $f = p(\cdot | \mathbf{x})$ when the risk function ℓ is a squared error, *i.e.*, $\ell(a, b) = (a - b)^2$.

19

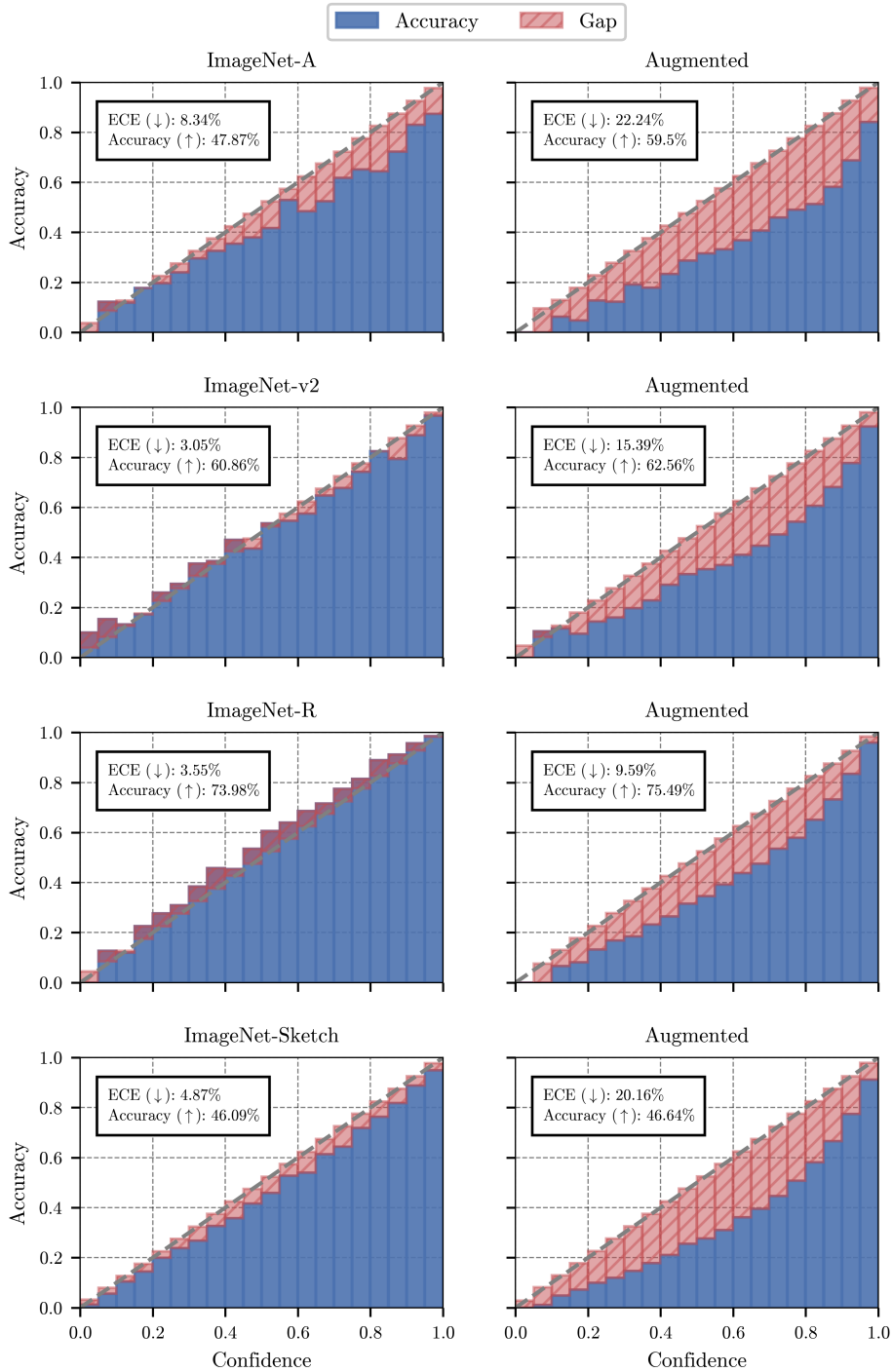



Figure 4: Reliability diagrams (20 bins) for `CLIP-ViT-B-16` on the 4 datasets for Natural Distribution Shifts. In each row, left is the ECE on the source dataset, right on the augmented and filtered version. Row 1: ImageNet-A [13]; Row 2: ImageNet-v2 [32]; Row 3: ImageNet-R [12]; Row 4: ImageNet-Sketch [46].

Here, we show that such a bound can be extended to any risk function ℓ that checks the triangular inequality. Specifically, note that if ℓ satisfies the triangular inequality, then:

$$\ell(y, \overline{p}(\mathbf{x})) \leq \frac{1}{N} \sum_{i=1}^{N} \ell(y, p(\mathbf{x}_i)). \tag{22}$$

20

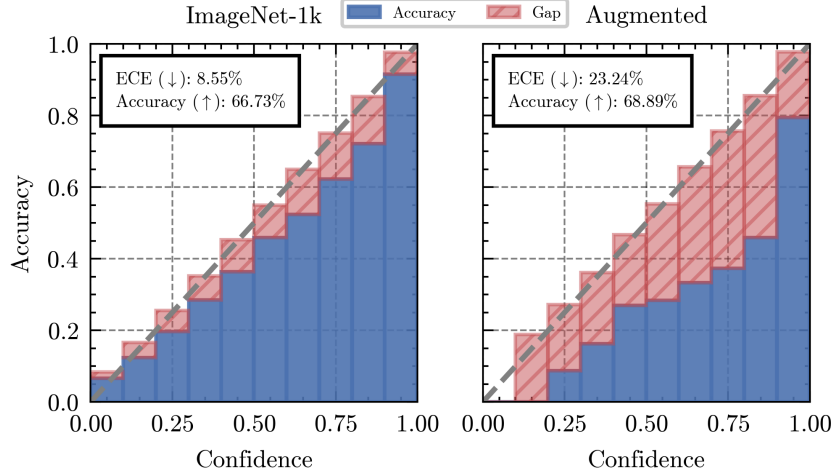



Figure 5: Reliability diagram (10 bins) for `CLIP-ViT-B-16` pretrained on LAION-2B when transferred zero-shot on ImageNet-1k. (left) Source Dataset, (right) Augmented version of the dataset.

The above inequality is obtained following these simple steps:

$$\|y - \overline{p}(\mathbf{x})\| = \|y - \frac{1}{N}\sum_{i=1}^{N}p(\mathbf{x}_i)\| = \|\frac{1}{N}\sum_{i=1}^{N}(y - p(\mathbf{x}_i))\| \leq \frac{1}{N}\sum_{i=1}^{N}\|(y - p(\mathbf{x}_i))\| \qquad (23)$$

Applying the expectation operator $\mathbb{E}$ over the joint distribution $P_{\mathcal{X}\mathcal{Y}}$ to both sides of Eq.(22) leads to:

$$\mathcal{R}(\overline{p}) \leq \frac{1}{N}\sum_{i=1}^{N}\mathcal{R}(p) = \mathcal{R}(p). \qquad (24)$$

Hence, the empirical risk of $\overline{p}$ lower-bounds that of p for any risk function ℓ satisfying the triangular inequality.

E Tie breaking with ZERO

A caveat of ZERO are ties, *i.e.*, cases with multiple classes having identical probability within the marginal probability distribution. This is clear to see when viewing ZERO as its equivalent paradigm of voting among confident views, simply because more than one class may have an equal amount of "votes". Throughout all the experiments of this work, ties are broken *greedily*. If a tie results from the top views, the procedure for breaking it follows these two steps: ① sort the remaining views by ascending entropy (most to least confident) and ② scan the views until a prediction is encountered that breaks the tie. Other than this, many alternative are possible, such as relying on the most confident prediction. Specifically, we have explored the following alternatives:

1. greedy tie breaking, as discussed above;
2. relying on the most confident prediction;
3. computing several marginal probabilities for $\overline{p}$, each by marginalizing over views with identical predictions, and picking the one with the lowest entropy for the final decision;
4. relying on the maximum logit (pre-Softmax);
5. using the averaged logits (pre-Softmax);
6. doing similar to point 2, using logits instead of probabilities;
7. random tie breaking;

and did not find consistent behaviour across all (fine-grained and NDS) datasets, suggesting this is indeed a minor component. We opted for greedy tie breaking due to its slightly better performance on the ImageNet validation set.

21

F A failure mode for TTA: satellite imagery

In our experiments, we find that an extremely OOD domain represents a consistent failure mode for TTA: satellite imagery. In all comparison groups, a zero-shot baseline largely outperforms *any* TTA strategy when evaluated on EuroSAT[11]:

- in *Group 1* the zero-shot baseline `CLIP-Ensemble` largely outperforms the best TTA strategy $\text{Z\textsc{ero}}_{+\text{Ensemble}}$;

- in *Group 2*, zero-shot `MaPLe` outperforms *PromptAlign*;

- in *Group 3* the best $\text{RLCF}_{t=1}^{\Theta_v}$ pipeline lies far behind the zero-shot teacher `CLIP-ViT-L-14`.

Here, we qualitatively and quantitatively report two main root causes for failures.

Qualitatively poor augmentations. In principle, TTA methods should rely on generic data augmentations, since not doing so would require going against the principles of the field by assuming some prior knowledge about the test data is available. As discussed in Sec.3, data augmentations are a doubled edged sword in TTA, and failing in crafting properly augmented views can potentially generate misleading or uninformative visual signals. We report some qualitative examples conveying this problem in Figure 7. In the Figure, we report three images from [11], together with the top-3 augmented views leading `CLIP-ViT-B-16` to its most confident predictions. Each source image is reported with the groundtruth, and all views are reported with both the prediction and the confidence of CLIP. Visually, one can perceive that the simple data augmentation scheme of cropping and flipping, which has largely been proven successful in [38, 33] and in our work, does not provide informative views, since most are alike one another.

Quantitatively high error. Augmentations are used by all TTA methods discussed in this paper, hence the previous discussion holds for TPT as much as it does for PromptAlign, RLCF or ZERO.

Nevertheless, we highlight an additional caveat about satellite imagery which is particularly detrimental for ZERO, and relates to the base model error over augmentations. Recall that, in ZERO, the usage of $\overline{p}$ is backed by theoretical motivations, and the manual adaptation of the temperature is supported by two concurrent observations: augmentations-induced overconfidence and a comparable error rate between source and augmented images. Simply put, the latter condition is not verified for satellite imagery. To show this phenomenon, we follow the experimental setup of Sec.3.1 and examine the reliability diagrams of EuroSAT[11] and of its augmented counterpart in Figure 6. As per Section 3.1, we display the ECE and the Top1-Accuracy on each version of the dataset. From this perspective, one can note that the base model error largely increases, in this domain, when augmented views are present. The accuracy on source images is 42.01%, dropping to 35.21% simply due to augmentations.

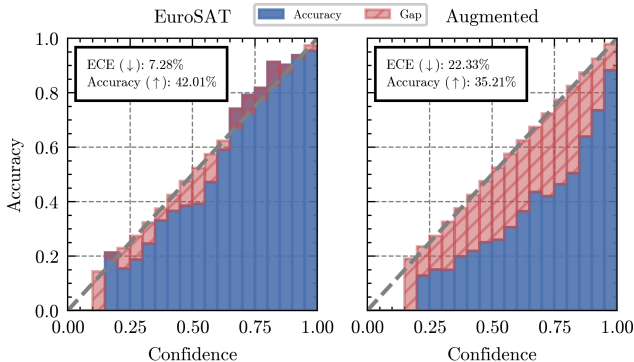



Figure 6: Reliability diagrams of `CLIP-ViT-B-16` for EuroSat and its augmented version, generated following Sec.3.1.

Both observations, combined, suggest that crafting augmentations for satellite imagery requires an ad-hoc treatment, which makes it a controversial benchmark for TTA.

G Natural Distribution Shifts vs Fine-grained Classification

Throughout the manuscript, one can observe that ZERO consistently provides larger improvements in Natural Distribution Shifts than it does in the Finegrained suite. We thus devote this section to digging deeper into this matter.

22

Table 10: Comparison among ① CLIP's zero-shot accuracy, ② CLIP's accuracy on the augmented counterpart of the dataset, and ③ ZERO. The augmented datasets are crafted following the protocol of Section 3.1 . "Gap" is defined as CLIP's zero shot accuracy *minus* its accuracy on the augmented dataset. "Improvement" is defined as the accuracy of ZERO *minus* that of zero-shot CLIP. Spearman's coefficient between "Gap" and "Improvement" equals -0.95: as the "Gap" decreases (*i.e.*, the lower the error on augmented views) ZERO provides more substantial improvements.

Method	FLWR	DTD	PETS	CARS	UCF	CAL	FOOD	SUN	AIR
① Zero-Shot	67.44	44.27	88.25	65.48	65.13	93.35	83.65	62.59	23.67
② Augmented	66.19	44.90	86.17	65.88	65.59	92.62	83.25	62.97	23.52
③ ZERO (perc = 0.1)	67.07	45.80	86.74	67.54	67.64	93.51	84.36	64.49	24.40
Gap = ① − ②	+1.25	−0.63	+2.08	−0.40	−0.46	+0.73	+0.40	−0.38	+0.15
Improvement = ③ − ①	−0.37	+1.53	−1.51	+2.06	+2.51	+0.16	+0.71	+1.90	+0.73

Perhaps unsurprisingly, we posit that ZERO improves over the zero-shot baseline if the zero-shot error rate of the model does not largely increase with augmented views. As Fig.1(b) displays, this is the case for all Natural Distribution Shifts datasets. To understand any different behaviors, we repeat the same experiment of Section 3.1 for the entire Fine-grained suite and report the results in Table 10. Please note that, in the table, the percentile for confidence-based in filtering in ZERO is set to 0.1, since the protocol for generating the augmented datasets follows the setup of TPT, which also uses a cutoff percentile of 0.1, and that we omit EuroSAT since an analogous experiment was presented in the previous Appendix.

Overall, we observe a strong correlation between the error gap and the improvement provided by Zero, with Spearman's coefficient being -0.95 across datasets. This result shows that the correlation is negative, *i.e.*, the lower the error gap, the larger the improvement (or, in other words, the better the zero-shot performance on augmented views, the larger the improvement of ZERO). This pattern is also consistent with the experiments on EuroSAT reported in the previous Appendix. Understanding why augmentations induce larger or smaller errors may be a case-by-case matter that relates to the nature of the datasets. Here, we pinpoint two possible reasons:

- The semantic space of the ImageNet variants of the Natural Distribution Shifts benchmark comprises many common categories, which may have appeared frequently during CLIP's pretraining. Hence, it seems reasonable that CLIP is robust *w.r.t.* augmented views of images belonging to these categories. In the Fine-grained classification suite, datasets such as SUN397 and Caltech101 also contain common object categories, which is consistent with the results shown above. In contrast, other datasets such as Flowers102 and Oxford-Pets span much less frequent concepts.

- Other than the semantic classification space, images' visual appearance also plays an important role. For example, datasets such as FGVC-Aircraft and Stanford Cars still contain rare concepts, but ZERO largely improves over the baseline nonetheless. Our augmentation setup is simple, and only contains random resized crops and random horizontal flips, which can constitute a "zoom-in" to a random portion of the image. For some benchmarks, this is useful as it may trigger CLIP's capabilities to recognize small details, such as logos, or even reading text, such as the car brand or the airline name. In contrast, more object-centric datasets such as Flower102, may lead to missing precious visual features (*e.g.*, the stem).

In our work we did not search for the best data augmentations but rather stuck to an established setting, using the same augmentations setup for all datasets. Nevertheless, the performance of ZERO is linked to the impact that data augmentations have on how the model perceives images, and we believe this is an interesting research direction to pursue.

H   Independence among views in the setup of Test-Time Adaptation

The theoretical framework of Section 2.3 models an ideal scenario, where independence holds among different inputs. To clarify, this means that the model's error on view x_i should not be correlated with the error on any other view x_j, which allows writing the compound error with a binomial distribution as in (6).

23

Table 11: Standard deviations of ZERO for Fine-grained classification. Each cell refers to Tab. 2.

Method	FLWR	DTD	PETS	CARS	UCF	CAL	FOOD	SUN	AIR	ESAT
CLIP-ViT-B-16										
ZERO	±0.12	±0.07	±0.06	±0.15	±0.24	±0.14	±0.01	±0.10	±0.12	±0.11
ZERO$_{+Ensemble}$	±0.07	±0.26	±0.16	±0.04	±0.07	±0.19	±0.04	±0.18	±0.47	±0.08
MaPLe										
ZERO	±0.33	±0.51	±0.41	±0.52	±0.66	±0.32	±0.10	±0.40	±0.33	±4.77
CLIP-ViT-B-16 + CLIP-ViT-L-14										
ZERO	±0.11	±0.09	±0.15	±0.19	±0.14	±0.04	±0.08	±0.03	±0.32	±0.19

In practice, achieving perfect independence is challenging, if not impossible. Hence, a suitable approximation strategy to mitigate this issue is to promote diversity. In classical ensembling theory, a well-established approach is to train different models on different subsets of the available data. Similarly, the augmentation scheme of random cropping aligns with this approach by presenting the model with different portions of the image each time.

Moreover, ideally, the augmentation pipeline should not change the underlying label of the original input and guarantee that the model's error rate on augmented views remains comparable to the error rate on the original inputs belonging to the same category. In practice, this entails that augmentations should not disrupt the visual appearance of the image, and, consequently, some views may result in a slight or moderate correlation, because some "parts" of the source image will overlap among them. An analogy with classical literature can be drawn also in this case. Specifically, when not enough data are available, overlaps among the training sets of different models are required to ensure convergence. Consequently, models producing slightly or moderately correlated predictions are more likely to emerge.

I   Additional Implementation Details

Standard deviations. To complement the results on Fine-grained classification, we report the standard deviation of ZERO computed over 3 runs with different seeds in Table 11. These are not reported together with the average top1-accuracy in Tab. 2 to avoid an excessively dense table. On average, standard deviations are very small, suggesting that regardless of the inherent randomness of data augmentations, ZERO is relatively stable. Note that standard deviations in *Group 2* (*i.e.*, with MaPLe) are slightly greater than those in the other groups. This fact does not stem from ZERO's or MaPLe's greater instability, but from an experimental detail which we report here for completeness: while only one set of weights is officially released for each CLIP version [28], Khattak et al. [15] released 3 sets of pretrained weights for MaPLe, varying on the seed. To avoid picking one, we associated a set of weights to each of our runs, hence results from slightly different initializations are computed to match the experimental setup of Samadh et al. [33] (PromptAlign).

Reproducibility of TTA methods. For section 4, we reproduced all methods using the source code provided by the authors with the hardware at our disposal. This was done to ensure that hardware differences did not interfere with a correct evaluation. We found that all TTA strategies are highly reproducible, with negligible differences (*i.e.*, $\Delta < 0.1$) which we omitted by reporting the numbers from the official papers. In case of larger differences, we reported reproduced results.

24

Forest	Highway or Road (0.76)	Highway or Road (0.73)	Highway or Road (0.75)
Herbaceous Vegetation Land	Sea or Lake (0.48)	Sea or Lake (0.44)	Sea or Lake (0.43)
Sea or Lake	Highway or Road (0.24)	Highway or Road (0.24)	Highway or Road (0.25)
(a) Source Image	(b) Most confident	(c) 2nd most conf.	(d) 3rd most conf.

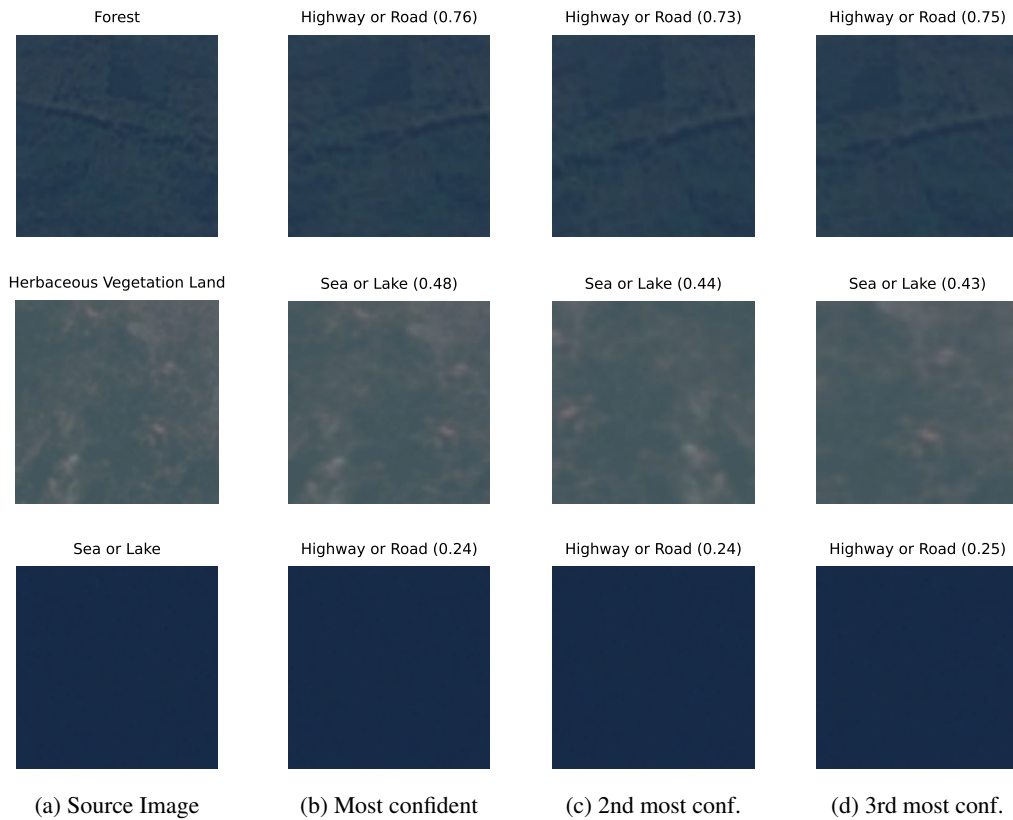



Figure 7: Examples of satellite imagery taken from EuroSAT[11], along with augmentations leading to high confident predictions. Column (a) reports source images with their label. Columns (b-d) report views sorted by entropy (lowest to highest), paired with the prediction and the confidence of `CLIP-ViT-B-16` [31].

25

NeurIPS Paper Checklist

1. **Claims**

    Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

    Answer: [Yes]

    Justification: This work studies TTA by MEM, and shows that a method obtained by manually adapting a single parameter is hidden therein. The method, called ZERO, is supported by theoretical justifications. Experiments show that ZERO compares favorably *w.r.t.* existing TTA methods while being faster and more memory friendly, which is what the abstract claims.

    Guidelines:

    - The answer NA means that the abstract and introduction do not include the claims made in the paper.
    - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
    - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
    - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

    Question: Does the paper discuss the limitations of the work performed by the authors?

    Answer: [Yes]

    Justification: The paper contains a "Limitations" section in the main body. It discusses theoretical assumptions and computational aspects. The paper discusses computational aspects also in the "Experiments" section of the main paper, especially concerning the dataset size (which is the size of the label space, in the scope of this paper). The Appendix provides evidence for the theoretical proposition of this paper and reports failure modes for the presented approach, as well as in broader TTA.

    Guidelines:

    - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
    - The authors are encouraged to create a separate "Limitations" section in their paper.
    - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
    - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
    - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
    - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
    - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
    - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

26

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: This paper contains a proposition, whose proof is in the Appendix. All other theoretical results are supported by empirical evidence.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper contains implementation details in both the "Experiments" section of the main body and the Appendix. A PyTorch-like implementation is provided alongside the proposed strategy in Section 3 of the main body.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

27

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The datasets we have used are all publicly available, including the data splits. The manuscript links to a public GitHub repository containing the implementation.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All details are explicitly reported in the "Experiments" section of the main paper, including hyperparameters and data splits.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

28

Answer: [Yes]

Justification: The "Experiments" section contains the standard deviations, computed over 3 runs, of the proposed method for one out of two benchmarks used. The standard deviations for the second benchmark are reported in the Appendix, to avoid formatting an excessively dense table, and are discussed. The empirical verification for the proposition of this paper is also executed for 3 runs, and standard deviations are reported in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: In Section "Experiments", we explicitly tell which and how many GPU devices were used. Computational aspects concerning runtime and peak memory consumption are also reported in the same section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We have carefully inspected the code of ethics, and believe our work conforms with it.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

29

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This paper theoretically analyzes a paradigm of Test-Time Adaptation with Vision-Language Models and presents a simple and effective method. We do not see a direct path from our method to a negative societal impact which is not inherently already related to VLMs.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We believe the content of this paper does not have a high risk of misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

30

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The original creators of all models, datasets and algorithms used in this work are properly credited, with citations around the manuscript. We used their material for the only non-commercial purpose of developing this research paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: There are no external assets related to this paper or its Appendix, except for the open-sourced implementation. The code is properly commented, on to guide readers into understanding it step-by-step, and a comprehensive "readme" file for installation is given.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This research does not involve crowdsourcing, nor the involvement of external human subjects except for the authors.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

31

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This research does not involve any participants except for the authors.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

32