AGI-TPT: ADAPTIVE GEOMETRY-INFORMED TEST-TIME PROMPT TUNING FOR CALIBRATED ZERO-SHOT VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time prompt tuning (TPT) adapts a vision–language model such as CLIP to each test image by lowering the entropy of its prediction. It improves accuracy but makes the model over-confident, and when the prompt is carried along a stream of test images it collapses. We give a mechanism: entropy minimization pulls together the classes that compete for an image. We propose AGI-TPT, a regularizer whose target adapts to the number of classes N and the feature dimension D, and whose push adapts to each pair's current state. It pushes apart only the pairs of class text features whose cosine is above the *attainable level* $c^\star(N, D)$, the smallest largest-cosine that N unit vectors in D dimensions can have, and it applies no force to pairs below it. The level is computed from N and D, not tuned. For free class vectors we prove that the zeros of the loss are the maximally separated configurations, that its low-rank local minima are global, and that merged classes are never local minima ($N \leq 2D - 1$). We also show that at a single TPT step, spread regularizers differ mainly in strength: at equal strength AGI-TPT coincides with O-TPT's penalty, and a random reversal of TPT's step of the same strength comes within 0.13 ECE of it. The geometry matters when adaptation continues. There, among the prompt-tuning methods with a single template that we compare, AGI-TPT gives the best accuracy and calibration. With ViT-B/16 on 11 datasets, online, it reaches 68.8% average accuracy at ECE 1.38; at a single step it lowers the ECE of A-TPT, the best-calibrated published method, from 2.61 to 1.81 under the same protocol, and under sustained adaptation from 2.9 (A-TPT with a drift remedy) to 1.4. On four ImageNet shifts it reaches 65.3% accuracy at ECE 1.40. It keeps more of CLIP's class-similarity structure than any baseline, and its calibration does not drift along the stream. Code will be made publicly available upon acceptance.

1 INTRODUCTION

A vision–language model such as CLIP (Radford et al., 2021) embeds images and short texts in one space. It classifies an image *zero-shot*: each class is described by a prompt such as "a photo of a {class}", and the class whose text feature is most similar to the image feature wins. Test-time prompt tuning (TPT; Shu et al., 2022) replaces the words "a photo of a" by learnable vectors and, for each test image, takes one gradient step that lowers the entropy of the prediction. TPT raises accuracy, and confidence even more. A classifier is *calibrated* if, among predictions made with confidence p, a fraction p is correct; the expected calibration error (ECE) measures the average gap. With CLIP ViT-B/16, TPT's ECE averaged over 11 datasets is 11.6, against 4.43 for zero-shot CLIP (Yoon et al., 2024, Table 1).

Recent methods restore calibration by pushing the text features of the classes apart: C-TPT (Yoon et al., 2024) disperses them around their centroid, O-TPT (Sharifdeen et al., 2025) pushes them toward orthogonality, A-TPT (Ahamed et al., 2026) widens each class's nearest-neighbour angle, and D-TPT (Han & Hwang, 2025) flattens the coordinates of each feature. Each is evaluated after TPT's single step. Our first result explains what that evaluation measures (section 2). At the first step of Adam every fixed-weight regularizer adds the *same* vector to every image's gradient, and it changes the prediction only through the coordinates where this vector overrides TPT's own step. The strength of this override, not the geometry behind it, sets the level of calibration. At equal strength AGI-TPT

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

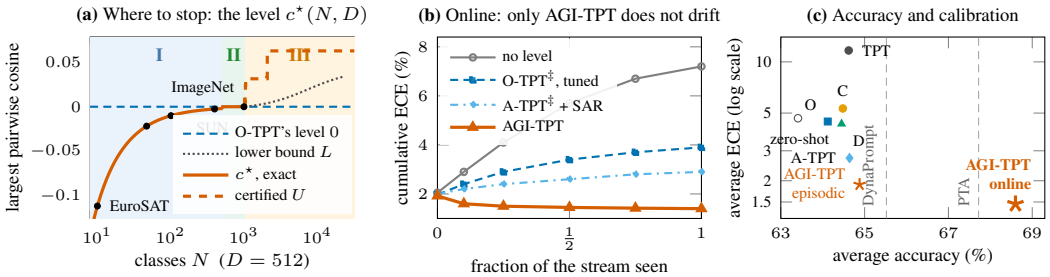


Figure 1: **Where to stop, and why it matters. (a)** The attainable level $c^\star(N, D)$, the smallest largest-pairwise cosine of N unit vectors in $\mathbb{R}^{512}$ (exact; certified upper bound U beyond $2D$). O-TPT's implicit level 0 is too high for $N \leq D{+}1$ and unreachable beyond $2D$. **(b)** Online (ViT-B/16, six datasets): TPT collapses (ECE 55.8, off scale), a push without a level drifts, tuned baselines drift more slowly; AGI-TPT does not drift. **(c)** Average accuracy and ECE on the 10 datasets every method reports (published episodic baselines, table 1); dashed: online methods that report accuracy only. The two AGI-TPT points are episodic (comparable to the baselines) and online (comparable only to the dashed lines).

and O-TPT's penalty coincide, most of the gap to C-TPT, D-TPT and A-TPT closes, and a random reversal of TPT's step with the same strength comes within 0.13 ECE of AGI-TPT (section 5.1). One-step calibration is therefore weak evidence for any geometry, including ours.

The geometry matters once adaptation continues, for example when the prompt is carried from one test image to the next (online TPT). There, unregularized TPT collapses: its accuracy on ImageNet-A falls to 6.96% (Xiao et al., 2025, Table 5). We show why (proposition 3). A step of entropy minimization pulls together the classes that compete for an image and pushes the other classes away from it. Over a stream, classes that often compete keep merging, and rarely predicted classes stop being predicted at all. A spread regularizer that counters this for thousands of steps must answer two questions that one step never raises: *where to stop*, and *which pairs of classes to move*. If it can never stop, it keeps pushing until the model becomes under-confident. If it pulls classes that are already far apart back together, it undoes part of its own work. Every prior regularizer fails one of these tests: C-TPT, O-TPT and D-TPT by where their minimizers lie, A-TPT because its push on a class's nearest pair moves that class toward other classes and never switches off (section 3). The concurrent SoC (Fillioux et al., 2026) also rejects orthogonality as the stopping point but keeps level 0 and a two-sided penalty (section 6).

We derive a regularizer that passes both (section 4). *Where to stop:* at the *attainable level* $c^\star(N, D)$, the smallest value that the largest pairwise cosine of N unit vectors in $\mathbb{R}^D$ can take (fig. 1a). It is known exactly up to $2D$ classes and certified beyond, and it is the most separated stopping point that the classes can actually reach. *Which pairs to move:* only the pairs whose cosine is above this level, through a ReLU on the excess, so that no pair that is already separated enough is pulled back. The loss is one line, $\frac{1}{N^2}\|\operatorname{ReLU}(S - G_{c^\star})\|_F^2$ on the Gram matrix S of the class features. We call the method AGI-TPT (Adaptive Geometry-Informed TPT): its target adapts to the number of classes and the dimension, and its push on each pair adapts to that pair's current state. The level is computed, never tuned; the weight λ is the one hyperparameter. On CLIP the level differs from O-TPT's 0 by less than 0.03 once $N \geq 37$, so it matters most for few classes and beyond $2D$; having a stop, and the ReLU, matter everywhere (section 5.5).

Contributions.

- *What one-step results measure.* At TPT's step every fixed-weight regularizer acts through one offset, so its one-step calibration is set mainly by strength (sections 2 and 5.1).
- *Why sustained adaptation needs a stopping rule.* Entropy minimization provably merges competing classes; every prior regularizer violates one of two requirements for a counter-force (section 3).
- *A regularizer with guarantees, for free class vectors.* AGI-TPT stops at the attainable level; its zeros are the maximally separated configurations, its low-rank local minima are global, and merged classes are never local minima for $N \leq 2D - 1$ (section 4).

- *Results and mechanism.* Among the single-template prompt-tuning methods we compare, AGI-TPT has the best accuracy and calibration: online, 68.8% on 11 datasets at ECE 1.38, and 65.3% at ECE 1.40 on four ImageNet shifts; like for like, it lowers A-TPT's ECE from 2.61 to 1.81 at one step and from 2.9 (with a drift remedy) to 1.4 online, and it does not drift. Causal split tests, a level test on 18 datasets in every regime, and a measure of preserved class structure isolate the ReLU and the level (section 5).

2 SETUP, AND WHAT ONE STEP CAN SHOW

Model and TPT. CLIP maps an image x to a unit vector v, and the prompt of class k, with learnable context vectors $\boldsymbol{\pi} \in \mathbb{R}^{d_\pi}$, to a unit *class feature* $t_k(\boldsymbol{\pi}) \in \mathbb{R}^D$. The matrix T stacks the N class features, $S = TT^\top$ holds their pairwise cosines s_{ij}, and $\bar{t} = \frac{1}{N}\sum_k t_k$ is their centroid. Predictions are $p(k \mid x) \propto \exp(z_k)$ with logits $z_k = \alpha\langle v, t_k\rangle$ and a fixed scale α. For every test image, TPT resets the prompt to $\boldsymbol{\pi}_0$ and takes one AdamW step (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) on $\mathcal{L}_{\text{TPT}}(\boldsymbol{\pi}; x)$, the entropy of the prediction averaged over the most confident augmented views (Shu et al., 2022). A spread regularizer $\mathcal{R}(S)$ is added with a weight λ. Calibration is measured by the ECE with 15 equal-width confidence bins (Naeini et al., 2015; Guo et al., 2017) for every number we compute; published baseline numbers are copied as printed and were computed with each paper's own evaluation code, whose bin count can differ (appendix F.5). We write gap for the mean confidence minus the accuracy; $\mathrm{ECE} \geq |\mathrm{gap}|$ for every predictor.

One step is one offset. At the first step, Adam's moment estimates for a gradient g are $\hat{m} = g$ and $\hat{v} = g^{\odot 2}$, so every coordinate moves by $-\eta\,\mathrm{sign}(g_j)$: the step is a *sign step*.

Proposition 1 (One-step offset)**.** *Let $\mathcal{R}$ depend only on $T(\boldsymbol{\pi})$, with a fixed weight λ. For every image x, the first bias-corrected AdamW step on $\mathcal{L}_{\text{TPT}}(\boldsymbol{\pi}; x) + \lambda\mathcal{R}$ from $\boldsymbol{\pi}_0$ is $\Delta\boldsymbol{\pi}(x) = -\eta[\mathrm{sign}(g_x + b_\mathcal{R}) + \omega\boldsymbol{\pi}_0]$ (as $\epsilon \to 0$), with $g_x = \nabla\mathcal{L}_{\text{TPT}}(\boldsymbol{\pi}_0; x)$, weight decay ω, and $b_\mathcal{R} = \lambda\nabla_{\boldsymbol{\pi}}\mathcal{R}(T(\boldsymbol{\pi}_0))$ the* same *for every image. Regularizers with parallel offsets make identical predictions once their strengths are matched.*

Consequently, the step reverses TPT's own step exactly on the *override set* $O_x = \{j : \mathrm{sign}(g_{x,j} + b_{\mathcal{R},j}) \neq \mathrm{sign}(g_{x,j})\}$, so two regularizers with the same override sets on every image make identical predictions, whatever their geometry (Proposition C.1).

Corollary 2 (Override and entropy)**.** *Let $\mu_x = \|g_x|_{O_x}\|_1 / \|g_x\|_1$ be the* override mass*. To first order in η the step changes the TPT loss by $-\eta\|g_x\|_1(1 - 2\mu_x) - \eta\omega\langle g_x, \boldsymbol{\pi}_0\rangle$. Apart from the shared weight-decay term, this is $(1 - 2\mu_x)$ times TPT's own change, so the step* ascends *the entropy once $\mu_x > \frac{1}{2}$.*

Proofs are in appendices C and E. A regularizer thus changes a one-step prediction only through *which* coordinates of TPT's step it reverses, and its override mass sets *how much* of TPT's confidence gain it undoes. At equal override mass (*equal strength*) AGI-TPT and O-TPT's penalty coincide and most of the gap to the other regularizers closes (section 5.1); the geometry matters when adaptation continues, which is where the rest of the paper operates.

3 WHY SUSTAINED ADAPTATION NEEDS A STOPPING RULE

Entropy minimization merges competing classes. To see what a spread regularizer must counter, consider the entropy step alone, acting directly on the class features (the prompt moves them jointly; section 5.3 tests the prompt case). For one image v, write $a_k = \langle v, t_k\rangle$ and $p = \mathrm{softmax}(\alpha a)$, and call class k *competitive* if $p_k > e^{-H(p)}$, i.e. more probable than a typical class. Unless p is uniform, the most probable class is competitive.

Proposition 3 (Entropy pulls competitors together; proof in appendix E.1)**.** *Under the gradient flow of the entropy $H(p)$ on the unit sphere of each class feature, $\dot{t}_k = \varepsilon_k(v - a_k t_k)$ with $\varepsilon_k = \alpha p_k(\log p_k + H)$. Hence* (i) *competitive classes move toward the image and the others move away from it: $\dot{a}_k = \varepsilon_k(1 - a_k^2)$;* (ii) *for two competitive classes with $a_i, a_j > 0$, the cosine s_{ij} increases whenever $s_{ij} < \min(a_i, a_j)/\max(a_i, a_j)$;* (iii) *for two non-competitive classes under the same condition, s_{ij} decreases.*

The condition in (ii) holds for the typical ambiguous image, whose top classes have nearly equal logits while their text features are not identical. Over a stream, pairs that often compete are pulled together again and again, and rarely competitive classes drift away from every image until they are no longer predicted: the collapse of online TPT (section 5.2). A counter-force that acts for thousands of steps must know where to stop.

Where a regularizer stops. Gradient descent on a regularizer can come to rest only at its critical points, and it rests stably only at its local minimizers. We therefore judge a spread regularizer by its minimizers, which is where its push ends, and by the force it applies to each pair on the way. For a pairwise regularizer $\mathcal{R} = \frac{1}{N^2}\sum_{i\neq j}\varphi(s_{ij})$ we call $F = \varphi'$ the *force* on a pair: a positive force pushes the pair apart, a negative one pulls it together. The largest pairwise cosine is the right measure of separation, because bounding it bounds every class at once:

Lemma 4 (Worst-pair certificate; proof in appendix C)**.** *If $s_{kj} \le c \le \frac{1}{2}$ for all $j \neq k$, then $\operatorname{dist}(t_k, \operatorname{conv}\{t_j\}_{j\neq k}) \ge \sqrt{1-2c}$; hence some unit direction v gives $\alpha\langle v, t_k - t_j\rangle \ge \alpha\sqrt{1-2c}$ for every $j \neq k$.*

A smaller largest cosine thus gives every class a larger margin in some direction; merged classes ($s = 1$) have none. This gives two requirements for a regularizer that must act for many steps:

- **(R1) Where to stop.** Its minimizers should exist for N classes in $\mathbb{R}^D$, contain no merged classes, and make the largest pairwise cosine as small as N classes in $\mathbb{R}^D$ allow.
- **(R2) Which pairs to move.** It should push only pairs that are not yet separated enough, and never pull a pair that is already separated back together.

Prior regularizers (table 3; proofs in appendix D). We analyze each objective as its paper writes it. *C-TPT* maximizes the mean distance of the class features from their centroid; its maximizers are all configurations with $\sum_k t_k = 0$, which for $N \ge 4$ include two coincident classes (Proposition D.3), making the observation of Sharifdeen et al. (2025, p. 4) exact. *O-TPT* penalizes $\|S - I\|_F^2 = \sum_{i\neq j} s_{ij}^2$ (its Eq. 2 is written with $\|\cdot\|_2^2$; we analyze the Frobenius reading, a sum over pairs), whose force $2s$ pulls every obtuse pair toward $90°$.

Its minimizers are the unit-norm tight frames, and for $N \ge D$ its landscape has no spurious local minima (Benedetto & Fickus, 2003). But for $D < N \le 2D$ a minimizer has largest cosine 0 only if $(N - D) \mid D$, which fails for ImageNet at $D = 512$ and 768: the most separated configuration is then a critical point of the penalty but not a minimizer. For $N \ge 2D$ some minimizers contain two coincident classes (Proposition D.2). O-TPT's failure is therefore where it stops, not how it descends. *A-TPT* maximizes $\mathrm{AD} = \frac{1}{N}\sum_i \min_{j\neq i}\theta_{ij}$, the mean over classes of the nearest-neighbour angle (Ahamed et al., 2026, Eq. 1), motivated by the Tammes problem of maximizing the *smallest* angle. The two coincide only when all nearest-neighbour angles are equal (Proposition D.4); AD's maximizers have no closed form, its push on the nearest pair is positive at every angle and so never switches off (Proposition D.5), and its gradient moves each class away from one neighbour only, which brings every class at an angle above $90°$ from that direction closer (Proposition D.4(iii)): it violates (R2). *D-TPT* adds $\frac{1}{N}\sum_k \mathrm{KL}(\sigma(t_k)\,\|\,\mathcal{U})$, which flattens the softmax σ of each feature's own coordinates (Han & Hwang, 2025, Eq. 3); no term involves two classes, and its minimizers place every class at $\pm\mathbf{1}/\sqrt{D}$, where all classes coincide up to sign (Proposition D.6).

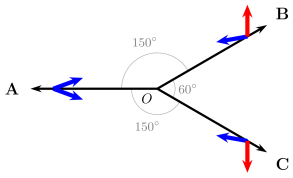
4 AGI-TPT: STOP AT THE ATTAINABLE LEVEL, PUSH ONLY WHAT IS ABOVE IT

We build the regularizer in two steps, one for each requirement, and then state what it guarantees.

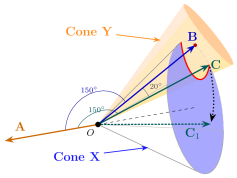
4.1 STEP 1 (R1): WHERE TO STOP

We want the regularizer to stop on a set of the form $Z(c) = \{T : \max_{i\neq j} s_{ij} \le c\}$, since by lemma 4 such a set certifies every class at once. The smallest useful c is the value $c^\star(N, D) = \min_T \max_{i\neq j} s_{ij}$ of the spherical-code (Tammes) problem: the best worst-pair cosine of N unit vectors in $\mathbb{R}^D$ (Tammes, 1930; Rankin, 1955). A configuration that attains it is an *optimal code*.

Lemma 5 (Attainability)**.** *$Z(c) \neq \emptyset$ if and only if $c \ge c^\star$. The set $Z(c^\star)$ is exactly the set of optimal codes, and for $c > c^\star$, $Z(c)$ also contains non-optimal configurations.*



(a) Angles $150°, 150°, 60°$ sum to 2π: widening BC must close AB or AC.



(b) Angles $150°, 150°, 20°$: moving C toward C_1 widens BC and keeps AC fixed.

Figure 2: **Which pairs to move, in the smallest case** ($N = 3$, level $c^\star = -\frac{1}{2}$). Blue arrows: attraction on the obtuse pairs AB, AC; red: repulsion on the close pair BC, under the two-sided residual. In (b) the two-sided residual still pulls AB and AC (force $1 - \sqrt{3} \approx -0.73$ each, together a third of its total force), although BC can be separated without closing them. The ReLU residual applies no force to AB, AC and pushes only BC (proposition 6; appendix D.5).

The lemma is elementary; its content is the value of $c^\star$. A level below $c^\star$ can never be reached: at every configuration some pair is still pushed, with force at least $2(c^\star - c)$, whatever the data (proposition 10); the push never *switches off*, although minimizers of the loss exist. A level above $c^\star$ can be reached but stops pairs early, and the margin of lemma 4 shrinks. We therefore use *the smallest level we can prove to be attainable*. Proposition 11 (appendix B.1) gives it in three regimes (fig. 1a). **I**, $N \leq D{+}1$: $c^\star = -\frac{1}{N-1}$, attained only by the regular simplex, whose vectors form a simplex equiangular tight frame (ETF), the geometry of neural collapse (Kothapalli, 2023). **II**, $D{+}2 \leq N \leq 2D$: $c^\star = 0$ (Rankin, 1955). Among the optimal codes, the one of least frame potential $\sum_{i,j} s_{ij}^2$ is unique: an orthogonal sum of $N{-}D$ regular simplices of nearly equal dimensions. It is not an ETF, and no ETF of $\mathbb{R}^D$ is an optimal code; a neural-collapse target is impossible here (Supp. S2.5). **III**, $N > 2D$: $c^\star > 0$ is not known in closed form in general. We use the smallest upper bound $U \geq c^\star$ certified by an explicit configuration of our closed-form family (antipodal simplices, Paley and Steiner ETF blocks, and known codes; appendix B.4), and a lower bound $L \leq c^\star$ from Levenshtein (1998). The bracket is wide just above $2D$, so beyond $2D$ we claim only that the level is attainable and of the right order. Every benchmark of table 1 has $N \leq 2D$, where $c^\star$ is exact. The level is computed from (N, D), never tuned.

4.2 Step 2 (R2): which pairs to move

For one pair with cosine s and a level c, the squared residual $(s - c)^2$ has the force $2(s - c)$, which is negative below c: it pulls pairs that are already farther apart back *up* to the level. The ReLU residual $\big[\mathrm{ReLU}(s - c)\big]^2$, with $\mathrm{ReLU}(x) = \max(x, 0)$ applied *before* squaring, has the force $2\,\mathrm{ReLU}(s - c)$ and applies no force there.

Proposition 6 (Why a ReLU on the residual; proof in appendix E.5)**.** *Let $c < 1$.* (i) *The pairs on which the ReLU residual applies no force are those with $s \leq c$; all pairs are at rest exactly on $Z(c)$, which is non-empty if and only if $c \geq c^\star$.* (ii) *All pair forces of the two-sided residual vanish only where all cosines equal c, which is impossible for every $c < 1$ when $N \geq D + 2$: on ImageNet with ViT-B/16 or ViT-L/14 its push never switches off (its minimizers exist, but at each of them some pair is still pulled or pushed).* (iii) *Even where both residuals have the same unique zero ($N \leq D + 1$, $c = c^\star$), the two-sided one pulls separated pairs back when this is not needed (fig. 2).*

4.3 The loss, and why each part is there

The two steps act on each entry of the Gram matrix $S = TT^\top$. Let $G_c = (1 - c)I + c\,\mathbf{1}\mathbf{1}^\top$, the matrix with ones on the diagonal and c elsewhere. The loss of AGI-TPT is

$$\mathcal{L}_{\text{AGI}}(T) = \frac{1}{N^2} \left\| \mathrm{ReLU}\left(S - G_{c^\star}\right) \right\|_F^2 = \frac{1}{N^2} \sum_{i \neq j} \Big[\mathrm{ReLU}\left(s_{ij} - c^\star\right) \Big]^2, \tag{1}$$

with ReLU acting entrywise ($\|A\|_F^2 = \sum_{i,j} a_{ij}^2$; the diagonal of $S - G_{c^\star}$ is 0). For $N > 2D$ the certified U replaces $c^\star$. AGI-TPT adds $\lambda \mathcal{L}_{\text{AGI}}$ to TPT's loss (Algorithm 1); the extra cost is one $N \times N$ Gram product per step. Each part has one job. The *level* implements (R1). The *ReLU*

implements (R2). The *square* makes the loss continuously differentiable, so the force grows smoothly from 0 at the level instead of jumping, and it makes the loss a convex function of S, which the landscape result below needs. The factor $\frac{1}{N^2}$ averages over the N^2 entries, so the loss stays in $[0, (1-c^\star)^2]$ for every N. O-TPT's penalty is the same expression with level 0, no ReLU and no $\frac{1}{N^2}$: $\|S-I\|_F^2 = \|S-G_0\|_F^2$. *Implementation.* Everything else is TPT's: 4 context tokens initialized to "a photo of a", 64 views, the entropy of the 10% most confident, one AdamW step per image (learning rate 5×10^{-3}); online, the prompt is carried across images. AGI-TPT adds the Gram matrix and $\lambda\mathcal{L}_{\text{AGI}}$ with $\lambda = 4$, selected once on a CIFAR-100 stream; λ sets the override mass and hence the one-step calibration (corollary 2, appendix F.4).

Guarantees (for free class vectors; proofs in appendix E).

Theorem 7 (Zeros and landscape)**.** (i) *The zeros of $\mathcal{L}_{\text{AGI}}$ are $Z(c^\star)$: the regular simplex for $N \le D+1$, and for $D+2 \le N \le 2D$ exactly the configurations with the largest possible one-vs-rest separation $\min_k \operatorname{dist}(t_k, \operatorname{conv}\{t_j\}_{j\neq k}) = 1$, the Softmax Codes of Jiang et al. (2024).* (ii) *Let the level be at least $-\frac{1}{N-1}$. Every local minimizer T of $\mathcal{L}_{\text{AGI}}$ with $\operatorname{rank} T < D$ is a zero of $\mathcal{L}_{\text{AGI}}$. In particular, for $N < D$ every local minimizer is global.*

Theorem 8 (Merged classes are never local minima)**.** *Let $N \le 2D-1$ and let two classes coincide, $t_i = t_j$. Then T is not a local minimizer of $\mathcal{L}_{\text{AGI}}$ at the level $c^\star$: splitting the pair symmetrically along a suitable direction lowers the loss at second order.*

Theorem 7(ii) is the rank-deficient analogue of what Benedetto & Fickus (2003) prove for O-TPT's penalty at every rank; its hypothesis holds for $c^\star$ and for every $U \ge c^\star$, and full-rank spurious minima for $N \ge D$ remain open. Theorem 8 matters because the push of pair (i,j) on t_i has length $\propto \operatorname{ReLU}(s_{ij}-c)\sin\theta_{ij}$, which vanishes as the pair merges: the merged set $\{t_i = t_j\}$ is invariant under the exact gradient flow, so the theorem gives instability under perturbation, not escape. *What transfers to a prompt.* $T(\boldsymbol{\pi})$ lies on a d_π-dimensional submanifold of $(S^{D-1})^N$, and theorems 7 and 8 do not transfer: with a 4-token prompt the largest ImageNet cosine converges to 0.44, so the zero set is out of reach (Supp. S5.5). Everything stated about forces does transfer by the chain rule: propositions 1 and 9 and corollary 2, zero force at or below the level (proposition 6(i)) and the positive force below $c^\star$ (proposition 10(ii)). Under a prompt AGI-TPT stops pairs one by one (section 5.3).

At one step. Every zero-shot CLIP pair lies far above the level (fig. S5), so the ReLU is inactive at TPT's single step and the loss reduces to a familiar form.

Proposition 9 (One-step form; proof in appendix E)**.** *If every pair satisfies $s_{ij} > c$, then $\nabla\mathcal{L}_{\text{AGI}} = \nabla\left[\frac{1}{N^2}\|S-I\|_F^2\right] - 2c\,\nabla\|\bar{t}\|^2$. For $c = 0$ the one-step offset of AGI-TPT is exactly that of O-TPT's Frobenius penalty; for $c < 0$ it adds a squared C-TPT dispersion push with weight $2|c|\lambda$.*

With proposition 1, at equal strength AGI-TPT and O-TPT's penalty (without its Householder step) give *identical* predictions on ImageNet with ViT-B/16 ($c^\star = 0$), and differ elsewhere only through the small dispersion term (section 5.1).

5 Experiments

Design. Each experiment tests a prediction of sections 2 to 4 against a named alternative. *Data and models*: the TPT benchmark of 11 datasets and four ImageNet shifts (Shu et al., 2022; Yoon et al., 2024); CLIP ViT-B/16 (D=512), RN50 (D=1024) and ViT-L/14 (D=768); and, to vary N/D, FER-2013, CIFAR-10, STL-10 and ImageNet-21k class subsets. *Settings*: *episodic*, one TPT step per image, then a reset (5 seeds); *online*, the prompt carried along the test stream (10 stream orders). *Weights*: $\lambda = 4$ everywhere, and every baseline's weight in our online runs, are chosen on the same held-out stream (CIFAR-100, never a benchmark); our ViT-L/14 runs use published weights and table 7 uses override-matched weights. *Statistics*: 95% paired bootstrap intervals; a difference is *resolved* when its interval excludes 0; the *noise floor* is the ECE that a perfectly calibrated predictor with the same confidences would show on the same test set. Published baseline numbers are copied from each paper; other baseline cells are our runs of the official code (‡; appendix F.5).

Table 1: **Main results** (CLIP ViT-B/16; accuracy and ECE in %). The episodic block (one step per image, then reset) and the online block (adaptation carried along the stream) are separate protocols and are not compared with each other; bold marks the best per column within each block. Baselines as published: zero-shot CLIP, TPT and C-TPT from Yoon et al. (2024, Tables 1–2), O-TPT (Sharifdeen et al., 2025, Tables 2, 4), D-TPT (Han & Hwang, 2025, Tables 2, 4), A-TPT (Ahamed et al., 2026, Tables 2–3), DynaPrompt (Xiao et al., 2025, Tables 1–2) and PTA (Huang et al., 2026, Table 13 for the 10 datasets, Table 2 for ImageNet and the shifts); "—": not reported. †: mean over the 10 datasets without ImageNet (AGI-TPT online on the same 10: 68.60; PTA's 11-dataset mean with its ensemble ImageNet cell: 67.96). §: PTA's 80-prompt ensemble (its only setting on ImageNet and the shifts); all other cells use "a photo of a". ECE: our rows (AGI-TPT) use 15 equal-width bins; published ECE rows are copied as printed and were computed with each paper's own code (C-TPT's uses 20 bins; appendix F.5); with 20 bins our averages are 1.88 (episodic) and 1.44 (online), so gaps of a few tenths across rows are not resolved. Episodic AGI-TPT: mean of 5 seeds (s.d. of the averages 0.05 acc. / 0.04 ECE); online: mean of 10 orders (appendix G.4).

		Fine-grained and generic classification												ImageNet shifts				
	Method	ImNet	DTD	Flowers	Food	SUN	Aircraft	Pets	Caltech	UCF	EuroSAT	Cars	**Avg.**	IN-A	IN-V2	IN-R	IN-S	**Avg.**
Accuracy ↑	Zero-shot CLIP	66.7	44.3	67.3	83.6	62.5	23.9	88.0	92.9	65.0	41.3	65.3	63.7	47.8	60.8	74.0	46.1	57.2
	TPT (episodic)	**69.0**	**46.7**	69.0	**84.7**	65.5	23.4	87.1	93.8	**67.3**	42.4	**66.3**	65.0	**52.6**	**63.0**	**76.7**	47.5	59.9
	C-TPT (episodic)	68.5	46.0	69.8	83.7	64.8	24.0	88.2	93.6	65.7	43.2	65.8	64.8	51.6	62.7	76.0	47.9	59.6
	O-TPT (episodic)	67.33	45.68	70.07	84.13	64.23	23.64	87.95	**93.95**	64.16	42.84	64.53	64.41	49.87	61.65	72.55	47.12	57.80
	D-TPT (episodic)	67.39	44.96	**71.54**	83.11	64.59	23.94	88.29	93.09	66.71	44.02	64.25	64.72	49.12	61.35	74.05	46.97	57.87
	A-TPT (episodic)	67.70	45.51	69.22	83.64	**66.04**	23.76	**88.33**	93.87	66.16	44.06	65.78	64.92	50.39	60.90	74.87	46.09	58.06
	AGI-TPT (episodic)	68.42	46.34	69.87	84.21	65.63	**24.12**	88.06	93.84	66.41	**44.38**	65.92	**65.20**	52.31	62.97	76.58	**48.06**	**59.98**
	DynaPrompt (online)	69.61	47.96	69.95	85.42	66.32	24.33	88.28	94.32	68.72	42.28	67.65	65.89	56.17	64.67	78.17	48.22	61.81
	PTA (online)	70.28§	48.64	**72.63**	86.11	67.22	24.84	89.42	93.83	70.63	**55.46**	**68.46**	67.72†	61.15§	64.85§	**80.79**§	51.00§	64.45§
	AGI-TPT (online)	**70.93**	**50.62**	72.47	**86.58**	**68.61**	**26.94**	**90.57**	**94.52**	**72.18**	55.09	68.38	**68.81**	**62.34**	**65.41**	80.52	**52.73**	**65.25**
ECE ↓	Zero-shot CLIP	2.12	8.50	3.00	2.39	2.53	5.11	4.37	5.50	3.59	7.40	4.25	4.43	8.61	3.01	3.58	4.95	5.04
	TPT (episodic)	10.6	21.2	13.5	3.98	11.3	16.8	5.77	4.51	13.0	21.5	5.16	11.6	16.4	11.1	4.36	16.1	12.0
	C-TPT (episodic)	3.15	11.9	5.04	3.43	5.04	4.36	1.90	4.24	2.54	13.2	1.59	5.13	8.16	6.23	1.54	7.35	5.82
	O-TPT (episodic)	1.96	7.88	3.87	1.46	4.93	3.68	1.90	3.80	2.34	12.98	1.78	4.23	7.22	3.97	1.46	6.87	4.88
	D-TPT (episodic)	2.73	5.20	2.82	4.07	4.50	3.74	2.38	5.43	2.65	9.78	2.62	4.18	6.35	2.66	3.85	**2.47**	3.83
	A-TPT (episodic)	1.45	4.76	3.61	1.37	3.28	3.14	**1.17**	2.76	2.12	3.92	**1.09**	2.61	6.45	2.96	1.39	4.87	3.92
	AGI-TPT (episodic)	**0.93**	**3.38**	**2.45**	**0.97**	**1.25**	**2.26**	1.44	**1.44**	**2.05**	**2.52**	1.19	**1.81**	**4.12**	**2.21**	**1.12**	3.27	**2.68**
	AGI-TPT (online)	**0.58**	**2.74**	**2.09**	**0.61**	**0.79**	**1.92**	1.19	**1.17**	**1.70**	**1.33**	**1.08**	**1.38**	**1.92**	**1.36**	**0.83**	**1.49**	**1.40**

5.1 Main results

Main results. The two blocks of table 1 are separate protocols and are compared only within a block. *Episodic*: AGI-TPT has the highest average accuracy (65.20%) and the lowest ECE, 1.81 against 2.61 for A-TPT, the best-calibrated published method (31% lower); most of this one-step gain is a gain of strength (below). *Online*: 68.81% at ECE 1.38 on the 11 datasets and 65.25% at 1.40 on the shifts, above the online prompt-tuning baselines where they report (68.60% vs. 67.72% for PTA on 10 datasets with the same template; 65.25% vs. 64.45% on the shifts, where PTA uses an 80-prompt ensemble), leading on 8 of 11 datasets and 3 of 4 shifts. Neither online baseline reports calibration; the online calibration comparison is table 9. We do not claim the state of the art for test-time adaptation of CLIP at large: cache- and prototype-based online methods with the 80-template ensemble report higher averages in PTA's tables (Huang et al., 2026, Table 1); they tune no prompt, report no calibration, and we did not run them. Online, the ECE is within 0.15 of the noise floor on every dataset (debiased: 0.50; appendix G.4). Our ECE uses 15 bins; published rows use each paper's code (appendix F.5), so gaps of a few tenths across rows are not resolved.

One step: calibration is set mainly by strength. Published weights differ in strength. At equal strength, propositions 1 and 9 predict that AGI-TPT and O-TPT's penalty without its Householder step make identical predictions on ImageNet ($c^\star = 0$), and nearly identical ones elsewhere. They do (table 7): they agree to the last digit on ImageNet and differ by 0.03 ECE over 11 datasets. Most of the gap to the other regularizers closes (C-TPT 2.38, D-TPT 2.26, A-TPT 2.57 against 1.81, with the differences resolved on ImageNet and, for A-TPT, on six datasets), and a random flip set of the same strength reaches 1.94, within 0.13 of AGI-TPT. One-step calibration is therefore weak evidence for any geometry, including ours; at one step AGI-TPT's edge over the flip set is mainly accuracy (65.20 vs. 64.30). Geometry pays off when adaptation continues.

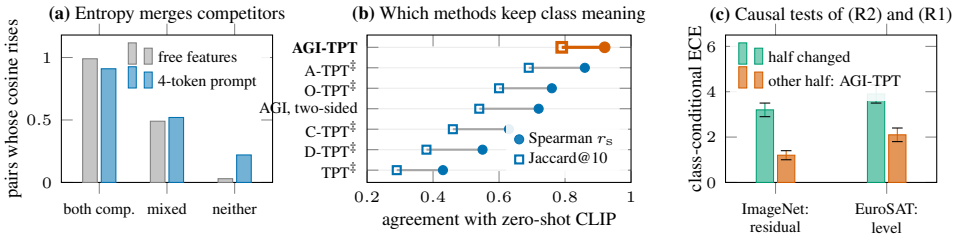


Figure 3: **Why AGI-TPT works** (ViT-B/16; our runs). **(a)** Test of proposition 3: after one entropy step, competitive pairs move together and non-competitive pairs separate. **(b)** How much of zero-shot CLIP's class-similarity structure survives the online stream (six datasets): Spearman correlation of all pairwise class cosines, and overlap of each class's 10 nearest classes. **(c)** Causal tests inside one online run: only the half of the classes whose loss is changed (two-sided residual on ImageNet; level 0 on EuroSAT) loses calibration.

5.2 Sustained adaptation

Prediction (propositions 3, 6 and 10): online, TPT collapses; a push whose forces never switch off, or a two-sided one, drifts; a push that stops at an attainable level does not. *Alternative*: any tuned spread regularizer or drift remedy does as well. Table 9 and fig. 1b support the prediction. TPT collapses (21.4% accuracy); resets and SAR leave ECE above 13; the tuned spread regularizers keep the accuracy but their ECE grows along the stream; the best, A-TPT with SAR, reaches 2.9. AGI-TPT reaches 1.4, lower on all six datasets (each resolved), with the highest accuracy (56.8%) and no drift (ECE_{end} 1.3). A neural-collapse ETF target (Kothapalli, 2023; Markou et al., 2024) reaches 4.0: unreachable for $N > D{+}1$, and two-sided (Supp. S2.5). The lead grows on harder streams: under label shift AGI-TPT reaches 2.1 against 5.4, and over three passes it stays at 1.5 while every other method keeps drifting (Supp. S6.7).

5.3 Why it works: the mechanism, tested directly

Entropy merges competitors (fig. 3a). After one small entropy step on ImageNet, the cosine of two competitive classes rises for 91% of such pairs with a 4-token prompt (99% with free features), and 78% of non-competitive pairs separate, as proposition 3 predicts. Along the stream the same mechanism precedes class absorption (Supp. S6.2).

The ReLU keeps what the classes mean (fig. 3b). A two-sided residual attracts every pair below the level, the most dissimilar ones hardest (Proposition C.2); the ReLU leaves them alone (proposition 6). We measure how much of zero-shot CLIP's similarity structure survives the stream: the Spearman correlation r_s of all pairwise class cosines, and the overlap of each class's 10 nearest classes (Jaccard@10). Structure is lost in the order of how much each method moves already-separated pairs: D-TPT keeps the least ($r_s = 0.55$), then C-TPT, O-TPT and A-TPT (0.63, 0.76, 0.86); AGI-TPT keeps the most (0.92). Replacing only its ReLU by the two-sided residual lowers r_s to 0.72 on every dataset, costs 1.1 accuracy points, and more than doubles the ECE (1.4 $\rightarrow$ 3.4; Supp. S6.3).

The ReLU and the level, tested causally (fig. 3c). Within one online run we split the classes at random into two halves sharing weight, prompt and stream, and change the loss only for pairs inside one half. With the two-sided residual there (R2, ImageNet), 27% of those pairs' crossings are pulled back above the level, against 2% in the other half, and their classes' ECE is 3.2 against 1.2. With level 0 instead of $c^\star = -\frac{1}{9}$ (R1, EuroSAT), those classes stop early and stay over-confident (3.9 against 2.1). A strength-based account predicts no difference between halves.

5.4 Each regime needs its own level

On all 18 datasets we moved the level to $c^\star + \xi$, $\xi \in \{0, \pm 0.02, \pm 0.05\}$, over 10 stream orders, predicting (P1) a minimum at $\xi = 0$ and (P2) that a level below $c^\star$ costs more than one equally far above it, since its push never switches off (proposition 10). Both hold on all 18 (fig. 4a, table 12; resolved against $\xi = -0.02$ on 14).

Datasets on both sides of each regime boundary (fig. 4b, appendix G.7): *where* $N \leq D{+}1$, level 0 instead of $c^\star = -\frac{1}{N-1}$ costs 3.2 ECE points on FER-2013 and 1.2–1.9 on the 10-class datasets, and matching its summed force to ours removes under half of this (table S17). *Where* $D{+}2 \leq N \leq 2D$, level 0 is exact and the simplex level costs 0.01, the two differing by at most $\frac{1}{D+1}$. *Beyond* $2D$ (ImageNet-21k subsets, three backbones), level 0 lies below $c^\star$ and costs up to 3.7 points while the certified U keeps ECE flat (appendix G.8).

5.5 Ablations

One component at a time (table 11): at one step all variants are within 0.4 ECE points and the two residuals coincide (Proposition C.2). Online the ingredients rank clearly. *Having a stop* matters most: no level costs 5.8 points, and annealing without a level recovers only part of this (2.2–2.4). *The one-sided hinge* is next: the two-sided residual costs 2.0. *Which stop* matters least here: against AGI-L0 (level 0 with our ReLU), $c^\star$ gains 0.4, almost all on EuroSAT; for $N \geq 37$ the levels differ by under 0.03 and level 0 costs at most 0.13. Nearest-pair and per-class variants are worse (Supp. S6.4).

6 Related work

Test-time adaptation of CLIP. TPT (Shu et al., 2022) tunes a prompt per image; DiffTPT (Feng et al., 2023), TDA (Karmanov et al., 2024), DPE (Zhang et al., 2024), PTA (Huang et al., 2026) and ZERO (Farina et al., 2024) change the views, keep caches or prototypes of past test images, or change the update. C-TPT (Yoon et al., 2024), O-TPT (Sharifdeen et al., 2025), D-TPT (Han & Hwang, 2025) and A-TPT (Ahamed et al., 2026) regularize the text geometry for calibration after one step; Murugesan et al. (2024) link miscalibration to logit ranges. Closest and concurrent is SoC (Fillioux et al., 2026): from a confidence bound it argues that driving the largest cosine to 0 over-confidences the model, and replaces O-TPT's penalty by a two-sided Huber penalty on min–max-normalized cosines with a data-dependent threshold and a per-image weight, evaluated after one step (not run by us). We share the view that level 0 is the wrong stop; our level is the attainable optimum computed from (N, D), our penalty is one-sided, our zero set and landscape are characterized, and our evidence is under sustained adaptation, where one-step comparisons cannot separate geometries (proposition 1). Drift remedies (Niu et al., 2023; Press et al., 2023; Wang et al., 2022) limit how far a model moves.

Class geometry. Uniformity objectives and fixed codes spread classes on the sphere (Zhang et al., 2017; Liu et al., 2018; Wang & Isola, 2020; Hoffer et al., 2018; Mettes et al., 2019; Kasarla et al., 2022; Yang et al., 2022). Trained classifiers collapse to a simplex ETF (Papyan et al., 2020), which needs $N \leq D{+}1$ (Kothapalli, 2023; Markou et al., 2024); beyond, the Softmax Codes of Jiang et al. (2024) are our zeros in regime II. The worst-pair problem is Tammes's (Tammes, 1930; Rankin, 1955; Levenshtein, 1998); the squared-hinge pair term and the codes are classical.

7 Limitations and conclusion

Limitations. At one step the calibration gain is mainly one of strength. The guarantees concern free class vectors (section 4.3). On CLIP the level's benefit over level 0 with our ReLU is concentrated on few-class datasets and beyond $2D$, where the level is only certified. A sustained push over-separates genuinely confusable classes (Supp. S7.1 and appendix G.9); encoder biases are re-arranged, not removed.

Conclusion. We showed that at Adam's first step every fixed-weight spread regularizer adds one image-independent offset, so one-step calibration is set mainly by the override mass (proposition 1 and corollary 2), and that entropy minimization pulls competing classes together, which is why online TPT collapses (proposition 3). A regularizer acting for thousands of steps is therefore judged by where it stops and which pairs it moves (R1, R2). AGI-TPT stops each pair at the attainable level $c^\star(N, D)$, exact up to $2D$ classes and certified beyond, through a ReLU residual that applies no force below it. For free class vectors its zeros are the optimal codes, its rank-deficient local minima are global, and merged classes are never local minima for $N \leq 2D - 1$ (theorems 7 and 8); at one step it reduces to O-TPT's penalty plus a dispersion term (proposition 9). Like for like it is the most accurate and best calibrated of the methods compared (ECE 1.81 vs. 2.61 at one step, 1.4 vs. 2.9 online), keeps the most class structure ($r_{\mathrm{s}} = 0.92$), and does not drift on label-shifted or repeated streams.

Ethics statement

This work proposes a regularizer that improves the calibration of vision–language models adapted at test time. It introduces no new data and no experiments on human subjects. Better-calibrated confidence supports safer abstention in decision-support settings; the method does not remove biases present in the pretrained text encoder (section 7).

Reproducibility statement

Appendices B–E prove every result stated in the main text, with two exceptions that are in the separate *Supplementary Material for AGI-TPT* (`AGI-TPT-supplementary`): the frame-potential uniqueness in proposition 11(ii) and the angle-sum lemma behind fig. 2. Appendices F–G give the method, the protocol and the main result tables; the supplement holds further certificates and experiments. The loss used for every result is eq. (1) with $\lambda = 4$ and the level of table 4; it has no other hyperparameter. Appendix F gives Algorithm 1, the level for every (N, D), and the full protocol: strength matching, held-out weight selection for every method, stream orders (seeds 0–9), the ImageNet-21k subsets, statistics and compute. The code will be made publicly available upon acceptance, together with the configuration of every run, per-seed and per-order results, the class-subset lists and the explicit certificate configurations of appendix B.4.

AI use statement

A large language model assisted with drafting and editing the text, with literature search, and with checking derivations and the arithmetic of the tables. The authors verified every mathematical statement, citation and result, and take full responsibility for the content.

References

Shihab Aaqil Ahamed, Udaya Sampath K. Perera Miriya Thanthrige, Ranga Rodrigo, and Muhammad Haris Khan. A-TPT: Angular diversity calibration properties for test-time prompt tuning of vision-language models. In *International Conference on Learning Representations*, pp. 69175–69199, 2026.

Abdo Y. Alfakih. On theorems of Šiňajová, Rankin and Kuperberg concerning spherical point configurations. *Applied Set-Valued Analysis and Optimization*, 5(2):155, 2023.

Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von Mises–Fisher distributions. *Journal of Machine Learning Research*, 6: 1345–1382, 2005.

John J. Benedetto and Matthew Fickus. Finite normalized tight frames. *Advances in Computational Mathematics*, 18(2):357–385, 2003. doi: 10.1023/A:1021323312367.

Jarosław Błasiok and Preetum Nakkiran. Smooth ECE: Principled reliability diagrams via kernel smoothing. In *International Conference on Learning Representations*, pp. 1679–1707, 2024.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 215–223, 2011.

Henry Cohn, David de Laat, and Nando Leijenhorst. Optimality of spherical codes via exact semidefinite programming bounds. *arXiv preprint arXiv:2403.16874*, 2024.

Philippe Delsarte, Jean-Marie Goethals, and Johan J. Seidel. Spherical codes and designs. *Geometriae Dedicata*, 6(3):363–388, 1977.

Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 129062–129093, 2024.

Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2704–2714, 2023.

Matthew Fickus and Dustin G. Mixon. Tables of the existence of equiangular tight frames. *arXiv preprint arXiv:1504.00253*, 2015.

Matthew Fickus, Dustin G. Mixon, and Janet C. Tremain. Steiner equiangular tight frames. *Linear Algebra and its Applications*, 436(5):1014–1027, 2012.

Leo Fillioux, Omprakash Chakraborty, Ismail Ben Ayed, Paul-Henry Cournède, Stergios Christodoulidis, Maria Vakalopoulou, and Jose Dolz. SoC: Semantic orthogonal calibration for test-time prompt tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2601.08617.

Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing (ICONIP)*, pp. 117–124, 2013.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017.

Jisu Han and Wonjun Hwang. D-TPT: Dimensional entropy maximization for calibrating test-time prompt tuning in vision-language models. *arXiv preprint arXiv:2510.09473*, 2025.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Fix your classifier: the marginal value of training the last weight layer. In *International Conference on Learning Representations*, 2018.

Zhaohong Huang, Yuxin Zhang, Wenjing Liu, Fei Chao, and Rongrong Ji. Prototype-based test-time adaptation of vision-language models. In *International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Jiachen Jiang, Jinxin Zhou, Peng Wang, Qing Qu, Dustin G. Mixon, Chong You, and Zhihui Zhu. Generalized neural collapse for a large number of classes. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14162–14171, 2024.

Tejaswi Kasarla, Gertjan J. Burghouts, Max van Spengler, Elise van der Pol, Rita Cucchiara, and Pascal Mettes. Maximum class separation as inductive bias in one matrix. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19553–19566, 2022.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Vignesh Kothapalli. Neural collapse: A review on modelling principles and generalization. *Transactions on Machine Learning Research*, 2023.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Włodzimierz Kuperberg. Optimal arrangements in packing congruent balls in a spherical container. *Discrete & Computational Geometry*, 37(2):205–212, 2007. doi: 10.1007/s00454-006-1303-6.

Vladimir I. Levenshtein. Universal bounds for codes and designs. In Vera S. Pless and W. Cary Huffman (eds.), *Handbook of Coding Theory*, volume 1, chapter 6, pp. 499–648. Elsevier, 1998.

Weiyang Liu, Rongmei Lin, Zhen Liu, Lixin Liu, Zhiding Yu, Bo Dai, and Le Song. Learning towards minimum hyperspherical energy. In *Advances in Neural Information Processing Systems*, volume 31, pp. 6222–6233, 2018.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Evan Markou, Thalaiyasingam Ajanthan, and Stephen Gould. Guiding neural collapse: Optimising towards the nearest simplex equiangular tight frame. In *Advances in Neural Information Processing Systems*, volume 37, pp. 35544–35573, 2024.

Pascal Mettes, Elise van der Pol, and Cees G. M. Snoek. Hyperspherical prototype networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Balamurali Murugesan, Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz. Robust calibration of large vision-language adapters. In *European Conference on Computer Vision (ECCV)*, pp. 147–165, 2024.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *AAAI Conference on Artificial Intelligence*, volume 29, 2015.

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*, 2023.

Raymond E. A. C. Paley. On orthogonal matrices. *Journal of Mathematics and Physics*, 12(1–4): 311–320, 1933.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Ori Press, Steffen Schneider, Matthias Kümmerer, and Matthias Bethge. RDumb: A simple approach that questions our progress in continual test-time adaptation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 39915–39935, 2023.

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*, pp. 8748–8763, 2021.

Robert A. Rankin. The closest packing of spherical caps in n dimensions. *Proceedings of the Glasgow Mathematical Association*, 2(3):139–144, 1955. doi: 10.1017/S2040618500033219.

Ashshak Sharifdeen, Muhammad Akhtar Munir, Sanoojan Baliah, Salman Khan, and Muhammad Haris Khan. O-TPT: Orthogonality constraints for calibrating test-time prompt tuning in vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19942–19951, 2025.

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. In *Advances in Neural Information Processing Systems*, volume 35, pp. 14274–14289, 2022.

Thomas Strohmer and Robert W. Heath Jr. Grassmannian frames with applications to coding and communication. *Applied and Computational Harmonic Analysis*, 14(3):257–275, 2003.

Mátyás A. Sustik, Joel A. Tropp, Inderjit S. Dhillon, and Robert W. Heath Jr. On the existence of equiangular tight frames. *Linear Algebra and its Applications*, 426(2–3):619–635, 2007.

Pieter Merkus Lambertus Tammes. On the origin of number and arrangement of the places of exit on the surface of pollen-grains. *Recueil des travaux botaniques néerlandais*, 27(1):1–84, 1930.

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)*, pp. 7191–7201, 2022.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939, 2020.

Lloyd R. Welch. Lower bounds on the maximum cross correlation of signals. *IEEE Transactions on Information Theory*, 20(3):397–399, 1974.

Zehao Xiao, Shilin Yan, Jack Hong, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiayi Shen, Qi Wang, and Cees G. M. Snoek. DynaPrompt: Dynamic test-time prompt tuning. In *International Conference on Learning Representations*, pp. 52489–52505, 2025.

Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? In *Advances in Neural Information Processing Systems*, volume 35, pp. 37991–38002, 2022.

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang D. Yoo. C-TPT: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *International Conference on Learning Representations*, pp. 8636–8660, 2024.

Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Dual prototype evolving for test-time generalization of vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 32111–32136, 2024.

Xu Zhang, Felix X. Yu, Sanjiv Kumar, and Shih-Fu Chang. Learning spread-out local feature descriptors. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 4595–4603, 2017.

ORGANIZATION OF THE APPENDIX AND THE SUPPLEMENTARY MATERIAL

The appendices below state and prove every result of the main text (appendices B to E), with two exceptions kept in the supplement for space, the frame-potential uniqueness of proposition 11(ii) and the angle-sum lemma behind fig. 2; they give the method and the full protocol (appendix F), and report the main experimental tables and the essential ablations (appendix G). Secondary material is in a separate PDF, *Supplementary Material for AGI-TPT* (`AGI-TPT-supplementary`): further certificates for the level, the Welch bound, exact configurations in $\mathbb{R}^3$, the three-vector angle lemma, the selection of λ, and further experiments (per-dataset tables for other backbones, one-step null controls, the mechanism along the stream, class-similarity structure, robustness, other vision–language models and a synthetic testbed). Its sections, tables, figures and results carry the prefix S (e.g. Supp. S2, table S16) and link to that file.

A NOTATION

Symbol	Meaning
N, D, d_π	number of classes, embedding dimension, number of prompt parameters
t_k, T, $S = TT^\top$, s_{ij}	unit class features, their matrix, Gram matrix, pairwise cosines
$\bar{t} = \frac{1}{N}\sum_k t_k$	mean class feature
$c^\star(N, D)$	attainable level: smallest largest-pairwise cosine of N unit vectors in $\mathbb{R}^D$, proposition 11
$U(N, D)$	certified upper bound on $c^\star$, used in place of $c^\star$ when $N > 2D$
$\bar{c}(N, D)$	the level in all regimes: $c^\star$ for $N \le 2D$, U for $N > 2D$ (used only in figures and tables of levels)
$Z(c)$	configurations whose largest cosine is at most c: $\{T : \max_{i\neq j} s_{ij} \le c\}$
G_c	$(1-c)I + c\,\mathbf{1}\mathbf{1}^\top$: ones on the diagonal, c elsewhere
$\|A\|_F^2$	squared Frobenius norm $\sum_{i,j} a_{ij}^2$
$\mathcal{L}_{\text{AGI}}$	AGI-TPT's penalty $\frac{1}{N^2}\|\,\mathrm{ReLU}(S - G_{c^\star})\|_F^2 = \frac{1}{N^2}\sum_{i\neq j}\left[\mathrm{ReLU}(s_{ij} - c^\star)\right]^2$ (eq. (1)), with U in place of $c^\star$ if $N > 2D$
φ, $F = \varphi'$	pair penalty and the force it applies to a pair
g_x, $b_{\mathcal{R}}$, O_x	entropy gradient, regularizer offset, override set
μ, ν	override mass $\mathbb{E}_x\|g_x\|_{O_x}\|_1/\|g_x\|_1$ (matched), override fraction $\mathbb{E}_x\|O_x\|/d_\pi$
gap	signed calibration gap: mean confidence minus accuracy
$\rho(T)$	one-vs-rest separation $\min_k \mathrm{dist}(t_k, \mathrm{conv}\{t_j\}_{j\neq k})$
d_k	distance from t_k to its nearest other class feature, $\min_{j\neq k}\|t_k - t_j\|$
ξ	offset of a tested level from $c^\star$: level $c^\star + \xi$ (section 5.4)
r_s	Spearman rank correlation of pairwise class cosines with zero-shot CLIP's (Supp. S6.3)
δ_{jk}	Kronecker delta

Each symbol has one meaning throughout. c denotes a generic level; $c^\star$ is the attainable level, and U replaces it only when $N > 2D$. Squares of a ReLU are always written with brackets, $\left[\mathrm{ReLU}(x)\right]^2$: the ReLU is applied first, then the result is squared.

Table 2: **The level $\bar{c}(N, D)$ used by AGI-TPT ($= c^\star$ up to $2D$, certified U beyond) and the geometry that attains it** (proposition 11; proofs, certificates and values for $D \in \{512, 768, 1024\}$ in appendix B). $W_\pm = W(\lceil N/2 \rceil, D)$ bounds every antipodal configuration; L is Levenshtein's lower bound on $c^\star$.

Regime	$\bar{c}(N, D)$	Geometry $T^\star$	Status	In this paper
I $N \le D{+}1$	$-\frac{1}{N-1}$	regular simplex (simplex ETF)	exact, unique	15 of 18 datasets (ViT-B/16); all (RN50)
II $D{+}2 \le N \le 2D$	0	orthogonal sum of $N{-}D$ regular simplices	exact; unique FP-optimum; not an ETF	ImageNet, -V2, -Sketch (ViT-B/16, ViT-L/14)
IIIa $2D < N \le 2D{+}2k$	$1/\lfloor D/k \rfloor$	$\oplus$ of k antipodal simplices	certified; U/L up to 512 at $N = 2D{+}2$	ImageNet-21k subsets
IIIb $N \le 2D{+}2d$	$(2d{-}1)^{-1/2}$	antipodal Paley ETF $\oplus$ cross-polytope	certified; 1.0–4.1 $W_\pm$	ImageNet-21k subsets
III Gerzon cells	$(D{+}2)^{-1/2}$	antipodal Gerzon ETF	exact	synthetic
IIIc $N \gtrsim 4D$	$D^{-1/2}, (2/D')^{1/2}$	Kerdock, Gold / dual BCH (*not* ETF)	certified; $U/L \approx 2$–11	ImageNet-21k

B THE ATTAINABLE LEVEL: CONSTRUCTION, CERTIFICATES AND PROOFS

B.1 STATEMENTS OF SECTION 4.1

The main text summarizes the following results; their proofs are in appendix B.5. In this appendix we write $\bar{c}(N, D)$ for the level that AGI-TPT uses in each regime: $\bar{c} = c^\star$ for $N \le 2D$, where $c^\star$ is known exactly, and $\bar{c} = U$, the certified upper bound on $c^\star$ of (iii), for $N > 2D$. $T^\star$ denotes a configuration of least frame potential whose largest cosine is $\bar{c}$.

Proposition 10 (No universal level; the cost of a level below $c^\star$)**.** *In the model of remark 12:* (i) $\mathcal{R}_c$ *is inactive at the Bayes-calibrated classifier with class means m if and only if $c \ge \max_{i \ne j} \langle m_i, m_j \rangle$. Every level $c < 1$ is therefore active at some calibrated classifier, and $c^\star$ is the smallest level that is inactive at at least one (those whose means form an optimal code).* (ii) *If $c < c^\star$, then at every configuration some pair receives a force of at least $2(c^\star - c) > 0$: the push never switches off, whatever the data.* (iii) *$c^\star(n, D) \le c^\star(N, D)$ for $n \le N$. A per-pair level $c \ge c^\star(N, D)$ thus never asks any subset of the classes for a configuration they cannot share, whereas every $c < c^\star(N, D)$ asks it of the full set.*

Proposition 11 (The attainable level)**.** (i) *If $N \le D{+}1$, then $\bar{c} = c^\star = -\frac{1}{N-1}$, and $T^\star$ is the regular simplex, the unique optimal code (the simplex ETF).* (ii) *If $D{+}2 \le N \le 2D$, then $c^\star = 0$, and among optimal codes* FP *is minimized exactly by the orthogonal sum of $N{-}D$ regular simplices in mutually orthogonal blocks of dimensions $\lfloor \frac{D}{N-D} \rfloor$ and $\lceil \frac{D}{N-D} \rceil$, unique up to an orthogonal transformation and the assignment of classes to blocks. It is never equiangular, hence not an ETF of $\mathbb{R}^D$. It is a tight frame if and only if $(N{-}D) \mid D$, which fails for ImageNet at $D = 512$ and 768. No ETF of $\mathbb{R}^D$ is an optimal code.* (iii) *If $N \ge 2D{+}1$, then $c^\star > 0$ and its value is open. Orthogonal sums of k antipodal simplex ETFs certify $c^\star \le 1/\lfloor D/k \rfloor$ for $N \le 2D{+}2k$. An antipodal Paley ETF in a d-dimensional block certifies $c^\star \le (2d-1)^{-1/2}$ for $N \le 2D{+}2d$ whenever $2d-1$ is a prime power $\equiv 1 \pmod 4$. Within orthogonal sums of such blocks, Steiner blocks and a cross-polytope, the smallest of these certificates is optimal (Proposition S2.1).* (iv) *For $D \in \{512, 768, 1024\}$, no real ETF of $N \ge D{+}2$ vectors exists in $\mathbb{R}^D$.*

This appendix proves lemma 5 and propositions 10 and 11 and gives the certificates behind table 2. Every object here is classical except where we say otherwise: the values up to $2D$ are Rankin's, and conference, Steiner, Gerzon and Kerdock codes are known. Our contributions are the frame-potential uniqueness in regime II, optimality within the ETF-built family (Proposition S2.1) and the non-existence result (Proposition S2.3). Throughout, $t_1, \dots, t_N$ are unit vectors in $\mathbb{R}^D$, $s_{ij} = \langle t_i, t_j \rangle$, $M(T) = \max_{i \ne j} s_{ij}$, $c^\star(N, D) = \min_T M(T)$, and

$$W(n, D) = \sqrt{\frac{n - D}{D(n-1)}} \qquad (n > D)$$

is the Welch value. Table S3 summarizes what is exact, what is certified, and which configurations are ETFs.

B.2 THE VALUE OF $c^\star$ FOR $N \le 2D$, AND THE BLOCK STRUCTURE

$N \le D + 1$. For unit vectors, $0 \le \|\sum_k t_k\|^2 = N + \sum_{i \ne j} s_{ij} \le N + N(N-1)M(T)$, so $M(T) \ge -\frac{1}{N-1}$. The same computation shows that the *mean* cosine is at least $-\frac{1}{N-1}$ for

every configuration. Equality forces $\sum_k t_k = 0$ and $s_{ij} = -\frac{1}{N-1}$ for all $i \neq j$. The Gram matrix $\frac{N}{N-1}I - \frac{1}{N-1}J$ is positive semidefinite of rank $N-1$, so it is realized in $\mathbb{R}^D$ if and only if $N-1 \leq D$, uniquely up to an orthogonal map. This is the regular simplex. Its N vectors are equiangular and tight in their span, with $|s| = \frac{1}{N-1} = W(N, N-1)$: a simplex ETF.

Lemma B.1 (Block structure)**.** *Let $s_{ij} \leq 0$ for all $i \neq j$, and join i and j when $s_{ij} < 0$.* (a) *The connected components span mutually orthogonal subspaces.* (b) *A linearly dependent component with m vectors spans exactly $m-1$ dimensions and has a dependence with strictly positive coefficients.* (c) *If r components are dependent, then $N \leq D + r$ and $r \leq D$. Hence $N \leq 2D$, with equality only for the cross-polytope $\{\pm e_1, \ldots, \pm e_D\}$ in some orthonormal basis.*

Proof. (a) Vectors in different components have $s = 0$. (b) Let $\sum_i a_i t_i = 0$ within a component C, with $a \neq 0$. Let $P = \{a_i > 0\}$ and $Q = \{a_i < 0\}$, and set $w = \sum_P a_i t_i = \sum_Q |a_i| t_i$. Then $\|w\|^2 = \sum_{P \times Q} a_i |a_j| s_{ij} \leq 0$, so $w = 0$. Pairing $\sum_P a_i t_i = 0$ with t_k for $k \in C \setminus P$ gives $\sum_P a_i s_{ik} = 0$, a sum of non-positive terms, so $s_{ik} = 0$ for all $i \in P$. By connectivity this forces $P = C$ or $P = \emptyset$, and likewise for Q, so every nonzero dependence has one sign and full support. Two independent such dependences would combine into a nonzero one with a zero coefficient; hence the dependences form a space of dimension at most one, and the rank is $m - 1$. (c) By (a) and (b), $D \geq \sum_C \dim \operatorname{span} C \geq N - r$. Each dependent component spans at least one dimension, so $r \leq D$. If $N = 2D$, then $r = D$, every dependent component spans exactly one dimension and contains two vectors (an antipodal pair), and there is no independent component. □

$D + 2 \leq N \leq 2D$. By Lemma B.1(b) with a single component, at most $D + 1$ vectors can be pairwise strictly obtuse (Rankin, 1955). So $M(T) \geq 0$, and N vertices of the cross-polytope attain 0: $c^\star = 0$. Every optimal configuration therefore decomposes, by Lemma B.1, into at least $N - D$ mutually orthogonal positive bases; see also Kuperberg (2007) and Alfakih (2023, Thm. 1.4).

$N \geq 2D + 1$. By Lemma B.1(c), $M(T) \leq 0$ is impossible, so $M(T) > 0$ for every configuration. By compactness the minimum is attained, and $c^\star > 0$.

B.3 Regime II: the unique ETF-built optimum

Proposition B.2 (No ETF is optimal in regime II)**.** *Let $D + 2 \leq N \leq 2D$. No equiangular set of N unit vectors, and in particular no ETF of $\mathbb{R}^D$, has $M \leq 0$. Every orthogonal sum of regular simplices with at most D dimensions in total has $M = 0$ and is optimal; each block is a simplex ETF of its span.*

Proof. An equiangular set with common value $-a < 0$ is pairwise strictly obtuse, so $N \leq D + 1$ by Lemma B.1. With common value $a > 0$, $M = a > 0$. Within a block of n vectors the cosines are $-\frac{1}{n-1}$, and across blocks they are 0. □

The proof of proposition 11(ii), frame-potential uniqueness in regime II, is in Supp. S2.1. An ETF (neural-collapse) target cannot be reached for $N \geq D + 2$ (Supp. S2.5).

B.4 Regime III: ETF-built certificates

For $N \geq 2D + 1$ no closed form for $c^\star$ is known in general (Cohn et al., 2024). An explicit configuration certifies an upper bound, and a linear-programming bound gives a lower one. Any subset of a certificate, or its embedding in a larger space, has at most the same largest cosine.

Proposition B.3 (Orthogonal sums of antipodal simplex ETFs)**.** *Split $\mathbb{R}^D$ into $k \leq D/2$ orthogonal blocks of dimensions $\lfloor D/k \rfloor$ and $\lceil D/k \rceil$, and place $\{\pm u_0, \ldots, \pm u_d\}$ in each block, where the u_i form a regular simplex of the block. The result has $2D + 2k$ unit vectors with largest cosine $1/\lfloor D/k \rfloor$.*

Proof. Within a block the cosines are $\langle u_i, u_j \rangle = -\frac{1}{d}$, $\langle u_i, -u_j \rangle = \frac{1}{d}$ and $\langle u_i, -u_i \rangle = -1$; across blocks they are 0. A block of dimension d contributes $2(d+1)$ vectors, so the total is $2D + 2k$. □

Proposition B.4 (Antipodal Paley ETF)**.** *Let $q \equiv 1 \pmod 4$ be a prime power and $d = \frac{q+1}{2} \leq D$. There are $2D + 2d$ unit vectors in $\mathbb{R}^D$ with largest cosine $q^{-1/2} = (2d-1)^{-1/2}$.*

Proof. The Paley conference matrix C of order $q+1$ is symmetric, with zero diagonal, off-diagonal entries ± 1, and $C^2 = qI$ (Paley, 1933). Hence $\Gamma = I + C/\sqrt{q}$ has eigenvalues 1 ± 1: it is positive semidefinite of rank $\frac{q+1}{2} = d$ with unit diagonal, the Gram matrix of $q+1$ unit vectors in $\mathbb{R}^d$ with all $|s| = q^{-1/2}$, and $\Gamma^2 = 2\Gamma$ makes them a tight frame. This is an ETF meeting the Welch bound, $W(q+1, d) = q^{-1/2}$ (Strohmer & Heath Jr., 2003; Fickus & Mixon, 2015). Its antipodal double has $4d$ vectors with cosines $\pm q^{-1/2}$ and -1, and placing it in $\mathbb{R}^d$ with a cross-polytope in the orthogonal $\mathbb{R}^{D-d}$ adds $2(D-d)$ vectors at cosine 0 or -1. □

The admissible Paley dimensions in CLIP's dimensions, further certificates, and Levenshtein's lower bound $L \le c^\star$ are in Supp. S2.2.

B.5 Proofs of lemma 5 and propositions 10 and 11

Proof of lemma 5. $Z(c) = \{M \le c\}$ is non-empty if and only if $c \ge \min_T M = c^\star$; the minimum is attained because M is continuous on the compact set of configurations. $Z(c^\star) = \arg\min_T M$ is the set of optimal codes. For $c > c^\star$, move one vector of an optimal code continuously onto another. M varies continuously from $c^\star$ to 1, so it takes a value in $(c^\star, c]$, at a configuration in $Z(c)$ that is not optimal. □

Proof of proposition 10. (i) By remark 12 the classifier with $t = m$ and $\alpha = \kappa$ is Bayes-calibrated. Every pair force $2\,\mathrm{ReLU}(s - c)$ vanishes at m if and only if $\langle m_i, m_j \rangle \le c$ for all $i \ne j$. For $c < 1$ choose means with two cosines above c: the level is active there. A level is inactive at some truth if and only if some configuration has $M \le c$, i.e. $c \ge c^\star$ by lemma 5, with equality exactly at optimal codes. (ii) At every configuration $M(T) \ge c^\star$, so the pair attaining M receives the force $2\,\mathrm{ReLU}(M - c) \ge 2(c^\star - c)$. (iii) Deleting $N - n$ vectors from an optimal N-code leaves an n-configuration with $M \le c^\star(N, D)$, so $c^\star(n, D) \le c^\star(N, D)$ and every subset can sit below $c^\star(N, D)$ at once. For $c < c^\star(N, D)$ the full set cannot, by lemma 5. □

Proof of proposition 11. (i) is appendix B.2. (ii) is proved in Supp. S2.1. (iii) combines $c^\star > 0$ (appendix B.2) with Propositions S2.2, B.3 and B.4. (iv) is Proposition S2.3. □

Proof of theorem 7. Both parts are proved in appendix E.2. Beyond $2D$, with a certified level $U \le \frac{1}{2}$, lemma 4 also gives $\rho \ge \sqrt{1 - 2U}$ at every zero. □

C One step, the class-geometry certificates, and the Bayes model

C.1 Certificates for a class

Proof of lemma 4. For $p = \sum_{j \ne k} \lambda_j t_j$ with λ in the simplex, $\langle t_k, p \rangle \le c$, so $\|t_k - p\|^2 = 1 - 2\langle t_k, p \rangle + \|p\|^2 \ge 1 - 2c$. Let p^* be the nearest point and $v = (t_k - p^*)/\|t_k - p^*\|$. The projection property of convex sets gives $\langle v, t_k - t_j \rangle \ge \|t_k - p^*\|$ for every $j \ne k$. □

C.2 Override determinism

Proposition C.1 (Override determinism)**.** *Let $O_x = \{j : \mathrm{sign}(g_{x,j} + b_{\mathcal{R},j}) \ne \mathrm{sign}(g_{x,j})\}$ be the* override set. *Up to ties of measure zero, the step of proposition 1 is $\Delta\boldsymbol{\pi}(x) = -\eta\,\mathrm{sign}(g_x) \odot (\mathbf{1} - 2\,\mathbf{1}_{O_x}) - \eta\omega\boldsymbol{\pi}_0$. Two regularizers with the same override sets on every image therefore make identical predictions, whatever their geometry.*

Proof of Proposition C.1. By proposition 1 the step is $-\eta[\mathrm{sign}(g_x + b_{\mathcal{R}}) + \omega\boldsymbol{\pi}_0]$. Off the ties $g_{x,j} + b_{\mathcal{R},j} = 0$, $\mathrm{sign}(g_{x,j} + b_{\mathcal{R},j})$ equals $\mathrm{sign}(g_{x,j})$ for $j \notin O_x$ and $-\mathrm{sign}(g_{x,j})$ for $j \in O_x$, which is the stated form. It depends on $b_{\mathcal{R}}$ only through O_x, so equal override sets give equal steps, equal prompts and equal predictions. □

C.3 THE BAYES MODEL: WHICH GEOMETRY IS CALIBRATED

Remark 12 (No universal calibrated geometry). If $v \mid y{=}k \sim \mathrm{vMF}(m_k, \kappa)$ with equal priors, then $p(y{=}k \mid v) = \mathrm{softmax}_k(\kappa\langle v, m_k\rangle)$ (Banerjee et al., 2005), so CLIP's form with $t_k = m_k$ and $\alpha = \kappa$ is Bayes-calibrated. A classifier (T, α) reproduces this posterior at every v if and only if $\alpha t_k - \kappa m_k$ is the same vector for all k. No single configuration is therefore calibrated for every mixture.

Here $\mathrm{vMF}(m, \kappa)$ is the von Mises–Fisher distribution on the unit sphere, with density proportional to $e^{\kappa\langle v, m\rangle}$.

Proof of remark 12. With equal priors and a common concentration, the von Mises–Fisher normalizer $C_D(\kappa)$ cancels:

$$p(y{=}k \mid v) = \frac{C_D(\kappa)e^{\kappa\langle v, m_k\rangle}}{\sum_j C_D(\kappa)e^{\kappa\langle v, m_j\rangle}} = \mathrm{softmax}_k\left(\kappa\langle v, m_k\rangle\right)$$

(Banerjee et al., 2005). If $\alpha t_k - \kappa m_k = b$ for all k, the two sets of logits differ by the common term $\langle v, b\rangle$, which the softmax ignores. Conversely, if the two softmaxes agree at every unit v, their log-ratios agree: $\alpha\langle v, t_k - t_j\rangle = \kappa\langle v, m_k - m_j\rangle$ for all $v \in S^{D-1}$ and all k, j. A linear function that vanishes on the sphere is zero, so $\alpha(t_k - t_j) = \kappa(m_k - m_j)$, i.e. $\alpha t_k - \kappa m_k = \alpha t_j - \kappa m_j$. Two mixtures whose means are not related in this way therefore need different configurations. □

C.4 THE TWO RESIDUALS AT THE LEVEL

Proposition C.2 (Two residuals at the attainable level)**.** *With $x = s - c$, the two-sided residual $(s-c)^2$ (AGI-Fixed) has the force $2x$ and the ReLU residual $\left[\mathrm{ReLU}(s-c)\right]^2$ (AGI-TPT) the force $2\,\mathrm{ReLU}(x)$. Below the level the first is attractive ($2x < 0$) and the second is zero. If no pair lies below the level, the two forces, and hence the two gradients, coincide.*

The proof is in appendix E. At TPT's single step no CLIP pair lies below the level (fig. S5), so the two residuals give identical one-step predictions; they can differ only once pairs reach the level.

D ANALYSES OF PRIOR DISPERSION OBJECTIVES

Throughout, class features are unit vectors, $N \geq 2$, and "largest cosine" means $\max_{i \neq j} s_{ij}$. Each subsection states the objective as its paper writes it, with the equation and page, and then our analysis. Table 3 summarizes the results against the requirements (R1) and (R2) of section 3.

Table 3: **Where each spread regularizer stops, and which pairs it moves** (for free class vectors; proofs in appendices D and E; SoC is discussed in section 6). "Stops at": the regularizer's global minimizers (maximizers for C-TPT and A-TPT). $c^\star(N, D)$: the attainable level of section 4.1; for $N > 2D$, AGI-TPT uses a certified upper bound $U \geq c^\star$.

Method	(R1) Stops at	Merged classes there?	(R2) Pairs it moves
C-TPT	every T with $\sum_k t_k = 0$	yes, for $N \geq 4$	all, through the centroid only
O-TPT	unit-norm tight frames; largest cosine above $c^\star$ if $N \leq D$, and if $D < N \leq 2D$ unless $(N{-}D) \mid D$	yes, for $N \geq 2D$	all; pulls obtuse pairs back to 90°
A-TPT	maximizers of a mean of nearest-neighbour angles; no closed form; the nearest-pair force never switches off	not characterized	each class's nearest pair only; moving away from it brings closer every class at an angle above 90° from that direction (fails R2)
D-TPT	$t_k = \pm\mathbf{1}/\sqrt{D}$ for every class	yes, all classes	none (no pair term)
AGI-TPT	$\{\max_{i \neq j} s_{ij} \leq c^\star\}$: the most separated configurations	never; not even local minima ($N \leq 2D{-}1$)	only pairs above $c^\star$; no force on a pair below it

D.1 O-TPT

O-TPT adds $\lambda\|EE^\top - I_C\|_2^2$ to the TPT loss, with E the matrix of normalized class features (Sharifdeen et al., 2025, Eq. 2), and applies a Householder transform "on top of EE^T" in its

implementation; its Table 8 reports 64.34 / 4.45 (accuracy / ECE, ViT-B/16) without this transform and 64.41 / 4.23 with it. We analyze the written objective; parts (iii)–(v) use the Frobenius reading $\|S - I\|_F^2 = \sum_{i \neq j} s_{ij}^2$, a sum over pairs, under which the pair penalty is $\varphi(s) = s^2$ and the force is $2s$.

Proposition D.1 (O-TPT). *(i) The target $S = I$ is attainable if and only if $N \leq D$, for any matrix norm. (ii) Read as a level on the largest cosine, its level 0 is sub-optimal for $N \leq D + 1$, equal to $c^\star$ for $D + 2 \leq N \leq 2D$, and unattainable for $N \geq 2D + 1$. (iii) The force $2s$ is negative on obtuse pairs, which it pulls toward $90°$. For $N \leq D$ the zero set is the set of orthonormal configurations, whose largest cosine 0 exceeds the optimum $-\frac{1}{N-1}$. (iv) For $N \geq D$ the minimizers of $\|S - I\|_F^2$ over unit vectors are the unit-norm tight frames (Benedetto & Fickus, 2003). For $D < N \leq 2D$, a unit-norm tight frame with largest cosine 0 exists if and only if $(N - D)$ divides D. (v) A unit-norm tight frame of N vectors in $\mathbb{R}^D$ can contain two coincident vectors if and only if $N \geq 2D$. Hence for $N \geq 2D$ some minimizers of O-TPT's penalty contain two merged classes.*

Proof. (i) $\operatorname{rank}(S) \leq D$, while $\operatorname{rank}(I) = N$. (ii) is proposition 11. (iii) follows from $\varphi'(s) = 2s$ and proposition 11(i). For (iv), $\|S - I\|_F^2$ is the frame potential minus N, whose minimizers over unit vectors for $N \geq D$ are exactly the unit-norm tight frames (Benedetto & Fickus, 2003). *If:* with $d = D/(N - D)$, take $N - D$ mutually orthogonal copies of a regular simplex in $\mathbb{R}^d$; this gives $(N - D)(d + 1) = N$ vectors with cosines $-\frac{1}{d}$ within a copy and 0 across copies, and frame operator $\frac{d+1}{d} I = \frac{N}{D} I$. *Only if:* join two vectors when their cosine is negative. Vectors in different connected components are orthogonal, so the components span mutually orthogonal subspaces V_k of dimensions d_k with $\sum_k d_k = D$, because a tight frame spans $\mathbb{R}^D$. Within a component with n_k vectors, let $\sum a_i v_i = 0$ with $a \neq 0$. Because off-diagonal cosines are non-positive, $\| \sum |a_i| v_i \|^2 \leq \| \sum a_i v_i \|^2 = 0$, so $|a|$ is also a dependence. If $a_i = 0$ for some i, then $0 = \sum_j |a_j| \langle v_i, v_j \rangle$ is a sum of non-positive terms, so $a_j = 0$ for every neighbour of i, and by connectivity $a = 0$. Every nonzero dependence therefore has full support, so the space of dependences has dimension at most 1 and $n_k \leq d_k + 1$. Tightness makes the frame operator equal to $\frac{N}{D} I$ on each V_k; taking traces, $n_k = \frac{N}{D} d_k > d_k$. Hence $n_k = d_k + 1$ and $d_k = D/(N - D)$ for every k, which must be an integer. (v) *Only if:* a tight frame of N unit vectors has frame operator $\frac{N}{D} I$; if it contains a vector v twice, then $\frac{N}{D} = v^\top (\frac{N}{D} I) v \geq 2 (v^\top v)^2 = 2$, so $N \geq 2D$. *If:* let $N \geq 2D$ and $M = N - D \geq D$. By Theorem 7.1 of Benedetto & Fickus (2003) the frame potential attains its minimum over M unit vectors in $\mathbb{R}^D$ and every minimizer is a unit-norm tight frame, so such a frame $\{x_m\}$ exists; rotate it so that $x_1 = e_1$. Adding the orthonormal basis $\{e_1, \ldots, e_D\}$ gives N unit vectors with frame operator $\frac{M}{D} I + I = \frac{N}{D} I$, a unit-norm tight frame in which e_1 appears twice. By (iv) it minimizes O-TPT's penalty. □

For ImageNet with ViT-B/16 ($N = 1000$, $D = 512$, $N - D = 488$) and with ViT-L/14 ($D = 768$, $N - D = 232$), $N - D$ does not divide D, so every minimizer of O-TPT's penalty has a pair with positive cosine, although the attainable optimum is 0. O-TPT's loss does attain its minimum, at a well-spread geometry (a unit-norm tight frame); what is unattainable for $N > D$ is its per-pair target. These are long-run properties: at one step O-TPT's force $2s$ is positive on every CLIP pair, so its target is dormant; there AGI-TPT's penalty equals O-TPT's up to a dispersion term (proposition 9).

Proposition D.2 (O-TPT, summary). (a) *The target $S = I$ is attainable if and only if $N \leq D$.* (b) *For $N \geq D$ the penalty has no spurious local minima, and its minimizers are the unit-norm tight frames (Benedetto & Fickus, 2003).* (c) *For $D < N \leq 2D$, a minimizer has largest cosine 0 if and only if $(N - D) \mid D$; for ImageNet with $D = 512$ or 768 this fails, so every minimizer has a pair with positive cosine although 0 is attainable, and the most separated configuration is a critical point of the penalty but not a minimizer.* (d) *For $N \geq 2D$ some minimizers contain two coincident classes.*

Proof. (a) is Proposition D.1(i); (b) and the first part of (c) are Proposition D.1(iv) and Proposition S4.1(i); the rest of (c) is Proposition S4.1(iii); (d) is Proposition D.1(v). □

D.2 C-TPT

C-TPT maximizes $\mathrm{ATFD} = \frac{1}{N} \sum_i \|\bar{t} - t_i\|_2$, the average distance of the class features from their centroid (Yoon et al., 2024). Sharifdeen et al. (2025, p. 4) observe that it "mainly affects the centroid's

position without ensuring adequate spacing between the individual features." The next proposition makes this exact for C-TPT's published, unsquared objective.

Proposition D.3 (C-TPT). *For unit vectors,* $\mathrm{ATFD} \le \sqrt{1-\|\bar{t}\|^2} \le 1$*, and* $\mathrm{ATFD} = 1$ *if and only if* $\sum_i t_i = 0$*. For* $N \ge 4$ *and* $D \ge 2$*, this maximizer set contains configurations with two coincident classes; for* $N = 3$ *it is the planar regular simplex.*

Proof. $\frac{1}{N}\sum_i \|t_i - \bar{t}\|^2 = \frac{1}{N}\sum_i (1 - 2\langle t_i, \bar{t}\rangle + \|\bar{t}\|^2) = 1 - \|\bar{t}\|^2$. By the Cauchy–Schwarz (or Jensen) inequality, $\mathrm{ATFD} \le (\frac{1}{N}\sum_i \|t_i - \bar{t}\|^2)^{1/2} = \sqrt{1-\|\bar{t}\|^2} \le 1$. Equality in the last step needs $\bar{t} = 0$. Conversely, if $\bar{t} = 0$ then $\|t_i - \bar{t}\| = 1$ for every i, so $\mathrm{ATFD} = 1$. For $N \ge 4$, set $t_1 = t_2 = u$; the remaining $N - 2 \ge 2$ unit vectors must sum to $-2u$, which is possible since a sum of $k \ge 2$ unit vectors in $\mathbb{R}^D$, $D \ge 2$, can have any norm in $[0, k]$ and any direction. For $N = 3$, $t_1 + t_2 + t_3 = 0$ with unit vectors forces pairwise cosines $-\frac{1}{2}$. □

Its squared form $\frac{1}{N}\sum_i \|\bar{t} - t_i\|_2^2 = 1 - \|\bar{t}\|^2$ has the same maximizers; in pair form it is $\frac{1}{N^2}\sum_{i\ne j} s_{ij}$ up to sign and a constant, which pushes every pair with the same constant force; this push stops only where the centroid vanishes, at the maximizers above.

D.3 A-TPT

A-TPT minimizes $-\mathrm{AD}$ with $\mathrm{AD} = \frac{1}{N}\sum_i \min_{j\ne i} \theta_{ij}$ and $\theta_{ij} = \arccos s_{ij}$ (Ahamed et al., 2026). It motivates this objective by the Tammes problem of maximizing the minimum pairwise angle $\theta_{\min} = \min_{i\ne j} \theta_{ij}$ (Tammes, 1930).

Proposition D.4 (A-TPT). *(i)* $\mathrm{AD} \ge \theta_{\min}$*, with equality if and only if every class has the same nearest-neighbour angle. (ii) Where nearest neighbours are unique, the gradient of* AD *involves only the pairs* $(i, j^*(i))$ *of each class and its nearest neighbour, at most* N *of the* $\binom{N}{2}$ *pairs. (iii) Let* t_i *move on the sphere with unit speed along the geodesic away from* t_j*. Then* $\frac{d}{d\epsilon}\theta_{ij} = 1$ *and, for every other class* k*,* $\frac{d}{d\epsilon}\theta_{ik} = \cos \angle_{t_i}(j,k)$*, where* $\angle_{t_i}(j,k)$ *is the angle at* t_i *between the geodesics to* t_j *and* t_k*. Classes with* $\angle_{t_i}(j,k) > \pi/2$ *come closer, and on a common great circle through* t_j*,* t_i *and* t_k *(in this order) the loss is one-for-one:* $\theta_{ij} + \theta_{ik} = \theta_{jk}$*.*

Proof. (i) A mean is at least its minimum. (ii) Each term $\min_{j\ne i} \theta_{ij}$ depends only on the nearest pair. (iii) Write $v_k = (t_k - s_{ik}t_i)/\sin\theta_{ik}$ for the unit tangent at t_i toward t_k. For a unit tangent u, $\frac{d}{d\epsilon}\cos\theta_{ik} = \langle u, t_k\rangle = \sin\theta_{ik}\langle u, v_k\rangle$, so $\frac{d}{d\epsilon}\theta_{ik} = -\langle u, v_k\rangle$. With $u = -v_j$ this gives 1 for $k = j$ and $\langle v_j, v_k\rangle = \cos\angle_{t_i}(j,k)$ otherwise. □

Proposition D.5 (A-TPT's force). *Where nearest neighbours are unique and* $s_{ij} \in (-1, 1)$*, the objective* $-\mathrm{AD}$ *acts on the nearest pair* $(i, j^*(i))$ *of each class with the force* $\frac{1}{N}(1 - s^2)^{-1/2}$ *in its cosine, and on no other pair. This force is positive for every* $s \in (-1, 1)$*. Hence A-TPT has no configuration at which all its pair forces vanish.*

Proof. $-\mathrm{AD} = -\frac{1}{N}\sum_i \arccos s_{ij^*(i)}$, and $\frac{d}{ds}(-\arccos s) = (1 - s^2)^{-1/2} > 0$ on $(-1, 1)$. □

The objective is not a sum of fixed pair potentials, because the nearest neighbour changes; between such changes each class feels a positive force from one neighbour only. A-TPT's objective is bounded and continuous, so its maximizers exist; they have no closed form, and whether they solve the Tammes problem is open.

D.4 D-TPT

D-TPT adds $\lambda\frac{1}{N}\sum_k \mathrm{KL}\big(\sigma(t_k)\,\|\,\mathcal{U}\big)$, where $t_k \in \mathbb{R}^D$ is the normalized feature of class k, σ is the softmax over its D coordinates, and $\mathcal{U}$ is the uniform distribution on the D coordinates, with $\lambda = 10^5$ for all datasets and backbones (Han & Hwang, 2025, Eq. 3 and p. 7). The following analysis is ours.

Proposition D.6 (D-TPT). *Let* $x \in \mathbb{R}^D$ *with* $\|x\| = 1$ *and* $p = \sigma(x)$*. (i)* $\mathrm{KL}(p\,\|\,U) = 0$ *if and only if* $x = \pm\mathbf{1}/\sqrt{D}$*. Hence every global minimizer of the D-TPT regularizer places each class feature at* $\pm\mathbf{1}/\sqrt{D}$*, so all pairwise cosines equal* ± 1*. (ii)* $\nabla_x \mathrm{KL}(p\,\|\,U) = p \odot \big(x - \langle p, x\rangle\mathbf{1}\big)$*: a descent*

step moves each coordinate toward the p-weighted mean of the coordinates. (iii) To second order in the coordinates, $\mathrm{KL}(p \,\|\, U) = \frac{1}{2D}\big(1 - \langle x, \mathbf{1}/\sqrt{D}\rangle^2\big) + O(\|x\|_\infty^3)$. *(iv) The regularizer is a sum of per-class terms and has no term that depends on two classes.*

Proof. (i) $\sigma(x)$ is uniform if and only if all coordinates of x are equal; with $\|x\| = 1$ this means $x = \pm\mathbf{1}/\sqrt{D}$ (one global sign, since softmax acts on signed values). (ii) With $\log p_j = x_j - \log Z$, $\partial_{x_j} \sum_i p_i \log p_i = p_j\big(\log p_j - \sum_i p_i \log p_i\big) = p_j\big(x_j - \sum_i p_i x_i\big)$. (iii) For small coordinates, $p_i = \frac{1}{D}\big(1 + x_i - \bar{x}\big) + O(\|x\|_\infty^2)$ with $\bar{x} = \frac{1}{D}\sum_i x_i$, and $\mathrm{KL}(p \,\|\, U) = \frac{D}{2}\sum_i (p_i - \frac{1}{D})^2 + O(\|p - U\|^3) = \frac{1}{2D}\sum_i (x_i - \bar{x})^2 + O(\|x\|_\infty^3)$; finally $\sum_i (x_i - \bar{x})^2 = 1 - \langle x, \mathbf{1}/\sqrt{D}\rangle^2$. (iv) is immediate. □

D.5 The smallest counterexample: a uniform target is not always good

AGI-Fixed, the two-sided residual $(s - c^\star)^2$, has the force $2(s - c^\star)$, which is attractive below the level (Proposition C.2). Its resting set, all cosines equal to $c^\star$, is empty for $N \geq D + 2$ (Proposition S1.1). A magnitude variant $(s^2 - c^2)^2$ targets $|s| = c$ with the Welch value $c = W$; its target is attainable only if a real ETF exists (Supp. S2.3).

A three-vector illustration. Figure 2 (main text) shows the difference between one- and two-sided residuals (proposition 6) in its smallest instance, with $N = 3$ unit vectors A, B, C, for which $c^\star = -\frac{1}{2}$ (the planar 120° simplex). The sum of their angles is at most 2π, with equality if and only if $0 \in \mathrm{conv}\{A, B, C\}$ (Lemma S4.2; proof in the supplementary material).

In fig. 2a the angles are $150°, 150°, 60°$. Their sum is 2π, so by Lemma S4.2 any increase of θ_{BC} must decrease θ_{AB} or θ_{AC}. Here the attraction that AGI-Fixed applies to AB and AC points where the geometry must go anyway, and AGI-Fixed and AGI-TPT have the same unique zero, the planar simplex. In fig. 2b the angles are $150°, 150°, 20°$ and their sum is $320° < 2\pi$. The directions at angle at least 150° from A form a cone (Cone X) with B and C on its boundary. Moving C along this boundary away from B keeps $\theta_{AC} = 150°$ and increases θ_{BC}, up to 60° when C is opposite B on the boundary circle. Separating B from C therefore does not require contracting AB or AC. AGI-Fixed nevertheless applies the attractive force $2(\cos 150° + \frac{1}{2}) \approx -0.73$ to AB and AC. The ReLU of AGI-TPT applies no force to these two pairs, which lie below the level, and pushes only on BC. Both objectives still share the same zero set, because three vectors can all be at 120° only in the planar simplex; they differ in the forces along the way.

The counterexample generalizes (proposition 6(iii)): even where the two residuals share their unique zero, a two-sided target such as a uniform (ETF) Gram matrix attracts pairs that are already separated, and for $N \geq D + 2$ it has no zero at all (proposition 6(ii)).

E Proofs for the method, one step and sustained adaptation

Throughout, c is the level used by the loss: $c^\star(N, D)$, or its certified upper bound U when $N > 2D$. The pair penalty of AGI-TPT is $\varphi(s) = \big[\mathrm{ReLU}(s - c)\big]^2$, which is continuously differentiable with force $F(s) = 2\,\mathrm{ReLU}(s - c)$. We write $f(G) = \sum_{i \neq j} \big[\mathrm{ReLU}(G_{ij} - c)\big]^2$ for a symmetric $N \times N$ matrix G, so that $\mathcal{L}_{\mathrm{AGI}}(T) = f(TT^\top)/N^2$. The function f is convex (a sum of convex non-decreasing functions of single entries) and continuously differentiable, with $\nabla f(G)_{ij} = 2\,\mathrm{ReLU}(G_{ij} - c)$ for $i \neq j$ and 0 on the diagonal.

E.1 Entropy pulls competitors together (proposition 3)

Proof of proposition 3. Let $p = \mathrm{softmax}(z)$ and $H = -\sum_j p_j \log p_j$. Since $\partial p_j / \partial z_k = p_j(\delta_{jk} - p_k)$,

$$\frac{\partial H}{\partial z_k} = -\sum_j (\log p_j + 1)\, p_j(\delta_{jk} - p_k) = -p_k(\log p_k + 1) + p_k \sum_j p_j(\log p_j + 1) = -p_k(\log p_k + H).$$

With $z_k = \alpha\langle v, t_k\rangle$, the Euclidean gradient of H with respect to t_k is $-\alpha p_k(\log p_k + H)\, v$. Its projection on the tangent space of the unit sphere at t_k is $-\alpha p_k(\log p_k + H)(v - a_k t_k)$, so the

gradient flow is $\dot{t}_k = \varepsilon_k(v - a_k t_k)$ with $\varepsilon_k = \alpha p_k(\log p_k + H)$. (i) $\dot{a}_k = \langle \dot{t}_k, v\rangle = \varepsilon_k(1 - a_k^2)$, and $\varepsilon_k > 0$ if and only if $\log p_k > -H$, i.e. $p_k > e^{-H}$. Since $H \geq -\log \max_k p_k$, the most probable class has $p_k \geq e^{-H}$, with equality only if p is uniform (a softmax has full support). Moreover $\sum_k \varepsilon_k = \alpha(-H + H) = 0$. (ii)–(iii) $\dot{s}_{ij} = \langle \dot{t}_i, t_j\rangle + \langle t_i, \dot{t}_j\rangle = \varepsilon_i(a_j - a_i s_{ij}) + \varepsilon_j(a_i - a_j s_{ij})$. Let $a_i \geq a_j > 0$ and $s_{ij} < a_j/a_i \leq 1$. Then $a_j - a_i s_{ij} > 0$, and $a_i - a_j s_{ij} > 0$ (if $a_i > a_j$ because $s_{ij} \leq 1$; if $a_i = a_j$ because $s_{ij} < 1$). Both brackets are positive, so $\dot{s}_{ij}$ has the common sign of ε_i and ε_j whenever they agree: positive for two competitive classes, negative for two non-competitive ones. □

Averaged views. TPT minimizes the entropy of the prediction $\bar{p} = \frac{1}{M}\sum_m p^{(m)}$ averaged over M views v_m. The same computation gives $\partial H(\bar{p})/\partial z_k^{(m)} = -\frac{1}{M} p_k^{(m)}\big(\log \bar{p}_k - \sum_l p_l^{(m)} \log \bar{p}_l\big)$, so each class moves toward the views on which it is more probable than the $p^{(m)}$-weighted typical class and away from the others; with one view this is proposition 3. In prompt space all class features move through the shared Jacobian of the text encoder, which is why fig. 3a tests the prompt case separately.

E.2 Zeros and landscape (theorem 7)

Proof of theorem 7(i). Off the diagonal, $\mathrm{ReLU}(s - c)$ vanishes if and only if $s \leq c$, so the zeros of $\mathcal{L}_{\text{AGI}}$ are $Z(c)$. For $N \leq D+1$ and $c = -\frac{1}{N-1}$, $Z(c)$ is the set of regular simplices (appendix B.2), the Softmax Codes for $N \leq D + 1$ (Jiang et al., 2024, Thm. 3.3). Let $D + 2 \leq N \leq 2D$ and $c = 0$, and let $\rho(T) = \min_k \mathrm{dist}(t_k, C_k)$ with $C_k = \mathrm{conv}\{t_j\}_{j\neq k}$. $Z(0) \subseteq \{\rho = 1\}$. If $\max s \leq 0$, lemma 4 with $c = 0$ gives $\mathrm{dist}(t_k, C_k) \geq 1$ for every k, so $\rho \geq 1$. Conversely $\rho \leq 1$ for every configuration with $N \geq D + 2$ (Jiang et al., 2024, Lemma C.9); for completeness: the N points are affinely dependent, $\sum_i a_i t_i = 0$ and $\sum_i a_i = 0$ with $a \neq 0$. Let $I_{\text{pos}} = \{i : a_i > 0\}$, $I_{\text{neg}} = \{j : a_j < 0\}$ and $\Lambda = \sum_{i\in I_{\text{pos}}} a_i > 0$. The Radon point $w = \sum_{i\in I_{\text{pos}}}(a_i/\Lambda)t_i = \sum_{j\in I_{\text{neg}}}(|a_j|/\Lambda)t_j$ lies in both hulls. Since $\|w\|^2 = \sum_{i\in I_{\text{pos}}}(a_i/\Lambda)\langle t_i, w\rangle$, some $i \in I_{\text{pos}}$ has $\langle t_i, w\rangle \geq \|w\|^2$, and $w \in C_i$ because w is a convex combination of $\{t_j\}_{j\in I_{\text{neg}}}$. Hence

$$\mathrm{dist}(t_i, C_i)^2 \leq \|t_i - w\|^2 = 1 - 2\langle t_i, w\rangle + \|w\|^2 \leq 1 - \|w\|^2 \leq 1.$$

$\{\rho = 1\} \subseteq Z(0)$. If $\rho = 1$, the display gives $1 \leq 1 - \|w\|^2$, so $w = 0$. Then $0 \in \mathrm{conv}\{t_i\}_{i\in I_{\text{pos}}}$ and $0 \in \mathrm{conv}\{t_j\}_{j\in I_{\text{neg}}}$. Every k lies outside one of these disjoint index sets, so $0 \in C_k$ for every k. As $\mathrm{dist}(t_k, C_k) \geq \rho = 1 = \|t_k - 0\|$, the point 0 is a nearest point of C_k to t_k, and the projection property of convex sets gives $\langle t_k - 0,\, t_j - 0\rangle \leq 0$ for every $t_j \in C_k$, i.e. $s_{kj} \leq 0$ for all $j \neq k$. So $T \in Z(0)$. Since $Z(0) \neq \emptyset$ (lemma 5) and $\rho \leq 1$ everywhere, $\{\rho = 1\}$ is the set of maximizers of ρ, which are the Softmax Codes (Jiang et al., 2024, Def. 2.1). □

Proof of theorem 7(ii). Let T be a local minimizer of $F(T) = f(TT^\top)$ on $(S^{D-1})^N$ with $\mathrm{rank}\, T < D$, and let $G = TT^\top$ and $\Phi = \nabla f(G)$. *First order.* The Euclidean gradient of F is $2\Phi T$. At a critical point on the product of spheres, each row $(\Phi T)_i$ is parallel to t_i: $(\Phi T)_i = \mu_i t_i$ with $\mu_i = (\Phi G)_{ii}$. Hence the symmetric matrix $\Lambda = \Phi - \mathrm{Diag}(\mu)$ satisfies $\Lambda T = 0$, so $\Lambda G = 0$ and $\langle \Lambda, G\rangle = 0$. *Second order.* Since $\mathrm{rank}\, T < D$, there is a unit $u \in \mathbb{R}^D$ with $Tu = 0$, i.e. $u \perp t_i$ for all i. For any $z \in \mathbb{R}^N$, let T_ϵ have rows $(t_i + \epsilon z_i u)/\sqrt{1 + \epsilon^2 z_i^2}$, a curve in $(S^{D-1})^N$ with $T_0 = T$. Its Gram matrix is

$$G_\epsilon = G + \epsilon^2\Big(zz^\top - \tfrac{1}{2}\big(\mathrm{Diag}(z^{\odot 2})G + G\,\mathrm{Diag}(z^{\odot 2})\big)\Big) + O(\epsilon^4) =: G + \epsilon^2 H_z + O(\epsilon^4),$$

because the cross terms vanish ($Tu = 0$) and the normalization contributes $-\frac{1}{2}\epsilon^2(z_i^2 + z_j^2)G_{ij}$. As f is continuously differentiable, $F(T_\epsilon) = F(T) + \epsilon^2\langle \Phi, H_z\rangle + o(\epsilon^2)$, and local minimality forces $\langle \Phi, H_z\rangle \geq 0$. Now $\langle \Phi, H_z\rangle = z^\top \Phi z - \sum_i z_i^2 (\Phi G)_{ii} = z^\top \Lambda z$. So $\Lambda \succeq 0$. *Global optimality over the elliptope.* Let G' be any positive semidefinite matrix with unit diagonal. By convexity and differentiability of f,

$$f(G') \geq f(G) + \langle \Phi, G' - G\rangle = f(G) + \langle \Lambda, G'\rangle - \langle \Lambda, G\rangle + \textstyle\sum_i \mu_i(G'_{ii} - G_{ii}) = f(G) + \langle \Lambda, G'\rangle \geq f(G).$$

Take for G' the Gram matrix of the regular simplex, $\frac{N}{N-1}I - \frac{1}{N-1}\mathbf{1}\mathbf{1}^\top$, whose off-diagonal entries equal $-\frac{1}{N-1} \leq c$. Then $f(G') = 0$, so $f(G) = 0$ and T is a zero. If $N < D$, every T has rank at most $N < D$, so every local minimizer is a zero. □

The argument uses only that f is convex and continuously differentiable in G, which is why the loss is a *squared* ReLU of the cosine; an angle-based penalty is not convex in G.

E.3 MERGED CLASSES (THEOREM 8)

Proposition 13 (Splitting a merged pair)**.** *Let $D \geq 2$, $c < 1$, and let $t_1 = t_2 = u$. If $(N-2)\big(\frac{1}{D-1} + \frac{c^2}{4}\big) < 2(1-c)$, then T is not a local minimizer of $\mathcal{L}_{\text{AGI}}$.*

Proof. For a unit $w \perp u$, let $t_1(\epsilon) = \cos\epsilon\, u + \sin\epsilon\, w$ and $t_2(\epsilon) = \cos\epsilon\, u - \sin\epsilon\, w$, keeping the other classes fixed, and let $\phi(\epsilon) = f(T(\epsilon)T(\epsilon)^\top)$. Swapping w and $-w$ swaps the two classes, so ϕ is even. We expand $\phi(\epsilon) - \phi(0)$ to second order. *The pair* $(1,2)$. $s_{12}(\epsilon) = \cos 2\epsilon = 1 - 2\epsilon^2 + O(\epsilon^4)$, so the two ordered terms change by $2\big[(1-c-2\epsilon^2)^2 - (1-c)^2\big] + O(\epsilon^4) = -8(1-c)\epsilon^2 + O(\epsilon^4)$. *A pair* $(1,j)$, $(2,j)$ *with* $j \geq 3$. Let $a = \langle u, t_j\rangle$ and $b = \langle w, t_j\rangle$. Then $s_{1j}, s_{2j} = a\cos\epsilon \pm b\sin\epsilon = a \pm b\epsilon - \frac{a}{2}\epsilon^2 + O(\epsilon^3)$. If $a < c$, both stay below c for small ϵ and contribute 0. If $a > c$, the four ordered terms change by $4\big(b^2 - a(a-c)\big)\epsilon^2 + O(\epsilon^3)$. If $a = c$, they change by at most $2b^2\epsilon^2 + o(\epsilon^2) \leq 4\big(b^2 - a(a-c)\big)\epsilon^2 + o(\epsilon^2)$. Pairs not involving classes $1, 2$ do not change. Hence $\phi(\epsilon) - \phi(0) \leq \kappa(w)\epsilon^2 + o(\epsilon^2)$ with $\kappa(w) = -8(1-c) + 4\sum_{j\geq 3:\, a_j \geq c}\big(b_j^2 - a_j(a_j - c)\big)$. Draw w uniformly from the unit sphere of $u^\perp$, which has dimension $D-1 \geq 1$. Then $\mathbb{E}\, b_j^2 = (1-a_j^2)/(D-1) \leq 1/(D-1)$, and $-a(a-c) \leq c^2/4$ for every real a. So $\mathbb{E}\,\kappa(w) \leq -8(1-c) + 4(N-2)\big(\frac{1}{D-1} + \frac{c^2}{4}\big) < 0$, and some w has $\kappa(w) < 0$. Along that curve $\phi(\epsilon) < \phi(0)$ for all small $\epsilon \neq 0$, so T is not a local minimizer. □

Proof of theorem 8. Apply proposition 13 with $c = c^\star$ (for $N = 2$ the sum over j is empty and the condition reads $0 < 2(1-c)$). *Regime I*, $N \leq D+1$, $c = -\frac{1}{N-1}$: $\frac{N-2}{D-1} \leq 1$ and $\frac{(N-2)c^2}{4} = \frac{N-2}{4(N-1)^2} < \frac{1}{4}$, so the left side is below $\frac{5}{4} < 2 < 2(1-c)$. *Regime II*, $D+2 \leq N \leq 2D-1$, $c = 0$: the left side is $\frac{N-2}{D-1} \leq \frac{2D-3}{D-1} < 2 = 2(1-c)$. □

For $N = 2D$ the condition holds with equality at $c = 0$, and beyond $2D$ it fails, so these cases are not covered.

E.4 THE TWO RESIDUALS, ONE STEP, AND THE RELU

Proof of Proposition C.2. $\frac{d}{ds}(s-c)^2 = 2(s-c)$. For $s > c$, $\frac{d}{ds}\big[\,\mathrm{ReLU}(s-c)\big]^2 = 2(s-c)$; for $s < c$ it is 0; at $s = c$ both one-sided derivatives are 0, so the derivative exists and equals $2\,\mathrm{ReLU}(s-c)$. Below the level $2(s-c) < 0$ while $2\,\mathrm{ReLU}(s-c) = 0$. If every pair has $s_{ij} \geq c$, the two forces agree on every pair, and since both penalties are sums of pair terms, their gradients with respect to the class features (and to the prompt) agree. □

Proof of proposition 1. AdamW first applies the decoupled decay $\boldsymbol{\pi} \leftarrow \boldsymbol{\pi} - \eta\omega\boldsymbol{\pi}$ and then the Adam step. At the first step $\hat{m} = g$ and $\hat{v} = g^{\odot 2}$, so the Adam step is $-\eta\, g/(|g| + \epsilon) \to -\eta\,\mathrm{sign}(g)$ coordinatewise as $\epsilon \to 0$. The gradient of $\mathcal{L}_{\text{TPT}} + \lambda\mathcal{R}$ at $\boldsymbol{\pi}_0$ is $g_x + b_{\mathcal{R}}$, and $b_{\mathcal{R}} = \lambda\nabla_{\boldsymbol{\pi}}\mathcal{R}(T(\boldsymbol{\pi}_0))$ does not involve x because $\boldsymbol{\pi}_0$ is the same for every image and $\mathcal{R}$ depends only on the class features. If $b_{\mathcal{R}} = \beta\, b_{\mathcal{R}'}$ with $\beta > 0$, multiplying the weight of $\mathcal{R}'$ by β makes the two offsets, hence the two steps and the two predictions, identical for every image. □

Proof of proposition 9. $\sum_{i\neq j}(s_{ij} - c)^2 = \sum_{i\neq j} s_{ij}^2 - 2c\sum_{i\neq j} s_{ij} + N(N-1)c^2$, and $\sum_{i\neq j} s_{ij} = N^2\|\bar{t}\|^2 - N$. Dividing by N^2,

$$\frac{1}{N^2}\sum_{i\neq j}(s_{ij} - c)^2 = \frac{1}{N^2}\sum_{i\neq j} s_{ij}^2 - 2c\|\bar{t}\|^2 + \frac{2c}{N} + \frac{(N-1)c^2}{N}.$$

If every pair satisfies $s_{ij} > c$, then $\mathcal{L}_{\text{AGI}}$ equals the left side in a neighbourhood of the current configuration (Proposition C.2), so the gradients agree. The last two terms are constant. $\sum_{i\neq j} s_{ij}^2 = \|S - I\|_F^2$ for unit vectors, which is O-TPT's penalty in the Frobenius reading, and for $c \leq 0$ the

term $-2c\|\bar{t}\|^2 = 2|c|\|\bar{t}\|^2$ is a positive multiple of the squared C-TPT dispersion term. For $c = 0$ the second term vanishes. □

Equal strength gives identical predictions on ImageNet. With $c = c^\star = 0$, the one-step offsets of AGI-TPT and of O-TPT's Frobenius penalty are parallel. Along the ray of weights the override mass is non-decreasing, so matching it matches the weight up to the ties that leave override sets unchanged, and Proposition C.1 gives identical steps and predictions.

Proof of corollary 2. To first order, the TPT loss changes by $\langle g_x, \Delta\boldsymbol{\pi}\rangle = -\eta \sum_j g_{x,j} \operatorname{sign}(g_{x,j} + b_{\mathcal{R},j}) - \eta\omega\langle g_x, \boldsymbol{\pi}_0\rangle$. For $j \notin O_x$ the sign equals $\operatorname{sign}(g_{x,j})$ and the term is $-|g_{x,j}|$; for $j \in O_x$ it is the opposite sign and the term is $+|g_{x,j}|$ (coordinates where $g_{x,j} + b_{\mathcal{R},j} = 0$ do not move and contribute 0). Summing gives $-\eta(\|g_x\|_1 - 2\|g_x|_{O_x}\|_1) = -\eta\|g_x\|_1(1 - 2\mu_x)$. A coordinate with $g_{x,j} \neq 0$ is in O_x only if $b_{\mathcal{R},j}$ has the opposite sign and $|b_{\mathcal{R},j}| > |g_{x,j}|$. □

The step of Proposition C.1 follows coordinate by coordinate from the same case analysis.

The level identity. With bins B_b of sizes n_b over n predictions, $\mathrm{ECE} = \sum_b \frac{n_b}{n}|\mathrm{acc}(B_b) - \mathrm{conf}(B_b)| \geq \left|\sum_b \frac{n_b}{n}(\mathrm{conf}(B_b) - \mathrm{acc}(B_b))\right| = |\mathrm{gap}|$ by the triangle inequality, with equality when the miscalibration has one sign in every bin; and $\mathrm{gap} - \mathrm{gap}_{\mathrm{ZS}} = \Delta\mathrm{conf} - \Delta\mathrm{acc}$ by definition.

Table S2 (supplementary material) lists each guarantee with its status and where it is tested.

E.5 Proof of proposition 6

Proof. (i) The ReLU residual $\left[\mathrm{ReLU}(s - c)\right]^2$ has force $2\,\mathrm{ReLU}(s - c)$, which vanishes exactly for $s \leq c$. So all pair forces vanish exactly on $Z(c)$, which is non-empty if and only if $c \geq c^\star$ by lemma 5. (ii) The two-sided force $2(s - c)$ vanishes only at $s = c$, so all pair forces vanish exactly when all cosines equal c. By Proposition S1.1, N unit vectors with all cosines equal to $c < 1$ exist only if $N \leq D + 1$. This concerns the pair forces, not the existence of minimizers: the two-sided loss is continuous on a compact set and attains its minimum (at $c = 0$ its minimizers are the unit-norm tight frames, Proposition D.1(iv)), but at every minimizer some pair force is non-zero, and the forces on separated pairs point inward. (iii) Take $D = 3$ and $c = c^\star(3, 3) = -\frac{1}{2}$; the unique zero of both residuals is the planar regular simplex. Let $A = (0, 0, 1)$ and, for $\psi \in [0, \pi]$, $C(\psi) = (\frac{1}{2}\cos\psi, \frac{1}{2}\sin\psi, -\frac{\sqrt{3}}{2})$ and $B = C(0)$. Then $\langle A, B\rangle = \langle A, C(\psi)\rangle = -\frac{\sqrt{3}}{2}$, i.e. both angles are $150°$ for every ψ, and $\langle B, C(\psi)\rangle = \frac{3+\cos\psi}{4}$. At $\psi_0 = \arccos(4\cos 20° - 3) \approx 40.6°$ the three angles are $(150°, 150°, 20°)$. There the two-sided forces are $2(\cos 150° + \frac{1}{2}) = 1 - \sqrt{3} \approx -0.732$ on AB and AC (attractive) and $2(\cos 20° + \frac{1}{2}) \approx 2.879$ on BC. The attractive part is $2(\sqrt{3} - 1)/(2(\sqrt{3} - 1) + 2.879) \approx 0.34$ of the total absolute force. The ReLU forces are 0, 0 and 2.879. As ψ increases from ψ_0 to π, $s_{BC} = \frac{3+\cos\psi}{4}$ decreases strictly from $\cos 20°$ to $\frac{1}{2}$ while s_{AB}, s_{AC} stay fixed. Since $s_{BC} > c$ throughout, both $(s_{BC} - c)^2$ and $\left[\mathrm{ReLU}(s_{BC} - c)\right]^2$ decrease strictly, so both penalties decrease along the curve, and the angle θ_{BC} rises from $20°$ to $60°$. □

What (iii) does not say. At $c = -\frac{1}{2}$ the only zero is the planar simplex, whose angles are all $120°$, so AB and AC must eventually close from $150°$ to $120°$. At $\psi = \pi$ the angles are $(150°, 150°, 60°)$ and sum to 2π, and Lemma S4.2 then forces any further widening of BC to close AB or AC. The difference between the residuals is therefore when the satisfied pairs are pulled. The two-sided residual pulls them from the start. The ReLU residual applies no force to them, so they move only as a side effect of separating BC. For $N \geq D + 2$ the difference is larger, since the two-sided residual has no zero at all (ii).

The push on a pair. The Riemannian gradient of $\left[\mathrm{ReLU}(s_{ij} - c)\right]^2$ with respect to t_i is $2\,\mathrm{ReLU}(s_{ij} - c)(t_j - s_{ij}t_i)$, whose length is $2\,\mathrm{ReLU}(s_{ij} - c)\sqrt{1 - s_{ij}^2} = 2\,\mathrm{ReLU}(s_{ij} - c)\sin\theta_{ij}$. The force in the cosine is proportional to the excess, but the geometric push also scales with $\sin\theta_{ij}$ and vanishes as the pair merges; theorem 8 shows that merged configurations are nevertheless never local minima.

F METHOD DETAILS

F.1 OVERVIEW AND ALGORITHM

Figure S1 (supplementary) shows where the regularizer enters test-time prompt tuning, and Algorithm 1 gives the procedure for one test image. The only changes to TPT are the Gram matrix of the class features and the added loss term.

Algorithm 1 AGI-TPT for one test image

Require: image x; prompt $\boldsymbol{\pi}$ initialized to "a photo of a"; class names; weight λ; learning rate η; augmentations $\mathcal{A}(x)$ (64 views)

$c \leftarrow c^\star(N, D)$, i.e. $-\frac{1}{N-1}$ if $N \le D+1$ and 0 if $N \le 2D$; the certified upper bound $U(N, D)$ if $N > 2D$ ▷ appendix B.4

$T \leftarrow \textsc{Normalize}\big(f_T(\boldsymbol{\pi}, \text{class names})\big) \in \mathbb{R}^{N \times D}$

$S \leftarrow TT^\top$

$G_c \leftarrow (1-c)I + c\,\mathbf{1}\mathbf{1}^\top$ ▷ ones on the diagonal, c elsewhere

$\mathcal{L}_{\text{AGI}} \leftarrow \frac{1}{N^2}\|\operatorname{ReLU}(S - G_c)\|_F^2$ ▷ ReLU entrywise; diagonal entries are 0

$\mathcal{L}_{\text{TPT}} \leftarrow$ entropy of the prediction averaged over the 10% most confident views of $\mathcal{A}(x)$

$\boldsymbol{\pi} \leftarrow \textsc{AdamW}\big(\boldsymbol{\pi},\ \nabla_{\boldsymbol{\pi}}(\mathcal{L}_{\text{TPT}} + \lambda\mathcal{L}_{\text{AGI}}),\ \eta\big)$ ▷ one step

return $\arg\max_k\ p(k \mid x; \boldsymbol{\pi})$

The diagonal of S and of G_c both equal 1, so the diagonal of $S - G_c$ is 0 and only the pairs $i \neq j$ contribute. This is the loss of eq. (1).

F.2 COMPLEXITY

The forward pass of the regularizer is one Gram product, $2N^2D$ floating-point operations (≈ 1.0 GFLOP at $N = 1000$, $D = 512$), plus $O(N^2)$ for the residual. The backward pass adds one product of the same size, $(R + R^\top)T$ with R the residual. Memory is dominated by the $N \times N$ Gram matrix (4 MB in single precision at $N = 1000$). Inference is unchanged, because the adapted prompt is used exactly as in TPT. For N in the tens of thousands the Gram product can be computed in blocks, or the loss estimated on sampled pairs, since it is a sum over pairs.

F.3 LEVELS USED IN THE EXPERIMENTS

Table 4 lists the level for every dataset and backbone. Every benchmark satisfies $N \le 2D$, so every level is exact. At one step the level is dormant: the smallest zero-shot cosine of every dataset exceeds $c^\star$ (fig. S5).

F.4 HELD-OUT SELECTION OF THE WEIGHT λ

We select λ on CIFAR-100 (Krizhevsky, 2009) ($N = 100$), a dataset that is not a benchmark, using the ECE of one TPT step on its test split and of an online pass over it, and freeze it for every benchmark, backbone, shift and setting. Both criteria select $\lambda = 4$ (table 5). The 11-benchmark rows were computed after the selection and are printed for completeness. Choosing λ per dataset with the test labels, an oracle unavailable in practice, would lower the average one-step ECE from 1.81 to 1.66 and the online one from 1.38 to 1.31. The same held-out selection over each baseline's weight grid ($\frac{1}{4}$ to 4 times its published weight) moves every baseline toward equal strength and improves it; AGI-TPT stays first.

F.5 HYPERPARAMETERS AND PROTOCOL

TPT. One AdamW step per image (learning rate 5×10^{-3}, weight decay $\omega = 0.01$, the PyTorch default used by TPT), 64 augmented views (random resized crops, with AugMix for the fine-grained datasets as in TPT), the 10% most confident views in the entropy, and the prompt initialized to "a photo of a" (4 context tokens of width 512, or 768 for ViT-L/14, so $d_\pi = 2048$ or 3072). ECE uses 15 equal-width bins for every number we compute (ours, our controls and every cell marked ‡).

Table 4: **Levels $c^\star(N, D)$ used in the experiments.** RN50 has $D = 1024$, ViT-B/16 $D = 512$ and ViT-L/14 $D = 768$. All datasets with $N \leq 397$ are in the simplex regime on every backbone. Caltech101 has 100 classes in the CoOp split used by TPT. The ImageNet-21k row gives the certified levels of appendix B.4, evaluated on class subsets in appendix G.8.

Dataset	N	RN50	ViT-B/16	ViT-L/14
FER-2013 (level-value test only)	7	-0.1667	-0.1667	-0.1667
EuroSAT; CIFAR-10, STL-10 (level-value test)	10	-0.1111	-0.1111	-0.1111
Oxford Pets	37	-0.0278	-0.0278	-0.0278
DTD	47	-0.0217	-0.0217	-0.0217
FGVC-Aircraft, Caltech101	100	-0.0101	-0.0101	-0.0101
Food101, UCF101	101	-0.0100	-0.0100	-0.0100
Flowers102	102	-0.0099	-0.0099	-0.0099
Stanford Cars	196	-0.0051	-0.0051	-0.0051
ImageNet-A, ImageNet-R	200	-0.0050	-0.0050	-0.0050
SUN397	397	-0.0025	-0.0025	-0.0025
ImageNet, -V2, -Sketch	1000	-0.0010	0	0
ImageNet-21k (certified U)	21,841	0.03125	0.0625	0.0625

Table 5: **Selecting λ on CIFAR-100** (ViT-B/16). Top: the sweep (ECE); the 11-benchmark rows were computed after the selection and not used for it. Bottom: every method with its weight selected on CIFAR-100 (multiplier of the published weight), one step, 11-dataset averages.

λ	0.25	0.5	1	2	**4**	8	16	32
CIFAR-100, one step	9.8	8.1	5.6	3.0	**1.6**	2.4	6.1	11.9
CIFAR-100, online	21.7	11.4	4.2	1.9	**1.3**	1.5	3.4	8.8
11 benchmarks, one step	7.9	6.6	4.6	2.6	**1.81**	2.5	5.9	11.2
11 benchmarks, online	19.4	10.2	3.9	1.77	**1.38**	1.52	3.3	8.1

Selected on CIFAR-100	C-TPT‡	O-TPT‡	D-TPT‡	A-TPT‡	**AGI-TPT**
Weight multiplier	$\times 2$	$\times 2$	$\times 2$	$\times 1$	$\lambda = 4$
11 benchmarks, Acc.	64.68	64.95	64.55	64.85	**65.20**
11 benchmarks, ECE	2.47	1.86	2.41	2.68	**1.81**

Published baseline cells are copied as printed and were computed with each paper's own evaluation code. C-TPT's released code computes ECE with 20 equal-width bins (`ECE_Loss(20, ...)` in its `tpt_classification.py`; read, not run), and the other baselines state that they follow C-TPT's evaluation settings, which we did not verify in their code. Equal-width ECE is not invariant to the bin count, so differences of a few tenths between a published cell and one of ours are not resolved; recomputed with 20 bins, our 11-dataset averages are 1.88 (one step) and 1.44 (online) instead of 1.81 and 1.38. Table 7, where every row is computed by us with 15 bins, is the like-for-like one-step comparison. For SigLIP, class probabilities are the softmax of the image–text logits including SigLIP's learned scale (its bias cancels in the softmax), and TPT's entropy uses the same probabilities.

Settings. *Episodic*: the TPT protocol, one step per image, reset between images, 5 seeds (the seeds vary the augmented views). *Episodic, k steps*: k AdamW steps per image, reset between images. *Online*: one step per image without reset, so the prompt accumulates updates along the test stream; 10 stream orders, drawn as seeded random permutations (seeds 0–9) and shared by all methods. Online, AdamW's decoupled weight decay is switched off ($\omega = 0$), because over 5×10^4 steps it alone would shrink $\boldsymbol{\pi}_0$ by a factor of about 0.08; the learning rate and λ are as in the episodic protocol. ECE is computed over all predictions of the stream. ECE_{end} is the ECE of the prompt frozen at the end of the stream, evaluated on the whole test set, so that it has the same noise floor as the cumulative ECE. The six datasets of table 9 (ImageNet, SUN397, Cars, Aircraft, DTD, EuroSAT) were fixed before the runs to span N from 10 to 1000 and TPT's weakest to strongest over-confidence; tables 1 and S13 cover the others.

Weights. AGI-TPT uses $\lambda = 4$, selected on CIFAR-100 (Krizhevsky, 2009), which is not among the benchmarks, both for one step and online, and frozen for every dataset, backbone, shift and setting

(appendix F.4). The published rows use each baseline's published weight: C-TPT 50, O-TPT 18, D-TPT 10^5 and A-TPT 80; on the ImageNet shifts C-TPT 20, O-TPT 2 and A-TPT 10. For our online runs of every baseline, the weight is selected on the same held-out online CIFAR-100 stream from $\{\frac{1}{4}, \frac{1}{2}, 1, 2, 4\}$ times the published weight, with the same criterion as ours.

Equal strength at one step. We match the quantity that corollary 2 identifies, the override mass $\mu = \mathbb{E}_x \|g_x|_{O_x}\|_1 / \|g_x\|_1$. On each dataset we measure it, without labels, on 500 images of the training split (the CoOp splits), so the test sets stay untouched, and rescale each method's weight until μ equals AGI-TPT's (0.58 on ImageNet). At their published weights, (ν, μ) on ImageNet, with $\nu = \mathbb{E}_x |O_x| / d_\pi$ the override fraction, are: C-TPT (0.58, 0.46); O-TPT (0.62, 0.52); D-TPT (0.66, 0.53); A-TPT (0.73, 0.61); AGI-TPT (0.71, 0.58).

Controls. *Random flip set*: for each image, a random order of the coordinates of g_x; TPT's step is reversed on the first coordinates of that order whose share of $\|g_x\|_1$ reaches AGI-TPT's override mass on that dataset; no regularizer; 20 draws. A fixed random-sign offset is not a valid control at this mass: with independent random signs each coordinate is overridden with probability at most $\frac{1}{2}$, so its expected override mass is at most $\frac{1}{2}$. The other one-step nulls are defined in Supp. S5.3. *Annealed* controls rescale AGI-NoLevel's weight after every image so that its offset norm, or its summed pair force, equals AGI-TPT's at the same point of the stream, or decay each pair's weight as $1/(1 + n_{ij}/1000)$ with n_{ij} the number of steps pair (i, j) has been pushed. *Push-mass-matched level* 0 rescales level 0's weight after every image so that its cumulative pair force equals AGI-TPT's. *Periodic reset* returns the prompt to $\boldsymbol{\pi}_0$ every 1000 images (Press et al., 2023). The *EMA teacher* predicts with an exponential moving average of the prompt (momentum 0.999) (Wang et al., 2022). *SAR* (Niu et al., 2023) keeps images whose entropy is below $0.4 \ln N$, takes a sharpness-aware step (radius 0.05) and resets when the moving-average entropy falls below 0.2.

ETF target. The neural-collapse control penalizes $\frac{1}{N^2} \|\bar{S} - \frac{N}{N-1}(I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top)\|_F^2$, where $\bar{S}$ is the Gram matrix of the class features after removing their mean and renormalizing; its zero is the simplex ETF of the centred features, the NC2 geometry (Papyan et al., 2020; Kothapalli, 2023). Its weight is selected like every other method's.

Split tests. Within one online run, the classes are split at random into two halves (2 partitions $\times$ 10 orders). The loss is changed only for pairs with both classes in the first half: the two-sided residual (R2, ImageNet) or level 0 (R1, EuroSAT). Pairs across halves keep AGI-TPT's loss. We report the class-conditional ECE of each half, the mean over its classes of the ECE of the predictions for that class.

Structure measures. Spearman's rank correlation r_s between the $\binom{N}{2}$ pairwise cosines of the class features at the end of the stream and those of zero-shot CLIP; Jaccard@k between each class's k nearest classes before and after ($k = 10$, and $k = 3$ on EuroSAT), averaged over classes; the order map of fig. S7b uses deciles of zero-shot cosine.

Diagnostics. *Force below* $c^\star$: the share of the summed absolute pair force on pairs with $s_{ij} \le c^\star$. *Reached*: a pair whose cosine is at or below the level at some step. *Absorbed classes*: ImageNet classes present in the last 5k images of the stream but predicted at less than 1% of their frequency.

Statistics. Intervals are 95% paired bootstrap intervals over test images and seeds (1000 resamples; all methods share the resamples), and over stream orders for online runs. The *noise floor* of a dataset is the expected 15-bin ECE of a perfectly calibrated predictor with the method's own confidences on the same test set, computed by resampling labels from the confidences. The debiased ECE follows Kumar et al. (2019); we also report the smooth ECE (Błasiok & Nakkiran, 2024). A per-dataset difference is *resolved* when its interval excludes 0.

Compute. Every episodic run and every online run on the benchmarks used one A100 (80 GB). The benchmark experiments, including all controls and 10 stream orders, took about 3,100 GPU-hours; the ImageNet-21k subsets took about 11,000, most of it at $N = 21{,}841$ (appendix G.8).

Results manifest. With the code we will release the configuration of every run (loss, λ, level, seed, stream order), per-seed and per-order results for every table cell, the class-subset lists of ImageNet-21k, and the explicit certificate configurations of appendix B.4.

Table 6: **Other backbones and distribution shift**, accuracy / ECE. (a) Averages over the 11 datasets with RN50 (D=1024) and ViT-L/14 (D=768); per dataset in appendix G.2. (b) ImageNet shifts with ViT-L/14. RN50 baselines as reported in their papers. ‡: our runs of the official code with published hyperparameters, in settings the method's paper does not report. Random flip set: our null control at AGI-TPT's override mass (20 draws). All rows: one TPT step.

(a) Backbones

Method	RN50	ViT-L/14‡
Zero-shot	55.9 / 5.61	71.1 / 5.0
TPT	57.7 / 11.7	71.9 / 14.7
C-TPT	57.8 / 6.20	71.9 / 9.8
O-TPT	57.47 / 5.45	71.3 / 7.6
D-TPT	57.13 / 5.91	71.7 / 5.8
A-TPT	57.53 / 2.92	72.0 / 4.3
Random flip set	56.62 / 2.71	71.7 / 3.6
AGI-TPT	**57.99 / 2.52**	**72.5 / 3.3**

(b) ImageNet shifts, ViT-L/14

Method	IN-A	IN-V2	IN-R	IN-Sketch	Avg.
Zero-shot‡	68.6 / **4.0**	67.7 / **3.8**	85.2 / 2.6	57.6 / 7.5	69.8 / 4.5
TPT‡	**73.6** / 13.4	70.2 / 16.1	**87.6** / 5.2	**59.5** / 21.8	**72.7** / 14.1
C-TPT‡	72.2 / 11.4	69.7 / 14.2	86.5 / 3.7	59.3 / 20.8	71.9 / 12.5
O-TPT‡	71.1 / 10.5	69.2 / 10.9	85.8 / 2.9	58.6 / 15.0	71.2 / 9.8
D-TPT‡	70.6 / 7.9	68.8 / 7.4	85.9 / 2.2	58.2 / 11.3	70.9 / 7.2
A-TPT‡	71.9 / 6.1	69.3 / 6.4	86.5 / 1.6	58.6 / 8.3	71.6 / 5.6
Random flip set	69.9 / 4.9	68.6 / 5.4	86.1 / 1.6	57.6 / 5.7	70.6 / 4.4
AGI-TPT	73.2 / 4.3	**70.3** / 4.9	**87.6** / **1.2**	59.3 / **5.1**	72.6 / **3.9**

G Experimental details and main result tables

All results use the protocol of section 5 and appendix F.5: ViT-B/16 unless stated, 5 seeds for episodic runs and 10 stream orders for online runs. Bracketed intervals are 95% paired bootstrap intervals over test images and seeds or orders (appendix F.5). Cells marked ‡ are our runs of a baseline's official code in a setting its paper does not report. Table S1 in the supplementary material lists which experiment tests which claim.

G.1 Baselines and their reproduction

Baseline numbers in the settings a paper reports are copied from that paper and never re-run; the papers use their own seed protocols. Where a setting is missing, we run the official implementation, unmodified, and mark the cell ‡. To show how close our runs are to the published ones, we also ran each baseline at its published weight in a setting its paper does report (ViT-B/16, 11 datasets, one step): average ECE 5.19 for C-TPT (published 5.13), 4.31 for O-TPT (4.23), 4.26 for D-TPT (4.18) and 2.68 for A-TPT (2.61), all within 0.08 (our 15-bin ECE against each paper's printed value); the accuracies are within 0.3 points. Without its Householder step, O-TPT‡ reaches 4.52 (O-TPT reports 4.45 in its own ablation). At its published weight each baseline overrides a different share of TPT's gradient, from 0.46 (C-TPT) to 0.61 (A-TPT), so published-weight comparisons partly compare strengths; table 7 compares them at equal strength.

G.2 Other backbones

Table 6 summarizes RN50 and ViT-L/14 and the ImageNet shifts with ViT-L/14; tables S6 and S7 give the per-dataset numbers. All are one TPT step (episodic). The ImageNet shifts with ViT-B/16 are in table 1.

G.3 One step at equal strength

Table 7 is the one-step comparison that proposition 1 and Proposition C.1 call for: every regularizer is run at the weight at which it overrides the same share of TPT's gradient as AGI-TPT. At equal strength the methods are close, and AGI-TPT and O-TPT's penalty without its Householder step agree to the last digit on ImageNet, as proposition 9 requires ($c^\star = 0$ there). The one-step gains in table 1 are therefore mostly gains of strength; the gains that geometry buys appear when adaptation continues (section 5.2).

G.4 Online results per dataset

Table 8 gives the per-dataset numbers behind the online rows of table 1, with the noise floor of each test set, the paired interval, the ECE of the prompt at the end of the stream and a debiased ECE.

Table 7: **One step, at equal strength** (CLIP ViT-B/16, ECE in %, 11 datasets). Every regularizer's weight is set so that it overrides the same share of TPT's gradient as AGI-TPT (override mass μ, measured without labels on held-out images; appendix F.5). *Noise floor*: the expected ECE of a perfectly calibrated predictor with AGI-TPT's confidences on that test set. ‡: our runs of the official code. *: difference to AGI-TPT resolved (paired 95% interval over images and seeds excludes 0). Published-weight results are in table 1.

Method	ImNet	DTD	Flowers	Food	SUN	Aircraft	Pets	Caltech	UCF	EuroSAT	Cars	ECE	Acc.
noise floor	*0.5*	*2.6*	*2.0*	*0.5*	*0.7*	*1.8*	*1.1*	*1.1*	*1.6*	*1.2*	*1.0*	*1.28*	—
Zero-shot	2.12	8.50	3.00	2.39	2.53	5.11	4.37	5.50	3.59	7.40	4.25	4.43	63.7
TPT	10.6	21.2	13.5	3.98	11.3	16.8	5.77	4.51	13.0	21.5	5.16	11.6	65.0
Random flip set	1.05*	3.62	2.66	1.02	1.30	2.41	1.58	1.62	2.23	2.48	1.39	1.94	64.30
C-TPT‡	1.44*	4.31*	2.95	1.32*	1.79*	3.02	1.55	1.94	2.52	4.08*	1.26	2.38	64.78
D-TPT‡	1.36*	3.85	2.86	1.41*	1.63*	2.71	1.66	1.90	2.36	3.61*	1.51	2.26	64.69
A-TPT‡	1.52*	4.70*	3.45*	1.33*	2.90*	3.05	1.30	2.70*	2.30	3.80*	1.22	2.57	64.95
O-TPT‡	0.96	3.44	2.41	0.98	1.27	2.21	1.52	1.47	2.08	2.81*	1.21	1.85	65.02
without Householder‡	0.93	3.41	2.46	0.97	1.25	2.27	1.49	1.44	2.04	2.79*	1.19	1.84	65.11
AGI-TPT	0.93	3.38	2.45	0.97	1.25	2.26	1.44	1.44	2.05	2.52	1.19	**1.81**	**65.20**

On every one of the 11 datasets the online ECE of AGI-TPT is within 0.15 of the noise floor, the ECE that a perfectly calibrated predictor with the same confidences would show on the same test set. Debiased for this finite-sample noise, its average calibration error is 0.50 on the 11 datasets and 1.02 on the four shifts. The end-of-stream ECE is at or below the cumulative one on every dataset: the prompt does not drift.

Table 8: **AGI-TPT online, per dataset** (ViT-B/16; 10 stream orders). Acc.: mean ± s.d. over orders. ECE [·]: 95% interval over orders and images. Floor: expected ECE of a perfectly calibrated predictor with the same confidences on the same test set. ECE_{end}: the prompt frozen at the end of the stream, evaluated on the whole test set. deb.: debiased ECE.

	ImNet	DTD	Flwr	Food	SUN	Airc	Pets	Cal	UCF	ESAT	Cars	Avg.	IN-A	IN-V2	IN-R	IN-S	Avg.
Test images	50,000	1,692	2,463	30,300	19,850	3,333	3,669	2,465	3,783	8,100	8,041		7,500	10,000	30,000	50,889	
Acc.	70.93	50.62	72.47	86.58	68.61	26.94	90.57	94.52	72.18	55.09	68.38	68.81	62.34	65.41	80.52	52.73	65.25
s.d.	0.08	0.46	0.31	0.07	0.12	0.38	0.21	0.17	0.29	0.52	0.18		0.34	0.22	0.10	0.09	
ECE	0.58	2.74	2.09	0.61	0.79	1.92	1.19	1.17	1.70	1.33	1.08	1.38	1.92	1.36	0.83	1.49	1.40
lower	0.52	2.41	1.81	0.53	0.70	1.67	1.00	0.95	1.49	1.19	0.96		1.72	1.20	0.75	1.40	
upper	0.64	3.07	2.37	0.69	0.88	2.17	1.38	1.39	1.91	1.47	1.20		2.12	1.52	0.91	1.58	
Floor	0.5	2.6	2.0	0.5	0.7	1.8	1.1	1.1	1.6	1.2	1.0	1.28	1.4	1.1	0.6	0.5	0.90
ECE_{end}	0.55	2.64	2.04	0.59	0.76	1.82	1.16	1.14	1.66	1.25	1.04	1.33	1.85	1.31	0.80	1.45	1.35
deb.	0.29	0.86	0.61	0.35	0.37	0.67	0.45	0.40	0.57	0.57	0.41	0.50	1.31	0.80	0.57	1.40	1.02

G.5 Online baselines per dataset

Table 9: **Sustained adaptation** (online, ViT-B/16; six-dataset averages, 10 stream orders; every weight selected on held-out CIFAR-100). ECE over the stream [95% interval]; ECE_{end}: final prompt on the whole test set; abs.: ImageNet classes no longer predicted; res.: datasets (of 6) where the gap to AGI-TPT is resolved. Noise floor 1.30. ‡: our runs.

Method	Kind	Acc. ↑	ECE ↓	ECE_{end} ↓	abs. ↓	res.
TPT‡	no regularizer	21.4	55.8 [54.3, 57.3]	76.9	861	6
+ periodic reset‡	drift remedy	48.3	16.2 [15.6, 16.8]	16.9	38	6
+ SAR‡	drift remedy	51.6	13.1 [12.6, 13.6]	13.4	17	6
DynaPrompt	prompt buffer	53.0	10.7‡ [10.2, 11.2]	11.1	0.4	6
C-TPT‡	spread, tuned	51.8	6.9 [6.5, 7.3]	8.8	2.9	6
D-TPT‡	spread, tuned	51.5	4.6 [4.3, 4.9]	5.7	2.0	6
O-TPT‡	spread, tuned	50.7	3.9 [3.6, 4.2]	4.8	1.0	6
A-TPT‡	spread, tuned	52.1	3.6 [3.4, 3.8]	4.4	0.8	6
+ SAR‡	best baseline + remedy	52.4	2.9 [2.7, 3.1]	3.3	0.6	6
ETF target (neural collapse)	spread, tuned	52.6	4.0 [3.7, 4.3]	5.2	1.5	6
AGI-TPT	spread, tuned	**56.8**	**1.4** [1.3, 1.5]	**1.3**	**0.3**	—

Table 10 gives every online row of table 9 per dataset. Every method's weight was selected on the same held-out online stream (CIFAR-100) from a grid of $\frac{1}{4}$ to 4 times its published weight: the selected multipliers are $\times 4$ for C-TPT, $\times 2$ for O-TPT, D-TPT and A-TPT. At the equal starting override mass instead, O-TPT‡ and A-TPT‡ reach 4.1 and 3.8, so the selection helps them. Drift remedies applied to the best baseline, A-TPT‡, give 3.2 with a periodic reset, 3.1 with an EMA teacher and 2.9 with SAR; SAR is the best and is the row shown in the main text. Every baseline is worse than AGI-TPT on every dataset, and each difference is resolved.

Table 10: **Online baselines per dataset** (ViT-B/16, ECE over the whole stream; mean over 10 stream orders). DynaPrompt's accuracies are published (Xiao et al., 2025); its ECE is our run‡. Every per-dataset difference to AGI-TPT is resolved.

Method	ImNet	SUN	Cars	Airc	DTD	ESAT	Avg. ECE	Avg. Acc.
noise floor	*0.5*	*0.7*	*1.0*	*1.8*	*2.6*	*1.2*	*1.30*	—
TPT‡	60.1	57.2	49.3	61.4	53.6	53.2	55.8	21.4
+ periodic reset‡	13.4	14.1	9.5	20.2	20.6	19.4	16.2	48.3
+ SAR‡	9.8	10.5	5.2	16.4	19.1	17.6	13.1	51.6
DynaPrompt	8.1‡	9.2‡	5.9‡	14.3‡	13.6‡	13.1‡	10.7	53.0
C-TPT‡	5.1	6.2	4.3	7.9	9.6	8.3	6.9	51.8
D-TPT‡	3.6	4.1	3.2	5.0	5.9	5.8	4.6	51.5
O-TPT‡	2.9	3.4	2.5	4.3	5.2	5.1	3.9	50.7
A-TPT‡	2.5	3.1	2.2	4.1	5.0	4.7	3.6	52.1
+ SAR‡	1.9	2.3	1.7	3.4	4.2	3.9	2.9	52.4
ETF target (neural collapse)	3.8	3.7	2.9	4.3	4.6	4.9	4.0	52.6
AGI-TPT	**0.58**	**0.79**	**1.08**	**1.92**	**2.74**	**1.33**	**1.41**	**56.8**

G.6 Ablations

Table 11 changes one component of AGI-TPT at a time, at the same weight; section 5.5 summarizes it.

Table 11: **Ablations: one component changed at a time** (ViT-B/16, six-dataset averages; every variant at AGI-TPT's weight). Below $c^\star$: share of the pair force spent on pairs already at or below the attainable level (attraction for the two-sided residual, over-pushing without a level). r_s: Spearman correlation between the class cosines at the end of the stream and those of zero-shot CLIP (section 5.3). abs.: ImageNet classes no longer predicted. At one step, AGI-Fixed equals AGI-TPT and AGI-L0 equals O-TPT without its Householder step, exactly (Proposition C.2 and proposition 9).

		One step		Online				
Variant	What changes	ECE↓	Acc.↑	Acc.↑	ECE↓	below $c^\star$ (%)	r_s ↑	abs.↓
AGI-TPT	level $c^\star$, ReLU residual	1.92	52.47	**56.8**	**1.4**	0	**0.92**	0.3
AGI-Fixed	two-sided residual $(s - c^\star)^2$	1.92	52.47	55.7	3.4	27	0.72	1.2
AGI-L0	level 0 (O-TPT's)	1.97	52.32	56.0	1.8	0	0.91	0.3
AGI-Lsimplex	level $-\frac{1}{N-1}$, ignores D	1.92	52.47	**56.8**	**1.4**	<0.1	**0.92**	0.3
AGI-NoLevel	no level ($c = -1$): never stops	2.05	52.21	53.3	7.2	37	0.80	4.3
annealed, offset norm	weight rescaled to our offset norm	—	—	55.2	2.4	26	0.84	1.3
annealed, push mass	weight rescaled to our summed force	—	—	55.4	2.3	25	0.85	1.1
annealed, per pair	each pair's weight decays over the stream	—	—	55.7	2.2	21	0.87	0.9
AGI-Angle	$\left[\mathrm{ReLU}(\theta^\star - \theta_{ij})\right]^2$ in angle	1.94	52.40	56.7	**1.4**	0	**0.92**	**0.2**
AGI-NN	force on each class's nearest pair only	2.19	52.05	55.5	2.4	0	0.88	0.7
AGI-PerClass	per-class term, no pair term	2.28	51.92	52.3	5.5	—	0.66	2.4

Stopping versus annealing. AGI-TPT's push shrinks as pairs reach the level, so part of its advantage over AGI-NoLevel could be an implicit annealing schedule. Three annealed versions of AGI-NoLevel match our offset norm after every image, our summed pair force, or decay each pair's weight over the stream (appendix F.5). They reach 2.2–2.4, against 1.4, and still spend 21–26% of their force on pairs already below $c^\star$ (table 11). The level does two things: it anneals, and it redirects the push from separated pairs to those that are not yet separated.

G.7 THE LEVEL ON EVERY DATASET

Design. Proposition 10 makes two predictions about levels near $c^\star$. A level $c < c^\star$ can never be reached, so some pair is always pushed; a level $c > c^\star$ can be reached but stops pairs early. Neither says how large the effect on calibration is, so we test them on all 18 datasets we use, which span N from 7 to 1000: the 11 benchmarks, the four ImageNet shifts, FER-2013 (Goodfellow et al., 2013), CIFAR-10 (Krizhevsky, 2009) and STL-10 (Coates et al., 2011). We run online AGI-TPT at $\lambda = 4$ with levels $c^\star + \xi$, $\xi \in \{-0.05, -0.02, 0, +0.02, +0.05\}$, and with the level of the other regime (0 where $N \le D + 1$, $-\frac{1}{N-1}$ where $D + 2 \le N \le 2D$), over 10 stream orders. Before running, we stated two predictions:

(P1) the minimum over ξ is at $\xi = 0$ on every dataset;

(P2) a level below $c^\star$ costs more than one equally far above it (proposition 10: it never stops);

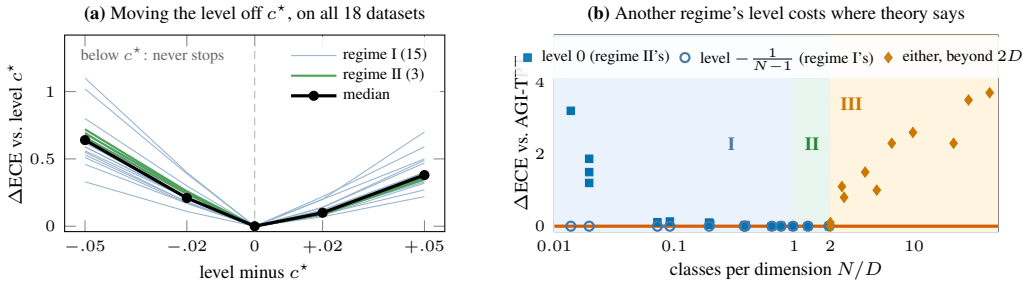


Figure 4: **Each regime needs its own level** (online, our runs). **(a)** ECE when the level is moved off $c^\star$ by ξ, on all 18 datasets: the minimum is at $c^\star$, and a level below $c^\star$, which can never be reached, costs more than one equally far above it. **(b)** Cost of another regime's level, relative to AGI-TPT (orange line). Level 0 is too high for $N \le D{+}1$, exact for $D{+}2 \le N \le 2D$ (where the simplex level differs from it by less than $\frac{1}{D+1}$), and unreachable beyond $2D$ (ImageNet-21k subsets, three backbones).

Results (table 12 and fig. 4a). (P1) holds on all 18 datasets. The minimum is resolved against $\xi = -0.02$ on 14 and against $\xi = +0.02$ on 10; the unresolved cases are the datasets with the smallest test sets. (P2) holds on all 18. The median cost over the 18 datasets is 0.21 at $\xi = -0.02$ against 0.10 at $\xi = +0.02$, and 0.64 against 0.38 at ± 0.05. The same minimum holds at $\lambda = 2$ and $\lambda = 8$ on the six datasets with fewest classes (table S17), so it is not an artefact of the weight.

The level of the other regime. Where $N \le D + 1$, replacing $c^\star = -\frac{1}{N-1}$ by level 0 costs in proportion to how far 0 is from $c^\star$: 3.2 ECE points on FER-2013 ($|c^\star| = \frac{1}{6}$), 1.2–1.9 on the three 10-class datasets, and at most 0.13 once $N \ge 37$. Where $D + 2 \le N \le 2D$, replacing $c^\star = 0$ by the simplex level $-\frac{1}{N-1}$ costs 0.01, as it should: the two levels differ by at most $\frac{1}{D+1} \approx 0.002$ there. On CLIP, where $D \ge 512$, the regime rule therefore matters most at few classes, where level 0 is too high, and beyond $2D$, where both levels are unreachable (appendix G.8).

G.8 THE REGIME BOUNDARY ON IMAGENET-21K

Levels. AGI-NoLevel ($c = -1$); $c = 0$; Levenshtein's bound L, which lies between 0 and $c^\star$ and so is predicted to cost no more than 0; $c^\star$, exact for $N \le 2D$, and the certified value $U \ge c^\star$ of appendix B.4 beyond; and Welch's value where it is feasible and above $c^\star$ ($D + 2 \le N \le 2D$). Beyond $2D$ we know only $L \le c^\star \le U$, so every statement there is about a certified level of the right order, not about $c^\star$ itself.

Results (table 13 and fig. S8).

- $N \le 2D$ *and* $N = 2D + 2$*:* level 0 is within 0.1 ECE of $c^\star$, as it should be where $c^\star \le 1/D$.
- *Beyond:* the cost of level 0 grows from 0.8 to 3.7 ECE points as the lower bound $L(N, D)$ grows, and its gap turns negative: the extra push produces under-confidence. The cost tracks the share f_0 of pairs that end the stream between 0 and U, the pairs that level 0 keeps pushing and U would stop; this relation is close to definitional and we do not use it as evidence for $c^\star$.

Table 12: **The level on all 18 datasets** (online AGI-TPT, ViT-B/16, $\lambda = 4$; ECE, mean over 10 stream orders). $*$: difference to $\xi = 0$ resolved (paired 95% interval over orders and images excludes 0). Other regime: level 0 for $N \le D + 1$ and $-\frac{1}{N-1}$ for $D + 2 \le N \le 2D$. Reached: pairs whose cosine reaches the level at some step. Bold: minimum per row.

				level $c^\star + \xi$					other regime's level		
Dataset	N	regime	$c^\star$	-0.05	-0.02	0	$+0.02$	$+0.05$	level	ECE	reached (%)
FER-2013	7	I	-0.1667	4.20	3.50*	**3.10**	3.30*	3.80	0	6.30	44
EuroSAT	10	I	-0.1111	2.35	1.72*	**1.33**	1.55*	1.92	0	3.20	38
CIFAR-10	10	I	-0.1111	2.20	1.70*	**1.40**	1.60*	1.90	0	2.90	36
STL-10	10	I	-0.1111	1.80	1.30*	**1.10**	1.20	1.50	0	2.30	33
Pets	37	I	-0.0278	1.52	1.30	**1.19**	1.26	1.41	0	1.30	12
DTD	47	I	-0.0217	3.30	2.95	**2.74**	2.86	3.12	0	2.87	11
Caltech101	100	I	-0.0101	1.63	1.34	**1.17**	1.26	1.44	0	1.21	15
Aircraft	100	I	-0.0101	2.61	2.18*	**1.92**	2.06	2.39	0	2.00	18
Food101	101	I	-0.0100	1.12	0.78*	**0.61**	0.69*	0.93	0	0.65	16
UCF101	101	I	-0.0100	2.29	1.91*	**1.70**	1.82	2.10	0	1.76	14
Flowers102	102	I	-0.0099	2.62	2.27	**2.09**	2.20	2.47	0	2.14	13
Cars	196	I	-0.0051	1.67	1.29*	**1.08**	1.18*	1.42	0	1.09	19
ImageNet-A	200	I	-0.0050	2.64	2.17*	**1.92**	2.06*	2.41	0	1.95	17
ImageNet-R	200	I	-0.0050	1.38	1.01*	**0.83**	0.93*	1.16	0	0.86	15
SUN397	397	I	-0.0025	1.41	1.00*	**0.79**	0.89*	1.15	0	0.80	22
ImageNet	1000	II	0	1.27	0.83*	**0.58**	0.67*	0.94	-0.0010	0.59	24
ImageNet-V2	1000	II	0	2.02	1.60*	**1.36**	1.45	1.71	-0.0010	1.37	21
ImageNet-Sketch	1000	II	0	2.21	1.75*	**1.49**	1.58*	1.88	-0.0010	1.50	23

- *Level L:* it never costs more than level 0, and it saves more the larger L is.
- *Certified level U:* its ECE stays flat from $N = 100$ to 21,841 on all three backbones. How far U lies above $c^\star$ is bracketed in table S4; table S18 shows that ECE varies by at most 0.2 (two steps of the printed resolution) for levels between $0.75U$ and $1.25U$.
- *Welch:* it stops early wherever it applies ($+0.4$ to $+0.6$ ECE).
- *AGI-NoLevel:* it costs 5.1 to 8.2 points everywhere.

Table 13: **Online TPT on ImageNet-21k class subsets, for several levels** (ECE; mean over stream orders). $c^\star$: exact value for $N \le 2D$; beyond, the bracket $[L, U]$, and the column "$c^\star$ or U" uses U. Acc.: accuracy at that level. gap: confidence minus accuracy with level 0. f_0: share of pairs ending the stream between level 0 and U under level 0. Welch's value is shown where it is feasible and above $c^\star$.

Backbone	D	N	$\frac{N}{2D}$	$c^\star$ or $[L, U]$	$c = -1$	$c = 0$	$c = L$	$c = c^\star$ or U	Welch	Acc. ($c^\star$/U)	gap ($c = 0$)	f_0 (%)
ViT-B/16	512	100	0.10	-0.0101	7.4	2.3	—	**2.2**	—	83.1	+0.6	0
		500	0.49	-0.0020	7.9	**2.1**	—	**2.1**	—	71.4	+0.3	0
		1000	0.98	0	8.3	**2.0**	—	**2.0**	2.6	66.2	+0.2	0
		1026	1.00	[$<10^{-4}$, 0.0020]	8.4	2.1	2.1	**2.0**	—	65.9	+0.1	0.3
		1280	1.25	[0.0010, 0.0313]	8.8	3.1	3.0	**2.0**	—	62.8	-1.9	3.3
		2000	1.95	[0.0037, 0.0313]	9.6	3.4	3.3	**1.9**	—	57.4	-2.7	6.4
		5000	4.88	[0.0136, 0.0625]	9.9	4.5	3.7	**1.9**	—	46.3	-3.9	13.2
		21,841	21.3	[0.0333, 0.0625]	10.2	5.7	3.0	**2.0**	—	29.8	-5.3	17.5
ViT-L/14	768	500	0.33	-0.0020	7.0	**1.9**	—	**1.9**	—	78.9	+0.3	0
		1000	0.65	0	7.6	**1.8**	—	**1.8**	2.2	73.8	+0.2	0
		1538	1.00	[$<10^{-4}$, 0.0013]	8.0	**1.8**	**1.8**	**1.8**	—	69.9	+0.1	0.2
		2000	1.30	[0.0008, 0.0259]	8.9	2.5	2.5	**1.7**	—	67.1	-1.2	2.6
		5000	3.26	[0.0057, 0.0625]	9.3	4.1	3.6	**1.8**	—	56.2	-3.6	10.4
		21,841	14.2	[0.0216, 0.0625]	9.6	5.3	3.2	**1.8**	—	38.7	-4.7	15.1
RN50	1024	1000	0.49	-0.0010	8.1	**2.4**	—	**2.4**	—	63.4	+0.5	0
		2000	0.98	0	8.6	**2.3**	—	**2.3**	2.8	55.9	+0.2	0
		2050	1.00	[$<10^{-4}$, 0.0010]	8.8	**2.3**	**2.3**	**2.3**	—	55.6	+0.2	0.1
		5000	2.44	[0.0028, 0.0313]	9.8	3.2	3.1	**2.2**	—	44.1	-2.0	6.0
		21,841	10.7	[0.0146, 0.0313]	10.3	4.5	3.2	**2.2**	—	27.5	-3.6	11.8

G.9 FAILURE CASES AND LIMITS

- *Where another method is better.* Online, PTA is more accurate on Flowers102, EuroSAT, Cars and ImageNet-R (table 1), and A-TPT is better calibrated on Pets (1.17 vs. 1.44 for AGI-TPT at one step and 1.19 online). With RN50 at one step, A-TPT has a lower ECE on DTD (6.41 vs. 7.43) and Food101 (1.11 vs. 1.94); with ViT-L/14, A-TPT‡ is lower on Food101, and zero-shot CLIP‡ is better calibrated than every adapted method on SUN397, IN-A and IN-V2. At one step with ViT-B/16, D-TPT is better calibrated on IN-Sketch (2.47 vs. 3.27).
- *Sampling noise.* On small test sets (DTD, Flowers, Caltech, Pets), ECE differences below about 0.2 are within the noise floor and are not resolved (table 12).
- *Under-confident starting points.* When TPT inflates confidence little, the override overshoots into under-confidence (ImageNet at $\alpha = 50$: 4.8 vs. TPT's 2.4; Supp. S5.3).
- *Genuinely confusable classes.* In the synthetic hierarchical control, where classes within a superclass truly overlap, every sustained spread push over-separates them and none beats zero-shot (Supp. S7.1). The attainable level bounds the damage but does not remove it.
- *Reachable geometry.* A 4-token prompt cannot reach the zero set of $\mathcal{L}_{\text{AGI}}$ (largest cosine 0.44 on ImageNet at convergence; Supp. S5.5). AGI-TPT stops pushing individual pairs, those that reach the level, rather than bringing the configuration to rest; the separation guarantee of lemma 4 is for free class vectors.
- *Certified levels.* Beyond $2D$ the level is a certified upper bound U on $c^{\star}$, which may stop some pairs early by up to the width of the bracket of table S4; near $2D$ that bracket is two orders of magnitude wide. ECE is flat for levels within 25% of U, but we cannot say where $c^{\star}$ lies inside the bracket, and tighter numerical codes are not used here.
- *One step.* At one step the random flip set reaches most of the calibration of every method we test (table S9); one-step calibration is therefore weak evidence for any geometry, including ours, which is why table 1's episodic block is a like-for-like comparison of published methods, not evidence for the level or the ReLU.