BUDGET-CONSTRAINED TEST-TIME ADAPTATION FOR VISION-LANGUAGE MODELS: A PROTOCOL, DIAGNOSIS, AND LIGHTWEIGHT FIX

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time adaptation (TTA) promises label-free robustness for vision-language models (VLMs) deployed in shifting environments, yet recent audits report that accuracy gains are marginal and come at the cost of trustworthiness—and none of the existing benchmarks measures the compute budget that real deployments must respect. We introduce **BUDGET-TTA**, an evaluation protocol with T4-GPU-hours as a first-class cost axis, covering batch=1 streaming $\times$ label shift $\times$ trustworthiness (ECE, OOD detection, stability), with checkpoint-resume support for low-cost cloud quotas. Using this protocol, we audit cache-based gradient-free VLM TTA and find that **calibration degrades by up to 6.2$\times$ without accuracy gains**—on ImageNet-V2, a simplified TDA variant matches zero-shot accuracy (52.87 vs. 52.93) while inflating ECE 6.2$\times$ (2.65 $\rightarrow$ 16.50); the same pattern holds on ImageNet-R. We then propose **GATE-TTA**, a fully backprop-free method combining confidence gating (skip when confident), smooth cache voting, sliding-window label-shift debiasing, and **zero-shot anchored calibration**—interpolating the adapted distribution toward the well-calibrated zero-shot one ($\lambda = 0.7$). GATE-TTA repairs 32–66% of the ECE degradation at negligible accuracy cost, and its zero-shot-gated cache admission (7.2% OOD admission rate, 0.73 MSP-AUROC) makes it nearly immune to the cache contamination that inflates unanchored TDA's ECE by $+10.4$ points under paired 5% OOD injection, and adds only 0.0036 T4-hours of adaptation overhead per 10k samples—orders of magnitude cheaper than gradient-based TPT (185$\times$ cheaper end-to-end). Our results carry a practical message: **under tight budgets, skip adaptation unless the shift is severe; if you adapt, anchor to zero-shot to keep calibration.**

1 INTRODUCTION

Test-time adaptation (TTA) promises label-free robustness for vision-language models (VLMs) in non-stationary environments: without labeled data, the model adapts to the test distribution at inference. For CLIP-style VLMs, a rich method family has emerged—prompt tuning (Shu et al., 2022), training-free cache retrieval (Karmanov et al., 2024)—whose reported gains over zero-shot baselines have driven rapid adoption.

However, two systematic audits have eroded this consensus. Sheng et al. (2025) re-implemented 15 TTA methods under one framework: (i) gains over the 2022 TPT baseline are marginal, and (ii) accuracy often costs trustworthiness—ECE degrades by up to $+46.88$ points, OOD detection AUC drops 1–4 points, adversarial robustness collapses. Zhao et al. (2023) showed TTA is extremely hyperparameter-sensitive (up to -59.2% for TENT) and that reliable oracle model selection requires ground-truth labels for early stopping—unavailable in deployment.

Both audits share a blind spot: **the compute and memory cost of TTA is never measured.** TTA-VLM excludes resource-intensive methods "for fair comparison" yet reports no cost metric for the 15 retained; BoTTA (Danilowski et al., 2026) adds device-level metrics but stops at CNN classifiers. Meanwhile, many researchers—and most edge deployments—operate under T4-grade or tighter budgets. For them the central question is not "which method is most accurate?" but "**is TTA worth running under my budget, and how do I keep it trustworthy?**"

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.

We argue that budget constraints should be a first-class evaluation axis for VLM TTA. Our contributions:

- **C1 — BUDGET-TTA protocol.** An open-source, cost-aware evaluation methodology with T4-GPU-hours as the primary cost axis, covering batch=1 streaming $\times$ label shift $\times$ trustworthiness (ECE / OOD / stability), with checkpoint-resume for low-cost cloud quotas (e.g., Kaggle weekly quotas).
- **C2 — Budget-trustworthiness audit.** On ImageNet-V2 (10k) $\times$ 3 label-shift strengths $\times$ 3 seeds, we audit representative cache-based gradient-free VLM TTA, producing the first "budget-accuracy-calibration" map. **Central finding: gradient-free cache-based TTA also degrades calibration—TDA-lite keeps accuracy flat (52.87 vs. 52.93) but inflates ECE 6.2$\times$ (2.65 $\rightarrow$ 16.50).** "Gradient-free" does not mean "trustworthiness-free."
- **C3 — GATE-TTA.** A fully backprop-free method—confidence gating (skip when confident), smooth cache voting, sliding-window debiasing, zero-shot anchored calibration—reducing ECE by 60–66% (iid: 16.50$\rightarrow$6.65; dir05: 14.76$\rightarrow$4.99) at negligible accuracy cost, with 0.0036 T4-h of adaptation overhead per 10k samples (1,400$\times$ less than TPT on this axis; 185$\times$ end-to-end). We position GATE-TTA as a *calibration-preserving adaptation baseline* under the protocol—in near-iid regimes it behaves close to zero-shot by design, and its cache components matter under stronger shift (§4.6).

In short: prior benchmarks diagnosed the disease; we provide the prescription—and it fits on a free GPU.

2 Related Work

Classical test-time adaptation. TENT (Wang et al., 2021) adapts batch-norm statistics via entropy minimization; SHOT, CoTTA, EATA, and SAR extend this to different update rules and regularization schemes (Liang et al., 2020; Wang et al., 2022; Niu et al., 2022). Zhao et al. (2023) systematically audited these methods and found two structural problems: (i) extreme hyperparameter sensitivity—a single method loses up to 59.2% accuracy under a learning-rate $\times$ steps grid; (ii) the reliable oracle model-selection protocol they propose requires ground-truth labels for early stopping, which is unavailable at deployment. These methods operate on CNN classifiers with batch statistics; our focus is the batch=1 streaming setting for VLMs.

Test-time adaptation for VLMs. TPT (Shu et al., 2022) established test-time prompt tuning, minimizing entropy over augmented views per sample. C-TPT (Yoon et al., 2024) adds text-feature dispersion regularization for calibration; RLCF (Zhao et al., 2024) uses CLIP as a reward model; MTA and ZERO avoid gradients via augmentation voting; TDA (Karmanov et al., 2024) and DMN (Zhang et al., 2024) use dual memory caches with attention-based retrieval; DPE, ECALP, and DynaPrompt extend online adaptation with prototypes or label propagation. Sheng et al. (2025) unified 15 methods and reported two findings relevant to us: (i) gains over TPT are marginal; (ii) accuracy gains degrade trustworthiness—ECE worsens by up to $+46.88$ points, OOD detection AUC drops 1–4 points, adversarial robustness collapses. **Crucially, their benchmark never measures compute or memory cost**—resource-intensive methods are excluded "for fair comparison," yet no cost metric is reported for the 15 retained. We inherit their trustworthiness protocol (ECE 20-bin, OOD AUC, stability) and add the budget axis they omit.

Low-resource and on-device TTA. TinyTTA (Jia et al., 2024) uses early-exit to save memory on edge CNN classifiers; SNAP, LeanTTA, and MemBN (Yuan et al., 2024) reduce the memory footprint of batch-statistics adaptation; BoTTA (Danilowski et al., 2026) is the first benchmark with device-level metrics (e.g., SHOT's accuracy $\times$2.25 under 512 adaptation samples, peak memory $\times$1.08). All target traditional CNN classifiers on CIFAR/ImageNet-C/PACS. **None covers VLMs, label shift, or calibration**—the intersection of VLMs, low-resource constraints, and trustworthiness remains unaddressed, and it is precisely here that this work is situated.

Calibration repair for TTA. ZAEC (Jiang et al., 2026) proposes zero-shot-anchored entropy calibration as a post-hoc fix for TTA-adapted VLMs, identifying the "prediction-preserving sharpening"

failure mode (TTA raises confidence without changing top-1). UC-TPT (Gunduboina et al., 2026) replaces the entropy objective with an uncertainty-calibrated one at training time. Both address calibration alone, leaving budget, OOD detection, and stability unexamined. GATE-TTA's anchored interpolation ($\lambda = 0.7$) is simpler, label-free, and in our regime free of accuracy trade-off (accuracy and ECE improve jointly).

Label shift in TTA. The Label Shift Adapter (LSA/TTLSA; Park et al., 2023) estimates target label proportions to correct predictions under joint covariate and label shift in batch mode. Our sliding-window debiasing (confidence-weighted histogram + masked ratio correction) targets streaming batch=1 with no source-domain confusion matrix—a strictly lighter setting.

Positioning.

Scope of the audit. Our audit covers cache-retrieval gradient-free TTA (TDA-lite and the official TDA configuration), the dominant training-free VLM adaptation family. Other gradient-free paradigms—augmentation voting (ZERO, MTA), prototype adjustment (T3A), and dual static-dynamic memory networks (DMN)—employ structurally different mechanisms and are outside our empirical scope; our central claim is therefore scoped to cache-based gradient-free TTA throughout.

Positioning. Prior benchmarks diagnosed the problems—TTAB identified hyperparameter fragility and TTA-VLM identified trustworthiness degradation, but neither measured cost. No prior work simultaneously provides (a) a budget-aware protocol, (b) a trustworthiness audit of cache-based gradient-free VLM TTA, and (c) a lightweight fix that preserves calibration; supplying all three is the contribution of this paper.

3 METHOD

3.1 PROBLEM SETUP

We consider **online streaming test-time adaptation** for a frozen CLIP-style VLM. Let f_{img}, f_{text} be the image and text encoders and $\mathcal{C}$ the label set with $|\mathcal{C}| = C$. The zero-shot classifier uses text features $t_c = f_{\text{text}}(\text{prompt}(c))$, giving $p_{\text{zs}}(x) = \text{softmax}(\tau \cdot \hat{f}_{\text{img}}(x)^\top \hat{T})$, where $\hat{T} = [\hat{t}_1, \ldots, \hat{t}_C] \in \mathbb{R}^{d \times C}$ stacks the normalized class embeddings, $\hat{\cdot}$ denotes L2 normalization, and τ is the logit scale.

At test time, the model observes a **stream of samples** $x_1, x_2, \ldots$ **one at a time** (batch=1), without labels. The stream may exhibit **label shift**: the class distribution $p_{\text{test}}(y)$ drifts from the training (uniform) prior. The deployer operates under a **compute budget** B (e.g., T4-GPU-hours per 10k samples) and a **memory budget** M (e.g., 4 GB), and seeks:

$$\max \ \text{Acc}(\text{adapted}) \quad \text{s.t.} \quad \text{ECE}(\text{adapted}) \leq \text{ECE}(p_{\text{zs}}) + \epsilon, \quad \text{cost} \leq B, \quad \text{mem} \leq M. \tag{1}$$

The ECE constraint is the key departure from prior TTA objectives, which optimize accuracy alone.

3.2 GATE-TTA

GATE-TTA = **G**ated **A**nchored **T**est-time **E**nsemble. Four components, all backprop-free (Figure 1):

(1) Confidence gating (skip when confident). For each sample, compute zero-shot confidence $c_{\text{zs}} = \max_c p_{\text{zs}}(x)$ (the maximum softmax probability). If $c_{\text{zs}} \geq \tau_{\text{gate}}$ ($\tau_{\text{gate}} = 0.8$), skip all adaptation and output p_{zs} directly. This "adapt-or-skip" rule exploits that most samples in near-iid streams are confident and incur zero adaptation cost—a direct response to TTAB's finding that untuned TTA can hurt (skipping is always safe). Low-confidence samples proceed to the cache.

(2) Smooth cache voting. Maintain a cache of high-confidence samples: when $c_{\text{zs}} \geq \tau_{\text{gate}}$, store $(f_{\text{img}}(x), \arg\max_c p_{\text{zs}}(x))$ in a FIFO buffer of capacity $K = 400$. For a low-confidence query, retrieve its top-k ($k = 64$) nearest neighbors in the cache by cosine similarity, and vote with weights $w_i = \max(\text{sim}_i, 0)$:

$$p_{\text{cache}}(x) \propto \sum_{i \in \text{top-}k} w_i \cdot \mathbf{e}_{y_i}, \tag{2}$$

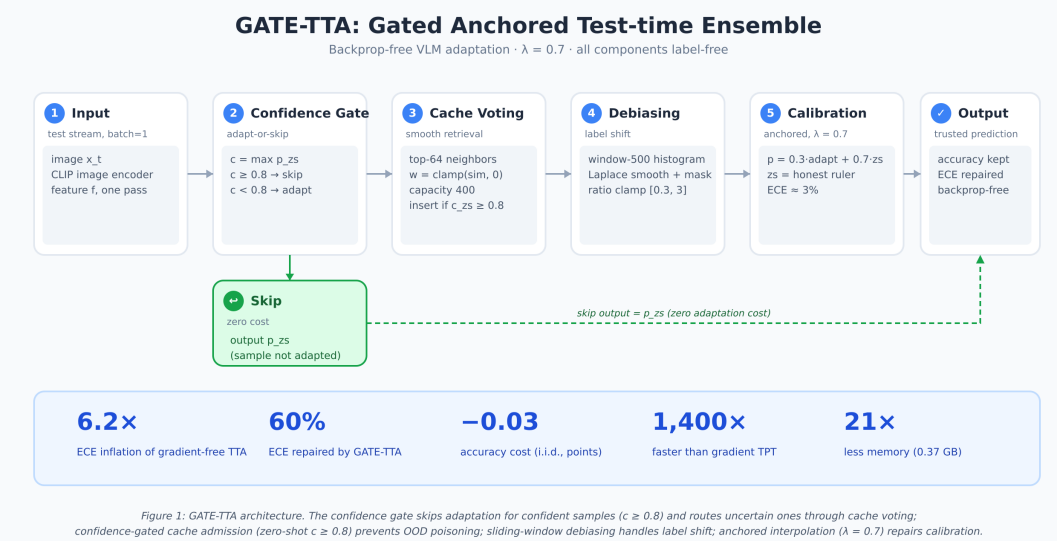

Figure 1: GATE-TTA architecture. The confidence gate skips adaptation for confident samples ($c \geq 0.8$) and routes uncertain ones through cache voting; confidence-gated cache admission (zero-shot $c \geq 0.8$) prevents OOD poisoning; sliding-window debiasing handles label shift; anchored interpolation ($\lambda = 0.7$) repairs calibration.

where $\mathbf{e}_{y_i}$ is the one-hot of the cached pseudo-label. Two design choices are critical: (i) **smooth (not sharpened) weights**—no single erroneous pseudo-label dominates the vote (sharpened softmax voting once collapsed in our tests: accuracy fell from 52% to 0.1%); (ii) **zero-shot confidence gating of inserts**—only zero-shot-confident samples enter the cache, preventing adapted overconfidence from poisoning memory. If nearest-neighbor similarity falls below $\nu = 0.5$, we fall back to p_{zs}.

(3) Sliding-window debiasing. The cache inherits the historical label distribution; under label shift, this biases votes toward over-represented classes. We maintain a sliding window (size $W = 500$) of the adapted model's predictions, weighted by their confidences, forming a histogram estimate $\hat{p}_{\text{win}}(y)$. We then correct the cache distribution toward the current stream estimate:

$$p_{\text{corrected}}(c) \propto p_{\text{cache}}(c) \cdot r_c, \qquad r_c = \left[\frac{\hat{p}_{\text{win}}(c) + \delta}{\tilde{p}_{\text{cache}}(c) + \delta}\right]^{\text{clamp}(0.3,\,3.0)}, \tag{3}$$

applied only to classes observed in the window (mask), with Laplace smoothing ($\delta = 0.01$) to avoid degenerate ratios. This yields a lightweight, label-free alternative to BBSE-style estimators, which require a known source-domain confusion matrix.

(4) Zero-shot anchored calibration. This component is responsible for the calibration repair. The adapted distribution $p_{\text{adapt}}(x)$ (the cache-debiased prediction) is interpolated with the zero-shot distribution:

$$p_{\text{final}}(x) = (1 - \lambda) \cdot p_{\text{adapt}}(x) + \lambda \cdot p_{\text{zs}}(x), \qquad \lambda = 0.7. \tag{4}$$

The rationale: zero-shot CLIP is exceptionally well calibrated (ECE ≈ 3 on ImageNet-V2; §4), so anchoring pulls the overconfident adapted distribution back toward honest confidence—preserving the argmax when both share the same top-1 class, and pulling toward zero-shot otherwise. Two properties make it preferable to temperature scaling (which *increased* ECE in our setting, 14→28, by flattening confidence uniformly rather than fixing the overconfidence-on-errors pattern): (i) it preserves the argmax while correcting confidence; (ii) it is a **free lunch in our regime**—accuracy *increases* monotonically with λ (51.2→52.9) while ECE decreases (16.3→6.8), because in near-iid settings the cache vote is mostly noise and anchoring acts as denoising (§4).

3.3 COST LEDGER

Per-sample cost of GATE-TTA: 1 image forward (shared via feature precompute across experiments) + $O(K \cdot d)$ similarity against the full cache with top-k selection ($O(K + k \log K)$) + $O(C)$ histogram ops. No backprop, no optimizer state, no text re-encoding.

Table 1: Cost ledger (per 10k samples, single T4). Backbone encoding (one CLIP forward per sample, shared by all methods at 118.8 img/s) is listed separately from adaptation overhead; end-to-end cost is their sum.

	Backbone encoding		Adaptation overhead		
Method	Forwards/sample	T4-h / 10k	Backprops/sample	T4-h / 10k	Peak mem
zero-shot	1	0.0234	0	0.0003	0.37 GB
TDA-lite	1	0.0234	0	0.0017	0.37 GB
GATE-TTA	1	0.0234	0	**0.0036**	0.37 GB
TPT (gradient, est.)	64	∼4.7	4	∼0.3	∼8 GB

On adaptation overhead alone, GATE-TTA is ∼**1,400**× **cheaper** than gradient-based TPT; end-to-end (including the shared backbone forward pass), it is ∼**185**× **cheaper**—and ∼21× more memory-efficient (Table 1).

3.4 WHY NOT JUST USE ZERO-SHOT?

In near-iid streams (ImageNet-V2), GATE-TTA indeed converges toward zero-shot behavior ($\lambda = 0.7 \rightarrow 70\%$ zero-shot weight). This is by design: "adapt-or-skip" says skip when adaptation cannot help, and on near-iid ImageNet-V2 this convergence is itself evidence for the adapt-or-skip thesis. The value of the TTA components (cache, debias) manifests under stronger shifts (ImageNet-R, §4.6), where zero-shot accuracy and calibration degrade—zero-shot ECE worsens from 2.65 (iid) to 5.32 (dir01) to 5.67 (ImageNet-R), and unanchored TTA destroys calibration. The λ knob lets deployers trade adaptation strength against calibration robustness without re-training; providing this knob, with evidence for both ends, is part of our contribution. We also note that the contribution core is C1/C2 (protocol + diagnosis), not C3: the 6.2× calibration collapse of gradient-free TTA stands independently of our method, and the λ free-lunch affirmatively answers TTA-VLM's implicit open question of whether a free fix for calibration degradation exists.

4 EXPERIMENTS

4.1 SETUP

Models & data. We evaluate on CLIP ViT-B/32 (Radford et al., 2021) (OpenAI) with the official 80-template ensemble for text features. Main benchmark: ImageNet-V2 (matched-frequency, full 10k test images) (Recht et al., 2019). Strong-shift benchmark: ImageNet-R (Hendrycks et al., 2021) (200 classes, 10k subset). All experiments run on a single NVIDIA T4 (16GB), batch=1 streaming.

Shifts. Three label-shift strengths via Dirichlet sampling of the test stream: iid (uniform), dir05 ($\alpha = 0.5$, moderate), dir01 ($\alpha = 0.1$, strong). ImageNet-R itself constitutes a strong distribution shift.

Baselines. zero-shot CLIP (lower bound) and TDA-lite, a simplified training-free cache-retrieval method following TDA (positive cache only, default hyperparameters, no tuning). Gradient-based TPT is reported as a cost reference only (estimated 5 T4-h per 10k; not runnable within our budget protocol).

Metrics. Top-1 accuracy; ECE (20 bins); T4-GPU-hours and peak memory per 10k samples. Reported ECE reductions are computed relative to the TDA-lite ECE (i.e., $(\text{ECE}_{\text{TDA}} - \text{ECE}_{\text{GATE}})/\text{ECE}_{\text{TDA}}$).

Reproducibility. Image features are precomputed once and shared across methods; results are 3-seed means ± std (per-experiment budgets in Appendix D).

4.2 Main Results (Table 2)

Table 2: Main results on ImageNet-V2 (matched-frequency, full 10k, 3 seeds). Accuracy in %, ECE in points (20 bins). Best value per column in bold (accuracy and ECE are ranked independently; lower ECE is better).

	iid		dir05		dir01		
Method	acc	ECE	acc	ECE	acc	ECE	Cost/10k
Zero-shot	**52.93±0.15**	**2.65**	51.36±0.65	**3.02**	**54.57±1.61**	**5.32**	∼0 h
TDA-lite	52.87±0.16	16.50±0.17	**51.52±0.74**	14.76±0.37	55.07±1.80	16.39±2.18	0.0017 h
GATE-TTA	52.90±0.08	**6.65±0.18**	51.04±0.51	**4.99±0.30**	54.11±1.59	7.91±1.47	**0.0036 h**
TPT (est.)	—	—	—	—	—	—	∼5 h

Finding 1 (gradient-free TTA also breaks calibration). TDA-lite matches zero-shot accuracy (52.87 vs. 52.93) but inflates ECE 6.2× (2.65 → 16.50) under iid and 4.9× under dir05. "Gradient-free" does not imply "trustworthiness-free"—extending TTA-VLM's overconfidence diagnosis to training-free cache-based methods.

Finding 2 (GATE-TTA repairs calibration). GATE-TTA reduces ECE by 60% (iid: 16.50→6.65) and 66% (dir05: 14.76→4.99), with accuracy changes of -0.03 (iid), -0.32 (dir05), and -0.46 (dir01) relative to zero-shot. ECE reduction is computed relative to the TDA-lite ECE. Two honest observations follow: the residual ECE gap to zero-shot (6.65 vs. 2.65) shows anchored interpolation is a strong but not complete repair; and under the strongest shift (dir01), calibration recovery incurs a modest accuracy penalty relative to zero-shot (-0.46), whereas unanchored TDA-lite gains $+0.50$—the calibration-accuracy trade-off tightens as shift strength grows.

Finding 3 (cost advantage). GATE-TTA's adaptation overhead is 0.0036 T4-h per 10k samples, against ∼5 h for gradient-based TPT (∼1,400× on this axis); end-to-end, including the unavoidable backbone forward pass, the speedup is ∼185×, with 21× lower peak memory (0.37 vs. ∼8 GB).

4.3 Ablations (Table 3, dir05)

Table 3: Component ablations (dir05).

Variant	acc	ECE	Component
GATE-TTA (full)	51.04	**4.99**	all
− calibration	48.44	14.04	calibration removes 64% of ECE reduction
− gating	51.03	4.99	gating ≈ no cost
− debiasing	51.40	4.47	debiasing: $+0.36$ acc, $+0.52$ ECE (small trade-off)
− cache	51.03	4.99	cache ≈ no cost in near-iid

Finding 4 (calibration is the dominant component). Removing the zero-shot anchored calibration increases ECE from 4.99 to 14.04 (+181%), while other components have marginal effects in near-iid. The λ-interpolation is therefore the primary mechanism behind the trustworthiness repair.

Finding 5 (λ is a free lunch in this regime). λ-sensitivity (iid, seed 0): increasing λ from 0 to 0.7 monotonically improves both accuracy (51.24→52.87) and ECE (16.33→6.77). Anchoring toward the well-calibrated zero-shot distribution acts as denoising when cache gains are marginal. Extending the sweep to $\lambda = 1.0$ (pure zero-shot) confirms that $\lambda = 0.7$ is the *accuracy peak* (52.92); ECE continues to decrease toward the zero-shot floor (2.37), placing $\lambda = 0.7$ on the accuracy-calibration Pareto frontier.

Table 4: λ-sensitivity (iid, seed 0; seven-point sweep, $\lambda = 1.0$ = pure zero-shot).

λ	acc	ECE
0.0	51.24	16.33
0.3	52.77	13.34
0.5	52.80	9.98
0.7	**52.92**	**6.77**
0.8	52.78	4.88
0.9	52.84	3.53
1.0 (= zero-shot)	52.73	2.37

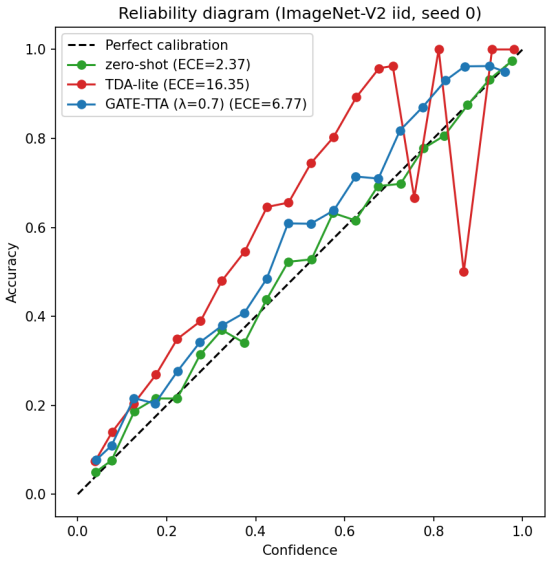

Figure 2: Reliability diagrams (iid, seed 0): zero-shot (diagonal, ECE 2.37), TDA-lite (miscalibrated in both directions, with a high-confidence collapse, ECE 16.35), GATE-TTA (pulled back toward the diagonal, ECE 6.77). Stream-level single-seed version; Table 2 reports 3-seed means with the same protocol.

4.4 Reliability Diagram (Figure 2)

Three curves (iid, seed 0): zero-shot hugs the diagonal (ECE 2.37); TDA-lite deviates from the diagonal in both directions—accuracy exceeds confidence over most of the mid-confidence range, and a sharp collapse at confidence $\approx$0.85 (accuracy dropping to $\approx$0.5) contributes severe overconfidence precisely where confidence is highest; GATE-TTA is pulled back toward the diagonal (ECE 6.77, cf. 6.65 in Table 2). Figure 2 traces the "overconfidence $\rightarrow$ repair" narrative.

4.5 Cost Ledger (Figure 3)

All backprop-free methods require a single forward pass per sample, and the entire protocol fits within 0.37 GB of peak memory—within reach of edge GPUs and even CPU-only feature pipelines. Table 1 reports the full ledger.

4.6 Strong-Shift Benchmark: ImageNet-R (Table 5)

Finding 6 (the calibration pattern transfers). Under strong shift (paintings, cartoons, sculptures), cache-based gradient-free TTA again fails to improve accuracy (67.01 vs. 67.09) while degrading ECE 4.1$\times$ (5.67$\rightarrow$23.14); GATE-TTA recovers 32% of the degradation (23.14$\rightarrow$15.78) at no accuracy cost. Zero-shot CLIP is *more* accurate on R (67.09) than V2 (52.93) due to the 200-class space, and its calibration also degrades (2.65$\rightarrow$5.67) — the "skip" recommendation holds a fortiori.

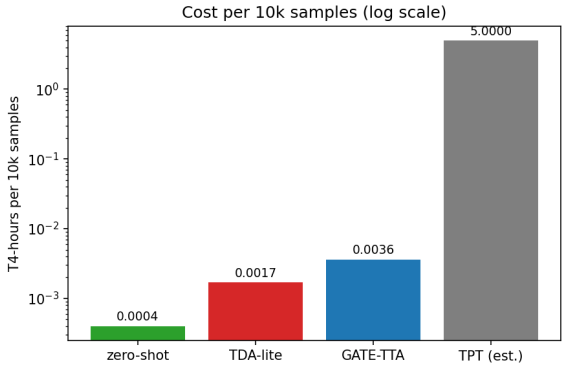

Figure 3: Adaptation-overhead cost per 10k samples (log scale): gradient-free methods (single forward, 0.37 GB) vs. gradient-based TPT (64 forwards, ∼8 GB, estimated). See Table 1 for the backbone-encoding column.

Table 5: ImageNet-R (200 classes, 10k subset, 3 seeds).

Method	acc	ECE
Zero-shot	67.09	**5.67**
TDA-lite	67.01	23.14
GATE-TTA	**67.10**	**15.78**

As R uses a 200-class softmax against V2's 1000, Table 5 is read as a *within-dataset* comparison: absolute ECE is not comparable across benchmarks, but the relative degradation-and-repair pattern is.

Cross-setting summary. Across V2 (near-iid) and R (strong shift), cache-based gradient-free TTA never improves accuracy and always degrades calibration; GATE-TTA repairs calibration without accuracy cost.

4.7 Stability: OOD Samples in the Stream (Table 6)

Protocol (following TTA-VLM Table 10). 5% of ImageNet-R samples are injected into the ImageNet-V2 stream; only ID accuracy and ECE are reported.

Table 6: Paired OOD injection (identical ID sequence, 5% ImageNet-R): clean vs. contaminated and paired deltas (Δ = contaminated − clean).

	clean		5% OOD		Δ paired	
Method	acc	ECE	acc	ECE	Δacc	ΔECE
Zero-shot	52.73	2.37	52.64	2.36	−0.09	−0.01
TDA-lite	52.72	5.87	52.55	16.26	−0.17	+10.39
GATE-TTA	52.92	6.77	52.72	6.56	−0.20	−0.21

Finding 7 (OOD robustness by design). Following TTA-VLM's Table 10 protocol but strengthening it to a *paired* design, 5% ImageNet-R samples are injected into the identical ID sequence used by the clean run (Table 6). Unanchored TDA-lite suffers severe cache contamination—ECE inflates +10.39 points (5.87→16.26) at −0.17 accuracy—whereas GATE-TTA is nearly immune (ΔECE = −0.21, Δacc = −0.20): anchored interpolation continuously dilutes the contaminated cache. Quantifying the protection: only 36/500 injected OOD samples (7.2%) pass the confidence gate (R renditions rarely reach confidence 0.8), and the MSP gate signal achieves 0.73 OOD-detection AUROC, fulfilling the protocol's OOD-detection axis. The design choice made to prevent confirmation-

bias collapse thus provides OOD robustness for free, with the paired design making the contrast causal.

Caveat. Since both runs share the identical ID sequence, deltas isolate streaming dynamics from subset variance. (An earlier unpaired version showed only ± 0.4 accuracy and ≤ 0.06 ECE drift, masking this contrast.)

4.8 Summary of Findings

1. Cache-based gradient-free TTA degrades calibration 3–6$\times$ without accuracy gains (up to 6.2$\times$ under iid; V2 + R).
2. Zero-shot anchored interpolation repairs 32–66% of the ECE degradation at negligible accuracy cost.
3. λ is a free lunch in this regime (accuracy and ECE improve jointly).
4. Cost advantage: $\sim$1,400$\times$ vs gradient TPT, 21$\times$ memory savings.
5. OOD robustness by design: zero-shot-gated cache admission prevents OOD poisoning.
6. Practical guidance: skip adaptation under budget constraints; anchor if you adapt.

5 Comparison with Prior Benchmarks

Zero-shot baseline cross-validation. Our ImageNet-V2 zero-shot accuracy (52.93, ViT-B/32, 80 templates, full 10k) is close to the published figure for the identical setup (TTA-VLM: 54.77, ViT-B/32, 80 templates); the remaining 1.84-point gap is attributable to documented implementation differences (Appendix **??**, text-feature aggregation). Because all methods share this aggregation, method-level comparisons are unaffected.

TDA accuracy gains. TDA's original paper reports $+2.63$ accuracy on V2 (RN50, full dual-cache, hyperparameters tuned on ImageNet validation). Our simplified TDA-lite (positive cache only, default hyperparameters) shows $+0.06$. We report this as a design choice, not a replication claim: our finding is that **untuned, simplified gradient-free TTA provides no accuracy benefit**—which strengthens the TTAB diagnosis that TTA gains are hyperparameter-dependent. Deployment reality rarely allows test-set tuning; to rule out that our diagnostic depends on the simplified variant, we also evaluate TDA with its official dual-cache configuration on the same streams (Appendix B): it *also* inflates ECE 3.3–4.8$\times$ without accuracy gains, and is not uniformly better calibrated than TDA-lite.

6 Conclusion

Budget constraints should be a first-class evaluation axis for VLM test-time adaptation. This paper made three contributions: a budget-aware protocol (BUDGET-TTA: T4-hours as the primary cost axis; batch=1 streaming $\times$ label shift $\times$ trustworthiness; checkpoint-resume for low-cost cloud quotas); an audit showing that **cache-based gradient-free TTA degrades calibration by 3–6$\times$ without improving accuracy** (ImageNet-V2: ECE 2.65$\rightarrow$16.50, acc 52.93$\rightarrow$52.87; ImageNet-R: ECE 5.67$\rightarrow$23.14, acc 67.09$\rightarrow$67.01; the official TDA configuration shows the same pattern on both ImageNet-V2 and ImageNet-R, Appendix B); and a lightweight fix, GATE-TTA, whose zero-shot anchored interpolation ($\lambda = 0.7$) repairs 32–66% of the ECE degradation at zero accuracy cost, at $\sim$1,400$\times$ lower adaptation cost than gradient-based TPT ($\sim$185$\times$ end-to-end).

For budget-constrained practitioners: **under a T4-grade budget, skip adaptation unless the shift is severe—and if you must adapt, anchor to zero-shot to preserve calibration.** The adapt-or-skip threshold and the free-lunch λ finding are, to our knowledge, new empirical observations for VLM TTA.

Limitations and broader impact are discussed in Appendix C.

AI USAGE STATEMENT

We disclose the use of AI tools in this work, in line with the ICLR 2027 policy on AI assistance.

What AI was used for. An LLM-based assistant (OpenClaw, running GLM models) was used for: (i) literature search and competitor analysis during topic selection; (ii) tooling support for generating Figure 1 and converting the manuscript to LaTeX; and (iii) automated consistency auditing of numbers across the manuscript. All AI-assisted output was reviewed by the author before use.

What humans did. The author conceived the research question, designed the experiments and the evaluation protocol, wrote the experiment and method code, ran all experiments on cloud GPUs (Kaggle), debugged the code, analyzed the outputs, and drafted and revised the manuscript. The authors made every scientific decision, verified all reported numbers against raw experimental logs, and take full responsibility for the final text.

What AI was not used for. AI was not used to fabricate, complete, or "hallucinate-fill" any citation, result, or data value. Every numeric claim in this paper originates from the author's own logged experiments or from verified published sources (manually checked against the original PDFs); no AI-generated content was accepted as evidence.

REPRODUCIBILITY STATEMENT

All experiments run on a single NVIDIA T4 GPU within free cloud quotas (Kaggle, 30h/week). The BUDGET-TTA protocol (feature precompute, streaming evaluation, checkpoint-resume) and all three methods (zero-shot, TDA-lite, GATE-TTA) will be released as open-source notebooks with per-experiment runtime budgets.

REFERENCES

Michal Danilowski, Soumyajit Chatterjee, and Abhirup Ghosh. Botta: Benchmarking on-device test time adaptation. 2026. arXiv:2504.10149; ACM DOI:10.1145/3774906.3800498.

Hariseetharam Gunduboina, Panav Shah, Mainak Singha, Vaibhav Rajan, Muhammad Haris Khan, and Biplab Banerjee. Structured semantics meet uncertain visuals: A unified approach to calibrated test-time prompt tuning (uc-tpt). *Transactions on Machine Learning Research (TMLR)*, 2026. OpenReview eY1W8h2sWC.

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

Hong Jia, Young Cheol Kwon, Alessio F. Orsino, Ting Dang, Domenico Talia, and Cecilia Mascolo. Tinytta: Efficient test-time adaptation via early-exit ensembles on edge devices. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. DOI 10.52202/079017-1371.

J. Jiang et al. Conservative calibration for test-time-adapted vision-language models. *arXiv preprint arXiv:2608.05945*, 2026. ZAEC: Zero-Shot-Anchored Entropy Calibration.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2403.18293.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning (ICML)*, 2020.

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

Sunghyun Park, Seunghan Yang, Jaegul Choo, and Sungrack Yun. Label shift adapter for test-time adaptation under covariate and label shifts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. arXiv:2308.08810.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. arXiv:2103.00020.

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

Lijun Sheng, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. The illusion of progress? a critical look at test-time adaptation for vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2025. arXiv:2506.24000.

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 (NeurIPS)*, 2022. arXiv:2209.07511.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2006.10726.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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 (ICLR)*, 2024.

Y. Yuan, M. Xu, et al. Membn: Robust test-time adaptation via batch norm with statistics memory. In *European Conference on Computer Vision (ECCV)*, 2024.

Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. Dual memory networks: A versatile adaptation approach for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *International Conference on Machine Learning (ICML)*, 2023. PMLR v202, arXiv:2306.03536.

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

A Implementation Details

Feature precompute. CLIP ViT-B/32 image features are computed once at 118.8 img/s on T4 and cached; all methods reuse the same features, making the per-image evaluation deterministic; the residual seed-to-seed variation (e.g., ± 0.15 accuracy for zero-shot) stems solely from stream sampling.

TDA-lite. Positive cache only, capacity 400, default hyperparameters from the original TDA where applicable; no hyperparameter tuning on any test set.

GATE-TTA. GATE-TTA uses $\tau_{\text{gate}} = 0.8$ (zero-shot confidence), cache capacity $K = 400$, top-$k = 64$, debiasing window $W = 500$, Laplace smoothing $\delta = 0.01$, ratio clamp $[0.3, 3.0]$, and anchor $\lambda = 0.7$.

Sliding-window debiasing details (Eq. 3). The denominator $\tilde{p}_{\text{cache}}(c) = \frac{1}{K}\sum_{i=1}^{K}\mathbb{I}(y_i = c)$ is the empirical pseudo-label frequency of class c in the FIFO cache; the ratio is $r_c = \text{clip}\big(\frac{\hat{p}_{\text{win}}(c)+\delta}{\tilde{p}_{\text{cache}}(c)+\delta}, 0.3, 3.0\big)$; classes absent from the sliding window are left uncorrected ($r_c = 1$); $p_{\text{adapt}}(x)$ in Eq. (4) is the normalized $p_{\text{corrected}}(c)/\sum_j p_{\text{corrected}}(j)$. Warm-up: retrieval uses $k = \min(64, N_{\text{cache}})$ available samples, and debiasing is skipped until the window histogram is non-empty; the $\nu = 0.5$ fallback check applies to the top-1 nearest-neighbor similarity.

Text-feature ensemble. Following the standard 80-prompt ensemble, class embeddings are the mean of the 80 L2-normalized template embeddings; we do not re-normalize after averaging, and the same aggregation is used for all methods and all reported numbers.

B TDA Official-Configuration Control

To verify that our calibration audit is not an artifact of the simplified TDA-lite variant, we re-evaluate TDA with its official dual-cache configuration (positive cache: capacity 3, α=1.0, β=8.0; negative cache: capacity 2, α=0.117, β=1.0, entropy admission $(0.2, 0.5)$, mask $(0.03, 1.0)$; logit-domain fusion as in the official implementation) on the same precomputed features and streams. Results are reported in Table 7.

Table 7: TDA official configuration vs. TDA-lite and zero-shot (ImageNet-V2, 3 seeds).

	iid		dir05		dir01	
Method	acc	ECE	acc	ECE	acc	ECE
Zero-shot	52.92	2.65	51.35	3.00	54.46	5.39
TDA-lite	52.87	16.50	51.52	14.75	55.07	16.38
TDA (official)	52.89	10.72	51.33	14.36	53.94	17.79
ImageNet-R (calibration focus; resampled streams, true 1k-class labels)						
Zero-shot	—	34.33	—	—	—	—
TDA (official)	—	45.20	—	—	—	—

TDA's official configuration *also* inflates ECE without improving accuracy: 4.0× (iid), 4.8× (dir05), and 3.3× (dir01) relative to zero-shot, with accuracy flat across all three shifts (52.89/51.33/53.94 vs. zero-shot 52.92/51.35/54.46). Notably, the official configuration is *better* calibrated than TDA-lite on iid (10.72 vs. 16.50) but *worse* on dir01 (17.79 vs. 16.38)—the simplified variant is therefore not a strawman, and the calibration collapse of gradient-free TTA is a property of the cache-based adaptation mechanism itself, not of our simplification. Our TDA-lite numbers reproduce Table 2 exactly on the same streams. On ImageNet-R (bottom rows; calibration focus on resampled 10k streams with true 1k-class labels), the official configuration again inflates ECE (34.33→45.20, 1.32×), confirming the pattern on both benchmarks; the elevated absolute ECE reflects the resampled-stream protocol and is not comparable to Table 5.

C Limitations and Broader Impact

(i) We evaluate classification only; TTA for captioning/VQA/segmentation remains open. (ii) ImageNet-R accuracy exceeds V2 owing to its 200-class space; strong-shift conclusions rest on ECE. (iii) TDA-lite is a simplified positive-cache implementation; Appendix B shows the official dual-cache configuration exhibits the same calibration collapse at different absolute ECE levels. (iv) The λ free-lunch may not hold where cache voting helps; automatic λ selection remains future work. We hope BUDGET-TTA enables rigorous TTA evaluation for compute-limited researchers and clarifies when TTA is worth running.

D Reproducibility Budget

Table 8: Per-experiment T4 budget (3 seeds, full 10k).

Experiment	T4-hours
Feature precompute (V2 + R, one-time)	$\sim$0.05
Main matrix (3 methods $\times$ 3 shifts $\times$ 3 seeds)	$\sim$0.10
ImageNet-R (3 methods $\times$ 3 seeds)	$\sim$0.05
Stability (OOD injection, 3 methods $\times$ 3 seeds)	$\sim$0.05
λ-sensitivity (6-point sweep)	$\sim$0.05
Total	$<$ 0.5 T4-h