# BEYOND ONLINE ACCURACY: WHAT PERSISTS AFTER CONTINUAL TEST-TIME ADAPTATION OF MULTIMODAL LLMS ON MULTIPLE-CHOICE VQA?

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Continual test-time adaptation (CTTA) lets multimodal large language models (MLLMs) update from unlabeled test streams, but it is usually evaluated only by online accuracy. This can confound the effect of adaptation with the inference procedure and does not reveal what the model retains after the stream. We introduce a two-stage benchmark for multiple-choice VQA that first measures online performance under paired presentation histories, then freezes each adapted model and evaluates it on shared held-out questions against a frozen model using the same inference procedure. Across ten published TTA and CTTA methods on Qwen2.5-VL, adapted models rarely outperform their matched frozen baselines, while temporary interface changes can leave persistent, order-dependent answer biases. Controlled updates further show that retained gains depend on the information available in the update target. Based on this finding, we propose Deliberate-then-Consolidate (DtC), which uses chain-of-thought (CoT) inference online and distills its CoT answer distribution into direct predictions. On Qwen2.5-VL-7B, DtC reaches the strongest frozen online inference and, on the benchmark split, improves clean held-out direct-answer accuracy by 2.1 points over frozen PA. Across four question splits, retained gains vary with the advantage of CoT over direct inference, a distinction that online accuracy alone does not reveal.

## 1 INTRODUCTION

Multimodal large language models (MLLMs) are deployed on streams of questions whose images and formats drift over time. Continual test-time adaptation (CTTA) offers a way to keep improving such a model without labels: the model answers each question and then updates itself from its own outputs (Wang et al., 2021; Niu et al., 2023; Wang et al., 2022). Recent work brings this idea to language and vision-language models through entropy minimization, input perplexity, augmentation consensus and test-time reinforcement learning (Chen et al., 2026a; Hu et al., 2025; Kaya et al., 2026; Singh et al., 2026). These methods are mostly compared by *online accuracy*, the accuracy of the answers given along the stream.

Relying on this criterion alone has two limitations. First, comparisons across methods mix the effect of the updates with the inference procedure. Permutation averaging (PA) averages answer probabilities across option orders; chain-of-thought (CoT) voting aggregates answers from sampled reasoning traces (Wei et al., 2022; Wang et al., 2023). Both can change accuracy without updating the model. Second, it does not show what the adapted model keeps. Multiple-choice questions make this sharper: each question comes with an answer interface (option order, option labels, instruction layout), and updates can absorb the interface itself. Figure 1 previews this distinction on Qwen2.5-VL-7B: Stage 1 measures performance on the adaptation stream, while Stage 2 asks what remains on shared held-out questions under matched inference.

To address both limitations, we build a benchmark that measures what persists (Section 3). The same four-option questions from eight datasets form four *paired histories* for Stage 1: *Standard* (S: standard interface, clean images), *Interface* (I: interface changes only), *Perceptual* (P: image corruptions only), and *Compound* (C: both). The questions and their order are identical across histories; only the presentation differs. In Stage 1 a model adapts on 768 questions and we record its



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.

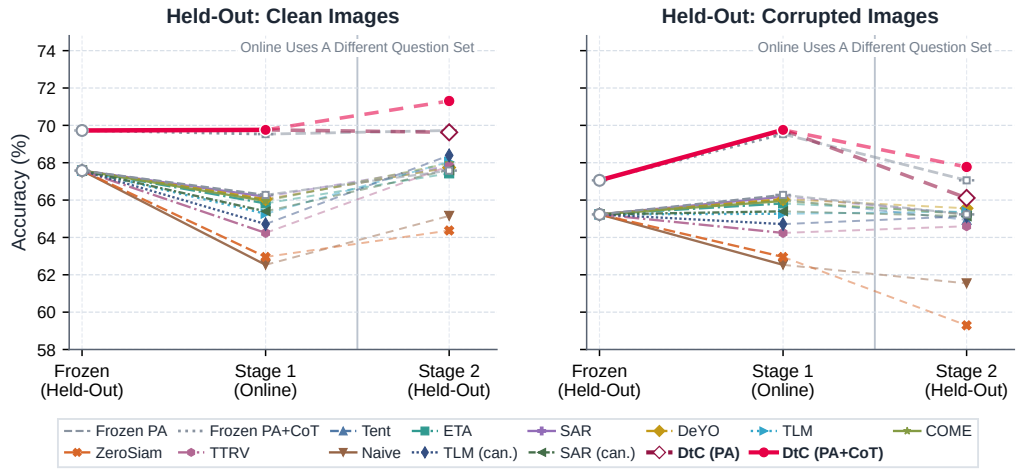



Figure 1: **Online accuracy does not show what persists after adaptation.** Qwen2.5-VL-7B results averaged over four histories and source orders O1/O2. Stage 1 reports accuracy on the adaptation stream; Stage 2 evaluates the resulting models without further updates on shared held-out questions with clean or corrupted images. Frozen points at left are matched held-out baselines; Stage 1 uses a different question set, so cross-stage slopes are visual guides. Published methods use PA at Stage 2; DtC uses PA+CoT online and both PA and PA+CoT after adaptation.

online answers. In Stage 2, adapted models stop updating and answer 768 shared held-out questions, excluded from their updates. Their accuracy difference from the unadapted (*frozen*) model under the same inference procedure is the *retained gain*. We report all inference and update costs.

Evaluating ten published adaptation methods (2021–2026) on two Qwen2.5-VL sizes in a fixed dataset-block order (O1) and its reverse (O2), we find:

- **Strong online performance does not mean a better model.** After adaptation, no published method is significantly more accurate than the frozen model with the same inference procedure, several are worse, and one collapses to a single option letter. Online, a frozen model with CoT voting is more accurate than every published method.
- **Presentation history leaves persistent, order-dependent effects.** Temporary interface changes can make an adapted model prefer one option letter on later standard questions: after the same history, the 7B model ends 3 points above the standard history in one source order and answers A to 91% of the held-out questions in the other.
- **Retained gains follow the information in the update target.** Training towards the model's direct predictions (without generated reasoning) leaves no significant gain; CoT targets leave more and true labels most, with gains varying across splits.

Building on these findings, we propose *Deliberate-then-Consolidate* (DtC; Section 5), which combines direct and CoT answers, then distills its own CoT answer distribution into direct predictions. On Qwen2.5-VL-7B, it achieves 69.8% online accuracy, close to frozen PA + CoT (69.5%) and 3.6–6.8 points above the eight published methods (a system-level comparison including CoT inference; Table 2). Canonicalization sorts options by text and fixes the template, removing the constructed interface differences before prediction and update. On the benchmark split, its direct answers generate no tokens and gain 2.1 points over frozen PA on clean held-out questions, approaching frozen PA + CoT accuracy at 0.4 rather than 7.1 seconds per question. Across four splits (O1), larger clean retained gains coincide with larger frozen-model CoT advantages over PA on the stream (mean retained gain $+0.6$; Table 15), consistent with Finding 3.

Our main contributions are as follows: (1) we identify two limitations of relying on online accuracy alone for CTTA and introduce a paired two-stage benchmark with matched frozen baselines, full costs and several splits of the questions (Section 3); (2) evaluating ten published methods, we show

Table 1: **Benchmark summary.** All questions have four options and one image; reference answers are used only for offline benchmark construction and evaluation and are never passed to the learner.

| | |
|---|---|
| Sources | MathVista, PhyX, AI2D, SEED-Bench-2-Plus, MMBench, PathMMU, A-OKVQA, MMMU (192 questions each) |
| Stage 1 | 768 questions, 96 per source in eight source blocks; order O1 and the reversed order O2 |
| Stage 2 | the other 768 questions, 96 per source; standard interface, clean or corrupted images |
| Interface changes | cyclic option permutation; numeric labels; moved instruction; skewed answer position |
| Image corruptions | Gaussian blur (radius 2.5 px); JPEG (quality 5); low light (brightness $\times 0.1$); grayscale |
| Assignment | 12 questions per source for each of the 16 interface-change $\times$ corruption pairs |
| Procedures | raw, canonical, PA, CoT voting, PA + CoT, each also for the frozen model |
| Reported | online accuracy; accuracy after adaptation; gain over the frozen model; updated parameters, update rate, tokens, time |

that adaptation rarely leaves a model better than frozen inference with the same procedure and that temporary presentation changes can leave persistent, order-dependent effects (Section 4); (3) we propose DtC, which turns the model's own reasoning into the most accurate online answers among the evaluated CTTA methods and, on the benchmark split, into a retained gain of its direct answers (Section 5).

## 2 Related work

**Test-time adaptation.** Tent adapts normalization layers by entropy minimization (Wang et al., 2021). Later methods select reliable samples and regularize updates (Niu et al., 2022; 2023; Lee et al., 2024), make entropy minimization conservative (Zhang et al., 2025), use teacher averaging and parameter restoration against error accumulation in continual streams (Wang et al., 2022; Yuan et al., 2023; Hoang et al., 2024), adjust learning rates or resets (Maharana et al., 2025; Lim et al., 2026), or prevent collapse by design (Chen et al., 2026a). Evaluation studies show that TTA results depend on hyperparameter selection and the adapted model (Zhao et al., 2023) and change when computation is accounted for (Alfarra et al., 2024), that adaptation can harm later predictions and collapse on long streams (Su et al., 2025; Hoang et al., 2024; Press et al., 2023), and that TTA methods for CLIP-style VLMs improve little over simple or early baselines (Farina et al., 2024; Sheng et al., 2025). We study what persists in MLLMs.

**Test-time learning for (M)LLMs.** TLM (Hu et al., 2025) minimizes the perplexity of test inputs with LoRA (Hu et al., 2022). TTRL and TTRV use agreement among sampled answers as a reward for reinforcement learning (Zuo et al., 2025; Singh et al., 2026), TTIQ rewards responses by their dependence on the image and the question instead (He et al., 2026), TTAdapt trains on the consensus of augmented views (Kaya et al., 2026), test-time consistency training aligns predictions across semantically equivalent inputs (Chou et al., 2026), SECL distills self-verification into the confidence of answers (Hedna et al., 2026), and VDA studies unsupervised continual post-training of MLLMs, noting that task-specific templates such as multiple-choice formats can cause cross-task interference (Li et al., 2026). Distilling CoT or other expensive procedures into direct answers is established offline, also without labels (Hinton et al., 2015; Yu et al., 2024; Huang et al., 2023), as is permutation debiasing for multiple-choice questions (Liusie et al., 2024; Zheng et al., 2024); D2I trains MLLMs with deliberate reasoning and lets them answer intuitively at test time (Yu et al., 2025). DtC combines these ingredients in the online, label-free setting.

**Evaluation of multiple-choice VQA.** Option order and format change accuracy and rankings (Pezeshkpour & Hruschka, 2024; Zheng et al., 2024; Liu et al., 2024; Sterz et al., 2025; Rosenthal et al., 2026). Gao et al. (2026) show that shared answer vocabularies inflate retention in supervised continual VQA, and Xu et al. (2026a) show that self-improvement gains of an LLM can vanish against a frozen control run through the same pipeline. We ask this for unlabeled test-time updates of MLLMs: which gains are real, and which persist. More related work: Appendix J.

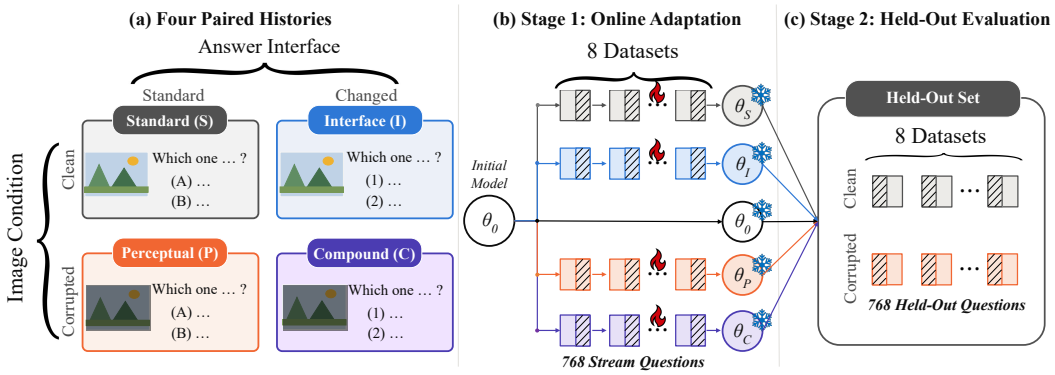



Figure 2: **Two-stage benchmark and protocol. (a) Four paired histories** form a $2 \times 2$ design over answer interface (standard/changed) and image condition (clean/corrupted): Standard (S), Interface (I), Perceptual (P), and Compound (C). **(b) Stage 1** runs each history from the same initial model $\theta_0$ on the same 768-question stream, recording each answer before updating; flames indicate updates and snowflakes indicate frozen terminal states. **(c) Stage 2** evaluates these states and the unadapted $\theta_0$, without further updates, on the same 768 held-out questions with clean and corrupted images under the same inference procedure.

## 3 A TWO-STAGE BENCHMARK FOR CTTA

**Questions and histories.** We sample 192 single-image, four-option questions from each of MathVista, PhyX, AI2D, SEED-Bench-2-Plus, MMBench, PathMMU, A-OKVQA and MMMU (Lu et al., 2024; Shen et al., 2025; Kembhavi et al., 2016; Li et al., 2024; Liu et al., 2024; Sun et al., 2024; Schwenk et al., 2022; Yue et al., 2024), 1,536 in total, and split each source evenly into a Stage-1 stream and a held-out set. Interface changes use four families: cyclic permutation of the options, numeric option labels, a moved instruction, and a skewed answer position (the correct option is moved to a favored letter for 75% of the questions). Image changes use four corruptions: Gaussian blur, JPEG compression, low light and grayscale. Each question receives one interface family and one corruption, balanced across sources (Table 1; Appendix E). These four Stage-1 histories form a $2 \times 2$ design crossing answer-interface changes with image corruptions (Figure 2a). All histories start from the same model and see the same questions in the same order, so any difference between them is caused by the presentation.

**Stage 1: online adaptation.** Each history is a stream of 768 questions in eight source blocks of 96 (order O1; the reversed block order O2 is used for robustness). For every question $x_t$ the method answers with its current state $s_{t-1}$, the answer is recorded, and only then may the method update:

$$\hat{y}_t = R(x_t; s_{t-1}), \qquad s_t = U(s_{t-1}, x_t). \tag{1}$$

No labels, source identities or later questions are available, and no state is reset at source boundaries. The state $s$ includes everything a method keeps, such as adapters, teachers or memory. Online accuracy is the accuracy of the recorded answers $\hat{y}_t$.

**Stage 2: what persists.** After Stage 1, each adapted model is frozen and answers the same 768 held-out questions, once with clean and once with corrupted images. The held-out questions use the standard interface, so the effect of an interface history is measured through the adapted models (I and C against S and P; Appendix H.1 evaluates changed interfaces). Because the questions are the same for all histories, differences between histories measure what each history left in the model. We report the *gain over the frozen model* under the same inference procedure $R$,

$$G_R(H) = \mathrm{Acc}_{\mathcal{Q}}(R; s_H) - \mathrm{Acc}_{\mathcal{Q}}(R; s_0), \tag{2}$$

where $s_H$ is the state after history $H$, $s_0$ the pretrained model and $\mathcal{Q}$ the held-out set. Online, we use the analogous gain over the frozen model on the stream.

**Inference procedures and cost.** We compare five inference procedures. The first three score option labels without token generation. *Raw*: score once as displayed. *Canonical*: sort options by text, fix the template and score once. *PA*: average option probabilities over four cyclic shifts of the canonical order, aligned by option content. *CoT voting*: sample four short rationales, one per shift, then take the majority of their scored answers, breaking ties by PA. *PA + CoT*: equally average PA with the mean answer distribution scored after the four rationales. We evaluate the frozen model with each procedure and report updated parameters, generated tokens and measured time per question (Appendix B).

**Statistics.** Histories and source orders are averaged within each question; 95% intervals are pointwise and come from 10,000 paired bootstrap resamples of questions, stratified by source (Appendix D adds a Holm correction (Holm, 1979)). The 1,536 core questions were used during development; separate panels of new questions are used only for reporting (Appendix F).

## 4 What persists after adaptation?

We evaluate at least one published method per year from 2021 to 2026: Tent (Wang et al., 2021), CoTTA (Wang et al., 2022), ETA (Niu et al., 2022), SAR (Niu et al., 2023), DeYO (Lee et al., 2024), TLM (Hu et al., 2025), COME (Zhang et al., 2025), ZeroSiam (Chen et al., 2026a), TTRV (Singh et al., 2026) and TTAdapt (Kaya et al., 2026), together with naive self-training on the model's own answers and PA self-distillation, which distills the frozen model's PA answer into the direct answer. The set spans entropy-based TTA (Tent, ETA, SAR, DeYO, COME, ZeroSiam), continual TTA with teacher averaging (CoTTA) and test-time learning for (M)LLMs (TLM, TTRV, TTAdapt). Each implementation keeps the method's update rule and loss hyperparameters (filters, thresholds and weights); Appendix A lists every change needed for four-option questions, such as entropy thresholds scaled by $\ln 4$, COME's prior set to the four options, and ZeroSiam's update scope, which follows its vision implementation. Methods that update normalization layers share one learning rate, chosen for Tent on development averages; TLM, TTRV and TTAdapt use the learning rates of their public code. Online, the published methods answer the displayed question with one forward pass, except CoTTA and TTAdapt, which average augmented views. All methods run under the same continual protocol: each answer is recorded before the update on that question, and no method restores its whole model between questions or datasets (CoTTA keeps its stochastic restoration of single weights). This changes two methods: TTAdapt originally adapts to each question before answering and then restores its weights, and SAR runs without its model reset (with the reset in Appendix I.2). ASR (Lim et al., 2026) is not included because its reset signal needs batches of questions (Appendix A). As controls for input preprocessing, we also run TLM and SAR with the canonical option order and template that DtC uses. Every method runs on both source orders (TTRV at 3B: O1 only); CoTTA and TTAdapt update all parameters and run on Qwen2.5-VL-3B only (Appendix C), so Table 2 has the other eight.

**Finding 1: strong online performance does not mean a better model.** Online, the frozen model gains 0.6 points from PA and 2.9 points from CoT voting over raw answers, and PA + CoT reaches 69.5%, more than every published method (at most 66.2%; Table 2 and Figure 3). After adaptation, with the same procedure, no published method is significantly more accurate than the frozen model on the held-out questions: the largest gains are $+0.5$ (TLM, clean) and $+0.3$ (DeYO, corrupted), while ZeroSiam loses 3.2 and 5.9 points. This holds with DtC's canonical inputs ($+0.8$ and $0.0$ for TLM and SAR), with other inference procedures (at most $+0.5$; Appendix B), at a tenth of the default learning rates (at most $+0.6$; ten times the default costs up to 43 points; Appendix A) and with a second seed (Appendix A). Only on held-out questions with the changed interface and corrupted images do SAR and DeYO gain 1.1 and 1.5 points (Appendix H.1). Online gains do not predict retained ones: SAR gains $+0.5$ online (significant) but $+0.3$ and $0.0$ after adaptation.

**Finding 2: presentation history leaves persistent, order-dependent effects.** Because the histories differ only in presentation, comparing I with S on the same held-out questions isolates what the interface changes left in the model. With naive self-training at learning rate $10^{-4}$, the 7B model ends the I history 3.3 points above the S history in order O1 ($[+1.0, +5.5]$) but 27.6 points below it in order O2 ($[-32.4, -22.8]$), where it answers A to 91% of the held-out questions although its online loss was only 7.3 points; the 3B model answers A to all of them (Appendix C). Interface

Table 2: **Two-stage results on Qwen2.5-VL-7B** (accuracy in %, averaged over four histories and O1/O2). Parentheses: gain over the frozen model with the same inference procedure; *: pointwise 95% interval excludes zero. Online uses each method's own procedure (published: raw; canonical-input controls: canonical; DtC: PA+CoT). After adaptation, adapted models use PA; DtC is also evaluated with PA+CoT. Tokens and seconds are per Stage-1 question. Bold: best learning method; last row: DtC minus the best published method. Details: Appendix B.

| | | | | | | After adaptation | |
|---|---|---|---|---|---|---|---|
| Method | Venue | Updated parameters | Tokens | Time(s) | Online | Clean | Corrupted |
| Frozen, raw | – | – | 0 | 0.1 | 65.7 | 66.7 | 64.3 |
| Frozen, canonical | – | – | 0 | 0.1 | 65.6 | 68.8 | 64.7 |
| Frozen, PA, displayed order | – | – | 0 | 0.4 | 66.5 | 67.7 | 66.3 |
| Frozen, PA | – | – | 0 | 0.4 | 66.3 | 67.6 | 65.2 |
| Frozen, CoT voting | – | – | 375 | 6.6 | 68.6 | 69.4 | 66.3 |
| Frozen, PA + CoT | – | – | 375 | 7.1 | 69.5 | 69.7 | 67.1 |
| Naive self-training | – | LoRA $q, v$ (2.5M) | 0 | 0.7 | 62.5 ($-3.2^*$) | 65.2 ($-2.4^*$) | 61.5 ($-3.7^*$) |
| Tent | ICLR'21 | norm (0.29M) | 0 | 0.6 | 66.0 (+0.3) | 67.8 (+0.2) | 64.9 ($-0.3$) |
| ETA | ICML'22 | norm (0.29M) | 0 | 0.3 | 65.8 (+0.1) | 67.4 ($-0.1$) | 65.3 (+0.1) |
| SAR | ICLR'23 | norm (0.29M) | 0 | 1.0 | 66.2 (+0.5*) | 67.8 (+0.3) | 65.2 (0.0) |
| DeYO | ICLR'24 | norm (0.29M) | 0 | 0.4 | 66.0 (+0.4) | 67.7 (+0.2) | 65.6 (+0.3) |
| TLM | ICML'25 | LoRA $q, v$ (2.5M) | 0 | 0.3 | 65.3 ($-0.4$) | 68.1 (+0.5) | 65.4 (+0.1) |
| COME | ICLR'25 | norm (0.29M) | 0 | 0.6 | 66.0 (+0.3) | 68.0 (+0.4) | 65.2 (0.0) |
| ZeroSiam | ICLR'26 | norm + predictor (13.1M) | 0 | 0.6 | 63.0 ($-2.7^*$) | 64.4 ($-3.2^*$) | 59.3 ($-5.9^*$) |
| TTRV | CVPR'26 | all weights | 100 | 9.0 | 64.2 ($-1.6^*$) | 67.8 (+0.2) | 64.6 ($-0.6$) |
| TLM, canonical inputs | control | LoRA $q, v$ (2.5M) | 0 | 0.3 | 64.7 ($-0.8$) | 68.4 (+0.8) | 65.1 ($-0.1$) |
| SAR, canonical inputs | control | norm (0.29M) | 0 | 1.0 | 65.4 ($-0.2$) | 67.6 (0.0) | 65.1 ($-0.1$) |
| DtC | ours | LoRA $q, v$ (2.5M) | 370 | 12.6 | **69.8 (+0.2)** | **69.6 (+2.1*)** | **66.1 (+0.9)** |
| DtC, answer with CoT | ours | LoRA $q, v$ (2.5M) | 370 | 12.6 | – | **71.3 (+1.6*)** | **67.8 (+0.7)** |
| vs. best published method | | | | | +3.6 | +1.6 | +0.6 |

changes thus leave substantial effects whose direction depends on the stream. InternVL3.5 (Wang et al., 2025) shows the same effect (Appendix H; Appendices H.2 and H.4 add a third model family and new questions). TTAdapt on the 3B model, run without its per-question reset, collapses to a single letter in every history (Appendix C). Sorting the options and fixing the template before every answer and update removes the constructed interface differences, which makes the canonical option order a sensible default for CTTA on multiple-choice questions.

**Finding 3: retained gains follow the information in the update target.** Why does no published method consistently leave a better model? Their update targets come from the model's own answers or inputs: entropy, pseudo-labels, augmentation or sampling consensus, input perplexity. Table 3 changes only the target of one LoRA update per question. Training the direct answer towards the model's own PA distribution retains $+0.7$ points ($[-0.6, +2.0]$): PA already knows this. Training it towards the CoT answer distribution of the frozen model retains $+2.2$ points ($[+0.7, +3.8]$), consistent with CoT using computation that the direct answer does not. A 32B teacher whose answers are 2.7 points more accurate than CoT retains a similar gain ($+2.1$ on clean and $+2.5$ on corrupted questions), and training towards the true answers with the same update, a supervised reference rather than a CTTA method, retains $+3.7$ ($[+1.9, +5.6]$). This also depends on the model: at 3B, CoT is only 0.9 points more accurate than PA on the stream, and distilling it does not improve held-out accuracy ($-1.2$, $[-2.6, +0.1]$), while the supervised reference retains $+2.3$. With the halves swapped, CoT is only 0.7 points more accurate than PA on the stream and its target retains $-0.4$; over both splits, the order holds ($+0.2$, $+0.9$, $+2.2$; Appendix D).

## 5 DELIBERATE-THEN-CONSOLIDATE

Finding 3 suggests a direct design: an MLLM can produce an informative target itself by reasoning before it answers. DtC uses this target both to answer and to learn (Figure 4).

Table 3: **What the update target adds and what persists** (controlled study with fixed targets from the frozen model or a 32B model; one LoRA $q, v$ step per question, learning rate $10^{-4}$, O1, answers with PA). Target gain: accuracy of the target's top answer on the Stage-1 stream minus that of the frozen model with PA. Retained gain: held-out accuracy after adaptation minus the frozen model's, with 95% intervals. The last row of each model is a supervised reference with the same update: it trains towards the true answers and uses labels.

| Model | Update target | Target gain | Retained, clean | Retained, corrupted |
|---|---|---|---|---|
| 7B | frozen-model PA answer | 0.0 | $+0.7$ $[-0.6, +2.0]$ | 0.0 $[-1.3, +1.3]$ |
| 7B | CoT answer of the frozen model | $+2.2$ | $+2.2$ $[+0.7, +3.8]$ | $+1.6$ $[-0.2, +3.3]$ |
| 7B | PA answer of a 32B model | $+4.9$ | $+2.1$ $[+0.5, +3.9]$ | $+2.5$ $[+0.6, +4.4]$ |
| 7B | true answer (oracle; uses labels) | $+33.7$ | $+3.7$ $[+1.9, +5.6]$ | $+3.2$ $[+1.2, +5.2]$ |
| 3B | frozen-model PA answer | 0.0 | $-0.3$ $[-1.6, +1.0]$ | $-0.3$ $[-1.8, +1.2]$ |
| 3B | CoT answer of the frozen model | $+0.9$ | $-1.2$ $[-2.6, +0.1]$ | $-0.7$ $[-2.2, +0.8]$ |
| 3B | PA answer of a 32B model | $+7.1$ | $+0.9$ $[-0.6, +2.4]$ | $+0.4$ $[-1.3, +2.1]$ |
| 3B | true answer (oracle; uses labels) | $+35.9$ | $+2.3$ $[+0.5, +4.2]$ | $+2.6$ $[+0.5, +4.7]$ |

**Method.** For question $x_t$ with options in canonical order, let $p_t$ be the direct answer distribution of the current model averaged over the four cyclic option orders $g \in \mathcal{G}$, and let $r_t$ be the average answer distribution after four sampled CoT rationales, one per option order, mapped back to the option contents. The model answers $\hat{y}_t = \arg\max_c \frac{1}{2} p_t(c) + \frac{1}{2} r_t(c)$, the answer is recorded, and one AdamW step on LoRA $q, v$ adapters (rank 8, learning rate $10^{-4}$) minimizes

$$\mathcal{L}_t = -\frac{1}{4} \sum_{g \in \mathcal{G}} \sum_{c=1}^{4} r_t(c) \, \log p_\theta\big(g(c) \mid g\, x_t\big), \tag{3}$$

where $g(c)$ is the position of option $c$ under order $g$. Both the rationales and the target come from the current model; nothing is reset, filtered or replayed. Because answers and updates use the canonical option order and template, the interface changes of this benchmark are removed before the model sees the question, and the I and C histories give exactly the inputs of S and P (a preference for positions within the canonical order can still be learned).

**Results.** DtC is the most accurate learning method in all three accuracy columns of Table 2 (Figure 5). Online, it answers 69.8% of the stream questions correctly, 3.6 to 6.8 points more than every

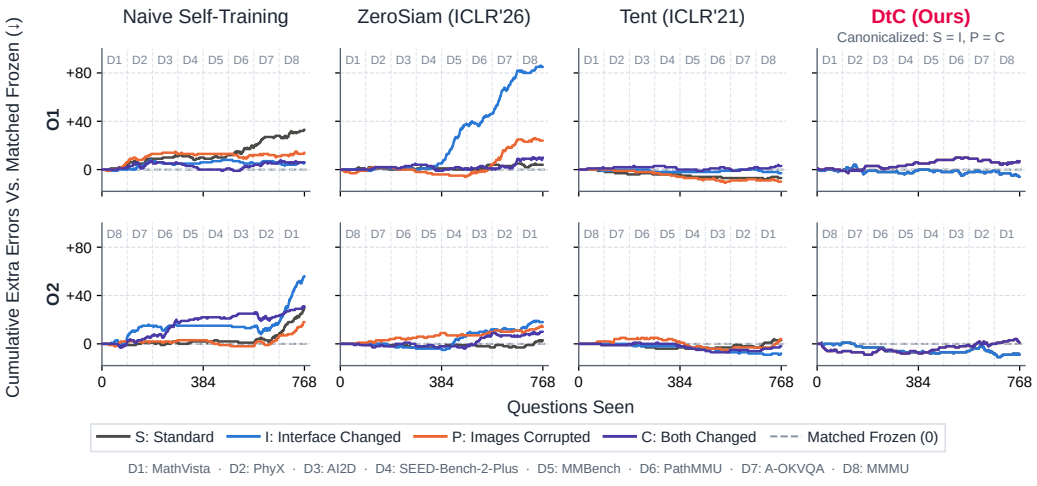



Figure 3: **Where errors accumulate during Stage 1** (Qwen2.5-VL-7B). Cumulative extra errors relative to the matched frozen model across four histories and source orders O1/O2; lower is better. D1–D8 denote dataset blocks. For DtC, canonicalization gives S=I and P=C.

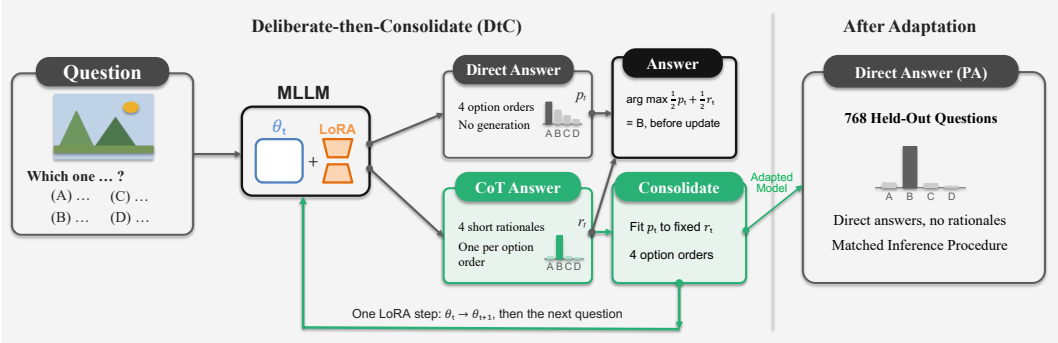



Figure 4: **DtC.** For each question, the current model computes a direct PA distribution $p_t$ over four option orders and a CoT distribution $r_t$ from four short rationales, one per option order. It records the PA+CoT answer before updating, then takes one LoRA step that fits $p_t$ to the fixed target $r_t$. After adaptation, the model answers held-out questions directly with PA, without generating rationales.

published method (significant after a Holm correction; a system-level comparison including the cost of CoT), and it is comparable to the strongest frozen policy, PA + CoT ($+0.2$ points, $[-1.1, +1.5]$), at a similar generation budget (370 versus 375 tokens per question). After adaptation, its direct answers generate no tokens and are 2.1 points more accurate than those of the frozen model with PA on clean held-out questions ($[+0.3, +3.8]$), 1.6 to 5.3 points above the published methods, and comparable to frozen PA + CoT (69.6% versus 69.7%). With CoT, the adapted model is 1.6 points more accurate than frozen PA + CoT ($[+0.2, +3.0]$, mean of two sampling seeds), the highest in the table. A second training seed gives similar gains in both orders ($+1.8$ and $+1.7$; pooled: 1.9, $[+0.2, +3.6]$). Over four splits of the core questions (order O1: the benchmark split, its role swap and two random re-splits fixed in advance; Table 15), the observed pattern is consistent with Finding 3: DtC retains $+2.3$ and $+1.7$ where frozen-model CoT is 2.2 and 1.6 points more accurate than PA on the stream, and $-1.0$ and $-0.5$ where it is only 0.7 and 0.8 points more accurate (mean $+0.6$); online, it stays at the level of frozen PA + CoT on every split. The extra computation is paid during adaptation (median 12.6 versus 7.1 s per question for frozen PA + CoT); afterwards DtC answers directly in 0.4 s.

**Ablations.** Table 4 varies one part of DtC at a time. Compared on the same questions, distilling the current model's PA answer instead of its CoT answer is 27.1 points worse on clean held-out questions ($[23.7, 30.5]$): its direct answers become nearly uniform, although its recorded online answers, which still use CoT, remain at 68.0%. A stronger inference procedure can thus hide a degraded direct predictor. Distilling into all four option orders is 1.6 points better than into one ($[0.3, 2.9]$), a target mixing PA and CoT performs similarly, and training on the rationale tokens as well (Zelikman et al., 2022) lengthens the rationales until they are truncated and collapse. Stronger updates, replay or an extra corrupted view do not improve on the fixed CoT target (Appendix D).

Table 4: **Ablations of DtC** (7B, O1; gains over the frozen model with PA). Each variant changes one part of DtC. Online: PA-only answers; Table 2 uses PA + CoT for DtC. $^*$: 95% interval excludes zero.

| Variant | Online | Clean | Corrupted |
|---|---|---|---|
| DtC: CoT target, four student option orders | 0.0 | $+2.3^*$ | $+0.8$ |
| current-model PA target instead | $-16.8^*$ | $-24.8^*$ | $-24.5^*$ |
| one student option order instead of four | $+0.5$ | $+0.7$ | $-0.1$ |
| target: recorded answer ($\frac{1}{2}$PA + $\frac{1}{2}$CoT) | $+0.4$ | $+2.0$ | $+1.4$ |
| also training on the rationale tokens | $-11.1^*$ | $-36.4^*$ | $-34.4^*$ |
| second training seed | $+0.3$ | $+1.8$ | $+0.5$ |

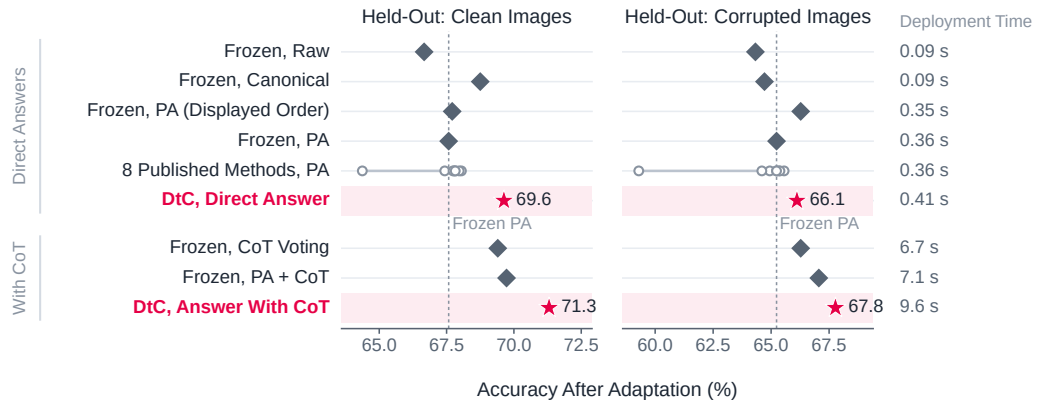



Figure 5: **Accuracy after adaptation and deployment time** (Qwen2.5-VL-7B, mean over four histories and source orders O1/O2). Diamonds show frozen inference procedures, open circles the eight published methods evaluated with PA, and stars DtC; dashed lines mark frozen PA. The upper block uses direct answers, while the lower block uses CoT at deployment. Times at right are median seconds per held-out question on one H200 after adaptation and exclude the adaptation cost.

## 6 Discussion and Limitations

Matched online gains measure updates on the seen stream, not what the adapted model retains; matched frozen baselines, held-out evaluation and full costs can change the conclusion. DtC further shows that retained gains can vary across splits even when online performance remains similar. We therefore recommend matched frozen baselines, held-out evaluation over several splits, paired presentation histories, canonical option order, and full compute reporting.

Our study has limits. It covers one task format, two Qwen2.5-VL sizes for the full comparison and two further MLLMs for the history effects; the image corruptions and interface changes are synthetic and chosen to be clearly visible; the 1,536 core questions were used during development; and the published methods are re-implemented for a new task and for our continual protocol (for example, TTAdapt originally resets after every question) that may not reflect their best settings, although we keep each method's own update rule and list every change. The retained gains of DtC depend on the split of the questions (direct answers: mean $+0.6$ over four splits; with CoT: positive on the benchmark split, not on its role swap), are largest for the inference procedure it distills into, are not significant after a Holm correction over our pre-specified comparisons, do not appear at 3B or on InternVL3.5 with the settings of the 7B model, and were not confirmed on a small panel of new questions (Appendix D). CoTTA and TTAdapt were run at 3B only, the InternVL3.5 check covers a subset of the methods with settings transferred from Qwen, and ASR was not included. We will release the benchmark, the protocol and all per-question predictions to test these limits.

## 7 Conclusion

Online gains and retained gains are different quantities. Our two-stage benchmark separates them, showing that adapted models rarely outperform matched frozen inference and that temporary presentation changes can leave persistent, order-dependent effects. On Qwen2.5-VL-7B, DtC achieves the highest online accuracy among evaluated CTTA methods, while its retained gains vary across splits with the frozen-model CoT advantage. Together, these results show why CTTA should be judged not only by online accuracy, but also by what persists after adaptation.

## AI USE STATEMENT

Generative AI assistants supported research ideation, literature search, experiment design, code implementation and debugging, statistical analysis, figure design and plotting code, and drafting and editing of this manuscript. All numbers are computed from recorded model predictions by documented scripts. Questions and reference answers come from the source datasets and were not generated or relabeled by an AI assistant. The authors reviewed the manuscript, the citations and the empirical claims, and take full responsibility for the content.

## ETHICS STATEMENT

The study reuses existing benchmark questions and source labels. Pathology questions are not a validation of clinical decision support. Image corruptions may remove information needed to answer; results should not be read as statements about human performance. Any released data must follow the licenses of the original sources.

## REPRODUCIBILITY STATEMENT

Section 3 and Appendices A, E and G specify the benchmark, protocol, update scopes and evaluation procedures. We record source identifiers, content hashes, configurations, predictions, model states and random-number states. In our checks, resumed runs reproduced uninterrupted runs exactly, and saved states reproduced their original answers when evaluated with the same precision and forward path.

The supplementary code provides a condensed implementation of the components that determine the experiments, including benchmark construction, the two-stage protocol, DtC, published-method adaptations, statistics and configurations. Routine data loading, paths, logging, job scripts and plotting are omitted. TTRV additionally uses its public training code with our continual-protocol hooks. Complete code and per-question predictions will be released with the paper. Source images, question text and model weights are not redistributed.

## REFERENCES

Motasem Alfarra, Hani Itani, Alejandro Pardo, Shyma Yaser Alhuwaider, Merey Ramazanova, Juan Camilo Perez, Zhipeng Cai, Matthias Müller, and Bernard Ghanem. Evaluation of test-time adaptation under computational time constraints. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 976–991, 2024. URL https://proceedings.mlr.press/v235/alfarra24a.html.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. arXiv:2502.13923, 2025. URL https://arxiv.org/abs/2502.13923.

Cheng Chen, Junchen Zhu, Xu Luo, Hengtao Shen, Lianli Gao, and Jingkuan Song. CoIN: A benchmark of continual instruction tuning for multimodel large language models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2403.08350.

Guohao Chen, Shuaicheng Niu, Deyu Chen, Jiahao Yang, Zitian Zhang, Mingkui Tan, Pengcheng Wu, and Zhiqi Shen. ZeroSiam: An efficient asymmetry for test-time entropy optimization without collapse. In *International Conference on Learning Representations*, 2026a.

Xiao Chen, Jiazhen Huang, Zhiming Liu, Qinting Jiang, Fanding Huang, Jingyan Jiang, and Zhi Wang. Test-time distillation for continual model adaptation. In *Findings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 7593–7604, 2026b.

Shih-Han Chou, Shivam Chandhok, James J. Little, and Leonid Sigal. Test-time consistency in vision language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7789–7798, 2026. URL https://arxiv.org/abs/2506.22395.

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, 2024. URL https://arxiv.org/abs/2405.18330.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.01412.

Zijian Gao, Zicheng Sun, Xingxing Zhang, Kele Xu, and Huaimin Wang. Re-evaluating Continual VQA: Toward Fair and Robust Evaluation for Multimodal Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 18021–18031, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Gao_Re-evaluating_Continual_VQA_Toward_Fair_and_Robust_Evaluation_for_Multimodal_CVPR_2026_paper.html.

John Gkountouras, Josip Jukić, and Ivan Titov. Consensus as privileged context for label-free self-distillation. arXiv:2607.13643, 2026. URL https://arxiv.org/abs/2607.13643.

Xinrui He, Ting-Wei Li, Junting Wang, Mengting Ai, Xinyu He, Hanghang Tong, and Jingrui He. Harnessing image question dependence for better VLM test-time reinforcement learning. arXiv:2609.13296, 2026. URL https://arxiv.org/abs/2609.13296.

Mohamed Rissal Hedna, Jan Strich, Martin Semmann, and Chris Biemann. Self-calibrating language models via test-time discriminative distillation. arXiv:2604.09624, 2026. URL https://arxiv.org/abs/2604.09624.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1903.12261.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015. URL https://arxiv.org/abs/1503.02531.

Trung-Hieu Hoang, Duc Minh Vo, and Minh N. Do. Persistent test-time adaptation in recurring testing scenarios. In *Advances in Neural Information Processing Systems*, 2024.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Jinwu Hu, Zitian Zhang, Guohao Chen, Xutao Wen, Chao Shuai, Wei Luo, Bin Xiao, Yuanqing Li, and Mingkui Tan. Test-time learning for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24823–24849, 2025.

Audrey Huang, Adam Block, Dylan J. Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan T. Ash, and Akshay Krishnamurthy. Self-improvement in language models: The sharpening mechanism. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2412.01951.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Mehmet Onurcan Kaya, Desmond Elliott, and Dim P. Papadopoulos. Efficient Test-Time Scaling for Small Vision-Language Models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ec3d49763c653ad7c8d587f52220c129-Abstract-Conference.html.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A Diagram Is Worth A Dozen Images. In *European Conference on Computer Vision*, pp. 235–251. Springer, 2016.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023. URL https://arxiv.org/abs/2309.06180.

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

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. SEED-Bench-2-Plus: Benchmarking Multimodal Large Language Models with Text-Rich Visual Comprehension. arXiv:2404.16790, 2024. URL https://arxiv.org/abs/2404.16790.

Kaichen Li, Zhilin Zhu, Jianhao Huang, Zhengqin Lai, Baochen Xiong, Zibo Shao, Yaguang Song, Linhui Xiao, Xiaoshan Yang, and Changsheng Xu. A Visual Dependence-Aware Framework for Multimodal Unsupervised Continual Post-Training. arXiv:2608.26095, 2026. URL https://arxiv.org/abs/2608.26095.

Taejun Lim, Joong-Won Hwang, and Kibok Lee. When and where to reset matters for long-term test-time adaptation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/79bdd6fe3f012befcc459ad13de65d13-Abstract-Conference.html.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Adian Liusie, Yassir Fathullah, and Mark J. F. Gales. Teacher-student training for debiasing: General permutation debiasing for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1376–1387, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

Sarthak Kumar Maharana, Baoming Zhang, and Yunhui Guo. PALM: Pushing adaptive learning rate mechanisms for continual test-time adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18):19378–19386, 2025. doi: 10.1609/aaai.v39i18.34133. URL https://ojs.aaai.org/index.php/AAAI/article/view/34133.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. SmolVLM: Redefining small and efficient multimodal models. In *Conference on Language Modeling*, 2025.

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

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. URL https://openreview.net/forum?id=g2YraF75Tj.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024. URL https://aclanthology.org/2024.findings-naacl.130/.

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. URL https://arxiv.org/abs/2306.05401.

Nikita Rajaneesh, Thomas Zollo, and Richard Zemel. Test-time warmup for multimodal large language models. arXiv:2509.10641, 2025. URL https://arxiv.org/abs/2509.10641.

Fabio Rosenthal, Sebastian Schmidt, Thorsten Graf, Thorsten Bagdonat, Stephan Günnemann, and Leo Schwinn. Unexplored flaws in multiple-choice VQA make benchmarking unreliable. In *Findings of the Association for Computational Linguistics: EMNLP 2026*, 2026. URL https://arxiv.org/abs/2511.22341.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge. In *European Conference on Computer Vision*, pp. 146–162. Springer, 2022.

Sheikh Shafayat, Fahim Tajwar, Ruslan Salakhutdinov, Jeff Schneider, and Andrea Zanette. Can large reasoning models self-train? arXiv:2505.21444, 2025. URL https://arxiv.org/abs/2505.21444.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv:2402.03300, 2024. URL https://arxiv.org/abs/2402.03300.

Hui Shen, Taiqiang Wu, Qi Han, Yunta Hsieh, Jizhou Wang, Yuyue Zhang, Yuxin Cheng, Zijian Hao, Yuansheng Ni, Xin Wang, Zhongwei Wan, Kai Zhang, Wendong Xu, Jing Xiong, Ping Luo, Wenhu Chen, Chaofan Tao, Zhuoqing Mao, and Ngai Wong. PhyX: Does Your Model Have the "Wits" for Physical Reasoning? arXiv:2505.15929, 2025. URL https://arxiv.org/abs/2505.15929.

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, Datasets and Benchmarks Track*, 2025. URL https://arxiv.org/abs/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*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/5bf2b802e24106064dc547ae9283bb0c-Abstract-Conference.html.

Akshit Singh, Shyam Marjit, Wei Lin, Paul Gavrikov, Serena Yeung-Levy, Hilde Kuehne, Rogerio Feris, Sivan Doveh, James Glass, and M. Jehanzeb Mirza. TTRV: Test-Time Reinforcement Learning for Vision Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2026.

Hannah Sterz, Jonas Pfeiffer, and Ivan Vulić. DARE: Diverse Visual Question Answering with Robustness Evaluation. *Transactions of the Association for Computational Linguistics*, 13:1121–1145, 2025. URL https://aclanthology.org/2025.tacl-1.52/.

Yongyi Su, Yushu Li, Nanqing Liu, Kui Jia, Xulei Yang, Chuan-Sheng Foo, and Xun Xu. On the adversarial risk of test time adaptation: An investigation into realistic test-time data poisoning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bb309cc1fbdb88ea755bd7cee4b310ec-Abstract-Conference.html.

Yuxuan Sun, Hao Wu, Chenglu Zhu, Sunyi Zheng, Qizi Chen, Kai Zhang, Yunlong Zhang, Dan Wan, Xiaoxiao Lan, Mengyue Zheng, Jingxiong Li, Xinheng Lyu, Tao Lin, and Lin Yang. PathMMU: A Massive Multimodal Expert-Level Benchmark for Understanding and Reasoning in Pathology. In *European Conference on Computer Vision*, 2024.

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*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

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.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv:2508.18265, 2025. URL https://arxiv.org/abs/2508.18265.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

Zhiquan Wen, Shuaicheng Niu, Ge Li, Qingyao Wu, Mingkui Tan, and Qi Wu. Test-time model adaptation for visual question answering with debiased self-supervisions. *IEEE Transactions on Multimedia*, 26:2137–2147, 2024. URL https://ieeexplore.ieee.org/document/10173554.

Cheng Xu, Nan Yan, Liming Chen, and M-Tahar Kechadi. Phantom gains: Auditing self-improvement against a measured null. arXiv:2608.20290, 2026a. URL https://arxiv.org/abs/2608.20290.

Weijia Xu, Alessandro Sordoni, Chandan Singh, Zelalem Gero, Michel Galley, Xingdi Yuan, and Jianfeng Gao. Test-time learning with an evolving library. arXiv:2605.14477, 2026b. URL https://arxiv.org/abs/2605.14477.

Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. Distilling System 2 into System 1. arXiv:2407.06023, 2024. URL https://arxiv.org/abs/2407.06023.

Yahan Yu, Yuyang Dong, and Masafumi Oyamada. Learning deliberately, acting intuitively: Unlocking test-time reasoning in multimodal LLMs. arXiv:2507.06999, 2025. URL https://arxiv.org/abs/2507.06999.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust Test-Time Adaptation in Dynamic Scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR*, pp. 15922–15932, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Yuan_Robust_Test-Time_Adaptation_in_Dynamic_Scenarios_CVPR_2023_paper.html.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, June 2024.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, 2022.

Marvin Zhang, Sergey Levine, and Chelsea Finn. MEMO: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.nips.cc/paper/2022/hash/fc28053a08f59fccb48b11f2e31e81c7-Abstract-Conference.html.

Qingyang Zhang, Yatao Bian, Xinke Kong, Peilin Zhao, and Changqing Zhang. COME: Test-time adaption by conservatively minimizing entropy. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.10894.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42058–42080, 2023. URL https://proceedings.mlr.press/v202/zhao23d.html.

Hongbo Zhao, Fei Zhu, Haiyang Guo, Meng Wang, Rundong Wang, Gaofeng Meng, and Zhaoxiang Zhang. MLLM-CL: Continual learning for multimodal large language models. arXiv:2506.05453, 2025. URL https://arxiv.org/abs/2506.05453.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large Language Models Are Not Robust Multiple Choice Selectors. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=shr9PXz7T0.

Puzhuo Zheng and Hasan Kurban. It's the decoding format, not the perturbation: Auditing consistency-based selection for vision-language test-time scaling. arXiv:2608.01207, 2026. URL https://arxiv.org/abs/2608.01207.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv:2504.10479, 2025. URL https://arxiv.org/abs/2504.10479.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. TTRL: Test-time reinforcement learning. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2504.16084.

# A PUBLISHED METHODS: IMPLEMENTATION AND UPDATE SCOPES

**Shared protocol.** Every method sees the same 768 questions per history in the same order, starts from the same pretrained weights, answers each question and records the answer before any update on that question, and never receives labels, source identities, later questions or resets. Forward passes use BF16; updated parameters keep FP32 master copies. All runs use seed 42 and one H200 GPU unless stated. We keep each method's central objective, loss weighting and filters, and list below the parameter scope, optimizer and learning rate of each implementation together with every change from the original setting: changes that the four-option task requires (entropy thresholds defined as fractions of $\ln C$ use $C = 4$), changes of scope (ZeroSiam follows its vision implementation, not its LoRA and AdamW setting for language models), and changes of timing and resets (TTAdapt, SAR). The comparisons therefore evaluate these continual implementations, not every method in its original setting.

Table 5: Published methods and baselines as implemented. Scope: updated parameters (norm: all normalization affine parameters, 0.29M at 7B; LoRA: rank-8 adapters on $q, v$, 2.5M at 7B; all: every weight). ZeroSiam's predictor adds 12.8M parameters at 7B. Upd.: fraction of Stage-1 questions with a non-zero gradient (7B, mean over histories); SAR steps its optimizer on every question, so its momentum also moves the weights after rejected questions.

| Method | Scope, optimizer, LR | Objective and changes for this setting | Upd. |
|---|---|---|---|
| Naive self-training | LoRA, AdamW, $10^{-4}$ | cross-entropy to the model's own raw answer | 100% |
| PA self-distillation | LoRA, AdamW, $10^{-4}$ | soft cross-entropy of the four cyclic option orders of the displayed question to the frozen model's PA distribution | 100% |
| Tent | norm, SGD, $10^{-4}$ | entropy of the four-option distribution | 100% |
| CoTTA (3B) | all, Adam, $10^{-6}$ | consistency with an EMA teacher, which averages 32 augmented views when the source model's confidence is below 0.92; stochastic restoration of 1% of the weights per step (Appendix I.1) | 100% |
| ETA | norm, SGD, $10^{-4}$ | entropy weighted by $e^{-(E-E_0)}$; reliable if $E < 0.4 \ln 4$; non-redundant if $\lvert\cos(p, \bar{p})\rvert < 0.05$ with $\bar{p}$ an EMA (0.9) of accepted outputs; this is ETA as defined by EATA, without EATA's Fisher regularizer, which is estimated on an additional set of in-distribution samples | 0.5% |
| SAR | norm, SAM ($\rho = 0.05$), $10^{-4}$ | reliable entropy ($E < 0.4 \ln 4$) with sharpness-aware updates; no model reset (a reset variant: Appendix I.2) | 48% |
| DeYO | norm, SGD, $10^{-4}$ | entropy weighted by confidence and PLPD; $E < 0.5 \ln 4$ and PLPD $> 0.2$ under a $4 \times 4$ patch shuffle that keeps the image size | 13.9% |
| TLM | LoRA, AdamW, $5 \times 10^{-5}$ | input perplexity, used if NLL $> 3$ and weighted by $0.1\, e^{\mathrm{NLL}-3}$; image placeholder tokens excluded from the perplexity | 10.9% |
| COME | norm, SGD, $10^{-4}$ | Tent with the entropy replaced by the entropy of COME's opinion over the four options and an uncertainty mass, prior $K = 4$ (1000, the number of ImageNet classes, in the public code); evidence from the four option logits, rescaled by their detached norm as in the public code | 100% |
| TLM, canonical inputs | LoRA, AdamW, $5 \times 10^{-5}$ | as TLM; predictor and update see the canonical option order and template | 10.6% |
| SAR, canonical inputs | norm, SAM ($\rho = 0.05$), $10^{-4}$ | as SAR; predictor and update see the canonical option order and template | 49.3% |
| ZeroSiam | norm + predictor, SGD, $10^{-4}$ | asymmetric Siamese entropy objective in the scope of its vision implementation; its language-model setting (LoRA, AdamW, entropy of generated tokens) is not used (Appendix I.3) | 100% |
| TTRV | all, AdamW, $5 \times 10^{-7}$ | GRPO (Shao et al., 2024) with frequency and entropy rewards over 32 sampled answers, 16 trained; one question per step; the two-GPU implementation pads each request to 64 generations and discards the padding before the reward (Appendix I.4) | 100% |
| TTAdapt (3B) | all, AdamW, $2 \times 10^{-6}$ | cross-entropy to the consensus of 16 augmented views; two pseudo-label rounds of six steps per question, with the optimizer re-created each round as in the public code; answers average the 16 views. The original method restores the initial weights after every question; the continual protocol keeps the updates | 100% |

Table 6: Learning-rate sensitivity of the published methods on Qwen2.5-VL-7B (order O1, four histories). Gain in points over the frozen model: online with each method's own answers vs frozen raw answers; after adaptation with PA vs frozen PA. * : 95% interval excludes zero. The default rows are the order-O1 runs behind Table 2.

| Method | Learning rate | Online | Clean (PA) | Corrupted (PA) |
|---|---|---|---|---|
| Tent | $1 \cdot 10^{-5}$ | +0.2 | +0.1 | +0.1 |
| | $1 \cdot 10^{-4}$ (default) | +0.6 | +0.3 | −0.5 |
| | $1 \cdot 10^{-3}$ | −26.2* | −40.9* | −39.2* |
| ETA | $1 \cdot 10^{-5}$ | −0.1 | −0.1 | −0.1 |
| | $1 \cdot 10^{-4}$ (default) | +0.2 | −0.2 | +0.1 |
| | $1 \cdot 10^{-3}$ | +0.2 | +0.1 | +0.1 |
| SAR | $1 \cdot 10^{-5}$ | +0.2 | +0.1 | +0.2 |
| | $1 \cdot 10^{-4}$ (default) | +0.6* | +0.1 | −0.2 |
| | $1 \cdot 10^{-3}$ | −4.9* | −11.9* | −14.0* |
| DeYO | $1 \cdot 10^{-5}$ | +0.2 | 0.0 | +0.2 |
| | $1 \cdot 10^{-4}$ (default) | +0.5 | +0.2 | +0.3 |
| | $1 \cdot 10^{-3}$ | −15.0* | −24.2* | −22.9* |
| TLM | $5 \cdot 10^{-6}$ | −0.1 | +0.6 | −0.2 |
| | $5 \cdot 10^{-5}$ (default) | +0.2 | +0.5 | +0.4 |
| | $5 \cdot 10^{-4}$ | −2.6* | −1.0 | −2.1 |
| COME | $1 \cdot 10^{-5}$ | +0.5* | +0.1 | +0.2 |
| | $1 \cdot 10^{-4}$ (default) | +0.6 | +0.5 | −0.2 |
| | $1 \cdot 10^{-3}$ | −27.0* | −42.7* | −39.1* |
| ZeroSiam | $1 \cdot 10^{-5}$ | +0.3 | +0.2 | −0.1 |
| | $1 \cdot 10^{-4}$ (default) | −4.0* | −4.7* | −6.8* |
| | $1 \cdot 10^{-3}$ | −37.3* | −41.8* | −40.2* |

**Filters on four-option questions.** The redundancy filter of ETA compares probability vectors over four options; their cosine similarity is rarely below 0.05, so ETA updates on only 0.5% of the questions and behaves almost like the frozen model. DeYO and TLM accept 13.9% and 10.9% of the questions. We keep the published thresholds rather than tuning them on our questions, and report these rates so that the comparison can be read accordingly.

**COME at the scale of language-model logits.** COME's uncertainty mass $K/(\sum_k e^{x_k} + K)$ is negligible at the scale of MLLM option logits: along our streams its loss differs from Tent's entropy by at most $4 \times 10^{-5}$. COME still differs from Tent because its logits are rescaled by their detached norm. Our implementation matches the public formula to $10^{-15}$ in value and gradient.

**Methods not included.** ASR (Lim et al., 2026) computes its reset signal from the mean entropy of a batch and the fraction of predictions that changed relative to a source model; in its public code this controller runs only when a batch holds more than one sample, whereas our streams present one question at a time.

**Learning rates.** Table 6 repeats the 7B runs of order O1 at a tenth and ten times each method's default learning rate, with everything else unchanged. At a tenth, no method changes held-out accuracy with PA by more than 0.6 points. At ten times, Tent, SAR, DeYO, COME and ZeroSiam lose 12 to 43 points, TLM loses up to 2.1 points, and ETA, whose filters admit few updates, stays unchanged. No learning rate gives a significant held-out gain; the grid is reported in full, so the best learning rate of each method can be read off post hoc, which favors the baselines.

**Seeds.** Naive self-training, PA self-distillation, DeYO (random patch shuffle) and TLM (LoRA initialization) depend on the seed; Tent, ETA, SAR, COME and ZeroSiam are deterministic given the stream. We repeated the four seed-dependent methods with seed 43 (7B, O1). None of them gains significantly after adaptation with either seed: naive self-training, PA self-distillation and

TLM change clean and corrupted held-out accuracy by $-0.8$ to $+0.7$ points with both seeds, and DeYO loses 2.1 and 2.2 points with seed 43 (both significant).

## B PER-HISTORY RESULTS AND COSTS

Table 7 lists every 7B method by history. Table 8 lists the cost of each procedure. Seconds are medians over the S and P histories on one H200 and include the update; they depend on implementation and are meant for relative comparison.

Table 7: Qwen2.5-VL-7B by history (accuracy in %, mean over the source orders O1 and O2). Online: each method's own procedure; after adaptation: PA. Frozen references (all histories averaged): online raw 65.7, PA 66.3, PA + CoT 69.5; after adaptation PA 67.6 (clean) and 65.2 (corrupted) for every history.

| | Online | | | | After adaptation, clean | | | | After adaptation, corrupted | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | S | I | P | C | S | I | P | C | S | I | P | C |
| Naive self-training | 64.6 | 62.8 | 62.0 | 60.7 | 66.4 | 59.4 | 67.5 | 67.3 | 61.6 | 55.4 | 64.3 | 64.9 |
| Tent | 69.0 | 67.5 | 64.5 | 63.1 | 68.1 | 67.8 | 67.4 | 67.8 | 65.0 | 65.0 | 64.6 | 65.1 |
| ETA | 68.8 | 67.3 | 63.8 | 63.4 | 67.3 | 67.5 | 67.4 | 67.6 | 65.6 | 65.2 | 65.2 | 65.2 |
| SAR | 69.1 | 67.4 | 64.8 | 63.3 | 67.8 | 67.8 | 67.8 | 68.0 | 64.9 | 65.5 | 64.8 | 65.7 |
| DeYO | 68.7 | 67.5 | 64.4 | 63.6 | 67.6 | 67.7 | 67.8 | 67.8 | 65.2 | 65.5 | 65.7 | 65.9 |
| TLM | 68.9 | 66.8 | 62.9 | 62.4 | 68.0 | 68.0 | 68.4 | 67.8 | 65.0 | 65.4 | 65.4 | 65.6 |
| COME | 68.9 | 67.1 | 64.6 | 63.2 | 68.3 | 67.4 | 67.8 | 68.3 | 65.2 | 65.2 | 65.3 | 65.2 |
| ZeroSiam | 68.3 | 60.1 | 61.6 | 61.8 | 67.3 | 59.4 | 66.4 | 64.4 | 63.5 | 53.8 | 62.7 | 57.2 |
| TTRV | 66.5 | 65.6 | 63.5 | 61.4 | 67.4 | 68.1 | 67.9 | 67.9 | 63.8 | 65.1 | 65.2 | 64.3 |
| TLM, canonical inputs | 66.8 | 66.8 | 62.6 | 62.6 | 68.4 | 68.4 | 68.4 | 68.4 | 65.0 | 65.0 | 65.2 | 65.2 |
| SAR, canonical inputs | 67.3 | 67.3 | 63.5 | 63.5 | 67.4 | 67.4 | 67.8 | 67.8 | 65.1 | 65.1 | 65.2 | 65.2 |
| DtC | 71.8 | 71.8 | 67.7 | 67.7 | 69.7 | 69.7 | 69.6 | 69.6 | 66.1 | 66.1 | 66.1 | 66.1 |

Table 8: Cost per Stage-1 question on Qwen2.5-VL-7B. Passes: forward / backward passes (sampled rationales counted separately); Upd.: questions with a non-zero gradient (SAR steps its optimizer on every question, so momentum also moves the weights after rejected questions); s: median seconds on one H200 including the update (TTRV: native multi-GPU trainer, wall time per question).

| Method | Updated parameters | Passes | Upd. | Tokens | s |
|---|---|---|---|---|---|
| Frozen, raw | – | 1 / 0 | – | 0 | 0.1 |
| Frozen, PA | – | 4 / 0 | – | 0 | 0.4 |
| Frozen, CoT voting | – | 4 + 4 samples / 0 | – | 375 | 6.6 |
| Frozen, PA + CoT | – | 8 + 4 samples / 0 | – | 375 | 7.1 |
| Naive self-training | LoRA (2.52M) | 2 / 1 | 100% | 0 | 0.7 |
| Tent | norm (0.29M) | 2 / 1 | 100% | 0 | 0.6 |
| ETA | norm (0.29M) | 2 / 0.01 | 0.5% | 0 | 0.3 |
| SAR | norm (0.29M) | 3 / 2 | 48% | 0 | 1.0 |
| DeYO | norm (0.29M) | 2.6 / 0.14 | 13.9% | 0 | 0.4 |
| TLM | LoRA (2.52M) | 2 / 0.11 | 10.9% | 0 | 0.3 |
| ZeroSiam | norm + predictor (13.1M) | 2 / 1 | 100% | 0 | 0.6 |
| TTRV | all | 32 samples / GRPO | 100% | 100 | 9.0 |
| COME | norm (0.29M) | 2 / 1 | 100% | 0 | 0.6 |
| TLM, canonical inputs | LoRA (2.52M) | 2 / 0.11 | 10.6% | 0 | 0.3 |
| SAR, canonical inputs | norm (0.29M) | 3 / 2 | 49% | 0 | 1.0 |
| DtC | LoRA (2.52M) | 12 + 4 samples / 4 | 100% | 370 | 12.6 |
| DtC after adaptation, direct | – | 4 / 0 | – | 0 | 0.4 |

**Sensitivity to the inference procedure.** Besides PA over the canonical order, we evaluated every adapted model with PA over the displayed order, canonical single forwards and raw answers. Pooled over both source orders, no published method gains more than 0.5 points under any of them. For DtC (order O1), the clean held-out gain is $+2.3$ with canonical PA, $+1.5$ with displayed-order PA, $+0.5$ with canonical single forwards and $+2.0$ with raw answers (corrupted: $+0.8$, $0.0$, $+0.8$, $+1.4$), so

the retained gain is largely specific to the procedure that DtC distills into. The best frozen procedure without generation differs by condition (canonical single forwards on clean held-out questions, 68.8%; PA over the displayed order online, 66.5%, and on corrupted held-out questions, 66.3%); against it, DtC's direct answers gain $+0.9$ ($[-1.3, +3.1]$) on clean and $-0.2$ ($[-2.2, +1.9]$) on corrupted questions, pooled over orders.

## C QWEN2.5-VL-3B

Table 9 gives both stages on Qwen2.5-VL-3B with the readouts of Table 2, pooled over both source orders. After adaptation no method is significantly more accurate than the frozen model with PA; naive self-training collapses in order O2 ($-40.7$ points on clean held-out questions), and Tent after the I history in order O1, where it answers 51.4% of the stream questions correctly but only 26.7% of the held-out questions. DtC answers 64.9% of the stream questions correctly, more than every published method (at most 62.1%), but at this size its updates do not help: it is 0.8 points below the frozen PA + CoT policy online and 1.0 point below frozen PA after adaptation (both not significant); CoT is only 0.9 points more accurate than PA on the stream (Table 3). Table 10 adds the methods that run only at 3B or with their own procedure. TTAdapt, whose original design restores the initial weights after every question, collapses to one option letter within the first 100 questions of every history when its updates are kept; its frozen reference with the same 16 augmented views reaches 59.1% online and 63.4% on clean held-out questions. With order O2 it collapses in the same way: $-31.7$ points online, $-38.4$ and $-35.0$ on clean and corrupted held-out questions, and one answer letter for every held-out question in every history.

Table 9: Two-stage results on Qwen2.5-VL-3B with the readouts of Table 2 (accuracy in %, mean over the four histories and the source orders O1 and O2; in parentheses: gain over the frozen model with the same procedure; $^*$: 95% interval excludes zero). Online: each method's own procedure (published: raw; canonical-input controls: canonical; DtC: PA + CoT). After adaptation: PA on the same held-out questions. Frozen online: raw 62.0, canonical 63.6, PA 64.1, PA + CoT 65.8; frozen PA after adaptation: 67.3 (clean), 63.8 (corrupted).

| Method | Online | Clean (PA) | Corrupted (PA) |
|---|---|---|---|
| Naive self-training | 51.1 ($-11.0^*$) | 44.5 ($-22.9^*$) | 42.5 ($-21.3^*$) |
| PA self-distillation | 62.3 ($+0.3$) | 67.3 (0.0) | 63.9 ($+0.1$) |
| Tent | 60.4 ($-1.6^*$) | 61.9 ($-5.5^*$) | 59.2 ($-4.6^*$) |
| ETA | 61.9 ($-0.1$) | 67.0 ($-0.3$) | 63.6 ($-0.2$) |
| SAR | 61.9 ($-0.2$) | 66.6 ($-0.7$) | 63.5 ($-0.3$) |
| DeYO | 61.9 ($-0.2$) | 66.5 ($-0.8$) | 63.2 ($-0.6$) |
| TLM | 62.1 ($+0.1$) | 66.2 ($-1.2$) | 63.0 ($-0.8$) |
| COME | 61.7 ($-0.4$) | 66.7 ($-0.7$) | 63.6 ($-0.2$) |
| ZeroSiam | 61.5 ($-0.5$) | 66.7 ($-0.6$) | 63.6 ($-0.2$) |
| TLM, canonical inputs | 62.9 ($-0.7$) | 66.2 ($-1.1$) | 63.1 ($-0.7$) |
| SAR, canonical inputs | 63.8 ($+0.2$) | 66.5 ($-0.8^*$) | 63.7 ($-0.1$) |
| DtC | 64.9 ($-0.8$) | 66.3 ($-1.0$) | 62.8 ($-1.0$) |

Table 10: Qwen2.5-VL-3B, order O1: methods run only at 3B or with their own procedure. Gain over the frozen model with the same procedure (points, 95% intervals, mean over the four histories). CoTTA: augmented EMA-teacher procedure; TTRV: its vLLM sampling procedure; TTAdapt: 16 augmented views.

| Method | Online | After adaptation, clean | After adaptation, corrupted |
|---|---|---|---|
| SAR with reset | $-0.07$ $[-0.33, +0.20]$ | $+0.13$ $[-0.62, +0.91]$ | $-0.62$ $[-1.37, +0.13]$ |
| CoTTA | $+0.07$ $[-0.23, +0.36]$ | $-0.03$ $[-0.46, +0.36]$ | $-0.52$ $[-1.11, +0.00]$ |
| TTRV (vLLM) | $-1.04$ $[-2.51, +0.42]$ | $-2.57$ $[-4.65, -0.46]$ | $-2.93$ $[-5.14, -0.72]$ |
| TTAdapt | $-30.27$ $[-33.40, -27.18]$ | $-38.41$ $[-41.70, -35.06]$ | $-35.03$ $[-38.44, -31.58]$ |

# D DtC DETAILS

**Prompts and sampling.** Rationales use the instruction "Give a brief explanation using the image and question, in at most 80 words. End with exactly 'Final answer: X', where X is A, B, C, or D." We sample with temperature 0.7 and top-$p$ 0.9, at most 256 new tokens, one rationale for each cyclic shift of the canonical option order. The seed of each sample is derived from the question content, the image hash and a global seed (42 or 43), so the frozen baseline and DtC use the same sampling seeds and decoding settings; the rationales themselves differ once the adapted weights differ. After each rationale, the probabilities of the four option letters are scored at the position following "Final answer:"; they are mapped back to option contents and averaged to give $r_t$. The answer text is never parsed to form the target.

**Cost.** Per stream question (both orders, S and P histories), DtC generates 370 tokens on average (the frozen model 375) and spends on average 0.5 s on the four PA forwards, 11.8 s on the four rationales with scoring and 2.4 s on the update (four backward passes) on one H200; the median total is 12.6 s. After adaptation, the direct answer needs only the four PA forwards.

**Source orders and seeds.** Table 11 reports DtC for both source orders and two training seeds. The second seed changes the initialization of the LoRA adapters and the sampling seed of every rationale (43 instead of 42); its frozen reference uses the same sampling seeds, and the data, the order of questions and all other settings are unchanged. The direct-answer gain is similar for both orders and seeds (pooled $+1.9$, $[+0.2, +3.6]$). With CoT after adaptation, the gain over the frozen PA + CoT policy is $+1.6$ ($[+0.2, +3.0]$, mean of two sampling seeds) with the first training seed and $+0.7$ ($[-0.7, +2.1]$) with the second (pooled $+1.1$, $[-0.2, +2.5]$).

Table 11: DtC on Qwen2.5-VL-7B with two source orders and two training seeds (gain over the matched frozen procedure in points, pointwise 95% intervals). Online: recorded answer (PA + CoT) vs frozen PA + CoT with the same sampling seeds. Direct: PA after adaptation vs frozen PA. CoT: PA + CoT after adaptation vs frozen PA + CoT (mean of two evaluation sampling seeds). Clean held-out questions.

| Order | Seed | Online | Direct (clean) | CoT (clean) |
|---|---|---|---|---|
| O1 | 42 | −0.1 [−1.6, +1.4] | +2.3 [+0.4, +4.2] | +1.2 [−0.4, +2.8] |
| O2 | 42 | +0.5 [−1.0, +2.1] | +1.8 [0.0, +3.6] | +2.0 [+0.5, +3.5] |
| O1 | 43 | −0.3 [−1.9, +1.2] | +1.8 [−0.1, +3.8] | +0.9 [−0.6, +2.5] |
| O2 | 43 | −0.3 [−1.8, +1.2] | +1.7 [0.0, +3.4] | +0.5 [−0.9, +1.9] |
| both | 42 | +0.2 [−1.1, +1.5] | +2.1 [+0.3, +3.8] | +1.6 [+0.2, +3.0] |
| both | 43 | −0.3 [−1.6, +1.0] | +1.8 [0.0, +3.5] | +0.7 [−0.7, +2.1] |
| both | both | 0.0 [−1.1, +1.0] | +1.9 [+0.2, +3.6] | +1.1 [−0.2, +2.5] |

**Multiple comparisons.** The main text reports pointwise intervals. For the comparisons of DtC that we specified before the final runs (online against each published method and naive self-training; after adaptation, direct answers against frozen PA and CoT answers against frozen PA + CoT, on clean and on corrupted questions), Table 12 gives paired sign-flip randomization $p$-values (100,000 permutations; the intervals elsewhere are bootstrap intervals) with a Holm correction (Holm, 1979). The online comparisons remain significant after correction; the clean gains after adaptation do not ($p = 0.023$ and $0.034$, corrected $0.09$ and $0.10$).

**Controlled study.** Table 13 lists every variant of the controlled study with fixed targets from the frozen model (7B, O1, one LoRA $q, v$ step per question; gains over the frozen model with PA). No variant is significantly better than the reference CoT target, and learning rate $5 \times 10^{-4}$ collapses the model. The 32B rows of Table 3 use a frozen Qwen2.5-VL-32B: its PA distribution on each stream question is released to the learner only after the learner's answer is recorded, and the student takes one AdamW step on LoRA $q, v$ at learning rate $10^{-4}$, averaging the cross-entropy over the four option orders. The teacher answers 73.70% of the S/I and 68.62% of the P/C stream questions and 73.83% of the clean held-out questions correctly.

Table 12: Pre-specified comparisons of DtC (Qwen2.5-VL-7B, seed 42, both source orders): gain in points, paired sign-flip randomization $p$-value (100,000 permutations) and Holm-adjusted $p$-value.

| Comparison | Gain | $p$ | Holm $p$ |
|---|---|---|---|
| online vs Naive self-training | +7.2 | <0.0001 | 0.0001 |
| online vs ZeroSiam | +6.8 | <0.0001 | 0.0001 |
| online vs TTRV | +5.5 | <0.0001 | 0.0001 |
| online vs TLM | +4.5 | <0.0001 | 0.0003 |
| online vs ETA | +3.9 | 0.0001 | 0.0011 |
| online vs COME | +3.8 | 0.0002 | 0.0013 |
| online vs DeYO | +3.7 | 0.0002 | 0.0015 |
| online vs Tent | +3.7 | 0.0003 | 0.0017 |
| online vs SAR | +3.6 | 0.0004 | 0.0020 |
| direct vs frozen PA, clean | +2.1 | 0.0234 | 0.0937 |
| CoT vs frozen PA + CoT, clean | +1.6 | 0.0342 | 0.1025 |
| CoT vs frozen PA + CoT, P | +0.7 | 0.3813 | 0.7626 |
| direct vs frozen PA, P | +0.9 | 0.4062 | 0.7626 |

Table 13: Controlled study with fixed targets from the frozen model (7B, O1). Online: PA-only answers; Table 2 uses PA + CoT for DtC. $^*$: 95% interval excludes zero.

| Variant | Online | Clean | Corrupted |
|---|---|---|---|
| Reference: frozen-model CoT target, learning rate $10^{-4}$ | $+0.5$ | $+2.2^*$ | $+1.6$ |
| frozen-model PA target instead | $+1.0^*$ | $+0.7$ | 0.0 |
| hard CoT vote as target | $-0.1$ | $+2.0^*$ | $+1.1$ |
| learning rate $5 \times 10^{-5}$ | $+0.4$ | $+1.0$ | $+0.7$ |
| learning rate $2 \times 10^{-4}$ | $+0.4$ | $+1.4$ | $+1.2$ |
| learning rate $3 \times 10^{-4}$ | $+0.1$ | $+0.3$ | $+0.1$ |
| learning rate $5 \times 10^{-4}$ | $-8.8^*$ | $-29.2^*$ | $-27.4^*$ |
| replay of 4 past questions per step | $+0.7$ | $+1.2$ | $+1.8$ |
| replay of 8 past questions per step | $+0.5$ | $+0.8$ | $+1.3$ |
| one extra corrupted view per step | $+0.8$ | $+2.5^*$ | $+1.8^*$ |

**CoT answers after the PA-target ablation.** With the PA target of Table 4, whose direct answers become nearly uniform, the adapted model's PA + CoT answers (one sampling seed) stay at the frozen level on clean held-out questions ($+0.1$, $[-2.1, +2.3]$) and drop on corrupted ones ($-3.2$, $[-5.4, -0.9]$): after adaptation, too, the CoT readout can hide a degraded direct predictor.

**Where the gain comes from.** The clean direct-answer gain is largest on PhyX ($+9.6$); without PhyX it is $+1.0$ ($[-0.6, +2.6]$). On the held-out PhyX questions, the correct answer is the third of the sorted options for 51 of 96 questions, and DtC moves its direct answers towards the later positions (S history, order O1); in the role swap this pattern is weaker (39 of 96) and the PhyX gain is $-5.7$. With CoT after adaptation, the gain is positive on all eight sources ($+1.6$ without PhyX).

**New-question panel.** Before running it, we registered a panel of 768 new stream questions and 192 new held-out questions from PhyX and PathMMU, the two sources for which unused questions remained. Both orders and the S and P histories were run. Online, DtC is $+0.7$ ($[-1.2, +2.5]$) over the frozen PA + CoT policy. After adaptation, the direct answer is $-0.4$ ($[-5.3, +4.6]$) and $-1.0$ ($[-6.4, +4.4]$) from frozen PA on clean and corrupted images, and the CoT answer is $-2.9$ ($[-8.2, +2.5]$) and $+2.7$ ($[-2.3, +7.9]$) from the frozen PA + CoT policy. The panel is small and covers two hard sources, so it neither confirms nor rules out the development gains.

**Role-swap replication.** Before running it, we registered a replication that swaps the roles of the two halves of the core questions: the 768 held-out questions become the stream (order O1, S and P histories) and the 768 stream questions become the held-out set. The primary endpoint, fixed in advance, is the gain of DtC's direct answers over frozen PA on clean held-out questions. Table 14 reports every method. The published methods behave as on the main split. The primary endpoint is

not replicated: $-1.0$ ($[-2.9, +0.8]$; sign-flip $p = 0.31$), with PhyX at $-5.7$, while DtC stays close to the frozen PA + CoT policy online ($+1.0$, $[-0.5, +2.6]$). Pooled over both splits (order O1, 1,536 held-out questions), the direct-answer gain is $+0.6$ ($[-0.7, +2.0]$) on clean and $+0.5$ ($[-1.0, +1.9]$) on corrupted questions.

Table 14: Role-swap replication on Qwen2.5-VL-7B (order O1; S and P histories; the stream is the held-out set of the main split and vice versa). Accuracy in % and, in parentheses, gain over the frozen model with the same procedure; $^*$: 95% interval excludes zero. Online: published methods raw, DtC PA + CoT; after adaptation: PA. Frozen: online raw 65.5, PA + CoT 67.8; after adaptation PA 68.4 (clean), 64.2 (corrupted).

| Method | Online | Clean (PA) | Corrupted (PA) |
|---|---|---|---|
| Naive self-training | 53.4 ($-12.1^*$) | 59.5 ($-8.9^*$) | 56.4 ($-7.8^*$) |
| PA self-distillation | 67.4 ($+2.0^*$) | 67.4 ($-0.9$) | 63.9 ($-0.3$) |
| Tent | 65.8 ($+0.3$) | 68.7 ($+0.3$) | 65.5 ($+1.3$) |
| ETA | 65.4 ($-0.1$) | 67.9 ($-0.5^*$) | 64.8 ($+0.7^*$) |
| SAR | 65.4 ($-0.1$) | 68.3 ($-0.1$) | 64.7 ($+0.5$) |
| DeYO | 66.0 ($+0.5$) | 68.4 ($+0.1$) | 64.5 ($+0.3$) |
| TLM | 65.5 (0.0) | 67.6 ($-0.7$) | 64.5 ($+0.3$) |
| COME | 65.9 ($+0.4$) | 68.4 (0.0) | 64.6 ($+0.4$) |
| ZeroSiam | 65.3 ($-0.2$) | 64.4 ($-4.0^*$) | 58.5 ($-5.7^*$) |
| DtC | 68.9 ($+1.0$) | 67.3 ($-1.0$) | 64.3 ($+0.1$) |

**Update targets in the role swap.** We repeated the fixed-target updates of Table 3 on the role swap. On its stream, the CoT answer of the frozen model is only 0.7 points more accurate than PA (2.2 on the main split), and distilling it retains $-0.4$ on clean and $+0.7$ on corrupted held-out questions. Even training towards the true answers retains only $+0.8$ on clean questions ($+3.7$ on the main split) and $+2.6$ on corrupted ones, so the held-out questions of the role swap are harder to improve with any target. Pooled over both splits, the order is kept: the own PA target retains $+0.2$, the CoT target $+0.9$ and the true answers $+2.2$ points on clean held-out questions (corrupted: $+0.1$, $+1.1$, $+2.9$).

**Split variability.** Table 15 repeats the main comparisons on two random re-splits of the core questions, drawn within every source, interface family and corruption before any result was seen. The first column gives the CoT advantage on each stream: the accuracy of the frozen model's CoT target minus that of its PA on the stream questions (S and P histories), as the target gain of Table 3. Over the four splits, the clean direct-answer gain of DtC ($+2.3$, $-1.0$, $+1.7$, $-0.5$; mean $+0.6$) and that of the fixed CoT target ($+2.2$, $-0.4$, $+1.4$, $+0.6$) follow this advantage ($+2.2$, $+0.7$, $+1.6$, $+0.8$), training towards the true answers retains a gain on every split ($+2.2$ and $+2.6$ on average), and online DtC stays at the level of frozen PA + CoT (mean $+0.5$). In the role swap, where the CoT advantage is smallest, order O2 and a second training seed do not help either ($-2.3$ and $-0.5$ for clean direct answers; with CoT, $-1.6$ to $-0.9$ on clean and $-2.1$ to $-0.4$ on corrupted questions).

**Rationales and learning rate.** Table 16 varies the number of rationales and the learning rate (order O1). One rationale per question retains $+2.6$ and $+1.8$ points on clean and corrupted questions at a quarter of the tokens, two rationales retain $+0.8$ and $+0.3$, and the retained gain moves from $+1.2$ at learning rate $5 \times 10^{-5}$ to $-0.7$ at $2 \times 10^{-4}$. Online, all variants stay within 1.2 points of the matched frozen policy. The retained gain of DtC is thus sensitive to settings that barely change its online behavior.

## E Benchmark construction

**Selection.** Questions are sampled with seed 42 from each source. Each has four distinct non-empty options and its source answer; source identifiers, revisions and content hashes fix the membership. Exact duplicate images are removed, keeping at most one question per image; semantic near-duplicates may remain. Questions and options are not paraphrased or translated.

Table 15: Retained gains over four splits of the core questions (Qwen2.5-VL-7B, order O1, seed 42, S and P histories; the two re-splits were fixed before any of their results was seen). Gains in points over the frozen model with the same procedure: DtC online against frozen PA + CoT with the same sampling seeds; after adaptation, PA against frozen PA for DtC's direct answers and for the fixed-target updates of Table 3. CoT adv.: accuracy of the frozen model's CoT target minus its PA on the stream questions. $^*$: 95% interval excludes zero.

| | | DtC | | | CoT target | | True answer (oracle) | |
|---|---|---|---|---|---|---|---|---|
| Split | CoT adv. | Online | Clean | Corr. | Clean | Corr. | Clean | Corr. |
| Benchmark split | $+2.2$ | $-0.1$ | $+2.3^*$ | $+0.8$ | $+2.2^*$ | $+1.6$ | $+3.7^*$ | $+3.2^*$ |
| Role swap | $+0.7$ | $+1.0$ | $-1.0$ | $+0.1$ | $-0.4$ | $+0.7$ | $+0.8$ | $+2.6^*$ |
| Random re-split 1 | $+1.6$ | $+0.8$ | $+1.7$ | $+0.7$ | $+1.4$ | $+1.5$ | $+1.7$ | $+1.8$ |
| Random re-split 2 | $+0.8$ | $+0.3$ | $-0.5$ | $-0.9$ | $+0.6$ | $+0.3$ | $+2.7^*$ | $+2.6^*$ |
| Mean | $+1.3$ | $+0.5$ | $+0.6$ | $+0.2$ | $+0.9$ | $+1.0$ | $+2.2$ | $+2.6$ |

Table 16: DtC with fewer rationales and other learning rates (7B, O1; gains over the frozen model with the same procedure, points; $^*$: 95% interval excludes zero). Online: recorded answer against the frozen model with the same $k$ rationales (same sampling seeds). After adaptation: direct answers (PA) against frozen PA. Tokens: generated per stream question (mean); s: median seconds per stream question on one H200.

| Variant | Tokens | s | Online | Clean | Corrupted |
|---|---|---|---|---|---|
| $k = 1$ rationale | 94 | 5.6 | $+0.5$ | $+2.6^*$ | $+1.8^*$ |
| $k = 2$ rationales | 189 | 8.5 | $-0.8$ | $+0.8$ | $+0.3$ |
| $k = 4$ rationales (DtC) | 373 | 12.5 | $-0.1$ | $+2.3^*$ | $+0.8$ |
| learning rate $5 \times 10^{-5}$ | 375 | 13.1 | $-0.3$ | $+1.2$ | $+1.3$ |
| learning rate $2 \times 10^{-4}$ | 393 | 13.3 | $-1.2$ | $-0.7$ | $-0.7$ |

**Prompt.** All prompts use the instruction "Answer the multiple-choice question. Return only the label shown before your chosen option." Each option is rendered as its label, a period and its original text.

**Assignment and balance.** Within a source, 12 questions fall into each interface-change $\times$ corruption cell, so every interface family meets all four corruptions. Each 768-question stream uses six questions per cell and source. In the I history, the correct letters are A/B/C/D = 174/105/115/182 and 1/2/3/4 = 48/48/48/48; the skewed-position family alone contributes A/B/C/D = 80/16/16/80. The strengths were chosen during development; JPEG at quality 5 and low light in particular are strong corruptions, and a retained label does not guarantee that a corrupted question is still answerable.

**Images.** Images keep their native size when possible (1,532 of 1,536): they are padded with white to multiples of 28 pixels, resized at most once, and the Qwen processor resize is disabled. Corruptions apply to the padded image.

**Option wording.** An outcome-blind text screen flagged one MMBench distractor, "All above are not right", whose meaning can depend on its position when options move. We keep the original question and label; this is a possible change of interpretation, not a verified label error.

Table 17: Question sources. Counts are distinct questions, each with one image and four options.

| Source | Content | Questions |
|---|---|---|
| MMBench English v1.1 (Liu et al., 2024) | General multimodal understanding | 192 |
| MMMU (Yue et al., 2024) | Multidisciplinary expert reasoning | 192 |
| A-OKVQA (Schwenk et al., 2022) | Visual questions requiring world knowledge | 192 |
| AI2D (Kembhavi et al., 2016) | Scientific diagrams | 192 |
| MathVista (Lu et al., 2024) | Mathematics in visual contexts | 192 |
| SEED-Bench-2-Plus (Li et al., 2024) | Text-rich visual understanding | 192 |
| PathMMU (Sun et al., 2024) | Pathology | 192 |
| PhyX (Shen et al., 2025) | Physical reasoning | 192 |
| Total | Eight equally sized sources | 1,536 |

Table 18: Interface changes. Each family is assigned to 48 questions per source; the I and C histories use the same assignment.

| Family | Specification |
|---|---|
| Option permutation | Cyclically shift the option contents by one, two or three positions, keeping A/B/C/D as labels; the answer moves with its content. |
| Numeric labels | Replace A/B/C/D with 1/2/3/4, keeping the original option order. |
| Moved instruction | Move the instruction from before the question to after the options, separated by a blank line. |
| Skewed answer position | Split the 48 questions into two phases of 24 with correct-position counts (18,2,2,2) and (2,2,2,18) for A/B/C/D; half of the sources use A then D, the other half D then A. The reference answer is used only to construct the history, never by the learner. |

Table 19: Image corruptions, similar in kind to those of common-corruption benchmarks (Hendrycks & Dietterich, 2019), applied to the prepared RGB image. One corruption with a fixed strength per question.

| Corruption | Specification |
|---|---|
| Gaussian blur | Pillow GaussianBlur with radius 2.5 pixels. |
| JPEG | One encode/decode cycle at quality 5 with 4:2:0 chroma subsampling; the result is stored losslessly. |
| Low light | Brightness multiplied by 0.1. |
| Grayscale | Conversion to grayscale and back to RGB. |

# F QUESTION SPLITS AND DEVELOPMENT EXPOSURE

Table 20: Evaluation units. "Held-out" means excluded from that run's updates, not excluded from all research inspection.

| Study | Stream | Held-out | Role |
|---|---|---|---|
| Main two-stage comparison | 768 | 768 | Online accuracy, then evaluation on the same held-out questions for every history. |
| Role-swap replication | 768 | 768 | The main held-out questions as the stream and the main stream questions as the held-out set (Appendix D). |
| Random re-splits | 768 | 768 | Two further random assignments of the core questions to stream and held-out set (Appendix D). |
| Registered panel | – | 416 | New questions for fixed SmolVLM2 states (Appendix H.3). |
| New-history replication | 768 | 192 | New stream and held-out questions from two sources (Appendix H.4). |
| New-question panel for DtC | 768 | 192 | New stream and held-out questions from two sources (Appendix D). |

All 1,536 core questions were used during development: the benchmark design, the learning rates and the settings of DtC were chosen while looking at them. Per-run disjointness of stream and held-out questions holds throughout, but results on the core questions are development results. The registered panel, the new-history replication and the new-question panel for DtC were fixed before any model was run on them and are used only for reporting.

The methods of Table 2 were run on Qwen2.5-VL-7B and -3B with both source orders and all four histories; DtC and the canonical-input controls run the S and P histories, whose canonical inputs are exactly those of I and C. TTRV was run on the 3B model with order O1 only. CoTTA and TTAdapt, which update all weights, were run on the 3B model only, with both source orders. On InternVL3.5-2B, the methods of Table 21 were run with order O1.

# G IMPLEMENTATION DETAILS

**Models.** The main models are Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct (Bai et al., 2025). SmolVLM2-2.2B-Instruct (Marafioti et al., 2025) adds a different decoder family (SmolLM2-1.7B); it keeps its native maximum edge of 1,536 pixels and 384-pixel tiles. InternVL3.5-2B (Wang et al., 2025) is a further model family; its image path is adapted to keep the prepared image instead of the default 448-pixel tiling, and its language backbone (Qwen3) limits claims of independence from the Qwen family.

**Answer scoring.** The logits of the four displayed option labels at the answer position define a four-way distribution. Answers are mapped back to option contents for evaluation and for averaging across option orders. Forward passes use BF16 with SDPA attention; updated parameters keep FP32 master copies.

**Update scopes.** LoRA methods adapt the language model's $q$ and $v$ projections with rank 8, scaling 16 and no dropout, using AdamW (Loshchilov & Hutter, 2019) without weight decay and gradient clipping at 1. Normalization methods update all normalization affine parameters of the vision encoder, the visual merger and the language model. The learning rate of the methods that update normalization layers ($10^{-4}$) was chosen for Tent from $\{10^{-3}, 10^{-4}\}$ on development averages and then used for ETA, SAR, DeYO and ZeroSiam; TLM, TTRV and TTAdapt use the authors' learning rates (Table 5).

**Metrics.** Cumulative extra errors after $t$ questions are $E_t = \sum_{i \le t}[\mathbf{1}(\hat{y}_i \neq y_i) - \mathbf{1}(\hat{y}_i^0 \neq y_i)]$, where $\hat{y}^0$ is the answer of the frozen model with the same procedure. The maximum letter frequency

is the largest fraction of answers given with one displayed label; it is read together with the true answer positions.

**Statistics.** Intervals resample the questions within each source (10,000 draws, seed 42); a resampled question keeps all its histories and image conditions, so the four histories of one question are never treated as independent samples. Intervals are conditional on the trained states and are not adjusted for the number of methods examined during development. Small differences need many questions: with a disagreement rate of 10%, a half-width of one point requires about 3,800 independent questions, so differences below about two points on 768 questions are usually inconclusive.

## H INTERFACE EFFECTS ACROSS MODELS

**InternVL3.5-2B.** With naive self-training, InternVL3.5-2B answers 56.38% of the S stream and 29.95% of the I stream correctly (frozen: 59.64% and 60.55%). On the same held-out questions (raw answers; frozen 60.03%), the S history leaves it at 59.11%, whereas after the I history it gives the same letter to every question (25.00%). The I−S contrast is $-34.11$ points ($[-38.41, -29.82]$) with raw answers and $-36.72$ points ($[-41.15, -32.42]$) when the held-out options are sorted. The collapsed model ignores the image: replacing it by a white image changes its answer distribution by a total variation of about $3.6 \times 10^{-10}$. The frozen probabilities agree across histories, so the effect lives in the updated parameters. Reversing the source order makes both histories collapse.

**Methods on InternVL3.5-2B.** Table 21 runs the published methods that update normalization layers or LoRA adapters, naive self-training, PA self-distillation and DtC on InternVL3.5-2B with the settings chosen for Qwen2.5-VL. No published method significantly improves held-out accuracy with PA, and Tent, COME, DeYO, TLM and naive self-training significantly lower it. With the same settings, DtC lowers held-out accuracy with PA by 10.9 points on clean and 11.1 on corrupted questions: its direct answers become flatter, without collapsing to one letter, and for the frozen model CoT is only 2.0 points more accurate than PA on the stream ($[-0.5, +4.4]$).

Table 21: InternVL3.5-2B (order O1, four histories; settings transferred from Qwen2.5-VL without tuning). Accuracy in % and, in parentheses, gain over the frozen model with the same procedure; $^*$: 95% interval excludes zero. Online: each method's own procedure (published methods: raw; DtC: PA + CoT, against the frozen model with the same sampling seeds). After adaptation: PA over the canonical order and raw answers on the same held-out questions.

| | | After adaptation, PA | | After adaptation, raw | |
|---|---|---|---|---|---|
| Method | Online | Clean | Corrupted | Clean | Corrupted |
| Frozen | 58.2 | 63.5 | 60.9 | 60.0 | 57.4 |
| Naive self-training | 35.6 ($-22.5^*$) | 53.8 ($-9.8^*$) | 52.0 ($-9.0^*$) | 33.5 ($-26.5^*$) | 32.7 ($-24.7^*$) |
| PA self-distillation | 60.4 ($+2.3^*$) | 63.6 ($+0.1$) | 61.3 ($+0.3$) | 62.2 ($+2.2$) | 59.8 ($+2.3^*$) |
| Tent | 45.2 ($-13.0^*$) | 57.6 ($-5.9^*$) | 54.9 ($-6.0^*$) | 36.7 ($-23.4^*$) | 35.0 ($-22.5^*$) |
| ETA | 58.0 ($-0.1$) | 63.3 ($-0.2$) | 61.5 ($+0.6$) | 60.1 (0.0) | 57.3 ($-0.1$) |
| DeYO | 56.9 ($-1.3^*$) | 61.0 ($-2.6^*$) | 57.9 ($-3.0^*$) | 58.1 ($-1.9$) | 55.2 ($-2.2$) |
| COME | 39.3 ($-18.9^*$) | 57.8 ($-5.8^*$) | 54.2 ($-6.8^*$) | 30.9 ($-29.1^*$) | 30.4 ($-27.1^*$) |
| TLM | 55.3 ($-2.9^*$) | 61.5 ($-2.1^*$) | 59.5 ($-1.4$) | 52.8 ($-7.2^*$) | 50.5 ($-6.9^*$) |
| DtC | 58.5 ($-5.7^*$) | 52.6 ($-10.9^*$) | 49.9 ($-11.1^*$) | 52.6 ($-7.4^*$) | 50.1 ($-7.3^*$) |

**Which interface change matters (Qwen2.5-VL-3B).** With naive self-training at a fixed learning rate, restoring one interface family at a time to the standard presentation shows that removing the permutation family rescues order O1 but not O2, and the standard history itself collapses in O2, so interface changes are not necessary for self-training to fail. Sorting the options during the stream avoids collapse in both orders but does not make the updates beneficial ($+0.52$ and $-3.91$ points against the frozen raw answers).

### H.1 HELD-OUT QUESTIONS WITH THE CHANGED INTERFACE

The main text evaluates the held-out questions in the standard interface, because the canonical procedures re-render every question and remove the constructed interface changes: canonical answers on the held-out questions in the I (C) presentation are correct exactly when they are on the S (P) presentation, for every adapted state (0 mismatches in 113,664 comparisons). Raw answers keep the displayed interface, so Table 22 evaluates every method that updates normalization layers or LoRA adapters on the held-out questions in all four presentations. The conclusions of Finding 1 hold with two exceptions on the corrupted questions with the changed interface, where SAR and DeYO gain 1.1 and 1.5 points on Qwen2.5-VL-7B; ZeroSiam loses 4 to 6 points in every presentation. The last two columns contrast the I and S histories. We do not observe a stable interface-specific effect: where the I history differs from the S history, it does so on I and on S questions alike (for example $+4.0$ and $+3.3$ for naive self-training on 7B).

Table 22: Held-out questions in all four presentations (raw answers). Columns S–C: gain over the frozen model on the held-out questions shown in the standard (S), changed-interface (I), corrupted (P) and both-changed (C) presentation, averaged over the four histories. Last two columns: accuracy of the I-history model minus the S-history model on I and on S questions. $^*$: 95% interval excludes zero.

| | Gain over frozen, held-out presentation | | | | I history − S history | |
|---|---|---|---|---|---|---|
| Method | S | I | P | C | on I | on S |
| *Qwen2.5-VL-7B* (frozen, raw: S 66.7, I 67.6, P 64.3, C 62.8) | | | | | | |
| Naive self-training | −0.4 | −2.2* | −1.0 | −0.3 | +4.0* | +3.3* |
| Tent | +0.3 | −0.4 | −0.2 | +1.1 | 0.0 | −0.7 |
| ETA | +0.1 | 0.0 | +0.2 | 0.0 | 0.0 | +0.3 |
| SAR | +0.1 | −0.4 | −0.2 | +1.1* | +0.3 | −0.1 |
| SAR with reset | −0.1 | +0.1 | 0.0 | 0.0 | +0.1 | −0.3 |
| DeYO | +0.1 | −0.1 | +0.5 | +1.5* | −0.4 | +0.1 |
| TLM | −1.3 | +0.3 | +0.1 | +0.7 | −0.1 | +0.7 |
| ZeroSiam | −4.1* | −5.7* | −5.9* | −5.7* | −11.8* | −11.7* |
| PA self-distillation | +1.7* | −0.7 | +1.3 | +1.0 | +0.1 | +0.8 |
| *Qwen2.5-VL-3B* (frozen, raw: S 65.8, I 65.0, P 63.8, C 61.2) | | | | | | |
| Naive self-training | −13.0* | −14.8* | −13.5* | −13.9* | −41.1* | −39.5* |
| Tent | −9.6* | −10.1* | −9.9* | −8.9* | −38.2* | −40.0* |
| ETA | +0.4 | −0.3 | −0.4 | 0.0 | +0.4 | +0.5 |
| SAR | +0.4 | −0.4 | −0.4 | +0.3 | −1.3* | +0.9 |
| SAR with reset | +0.1 | −0.8* | −0.6 | −0.4 | +0.1 | −0.5 |
| DeYO | +0.3 | −1.0 | 0.0 | +0.9 | +0.9 | +0.1 |
| TLM | −1.1 | −2.6* | −2.9* | −1.2 | 0.0 | −1.2 |
| ZeroSiam | −0.1 | −1.4* | −1.3 | +0.1 | −1.8* | −0.3 |
| PA self-distillation | −0.5 | −1.1 | −2.0* | −0.3 | −0.1 | 0.0 |

### H.2 SMOLVLM2 AND THE HISTORY INTERACTION

Table 23: SmolVLM2 with naive self-training (LoRA, learning rate $10^{-4}$): canonical held-out accuracy (%). The maximum letter frequency refers to the canonical order.

| History | Clean | Corrupted | Max letter, clean |
|---|---|---|---|
| S | 52.08 | 48.57 | 37.50 |
| I | 40.23 | 39.06 | 71.48 |
| P | 50.39 | 48.83 | 64.71 |
| C | 52.99 | 50.65 | 40.62 |
| Frozen | 52.73 | 51.56 | – |

On clean canonical held-out questions, I−S $= -11.85$ points and C−P $= +2.60$ points, an interaction of $+14.45$ points ($[+8.46, +20.44]$; corrupted: $+11.33$, $[+5.47, +17.19]$). The interaction describes these fixed histories; it does not mean that corruption helps.

### H.3 Registered panel of new questions

Before running any model on them, we built a panel of new clean four-option questions, excluding every question and image used elsewhere in this work (including near-duplicate images within Hamming distance four of a difference hash). A fixed hash order selected at most 128 questions per source; sources with fewer than 16 eligible images were omitted by this rule. The panel has 416 questions: 32 from MMMU and 128 each from SEED-Bench-2-Plus, PhyX and PathMMU. We registered the fixed SmolVLM2 frozen, S and I states, the clean canonical I−S contrast as primary endpoint, equal source weights, and 10,000 bootstrap draws and 100,000 sign permutations with seed 42. With equal source weights, the clean canonical accuracy is 43.16% for the frozen model, 39.26% after the S history and 26.17% after the I history, a registered contrast of $-13.09$ points ($[-20.70, -5.47]$; sign-permutation $p = 0.0011$); with question weights it is $-16.11$. With raw answers it is $-4.49$ ($[-12.30, +3.32]$). The panel confirms the effect of the existing states on new questions; it does not test new histories.

### H.4 New histories and new held-out questions

Before running any model, we fixed SmolVLM2, two learning rates ($10^{-4}$, $10^{-5}$), two source orders, and the S, I and canonical-during-stream histories. The new pool has 960 questions from PhyX and PathMMU (384 stream and 96 held-out questions per source), excluding all recorded project questions, the registered panel and near-duplicate images. Only these two sources had enough unused questions. O2 reverses the two source blocks. Averaged over orders, the registered high-learning-rate I−S contrast is $+8.85$ points ($[+3.39, +14.32]$; Holm-adjusted $p = 0.007$), the opposite sign of the core result, and the low-minus-high learning-rate interaction is $-14.58$ ($[-21.09, -8.07]$). The history effect is therefore real but its direction depends on the stream and the update strength. Source mix, difficulty and block length all differ from the core benchmark, so the replication does not isolate a single cause.

Table 24: New-history replication: clean canonical held-out accuracy (%) on the same 192 questions (frozen: 35.94%).

| Order | Learning rate | S | I | Canonical during stream | I−S [95% CI] |
|---|---|---|---|---|---|
| O1 | $10^{-4}$ | 21.35 | 35.42 | 29.69 | $+14.06$ $[+5.73, +22.40]$ |
| O1 | $10^{-5}$ | 35.42 | 25.52 | 25.00 | $-9.90$ $[-16.67, -2.60]$ |
| O2 | $10^{-4}$ | 25.00 | 28.65 | 25.00 | $+3.65$ $[-3.65, +10.42]$ |
| O2 | $10^{-5}$ | 29.69 | 28.12 | 27.08 | $-1.56$ $[-5.21, +2.08]$ |

## I Implementation details for CoTTA, SAR, ZeroSiam and TTRV

### I.1 CoTTA

Our implementation keeps CoTTA's mechanics (Wang et al., 2022): an anchor-confidence threshold of 0.92, 32 augmented views with averaged logits, prediction with the EMA teacher before the update, soft cross-entropy, Adam, EMA coefficient 0.999 and stochastic restoration of source weights with probability 0.01. All 3.75B parameters of Qwen2.5-VL-3B are updated, with FP32 master weights and BF16 forward passes; the learning rate is $10^{-6}$ (chosen from $\{10^{-6}, 10^{-5}\}$ on a development stream). Flip and affine augmentations may hurt text-heavy questions; the frozen reference uses the same augmented procedure. With its own procedure, CoTTA changes online accuracy by $+0.07$ points and held-out accuracy by $-0.03$ (clean) and $-0.52$ (corrupted) points (Table 10). Single-forward answers of the EMA model gain $+0.78$ and $+1.01$ points on clean held-out questions in the two orders, but lose $0.65$ and $0.49$ points against a common BF16 forward pass whose frozen reference is 1.3 points more accurate, so this small effect depends on the numerical execution of the reference.

### I.2 SAR

We keep the published SAR update and SAM optimizer (Niu et al., 2023; Foret et al., 2021): entropy filter $0.4 \ln 4$, SAM radius 0.05, SGD momentum 0.9, entropy EMA 0.9 and the reset criterion 0.2. The main comparison disables SAR's model reset, because resets are outside the no-reset protocol; the variant with reset is reported separately. With the reset, SAR resets 433–593 times per history at 3B and 516–599 times at 7B, and two 7B histories end exactly at the initial weights. Without it, SAR accepts a non-zero gradient on 237–284 questions at 3B and 339–404 at 7B; its optimizer still steps on every question, so momentum also moves the weights after rejected questions. Neither variant improves the held-out accuracy (Table 2, Table 7).

### I.3 ZEROSIAM

ZeroSiam (Chen et al., 2026a) regularizes entropy minimization with an asymmetric predictor and a stopped original branch. Our implementation uses the final hidden state of the answer position as the feature, restricts the frozen output head to the four option labels, and updates the normalization parameters and an identity-initialized bias-free predictor (0.29M and 12.8M parameters at 7B) with separate SGD optimizers (momentum 0.9, learning rate $10^{-4}$). The loss is the predictor-branch entropy plus a symmetric KL term with weight one; answers use the original branch. The paper's language experiments use LoRA, AdamW and entropy over the first eight generated tokens with learning rates from $5 \times 10^{-6}$ to $1.25 \times 10^{-5}$; we did not find their executable branch in the public code, so our implementation follows the vision implementation. Our negative results do not refute the native language results.

### I.4 TTRV

**Native check.** We first ran the public InternVL3-2B (Zhu et al., 2025) MathVista setup of TTRV (Singh et al., 2026): 20 support questions for two epochs, 32 sampled answers, 16 trained responses per question, the frequency and entropy reward (entropy weight 0.75), GRPO (Shao et al., 2024) with reference KL 0.001, AdamW at $5 \times 10^{-7}$, and the native scope (vision frozen). Accuracy on the 490 native evaluation questions rises from 60.00% to 63.27% ($+3.27$, $[+1.22, +5.51]$), and to $+3.01$ ($[+0.86, +5.16]$) after excluding the support questions and their images. The method therefore works in its native episodic setting.

**Continual setting.** In our protocol each question is answered and recorded before one GRPO step on that question, with the same reward, 32 samples, 16 trained responses and reference KL, and no resets. We ran it on both model sizes with the native learning rate $5 \times 10^{-7}$, updating all weights, with its own vLLM inference (Kwon et al., 2023) (two GPUs per run at 7B); the frozen reference uses the same vLLM procedure. At 3B, the held-out accuracy changes by $-2.57$ ($[-4.65, -0.46]$) with raw and $-1.01$ ($[-3.09, +1.07]$) with canonical answers on clean questions; at 7B, by $-0.9$ with canonical answers from its vLLM inference (order O1) and by $+0.2$ (clean) and $-0.6$ (corrupted) with PA over both orders (Table 2). The same frozen 3B model with PA reaches 67.32% on the clean held-out questions, 3.7 points above TTRV's 63.57%.

## J FURTHER RELATED WORK

**Test-time adaptation beyond image classification.** Test-time adaptation has been applied to earlier VQA models, where debiased self-supervision counters language priors (Wen et al., 2024). For CLIP-style models, test-time prompt tuning minimizes the entropy of the prediction averaged over augmented views (Shu et al., 2022), as MEMO does for image classifiers (Zhang et al., 2022). Distillation from a frozen external model has been used for CTTA of vision models (Chen et al., 2026b). For MLLMs, test-time warmup adapts to each question with weakly supervised auxiliary tasks (Rajaneesh et al., 2025); EvoLib keeps knowledge in a library instead of the weights (Xu et al., 2026b).

**Self-improvement without labels.** Self-improvement can be described as sharpening towards what a more expensive inference procedure would select, which amortizes it (Huang et al., 2025);

the CoT target of DtC plays this role. Prolonged self-rewarded reinforcement learning with majority-vote rewards can collapse to consistent but wrong outputs (Shafayat et al., 2025), a counterpart of the single-letter collapse of naive self-training in Finding 2. Concurrent work uses the majority answer as privileged context for token-level self-distillation of language models (Gkountouras et al., 2026), and another concurrent study finds that gains of consistency-based selection for VLM test-time scaling largely disappear against a control with the same decoding format (Zheng & Kurban, 2026), in line with our comparisons against frozen models with the same inference procedure.

**Continual learning of MLLMs.** Continual instruction tuning benchmarks train MLLMs on labeled task sequences (Chen et al., 2024; Zhao et al., 2025); our setting has no labels, and Stage 2 measures what unlabeled test-time updates leave in the model.