Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Training-seed variance in single-run comparisons of frozen-CLIP adaptation: evidence from cross-dataset facial expression recognition

Zijun Jiao

2024212885@mail.hfut.edu.cn

Hefei University of Technology

Disheng Hong

Hefei University of Technology

Hongyu Liu

Hefei University of Technology

Mingyuan Li

Hefei University of Technology

Liuhao Chen

Hefei University of Technology

Research Article

Keywords: frozen vision-language models, facial expression recognition, test-time adaptation, token pooling, pre-specification, cross-dataset evaluation

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11340298/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Training-seed variance in single-run comparisons of frozen-CLIP adaptation: evidence from cross-dataset facial expression recognition

Zijun Jiao*
Hefei University of Technology, Hefei, China
ORCID 0009-0004-1717-8398

Disheng Hong
Hefei University of Technology, Hefei, China
ORCID 0009-0001-9539-9717

Hongyu Liu
Hefei University of Technology, Hefei, China
ORCID 0009-0007-7041-3758

Mingyuan Li
Hefei University of Technology, Hefei, China
ORCID 0009-0009-3158-2916

Liuhao Chen
Hefei University of Technology, Hefei, China
ORCID 0009-0001-6115-2425

Abstract

Comparisons between adaptation routes for frozen vision-language models are usually run once per arm. We ask whether the resulting contrasts are stable across training seeds for cross-dataset facial expression recognition, with matched readout capacity and frozen evaluation sets. We retrain three source-domain arms - token attention pooling, CLIP-Adapter and a parameter-matched global head - under a seed-controlled recipe (ten seeds at ViT-B/16, fifteen at ViT-L/14) and re-evaluate them on three pre-specified axes. Five of the six contrasts a single run reports are not recovered. The ViT-B/16 margin over CLIP-Adapter on the crossed grid moves from +3.88 to -1.61 points; none of those five intervals excludes zero. One contrast survives: at ViT-L/14 the token route has a higher accuracy than CLIP-Adapter on the pre-specified corruption axis, by +3.50 points (95% interval [+2.56, +4.43]), positive in all fifteen seeds and alone surviving family-wise correction. A second batch of fifteen seeds, fixed and hashed in advance, reproduces it at +3.77 points ([+2.99, +4.55]); pooled over thirty seeds, +3.63 points ([+3.06, +4.21]). Re-selecting the learning rate inside each seed moves one unstable contrast by 3.3 points, the size of its effect. Four of the six contrasts then remain unrecovered: the prospective batch also lifts the ViT-B/16 degradation margin above zero but its sign test does not confirm it, and the surviving contrast does not extend to other degradation families, a third encoder, or a degradation-specific increment on the JAFFE posed-expression corpus. Single-run comparisons in frozen-CLIP adaptation can report effects that do not replicate.

Keywords: frozen vision-language models; facial expression recognition; test-time adaptation; token pooling; pre-specification; cross-dataset evaluation

1 Introduction

Frozen vision-language encoders such as CLIP are now a default backbone for open-vocabulary recognition [19], yet their zero-shot accuracy degrades whenever the deployment domain differs from

*Corresponding author: Zijun Jiao, 2024212885@mail.hfut.edu.cn

the pre-training distribution [12]. Facial expression recognition (FER) is a demanding instance of that problem: expressions are labelled by convention rather than by appearance, and a model trained on one corpus loses accuracy on another because subjects, capture devices, illumination and annotation protocols change at once [14, 6, 1, 15, 11].

Adaptation that leaves the encoder frozen has become the practical response. Two families dominate. The first adapts the pooled global image embedding: a source-supervised linear probe, a bottleneck residual adapter such as CLIP-Adapter [5], a prior-refinement method over the CLIP cache [28], or a cache of labelled support features scored against the global vector [27]. The second operates below the pooled vector and learns how patch tokens are aggregated or attended [8]. The token route is often assumed to carry more class information than a single pooled vector, and comparisons between the two routes are reported as if a single training run settled the question.

That assumption is the object of this paper, but not in the usual way. Rather than asking which route is better, we ask whether the answer a single run produces is stable. Adaptation arms for a frozen encoder have few trainable parameters and are usually fitted once; the resulting contrasts are then reported with intervals that cover evaluation sampling but not the randomness of training itself. If the between-seed spread of a contrast is comparable to the contrast, a single run can produce a directional claim that a rerun would reverse.

We test this on a design that was built to make the comparison fair in every other respect: a parameter-matched global control, two frozen evaluation sets, two encoder scales, and a pre-specified degradation axis. We then retrain the three source-domain arms with fresh seeds, using a recipe in which the initial weights, the batch order and the learning rate all follow from the seed, and re-evaluate them on exactly the same confirmation cells. The answer is bounded and, for this literature, uncomfortable: five of the six headline contrasts fail the replication criterion fixed in Section 4 and one passes it; a prospectively fixed second batch of seeds reproduces the survivor, and pooling the two batches also clears the criterion for the ViT-B/16 degradation contrast, leaving four of the six unsupported. The between-seed spread is itself strongly encoder- and axis-dependent.

1.1 Contributions

1. **A multi-seed retraining protocol on a pre-specified comparison.** Three source-domain arms (token attention pooling, CLIP-Adapter, parameter-matched global head) are retrained from the seed alone - initial weights, batch order and learning rate - with 10 seeds at ViT-B/16 and 15 at ViT-L/14, and re-evaluated on the same 50-cell crossed grid, second-source targets and 12-condition corruption axis.

2. **A negative methodological result.** On the crossed grid the ViT-B/16 contrast between the token route and CLIP-Adapter, reported as +3.88 points by the single run, averages -1.61 points over ten retrained seeds with a between-seed standard deviation of 4.75 points, and only 2 of the 10 seeds are positive; resampling the cells inside a fixed training seed gives [-2.83, -0.37] points, so the sign reverses even with training held fixed. The ViT-L/14 cross-database contrast moves from -1.96 to -0.19 points (SD 2.42). On the second support the ViT-B/16 margin moves from +3.47 to +0.34 points (SD 3.65) and the ViT-L/14 margin from -1.37 to +0.59 points (SD 1.51). On the corruption axis the ViT-B/16 margin moves from +1.81 to +0.49 points (SD 1.18). A calibration against the original checkpoints shows that the same evaluator agrees with the single-run numbers to within 0.0002, the largest single deviation being 0.00014 on the second-source ViT-L/14 contrast, so the disagreement is not an evaluation artefact.

3. **One contrast replicates, in and out of domain.** At ViT-L/14 the token route has a higher accuracy than CLIP-Adapter on the pre-specified corruption axis, by +3.50 points

(SD 1.69, 15/15 seeds positive, 95% interval [+2.56, +4.43]) and a higher accuracy than the parameter-matched global head, by +2.73 points (SD 1.31, 14/15); it is the only one of the six contrasts that survives Holm-Bonferroni or Benjamini-Hochberg correction over the family (adjusted p = 0.0004). The same contrast was then re-tested on JAFFE, a corpus that had not been used anywhere in this work, under a hypothesis, degradation grid and pass criterion hashed before the corpus was opened: it is +2.75 points (SD 1.76, [+1.78, +3.72], 15/15 seeds), so the direction replicates out of sample. Clean controls show where the two axes differ - in domain the margin is a degradation effect (clean FER2013 test: -0.17 points), whereas out of domain it is a cross-domain accuracy advantage (clean JAFFE: +2.04 points) whose degraded-minus-clean increment on the same seeds is +0.70 points with a paired interval of [-0.38, +1.79] that includes zero. A second batch of fifteen seeds, fixed and hashed before it ran, reproduces the surviving contrast (+3.77 points, interval [+2.99, +4.55]; pooled over thirty seeds +3.63 points [+3.06, +4.21]).

4. **Two axes of instability.** Re-selecting the learning rate inside each seed moves the ViT-B/16 crossed-grid contrast from -1.61 to +1.66 points without making it significant, so the tuning convention is a second source of variance of the same order as the seed. The between-seed standard deviation of the same contrast is about twice as large at ViT-B/16 as at ViT-L/14 on the crossed grid (4.75 against 2.42 points) but larger at ViT-L/14 on the corruption axis (1.69 against 1.18), so the instability is axis-specific rather than a simple property of encoder size.

5. **Reusable evidence.** The 150 retrained arm checkpoints (three arms, ten or fifteen seeds, two learning rates), the aligned two-view arm sets, the pre-specification and its hash for the fresh corpus, per-cell records for every cell and condition, and the scripts that recompute every number in this paper from the released records are published with the paper. The whole retraining run is reproducible from its seed list: rebuilding an arm from its seed gives a bit-identical checkpoint within the same environment and hardware.

2 Related Work

Frozen-backbone adaptation of vision-language models. Adaptation of a frozen CLIP encoder [19] has converged on two design choices. The first keeps the pooled global image embedding and changes how it is read out: a source-supervised linear probe, a bottleneck residual adapter [5], a per-class linear probe that is competitive with much more elaborate adapters [10], a prior-refinement method that decouples domain-specific knowledge from the CLIP cache [28], a low-rank adapter [26], or a closed-form kernel view of the same linear family [2]. The second stores the labelled support set and scores the target against it, most simply as a training-free cache with max-affinity fusion [27], with prototype classifiers as the underlying few-shot baseline [23]. These routes are often presented as a progression from simple to sophisticated, yet the comparisons supporting that ordering are typically run at a single encoder with readouts that differ in capacity. What is missing is a comparison in which the adaptation locus and the parameter budget are matched, and in which any remaining differences between the two routes are measured and reported rather than assumed away.

Test-time adaptation and prior correction. A complementary literature adapts to the target without labels: by minimising prediction entropy [25], by estimating the shifted class prior [21, 13], or, in the broader distribution-shift setting, through the family of methods catalogued in a test-time adaptation survey [12]. Controlled evaluations that pre-specify the selection rule and report every condition remain rare: recent controlled studies benchmark TTA4CLIP methods across unified update paradigms [9] but do not compare the adaptation locus under matched readout capacity. Our

comparisons therefore carry both a training-free prior-correction arm and a support-memory arm as reference columns, so that every reported difference is measured against the stronger of the two rather than against a single baseline.

Token-level and spatial adaptation. A third line operates below the pooled vector, keeping spatial or token-level evidence and learning which parts of the image to use. Attention-guided test-time prompt tuning selects semantically meaningful regions to preserve fine-grained discriminability [8], while test-time prompt tuning adapts the text side under multi-view consistency [22]. The token-level route has become a common design in frozen-CLIP work. The assumption behind it - that per-token evidence carries more class information than a single pooled vector - is usually tested against a differently tuned global head, which confounds the adaptation locus with the readout. Whether the assumption survives a test that matches readout capacity and fixes the protocol in advance is the question this paper answers.

Cross-database facial expression recognition. FER relies on a small set of standard corpora: FER2013 [6], CK+ [14], FER+ [1] and KDEF [15], with surveys documenting how much cross-corpus variation is unrelated to expression [11]. Recent work in this journal addresses cross-domain FER directly, including a source-free setting with relational contrastive learning [7]. That literature has largely optimised accuracy under a fixed protocol; it has not treated the evaluation protocol itself as an object of study, which is where the cross-database and robustness axes can diverge.

Evaluation methodology. The final thread is methodological. pre-specification is standard practice in several empirical sciences [18] and false-discovery control is routine when many comparisons are made [3], yet frozen-CLIP adaptation benchmarks seldom pre-specify their decision rules or report a noise floor. Work on prototype blending has shown that the quantity being optimised can be the wrong target [24] - the same failure mode we probe here, approached from the comparison side rather than the objective side. We therefore report a measured seed-noise floor beside every difference and fix the decision rule before each confirmation run.

Unreliable single-run comparisons. A separate line of work has established that single-run comparisons in machine learning are unreliable and has argued for reporting score distributions rather than point estimates, first for sequence-tagging models [20] and then as general reporting guidance [4]. That literature treats the reporting protocol itself. In the work we cite, the hyper-parameter convention is not separated from the training seed, and no contrast between two adaptation routes for a frozen encoder is measured under matched capacity. The present study does both, which is why the recommendation it ends with is a pair of quantities to report rather than a new entry in an adapter leaderboard. The design reported here combines matched readout capacity with a protocol fixed in advance, seed retraining, and a re-test on a fresh corpus; we present that combination as this paper's contribution rather than as a priority claim over prior work.

3 Method

3.1 Task formulation

We study cross-dataset facial expression recognition (FER) with a frozen CLIP image encoder. Each sample is assigned one of seven expression classes, and every reported score is the unweighted average recall (UAR) over those seven classes. The classifier is fitted on a labelled source domain and deployed to targets whose labels are never used at any stage, including model and hyper-parameter selection. The source domain is FER2013 (28,709 training and 7,178 test images in total), split once into a probe-train subset of 22,968 images used to fit every trainable arm and a disjoint probe-validation subset of 5,741 images used only to select the learning rate. The same split file is used at both encoder scales. Two frozen CLIP encoders are compared: ViT-B/16 and ViT-L/14.

The question is narrow by design. Global-feature adaptation and token-level adaptation are two competing routes for a frozen encoder, and the token route is widely assumed to be stronger. We test that assumption under one readout budget and one selection protocol, at two encoder scales, so that any observed difference cannot be explained by unmatched readout capacity or by unequal tuning alone, and the residual difference is reported together with the feature-view convention that produced it.

3.2 Overview

Every arm shares three stages: a frozen CLIP encoder produces a pooled global embedding and a sequence of patch tokens; one adaptation route consumes one of the two; a small classifier maps the adapted vector to seven logits. Two routes are compared (Fig. 1). The *global-feature route* adapts the pooled embedding with a linear probe, a bottleneck residual adapter, or a multi-layer head whose parameter count is matched to the token route. The *token route* is TAP, which learns a scalar attention weight per patch token and aggregates before discriminating. A fourth, training-free reference column combines online prior correction and a support-feature cache; it is not trained and takes no part in the capacity comparison. Sections below cover the feature conventions, the three global arms, TAP's motivation, mechanism and notation, and the shared training and reference protocol.

3.3 Frozen encoder features and the two-view convention

Both encoders are used in inference mode with all parameters frozen. For ViT-B/16 the patch sequence is 196 tokens of width 768 (a 14×14 grid) and the pooled embedding is 512-dimensional; the frozen checkpoint is 350.8 MB. For ViT-L/14 the patch sequence is 256 tokens of width 1,024 and the pooled embedding is 768-dimensional; the checkpoint is 893 MB and its SHA-256 was verified against the publisher digest before use. Patch tokens are read from the encoder's visual tower with output_tokens = True, which returns the patch sequence with the class token already removed by the internal pooling step; at ViT-B/16 the resulting tokens match the pre-existing 512-dimensional feature cache to cosine 0.9999998–1.0000002 on 200 CK+ and 200 KDEF images.

The global feature cache stores **two views per image** — the image and its horizontal mirror — as unit-norm vectors, and the routed global vector is the L2-normalised mean of the two views (which averages the view axis and re-normalises). The patch-token caches, by contrast, store the **original view only**; this was a deliberate storage decision to halve disk use, and the token arms therefore train and are evaluated on the original view. We state this asymmetry explicitly because it is a protocol difference between the two routes. It also means that the comparison reported here is between each route in its native feature configuration, not a symmetric comparison in which the adaptation locus is the only variable. We therefore ran the symmetric version as a post-hoc sensitivity: a token arm retrained on the spatially aligned two-view mean of the image and its mirror. Alignment matters, because a left-right flip does not preserve patch position: the mean per-position cosine between the raw and mirrored token sequences is 0.65-0.78, rising to 0.89-0.92 after the mirror is un-flipped in the token grid. The aligned two-view token arm scored 0.6182 on Set A and 0.6344 on Set B, i.e. 6.02 and 3.78 points below the single-view token arm used for the main comparison, and on Set A it fell 2.14 points behind CLIP-Adapter with a 95% interval excluding zero. The single-view token configuration is therefore the better one for the token route, but the cross-database margin over CLIP-Adapter depends on the convention; the cross-database result is reported as configuration-specific, and the sensitivity is carried in the Results and Discussion.

3.4 Global-feature route and the capacity-matched head

Three trained arms consume only the pooled global embedding; all three take the 512-dimensional ViT-B/16 or 768-dimensional ViT-L/14 vector as input and differ in the readout.

The **linear probe** is a single Linear($d_{\text{glob}} \to 7$) layer trained on the source split; it has 3,591 trainable parameters.

CLIP-Adapter adds a bottleneck residual adapter to the global feature and classifies the adapted feature with the frozen text prototypes. With input x, it computes $f = 0.2\, f_{c2}(\text{ReLU}(f_{c1}(x))) + 0.8\, x$, L2-normalises f, and scores it against the frozen class prototypes at logit scale $s = 100$; the bottleneck ratio is 4, so $h_{\text{id}} = d_{\text{glob}}/4 = 128$ at ViT-B/16 and 192 at ViT-L/14, giving 131,712 and 295,872 trainable parameters respectively.

The **capacity-matched global head** is the control that makes the token comparison informative. It is a two-hidden-layer MLP on the pooled feature, Linear($d_{\text{glob}} \to h$) $\to$ tanh $\to$ Linear($h \to h$) $\to$ tanh $\to$ Linear($h \to 7$), and its width h is solved so that its parameter count matches TAP's. At ViT-B/16 the solution is $h = 256$, giving 198,919 parameters against TAP's 202,504 — a gap of 3,585 parameters (about 1.8%); at ViT-L/14 the solution is $h = 260$, giving 269,627 parameters against TAP's 269,832 — a gap of 205 parameters (0.076%). We therefore describe this arm as **parameter-matched** or **capacity-matched**, not exactly equal; the residual gap is larger at ViT-B/16 than at ViT-L/14 and is reported alongside every comparison.

3.5 TAP: motivation

The route tested here starts from the hypothesis that per-token evidence carries more class information than a single pooled vector, so that deciding *how* to aggregate the patch tokens is a better use of the adapter than adapting the already-pooled embedding. This is the premise behind the token-level family, and it is the premise this study puts under a matched comparison.

The obvious alternative — comparing a token module against a global feature head — does not test the premise, because any observed advantage may come from the readout: a larger head, a different nonlinearity, or a different tuning budget. The two nearest existing families both adapt the pooled embedding: global bottleneck adapters such as CLIP-Adapter, and cache/key–value adapters such as Tip-Adapter that score a support-feature cache against the global vector. TAP differs from both in one respect that can be stated without a novelty claim: adaptation happens **before** pooling, so the weighted token aggregate — not the encoder's own pooled vector — is what the classifier sees. The capacity-matched global head above exists precisely to separate "the token aggregate helps" from "a stronger readout helps"; the gap between TAP and that head is the quantity this paper reports as the net token contribution.

3.6 TAP: mechanism and notation

Let $x_1, \ldots, x_N \in \mathbb{R}^d$ be the patch tokens of one image, with $N = 196$, $d = 768$ at ViT-B/16 and $N = 256$, $d = 1024$ at ViT-L/14. TAP computes a scalar score per token with a two-layer scorer, converts the scores to a distribution over the N tokens, aggregates the tokens with that distribution, and classifies the result:

$a_1 : \mathbb{R}^d \to \mathbb{R}^h, \quad a_2 : \mathbb{R}^h \to \mathbb{R}$	scorer, $h = 256$
$u_k = a_2(\tanh(a_1 x_k))$	one scalar score per token
$w = \mathrm{softmax}_k(u), \quad \sum_k w_k = 1$	per-sample attention over the N tokens
$z = \sum_k w_k x_k$	weighted aggregate, computed before discrimination
$\mathrm{logits} = \mathrm{Linear}_{d \to 7}(z)$	linear classifier, cross-entropy loss

The scorer width is $h = 256$ at both encoders. The module has 202,504 trainable parameters at ViT-B/16 and 269,832 at ViT-L/14. Because the aggregate is a softmax-weighted mean, the pooled vector is a convex combination of the token features: the adapted representation cannot leave the convex hull of the tokens and carries no residual path back to the encoder's pooled embedding. This property is a candidate explanation for the encoder-dependent behaviour reported later and is taken up in the Discussion; it is stated here as a structural fact, not as an interpretation of the results.

3.7 Training, selection, and frozen reference arms

Training and selection protocol. Every trainable arm (TAP, CLIP-Adapter, capacity-matched head, linear probe) uses Adam, 8 epochs, a batch size of 128, and a learning-rate grid of {1e-3, 3e-3}. Fitting uses only the FER2013 probe-train split; the learning rate is chosen by UAR on the disjoint probe-validation split; no target label enters either step, and the grid is not expanded after selection. The selected values are: ViT-B/16 — TAP at 3e-3 (probe-val UAR 0.6351), CLIP-Adapter at 1e-3 (0.6172), capacity-matched head at 3e-3 (0.6144); ViT-L/14 — TAP at 1e-3 (0.6754), CLIP-Adapter at 1e-3 (0.6799), capacity-matched head at 1e-3 (0.6720).

Frozen reference arms (no training). Two label-free references define the comparison floor, and a cell's reference value is the better of the two. *PFB* applies online prediction-frequency balancing to the source linear probe: it subtracts $\min(\lambda, \mathrm{prior_clip}) \cdot \log(\hat{p})$ once at least min_prior_samples target predictions have accumulated, with $\lambda = 0.2$, prior_clip $= 1.0$, and min_prior_samples $= 32$, processing the target in seed order. *Support memory* scores each image against a labelled source support set: $s \cdot \cos(f, \mathrm{prototypes}) + \beta \cdot$ per-class-max affinity, with logit scale $s = 100$, affinity rule max, and $\beta = 256$, where the support features are the source probe-train global vectors. The max(PFB, support memory) column is carried in every table so that no trained arm is compared only against a weak internal baseline.

4 Experimental Setup

4.1 Datasets and roles

FER2013 (source). All training and hyper-parameter selection used FER2013 only: 28,709 training images and 7,178 test images. The training set was split once into a probe-train subset of 22,968 images and a disjoint probe-validation subset of 5,741. Every trainable arm was fitted on probe-train and its learning rate chosen on probe-validation; the FER2013 test split is a target, never a training source.

CK+ and KDEF (cross-database targets). CK+ contributes 902 test images and KDEF 2,938 test images. Both are used only as targets.

FER+ (a label layer, not a separate corpus). FER+ has no image manifest of its own: its images are the FER2013 test images, and the FER+ labels are a re-labelling of that same image set, so it required no separate extraction. We therefore do not describe FER+ as a fourth independent corpus.

KDEF-source (second source domain). A subject-held-out split of KDEF supplies a second labelled support source; the trainable arms themselves remain the FER2013 arms, so Set B varies the support set and the target, not the source domain of the trained arms. A subject-held-out split of KDEF serves as a second labelled source: 2,056 training images whose seven classes are near-balanced at 293–294 images each, plus a 441-image held-out test split. Support features and the prior for this source come only from its training split.

4.2 Two frozen evaluation sets

Two evaluation sets were fixed before the arms were compared, held unchanged through the reported runs, and were never used for training or for learning-rate or threshold selection. One of them carries an earlier selection history, which we state here rather than leave to a footnote: the crossed grid is the grid on which four pre-specified token modules (T1-T4) were compared before the reported arms were built, so it is a frozen evaluation set rather than an untouched confirmation set (see the order-of-use note in the next subsection). Set B was constructed afterwards and played no part in that module selection.

Set A — crossed class-prior grid (50 cells). Two target experiments, CK+ and KDEF, each crossed with a class-imbalance ratio of 0/1/2/5/10 applied to a nominated majority class, times five seeds (0–4): $2 \times 5 \times 5 = 50$ cells. Ratio 0 uses the full target and is therefore identical across the five seeds by design, which keeps the cell count at 50 exactly as in the original confirmation protocol.

Set B — second support source (15 cells). The KDEF-source split supplies the support set and the prior; the targets are (441), (7,178) and (902), each scored under five seeds (0–4): $3 \times 5 = 15$ cells. Target labels never enter training or selection in either set.

4.3 Pre-specified corruption axis

The third evaluation axis is a graded synthetic-corruption sweep of the FER2013 test set, fixed in a written pre-specification before any corrupted image, token or global feature was produced. It comprises **4 degradation families × 3 severity levels = 12 conditions, each applied to all 7,178 FER2013 test images** with unchanged labels: blur at severity 0.8 / 1.5 / 2.5; JPEG compression at quality 15 / 30 / 50; low-light at factor 0.3 / 0.5 / 0.7; and additive noise at level 10 / 25 / 40.

The degradations are generated by a single released script that defines the severity levels and uses a deterministic per-image corruption seed of $1000 + \text{row_index}$, so a given image in a given condition is reproducible from the manifest alone. The chain from the corruption script through the per-condition manifests and on to the condition caches is recorded in the repository manifest. Four legacy cache files were excluded: two are element-wise duplicates of sweep levels and two come from a different generator script, one of which routes through grayscale. No class-imbalance ratio is applied on this axis; it is a within-domain corruption axis, not a cross-database accuracy axis, and it is reported as such.

4.4 Evaluation protocol and statistics

Primary metric. Every cell is scored by unweighted average recall (UAR) over the seven classes; per-cell UAR is reported directly, and a set's summary is the mean UAR over its cells.

Two kinds of interval, kept apart. Every interval in this paper belongs to one of two units and we label them accordingly. An *evaluation interval* resamples target subsets or query units while holding the trained arm fixed, so it is conditional on that arm; these are the intervals attached to the single-run numbers and to the per-seed evaluation inside one seed. A *seed interval* resamples the

training seeds themselves and is only available from the retraining analysis; these are the intervals in the tables of Section 5 and in the prospective replication. The two are never pooled into one interval, and a contrast has an evaluation interval, a seed interval, or both, as stated in each case.

Paired shared-unit bootstrap. Differences between arms are assessed with a paired bootstrap using **2,000 draws** and a hierarchical, shared-unit design. On Sets A and B the bootstrap resamples seeds and target cells with replacement; on the corruption axis it resamples seeds (0–4) and conditions (12) with replacement. Within every sampled cell the query indices are resampled with replacement and the **same resampled indices are used for all arms**, so each arm is scored on identical units. The reported statistic is the mean over sampled cells, and intervals are 2.5th–97.5th percentiles.

What each interval covers. For Sets A and B the interval covers seed (subset sampling) and query sampling. On the corruption axis the four arms are frozen, so seeds 0–4 move only the online prior-correction order of the PFB reference; the interval therefore covers query and condition sampling only, and the axis must not be described as five independent seed repeats

What the interval does not cover. Neither the Set A/B interval nor the corruption interval covers training-seed variance: each trained arm was fitted once per encoder on the source split, and the seed axis inside the bootstrap resamples the target subset and the query units, not the training run. The single-run results are therefore scoped to the frozen trained instances that were evaluated, and the token-arm noise floor of the source diagnostic - between-seed standard deviation **0.0116 at ViT-B/16** and **0.0134 at ViT-L/14** - is retained as the magnitude criterion of the single-run decision rule. The retraining protocol described below measures the missing component directly.

Two features of the bootstrap change how its numbers should be read. First, in Set A the ratio-0 cells use the full target and are identical across the five seeds by construction, so they enter the resampling as repeated, correlated units. Treating correlated units as independent narrows an interval rather than widening it, so the Set A interval is, if anything, optimistic. The same applies to Set B, where the trained arms are frozen and their per-target predictions are identical across the five seed units, so those five units carry one training instance rather than five. Second, because the bootstrap resamples query indices within each cell, a contrast is computed on resampled query sets while UAR is an average of per-class recalls. The bootstrap mean of a contrast therefore need not equal the difference of the marginal full-set means reported for the single run, and the object of the decision rule is the interval, not the point estimate.

Noise floor. The training-variance floor is the between-seed standard deviation of the token arm, **0.0116 at ViT-B/16** and **0.0134 at ViT-L/14**, taken from the source-domain diagnostic across seeds. The retraining measures the same quantity directly and finds 0.0141 at ViT-B/16 and 0.0161 at ViT-L/14 (Table 3), so the two agree to about a fifth of a point of UAR. This floor is carried by the corruption-axis decision rule, because its bootstrap does not cover training-seed variance.

What the seed counts can resolve. At the between-seed standard deviations measured here (1.18 points at ViT-B/16 over ten seeds, 1.69 points at ViT-L/14 over fifteen), a seed-level paired test at the two-sided 0.05 level reaches 80% power for a true contrast of about 1.2 and 1.3 points respectively, computed by simulation of the paired test at those standard deviations. Contrasts smaller than that cannot be resolved at these seed counts.

4.5 Retraining protocol (this paper)

The confirmation runs reported below were each executed once per arm, which is the convention in this literature. To measure what that convention costs, we retrained the three source-domain arms with fresh seeds and re-evaluated them on exactly the same confirmation cells. Each arm is constructed and trained inside a seeded context, so its initial weights, its batch order and - in the nested condition - the learning rate it is given all follow from the seed alone, and a rebuilt run

reproduces its checkpoints bit for bit. Ten seeds (5-14) are used at ViT-B/16 and fifteen (5-19) at ViT-L/14, and every arm is trained at both points of the original learning-rate grid (1e-3 and 3e-3) so that each contrast can be reported both at the rate the original protocol selected and under selection inside the seed. No learning rate, threshold or variant was chosen on any confirmation cell.

Pre-specified subset and the seed-count extension. The decision rule was fixed before any seed was run, and the first pass used five seeds at ViT-L/14. On those five the surviving contrast was **+2.43 points** (SD 2.06, seed-level 95% interval [-0.13, +4.99], five of five seeds positive): the direction was already consistent, but five seeds cannot exclude zero here, and an exact two-sided sign test on five units cannot go below 0.0625 and therefore cannot reach the 0.05 level at that count. The seed count was raised to fifteen with no other quantity changed and with the decision taken before any of the additional seeds was trained or evaluated. The extension is outcome-aware by construction, because the direction was visible in the pre-specified subset, so the fifteen-seed interval is reported as a precision estimate rather than as a prospective test. Both counts are reported throughout, and the leave-one-seed-out analysis in Section 5 shows that no single seed drives the verdict.

The per-seed statistic reported in the Results is the mean over cells (Set A), over the three targets (Set B) or over conditions (the corruption axis) of the unweighted average recall, differenced between two arms within the same seed. Every retrained arm was scored on all three axes with the evaluation code used for the original runs. For the original single run this statistic reproduces the marginal-mean differences reported for that run exactly, and that calibration is reported first in the Results. The retraining analysis is a post-hoc variance probe: it re-uses confirmation cells that are already consumed and selects nothing on them.

Primary decision rule. Where a criterion is stated for a contrast, the seed-level interval is the primary rule: a contrast whose interval excludes zero in the expected direction is reported as replicated, and one whose interval includes zero is reported as attenuated. The exact sign test and the seed-level bootstrap are secondary diagnostics and are reported alongside it. Where a secondary diagnostic disagrees with the interval, both are stated and the disagreement is left open rather than resolved in favour of either; the interval and the sign test address different properties of the same seed distribution.

Replication criterion. We use three terms with fixed meanings. A contrast *replicates* if the mean of its per-seed differences has the same sign as the single-run value and the seed-level 95% interval excludes zero; it is *attenuated* if the mean keeps that sign but the interval includes zero; it is *reversed* if the mean has the opposite sign and the interval excludes zero; and we describe a contrast whose interval includes zero but whose point estimate has changed sign as *attenuated with a sign change*, which is what the ViT-B/16 crossed-grid contrast does here. A non-replicating contrast is therefore one whose single-run value is not a reliable estimate of the between-seed mean; an interval that includes zero is not evidence that the effect is absent, and we do not read it as such. The phrase *variance dominates* is used in that operational sense only: for the six contrasts defined here, at the two encoders studied, the between-seed spread is of the same order as the reported effect for five of them and not for the sixth.

4.6 Pre-specification and confirmation-set discipline

Table 1 lists every stage of this analysis with the date on which it was fixed and its status (pre-specified, post-hoc, outcome-aware or prospective); the text below follows that order.

The protocol for each confirmation run (hypothesis, arms, learning-rate grid, decision rule) was written to a dated file inside the project and frozen before the run. It was not deposited in a public pre-specification registry; we therefore use the term pre-specified rather than pre-registered.

Hypotheses, arms, the learning-rate grid and the decision rules were frozen before each run

and were not revised afterwards. No confirmation set may be used for model selection, learning-rate selection, threshold choice or variant design, before or after evaluation; the corruption axis in particular became a consumed confirmation set the moment its pre-specification was written.

Every comparison table must carry four columns: max(PFB, support memory), the capacity-matched global head, CLIP-Adapter, and TAP. Positive and negative findings are reported with the same detail, and a method variant designed after a confirmation result requires a fresh untouched set; the corruption-axis arms were reused from the frozen source-domain training, with no re-training or re-tuning on the new conditions

Order of use and consumed sets. Set A and Set B were frozen before the arms were compared and were each evaluated once for the crossed-grid and second-source comparisons; the corruption axis was frozen before any corrupted feature was produced and evaluated once on the corruption axis. All three are now consumed. A separate search over 13 header/fusion configurations was later screened against the same crossed protocol and is reported in full in the negative-results section; no configuration from that screen was adopted, and the arms, thresholds and decision rules reported in this paper were not revised after it. The sample-, degradation- and global-gating pilots and the RAF-DB gate are likewise reported in full and contributed no arm to the main comparison. Four token modules (T1-T4) were specified together at the start of the project, and all four were evaluated on the crossed grid; only T1 - the arm reported here as TAP - passed the pre-specified threshold, with T2, T3 and T4 reported in the negative-results section. The crossed grid therefore served both to choose among these four pre-specified modules and to carry the TAP-versus-comparator comparison, and its interval should be read with that in mind rather than as an independent confirmation. Any module search carried out after these results would need a fresh untouched set.

4.7 Reproducibility

The frozen token caches, source-domain arms and evaluation scripts are released as artifacts. Every derived cache can also be regenerated from the public corpora: each cache records its corpus, split, preprocessing and encoder checkpoint in a manifest, and the extraction entry points are deterministic functions of that manifest, so a reader who cannot obtain the caches directly can rebuild them or run the released evaluation scripts against a governed access copy. The patch-token caches hold $n \times 196 \times 768$ fp16 tensors for ViT-B/16 (about 11.7 GB across FER2013-train, FER2013-test, CK+ and KDEF) and $n \times 256 \times 1024$ fp16 tensors for ViT-L/14 (about 21.9 GB across the same sets plus the KDEF-source support), each accompanied by a small meta file holding ids and labels.

Training and evaluation entry points, per-cell result tables, bootstrap outputs and a run manifest recording commands, input hashes, exit codes and seeds are all released with the artifacts; the manifest maps each number reported below to the script and result file that produced it.

Extraction is gated on agreement with pre-existing caches rather than on the new caches alone. For ViT-B/16 the clean globals were recomputed and matched the stored FER2013-test cache at cosine 1.000000, and one corrupted condition was recomputed and matched the stored cache at cosine 1.000000 over all 7,178 rows. For ViT-L/14 there is no pre-existing corrupted cache to match against, so the gate is closed by an independent script that recomputed 512 clean FER2013-test rows (global and patch-token cosine 1.000000) and re-verified every condition cache; an independent recomputation of all 24 condition caches (12 conditions $\times$ 2 encoders) passed 24/24, with patch tokens reproducing to 0.12–1.5 fp16 ULP. The ViT-L/14 extraction therefore rests on exact agreement with the clean L/14 caches plus per-condition recomputation, without an independent degraded reference; because ViT-L/14 is the secondary encoder and does not enter the pre-specified decision, this is recorded as a lower evidence tier rather than as an equivalent gate.

Table 1: Analysis chronology and inferential status. *Fixed* means that the hypothesis, the arms, the grid and the decision rule were written to a dated file and hashed before any data for that stage were produced. The family of six is the six cross-encoder, cross-axis contrasts retained from the original protocol; the correction applied to it is reported in Section 5.6.

Stage (and when it was fixed)	Seeds and evaluation units	Primary test	Permissible claim
Crossed grid (Set A), 50 cells; fixed before retraining, with an earlier module-selection history on this grid	10 (ViT-B/16) or 15 (ViT-L/14) retrained seeds, 50 cells per seed	Seed-level t interval; family of six	Single-run value against the retrained distribution; not a causal attribution
Second support source (Set B), 3 targets; fixed before retraining, conditional on those targets	same seeds, 3 targets	Seed-level t interval; family of six	Inference conditional on those three targets
Corruption axis, 12 conditions; fixed before any corrupted feature	same seeds, 12 conditions	Seed-level t interval; family of six	Within-domain corruption only
Retraining recipe (this paper); post-hoc probe on consumed cells	10 or 15 seeds at 2 learning rates	Seed-level interval, estimation	Between-seed variability, not a confirmatory test
ViT-L/14 seed count extended 5 to 15; outcome-aware, decided before the extra seeds ran	15 seeds	Seed-level interval, precision	Precision estimate only
JAFFE axis; hashed 2026-10-08 before the corpus was opened	15 seeds, 20 conditions	Seed-level interval	Direction of the accuracy advantage; no degradation-specific increment
FER+ relabelling; fresh degradation families; hashed before running	10 or 15 seeds	Seed-level interval	Annotation and family dependence
Third encoder (LAION-2B ViT-B/32); fixed before running	10 seeds	Seed-level interval	Absence of the effect at that encoder
Prospective batch, seeds 20-34 (ViT-L/14) and 15-24 (ViT-B/16); hashed 2026-10-09	15 and 10 fresh seeds	Seed-level interval against a criterion frozen in advance	Reproducibility of the seed distribution on the fixed axes

5 Results

Per-seed records appear in Tables 3 to 7; the two extension axes, the summary across axes and the two sensitivity conditions follow in Tables 8, 9, 10 and 11–12.

5.1 Calibration: the single run is reproduced exactly

Before comparing runs we verified that the retraining evaluator returns the originally reported numbers when it is given the original seed-0 arms. On the 50 crossed-grid cells it returns TAP 0.67837, CLIP-Adapter 0.63961 and capacity-matched head 0.61583 at ViT-B/16, i.e. contrasts of +0.03876 and +0.06253 against the two comparators, and at ViT-L/14 TAP 0.757028, CLIP-Adapter 0.776653 and capacity-matched head 0.768073, i.e. -0.01963 and -0.01104. On the 12 corruption conditions it returns 0.47402, 0.45588 and 0.45227 at ViT-B/16 (+0.01814, +0.02176) and 0.536038, 0.504038 and 0.519003 at ViT-L/14 (+0.03200, +0.01704). On the three Set B targets it returns 0.672229, 0.637543 and 0.620185 at ViT-B/16 (+0.03469, +0.05204) and 0.740481, 0.754325 and 0.741203 at ViT-L/14 (-0.01384, -0.00072). Every contrast agrees with the value reported for the single run to within 0.0002; the largest single deviation is 0.00014, on the second-support ViT-L/14 contrast (Table 2). Two of the Set B entries differ in the last digit shown because the manuscript reports the bootstrap mean there while this table reports the difference of marginal means, and the two capacity-matched entries in the last column were not reported for the single run. The

Table 2: Calibration of the retraining evaluator against the single run (seed 0).

encoder	axis	contrast	this evaluator	single run
ViT-B/16	Set A (50 cells)	TAP − CLIP-Adapter	+0.0388	+0.0388
ViT-B/16	Set A (50 cells)	TAP − capacity-matched head	+0.0625	+0.0625
ViT-B/16	corruption (12 conditions)	TAP − CLIP-Adapter	+0.0181	+0.0181
ViT-B/16	corruption (12 conditions)	TAP − capacity-matched head	+0.0218	—
ViT-L/14	Set A (50 cells)	TAP − CLIP-Adapter	−0.0196	−0.0196
ViT-L/14	Set A (50 cells)	TAP − capacity-matched head	−0.0110	−0.0110
ViT-L/14	corruption (12 conditions)	TAP − CLIP-Adapter	+0.0320	+0.0320
ViT-L/14	corruption (12 conditions)	TAP − capacity-matched head	+0.0170	—
ViT-B/16	Set B (3 targets)	TAP − CLIP-Adapter	+0.0347	+0.0347
ViT-B/16	Set B (3 targets)	TAP − capacity-matched head	+0.05204	+0.0521
ViT-L/14	Set B (3 targets)	TAP − CLIP-Adapter	−0.01384	−0.0137
ViT-L/14	Set B (3 targets)	TAP − capacity-matched head	−0.0007	−0.0007

disagreement reported in the rest of this section is therefore a property of training, not of evaluation.

5.2 Retraining is reproducible from the seed list

The retraining recipe is fixed and fully determined by the seed: the initial weights, the batch order and, for the nested arm below, the choice of learning rate all follow from it, so a rebuilt run reproduces its checkpoint bit for bit within the same environment and hardware. We check that directly by rebuilding one arm twice inside the same run, and the two copies are identical. This matters for the present argument, because an earlier pass of this analysis left the initial weights to the process default and therefore did not reproduce its own token-arm checkpoints; every number below comes from the recipe that does. Every arm is trained at both grid points (1e-3 and 3e-3) so that each contrast can be reported both at the learning rate the original protocol selected and under selection inside the seed.

Table 3: Source probe-validation accuracy per training seed, at the rate the original protocol selected, with the rate each arm's own probe split prefers in the nested condition.

encoder	seed	TAP	CLIP-Adapter	capacity-matched head	nested rate TAP / CA / GH
ViT-B/16	5	0.5945	0.6062	0.6157	1e-3 / 3e-3 / 3e-3
ViT-B/16	6	0.6277	0.6347	0.6419	3e-3 / 1e-3 / 3e-3
ViT-B/16	7	0.6281	0.6394	0.5935	3e-3 / 1e-3 / 1e-3
ViT-B/16	8	0.6097	0.6319	0.5929	1e-3 / 1e-3 / 3e-3
ViT-B/16	9	0.6243	0.6333	0.6262	1e-3 / 3e-3 / 1e-3
ViT-B/16	10	0.6057	0.6039	0.5730	3e-3 / 3e-3 / 1e-3
ViT-B/16	11	0.6260	0.6422	0.6391	1e-3 / 3e-3 / 3e-3
ViT-B/16	12	0.5989	0.6088	0.6186	1e-3 / 1e-3 / 3e-3
ViT-B/16	13	0.6361	0.6168	0.6321	3e-3 / 3e-3 / 3e-3
ViT-B/16	14	0.6118	0.6192	0.6273	1e-3 / 3e-3 / 3e-3
ViT-L/14	5	0.6413	0.6601	0.6424	1e-3 / 1e-3 / 3e-3
ViT-L/14	6	0.6655	0.6700	0.6725	1e-3 / 3e-3 / 3e-3
ViT-L/14	7	0.6534	0.6809	0.6726	3e-3 / 1e-3 / 1e-3
ViT-L/14	8	0.6604	0.6720	0.6467	1e-3 / 1e-3 / 3e-3
ViT-L/14	9	0.6922	0.6840	0.6753	1e-3 / 1e-3 / 1e-3
ViT-L/14	10	0.6488	0.6588	0.6372	1e-3 / 3e-3 / 1e-3
ViT-L/14	11	0.6956	0.6901	0.6771	1e-3 / 1e-3 / 1e-3
ViT-L/14	12	0.6526	0.6550	0.6468	1e-3 / 1e-3 / 1e-3
ViT-L/14	13	0.6776	0.6459	0.6556	1e-3 / 3e-3 / 3e-3
ViT-L/14	14	0.6842	0.6751	0.6670	1e-3 / 3e-3 / 3e-3
ViT-L/14	15	0.6661	0.6633	0.6699	1e-3 / 3e-3 / 1e-3
ViT-L/14	16	0.6559	0.6683	0.6646	3e-3 / 1e-3 / 1e-3
ViT-L/14	17	0.6587	0.6791	0.6687	3e-3 / 3e-3 / 1e-3
ViT-L/14	18	0.6639	0.6679	0.6641	1e-3 / 1e-3 / 3e-3
ViT-L/14	19	0.6799	0.6682	0.6628	1e-3 / 3e-3 / 3e-3

5.3 Source-domain training variance

The three arms do not separate on the source split. At ViT-B/16 the token arm's probe-validation UAR ranges from 0.5945 to 0.6361 (mean 0.6163, SD 0.0141), CLIP-Adapter from 0.6039 to 0.6422 (0.6236, 0.0144) and the capacity-matched head from 0.5730 to 0.6419 (0.6160, 0.0226). At ViT-L/14 the corresponding ranges are 0.6413-0.6956 (0.6664, 0.0161), 0.6459-0.6901 (0.6693, 0.0117) and 0.6372-0.6771 (0.6616, 0.0127). The seed-to-seed spread of a single arm is therefore 1.2 to 1.6 points of UAR at both scales, which is the same order as the contrasts the confirmation cells are asked to resolve.

5.4 Cross-database contrast on the crossed grid does not replicate

The contrast that the single run reports as +3.88 points at ViT-B/16 (Fig. 2a) averages **-1.61 points** over the ten retrained seeds, with a between-seed standard deviation of 4.75 points, and only 2 of the 10 seeds are positive; the interval over seeds is [-5.00, +1.79] points. The point estimate has changed sign and the interval covers zero. Resampling the 50 crossed cells while holding one training seed fixed gives [-2.83, -0.37] points. That is a conditional statement about scoring units given a single trained arm set, not a between-seed result, and we read it only as showing that the negative sign is not an artefact of averaging over seeds: with the training run held fixed the token route is worse than CLIP-Adapter on this grid. At ViT-L/14 the contrast the single run reports as -1.96 points averages -0.19 points over fifteen seeds (SD 2.42, 6/15 positive, interval [-1.53, +1.15]).

Table 4: Crossed grid (50 cells), per training seed.

encoder	seed	TAP	CLIP-Adapter	matched head	TAP - CLIP-Adapter (pp)	TAP - matched head (pp)
ViT-B/16	5	0.5773	0.6430	0.6430	-6.57	-6.57
ViT-B/16	6	0.6315	0.6378	0.6439	-0.63	-1.24
ViT-B/16	7	0.5744	0.6422	0.6487	-6.78	-7.43
ViT-B/16	8	0.6335	0.6496	0.6352	-1.61	-0.17
ViT-B/16	9	0.6130	0.6337	0.6196	-2.07	-0.66
ViT-B/16	10	0.5797	0.6314	0.6206	-5.17	-4.09
ViT-B/16	11	0.7000	0.6411	0.6420	+5.88	+5.79
ViT-B/16	12	0.6253	0.6474	0.6537	-2.20	-2.84
ViT-B/16	13	0.7086	0.6379	0.6394	+7.07	+6.92
ViT-B/16	14	0.5993	0.6394	0.6525	-4.01	-5.32
ViT-L/14	5	0.7649	0.8039	0.7800	-3.89	-1.50
ViT-L/14	6	0.7796	0.7983	0.7657	-1.88	+1.39
ViT-L/14	7	0.7913	0.7905	0.7678	+0.08	+2.35
ViT-L/14	8	0.7815	0.7888	0.7683	-0.73	+1.33
ViT-L/14	9	0.8109	0.7853	0.7522	+2.56	+5.87
ViT-L/14	10	0.7712	0.7857	0.7788	-1.45	-0.76
ViT-L/14	11	0.8254	0.7850	0.7679	+4.04	+5.74
ViT-L/14	12	0.7583	0.7750	0.7516	-1.67	+0.67
ViT-L/14	13	0.8204	0.7773	0.7811	+4.31	+3.92
ViT-L/14	14	0.8129	0.7966	0.7811	+1.63	+3.18
ViT-L/14	15	0.7533	0.7737	0.7618	-2.04	-0.84
ViT-L/14	16	0.8070	0.7950	0.7813	+1.20	+2.57
ViT-L/14	17	0.7906	0.8012	0.7776	-1.06	+1.29
ViT-L/14	18	0.7699	0.7901	0.7696	-2.03	+0.03
ViT-L/14	19	0.7801	0.7990	0.7864	-1.88	-0.63

5.5 The second-support contrast does not replicate either

On the second support source the two margins move from +3.47 to **+0.34 points** at ViT-B/16 (SD 3.65, 3/10 positive, interval [-2.27, +2.95]) and from -1.37 to **+0.59 points** at ViT-L/14 (SD 1.51, 9/15 positive, interval [-0.25, +1.42]). Neither is distinguishable from zero. Because the trained arms do not depend on the support set, the five cell seeds of a given target carry identical arm predictions, so the per-seed statistic is the mean over the three targets; a three-unit resample is not interpretable as a cell-level interval and we do not report one for this axis. Every interval we quote for it is therefore conditional on those three targets as well as on the training seeds, and is not an estimate of uncertainty over a target population.

5.6 The degradation contrast replicates at the larger encoder in the first batch, and only there

At ViT-B/16 the contrast that the single run reports as +1.81 points averages **+0.49 points** over ten seeds (SD 1.18, 6/10 positive, interval [-0.35, +1.33]). At ViT-L/14 the same contrast averages **+3.50 points** over fifteen seeds with a standard deviation of 1.69 points; all fifteen seeds are positive, the smallest per-seed margin is 0.0 points and the largest 6.1 points, and the interval over seeds is [+2.56, +4.43] points. It is the only one of the six contrasts whose interval lies entirely above zero in this batch. On the five seeds of the first, pre-specified pass the same contrast was +2.43 points (SD 2.06, interval [-0.13, +4.99], five of five seeds positive): the direction was already consistent at that count, but five seeds cannot exclude zero here and the exact sign test cannot reach 0.05 with five units. The fifteen-seed figure is the powered estimate; Section 4.5 records how the extension was decided. A second batch of fifteen seeds (seeds 20-34) was then fixed, hashed and run prospectively with the recipe and the axes unchanged: it gives +3.77 points (SD 1.41, interval [+2.99, +4.55],

Table 5: Second support source (three targets), per training seed.

encoder	seed	TAP	CLIP-Adapter	matched head	TAP - CLIP-Adapter (pp)	TAP - matched head (pp)
ViT-B/16	5	0.6047	0.6253	0.6326	-2.06	-2.79
ViT-B/16	6	0.6474	0.6358	0.6490	+1.16	-0.17
ViT-B/16	7	0.5921	0.6333	0.6269	-4.11	-3.48
ViT-B/16	8	0.6424	0.6431	0.6236	-0.07	+1.88
ViT-B/16	9	0.6269	0.6272	0.6217	-0.03	+0.52
ViT-B/16	10	0.5957	0.6220	0.6039	-2.63	-0.82
ViT-B/16	11	0.6967	0.6376	0.6381	+5.90	+5.86
ViT-B/16	12	0.6228	0.6368	0.6422	-1.40	-1.94
ViT-B/16	13	0.6972	0.6237	0.6339	+7.35	+6.33
ViT-B/16	14	0.6252	0.6326	0.6423	-0.73	-1.71
ViT-L/14	5	0.7477	0.7591	0.7421	-1.14	+0.55
ViT-L/14	6	0.7697	0.7694	0.7430	+0.03	+2.67
ViT-L/14	7	0.7519	0.7672	0.7436	-1.53	+0.83
ViT-L/14	8	0.7570	0.7558	0.7348	+0.12	+2.22
ViT-L/14	9	0.7786	0.7590	0.7349	+1.96	+4.38
ViT-L/14	10	0.7580	0.7464	0.7417	+1.17	+1.63
ViT-L/14	11	0.7900	0.7569	0.7431	+3.31	+4.69
ViT-L/14	12	0.7478	0.7421	0.7247	+0.57	+2.32
ViT-L/14	13	0.7836	0.7496	0.7485	+3.40	+3.51
ViT-L/14	14	0.7845	0.7678	0.7492	+1.66	+3.53
ViT-L/14	15	0.7441	0.7450	0.7337	-0.09	+1.04
ViT-L/14	16	0.7769	0.7654	0.7474	+1.15	+2.94
ViT-L/14	17	0.7698	0.7723	0.7470	-0.25	+2.28
ViT-L/14	18	0.7428	0.7512	0.7380	-0.83	+0.48
ViT-L/14	19	0.7628	0.7699	0.7510	-0.71	+1.18

fifteen of fifteen seeds positive), and pooled over the thirty seeds the contrast is +3.63 points [+3.06, +4.21]. The criterion fixed and hashed in advance was met; the subsection below reports it. The margin against the capacity-matched head behaves differently: at ViT-L/14 it is positive on average on all three axes (+1.64 points on the crossed grid, +2.28 on the second support, +2.73 on the corruption axis), whereas the single run reported -1.10 and -0.07 points on the two cross-database sets. CLIP-Adapter is the comparator of record because it is the comparator the original pre-specified comparison used, fixed before any of the retrained numbers existed; the matched-head contrast is reported alongside it and is not used to choose the family, and the six contrasts were fixed by the original three axes crossed with the two encoders rather than selected from the retraining outcome.

5.7 Prospective replication of the seed distribution

The objection this subsection answers is that the fifteen-seed estimate was produced after the direction was already visible in the pre-specified five, so its interval is a precision estimate rather than a prospective test. To answer that with evidence, a second batch of seeds was fixed, written down and hashed before any of it was trained or evaluated (ViT-L/14 seeds 20-34, ViT-B/16 seeds 15-24), with the recipe, the arms, the two learning rates, the source split and the evaluation code all unchanged, and with the criterion fixed in the same document: for the ViT-L/14 corruption axis at the published rate, TAP minus CLIP-Adapter must have a positive batch mean and a seed-level 95% interval that excludes zero.

The criterion was met. Batch 1 gives +3.50 points (SD 1.69, [+2.56, +4.43], 15 of 15 seeds positive); batch 2, on seeds it had never seen, gives **+3.77 points** (SD 1.41, [+2.99, +4.55], 15 of 15); pooled over the thirty seeds the estimate is +3.63 points [+3.06, +4.21]. The two batch means differ by 0.27 points, less than either batch's own standard deviation, and the margin against the

Table 6: Corruption axis (12 pre-specified conditions), per training seed.

encoder	seed	TAP	CLIP-Adapter	matched head	TAP - CLIP-Adapter (pp)	TAP - matched head (pp)
ViT-B/16	5	0.4564	0.4413	0.4523	+1.51	+0.40
ViT-B/16	6	0.4544	0.4441	0.4627	+1.03	-0.83
ViT-B/16	7	0.4569	0.4604	0.4424	-0.35	+1.45
ViT-B/16	8	0.4470	0.4511	0.4428	-0.41	+0.42
ViT-B/16	9	0.4638	0.4630	0.4580	+0.09	+0.58
ViT-B/16	10	0.4419	0.4397	0.4241	+0.23	+1.79
ViT-B/16	11	0.4702	0.4679	0.4662	+0.24	+0.40
ViT-B/16	12	0.4458	0.4518	0.4511	-0.60	-0.54
ViT-B/16	13	0.4706	0.4377	0.4466	+3.28	+2.39
ViT-B/16	14	0.4546	0.4555	0.4520	-0.09	+0.27
ViT-L/14	5	0.5199	0.5013	0.5006	+1.87	+1.93
ViT-L/14	6	0.5136	0.5136	0.5146	+0.00	-0.10
ViT-L/14	7	0.5335	0.5198	0.5176	+1.38	+1.59
ViT-L/14	8	0.5330	0.4973	0.4984	+3.58	+3.46
ViT-L/14	9	0.5616	0.5083	0.5221	+5.33	+3.95
ViT-L/14	10	0.5233	0.4884	0.4965	+3.49	+2.68
ViT-L/14	11	0.5650	0.5137	0.5217	+5.13	+4.33
ViT-L/14	12	0.5224	0.5026	0.5036	+1.98	+1.89
ViT-L/14	13	0.5452	0.4837	0.5011	+6.15	+4.41
ViT-L/14	14	0.5534	0.5175	0.5195	+3.59	+3.39
ViT-L/14	15	0.5314	0.4915	0.5189	+3.99	+1.25
ViT-L/14	16	0.5461	0.5086	0.5140	+3.75	+3.21
ViT-L/14	17	0.5287	0.5022	0.5103	+2.66	+1.84
ViT-L/14	18	0.5448	0.5034	0.5160	+4.14	+2.88
ViT-L/14	19	0.5499	0.4956	0.5079	+5.43	+4.21

capacity-matched head replicates in the same way (batch 1 +2.73, batch 2 +2.65; pooled +2.69 [+2.21, +3.18]).

The negative results replicate with equal fidelity, which is the point of the exercise. On the crossed grid at ViT-L/14 the TAP minus CLIP-Adapter contrast is -0.19 points in batch 1 and -0.21 in batch 2, both intervals covering zero ([-2.04, +1.62] in batch 2); on the second support it is +0.59 and +0.61 points, again both covering zero ([-0.55, +1.77]). The instability documented above is therefore a property of the contrast rather than of one seed draw.

One secondary result changes with twenty seeds, and it is reported rather than buried. At ViT-B/16 the corruption-axis contrast at the published rate was +0.49 points ([-0.35, +1.33]) over the first ten seeds; the prospective batch gives +1.08 points ([+0.11, +2.05]) over seeds 15-24, and pooling the twenty seeds gives +0.79 points ([+0.19, +1.38]), which excludes zero, with an exact sign test of $p = 0.2632$ over the twenty seeds. The earlier text treated this margin as established only under rate re-selection; with the second batch the published-rate margin also excludes zero. It remains roughly a fifth of the ViT-L/14 margin, and the encoder ordering is unchanged. This contrast is one of the six; what the prospective batch changes is its verdict, from attenuated in batch 1 to an interval that excludes zero, while its exact sign test remains non-significant. We therefore report it as additional estimation evidence for the same encoder-axis contrast rather than as a second multiplicity-controlled discovery: the family-wise correction of Section 5.6 was applied to the first batch, and we do not claim a corrected confirmation here.

The protocol file, fixed and hashed in advance (sha256 1f0554921efe8c3d…), its freeze record and the per-seed records of both batches are archived with the release, and the full digest is in the release manifest. Because the three evaluation axes were already consumed, this replication upgrades the inferential status of the existing result; it is not a new confirmation set and is not presented as one.

Table 7: The surviving contrast on the fresh JAFFE axis, with the clean controls that determine how it should be read.

data	encoder	condition	TAP - CLIP-Adapter	TAP - matched head	seeds positive (TAP - CLIP-Adapter)
JAFFE, corrupted (pre-specified axis)	ViT-L/14	rate fixed	+2.75 (SD 1.76)	+3.75 (SD 1.86)	15/15
JAFFE, corrupted (pre-specified axis)	ViT-L/14	rate re-selected	+3.20 (SD 1.63)	+3.59 (SD 1.78)	14/15
JAFFE, corrupted (pre-specified axis)	ViT-B/16	rate fixed	-1.72 (SD 1.70)	-1.54 (SD 1.26)	3/10
JAFFE, clean (control)	ViT-L/14	rate fixed	+2.04 (SD 3.30)	+3.58 (SD 4.47)	11/15
JAFFE, clean (control)	ViT-B/16	rate fixed	-4.61 (SD 4.63)	-5.80 (SD 3.06)	1/10
FER2013 test, corrupted (the corruption axis of Section 5)	ViT-L/14	rate fixed	+3.50 (SD 1.69)	+2.73 (SD 1.31)	15/15
FER2013 test, clean (control)	ViT-L/14	rate fixed	-0.17 (SD 1.16)	+1.00 (SD 1.36)	7/15
FER2013 test, clean (control)	ViT-B/16	rate fixed	+0.40 (SD 1.35)	+1.05 (SD 2.09)	6/10

5.8 The surviving contrast on a fresh corpus

The degradation result at ViT-L/14 is the one claim this study leaves standing, so it was tested once more on a corpus that had not been used anywhere in this work: JAFFE [17], used here through the archived copy at `https://doi.org/10.5281/zenodo.14974867` [16], 213 images from ten models in seven expressions. The hypothesis, the preprocessing (grayscale, resampled to the 48x48 native resolution of FER2013), the 20-condition degradation grid (four families x five severities, the original twelve extended by one step in each direction) and the pass criterion were written and hashed **before** the corpus was opened. The amendment corrects a factual claim about the corpus structure that the first version got wrong; the hypothesis, the grid, the arms and the decision rule are unchanged by it. The source-domain arms were re-used unchanged, so nothing was trained, tuned or selected on the new axis.

Chronology. The archive was fetched on 2026-10-04 and its sha256 and md5 were recorded at that moment, in an acquisition note that already described it as a candidate untouched confirmation corpus and stated that it had not been extracted, trained on, tuned or evaluated. No image was opened until after the hypothesis file had been written and hashed on 2026-10-08 at 12:23:35, a timestamp carried in the freeze record, and after amendment 1 at 12:24:12; image extraction began after both, and the modification times of the extracted images and manifests are later than both hashes. Before the freeze the only JAFFE information in existence was the archive hash and the archive's file listing. We state this explicitly because the archive date and the protocol date differ and the difference is otherwise easy to misread.

The pre-specified verdict on this axis is confirmed and strengthened, subject to the accuracy-versus-robustness reading given below. On the fresh degraded axis the ViT-L/14 margin over CLIP-Adapter is +2.75 points with a seed-level 95% interval of [+1.78, +3.72], positive in all fifteen seeds, and the margin over the capacity-matched head is +3.75 points with fourteen of fifteen seeds positive. Under rate re-selection inside the seed the same contrast is +3.20 points [+2.30, +4.10]. The direction of the surviving claim therefore replicates on data that were never touched, under a criterion that was fixed in advance.

The clean rows in the same table are what keep that result from being over-read. On clean JAFFE

Table 8: Two further pre-specified axes. Contrasts are TAP minus the comparator; positive means the token route is better.

axis	encoder	TAP - CLIP-Adapter	TAP - matched head	seeds positive	verdict
B: FER+ annotation	ViT-L/14	+4.38 (SD 1.94)	+2.77 (SD 1.34)	15/15	confirmed
B: FER+ annotation	ViT-L/14 (clean control)	-0.32 (SD 1.85)	not reported	6/15	control
B: FER+ annotation	ViT-B/16	-0.52 (SD 1.09)	-0.26 (SD 0.67)	3/10	not confirmed
C: fresh families	ViT-L/14	+0.75 (SD 1.36)	+1.58 (SD 1.50)	11/15	not confirmed (interval touches zero)
C: fresh families	ViT-B/16	+1.67 (SD 1.68)	+2.64 (SD 1.45)	7/10	confirmed
D: third encoder, pre-specified families	ViT-B/32 (LAION-2B)	-0.83 (SD 1.30)	+0.52 (SD 1.19)	2/10	not confirmed
D: third encoder, fresh families	ViT-B/32 (LAION-2B)	-0.23 (SD 1.84)	+1.10 (SD 1.85)	4/10	not confirmed

the ViT-L/14 margin is already +2.04 points with an interval that excludes zero, so most of the fresh-axis advantage is a cross-domain accuracy advantage rather than one specific to the corruption, and the degraded-minus-clean increment on the same seeds is +0.70 points with a paired interval of [-0.38, +1.79] that includes zero, so on this corpus the advantage is a cross-domain accuracy advantage with no degradation-specific component that fifteen seeds can resolve. On the original within-domain axis the decomposition is the other way round: on clean FER2013 test the same arms differ by -0.17 points with an interval covering zero, and it is the pre-specified corruption that produces the +3.50 points. The compact reading is that at ViT-L/14 the token route's margin over CLIP-Adapter is a degradation effect in domain and a domain effect out of domain, with the out-of-domain increment not distinguishable from zero, and that the direction replicates in both cases. The encoder dependence replicates as well, and more sharply than on FER2013: at ViT-B/16 the token route is worse than CLIP-Adapter on the fresh degraded axis (-1.72 points, interval excluding zero) and worse still on clean JAFFE (-4.61 points). JAFFE is consumed by this analysis and is not used for anything else.

5.9 Two further pre-specified axes

Everything above varies something inside the same pipeline: the training seed, the learning-rate convention and the feature-view convention. Two further pre-specified axes change something outside it, and both were written into a hashed pre-specification before they were run; neither introduces new images, and the relabelling axis reuses the FER2013 test images as stated in Section 4.1. Nothing was trained, tuned or selected on either axis; the frozen arms of the retraining analysis were re-scored.

Axis B changes the annotation. FER+ is a relabelling of the same images, and this study had explicitly not used it as an axis. Labels come from the FER+ vote vectors with a rule fixed in advance: keep an image when one of the eight emotion columns is strictly the largest and is not contempt, and map the rest onto the seven classes used throughout. 6,796 of the 7,178 test images survive that rule.

Axis C changes the degradation. Three families that are absent from the pre-specified grid - occlusion (a centred black square of side 8, 16 or 24 pixels), contrast (factors 0.3, 0.5, 0.7) and rotation (10, 20, 30 degrees) - give nine further conditions on the same images and labels.

Axis B confirms the surviving contrast under a different annotation layer: +4.38 points over

CLIP-Adapter with an interval of [+3.31, +5.46] and fifteen of fifteen seeds positive, while the clean control under the same labels is -0.32 points, so the decomposition of the original axis survives the relabelling too. At ViT-B/16 the same labels give -0.52 points, so the encoder dependence survives it as well.

Axis C does not confirm it, and the way it fails is informative. At ViT-L/14 the contrast is +0.75 points with an interval of [-0.00, +1.51], which touches zero: by the rule fixed in advance this is **not confirmed**, and the margin is roughly a fifth of the one the original families produce. At ViT-B/16 the same nine conditions give +1.67 points with an interval of [+0.47, +2.87], which excludes zero. The ordering of the two encoders therefore inverts with respect to the original axis. Two conclusions follow, and we state both: the surviving effect does not extend to arbitrary degradation families, it is a property of the families that were pre-specified; and the sentence "the effect lives at the larger encoder" is itself axis-dependent rather than a stable property of the encoder, which is the same lesson this paper is making about single comparisons, applied to our own result.

Axis D adds a third encoder. The two encoders above are both OpenAI models; a third one was added that differs in architecture and in pretraining data at the same time: LAION-2B ViT-B/32, whose checkpoint was taken from ModelScope with its published sha256 verified before use. Its patch grid is 7x7 rather than the 14x14 or 16x16 of the others, and it is smaller than both. The three arms were re-trained for it with the same seed-controlled recipe on the FER2013 probe-train split, with that encoder's own global features and its own text prototypes, since mixing an encoder's token route with another encoder's global route would not be a test of anything. The result is negative and is reported as such: on the pre-specified families the contrast is -0.83 points with an interval of [-1.76, +0.10] and two of ten seeds positive, on the fresh families it is -0.23 points, and on clean data it is -1.45 points with an interval that excludes zero in the negative direction. The token route therefore has no advantage at this encoder, on either axis. Read together with the two OpenAI encoders, the honest statement is that the degradation-axis effect is largest at ViT-L/14, is also above zero at ViT-B/16 once the prospective batch is pooled over twenty seeds, and does not appear at this third encoder at all: it is encoder-specific rather than a general property of the token route, which is what the rest of this paper argues about single comparisons.

A third planned axis - re-extracting the tokens with an independent implementation of the encoder - could not be executed: the two routes to an independent implementation were blocked (the model hub was unreachable from this machine, and the only locally cached copy of the weights turned out to be the same implementation already in use). It is recorded as not executed rather than as a failure, and no number is reported from it.

5.10 What the single run would have reported

The intervals in Table 9 are ordinary two-sided t intervals on the per-seed differences. Five of the six straddle zero in this batch, and the crossed-grid contrast at ViT-B/16 - the one a single run reports as its headline positive - has moved to a point estimate of the opposite sign. With the prospective second batch of Section 5.7 the count falls to four: the ViT-B/16 degradation contrast has a pooled interval of [+0.19, +1.38] over twenty seeds, although its exact sign test over those seeds gives $p = 0.26$, so the interval and the sign test disagree and both are reported.

5.11 Does the rate-selection convention matter as much as the seed?

The choice is not degenerate: at ViT-L/14 the token arm prefers 1e-3 in 12 of the 15 seeds, CLIP-Adapter in 9 of 15 and the capacity-matched head in 6 of 15; at ViT-B/16 the token arm prefers 1e-3 in 6 of 10. Re-selecting the rate inside the seed moves the ViT-B/16 crossed-grid contrast from -1.61

Table 9: Summary across the three confirmation axes and two encoder scales. The single-run column is the number a one-run study would report. The retrained mean, SD and interval describe the distribution over the retrained seeds at the rate the original protocol selected; the prospective row uses the second, separately frozen batch of fifteen ViT-L/14 seeds, and its intervals are computed within that batch.

axis	encoder	seeds	single run (pp)	retrained mean (pp)	SD (pp)	95% interval over seeds (pp)	sign-test p	verdict
crossed grid	ViT-B/16	10	+3.88	**-1.61**	4.75	[-5.00, +1.79]	0.1094	attenuated, sign changed
crossed grid	ViT-L/14	15	-1.96	**-0.19**	2.42	[-1.53, +1.15]	0.6072	attenuated
second support	ViT-B/16	10	+3.47	**+0.34**	3.65	[-2.27, +2.95]	0.3438	attenuated
second support	ViT-L/14	15	-1.37	**+0.59**	1.51	[-0.25, +1.42]	0.6072	attenuated, sign changed
corruption	ViT-B/16	10	+1.81	**+0.49**	1.18	[-0.35, +1.33]	0.7539	attenuated
corruption	ViT-L/14	15	+3.20	**+3.50**	1.69	[+2.56, +4.43]	0.0001	replicates
corruption	ViT-L/14	15 (prospective)	+3.20	**+3.77**	1.41	[+2.99, +4.55]	0.0001	replicates

Table 10: The contrast under a fixed learning rate and under selection inside the seed. Each arm is trained at both grid points and the seed's own probe-validation split chooses; the contrast is then formed from that seed's selected arms.

axis	encoder	rate fixed (95% interval)	rate re-selected inside the seed (95% interval)	change
crossed grid	ViT-B/16	-1.6076 [-5.0035, +1.7882]	+1.6569 [-1.7033, +5.0170]	+3.2645
crossed grid	ViT-L/14	-0.1870 [-1.5281, +1.1542]	-0.2256 [-1.8430, +1.3917]	-0.0387
second support	ViT-B/16	+0.3378 [-2.2703, +2.9459]	+2.4434 [-0.1342, +5.0209]	+2.1056
second support	ViT-L/14	+0.5880 [-0.2474, +1.4233]	+0.6514 [-0.3962, +1.6989]	+0.0634
corruption	ViT-B/16	+0.4915 [-0.3504, +1.3334]	+0.9401 [+0.0254, +1.8549]	+0.4487
corruption	ViT-L/14	+3.4976 [+2.5627, +4.4324]	+3.6783 [+2.9554, +4.4012]	+0.1807

to +1.66 points, a swing of 3.27 points that is the size of the effect the single run reported there, without making it significant (interval [-1.70, +5.02]); on the second support the same manipulation moves the ViT-B/16 margin from +0.34 to +2.44 points. At ViT-L/14 the one surviving contrast is unaffected: +3.50 points with the rate fixed and +3.68 points with it re-selected, both intervals excluding zero. The convention used for the learning rate is therefore a second axis of unreliability, of the same order as the training seed for the contrasts that are already unstable at ViT-B/16, and it neither creates nor removes the effect at ViT-L/14. The only other contrast whose interval excludes zero under any convention is the ViT-B/16 degradation margin under re-selection, +0.94 points (interval [+0.03, +1.85]); its sign test gives p = 0.109 and the unit-level resample includes zero ([-0.44, +2.43]), so we do not treat it as an established effect on the first batch. With the prospective second batch the published-rate margin also excludes zero at ViT-B/16 (+1.08 points, [+0.11, +2.05] over seeds 15-24; +0.79 points, [+0.19, +1.38] over the twenty seeds pooled); that margin is about a fifth of the ViT-L/14 one and is reported with the batch-2 numbers in Section 5.7.

Non-parametric checks and multiplicity. The t intervals above assume approximate normality, which ten or fifteen seeds cannot establish, so every contrast is also reported with an exact

two-sided sign test and a seed-level bootstrap (10,000 resamples), and the family of six is corrected. The surviving ViT-L/14 degradation margin has 15 of 15 seeds positive, which gives a sign-test p of 0.0001 and a bootstrap interval of [+2.9, +4.4] points; it is the only member of the family that survives Holm-Bonferroni, which controls the family-wise error rate (adjusted p = 0.0004), and Benjamini-Hochberg, which controls the false-discovery rate (adjusted p = 0.0004) over the same six contrasts. The remaining five give sign-test p between 0.11 and 0.75, adjusted 0.33 to 1.00. The smallest of the fifteen surviving margins is +0.00004 UAR (seed 6), positive by 0.004 points; if that one were counted as a tie and dropped, the exact two-sided sign test would give 0.00012 and its Holm-adjusted value 0.0007 instead of 0.0001 and 0.0004, so the verdict does not depend on that convention. The correction matters here: with the five seeds this analysis originally used, the exact sign test cannot go below 0.0625 and nothing could have survived it. Resampling the evaluation units inside a fixed training seed, which isolates scoring variance from training variance, leaves the survivor intact ([+1.73, +5.87] points over the 12 conditions) and turns the ViT-B/16 crossed-grid contrast negative ([-2.83, -0.37] over the 50 cells). *Leave-one-seed-out.* No verdict rests on a single training seed. Recomputing every contrast with one seed removed in turn, the surviving ViT-L/14 degradation margin keeps a mean between +3.31 and +3.75 points and all fifteen leave-one-out intervals exclude zero; the fresh-axis margin keeps a mean between +2.57 and +2.94 points with all fifteen intervals excluding zero; and among the five contrasts that fail the criterion, none of the sixty leave-one-out intervals excludes zero, so their failure is not the work of any single influential seed. All of these quantities are recomputed from the released records by the released analysis script.

5.12 Negative results

We report the negative findings with the same detail as the positive ones.

5.12.1 Encoder dependence: the ViT-B/16 margins do not replicate at ViT-L/14

Both key ViT-B/16 quantities fail to replicate at the larger encoder. TAP − CLIP-Adapter is −1.96 points on Set A (CI [-2.59, -1.31]) and −1.37 points on Set B (CI [-2.38, -0.39]); TAP − the capacity-matched head is −1.10 points on Set A (CI [-1.82, -0.38]) and −0.07 points on Set B (CI [-1.08, +0.92], covering zero). The token branch is not broken at ViT-L/14 — TAP still exceeds max(PFB, memory) by 9.03 and 6.18 points on the two sets — but it stops being better than the strongest trained alternative, and on the source split CLIP-Adapter already ranked first (probe-val UAR 0.6799 vs TAP 0.6754).

5.12.2 Source-domain diagnostic: under matched nonlinear readouts no difference was detected

An independent diagnostic on the source split (Fig. 3) compared frozen and trained readouts of the global feature, of mean-pooled tokens, and of attention-pooled tokens. With a linear readout the token route led by about 3.4 points (R2 − R1 = 0.0336 at ViT-B/16 and 0.0341 at ViT-L/14). With a nonlinear MLP head of matched capacity no difference was detected: R3 (attention pooling) reached 0.6200 against 0.6201 for the global feature with an MLP head at ViT-B/16, and 0.6724 against 0.6714 at ViT-L/14. Adding a global passthrough to the token route gave only R4a − R2 = 0.0077 (ViT-B/16) and 0.0049 (ViT-L/14), and handing the global vector directly to the classifier gave R4b − R3 = −0.0096 (ViT-B/16) and −0.0046 (ViT-L/14). The source-side advantage of the token route therefore did not reproduce once the readout was matched.

5.12.3 RAF-DB confirmation blocked by the availability gate

The planned RAF-DB confirmation was stopped by its pre-specified availability gate and consumed nothing: no token extraction, no training, no evaluation. The only local copy is a third-party re-upload whose provenance file marks it QUARANTINED and whose content is described as a preprocessed FER2013/RAF-DB mixture; there is no official split or label file, the four classes are exactly equal at 5,920 images each, and the pre-specified ratios 2, 5 and 10 are not realisable. Four of the eight gate checks failed (provenance is not a quarantine notice; an official RAF-DB split/label file exists on disk; class counts are not an artificial balancing signature; all pre-specified ratios are realisable), so the run was stopped as pre-specified and H1 has no outcome.

5.12.4 Header/fusion module pool: 0 of 13 configurations passed

Thirteen header and fusion configurations were screened under the same confirmation protocol. None passed the threshold of being at least as good as both families' best in two-thirds of the confirmation cells. The best selection-set candidate, a scalar linear head scoring 0.6461 against a canonical 0.6431 (a difference of 0.0030), failed the crossed protocol in 10 of 10 cells with a worst single-cell deficit of 14.66 points; a small MLP collapsed during training (selection-set UAR 0.1921). We therefore did not name a new module from this pool.

5.12.5 T2, T3 and T4 do not hold

Of the four initially considered token modules, only T1 (TAP) met the confirmation threshold. On the crossed grid the mean margin over the two families' best was 0.0179 for T2, 0.0553 for T3 and 0.0292 for T4, against 0.1121 for T1; on the KDEF-source set T3 was -0.0927 against the capacity-matched head. The parameter-matched 1×1 control for T2 is recorded in the released per-configuration records.

5.12.6 Three gating families all failed

Three source-calibrated gating designs were tested and none transferred. A sample-level reliability fusion reached at least the two-family best in only 1 of 18 natural/prior cells (threshold 2/3), with a worst single-cell loss of 15.7 points; it failed because the source-side reliability weight did not transfer. A degradation-severity gate separated all 12 conditions under a per-condition oracle threshold, but the source-calibrated entropy rule hit only 10 of 12 sweep conditions and lost 0.0805 UAR on average on the natural grid. A global regime gate separated 18 of 18 cells under a target oracle threshold, but the same threshold hit only 12 of 18 cells on real targets, mis-selected all six CK+ cells, and cost up to 0.1083 UAR in a single domain; those 18 cells come from only four independent domains, so they should be read as four domain-level checks rather than 18 independent ones.

5.13 Feature-convention sensitivity (post-hoc)

The cross-database comparison in the crossed-grid and second-support subsections is between each route in its native feature configuration: the global route uses the L2-normalised mean of the image and its mirror, whereas the token caches hold the original view only. We therefore ran the symmetric version as a post-hoc sensitivity, retraining the token arm on the spatially aligned two-view mean, and report it so that the cross-database margin can be read against the convention it depends on. Mirror averaging requires care, because a left-right flip does not preserve patch position: the mean per-position cosine between the raw and mirrored token sequences is 0.65-0.78, rising to 0.89-0.92 once the mirror is un-flipped in the token grid.

Table 11: Feature-convention sensitivity of the token route (post-hoc, not confirmatory).

token-arm feature convention	Set A mean UAR	Set B mean UAR	vs CLIP-Adapter, Set A	vs CLIP-Adapter, Set B
single view (as in those subsections)	0.6784	0.6722	+0.0388	+0.0347
two views, spatially aligned mean	0.6182	0.6344	-0.0214	-0.0031

Two things follow. First, the single-view configuration is the better one for the token route, by 6.02 points on Set A and 3.78 on Set B. The convention is therefore not an accident of tuning, but neither is it neutral for the comparison: it is the more favourable of the two conventions for the token route, so the +3.88 and +3.47 point margins are upper bounds rather than convention-free estimates. Second, the cross-database margin over CLIP-Adapter is nevertheless convention-dependent: under the aligned two-view convention it reverses to -2.14 points on Set A with an interval excluding zero.

The aligned two-view arm on the degradation axis. The degradation axis carries the one contrast that survived retraining, so we repeated the same construction there for both encoders (Table 12). At ViT-L/14, where the surviving contrast lives, the convention makes no difference at fifteen seeds: the aligned two-view arm reaches **+3.85 points** over CLIP-Adapter (SD 1.20, seed-level 95% interval [+3.19, +4.52]) and **+3.08 points** over the capacity-matched head (SD 0.82, [+2.63, +3.54]), with fifteen of fifteen seeds positive in both cases, and the aligned-minus-single-view difference is +0.36 points (interval [-0.27, +0.98], i.e. indistinguishable). A seed-level bootstrap gives [+3.26, +4.45] points for the aligned arm against CLIP-Adapter, and a unit-level resample over the 12 conditions is not reported for this extension. The surviving margin is therefore reproduced under the symmetric convention, and here it costs nothing in precision: the between-seed spread is 1.20 points under the aligned convention against 1.69 for the single-view arm. At ViT-B/16 the convention matters, and the direction is the opposite of the cross-database result: the aligned two-view arm reaches +1.18 points over CLIP-Adapter (SD 1.00, interval [+0.47, +1.90], ten of ten seeds positive) and +1.33 points over the capacity-matched head, where the single-view arm averages +0.49 points (SD 1.18, 6/10). That is the same lesson one level down: on this axis the ViT-B/16 comparison moves by 0.7 points with the convention and its single-view interval covers zero, whereas at ViT-L/14 the margin is positive, large and interval-separated from zero under both. The aligned ViT-L/14 arms use the rate the source probe split selected (1e-3). The aligned ViT-B/16 row in the table is the one held at the rate the original protocol selected (3e-3), which gives +1.18 points; re-selecting the rate on the source probe split, which prefers 1e-3, gives +1.24 points. The two therefore agree in sign and magnitude, and the convention effect is not a learning-rate effect.

This analysis is post-hoc, carries no confirmatory weight, and consumed no new confirmation set.

6 Discussion

6.1 Anchor

The comparison this paper set out to make was a controlled one: a parameter-matched global control, two frozen evaluation sets, two encoder scales and a pre-specified degradation axis. Retraining shows that the control design was not the binding constraint. On the 50-cell crossed grid the ViT-B/16 contrast between the token route and CLIP-Adapter, reported as +3.88 points by the single run, averages -1.61 points with a between-seed standard deviation of 4.75 points and only 2 of 10 seeds positive; a conditional resample of the cells inside one fixed seed is also negative ([-2.83, -0.37]),

Table 12: Feature-convention sensitivity of the degradation-axis contrast (post-hoc, not confirmatory). The single-view rows are the phase-12 arms. The aligned ViT-L/14 row extends the phase-9 construction from five to fifteen seeds at the rate phase 9 selected on the source probe split (1e-3). The aligned ViT-B/16 row is trained with the rate held at the value the original protocol selected (3e-3) and gives +1.18 points; re-selecting that rate on the source probe split, which prefers 1e-3, gives +1.24 points.

encoder	token-arm feature convention	TAP - CLIP-Adapter	TAP - capacity-matched head	seeds positive
ViT-B/16	single view (as on the corruption axis)	+0.49 (SD 1.18)	+0.63 (SD 0.99)	6/10
ViT-B/16	two views, spatially aligned mean	+1.18 (SD 1.00)	+1.33 (SD 0.84)	10/10
ViT-L/14	single view (as on the corruption axis)	+3.50 (SD 1.69)	+2.73 (SD 1.31)	15/15
ViT-L/14	two views, spatially aligned mean	+3.85 (SD 1.20)	+3.08 (SD 0.82)	15/15

which shows that the sign is not an artefact of averaging over seeds rather than establishing anything between seeds. On the corruption axis the same contrast at the same encoder moves from +1.81 to +0.49 points. The ViT-L/14 cross-database contrast collapses to -0.19 points (SD 2.42), and the second-source contrasts behave the same way: +3.47 to +0.34 points (SD 3.65) at ViT-B/16 and -1.37 to +0.59 points (SD 1.51) at ViT-L/14. Of the six contrasts the single run would have reported across the three confirmation axes, one is recovered and it is then recovered again on a fresh corpus: the degradation-axis margin at ViT-L/14, +3.50 points over CLIP-Adapter (SD 1.69) and +2.73 over the parameter-matched head (SD 1.31) on FER2013, and +2.75 points over CLIP-Adapter (15/15 seeds) on the hash-frozen JAFFE axis.

6.2 Position

The nearest comparison is with work that reports a single fit per arm. Both CLIP-Adapter and the token adapter here were trained on the source split under the same budget and the same learning-rate grid, which is the protocol-consistent form of each method but not the form used in the original papers, where the adapter is fitted on the target few-shot support. What this paper adds to that comparison is not a further point on the adapter leaderboard but a measurement of the noise floor that a leaderboard comparison of this kind carries. The measurement is specific: it applies to a 50-cell cross-database grid, a 12-condition degradation axis, two encoder scales, and three arms with roughly 0.2M trainable parameters each. Its practical reading is that a single-run directional claim from a design of this size is a noisy estimate of one draw rather than a reliable estimate of the between-seed mean, and that the two can differ in sign.

6.3 Interpretation and contribution

The first contribution is the measurement itself, and it is directly established by the design. Holding the architecture, the data, the learning rates and the evaluation code fixed, the only thing that changes between the single run and the retrained arms is the seed. Under that manipulation the ViT-B/16 cross-database contrast changes sign on average and the ViT-B/16 degradation contrast vanishes, while the ViT-L/14 degradation contrast is reproduced in every seed. The single-run

numbers are not artefacts of a mis-specified evaluator: the same code returns the reported values to four decimals when it is given the original checkpoints.

The second contribution is the surviving effect, its shape, and the fact that it survives a second test. At ViT-L/14 the token route beats CLIP-Adapter and the parameter-matched global head at the same time, by margins that are about twice their between-seed standard deviation, and it is the only member of the six-contrast family that survives a family-wise correction. The parameter-matched control is beaten as well, so the margin is not a signature of a stronger readout layer, and the convention sensitivity shows that it is not a signature of the token route's feature-view convention either. It also reappears on JAFFE, which was never touched before the pass criterion for it was written down. Within the designs tested here the margin tracks where the adaptation is applied - before pooling rather than after it - but the three arms also differ in structure and in activation, so we treat that as the reading the design supports rather than as a decomposition of the effect.

The third contribution is the shape of the instability across axes and encoders. On the crossed grid the between-seed standard deviation of the same contrast is 4.75 points at ViT-B/16 against 2.42 at ViT-L/14, which is the opposite of the intuition that a smaller model is easier to fit reliably; on the corruption axis the ordering reverses (1.18 against 1.69 points). The instability is therefore a property of the axis as much as of the encoder, and a single measure of the noise floor does not transfer between them.

6.4 Boundary and limitations

Scoring detail on the second support source. Set B rebuilds the support set and the prior from the KDEF-source split but re-uses the trained arms as-is, so the trained-arm predictions are identical across the five cell seeds of each target; the per-seed statistic there is the mean over the three targets, which makes the Set B variance estimate rest on three rather than fifteen independent scoring units.

Precision of the variance estimates. With fifteen seeds the between-seed standard deviation is itself uncertain at roughly 18% relative, and at ten seeds closer to 24%. The qualitative conclusion - that five of the six contrasts move to values indistinguishable from zero while the ViT-L/14 degradation margin does not - does not depend on that precision, but the exact standard deviations should be treated as estimates.

Seed-count extension. The first pass of the retraining used five seeds at ViT-L/14, and the count was raised to fifteen after the direction was already visible in those five, so that fifteen-seed interval is a precision estimate rather than a prospective test. A second batch of fifteen seeds was subsequently fixed and hashed in advance and run with no change to the recipe, the arms, the rates or the axes; it reproduces the contrast (+3.77 against +3.50 points, pooled +3.63 [+3.06, +4.21]), and four of the five contrasts that failed in batch 1 also fail in batch 2. The fifth, the ViT-B/16 degradation margin at the published rate, moves to +1.08 points over the new seeds and +0.79 points over the twenty pooled seeds, an interval that excludes zero although its exact sign test does not ($p = 0.26$); both are reported and the contrast is treated as replicated on the interval criterion only. The prospective batch does not turn the consumed evaluation axes into a fresh confirmation set, and the five-seed subset that preceded the extension is still reported in full.

Multiplicity. The six core contrasts are one family and are corrected as such. With ten or fifteen seeds the exact sign test is no longer floored at 0.0625, and one contrast survives both Holm-Bonferroni (family-wise error rate) and Benjamini-Hochberg (false-discovery rate) correction while the other five have adjusted p-values between 0.33 and 1.00. At five seeds - the size of the first pass of this analysis - the exact test cannot reach 0.05 at all, so a five-seed study could not have made that distinction.

Rate selection inside the seed. The nested condition re-selects the learning rate for every arm inside every seed from the original grid, so its distribution covers selection as well as fitting. It leaves

the survivor unchanged (+3.68 against +3.50 points) but moves the ViT-B/16 crossed-grid contrast by 3.27 points, which is the size of the effect the single run reported there. The variance of these contrasts therefore depends on the tuning convention as well as on the seed, and a design that fixes the rate at a value chosen once understates that component.

Three arms only. The training-free reference columns (prediction-frequency balancing and the support-feature cache) have no training seed and were not part of the variance analysis.

Task and encoder scope. The study covers cross-dataset facial expression recognition at two CLIP scales. Whether the same variance structure appears in other tasks or in other encoders is untested.

Two-view convention. The global route consumes the mean of the image and its mirror while the token route consumes the original view only. The retraining analysis inherits that convention, since each route is measured in its own native configuration; a fully symmetric comparison gives the token route the aligned two-view mean, and the post-hoc sensitivity reported below covers both outcomes: on the ViT-B/16 cross-database sets the aligned convention lowers the token route and reverses its margin over CLIP-Adapter, whereas on the degradation axis at ViT-L/14 the aligned and single-view conventions are indistinguishable and the surviving margin is unchanged.

Negative results retained. The module pool, the three gating families and the blocked RAF-DB confirmation are reported in full below and contributed no arm to the retraining analysis.

Confirmatory versus estimation evidence. What is pre-specified in this paper is a set of contrast definitions together with four extension axes: the six cross-encoder, cross-axis contrasts at their original rates, and the JAFFE, relabelled, fresh-degradation-family and third-encoder axes, each judged against the criterion frozen before it was run. The retraining analysis that re-evaluates the six contrasts on retained cells is not itself a confirmatory test: it is the post-hoc variance probe described in Section 4.5, extended once for power in an outcome-aware way, and it is reported as estimation evidence with its own seed intervals. The nested learning-rate condition, the aligned-view sensitivity, the leave-one-seed-out analysis and the pooled two-batch estimates are estimation evidence: they characterise the same contrasts under other conventions, they are reported in full, and they are not part of the family-wise correction of Section 5.6.

What this paper does not claim. We do not claim that token-level adaptation is generally superior to adaptation on the pooled global feature, that the difference between the two routes is caused by the adaptation location alone, or that the surviving contrast is a general robustness property. It is an accuracy advantage on the twelve pre-specified corruption conditions at one encoder; it does not extend to other degradation families or to a third encoder, and on the fresh corpus the degradation-specific increment is not distinguishable from zero. Neither do we claim that the five contrasts which fail the criterion are absent effects: an interval that includes zero bounds the precision of a single-run estimate, and nothing more.

6.5 Extension

The fresh-axis test reported above was the most direct extension, and it passed under a criterion fixed in advance; what remains is that the corpus is small (213 images from ten models, posed expressions, studio capture) and that on it the token route's advantage is largely a cross-domain advantage rather than a degradation-specific one, so the in-domain decomposition and the out-of-domain decomposition should not be conflated. A second extension is to nest the remaining tuned quantities inside the seed, which we did here for the learning rate and which moved an unstable contrast by the size of its reported effect; a third is to re-run the second source domain with its own training arms rather than re-using the FER2013 arms. The practical recommendation that follows from the present results is modest and cheap: when a paper compares adaptation routes for a frozen encoder, report a between-seed spread for the contrast, because a single run of this size can produce a directional

result that a rerun reverses. Reporting the best of several seeds instead is a screening device rather than a substitute for that distribution, since it selects a favourable draw and is biased upwards, and we do not propose it as an estimate of the contrast.

7 Conclusion

We asked whether a single training run is enough to support a comparison between token-level and global-feature adaptation of a frozen CLIP encoder. It is not. Retraining the same three arms from the seed alone - initial weights, batch order and learning rate - with fifteen seeds at ViT-L/14 and ten at ViT-B/16 leaves five of the six contrasts that a single run would have reported unsupported; a prospectively fixed second batch of seeds reproduces the surviving contrast and, pooled over the twenty ViT-B/16 seeds, also brings the degradation contrast's interval above zero, leaving four of the six unsupported. The ViT-B/16 cross-database margin over CLIP-Adapter, reported as +3.88 points, averages -1.61 points with a between-seed standard deviation of 4.75 points and is negative even when the cells are resampled inside a fixed seed; the ViT-B/16 degradation margin moves from +1.81 to +0.49 points; the ViT-L/14 cross-database reversal moves from -1.96 to -0.19 points; and the second-source margins move from +3.47 to +0.34 and from -1.37 to +0.59 points. The same evaluator agrees with every one of the original numbers to within 0.0002 at worst (0.00014) from the original checkpoints, so the disagreement is about training, not about measurement.

One contrast survives, and it survives a second test. At ViT-L/14 the token route has a higher accuracy than CLIP-Adapter on the pre-specified corruption axis, by +3.50 points (SD 1.69), and a higher accuracy than the parameter-matched global head by +2.73 points (SD 1.31), positive in 15 and 14 of 15 seeds, with a seed-level interval of [+2.56, +4.43] points; it is the only member of the six-contrast family that survives Holm-Bonferroni or Benjamini-Hochberg correction, and it is unchanged when the learning rate is re-selected inside each seed or when both routes are measured under the same two-view feature convention. The same contrast, re-tested on the untouched JAFFE corpus under a criterion hashed in advance, is +2.75 points with all fifteen seeds positive. Its clean controls separate the two settings honestly: in domain the margin is produced by the degradation, out of domain it is mostly a cross-domain accuracy advantage that corruption only slightly widens.

Two further observations follow. Re-selecting the learning rate inside each seed moves one of the unstable contrasts by 3.3 points, the size of the effect it was reported to have, so the tuning convention is a second axis of instability rather than a remedy for the first. And the between-seed spread is axis-specific rather than a simple function of encoder size: it is twice as large at ViT-B/16 as at ViT-L/14 on the crossed grid, and larger at ViT-L/14 on the degradation axis. Comparisons of frozen-encoder adaptation routes should therefore report a training-variance estimate and state the tuning convention, rather than relying on a single fit.

Statements and Declarations

Competing interests: The authors declare no competing interests.

Funding: The authors received no specific funding for this work.

Ethics approval and consent to participate: Not applicable. All corpora used here (FER2013/FER+, CK+, KDEF, JAFFE) are public research benchmarks; no new participants were recruited and no new human-subjects data were collected. JAFFE is used under its own terms - non-commercial scientific research, no redistribution and no posting on websites or social media, citation of both creator-required articles, and no more than ten sample images in any publication; the archived copy used is identified by its sha256 in the release, and none of its images are reproduced here.

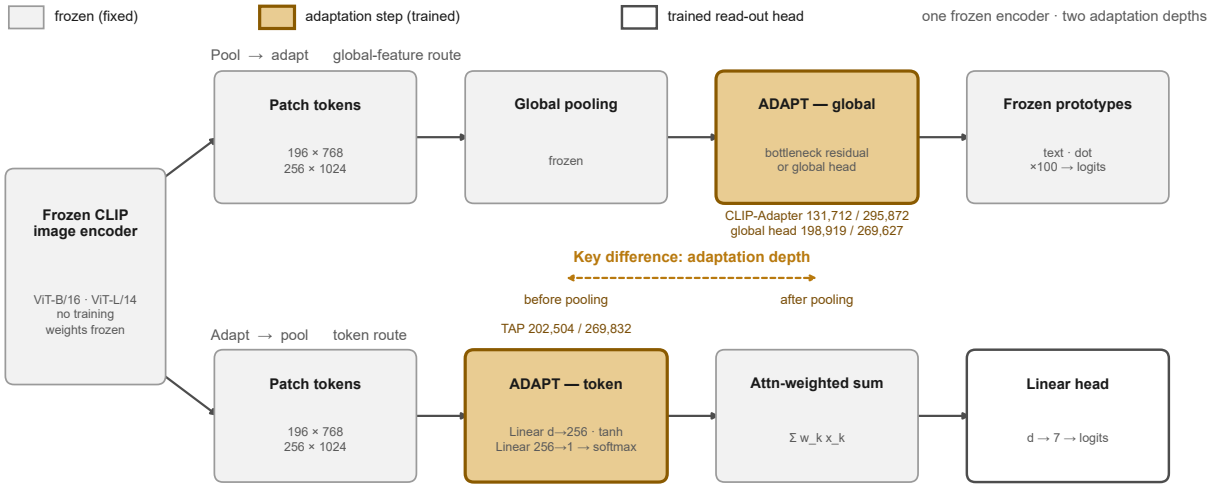


Figure 1: Adaptation before versus after pooling. Both routes share one frozen CLIP image encoder. The global route pools the patch tokens first and adapts the pooled vector; the token route adapts before pooling, aggregates the tokens with a learned attention distribution and classifies the aggregate. The only structural difference is the depth at which the adapter sits, and the trainable parameter counts are given for each adapter and encoder scale

The corpora were used under the terms attached to their public releases and were not re-identified or linked to any individual.

Consent for publication: Not applicable. The study uses only public research benchmarks and reports no individual-level data.

Data availability and reuse conditions: The image corpora (FER2013/FER+, CK+, KDEF) are public and are not redistributed with this paper. What is released is derived numerical data rather than face images: split and label-index files, extraction, training and confirmation scripts, per-cell result CSVs, paired-bootstrap JSONs, the per-seed records of the training-variance re-runs, the trained adapter weights, and a SHA256 manifest recording commands, hashes, exit codes and seeds. The full frozen patch-token and global-feature caches - tens of gigabytes of fp16 arrays - are available from the authors for research use on request, because they are embeddings of face images computed from public research corpora and are therefore released for research use only and under the terms of the corpus each cache derives from; the manifest records that provenance so a downstream user can check the intended use against the relevant corpus licence. One residual risk deserves to be stated explicitly: face embeddings, unlike the images themselves, can in principle support re-identification or attribute inference if they are matched against an external gallery. The caches we circulate carry no identity labels beyond the seven emotion categories of the source corpora, and we ask users not to attempt re-identification. The caches can also be regenerated from the public corpora with the released extraction scripts and manifests, so access does not depend on a bespoke transfer and the evaluation can be repeated against a governed copy. No restricted, clinical or otherwise non-public data are included. The released numerical artifacts are at `https://github.com/CIUMZY/tap-fer-clip` (release tags v1.0.0 to v1.5.0) and is archived at `https://doi.org/10.5281/zenodo.23237320`, with the version used here at `https://doi.org/10.5281/zenodo.23237321` (tag v1.5.0).

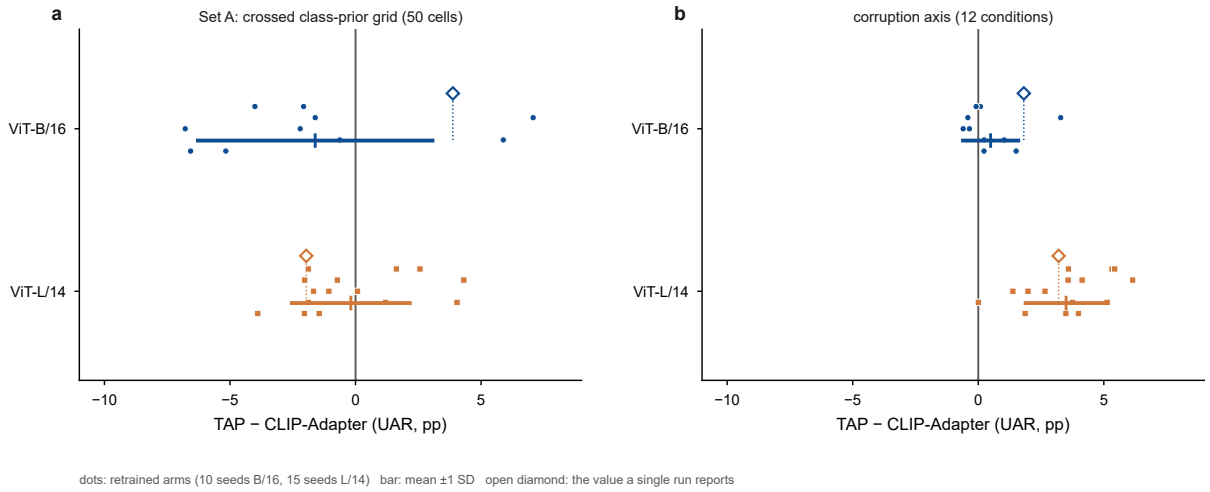


Figure 2: Training-seed variance of the TAP minus CLIP-Adapter contrast. Dots are the retrained arms, the bar is the between-seed mean plus or minus one standard deviation, and the open diamond is the value a single run reports. Panel a: the 50-cell crossed class-prior grid, where the ViT-B/16 contrast reverses sign on average. Panel b: the 12-condition corruption axis, where only the ViT-L/14 contrast survives retraining

Code availability: The extraction, training, retraining and evaluation scripts, the run manifest and the trained adapter weights are released together with the evidence files, under an MIT licence for code and research-use terms for the evidence files, at `https://github.com/CIUMZY/tap-fer-clip` (release tags v1.0.0 to v1.5.0). (archived at `https://doi.org/10.5281/zenodo.23237320`; the version used here is `https://doi.org/10.5281/zenodo.23237321`, tag v1.5.0)

Use of AI-assisted tools: During the preparation of this work the authors used an AI-based assistant for code development, data processing and language editing. The authors reviewed and edited all content produced with it, recomputed every reported number from the released records, and take full responsibility for the content of the publication.

Author contributions: All authors contributed to the conception and design of the study, to the implementation and running of the experiments, to the analysis and interpretation of the results, and to the writing and critical revision of the manuscript. All authors read and approved the final manuscript.

References

[1] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016. doi: 10.1145/2993148.2993165. URL `https://doi.org/10.1145/2993148.2993165`.

[2] Yassir Bendou, Amine Ouasfi, Vincent Gripon, and Adnane Boukhayma. Proker: A kernel perspective on few-shot adaptation of large vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025. doi: 10.1109/cvpr52734.2025.02336. URL `https://doi.org/10.1109/cvpr52734.2025.02336`.

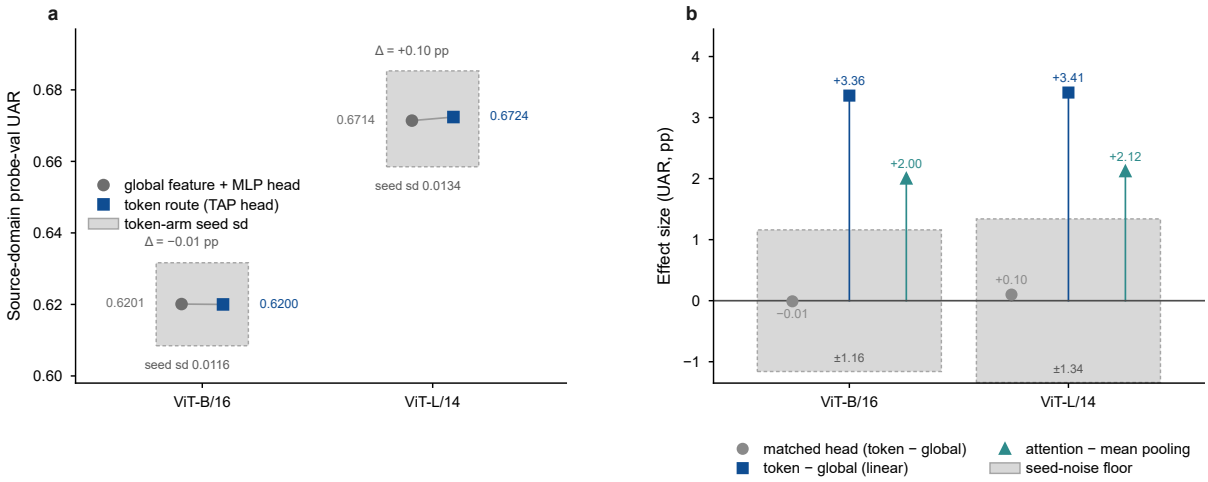


Figure 3: Matched-capacity attribution diagnostic. Panel a: source-domain probe-validation accuracy of the global feature with a matched nonlinear head against attention-pooled tokens, at both encoder scales; the shaded band is the between-seed standard deviation of the token arm and both differences fall inside it. Panel b: the same differences in percentage points against that band for three readouts, the matched nonlinear head, a linear readout, and attention pooling against uniform mean pooling; only the last two exceed the band

[3] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL `https://doi.org/10.1111/j.2517-6161.1995.tb02031.x`.

[4] Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194. Association for Computational Linguistics, 2019. doi: 10.18653/v1/d19-1224. URL `https://doi.org/10.18653/v1/d19-1224`.

[5] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 2023. doi: 10.1007/s11263-023-01891-x. URL `https://doi.org/10.1007/s11263-023-01891-x`.

[6] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 2015. doi: 10.1016/j.neunet.2014.09.005. URL `https://doi.org/10.1016/j.neunet.2014.09.005`.

[7] Zhe Guo, Yi Liu, Kui Wang, Bingxin Wei, and Yi Wang. Enhancing cross-domain facial

expression recognition with expression relationship contrastive learning in a source-free setting. *The Visual Computer*, 2026. doi: 10.1007/s00371-026-04424-y. URL https://doi.org/10.1007/s00371-026-04424-y.

[8] Jia-Wei Hai, Yijun Wang, and Xiu-Shen Wei. Towards fine-grained robustness: Attention-guided test-time prompt tuning for vision-language models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 39451–39467. PMLR, 2026. URL https://proceedings.mlr.press/v306/hai26a.html.

[9] Jiazhen Huang, Xiao Chen, Zhiming Liu, Yaru Sun, Jingyan Jiang, and Zhi Wang. What drives test-time adaptation for clip? a controlled empirical study from an update perspective. 2026. doi: 10.48550/arXiv.2606.14299. URL https://doi.org/10.48550/arXiv.2606.14299. Introduces the TTABC benchmark; arXiv:2606.14299; verified via doi.org CSL JSON 2026-10-07.

[10] Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. Lp++: A surprisingly strong linear probe for few-shot clip. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024. doi: 10.1109/cvpr52733.2024.02244. URL https://doi.org/10.1109/cvpr52733.2024.02244.

[11] Shan Li and Weihong Deng. Deep facial expression recognition: a survey. *IEEE Transactions on Affective Computing*, 13(3):1195–1215, 2022. URL https://arxiv.org/abs/1804.08348. provenance: verified in the previous project bibliography; arXiv:1804.08348.

[12] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 2024. URL https://arxiv.org/abs/2303.15361. provenance: verified in the previous project bibliography; arXiv:2303.15361.

[13] Zachary C. Lipton, Yu-Xiang Wang, and Alex Smola. Detecting and correcting for label shift with black box predictors. 2018. doi: 10.48550/arxiv.1802.03916. URL https://doi.org/10.48550/arxiv.1802.03916.

[14] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. IEEE, 2010. doi: 10.1109/cvprw.2010.5543262. URL https://doi.org/10.1109/cvprw.2010.5543262.

[15] D. Lundqvist, A. Flykt, and A. Öhman. Karolinska directed emotional faces. *PsycTESTS Dataset*, 1998. doi: 10.1037/t27732-000. URL https://doi.org/10.1037/t27732-000.

[16] Michael Lyons, Miyuki Kamachi, and Jiro Gyoba. The japanese female facial expression (jaffe) dataset, 1997. URL https://doi.org/10.5281/zenodo.14974867. Archived copy used as the untouched confirmation corpus in this work; the archive sha256 is recorded in the release. Accessed 2026-10-04.

[17] Michael J. Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 1998. doi: 10.1109/afgr.1998.670949. URL https://doi.org/10.1109/afgr.1998.670949.

[18] Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 2018. doi: 10.1073/pnas.1708274114. URL https://doi.org/10.1073/pnas.1708274114.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html. provenance: verified in the previous project bibliography.

[20] Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348. Association for Computational Linguistics, 2017. doi: 10.18653/v1/d17-1035. URL https://doi.org/10.18653/v1/d17-1035.

[21] Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 2002. doi: 10.1162/089976602753284446. URL https://doi.org/10.1162/089976602753284446.

[22] 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. 2022. doi: 10.48550/arxiv.2209.07511. URL https://doi.org/10.48550/arxiv.2209.07511.

[23] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. 2017. doi: 10.48550/arxiv.1703.05175. URL https://doi.org/10.48550/arxiv.1703.05175.

[24] Chris Vorster, Mayug Maniparambil, Noel E. O'Connor, Noel Murphy, and Derek Molloy. Hold-one-shot-out (hoso) for validation-free few-shot clip adapters. 2026. doi: 10.48550/arXiv.2603.04341. URL https://doi.org/10.48550/arXiv.2603.04341. arXiv:2603.04341; verified via doi.org CSL JSON 2026-10-07.

[25] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. 2020. doi: 10.48550/arxiv.2006.10726. URL https://doi.org/10.48550/arxiv.2006.10726.

[26] Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2024. doi: 10.1109/cvprw63382.2024.00166. URL https://doi.org/10.1109/cvprw63382.2024.00166.

[27] Renrui Zhang, Zhang Wei, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. 2022. doi: 10.48550/arxiv.2207.09519. URL https://doi.org/10.48550/arxiv.2207.09519.

[28] Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2605–2615. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00246. URL https://doi.org/10.1109/ICCV51070.2023.00246.