USEFUL DIRECTIONS AND USEFUL PARTNERS: AUDITING PAIRED GRADIENT SUBSPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Pairing related inputs can suggest directions for adapting a classifier. To measure what the original matching contributes, we hold each input's loss gradient and label fixed, shuffle partners within labels, and refit the subspace of gradient differences. Random directions provide a different comparison. We show that identical gains over random references can accompany positive, zero or negative pairing effects on retained target-gradient energy. This ambiguity remains even when random directions lie in the space where gradients can occur. Across visual and multilingual dialogue settings, energy advantages over this restricted random reference are common. Evidence that the original partners improve energy retention is selective, and accuracy gains are rarer. Compared with support-matched random directions, dialogue adaptation also incurs class-balanced prediction costs that ordinary accuracy can hide. The audit separates the quality of the directions, the contribution of the original partners, and the value of updating along those directions.

1 INTRODUCTION

Pairing an image with a changed view offers a way to choose candidate directions for adapting a classifier. Their gradient difference describes how the two losses respond differently to a small change in the adapted variables. The fitted subspace retains the greatest total squared length of these differences. We use this construction to study what the original image–view matching adds: compare its subspace with one fitted after exchanging partners. This comparison tests whether a gain over random directions depends on the original matching.

We test pairing by shuffling partners within labels while keeping both gradient collections fixed. This preserves the inputs and label agreement. Each shuffle yields a subspace with the same number of directions, scored on the same held-out target gradients. Comparing its score with that of the original fit measures what the original partners add beyond the two input collections. We call this comparison a *partner-shuffle audit*.

Why is a random subspace insufficient? It discards both the collected gradient structure and the pairing. It may even point where the classifier's gradients cannot occur. Restricting random directions to the gradient support removes this geometric disadvantage. It still leaves the matching untested: the same positive gains over both random references can arise when pairing helps, has no effect, or hurts. Random-reference scores alone therefore cannot settle the pairing question.

Across visual shifts and multilingual prompt–reply pairs, fitted directions often retain more gradient energy than random directions in the same support. Evidence that the original partners add further benefit is selective, and pairing-specific accuracy gains are rarer. Comparisons with support-matched random directions also expose class-balanced prediction costs that ordinary accuracy can hide. The audit separates useful directions from useful partners. The theory explains why random references cannot replace this comparison; the experiments show why prediction quality needs a separate test.

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

2 Related Work

Learning useful gradient directions. Low-dimensional gradient structure can help a model learn from related problems. SubGD identifies update directions shared across tasks to improve few-shot learning (Gauch et al., 2022). PGrad aggregates principal directions from optimization rollouts across training domains to improve generalization to unseen domains (Wang et al., 2023). Memory-efficient optimizers use this structure for another purpose: GaLore projects gradients to reduce optimizer memory, while LDAdam maintains adaptive statistics as the low-dimensional space changes (Zhao et al., 2024; Robert et al., 2025). Paired-input differences introduce an additional choice: which two gradients are subtracted. A useful subspace may reflect the member gradients, their original matching, or both. Determining what the matching adds therefore requires changing the assignment while preserving the gradients that define the construction.

Testing what a score reflects. Control studies show why a successful score need not establish its intended explanation. Some saliency maps are insensitive to the learned model and data-generating process (Adebayo et al., 2018). Input changes that preserve model behavior can also change attribution (Kindermans et al., 2019). For representation probes, Hewitt and Liang use random word-type control tasks to separate linguistic information from probe memorization (Hewitt & Liang, 2019). Lan et al. compare matched feature similarity with random pairings of the same features (Lan et al., 2024). Within-label partner shuffling applies this principle to paired gradients. It changes the matching while preserving the stored gradients and label agreement.

From retained directions to better predictions. Useful matching information may help find directions, but prediction gains also depend on the update rule. Test-time training uses self-supervised objectives on unlabeled target inputs (Sun et al., 2020), while Tent minimizes prediction entropy (Wang et al., 2021). Under continual shifts, CoTTA addresses unreliable pseudo-labels and forgetting through averaged predictions and weight restoration (Wang et al., 2022). EATA selects reliable, nonredundant samples and regularizes important weights (Niu et al., 2022). Gradient alignment alone cannot settle prediction quality. Song et al. find strong alignment with dominant Hessian directions in which projected SGD can cease to reduce training loss (Song et al., 2025). We therefore compare subspaces under one fixed hard-pseudo-label update rule and measure both gradient energy and prediction quality.

3 What a Random Reference Can Establish

Consider an illustrative two-dimensional case in which all gradients lie along a horizontal line. Any nonzero difference between two gradients also lies on that line, regardless of who is paired with whom. Fitting one direction therefore recovers the horizontal line whenever the differences are not all zero. It retains all the squared length of a horizontal target gradient. A random line in the plane retains only half on average: tilting away from the horizontal gradient shortens its projection. The fitted direction can therefore beat this random reference without gaining anything from the original partners.

The same reasoning applies when gradients occupy a small part of a larger space. Let r be the dimension of this gradient *support*, and let d be the full dimension. We choose $k \leq r$ directions. Their span S receives a score equal to the expected squared length of the projected target gradient: $\mathrm{PE}(S) = \mathbb{E}\|P_S v\|_2^2 = \mathrm{tr}(P_S T)$. Here P_S is the orthogonal projector and $T = \mathbb{E}[vv^\top]$ is the target second moment. Total target-gradient energy is $\mathrm{tr}(T)$. Uniform random directions retain a fraction k/d on average. Within the support, this fraction is k/r:

$$\mathbb{E}[\mathrm{PE}(S_{\mathrm{amb}})] = \frac{k}{d}\,\mathrm{tr}(T), \qquad \mathbb{E}[\mathrm{PE}(S_{\mathrm{sup}})] = \frac{k}{r}\,\mathrm{tr}(T). \tag{1}$$

Here target gradients lie in the support. The horizontal-line example has $k = r = 1$ and $d = 2$: the two fractions are one half and one. This advantage depends on where gradients can point, without using their partners.

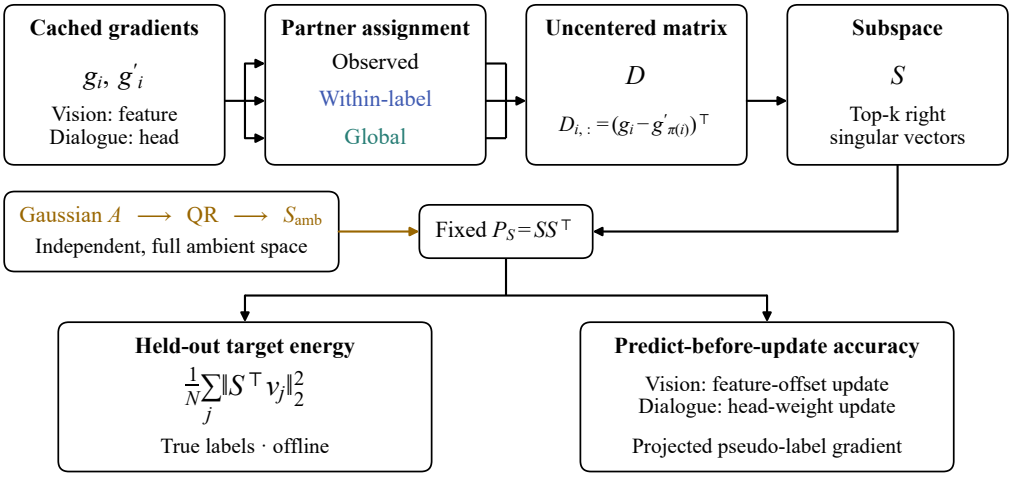


Figure 1: Each partner assignment defines its own difference matrix D and subspace S from the same cached gradients. A Gaussian basis supplies an additional reference. We compare projected energy and predict-before-update accuracy on held-out targets.

The geometric and pairing contributions can be separated by changing the reference in steps. Begin with random directions in the full space, then restrict them to the gradient support. Next fit directions after shuffling partners, then restore the original assignment. These changes in score sum to the observed gain over the ambient reference. Write S_{obs} for the original fit and S_{cond} for a fit after a within-label shuffle. The gain is $\Delta_b = \text{PE}(S_{\text{obs}}) - \mathbb{E}[\text{PE}(S_b)]$ for reference b. Then

$$\Delta_{\text{amb}} = \underbrace{\left(\frac{k}{r} - \frac{k}{d}\right)\text{tr}(T)}_{\text{support geometry}} + \underbrace{\mathbb{E}_\pi[\text{PE}(S_{\text{cond}})] - \frac{k}{r}\text{tr}(T)}_{\text{structure retained after shuffling}} + \underbrace{\Delta_{\text{cond}}}_{\text{original assignment}} . \tag{2}$$

The last term asks whether restoring the original partners helps. The shuffled fit still uses the same member gradients, so it may retain useful structure. Its advantage over support-random directions is the middle term. This can be positive even when the original assignment adds nothing or reduces energy. The terms are signed score differences, not nonnegative causal shares.

The ambiguity also occurs when gradients fill their support. Give both members of a pair the same nuisance vector. Subtraction cancels it, leaving the observed differences unchanged. Reassignment breaks the cancellation. Varying the nuisance covariance can then steer the shuffled fit toward directions with less, equal or more target energy. The observed subspace and its random-reference scores remain fixed. This gives the following result.

Proposition 1 (Random-reference nonidentification) *For every $k \geq 1$ and $d \geq 3k$, there exist three zero-mean paired-vector distributions with full marginal support of dimension $3k$. They have identical observed-difference second moments and identical target second moments. Both random-reference gains are positive and identical across distributions, while the population within-label pairing effects are positive, zero, and negative, respectively.*

The member covariances differ across the three examples; within each example, reassignment preserves its own marginals. Thus the result concerns inference from the random scores. The shuffle comparison uses additional information from the member gradients to distinguish these cases. Appendix N gives the construction and proof.

4 AUDITING THE PARTNER ASSIGNMENT

Choose the gradient coordinates. The relevant directions depend on which part of the classifier will adapt. Write its feature vector as $\phi(x) \in \mathbb{R}^{d_h}$, logits as $W\phi(x) + b$, and class probabilities as $p(x)$. For cross-entropy loss with label y, the residual $\rho(x, y) = p(x) - e_y$ measures the prediction error in class coordinates. In vision, adaptation adds a shared feature offset, so we use the feature gradient. In dialogue, adaptation changes the linear head weights, so we use the flattened weight gradient:

$$g_h(x, y) = W^\top \rho(x, y), \qquad g_W(x, y) = \mathrm{vec}_{\mathrm{row}}\big(\rho(x, y)\phi(x)^\top\big). \tag{3}$$

These coordinates impose different supports. The class residual sums to zero, so g_h lies in $\mathcal{A}_h = \{W^\top \rho : \mathbf{1}^\top \rho = 0\}$. Its dimension is at most $\min(d_h, C - 1)$ for C classes. In binary dialogue, the gradient rows are negatives of one another. The support therefore has dimension at most d_h inside $\mathbb{R}^{2d_h}$. Head biases remain fixed.

Change the partners. For each construction pair, cache gradients g_i and g'_i at the same source model. The original assignment and a permutation π produce difference rows

$$d_i^{\mathrm{obs}} = g_i - g'_i, \qquad d_i^{\pi} = g_i - g'_{\pi(i)}. \tag{4}$$

Our primary permutation is uniform within each label stratum. It preserves both gradient collections and label agreement, changing only which eligible members are paired. A second, global permutation allows assignments across labels, with each gradient retaining its original label. This distinction is substantive for feature gradients: within a label, the two label terms cancel,

$$d_i^{\mathrm{cond}} = W^\top [p(x_i) - p(z_{\pi(i)})], \tag{5}$$

whereas a global shuffle also introduces $W^\top(e_{y_{\pi(i)}} - e_{y_i})$. Global and within-label contrasts therefore answer different questions. Both use ordinary permutations, including possible fixed points; dialogue collections are permuted separately by language.

Refit at the same rank. For each assignment, we choose the k directions that retain the greatest sum of squared components of the difference vectors. Using the same fitting rule and rank makes the partner assignment the only change between these fits. Stack the differences into $D \in \mathbb{R}^{n \times d}$. Its leading right singular vectors give these directions:

$$D = U\Sigma V^\top, \qquad S = \mathrm{span}(V_{[:,1:k]}), \qquad P_S = V_{[:,1:k]} V_{[:,1:k]}^\top. \tag{6}$$

We use raw differences without centering, so both the mean and covariance can affect the directions. Rank k stays fixed across assignments; fitting uses no numerical-rank cutoff.

Ambient random references orthonormalize a Gaussian $d \times k$ matrix. Support-matched references use an orthonormal basis Q of the classifier support. This maps a Gaussian random rank-k subspace from $\mathbb{R}^r$ into the support. In vision, Q spans $W^\top(I - \mathbf{1}\mathbf{1}^\top/C)$; in binary dialogue it is $[I; -I]/\sqrt{2}$. These supports follow from head geometry. They use no held-out target labels and do not estimate a smaller intrinsic feature support.

Measure the assignment effect. Every fitted subspace is scored on the same held-out target gradients v_j, computed at the source snapshot. The score is the average squared gradient component retained by the projector:

$$\mathrm{PE}(S) = \frac{1}{|\mathcal{T}|} \sum_{j \in \mathcal{T}} \|P_S v_j\|_2^2, \tag{7}$$

$$\Delta_b = \mathrm{PE}(S_{\mathrm{obs}}) - \mathbb{E}[\mathrm{PE}(S_b)], \qquad b \in \{\mathrm{cond}, \mathrm{glob}, \mathrm{amb}, \mathrm{sup}\}. \tag{8}$$

The within-label contrast Δ_{cond} measures what the original assignment adds beyond reassignment of the stored collections. We fit each shuffled subspace separately and average its score over 12 draws. Fitting to the expected shuffled second moment gives a different quantity (Appendix K). Energy uses true target labels for offline evaluation; adaptation uses only predictions. Raw energy is unnormalized. We therefore compare scores within a model, target, seed, construction size, and rank, without pooling magnitudes across models.

5 EXPERIMENTAL DESIGN

We ask whether fitted directions retain more target-gradient energy, whether the original partners help find those directions, and whether they improve predictions. Random references, including support-matched directions, test energy retention. Partner shuffles test the original assignment. A common adaptation procedure tests prediction quality, with no update as the reference for adaptation itself.

Each run trains a source classifier and fits rank-four subspaces from labeled pairs. Vision uses source images; dialogue uses target-language texts. Separate held-out targets provide true labels for scoring energy and predictions. Online updates use only predictions. Energy is measured at the source checkpoint, and accuracy along an adapting stream. Each modality supplies three models and three targets: 18 settings in total. Within each setting, comparisons share the model, construction members, rank and evaluation examples.

Vision: an image and its altered view. The visual task is to recognize the same ten object classes when their context, shape or weather conditions change. We train on clean OOD-CV images and evaluate these three shifts in the phase-one benchmark (Zhao et al., 2022). To construct a pair, we change a training image's brightness and color while retaining its class label. The original assignment links each image to its own altered view; a within-label shuffle links it to another image's view from the same class. This tests whether the individual image–view match adds useful directions beyond the two image collections. For image values decoded once into $[0, 1]$, the altered view is

$$z = \mathrm{clip}\big(x^{0.72} \odot (1.05, 0.94, 0.88), 0, 1\big)\,. \tag{9}$$

The models are an ImageNet-initialized ResNet-18 with 512 features (He et al., 2016), a small CNN with 128 features, and a descriptor MLP with 64 features. Each of three training seeds supplies construction sets of 64 and 128 pairs. The loader divides image intensities by 255 once; partner views then follow the transform above. Sensitivity to preprocessing is examined below and in Appendix G.

Dialogue: a prompt and its reply. The dialogue task predicts a reply's quality from reply text alone. We train on English replies from OpenAssistant (Köpf et al., 2023) and evaluate German, French and Russian separately. For construction, the same classifier processes a prompt and its actual reply separately. Both gradients use the reply's quality label. The prompt gradient thus measures the classifier's response to prompt text under that label. Both gradients use a common loss and gradient coordinate system; the label still describes the reply.

Within each language, a within-label shuffle exchanges replies among prompts whose original replies share a quality label. This tests the recorded links in the gradient construction, while prediction remains a reply-only task.

Construction and evaluation use disjoint message trees in each target language. Tree-hash partitions also separate them from training. Eligible replies have a released aggregate quality score and at least three message reviews. A score of at least 0.5 defines the positive class. Separate class caps leave unequal class sizes. Frozen MiniLM and OPT-125M representations (Wang et al., 2020; Zhang et al., 2022) support linear heads; a lexical-statistics MLP supplies the third model. Appendix D gives counts, class balance and label checks.

From fixed directions to changing predictions. For each fitted subspace, we start from the same source classifier and run the same target batches in the same order. We score a batch's predictions before updating on it, so an update can affect only later scored batches. The model's predicted classes supply hard pseudo-labels; true target labels are used for scoring, not for these updates. At each step, we project the pseudo-label loss gradient onto the fitted subspace and move along that projected direction. The subspace stays fixed throughout the stream.

In vision, the updated parameter is a shared feature offset u, giving logits $W(\phi(x) + u) + b$. In dialogue, it is a head-weight offset M, giving logits $(W + M)\phi(x) + b$. Both offsets start at zero. Writing the appropriate vector of offsets as θ, each update is

$$\theta \leftarrow \theta - \eta P_S \nabla_\theta \frac{1}{|B|} \sum_{x \in B} \ell\big(\theta; x, \hat{y}_\theta(x)\big), \qquad \hat{y}_\theta(x) = \arg\max_c p_\theta(c \mid x). \tag{10}$$

Table 1: Rank-four support comparison, with dimensions shown as ambient/support. Energy fractions (%) average all three targets per model. d_P^* is the median distance between observed subspaces refitted on resampled construction data and the full-data fit; smaller values indicate less sampling variation. Dialogue uses the support-matched replication (Appendix I).

Model	d/r	Observed	Ambient	Support	d_P^*
ResNet-18	512/9	46.4	0.8	44.2	0.488
tiny CNN	128/9	63.2	3.2	44.9	0.235
descriptor MLP	64/9	79.1	6.0	44.9	0.060
MiniLM	768/384	28.2	0.5	1.0	0.605
OPT-125M	1536/768	80.9	0.3	0.5	0.419
lexical MLP	128/64	99.3	2.9	6.3	0.051

Thus changing P_S changes the allowed update directions while preserving the update rule. We recompute pseudo-labels before each step and hold them fixed during differentiation. The default is two steps per batch, learning rate 0.08 and batch size 32, without momentum or weight decay. The unconstrained reference uses $P_S = I$; no update uses $P_S = 0$. Accuracy weights batches by their actual sizes, including the final incomplete batch.

Reading the comparisons. Contrasts subtract the named reference from the original construction within each model–target setting; positive values favor the original. Energy measures gradient retention; accuracy differences are in percentage points. Estimates average three training seeds and 12 reference draws per seed, with equal weights for the two visual construction sizes. Draws vary reference subspaces, not target datasets. In the plots, Conditioned denotes within-label shuffling and Gaussian denotes the ambient random reference.

We report hierarchical percentile 95% intervals. Vision resamples seeds, then sizes and draws within them; dialogue resamples seeds and draws. The primary visual analysis and extensions use 6,000 bootstrap resamples. The original dialogue energy and accuracy comparisons use 10,000. Intervals wholly above or below zero resolve the descriptive contrast on these target samples. Intervals containing zero remain unresolved, rather than showing equivalence. Target items are not resampled, and no multiple-comparison adjustment is applied. Appendix E gives the full hierarchy.

6 Results

Useful directions remain after matching the reference space. Table 1 compares retained target-gradient energy with both random references. All 18 settings favor the observed subspace over ambient random directions. For ResNet-18, ambient directions retain 0.8% of target-gradient energy on average. Restricting them to the support raises this to 44.2%, close to the observed 46.4%. The choice of reference space accounts for much of this apparent advantage.

An advantage remains after support matching in 16 of 18 energy intervals. The two intervals containing zero are ResNet-18 under Shape and Weather shifts (Table 15, Appendix I).

The original partners rarely show a clear energy advantage. Shuffling changes the partners but keeps the gradients. Useful directions may therefore survive the shuffle, as the middle term in Equation 2 shows. Against within-label shuffling, 2 of 18 energy intervals favor the original partners, 4 favor reassignment, and 12 contain zero (Figure 2). The positive cases are German MiniLM and Weather tiny CNN.

Context descriptor MLP shows why the reference matters. Its raw energy gain is 1.2241 over ambient random directions, but -0.0616 over shuffled partners, with interval $[-0.1024, -0.0167]$. This reversal separates gradient structure from original matching, as in Equation 2.

Global shuffling gives higher energy than the original assignment in all nine dialogue settings. It also changes label agreement, so it tests a different intervention. Appendix A lists all contrasts.

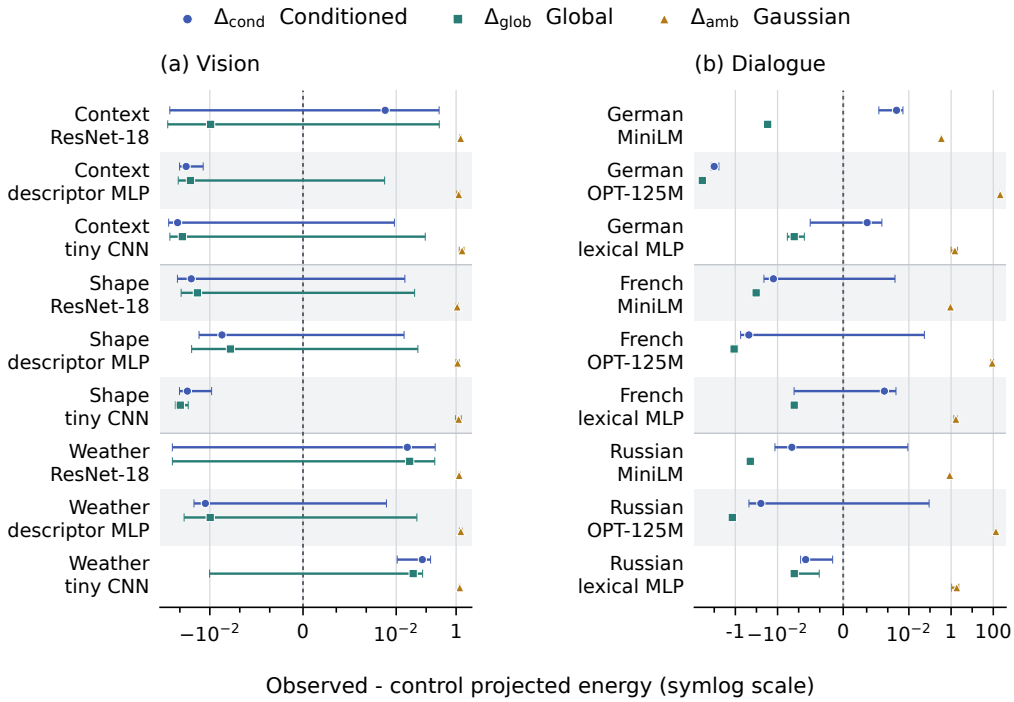


Figure 2: Observed minus control energy at rank four for nine corrected-preprocessing visual and nine dialogue comparisons. Conditioned denotes within-label shuffling; Gaussian denotes ambient random directions. Whiskers are descriptive 95% intervals. Each modality has a symmetric-log axis, linear within 10^{-4} of zero; raw magnitudes are comparable within a model and target.

An energy advantage does not ensure better predictions. Energy measures alignment with true-label gradients at the source model. Prediction gains also depend on updates driven by the model's predicted labels. Against within-label shuffling, the two positive energy cases differ in accuracy. German MiniLM has no measured difference. Weather tiny CNN gains 0.147 percentage points, with interval $[0.004, 0.281]$. This is the only positive pairing-accuracy interval among the 18 settings; the other 17 contain zero (Figure 3). Point estimates range from -0.094 to 0.330 points.

Adaptation nevertheless changes predictions. In vision, observed subspaces change 0.453% to 10.082% of predictions relative to no update across model–target means.

Overall accuracy can hide costs to one class. Overall accuracy gives more weight to the more common reply-quality label. Balanced accuracy instead averages recall equally over the two classes. For original versus shuffled pairing, all nine dialogue intervals contain zero. French MiniLM illustrates the role of class weighting: its pairing estimate is $+0.055$ ordinary accuracy points but -0.545 balanced-accuracy points (Table 12). Both remain unresolved.

The reference changes when assessing adaptation itself. Against support-matched random directions, seven ordinary accuracy intervals favor the observed subspace. Yet six balanced-accuracy intervals favor the reference, and none favor observed. Against no update, observed balanced accuracy is lower in all nine three-seed means. Energy retention and overall accuracy therefore give an incomplete picture of class outcomes.

Increasing the update budget does not consistently reveal a pairing gain. To test whether small updates hide pairing effects, eleven fixed budgets vary learning rate, step count and batch size while keeping models, streams and subspaces fixed.

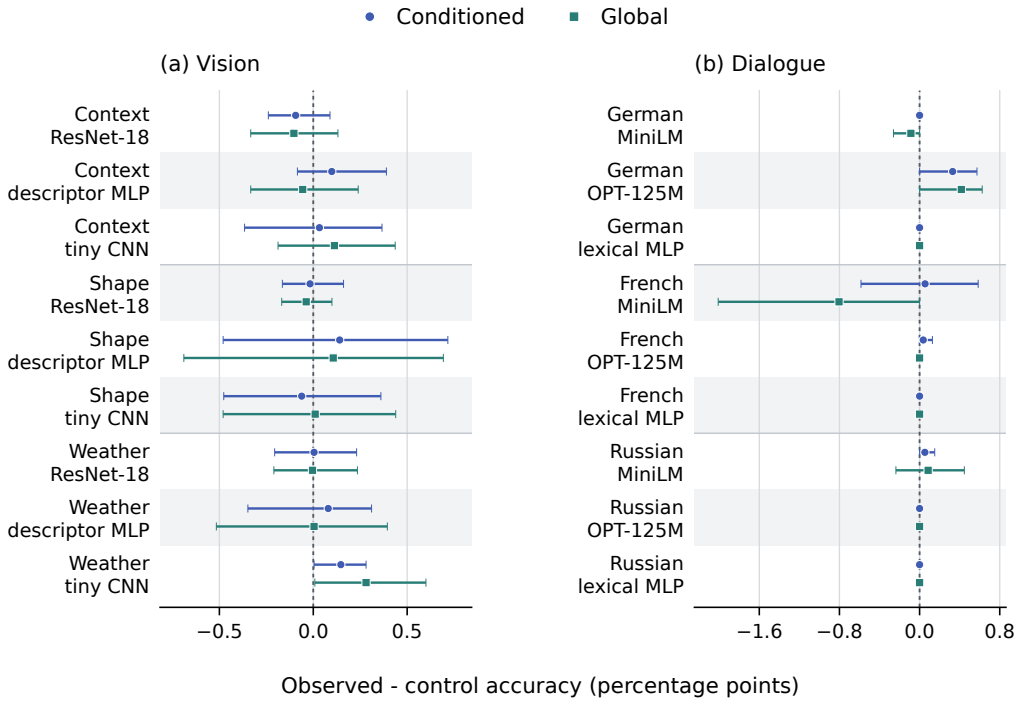


Figure 3: Accuracy difference from partner shuffles at rank four, in percentage points. Conditioned shuffles within labels; Global allows reassignment across labels. Positive values favor observed pairing; whiskers show descriptive 95% intervals. Vision uses corrected preprocessing; dialogue is unchanged.

At every budget, 15–18 pairing-accuracy intervals contain zero, with at most two positive and one negative interval (Appendix M). Larger updates can have substantial effects: the largest budget reduces visual accuracy by up to 6.91 percentage points relative to no update. Pairing gains remain uncommon across this grid even when the updates are strong enough to harm predictions.

The shuffle must actually change the fitted directions. A small pairing effect may arise because shuffling barely changes the subspace. When partner gradients are nearly constant within a label, exchanging them barely changes the difference matrix. An earlier visual implementation approached this regime by dividing already scaled partner images by 255 again. Comparing it with corrected views holds models and samples fixed.

The normalized distance between original and shuffled projectors rises from a median of 0.0026 to 0.5231 (Table 13). Zero distance means identical subspaces. The corrected audit therefore compares distinct directions.

A second appearance change, $0.8x + 0.1$, retains all nine ambient-random energy advantages. Four within-label contrasts are negative and five are unresolved. Thus mixed pairing effects persist with diverse partner outputs.

Numerical precision is a separate concern. Resampling construction examples changes the subspaces much more than switching SVD to float64 (Appendix I).

Known useful partners provide a detection check. The calibration creates a known pairing benefit using the nuisance-cancellation mechanism behind Proposition 1. Subtracting the correct partner cancels this nuisance and exposes a signal. Shuffling breaks the cancellation. Since held-out vectors lie mostly in the signal space, correct pairing should retain more energy.

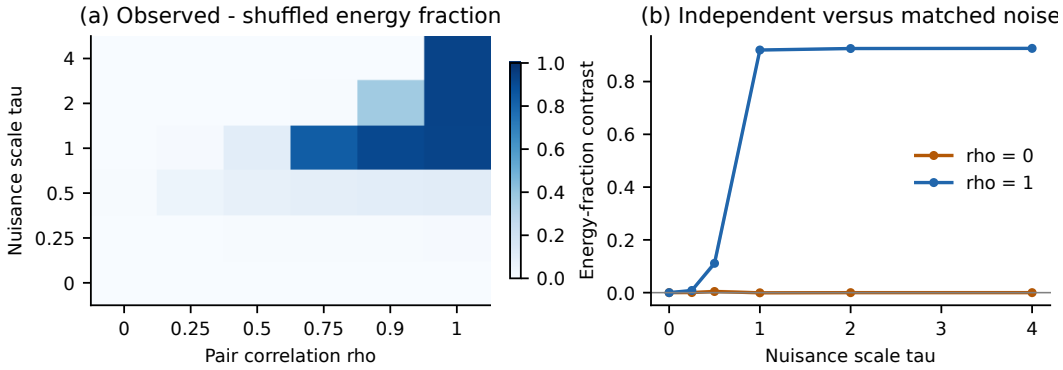


Figure 4: Known-truth calibration across the complete nuisance/correlation grid. Energy is normalized by held-out total energy. Panel (b) shows 95% intervals over twenty construction seeds, averaging twelve shuffles within each. Exact matched noise reveals signal; independent noise gives small contrasts. Appendix L includes finite-sample null behavior and the exact negative control.

Let s occupy a four-dimensional signal space and independent h_1, h_2 an orthogonal 28-dimensional nuisance space, with projectors P_s and P_h. Draw these components independently with zero mean and unit covariance within their spaces. In $a = s + \tau h_1$ and $b = \tau(\rho h_1 + \sqrt{1-\rho^2} h_2)$, τ sets nuisance strength and ρ sets how much the partners share. The paired and independently reassigned population second moments are

$$M_{\text{obs}} = P_s + 2\tau^2(1-\rho)P_h, \qquad M_{\text{shuffle}} = P_s + 2\tau^2 P_h. \tag{11}$$

When $2\tau^2(1-\rho) < 1 < 2\tau^2$, the original pairing selects signal directions while reassignment selects nuisance directions. At $\tau = 2, \rho = 1$, their measured energy fractions are 0.9352 and 0.0100, giving a contrast of 0.9252 with interval $[0.9249, 0.9255]$ (Figure 4).

With partners constant within labels, every difference stays unchanged. All 240 shuffles then give exactly the same subspace. Setting $\rho = 0$ gives a different check: only the population moments agree. Two of six finite-sample intervals remain slightly positive. The audit detects the designed benefit and respects exact invariance, but population equality need not give a zero measured contrast (Appendix L).

7 Scope and Limitations

Utility depends on the adaptation coordinates and objective. A visual feature offset acts through the shared logit bias Wu; dialogue updates change the final head weights. These are restricted updates, rather than feature learning throughout the network. Prompt–reply pairs supply gradients to a reply-quality classifier, rather than training a prompt-conditioned reward model.

Energy findings are conditional on the sampled member collections and fixed rank. Uncertainty over new target populations remains unmeasured. Within this scope, the shuffle contrast tests the original assignment. Extending it to learned, evolving subspaces or other paired learning settings requires specifying which information reassignment should preserve.

8 Conclusion

Useful gradient directions and useful original partners are different empirical claims. Random references establish an advantage of the fitted directions, but even support matching cannot determine the sign of the pairing effect. Holding member gradients fixed and changing their assignment supplies that missing comparison. Across the tested natural shifts, substantial energy advantages coexist with selective correspondence benefits and still rarer accuracy gains. Evaluating the directions, the assignment, and the resulting predictions separately gives a clearer account of what a paired construction contributes.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

Martin Gauch, Maximilian Beck, Thomas Adler, Dmytro Kotsur, Stefan Fiel, Hamid Eghbal-zadeh, Johannes Brandstetter, Johannes Kofler, Markus Holzleitner, Werner Zellinger, Daniel Klotz, Sepp Hochreiter, and Sebastian Lehner. Few-shot learning by dimensionality reduction in gradient space. In *Proceedings of The 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pp. 1043–1064, 2022. URL https://proceedings.mlr.press/v199/gauch22a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of CVPR*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 267–280. Springer, 2019. doi: 10.1007/978-3-030-28954-6_14. URL https://doi.org/10.1007/978-3-030-28954-6_14.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. OpenAssistant conversations - democratizing large language model alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 47669–47681, 2023. doi: 10.52202/075280-2064. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/949f0f8f32267d297c2d4e3ee10a2e7e-Abstract-Datasets_and_Benchmarks.html. Datasets and Benchmarks Track.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. Sparse autoencoders reveal universal feature spaces across large language models. *arXiv preprint arXiv:2410.06981*, 2024. doi: 10.48550/arXiv.2410.06981. URL https://arxiv.org/abs/2410.06981v1.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 16888–16905, 2022. URL https://proceedings.mlr.press/v162/niu22a.html.

Thomas Robert, Mher Safaryan, Ionut-Vlad Modoranu, and Dan Alistarh. LDAdam: Adaptive optimization from low-dimensional gradient statistics. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/0e6eb6f6001748b66348b9b53ad434db-Abstract-Conference.html.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. doi: 10.48550/arXiv.1910.01108. URL https://arxiv.org/abs/1910.01108.

Minhak Song, Kwangjun Ahn, and Chulhee Yun. Does SGD really happen in tiny subspaces? In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1757af1fe1429801bdf3abf5600f8bba-Abstract-Conference.html.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 9229–9248, 2020. URL https://proceedings.mlr.press/v119/sun20b.html.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of CVPR*, pp. 7201–7211, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Continual_Test-Time_Domain_Adaptation_CVPR_2022_paper.html.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Zhe Wang, Jake Grigsby, and Yanjun Qi. PGrad: Learning principal gradients for domain generalization. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=CgCmwcfgEdH.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. doi: 10.48550/arXiv.2205.01068. URL https://arxiv.org/abs/2205.01068.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.

Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. OOD-CV: A benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. In *Computer Vision – ECCV 2022*, pp. 163–180, 2022. doi: 10.1007/978-3-031-20074-8_10. URL https://doi.org/10.1007/978-3-031-20074-8_10.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61121–61143, 2024. URL https://proceedings.mlr.press/v235/zhao24s.html.

A Full Energy Contrasts

The complete energy results make the reference dependence visible in every setting. Tables 2 and 3 use rank four and report observed minus reference, so positive values favor the original assignment. Interval signs are determined before rounding. All eighteen model–target settings enter the summary.

Table 2: Corrected-preprocessing visual projected-energy differences. Entries are estimates [95% intervals]; each row uses one common gradient scale.

Model	Δ_{cond}	Δ_{glob}	Δ_{amb}
Context			
ResNet-18	0.0043 [−0.2167, 0.2704]	−0.0094 [−0.2542, 0.2740]	1.4159 [1.3161, 1.4943]
descriptor MLP	−0.0616 [−0.1024, −0.0167]	−0.0443 [−0.1134, 0.0041]	1.2241 [1.0640, 1.3266]
tiny CNN	−0.1192 [−0.2363, 0.0087]	−0.0826 [−0.2157, 0.0932]	1.5867 [1.2699, 1.8283]
Shape			
ResNet-18	−0.0417 [−0.1206, 0.0195]	−0.0259 [−0.0906, 0.0411]	1.0875 [1.0120, 1.1639]
descriptor MLP	−0.0039 [−0.0229, 0.0183]	−0.0021 [−0.0405, 0.0529]	1.1202 [0.9914, 1.2668]
tiny CNN	−0.0561 [−0.1030, −0.0087]	−0.0966 [−0.1418, −0.0526]	1.2122 [0.9512, 1.4886]
Weather			
ResNet-18	0.0235 [−0.1779, 0.2000]	0.0281 [−0.1778, 0.1941]	1.2598 [1.1945, 1.3430]
descriptor MLP	−0.0141 [−0.0338, 0.0047]	−0.0097 [−0.0722, 0.0491]	1.4347 [1.2937, 1.5780]
tiny CNN	0.0746 [0.0108, 0.1404]	0.0369 [−0.0102, 0.0762]	1.3237 [1.2646, 1.3837]

Table 3: Dialogue projected-energy differences, reported as estimates [95% intervals]. All nine settings enter the primary summary.

Model	Δ_{cond}	Δ_{glob}	Δ_{amb}
German			
MiniLM	0.00261 [0.00038, 0.00523]	−0.0295 [−0.0324, −0.0268]	0.3473 [0.3405, 0.3593]
OPT-125M	−10.09 [−13.93, −6.02856]	−36.70 [−48.28, −29.33]	212.47 [200.28, 221.47]
lexical MLP	0.00011 [−0.00028, 0.00053]	−0.0016 [−0.0034, −0.0005]	1.5158 [1.0061, 1.9960]
French			
MiniLM	−0.01569 [−0.04425, 0.00221]	−0.1033 [−0.1264, −0.0885]	0.9322 [0.9068, 0.9575]
OPT-125M	−0.22926 [−0.56179, 0.05422]	−1.1443 [−1.3164, −0.9625]	89.04 [76.45, 102.80]
lexical MLP	0.00069 [−0.00165, 0.00245]	−0.0016 [−0.0020, −0.0013]	1.6848 [1.3578, 1.9273]
Russian			
MiniLM	−0.00211 [−0.01332, 0.00895]	−0.1958 [−0.2032, −0.1839]	0.8631 [0.8477, 0.8760]
OPT-125M	−0.06188 [−0.22723, 0.09096]	−1.3863 [−2.1026, −0.9252]	132.95 [126.13, 142.54]
lexical MLP	−0.00047 [−0.00080, −0.00005]	−0.0016 [−0.0024, −0.0001]	1.8522 [1.0636, 2.3197]

B Downstream Accuracy

Prediction outcomes are reported in both relative and absolute terms. Tables 4 and 5 give observed-minus-shuffle accuracy in percentage points, with descriptive 95% intervals. Table 6 gives absolute dialogue accuracy, averaging the three seeds and, for randomized arms, twelve draws per seed.

At the original budget, corrected visual adaptation flips 0.453% to 10.082% of predictions relative to no update across means for each model and target; mean logit displacements range from 0.062 to 0.282. Against global shuffling, Weather tiny CNN gains 0.281 accuracy points $[0.008, 0.600]$. None of the eighteen dialogue accuracy intervals against either shuffle has a strictly positive lower endpoint; ten contrasts are identically zero at the run level.

Table 4: Corrected-preprocessing visual accuracy differences in percentage points, reported as estimates [95% intervals]. The no-update comparison measures adaptation itself.

Model	Within-label	Global	No update
Context			
ResNet-18	-0.094 $[-0.239, 0.089]$	-0.103 $[-0.333, 0.131]$	-0.225 $[-0.338, -0.056]$
descriptor MLP	0.099 $[-0.084, 0.389]$	-0.056 $[-0.333, 0.239]$	0.394 $[0.000, 1.126]$
tiny CNN	0.033 $[-0.366, 0.366]$	0.113 $[-0.188, 0.436]$	0.113 $[-0.338, 0.450]$
Shape			
ResNet-18	-0.017 $[-0.165, 0.161]$	-0.038 $[-0.168, 0.099]$	0.082 $[0.000, 0.329]$
descriptor MLP	0.141 $[-0.480, 0.717]$	0.106 $[-0.689, 0.693]$	0.000 $[-0.617, 0.741]$
tiny CNN	-0.062 $[-0.477, 0.360]$	0.010 $[-0.480, 0.439]$	-0.082 $[-0.617, 0.576]$
Weather			
ResNet-18	0.004 $[-0.206, 0.231]$	-0.004 $[-0.210, 0.235]$	-0.050 $[-0.352, 0.151]$
descriptor MLP	0.080 $[-0.348, 0.311]$	0.004 $[-0.516, 0.394]$	0.050 $[-1.208, 1.057]$
tiny CNN	0.147 $[0.004, 0.281]$	0.281 $[0.008, 0.600]$	-0.050 $[-0.352, 0.201]$

Table 5: Dialogue observed-minus-shuffle accuracy in percentage points, reported as estimates [95% intervals].

Model	Within-label	Global
German		
MiniLM	0 $[0, 0]$	-0.087 $[-0.260, 0]$
OPT-125M	0.330 $[0, 0.573]$	0.417 $[0, 0.625]$
lexical MLP	0 $[0, 0]$	0 $[0, 0]$
French		
MiniLM	0.055 $[-0.585, 0.585]$	-0.804 $[-2.010, 0]$
OPT-125M	0.037 $[0, 0.128]$	0 $[0, 0]$
lexical MLP	0 $[0, 0]$	0 $[0, 0]$
Russian		
MiniLM	0.053 $[0, 0.151]$	0.085 $[-0.236, 0.447]$
OPT-125M	0 $[0, 0]$	0 $[0, 0]$
lexical MLP	0 $[0, 0]$	0 $[0, 0]$

Table 6: Mean dialogue accuracy (%) under no update and three adaptation procedures.

Target	Model	No update	Unconstrained	Observed	Gaussian
German	MiniLM	79.79	81.67	81.25	79.97
	OPT-125M	18.12	14.58	15.00	17.92
	lexical MLP	67.71	74.58	74.58	68.00
French	MiniLM	75.44	79.82	77.63	75.38
	OPT-125M	69.52	82.89	82.68	70.63
	lexical MLP	69.30	71.93	71.93	69.94
Russian	MiniLM	77.38	78.09	78.17	77.37
	OPT-125M	55.95	59.10	59.26	56.73
	lexical MLP	66.04	68.95	68.95	66.40

C Algebraic Intervention and Rank Sensitivity

Algebraic diagnostic. The supplementary algebraic diagnostic replaces the first gradient in each pair by twice its partner. The observed input becomes $2g'_i - g'_i = g'_i$, while a global shuffle produces $2g'_i - g'_{\pi(i)}$. Table 7 gives the resulting rank-four energy contrast. All six transformer cases have positive intervals, and the three lexical cases have intervals containing zero.

If the partner collection has mean μ and covariance Σ (divisor n), the observed second moment is $\Sigma + \mu\mu^\top$, and the expected globally shuffled second moment is $5\Sigma + \mu\mu^\top$. When $\mu = 0$, these matrices have the same eigenspaces. The diagnostic can thus preserve the population directions even though it changes the difference vectors. Its observed contrast describes this algebraic intervention; the known-correspondence construction in Appendix L supplies a separate sensitivity check.

Table 7: Algebraic diagnostic: observed-minus-global projected energy at rank four, reported as estimates [95% intervals].

Target	Model	Estimate [95% interval]
German	MiniLM	$0.01424\ [0.01208, 0.01669]$
German	OPT-125M	$0.35659\ [0.30513, 0.42009]$
	lexical MLP	$0.00011\ [-0.00011, 0.00035]$
French	MiniLM	$0.01057\ [0.00797, 0.01348]$
	OPT-125M	$0.13566\ [0.06059, 0.20836]$
	lexical MLP	$0.00005\ [-0.00012, 0.00018]$
Russian	MiniLM	$0.00221\ [0.00114, 0.00331]$
	OPT-125M	$0.04280\ [0.01284, 0.07522]$
	lexical MLP	$0.00012\ [-0.00005, 0.00044]$

Dialogue rank sensitivity Table 8 gives dialogue means at ranks 1, 2, 4, and 8. German MiniLM has positive within-label means throughout; French MiniLM changes from 0.01162 at rank two to -0.01569 at rank four. Gaussian means stay positive throughout, while global-shuffle means are negative for every primary transformer case. Thus persistent random-basis advantages coexist with rank-dependent pairing effects. Visual draws at ranks 2, 4, 6, and 8 are included in the supplement.

Table 8: Mean observed-minus-within-label energy across dialogue ranks. All nine dialogue settings are included.

Target	Model	$k = 1$	$k = 2$	$k = 4$	$k = 8$
German	MiniLM	0.01671	0.01228	0.00261	0.00513
German	OPT-125M	−9.25698	3.07019	−10.09	−6.55044
	lexical MLP	−0.00121	−0.02651	0.00011	−0.00003
French	MiniLM	−0.00025	0.01162	−0.01569	−0.03036
	OPT-125M	−1.53871	−1.50229	−0.22926	−0.79101
	lexical MLP	0.00403	−0.15879	0.00069	−0.00007
Russian	MiniLM	0.02438	0.00535	−0.00211	−0.00081
	OPT-125M	−0.18338	−0.12162	−0.06188	−0.20364
	lexical MLP	−0.00160	0.00210	−0.00047	0.00005

D Labels and Utility Across Classes

An independent reconstruction from the released parquet fields matches every ordered source, construction, and evaluation manifest for every language, model, and seed. Label 1 means aggregate `quality`≥ 0.5; eligibility uses message-level `review_count`≥ 3. All nine model–language settings use this same label mapping and enter the analysis, including German OPT-125M. Low target accuracy does not alter a label or remove a setting.

Table 9 gives the complete selected counts. Seeds change order but not the evaluation class totals; class caps do not enforce balance. The released quality-specific rating count is recorded separately

Table 9: Selected class counts for every language, partition, and seed. Grouped seed rows have identical counts in each seed. Majority is the constant-class accuracy within that row's partition; evaluation rows supply the task baseline.

Language	Split	Seed	Class 0	Class 1	Majority (%)
English	source training	0	2020	2500	55.31
German	construction	1001–1003	31	128	80.50
German	evaluation	1001–1003	21	139	86.88
French	construction	1001–1003	37	128	77.58
French	evaluation	1001–1003	22	130	85.53
Russian	construction	1001–1003	108	128	54.24
Russian	evaluation	1001–1003	91	332	78.49

Table 10: German OPT confusion counts; N_{ab} has true class a and predicted class b. Summing columns gives the complete prediction distribution. BA is balanced accuracy (%).

Seed	Arm	N_{00}	N_{01}	N_{10}	N_{11}	BA
1001	no update	21	0	129	10	53.60
1001	paired	21	0	136	3	51.08
1002	no update	21	0	131	8	52.88
1002	paired	21	0	136	3	51.08
1003	no update	21	0	133	6	52.16
1003	paired	21	0	136	3	51.08

from message review count in the audit artifacts. Source sampling uses seed zero; model training uses seeds 1001–1003.

German OPT-125M predicts class 0 for 150, 152, and 154 of 160 examples before adaptation (Table 10), although class 1 is the majority. Adaptation in the observed subspace increases the count for class 0 to 157 in each seed. Mean accuracy changes from 18.125% to 15%; balanced accuracy changes from 52.88% to 51.08%. The independently verified labels locate this failure in the predictions: the model favors the minority class under the German shift despite its higher English source accuracy.

Table 11 complements ordinary accuracy with the mean recall of the two classes. At the original budget, no observed dialogue arm exceeds its language's majority accuracy in the three-seed mean. Several retain balanced accuracy above 50%, showing why both summaries matter. Against support-random, seven ordinary accuracy intervals favor observed, but six balanced accuracy intervals favor the reference and none favor observed. These comparisons establish class-dependent behavior rather than a uniform quality-prediction improvement. Complete seed-level confusion matrices, class distributions, macro-F1, and metrics for all control draws are retained.

Class-balanced pairing effects. Table 12 evaluates the primary correspondence comparison itself, keeping the original budget and ordinary accuracy. It reaggregates the saved confusion matrices, with no additional training or target selection. Ordinary accuracy weights each class by its frequency, whereas balanced accuracy gives the classes equal weight. Improving the more frequent class can therefore raise ordinary accuracy even if performance averaged equally across classes falls. Let R_c be recall for class c, π_c its fixed target proportion, and Δ denote observed minus within-label shuffle. Then

$$\Delta\mathrm{Acc} = \pi_0 \Delta R_0 + \pi_1 \Delta R_1, \qquad \Delta\mathrm{BA} = \tfrac{1}{2}(\Delta R_0 + \Delta R_1). \tag{12}$$

French MiniLM has mean recall changes of -1.389 points for class 0 and $+0.299$ for class 1, producing opposite signs for accuracy and BA. Its BA contrast is -0.545 points with interval $[-2.020, 0.342]$: the observed direction reverses, but neither sign is resolved. Across all nine settings, BA point estimates range from -0.545 to 0.190 points and every interval contains zero. Five settings have identical confusion counts by class in every matched comparison of seeds and draws, yielding exact zero contrasts for both metrics on these streams.

The no-update columns separate adaptation from correspondence. Observed BA falls below no-update BA in all nine three-seed means, even though ordinary accuracy rises in eight. A small pairing contrast can therefore coexist with a deterioration in class-balanced utility shared by the compared

Table 11: Dialogue utility at the original budget, averaged over three seeds. Accuracies and BA are percentages; the final column is observed minus support-random BA in points with a descriptive 95% interval. Target majority baselines are German 86.88%, French 85.53%, Russian 78.49%.

Setting	No update	Observed	BA	ΔBA vs support
German / MiniLM	79.79	81.25	62.26	-0.57 [-1.19,0.44]
German / OPT-125M	18.12	15.00	51.08	-1.62 [-2.23,-1.08]
German / lexical MLP	67.71	74.58	67.85	0.33 [-0.89,1.44]
French / MiniLM	75.44	77.63	57.34	-3.10 [-4.90,-1.08]
French / OPT-125M	69.52	82.68	57.14	-1.94 [-3.07,-0.18]
French / lexical MLP	69.30	71.93	69.74	-1.27 [-2.40,-0.31]
Russian / MiniLM	77.38	78.17	50.20	-1.16 [-1.55,-0.52]
Russian / OPT-125M	55.95	59.26	50.91	-3.13 [-4.07,-2.15]
Russian / lexical MLP	66.04	68.95	68.39	-1.25 [-4.59,0.54]

Table 12: Primary dialogue pairing contrasts at the original budget ($\eta = 0.08$, two steps per batch, batch size 32). BA columns are percentages averaged over three seeds; differences are observed minus within-label shuffle in percentage points. Brackets give descriptive 95% intervals from 6,000 hierarchical resamples of three seeds and twelve draws per seed. Both metrics use the same resamples; target examples and trees are fixed, with no multiplicity correction.

Setting	BA no update	BA observed	ΔAccuracy	ΔBA
German / MiniLM	62.77	62.26	0.000 [0.000,0.000]	0.000 [0.000,0.000]
German / OPT-125M	52.88	51.08	0.330 [0.000,0.573]	0.190 [0.000,0.330]
German / lexical MLP	67.94	67.85	0.000 [0.000,0.000]	0.000 [0.000,0.000]
French / MiniLM	60.47	57.34	0.055 [-0.585,0.585]	-0.545 [-2.020,0.342]
French / OPT-125M	58.89	57.14	0.037 [0.000,0.128]	0.021 [0.000,0.075]
French / lexical MLP	70.72	69.74	0.000 [0.000,0.000]	0.000 [0.000,0.000]
Russian / MiniLM	51.56	50.20	0.053 [0.000,0.151]	0.033 [0.000,0.096]
Russian / OPT-125M	57.45	50.91	0.000 [0.000,0.000]	0.000 [0.000,0.000]
Russian / lexical MLP	69.46	68.39	0.000 [0.000,0.000]	0.000 [0.000,0.000]

adaptation arms. The matched intervals below use the extension's 6,000 common bootstrap resamples for both metrics; the historical accuracy intervals in Table 5 retain their original 10,000 resamples.

E Implementation and Reproducibility

Models, preprocessing, and optimization Source training uses AdamW with learning rate 0.003, weight decay 10^{-4}, and batch size 64. Training lasts for 10 epochs in vision and 12 in dialogue. Learning rate and weight decay are explicitly supplied. Omitted optimizer arguments use the recorded library defaults, including $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and AMSGrad disabled.

Target adaptation uses direct gradient descent with learning rate 0.08, batch size 32, and two steps per batch, following Equation (10). Feature extraction uses batches of 128 images or 32 texts, distinct from the training and adaptation batches. Gradients are computed in float32. Experiments run on CPU with Python 3.12 and PyTorch 2.12.1+cpu.

Visual models and intensity handling. The ResNet-18 encoder uses pretrained ImageNet weights. Convolutional weights are fixed during source training, while batch normalization buffers update in training mode and freeze for feature extraction. The small CNN uses convolutional channels 32, 64, and 128, ReLU activations, two max-pooling stages, and global average pooling. The descriptor MLP maps 18 channel statistics through layers of width 96 and 64. The statistics are the mean, standard deviation, and four intensity quantiles per channel. Both small encoders are trained with their heads.

Mean accuracies on the source training set across targets are approximately 0.933 for ResNet-18, 0.677 to 0.696 for the small CNN, and 0.317 to 0.320 for the descriptor MLP.

Images are decoded with automatic orientation and resized to 96×96 using ImageMagick. The uint8 loader divides intensities by 255 once and performs no normalization using the ImageNet mean or

standard deviation. Corrected partner views follow Equation (9) and bypass the loader's conversion. The double-scaled diagnostic explicitly uses $z/255$; the additional contrast transform is $0.8x + 0.1$.

The matched visual study retrains source models with the original settings, samples, and seed order, then evaluates all three transforms within each trained model. It uses four CPU threads per seed rather than the historical 24. Historical trained checkpoints were not saved; the new tiny-CNN results are not bitwise reproductions of the earlier training runs. The legacy gradient branch is checked for exact equality with the original implementation on each newly trained model. Each new checkpoint and the complete sample identities are retained.

All 162 observed-pair difference matrices have numerical rank nine under the diagnostic tolerance $\max(n, d)\epsilon_{32}\sigma_1$. This tolerance never changes the fitted basis.

Dialogue encoders and labels. MiniLM and OPT-125M truncate tokenized inputs to 128 tokens and use mean-pooled representations with 384 and 768 coordinates, respectively. The lexical MLP maps eight statistics through widths 96 and 64: log character count, log token count, uppercase fraction, digit fraction, the fraction of `!?`, the fraction of `.,;:`, alphabetic fraction, and newline fraction. It uses ReLU activations and a binary linear head.

Transformer pooling is weighted by the attention mask. Accuracies on the source training set are 0.69 to 0.70 for MiniLM, 0.74 to 0.75 for OPT-125M, and about 0.67 for the lexical MLP, against a majority rate of 0.553.

Labels use the released aggregate quality score and each message's review count; this does not require three ratings specifically for quality. We remove deleted or empty messages, pairs whose parent is not a prompter or whose child is not an assistant, and pairs with mismatched language tags. The class cap is applied separately to each label and can leave unequal counts. The matched replication retains each ordered tree identifier, text hash, and label, and the independent audit verifies the mapping and reports all selected class counts in Appendix D.

Sampling and uncertainty The dialogue partition uses the first 16 hexadecimal digits of SHA-256 of the message tree identifier, modulo ten. Source sampling uses seed zero, target construction uses the training seed, and target evaluation uses that seed plus one. Draw seeds are derived deterministically from the training seed, rank, and draw index, with distinct offsets for each arm.

Visual construction samples without replacement from the source training pool. Both permutation schemes are uniform permutations, not derangements. Global permutations have one expected index fixed point per collection; conditioned permutations have one per nonempty stratum on average.

For visual intervals, each resample first draws three seed clusters with replacement. It then draws two construction sizes within each selected seed and twelve control draws within each selected size. The reported estimate gives the two sizes equal weight; the interval additionally varies their empirical mixture. Dialogue resamples the three seed clusters and then twelve draws within each selected seed.

Control draws approximate the reference expectation; they are not independent target datasets. Neither hierarchy resamples target examples or message trees. Bootstrap intervals over target items stored inside individual dialogue draws are separate diagnostics. They are not substituted for the reported hierarchical contrasts.

The interval endpoints are the 2.5th and 97.5th bootstrap percentiles. We classify signs using unrounded endpoints: lower endpoint above zero, upper endpoint below zero, or neither. The matched visual analysis uses bootstrap seed 20260923 and common resamples across transforms, with the same hierarchy of seeds, sizes and draws. A zero-width accuracy interval means every contributing aggregate difference is identical; it does not prove equality on new samples or identical predictions on every item.

Preserved research artifacts The research package contains configurations checked against pre-execution SHA-256 records, execution receipts, model revisions, dataset checksums, and results for each seed, rank, and control draw. The archived inputs distinguish corrected primary results from the historical visual branch. The visual correction adds sample manifests, probabilities, gradients, spectra, predictions, and checkpoint hashes.

The support-matched extension reuses the visual checkpoints. It saves newly fitted dialogue heads, lexical encoders, features, gradients, ordered tree identifiers, text hashes and labels, complete spectra, and bootstrap distances. Independent NumPy checks verify all 81 observed spectra and numerical ranks and all 972 sampled support-reference energies. Historical dialogue results remain separate. Their common-arm accuracies match the replication exactly; the largest raw-energy discrepancy is 0.00048.

The fixed-budget extension reuses these models and exactly reproduces reference-budget accuracy in all 81 seed, size and setting records. It retains 11 budgets, all arm confusion matrices, independent label reconstructions from data fields and the complete known-truth calibration. A separate calculation verifies 24,651 metrics from confusion matrices. These extensions reuse previously evaluated targets; they are not a newly sampled benchmark.

F Exploratory AG News Comparison

We use a publicly available DistilBERT classifier (Sanh et al., 2019) fine-tuned on AG News (Zhang et al., 2015).[1] This exploratory comparison uses 240 construction articles, 600 held-out articles, rank 32, and 32 Gaussian bases, and is excluded from the primary summary. Character swaps, word drops, and token masking define three injected shifts, each compared with a global partner shuffle.

Clean accuracy is 0.942, and corrupted accuracy ranges from 0.917 to 0.922. Observed-minus-Gaussian energy is 0.3281 $[0.3177, 0.3389]$ for character swaps, 0.2344 $[0.2254, 0.2438]$ for word drops, and 0.3092 $[0.2994, 0.3189]$ for token masking. The corresponding observed-minus-global values are 0.0023 $[-0.0024, 0.0071]$, -0.0376 $[-0.0419, -0.0329]$, and -0.0133 $[-0.0180, -0.0086]$.

G Visual Preprocessing Sensitivity

Table 13 tests whether near-invariance to partner shuffling comes from similar partner outputs. Let V_p be the trace variance of partner probabilities divided by that of clean probabilities, and V_g the analogous ratio of pooled within-label gradient variances. The double-scaled view suppresses both, reducing the expected perturbation of D under shuffling (Equation (13)). Its median normalized projector distance $d_P = \|P_{\mathrm{obs}} - P_{\mathrm{cond}}\|_F/\sqrt{2k}$ is only 0.0026. Correcting the normalization restores output diversity: every probability-variance ratio is at least 0.715, and the median cell-mean projector distance becomes 0.5231.

The affine-contrast view retains diverse outputs and distinct subspaces while preserving the random-basis advantage in all nine settings. Its within-label energy intervals favor shuffling in four settings and include zero in five. Context tiny CNN nevertheless has a positive accuracy contrast of 0.258 points with interval $[0.094, 0.422]$. These matched views separate preprocessing-induced near-invariance from mixed pairing effects with noncollapsed outputs, and separate energy from downstream utility.

Table 14 reports every double-scaled and affine-contrast cell from the matched rerun; corrected primary cells appear in Tables 2 and 4. The double-scaled rows are recomputed with the same source models as the corrected rows. Earlier historical tiny-CNN values remain separately archived.

The size of the shuffle perturbation can be related directly to partner diversity. For label c, let n_c be the number of pairs, $w_c = n_c/n$, and Σ'_c the covariance of partner gradients computed with divisor n_c. For the fixed collection, averaging over within-label permutations gives

$$\mathbb{E}_\pi \|D_\pi - D_{\mathrm{obs}}\|_F^2 = 2n \sum_c w_c \operatorname{tr}(\Sigma'_c). \tag{13}$$

Indeed, each partner in class c is uniformly assigned among its n_c members, and the sum of their mean squared pairwise distances is twice the sum of squared distances to the class mean. This also applies to permutations with fixed points. If partner gradients are constant within each class, every conditioned shuffle leaves D unchanged. For visual gradients, near-constant partner probabilities can induce this regime. Small nonzero variance alone does not guarantee a stable leading subspace without a spectral gap; Table 13 therefore reports measured projector distances as well as variance ratios.

[1] Public checkpoint provided by TextAttack.

Table 13: Matched visual preprocessing comparison at rank four. V_p, V_g are medians over 54 seed–size–model–target records; d_P is the median of nine cell means. Final columns count positive/negative/zero-containing within-label energy intervals and positive Gaussian intervals.

Partner view	V_p	V_g	d_P	Within-label	Gaussian
Double-scaled	4.27×10^{-5}	7.79×10^{-5}	0.0026	0/2/7	9/9
Corrected	0.871	1.015	0.5231	1/2/6	9/9
Affine contrast	0.915	0.871	0.5163	0/4/5	9/9

Table 14: Matched sensitivity results at rank four. Energy and accuracy columns are observed minus within-label shuffle, with 95% intervals; accuracy is in percentage points. d_P is mean normalized projector distance.

Target	Model	Energy	Accuracy	d_P
Double-scaled				
Context	ResNet-18	-0.00056 $[-0.00090, -0.00016]$	0.000 $[0.000, 0.000]$	0.0026
Context	descriptor MLP	-0.00028 $[-0.00051, -0.00004]$	0.000 $[0.000, 0.000]$	0.0027
Context	tiny CNN	-0.00048 $[-0.00130, 0.00008]$	0.000 $[0.000, 0.000]$	0.0045
Shape	ResNet-18	-0.00001 $[-0.00021, 0.00014]$	0.000 $[0.000, 0.000]$	0.0029
Shape	descriptor MLP	0.00001 $[-0.00013, 0.00011]$	-0.003 $[-0.017, 0.000]$	0.0023
Shape	tiny CNN	0.00024 $[-0.00010, 0.00065]$	0.000 $[0.000, 0.000]$	0.0026
Weather	ResNet-18	-0.00009 $[-0.00042, 0.00016]$	0.000 $[0.000, 0.000]$	0.0026
Weather	descriptor MLP	0.00007 $[-0.00011, 0.00020]$	0.000 $[0.000, 0.000]$	0.0017
Weather	tiny CNN	0.00004 $[-0.00006, 0.00015]$	0.000 $[0.000, 0.000]$	0.0012
Affine contrast				
Context	ResNet-18	-0.00479 $[-0.25262, 0.32384]$	-0.038 $[-0.113, 0.028]$	0.5410
Context	descriptor MLP	-0.18123 $[-0.23780, -0.11869]$	-0.089 $[-0.760, 0.450]$	0.4261
Context	tiny CNN	-0.19197 $[-0.35715, 0.00033]$	0.258 $[0.094, 0.422]$	0.5163
Shape	ResNet-18	-0.01467 $[-0.09749, 0.05536]$	0.014 $[-0.089, 0.113]$	0.5352
Shape	descriptor MLP	-0.05137 $[-0.06780, -0.03807]$	0.175 $[-0.398, 0.662]$	0.4304
Shape	tiny CNN	-0.05183 $[-0.10037, -0.00822]$	0.137 $[-0.158, 0.432]$	0.5547
Weather	ResNet-18	0.12214 $[-0.03760, 0.28479]$	-0.117 $[-0.537, 0.239]$	0.5037
Weather	descriptor MLP	-0.05524 $[-0.12701, -0.01271]$	0.126 $[-0.063, 0.319]$	0.3996
Weather	tiny CNN	-0.03677 $[-0.10607, 0.02123]$	-0.147 $[-0.382, 0.076]$	0.5307

H Shuffle Symmetry

Shuffles with unchanged geometry. Some reassignments change only the order of the difference rows, leaving their geometry intact. To see this, let π move partners only among indices with the same first gradient and label. Then $g_i = g_{\pi(i)}$, so

$$d_i^{\pi} = d_{\pi(i)}^{\mathrm{obs}}, \qquad D_\pi = R_\pi D_{\mathrm{obs}}, \qquad D_\pi^\top D_\pi = D_{\mathrm{obs}}^\top D_{\mathrm{obs}}, \tag{14}$$

where R_π is a row-permutation matrix. For $k < d$, a gap between the kth and $(k+1)$st largest eigenvalues of this common Gram matrix makes the leading k-dimensional projector unique, so the projected target energy is unchanged. At a tie, the Gram matrix is still invariant, but the selected subspace need not be unique.

Repeated same-label prompts can produce such groups because their cached first gradients coincide. This gives an exact mechanism by which changing partner indices leaves the difference Gram matrix unchanged. The argument applies to permutations confined to these groups; their frequency among the sampled shuffles was not measured.

I Support References and Stability

The extension evaluates rank four with twelve random draws and the original three seeds per modality. Vision uses both construction sizes and the corrected transform. It reuses the saved models and gradient arrays, reproducing every common-arm energy and accuracy exactly. Dialogue follows the

original training configuration and model/domain order, retaining a matched comparison among all five controls within each new source model. Frozen source features are cached without changing their ordered inputs. Source checkpoints, gradient arrays, and all intermediate spectra are retained.

For vision, an SVD of the row-centered head weights in float64 constructs Q, with threshold $\max(C, d_h)\epsilon_{64}\sigma_1$; all resulting supports have dimension nine. For dialogue, $Q = [I; -I]/\sqrt{2}$ follows directly from binary softmax. Relative Frobenius residuals of all left, right, and target gradient arrays outside these spans are checked below 10^{-5}. Support-random draws use a distinct seed offset from the ambient reference. No target label or target covariance enters either reference construction.

Table 15 reports every matched-reference contrast, including German OPT-125M. Its intervals use 6,000 bootstrap resamples with seed 20260923. They follow the primary hierarchy: seeds, sizes and draws for vision, or seeds and draws for dialogue. Historical dialogue tables retain their original 10,000-resample intervals.

Within-label energy counts in the matched extension remain one positive, two negative, and six zero-containing for vision. They are also one positive, two negative, and six zero-containing for all nine dialogue settings. Support and partner comparisons thus use the same fitted models within this extension.

Table 15: Observed minus support-matched random reference: raw gradient energy and prequential accuracy (percentage points), with descriptive 95% intervals. Each row has its own energy scale. All eighteen settings are included.

Target / model	Energy difference [95% interval]	Accuracy pp [95% interval]
Context / ResNet-18	0.0674 [0.0151, 0.1193]	−0.066 [−0.188, 0.070]
Context / descriptor MLP	0.5718 [0.4844, 0.6445]	0.253 [−0.080, 0.680]
Context / tiny CNN	0.4342 [0.2258, 0.6292]	0.206 [−0.099, 0.530]
Shape / ResNet-18	0.0434 [−0.0712, 0.1099]	−0.027 [−0.158, 0.117]
Shape / descriptor MLP	0.5206 [0.4026, 0.6585]	0.041 [−0.741, 0.545]
Shape / tiny CNN	0.3734 [0.2205, 0.5043]	−0.014 [−0.425, 0.326]
Weather / ResNet-18	0.0640 [−0.0120, 0.1638]	0.042 [−0.164, 0.235]
Weather / descriptor MLP	0.6863 [0.5271, 0.8057]	−0.227 [−0.810, 0.290]
Weather / tiny CNN	0.4460 [0.3990, 0.4941]	0.084 [−0.147, 0.432]
German / lexical MLP	1.4731 [0.9934, 1.9160]	6.337 [0.677, 13.038]
German / MiniLM	0.3343 [0.3253, 0.3459]	1.250 [0.799, 1.562]
German / OPT-125M	211.6582 [199.7634, 220.3155]	−2.813 [−3.872, −1.875]
French / lexical MLP	1.6084 [1.2872, 1.8366]	1.864 [1.352, 2.303]
French / MiniLM	0.9130 [0.8878, 0.9370]	2.230 [0.786, 4.605]
French / OPT-125M	88.7289 [76.2447, 102.3216]	10.764 [7.456, 15.260]
Russian / lexical MLP	1.7923 [1.0190, 2.2439]	2.338 [1.248, 3.441]
Russian / MiniLM	0.8537 [0.8395, 0.8668]	0.880 [0.604, 1.064]
Russian / OPT-125M	132.5707 [125.8969, 142.0530]	1.957 [−8.366, 8.432]

Spectra and numerical precision We retain complete spectra for each observed matrix and all twelve global and conditioned assignments: 1,350 visual and 675 dialogue matrices. Numerical rank counts singular values above $\tau = \max(n, d)\epsilon_{32}\sigma_1$; stable rank is $\sum_j \sigma_j^2/\sigma_1^2$. This threshold diagnoses the stored float32 gradients and never changes the fitted projector.

Figure 5 displays normalized observed spectra. Table 16 gives the rank-four relative gap $(\sigma_4 - \sigma_5)/\sigma_4$. It also compares full ambient-space float32 and float64 SVDs using d_P. Small numerical differences coexist with variable gaps. Rank sufficiency and precision stability are therefore distinct from sampling stability.

Resampling stability For each observed matrix and the fixed conditioned assignment from draw zero, we construct 32 bootstrap samples. Each sample yields a new rank-four fit in float64 support coordinates. Vision resamples complete pairs within source labels. Dialogue resamples complete construction trees with replacement, keeping every edge of each selected tree. Each assignment's full-sample subspace serves as the reference, rather than its bootstrap mean. We compare orthonormalized

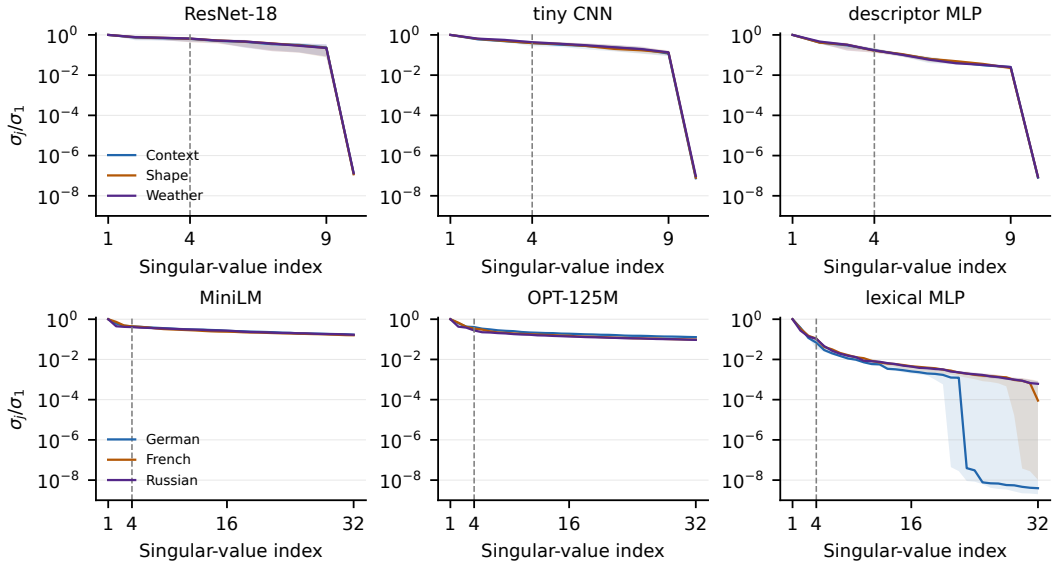


Figure 5: Normalized observed singular spectra. Lines are medians and shaded bands are minima to maxima over seeds and, in vision, construction sizes; bands are not confidence intervals. The dashed line marks rank four. Visual panels show ten singular values and dialogue panels show the first 32; complete spectra are retained. German OPT is included in the primary analysis.

bases using

$$d_P(U,V) = \frac{\|UU^\top - VV^\top\|_F}{\sqrt{2k}} = \frac{\|(I - VV^\top)U\|_F}{\sqrt{k}}.$$

This diagnostic varies the construction sample while holding the target stream fixed. It neither permutes partners anew inside each resample nor supplies population-level uncertainty for target accuracy. Table 16 reports observed-bootstrap medians; all 5,184 observed and conditioned bootstrap distances are supplied. No stability threshold is used to exclude settings or select a rank.

Table 16: Numerical and sample stability. Numerical rank is the minimum over all natural assignments; stable rank and median gap summarize observed matrices. Minimum gap and maximum precision distance cover all assignments. Bootstrap distance is the median over observed constructions and resamples.

Target / model	**Rank**	**Stable rank**	**Min gap**	**Med. gap**	**Max $d_P^{32,64}$**	**Boot. d_P**
Context / ResNet-18	9	3.18	0.006	0.122	4.1e-06	0.489
Context / descriptor MLP	9	1.32	0.019	0.468	1.4e-06	0.053
Context / tiny CNN	9	2.06	0.006	0.113	3.6e-06	0.227
Shape / ResNet-18	9	3.31	0.019	0.122	3.8e-06	0.494
Shape / descriptor MLP	9	1.33	0.024	0.296	3.5e-06	0.071
Shape / tiny CNN	9	2.16	0.025	0.120	6.3e-06	0.267
Weather / ResNet-18	9	3.31	0.023	0.108	3.5e-06	0.481
Weather / descriptor MLP	9	1.39	0.032	0.463	3.6e-06	0.060
Weather / tiny CNN	9	2.28	0.014	0.161	4.5e-06	0.210
German / lexical MLP	20	1.16	0.148	0.369	1.2e-06	0.077
German / MiniLM	159	4.55	0.025	0.074	3.7e-06	0.616
German / OPT-125M	159	3.33	0.022	0.136	4.4e-06	0.403
French / lexical MLP	28	1.15	0.513	0.606	4.6e-07	0.042
French / MiniLM	162	4.33	0.032	0.083	2.6e-06	0.552
French / OPT-125M	162	2.70	0.014	0.139	4.6e-06	0.448
Russian / lexical MLP	32	1.11	0.510	0.565	1.0e-06	0.048
Russian / MiniLM	236	4.11	0.006	0.064	1.4e-05	0.628
Russian / OPT-125M	236	2.32	0.018	0.152	2.4e-06	0.425

J Second Moments and Random References

Reassignment preserves the separate gradient collections while changing their cross terms. Expanding the second moment whose leading eigenvectors define the fitted subspace makes this distinction explicit. Write $M = \mathbb{E}_n[gg^\top]$, $M' = \mathbb{E}_n[g'g'^\top]$, and $C_\pi = \mathbb{E}_n[gg'^\top_\pi]$, where $\mathbb{E}_n$ is the average over construction pairs. Then

$$\frac{D_\pi^\top D_\pi}{n} = M + M' - C_\pi - C_\pi^\top. \tag{15}$$

Permuting the partners preserves M and M' and changes the cross-moment. It also preserves the mean difference,

$$\bar{d}_\pi = \frac{1}{n}\sum_i (g_i - g'_{\pi(i)}) = \bar{g} - \bar{g}'. \tag{16}$$

The raw SVD therefore retains a common mean contribution alongside assignment-dependent covariance. The exact expectation in Appendix K identifies the within-label covariance and mean structure retained by a shuffle. Let $T = \mathbb{E}_{j\in\mathcal{T}}[v_j v_j^\top]$. Since $\mathrm{PE}(S) = \mathrm{tr}(P_S T)$, distinct subspaces can capture similar target energy; a small contrast alone does not imply equal projectors.

The Gaussian reference has a separate geometric interpretation. Rotational symmetry gives $\mathbb{E}[P_{\mathrm{amb}}] = (k/d)I$, hence

$$\mathbb{E}[\mathrm{PE}(S_{\mathrm{amb}})] = \frac{k}{d}\,\mathrm{tr}(T). \tag{17}$$

When target gradients lie in the r-dimensional support of Q, the matched reference instead has expectation $(k/r)\,\mathrm{tr}(T)$. The ratio d/r is 56.9, 14.2, and 7.1 for the visual models and two for the dialogue heads. Table 1 measures the remaining alignment advantage inside that support; partner shuffles then test the additional role of correspondence.

K Expected Second Moment under Partner Shuffling

The expected raw second moment characterizes the within-label variation and differences between label-wise means retained after partner randomization.

Using covariance with divisor n, the second moment in Equation (15) is $\mathrm{Cov}_n(g - g'_\pi) + \bar{d}\bar{d}^\top$. The covariance identity

$$\mathrm{Cov}_n(g - g'_\pi) = \mathrm{Cov}_n(g) + \mathrm{Cov}_n(g') - \mathrm{Cov}_n(g, g'_\pi) - \mathrm{Cov}_n(g'_\pi, g) \tag{18}$$

describes the assignment-dependent terms, while the raw SVD retains the common mean contribution.

For each nonempty label stratum c, let $w_c = n_c/n$. Let μ_c, μ'_c and Σ_c, Σ'_c be the two gradient means and covariances within that stratum, using divisor n_c. For the uniform within-label permutations used here,

$$\mathbb{E}_{\pi\sim\Pi_{\mathrm{cond}}}\left[\frac{D_\pi^\top D_\pi}{n}\right] = \sum_c w_c\left[\Sigma_c + \Sigma'_c + (\mu_c - \mu'_c)(\mu_c - \mu'_c)^\top\right]. \tag{19}$$

Proof. For each i in stratum c, uniform sampling gives $\mathbb{E}_\pi[g'_{\pi(i)}] = \mu'_c$. The first gradients are fixed, so $\mathbb{E}_\pi[C_\pi] = \sum_c w_c \mu_c \mu_c'^\top$. Substituting this expression and the within-stratum decompositions of M and M' into Equation (15) proves the identity. No independence between different rows is required. □

A global shuffle is the single-stratum case. The conditioned reference retains the within-label marginal covariances and the differences between label-specific gradient means even as it randomizes pair identities. This identifies structure that can support a shuffled subspace without relying on the original partners. Equation (19) concerns an expected second moment; its leading eigenspace does not in general give the mean energy of the separately fitted shuffled subspaces.

L Calibration with Known Correspondence

The signal and nuisance projectors in Section 6 are orthogonal and sum to I_{32}. Each seed draws 512 independent rows of signal s in four dimensions and nuisances h_1, h_2 in 28 dimensions, then applies

Table 17: Complete fixed grid. $+/-/0$ counts observed-minus-within-label accuracy intervals by sign over nine settings per modality. Flip ranges are observed-versus-no-update model–target means in vision (%). All 198 setting–budget estimates and intervals, balanced accuracy, macro-F1, update norms, and logit changes are in the saved metrics.

η	Steps	Batch	Vision $+/-/0$	Dialogue $+/-/0$	Visual flips
0.02	1	32	0/0/9	0/0/9	0.00–0.70
0.02	2	32	0/1/8	0/0/9	0.06–1.98
0.02	4	32	1/1/7	0/0/9	0.08–3.91
0.08	1	32	0/1/8	0/0/9	0.08–3.99
0.08	2	32	1/0/8	0/0/9	0.45–10.08
0.08	4	32	1/1/7	1/0/8	1.11–24.49
0.32	1	32	1/0/8	1/0/8	1.15–23.91
0.32	2	32	0/0/9	1/0/8	1.69–49.67
0.32	4	32	1/0/8	1/0/8	3.21–63.05
0.08	2	16	0/1/8	1/0/8	1.11–26.01
0.08	2	64	1/0/8	0/0/9	0.04–2.92

a fixed orthogonal rotation. A separate set of 4,096 held-out vectors is $v = s_{\text{test}} + 0.1\epsilon$, with isotropic standard Gaussian ϵ. The ideal signal-aligned energy fraction is $(4 + 0.01 \cdot 4)/(4 + 0.01 \cdot 32) = 0.935185$.

We use the leading four eigenvectors of $D^\top D$, equivalent to raw right singular vectors without mean subtraction. The grid is $\tau \in \{0, 0.25, 0.5, 1, 2, 4\}$ and $\rho \in \{0, 0.25, 0.5, 0.75, 0.9, 1\}$. Twenty seeds 4100–4119 and twelve uniform shuffles per seed give 720 constructions and 8,640 control draws.

At $\rho = 0$, observed and independently shuffled population second moments agree. This is not an exact finite-sample invariance. The mean contrasts at $\tau = 0.25$ and 0.5 are 0.000375 $[0.000051, 0.000702]$ and 0.004296 $[0.000398, 0.008117]$, respectively; the other four $\rho = 0$ intervals include zero. We retain these small positive intervals rather than treating the grid as validation of type I error control. At $\rho = 1$, the contrasts for $\tau = 0.25, 0.5, 1, 2, 4$ are $0.00818, 0.11113, 0.91951, 0.92516, 0.92574$, respectively. The response follows the increasing benefit of removing shared nuisance.

For the exact negative control, two labels define two fixed partner vectors $b_i = c_{y_i}$. Every within-label permutation leaves each difference $a_i - b_i$ unchanged. All twenty seeds and twelve draws give bitwise identical difference matrices and zero projector distance. This exact symmetry is distinct from the population-covariance null above. Calibration intervals use 6,000 bootstrap resamples of the twenty seeds after averaging their twelve draws; draws are not counted as independent constructions. The earlier algebraic intervention remains a separate, nonguaranteed diagnostic.

M Fixed Grid of Adaptation Budgets

The grid was fixed before execution: learning rates $\{0.02, 0.08, 0.32\}$ crossed with $\{1, 2, 4\}$ steps at batch size 32, plus sizes 16 and 64 at the original $(0.08, 2)$ setting. Models, streams, subspaces and twelve control draws remain fixed. At the largest budget, visual prediction flips reach 63.0% and accuracy losses reach 6.91 points relative to no update.

Table 17 includes every frozen budget. All 54 visual records indexed by seed and size and 27 dialogue seed records are evaluated with their existing source models, construction gradients, target order, and random seeds. Every record reproduces the original reference-budget accuracy exactly for observed, conditioned-shuffle, and support-random arms. Unconstrained and no-update results are also retained. This is a sensitivity study, not hyperparameter selection using target labels. Batch size changes both the stream grouping and the number of updates, so it is not interpreted as an isolated effect of learning rate.

Each budget has at least fifteen accuracy intervals containing zero for paired versus shuffled comparisons across the eighteen settings. These effects remain unresolved; the intervals do not establish equivalence. Individual contrast magnitudes reach 2.36 percentage points elsewhere in the grid. Large budgets can amplify prediction changes while reducing accuracy. The findings support persistence of sparse pairing-specific gains, not stable utility or a preferred budget.

Hierarchical intervals use 6,000 resamples with seed 20260923. They follow the matched extension's hierarchy of seeds, sizes and draws (vision) or seeds and draws (dialogue). The grid adds no resampling of target populations or shared trees, or multiplicity correction.

N IDENTICAL RANDOM GAINS WITH DIFFERENT PAIRING EFFECTS

Shared nuisance cancels in observed differences but reappears after reassignment. Varying its covariance steers the shuffled subspace toward lower, equal or higher target energy, while preserving the observed subspace and random-reference scores.

Proof of Proposition 1. Decompose a $3k$-dimensional support into three mutually orthogonal rank-k projectors P_1, P_2, P_3, and write $I_{\mathcal{A}} = P_1 + P_2 + P_3$. Embed the support in $\mathbb{R}^d$ with $d \geq 3k$. In all three constructions, let $s \sim \mathcal{N}(0, P_1)$ and set

$$a = s + h, \qquad b = h, \qquad s \perp h. \tag{20}$$

All pairs belong to one label stratum. Use the three nuisance covariances

$$H_+ = I_{\mathcal{A}} + P_3, \qquad H_0 = I_{\mathcal{A}}, \qquad H_- = I_{\mathcal{A}} + P_2, \qquad h \sim \mathcal{N}(0, H). \tag{21}$$

Each H is positive definite on the support, so both member marginals are nondegenerate there. The observed difference remains $a - b = s$, with second moment P_1 and a unit gap after the leading k eigenvalues. Its rank-k projector is therefore uniquely P_1 in every construction.

Choose the same held-out target second moment

$$T = 2P_1 + 3P_2 + \tfrac{1}{10}P_3. \tag{22}$$

Observed energy is $2k$ and total energy is $51k/10$. A uniformly random rank-k projector in the support has expected energy $17k/10$; an ambient random projector has expected energy $51k^2/(10d)$. Thus every construction has the same two positive gains,

$$\Delta_{\mathrm{sup}} = \tfrac{3}{10}k, \qquad \Delta_{\mathrm{amb}} = 2k - \frac{51k^2}{10d} > 0. \tag{23}$$

Population within-label reassignment replaces b by an independent copy $\tilde{h}$ of the same partner marginal, giving $a - \tilde{h} = s + h - \tilde{h}$ and second moment $P_1 + 2H$. For H_+ the leading rank-k projector is P_3: its eigenvalue is four, compared with three on P_1 and two on P_2. For H_0 the leading projector is P_1; for H_- it is P_2. Each has a strictly positive truncation gap. Consequently,

$$\Delta^{(+)}_{\mathrm{cond}} = 2k - \tfrac{1}{10}k = \tfrac{19}{10}k, \qquad \Delta^{(0)}_{\mathrm{cond}} = 0, \qquad \Delta^{(-)}_{\mathrm{cond}} = 2k - 3k = -k. \tag{24}$$

The observed second moment, target geometry, support and random reference scores are identical while the pairing contrast takes all three signs. This proves the proposition. □

The member marginal covariances differ between constructions. The result therefore concerns identifiability from the scores, not the full joint distribution.

Population reassignment uses independent partners. For independent samples, observed and uniformly shuffled empirical second moments converge to their population counterparts. Strict truncation gaps ensure projector convergence. The finite-sample audit averages energies of separately fitted shuffled projectors, rather than fitting to an averaged second moment. This paired-vector example is neither a classification benchmark nor a universal calibration guarantee.

Derivation of the attribution identity. For an orthonormal support basis Q, a uniform rank-k support projector has expectation $(k/r)QQ^\top$. When $T = QQ^\top TQQ^\top$, expected energy is $(k/r)\,\mathrm{tr}(T)$ within the support and $(k/d)\,\mathrm{tr}(T)$ in the ambient space. Adding and subtracting the support-random and within-label shuffle expectations from $\mathrm{PE}(S_{\mathrm{obs}}) - (k/d)\,\mathrm{tr}(T)$ gives Equation (2). Only the support term is necessarily nonnegative. The other terms are signed, shuffle-dependent contrasts, not causal percentages. The identity uses exact expectations; random-reference estimates use twelve draws.

REPRODUCIBILITY AND DRAFTING RECORD

The accompanying provenance record identifies the experiment artifacts used here. Generative AI tools assisted with drafting, editing, and formatting.