A CONTRASTIVE TARGET IS A DEMAND, NOT A LABEL

Anonymous authors
Paper under double-blind review

ABSTRACT

An InfoNCE target is usually read as a label: it says which samples belong together. We show that it acts as a demand, and that what an encoder learns depends on how the demand can be met, not only on what it says. We write the target of a batch as a matrix with two parameters, sharpness and specificity, and build targets that agree exactly on class correctness, sharpness and the class composition of every batch, with the temperature tuned for every arm. The cleanest test compares two targets that name a random class-mate as each row's positive and carry identical information. When the class-mate is fixed, encoders memorise the pairing. When it is redrawn at every step, they cannot, and held-out classes become tighter by 30 to 50% in angular spread and more accurate by 1.5 to 4.6 points of within-view class retrieval. Against the standard target, which names the true partner, the redrawn class-mate gains 3.4 and 3.3 points of class structure on two image settings and gives up 44 and 82 points of instance retrieval, a trade that reappears at larger scale on CIFAR-100 with a ResNet-18. The second parameter, sharpness, behaves as a second temperature. Over a 7×8 grid with every optimum interior, the best temperature falls $40\times$ as sharpness grows $73\times$, the best smoothed target is within 2 points of the best-tempered one-hot target in three of four settings, and a closed form of the loss predicts the direction of this ridge. With the temperature free, sharpness is harmless on a correct target, and most of the damage a sharp wrong target does in a fixed-temperature loss disappears. Specificity, the axis the temperature cannot reach, is where the target's content lives.

1 INTRODUCTION

If we write the target of an InfoNCE batch as a matrix T instead of as an index, two decisions become visible that the usual notation hides. The first is how much mass each row places on its positive: in standard InfoNCE the matrix T is one-hot, but this is not required. The second is which column that mass sits on: in standard InfoNCE the matrix T points at the row's own augmented view, and this is not required either.

Both decisions are defaults inherited from instance discrimination (Wu et al., 2018; van den Oord et al., 2018), and neither has been studied as a variable of its own. We study both and find that they are different kinds of object. Sharpness is largely a second temperature. Specificity is a genuine design axis, and standard InfoNCE fixes it at one end without ever exposing the rest.

We construct targets that point every row at a sample of its own class, so the class correctness is exactly 1, and that keep T a bijection, so the class composition of every batch and the number of same-class negatives per row are identical across arms. The only thing that moves is which class-mate is named: the row's own other view ($\varphi = 1$), a class-mate fixed before training, or a class-mate redrawn at every step. The temperature is tuned for every arm and readout.

The fixed and redrawn class-mate targets carry exactly the same information, so any difference between them cannot come from what the target says. The difference is large. Encoders memorise a fixed pairing, and what they carry to held-out data is then barely more class structure than standard InfoNCE gives. Redrawn, the same target cannot be memorised, and it produces the tightest classes of all three, beating standard InfoNCE on class structure on both datasets (Table 1). The one-line summary is the title: a contrastive target is a demand, and the geometry an encoder learns is set

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.

by how the demand can be met. Instance retrieval falls by 45 to 84 points once the positive is a class-mate, as it must, since only $\varphi = 1$ names the true partner.

This view has a direct consequence for methods that choose positives from outside the pair: supervised contrastive learning, nearest-neighbour positives (Dwibedi et al., 2021), and clustering or pseudo-label targets (Caron et al., 2020). Whether such a positive is held fixed or resampled decides whether its class content is learned or memorised. We measure this under granted labels; the same question applies wherever positives are assigned rather than given.

The same view predicts that sharpness only sets how hard a demand is imposed, which InfoNCE already controls through its temperature. On a 7×8 grid the optimum is a single curve, with the best temperature falling $40\times$ as sharpness grows $73\times$ (Figure 4), and the best smoothed target exceeds the best-tempered one-hot target by $+0.83/ + 7.50/ + 0.81/ + 1.60$ pp in the four settings. A smoothing gain of $+33$ pp that we had measured in a fixed-temperature loss is, on this evidence, a temperature effect.

Contributions.

- **A controlled specificity axis** (§3). Targets matched exactly on class correctness, sharpness and batch composition, with the temperature tuned per arm, differ by 45 to 84 pp in instance retrieval and up to 4.6 pp in class structure, depending only on which same-class sample is the positive.
- **A mechanism, tested by intervention** (§4). A fixed class-mate pairing is memorised; redrawing it changes no class-level quantity and cuts held-out class spread by 28 to 50% on both datasets, across two independent implementations. The instance-to-class trade lies between the true pairing and an unmemorisable class pairing, the regime of CIFAR-100 (Appendix H).
- **Sharpness as a second temperature** (§5). The optimum over sharpness and temperature is a one-dimensional ridge with every optimum interior, and a closed form of the loss gives its direction.
- **The cost of a wrong target** (§6, §7). Sharpness is free on a correct target and costs in proportion to the target's class error; with the temperature free, that cost falls from 52 to 11 pp on digits, and a readout audit separates metric distortion from lost information.

2 PRELIMINARIES

The loss. For a batch of b paired samples with L2-normalised embeddings z^x, z^y, let L^{row} and L^{col} be the row-wise and column-wise log-softmax of $z^x (z^y)^\top / \tau$. For any target T with $\sum_{ij} T_{ij} = 1$,

$$\mathcal{L}(T, \tau) \;=\; -\tfrac{1}{2}\Big[\textstyle\sum_{ij} T_{ij} L^{\text{row}}_{ij} \;+\; \sum_{ij} T_{ij} L^{\text{col}}_{ij}\Big]. \tag{1}$$

With $T = \frac{1}{b} I$ this is standard symmetric InfoNCE. We parameterise T by two scalars:

$$T_{ij} \;=\; \frac{1-\rho}{b^2} \;+\; \frac{\rho}{b}\mathbb{1}[\, j = \pi(i)\,], \tag{2}$$

where $\rho \in (0, 1]$ is the **sharpness**, that is, the excess mass that each row places on its positive above the uniform floor, so that $\rho = 1$ is one-hot. The permutation π carries the **specificity**. We write φ for the fraction of rows with $\pi(i) = i$. The value $\varphi = 1$ is standard InfoNCE, and at $\varphi = 0$ every row points at a different sample of its own class. The remaining rows are deranged *within* each class, so π is a bijection and the class correctness is exactly 1 at every φ. We also write c for the class correctness of the target where it is varied deliberately, and $E = \rho(c - s)$ for the excess correct mass, with $s = 1/b$ the chance that a row names a given column ($b = 1024$ in the full-batch settings, so $E = 0.999$ at $\rho = c = 1$).

Testbeds. We use two two-view settings in which pair identity and class labels both exist and are used openly as instruments: **digits** (8×8 images split into even and odd pixel columns, 32 dimensions per view) and **Fashion-MNIST** (28×28 images, same split, 392 dimensions per view). Both use 1024 training and 384 validation samples, 16-dimensional latents, MLP encoders, and a fixed step budget, and all comparisons are paired t-tests (Appendix F). The y view is permuted so that no index correspondence survives.

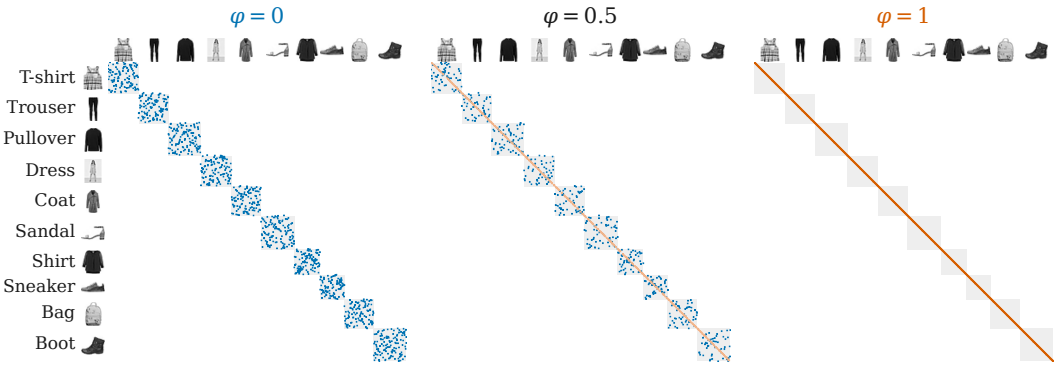

Figure 1: **Targets used in training on Fashion-MNIST.** Each panel is a full 1024×1024 target, one row per view-x sample and one column per view-y sample, with a randomly chosen training image of each class on the axes. Rows are sorted by class, and each column sits under the row of its true partner, so the true pairing is the diagonal. Each grey square is one class, and each coloured dot is the view-y sample a row is told to match. Every target keeps every dot inside the row's own class. At $\varphi = 1$ the dot is the row's own partner; at $\varphi = 0$ it is another image of the same class; at $\varphi = 0.5$ half of the rows are of each kind, with the dots on a row's own partner in light orange and the others in blue.

Table 1: **Three targets with the same class content.** Every target is a class-correct bijection, and every batch has the same class composition under all three. Readouts are the best cell of each arm over sharpness and temperature, mean ± sd over ten runs; the cells are listed in Table G.1. Memorisation is the share of training rows whose nearest view-y embedding is the assigned class-mate, at the arm's best within-view cell.

	digits			Fashion-MNIST		
positive is	own view $\varphi = 1$	class-mate fixed	class-mate redrawn	own view $\varphi = 1$	class-mate fixed	class-mate redrawn
matched by construction						
target class correctness		1.000			1.000	
batch class composition		identical			identical	
readouts (%)						
within-view 5NN	91.3 ± 0.6	93.2 ± 0.6	$\mathbf{94.7} \pm 0.3$	78.3 ± 1.9	77.0 ± 1.4	$\mathbf{81.6} \pm 0.7$
cross-view class	91.7 ± 1.2	91.7 ± 0.7	$\mathbf{93.8} \pm 0.7$	$\mathbf{98.1} \pm 0.5$	70.4 ± 1.6	78.1 ± 1.8
instance top-1	$\mathbf{48.4} \pm 2.3$	3.7 ± 0.5	3.9 ± 1.1	$\mathbf{88.6} \pm 1.0$	4.3 ± 0.8	6.2 ± 0.7
memorised pairing		35.9	0.9		91.2	1.0

Readouts. Within-view class 5NN, cross-view class agreement (1-NN), instance top-1 (the true partner found across views) and effective rank; §7 adds a linear probe, a whitened (ZCA) 5NN and cross-view linear transfer. Apart from constructing the targets, class labels appear only in readouts. Appendix F defines each one.

Supervision. The quantities c, ρ and φ are set using labels and known pairs, and every arm that reads the true pairing is a *granted-true-pairs diagnostic control*. Nothing here is a method that could run without that supervision; the contribution is an explanation, not an algorithm.

Protocol. Predictions and failure conditions were written down before each stage was run. The eight that failed are marked where they occur or listed in Appendix B, and the one rule fixed mid-stage is described in Appendix B.

3 Instance-Level and Class-Level Positives

We train with Eq. 1 and give every arm its own best (ρ, τ) for every readout, over $\rho \in \{0.055, 0.254, 1\}$ and eight temperatures (Table 1). Class correctness is 1 by construction, and

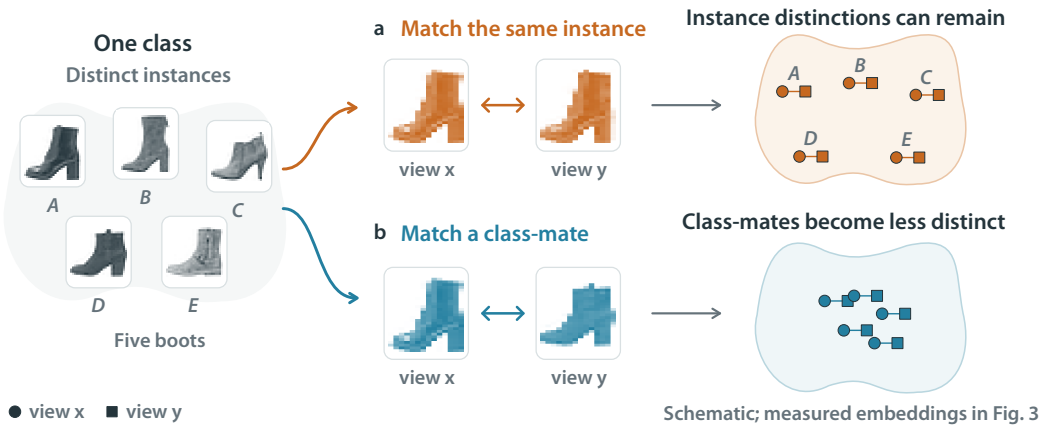

Figure 2: **The two targets make different demands on one class.** Five Fashion-MNIST ankle boots under two targets; a tile shows one view (even or odd pixel columns). (a) Each view is matched to the other view of the same instance ($\varphi = 1$). The pairs can be co-located while the five instances stay apart. (b) Each view is matched to the other view of a different class-mate ($\varphi = 0$), here boot A to boot D. The two share only the class, so what carries over to new samples is class structure.

because each target maps a class onto itself, every batch has the same class composition, and so the same number of same-class negatives per row, under all three.

Moving the positive from the row's own view to a class-mate costs 44.7 pp of instance retrieval on digits and 84.2 pp on Fashion-MNIST ($p = 9\times10^{-13}$ and 3×10^{-19}). Only $\varphi = 1$ names the true partner, and at $\varphi = 0$ the true partner sits among the negatives. What matters is what the class-mate target returns in class structure, and here the two class-mate targets part ways. With the class-mate fixed before training, the return is small: within-view class structure is 2.0 pp better on digits ($p = 6\times10^{-6}$) and 1.3 pp worse on Fashion-MNIST ($p = 0.12$), where cross-view class agreement is also 27.8 pp worse (a predicted class gain that **failed** on Fashion-MNIST; in the fixed-temperature loss the fixed class-mate had won on both, Table A.1).

With the class-mate redrawn at every step, the same target gains on class structure on both datasets: +3.4 pp within view over $\varphi = 1$ on digits ($p = 3\times10^{-8}$) and +3.3 pp on Fashion-MNIST ($p = 1.4\times10^{-3}$), while losing instance retrieval by 44.5 and 82.3 pp. This comparison is the core of the paper. The fixed and redrawn targets carry exactly the same information: the same classes, the same batches, and no instance information in either. They differ only in whether the pairing can be memorised, and that alone moves within-view class structure by 1.5 and 4.6 pp ($p = 3\times10^{-5}$ and 1×10^{-6}). No account that reads the target as a label can produce this difference.

The usual suspects do not separate these arms. Same-class negatives, which the debiasing literature treats as the defect to correct (Chuang et al., 2020; Robinson et al., 2020; Huynh et al., 2020), are identical by construction. Effective rank does vary between cells here, but in the fixed-temperature loss the class-mate and own-view targets reached ranks within 0.06 of each other on digits and still differed by 17.6 pp in instance retrieval, so dimensional collapse (Jing et al., 2021) is not the explanation. What remains is *which sample each row is told to be close to*, and whether that choice can be learned by rote.

4 Contrastive Targets as Constraints

Our reading is shown in Figure 2. Consider the rows of one class. At $\varphi = 1$ each row is asked to be close to the other half of the same image, and the two halves share almost all of their content. The encoders can therefore meet the demand with a rule that also holds for images they have not seen, and that rule keeps instances apart.

At $\varphi = 0$ each row is asked for "z_i^x close to $z_{\sigma(i)}^y$", where σ is a derangement inside the class. With two independent encoders this does not force the class to a point: nothing ties z_j^x to z_j^y, so giving

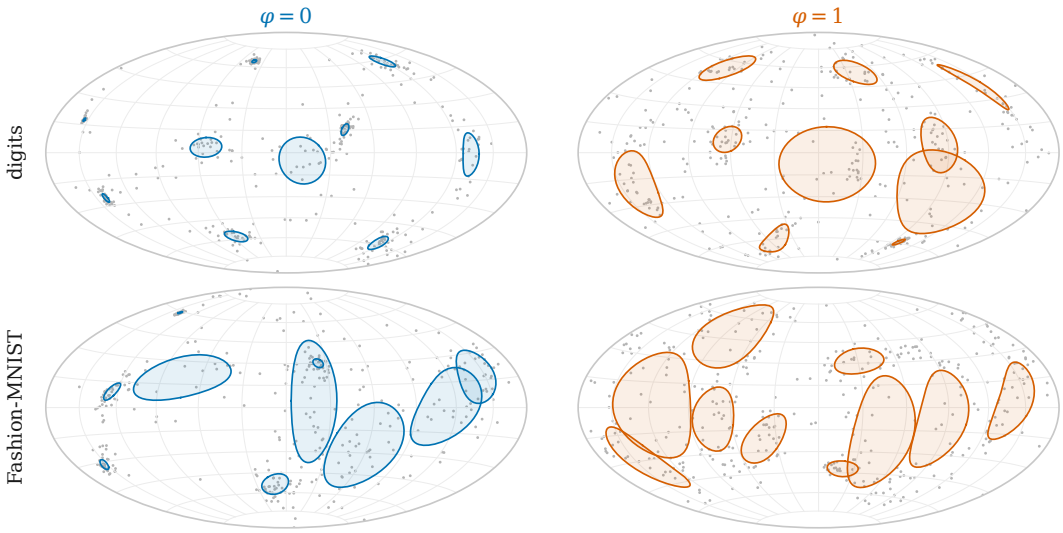

Figure 3: **The geometry the two demands produce**, in the fixed-temperature loss of Appendix A. The embeddings lie on a sphere (their distance from the centre varies by 1 to 4%), so we draw the top three principal directions as an equal-area map, on which every validation point of view x appears once. Each ellipse is one class at one standard deviation, fitted on the sphere to the 75% of the class nearest its mean direction. At $\varphi = 0$ the classes are small, tight regions; at $\varphi = 1$ they are wide. Within-class angular spread over the mean angle between class means, in the full embedding space and averaged over runs, rises from 0.15 to 0.34 on digits and from 0.27 to 0.42 on Fashion-MNIST.

every pair $(i, \sigma(i))$ its own location satisfies all the constraints with distinct instances. What x_i and $y_{\sigma(i)}$ share, however, is only their class, because σ is random. If σ is fixed, the instance part of the demand can be met by memorising it, which does not carry over to new images. If σ is redrawn at every step, it cannot, and the only way to meet the demand is class-level alignment. Held-out classes should therefore be tighter when σ is redrawn.

Both halves are measured, in a stage whose predictions were written before it ran. On the training set the nearest view-y embedding of x_i is its assigned class-mate for 35.9% of rows on digits and 91.2% on Fashion-MNIST at the fixed arm's best cell, against 0.9% and 1.0% when σ is redrawn, so the encoders do memorise a fixed σ. At that same cell, held-out within-class spread (Figure 3) falls from 0.159 to 0.112 on digits and from 0.499 to 0.247 on Fashion-MNIST when σ is redrawn (one-sided paired $p = 2\times10^{-8}$ and 3×10^{-14}). The same intervention, predeclared separately in the fixed-temperature loss, gave the same result (Appendix I), so the effect holds across two independent implementations. (Our declared memorisation check held at the compared cells and **failed** at $\tau \geq 4$, where a soft row cannot single out one class-mate.)

The picture is therefore precise. The trade between instance and class structure lies between the true pairing and a class pairing that cannot be memorised; a fixed class pairing is largely absorbed by memorisation. Standard InfoNCE sits at the $\varphi = 1$ end, and nothing in the loss exposes the other end. The intermediate values of φ are in Table G.1.

Both settings above are small and both build their views by splitting one image. We repeated the comparison on CIFAR-100 with a ResNet-18, 45,000 images and two independent augmentations as the two views (Appendix H). There, batches are redrawn at every step, so the class-mate changes from step to step: the $\varphi = 0$ target is the redrawn one. The result is the redrawn pattern at larger scale: instance retrieval falls from 81.18 to 13.56 while every class readout improves. Its 100 fine classes are nested in 20 coarse ones, and a class-mate drawn from the coarse class costs more instance retrieval than one drawn from the fine class (72.22 against 67.93 pp over five runs, one-sided paired $p = 3\times10^{-6}$). A coarse class-mate is five times less specific but also of the wrong fine class, and this comparison cannot separate the two; the coarse target also degrades fine-class retrieval below training without labels, which is class collapse (Graf et al., 2021; Chen et al., 2022; Xue et al., 2023).

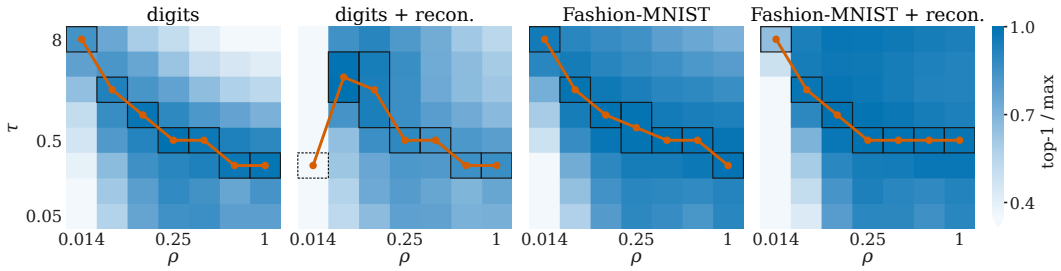

Figure 4: **The two-dimensional optimum is one-dimensional.** Instance top-1 retrieval of symmetric soft-target InfoNCE over target sharpness ρ and temperature τ, shaded relative to each panel's maximum (48.1, 55.1, 89.6 and 90.0%). Boxes mark the τ within one standard error of each column's best, and the line joins their centres, $\tau^\star$: as ρ grows $73\times$, $\tau^\star$ falls by up to $40\times$, and $\rho\,\tau^\star$ stays within $4.6\times$ right of $\rho = 0.014$. Every column's optimum is interior once the two leftmost columns are extended to $\tau = 64$ (Appendix C). The dashed box is the one excursion, the leftmost column of digits with reconstruction, where no cell exceeds 17.7%; there the reconstruction term swamps a near-uniform target (Appendix J).

The temperature rescales the logits and changes how peaked the model's Gibbs rows are, but it cannot change which column a target row points at, so no temperature setting substitutes for specificity. The same argument implies that the other axis, sharpness, should be reachable by the temperature and therefore redundant. The next section tests this.

5 Target Sharpness and Temperature

If the mechanism of §4 is correct, then target sharpness concerns how strongly a demand is imposed and not what the demand asks for, and InfoNCE already contains a control of strength. We therefore predict that ρ and τ are largely interchangeable. (The grid below was designed to test an earlier smoothing result, so this prediction is post hoc to it.)

In a loss whose temperature is fixed, ρ is the only control over peakedness, so any measured benefit of smoothing is confounded with a mis-setting of the temperature. In our own earlier experiments, in a loss with τ hard-coded to 1 and the logit scale set by a median pairwise distance, moving from $\rho = 1$ to $\rho \approx 0.05$ gained $+32.9$ pp on digits ($p = 3.1\times10^{-8}$) and $+24.5$ pp on Fashion-MNIST. It survived an alignment-weight control and a $4\times$ budget control, which makes it a strong test case for the prediction.

Implementation. We reimplemented Eq. 1 from scratch instead of reusing our own trainer. At $\rho = 1$ it agrees with $\frac{1}{2}[\mathrm{CE}(\mathrm{logits}, \mathrm{arange}(b)) + \mathrm{CE}(\mathrm{logits}^\top, \mathrm{arange}(b))]$ to a deviation of **0.0**. At $\rho < 1$ it agrees with a brute-force construction of T and an explicit double sum to 8.9×10^{-16}.

Design and stopping rule. We use $\rho \in \{0.0137, 0.0547, 0.127, 0.254, 0.5, 0.75, 1\}$ against $\tau \in \{0.05, 0.1, 0.2, 0.5, 1, 2, 4, 8\}$, on both datasets, with and without a reconstruction anchor. The test $J1'$ asks whether the best $(\rho < 1, \tau)$ cell exceeds the best $(\rho = 1, \tau)$ cell by more than 5 pp in instance top-1 retrieval. The grid was extended in τ until both compared cells were interior; that stopping rule was fixed after the first extension (Appendix B). A later predeclared stage extended the two leftmost columns to $\tau \in \{16, 32, 64\}$; every column optimum is now interior, including the three at $\tau = 8$, and no panel maximum moved.

Result. The prediction holds (Figure 4, Table 2). The optimum is a curve and not a region: as ρ ranges over $73\times$ the best temperature moves by $40\times$ in the opposite direction, and $\rho\,\tau^\star$, with $\tau^\star$ the geometric centre of the temperatures within one standard error of the best, stays within a factor of 4.6 across all four settings, and within 2.3 on digits without the anchor. One column is an exception: at $\rho = 0.0137$ with the anchor on digits no temperature reaches 17.7% top-1, so its $\tau^\star$ is not a ridge point, and excluding that degenerate column is the only exclusion we make. The column is not unlearnable: weighting reconstruction ten times less lifts its $\tau = 8$ cell to 53.7% (Appendix J).

Table 2: Best smoothed target against best-tempered one-hot target. Differences are paired. One of the four settings clears the predeclared bar of 5 pp.

Dataset	Anchor	best (ρ, τ)	one-hot best τ	Δ top-1	Δ transfer
digits	none	$(0.254, 0.5)$	0.2	$+0.83$ (p=0.39)	$+0.95$
digits	recon 0.5	$(0.055, 2.0)$	0.2	$+\mathbf{7.50}$ (p=4.6×10^{-7})	$+1.97$
fashion	none	$(0.500, 0.5)$	0.2	$+0.81$ (p=0.073)	$+1.13$
fashion	recon 0.5	$(0.127, 1.0)$	0.5	$+1.60$ (p=0.013)	$+2.28$

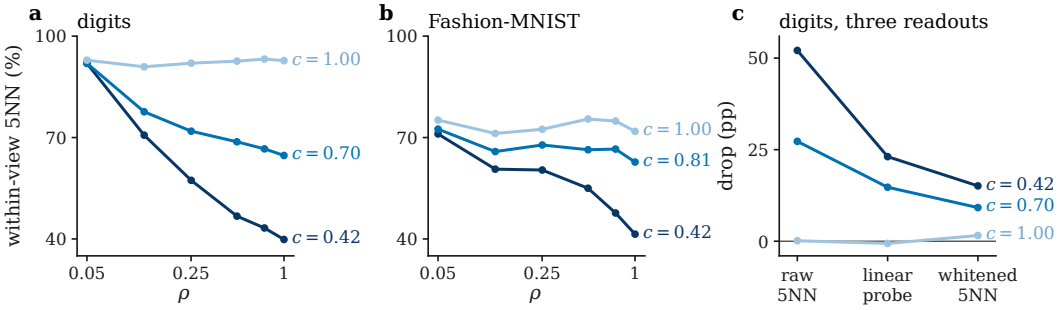

Figure 5: **Sharpness costs in proportion to the class error of the target**, in the fixed-temperature loss; Table G.2 repeats (a, b) with a free temperature. (a, b) Within-view class 5NN against ρ at three levels of target class correctness c; darker is more wrong. The cost grows with the error and vanishes at $c = 1$. (c) The same drop on digits under three readouts of one frozen embedding: raw 5NN overstates it, and whitening recovers 66% to 84% of it.

The ridge is also not a constant product everywhere: a log-log fit of $\tau^\star$ on ρ over $0.055 \le \rho \le 0.75$ has slope -0.79 and -1.00 on digits and -0.54 and -0.55 on Fashion-MNIST. The test $J1'$ holds in **one of four** settings; the other three show differences below 2 pp, one of them not significant.

A closed form for the direction. Row i of Eq. 1 is a cross entropy between the target row and a softmax over s_{ij}/τ, with s a cosine. It is minimised when the softmax equals the target, that is when the cosine gap between the positive and every negative is

$$\Delta \;=\; \tau \,\log\!\Big(1 + \frac{\rho\, b}{1-\rho}\Big).$$

If an encoder can realise a gap of roughly fixed size, the best temperature scales as $1/\log(1 + \rho b/(1-\rho))$, which falls as ρ grows; this is the uniform-floor analogue of the bound on logits under label smoothing (Müller et al., 2019). It gives the ridge its direction, but over our range it predicts a slope of -0.33, and the measured slopes are -0.54 to -1.00. At $\rho = 1$ it has no finite solution at all, yet the one-hot optimum sits at $\tau = 0.2$ or 0.5. So the algebra of the target explains why a ridge exists, and something outside it, which the held-out readout sees and the training loss does not, sets how steep it is.

The $+33$ pp therefore does not survive a free temperature, and its explanation is general: *when the temperature is held fixed, the shape of the target does the temperature's work*. A smoothing benefit reported without a temperature sweep is evidence about the temperature, and any study of soft contrastive targets needs the joint sweep. The reconstruction anchor widens the margin on both datasets ($0.83 \to 7.50$ and $0.81 \to 1.60$); the direction is consistent and the magnitude is not, so it deserves its own predeclared stage.

6 Sharpness under Incorrect Targets

The mechanism says that sharpness controls how forcefully a demand is imposed. It follows that sharpness should cost nothing when the demand is correct, and should cost in proportion when the demand is wrong. We vary ρ against targets of known class correctness c, first in the fixed-temperature loss (Figure 5) and then with the temperature free.

Fixed temperature. Raising ρ from 0.055 to 1 costs within-view class 5NN $52.1/27.3/0.1$ pp at $c = 0.42/0.70/1.00$ on digits, and $29.7/9.8/3.4$ pp on Fashion-MNIST, and the gap between correctness levels grows with ρ (all $p \leq 5 \times 10^{-3}$). A predeclared out-of-sample prediction about where the optimum moves also held (Appendix D). Our prediction that one-hot is harmful even with a correct target **failed**: at $c = 1$ the curve is flat over an $18\times$ range in ρ.

Free temperature. By §5, these costs could be a mis-set temperature. We reran the comparison in Eq. 1 with the temperature tuned for every (c, ρ), using targets in which a fraction c of rows point to the true partner and the rest to a sample of another class (Table G.2). The effect keeps its direction and shrinks. The cost at $c = 0.42$ is 10.7 pp on digits and 3.8 pp on Fashion-MNIST, against 0.7 and -0.2 pp at $c = 1$ (one-sided paired $p = 2\times10^{-7}$ and 3×10^{-5}), and the ordering over c is monotone on both. We had required the difference to exceed 5 pp on both datasets, so this prediction **failed** on Fashion-MNIST. Most of the fixed-temperature cost was the temperature; what remains is the clean form of the claim, that sharpness is free on a correct target and costs on a wrong one.

7 Dependence on the Evaluation Metric

Nearest-neighbour accuracy is the default readout here and it is sensitive to the geometry of an embedding, not only to the information in it. On one frozen embedding set of the fixed-temperature loss we measured 51.78% for the full 5NN, 35.98% from the top two principal components alone, and 81.99% from the residual after removing them, a residual carrying 0.16% of the variance and rising to 86.59% after whitening. So "class structure was destroyed" can mean two things. We tested this without retraining, by reloading the stored encoders and applying four readouts to identical frozen embeddings, with a maximum deviation from the logged values of 0.0 over 400 runs.

Across 18 grid arms, the raw 5NN and the linear probe rank the arms almost identically (Spearman $+0.967$, $p = 6.4\times10^{-11}$), so the ordering in §6 is not an artefact of kNN.

The size of the drop, however, depends on the readout. Whitening recovers 66% to 84% of the raw drop (Figure 5c), all in the fixed-temperature loss: a raw drop of -52.1 pp becomes -23.1 pp under a linear probe and -15.1 pp after whitening. The drop therefore has two parts, a large distortion of the metric and a smaller but real loss of linearly decodable class information, and both scale with the class error. We report all three readouts for every sharpness effect here and recommend the same practice elsewhere.

Under a cross-view linear readout, which uses no nearest neighbour, the predicted sign flip **failed**: $\rho : 0.055 \to 0.75$ gains at both $c = 0.42$ and $c = 1$ ($+11.09$ and $+16.73$ pp). The claim of §6 is about nearest-neighbour retrieval.

8 Related Work

Choice of positive. Instance discrimination (Wu et al., 2018; van den Oord et al., 2018; He et al., 2019; Chen et al., 2020) defines the positive as the other view of the same instance; supervised contrastive learning (Khosla et al., 2020) spreads it over class membership; clustering methods (Caron et al., 2018; Asano et al., 2019; Caron et al., 2020; Li et al., 2020) replace it with an inferred assignment; nearest-neighbour positives (Dwibedi et al., 2021) and mixed soft targets (Lee et al., 2021) move it within the data; work on debiasing and false negatives (Chuang et al., 2020; Robinson et al., 2020; Huynh et al., 2020) corrects the negatives instead. Weighted sums of the self-supervised and supervised losses (Islam et al., 2021) traverse the same instance-to-class range and report the same trade in transfer, and recent theory argues that the two objectives are closer than they appear (Luthra et al., 2025). These methods move along the specificity axis, but each also changes the number of positives per row, the mass on the true partner or the set of negatives. Our construction swaps which single sample is the positive and holds those fixed, which is what lets §4 attribute the effect to memorisation of the pairing. That supervised positives collapse finer structure than the label names is known as class collapse (Graf et al., 2021; Chen et al., 2022; Xue et al., 2023); the coarse-label result of Appendix H is an instance of it. Appendix H measures the supervised target beside ours: it keeps the true partner but gives it one eighth of the row, and instance retrieval collapses as far as when the partner is removed. Our prediction that keeping the partner prevents this **failed**.

Soft targets. Label smoothing (Szegedy et al., 2015; Müller et al., 2019) and distillation (Hinton et al., 2015) establish that softening a classification target can help, and optimal-transport distillation carries this into image and text training (Wu et al., 2021). Section 5 cautions against transferring it. In classification target sharpness has no competitor, whereas in InfoNCE the temperature is already a sharpness control (Wang & Liu, 2020), studied as uncertainty (Zhang et al., 2021) and as a schedule (Kukleva et al., 2023), and the two are largely the same one. We are not aware of a previous study that sweeps both jointly, which is what separates them.

Evaluating representations. The kNN readout and the linear probe are standard (Wu et al., 2018; Caron et al., 2021), and transfer studies find their rankings are not interchangeable across downstream tasks (Ericsson et al., 2020). Dimensional collapse (Jing et al., 2021) supplies the mechanism for the distortion we measure, but not for the rank-matched comparison of §3. Our contribution is to show that one frozen embedding set yields a damage estimate differing by a factor of two depending on the readout, and that the *location* of an optimum can move by five grid steps.

9 LIMITATIONS

Scale and architecture. The small settings use 1024 samples, 16-dimensional latents, MLP encoders and a fixed step budget, with no convergence claim. Appendix H carries §3 and §4 to a ResNet-18 on CIFAR-100, but §5 is not tested at that scale. We predeclared a sweep over batch size and latent dimension and did not run it.

Two implementations. Sections 3, 4 and 5 and the free-temperature half of §6 use Eq. 1 with a tuned temperature. The fixed-temperature half of §6, the readout audit of §7, Figure 3 and Appendices A, D and I use the loss this project started from, and where we reran them the magnitudes shrank. Seeds vary initialisation and batch order over one fixed data split.

View construction. The small settings split one image into even and odd pixel columns, which is what makes the matched controls of §3 possible. The CIFAR-100 stage uses two independent augmentations and the effect is larger there, so it is not a product of the column split. Two constructions are still two.

What is measured and what is derived. A tempting argument is that the $\varphi = 0$ constraints compose around a cycle and admit only a collapsed class; with independent encoders it is false, since distinct instances can satisfy every pair. The account that replaces it rests on measured memorisation and a predeclared intervention, run in both implementations, which confirmed its main prediction both times (§4, Appendix I). We have not derived the held-out geometry from the loss. The closest result is a sufficient condition for total collapse under a uniform-floor target, $\beta \leq 1 - \rho$ in a β-weighted variant of Eq. 1, which concerns a different regime.

Scope of the negative result. Section 5 shows redundancy *in this loss family, at this scale, and over the grid searched*. It does not show that no target shape helps, only that the one-parameter uniform-floor family does not help once the temperature is free.

Supervision. Every number here is a diagnostic control; we make no unsupervised claim.

10 CONCLUSION

An InfoNCE target looks like a label but behaves like a demand. Three targets with the same class content, the true partner, a fixed class-mate and a redrawn class-mate, produce three different geometries, and the last two differ only in whether the pairing can be memorised. Sharpness, by contrast, is mostly a second temperature, and a closed form of the loss gives the direction of that redundancy. The content of a contrastive target lives on the axis the temperature cannot reach. Standard InfoNCE sits at one end of that axis by construction; methods that assign positives, from labels, neighbours or clusters, move along it, and our results suggest that resampling such positives, rather than fixing them, is what lets their class content be learned rather than memorised.

THE USE OF LARGE LANGUAGE MODELS

The authors used a large language model for three things. It helped find related work. It polished the prose. It wrote and revised the training, evaluation and plotting code, including the from-scratch reimplementation of Eq. 1 used in §5 and the CIFAR-100 stage of Appendix H.

Responsibility. Every prediction was written down before the corresponding runs, with the exception stated in Appendix B, and every reported number was regenerated from stored artefacts. The authors verified the implementation checks of §5 and are responsible for every claim in this paper.

REFERENCES

Yuki M. Asano, C. Rupprecht, and A. Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019. doi: 10.48550/arXiv.1911.05371.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep Clustering for Unsupervised Learning of Visual Features. *arXiv preprint arXiv:1807.05520*, 2018. doi: 10.48550/arXiv.1807.05520.

Mathilde Caron, Ishan Misra, J. Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *arXiv preprint arXiv:2006.09882*, 2020. doi: 10.48550/arXiv.2006.09882.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, J. Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. *arXiv preprint arXiv:2104.14294*, 2021. doi: 10.48550/arXiv.2104.14294.

Mayee F. Chen, Daniel Y. Fu, Avanika Narayan, Michael Zhang, Zhao Song, Kayvon Fatahalian, and Christopher Ré. Perfectly balanced: Improving transfer and robustness of supervised contrastive learning. In *International Conference on Machine Learning (ICML)*, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv preprint arXiv:2002.05709*, 2020. doi: 10.48550/arXiv.2002.05709.

Ching-Yao Chuang, Joshua Robinson, Lin Yen-Chen, Antonio Torralba, and Stefanie Jegelka. Debiased Contrastive Learning. *arXiv preprint arXiv:2007.00224*, 2020. doi: 10.48550/arXiv.2007.00224.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How Well Do Self-Supervised Models Transfer? *arXiv preprint arXiv:2011.13377*, 2020. doi: 10.48550/arXiv.2011.13377.

Florian Graf, Christoph D. Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 3821–3830, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. *arXiv preprint arXiv:1911.05722*, 2019. doi: 10.48550/arXiv.1911.05722.

Geoffrey E. Hinton, O. Vinyals, and J. Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015. doi: 10.48550/arXiv.1503.02531.

Tri Huynh, Simon Kornblith, Matthew R. Walter, Michael Maire, and Maryam Khademi. Boosting Contrastive Self-Supervised Learning with False Negative Cancellation. *arXiv preprint arXiv:2011.11765*, 2020. doi: 10.48550/arXiv.2011.11765.

Ashraful Islam, Chun-Fu Chen, Rameswar Panda, Leonid Karlinsky, Richard Radke, and Rogerio Feris. A broad study on the transferability of visual representations with contrastive learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding Dimensional Collapse in Contrastive Self-supervised Learning. *arXiv preprint arXiv:2110.09348*, 2021. doi: 10.48550/arXiv.2110.09348.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised Contrastive Learning. *arXiv preprint arXiv:2004.11362*, 2020. doi: 10.48550/arXiv.2004.11362.

Anna Kukleva, Moritz Böhle, Bernt Schiele, Hilde Kuehne, and Christian Rupprecht. Temperature schedules for self-supervised contrastive methods on long-tail data. In *International Conference on Learning Representations (ICLR)*, 2023.

Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, and Honglak Lee. i-mix: A domain-agnostic strategy for contrastive representation learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Junnan Li, Pan Zhou, Caiming Xiong, R. Socher, and S. Hoi. Prototypical Contrastive Learning of Unsupervised Representations. *arXiv preprint arXiv:2005.04966*, 2020. doi: 10.48550/arXiv.2005.04966.

Achleshwar Luthra, Tianbao Yang, and Tomer Galanti. Self-Supervised Contrastive Learning is Approximately Supervised Contrastive Learning. *arXiv preprint arXiv:2506.04411*, 2025. doi: 10.48550/arXiv.2506.04411.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When Does Label Smoothing Help? *arXiv preprint arXiv:1906.02629*, 2019. doi: 10.48550/arXiv.1906.02629.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive Learning with Hard Negative Samples. *arXiv preprint arXiv:2010.04592*, 2020. doi: 10.48550/arXiv.2010.04592.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. *arXiv preprint arXiv:1512.00567*, 2015. doi: 10.48550/arXiv.1512.00567.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748*, 2018. doi: 10.48550/arXiv.1807.03748.

Feng Wang and Huaping Liu. Understanding the Behaviour of Contrastive Loss. *arXiv preprint arXiv:2012.09740*, 2020. doi: 10.48550/arXiv.2012.09740.

Bichen Wu, Rui Cheng, Peizhao Zhang, Tian-Ren Gao, Peter Vajda, and Joseph Gonzalez. Data Efficient Language-supervised Zero-shot Recognition with Optimal Transport Distillation. *arXiv preprint arXiv:2112.09445*, 2021. doi: 10.48550/arXiv.2112.09445.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised Feature Learning via Non-parametric Instance Discrimination. *arXiv preprint arXiv:1805.01978*, 2018. doi: 10.48550/arXiv.1805.01978.

Yihao Xue, Siddharth Joshi, Eric Gan, Pin-Yu Chen, and Baharan Mirzasoleiman. Which features are learnt by contrastive learning? on the role of simplicity bias in class collapse and feature suppression. In *International Conference on Machine Learning (ICML)*, 2023.

Oliver Zhang, Mike Wu, Jasmine Bayrooti, and Noah Goodman. Temperature as uncertainty in contrastive learning. *arXiv preprint arXiv:2110.04403*, 2021.

A Interpolating Between Instance and Class Positives

Table A.1 is the sweep of the specificity axis in the loss this project started from, whose temperature is fixed (Appendix F); Table G.1 is the same axis with the temperature free, which is what §3 reports. In the fixed-temperature loss the fixed class-mate won within-view class structure on both datasets, and at $\rho = 0.254$ it lost every readout to $\varphi = 1$; the free-temperature rerun settles which of these

holds. We had predicted that the class-mate target would also win cross-view class agreement on both datasets; it did on digits and **failed** on Fashion-MNIST, where the views start poorly aligned. Two things in it are worth stating explicitly. First, the direction of every effect is the same on both datasets, and the exchange rate is not: on digits cross-view class agreement falls with φ and on Fashion-MNIST it rises. Second, monotonicity is not universal. Five of the six columns are monotone in φ, and within-view class 5NN on Fashion-MNIST is not: it rises from 71.7 to 74.0 between $\varphi = 0$ and $\varphi = 0.5$ before falling to 69.5. The endpoint difference of -2.11 pp ($p = 0.0097$) is therefore the difference between the ends of a curve that turns, not a summary of a monotone decline.

Table A.1: Instance specificity swept at two sharpness levels in the fixed-temperature loss, mean and standard deviation. Class correctness is 1 and the per-row same-class negative count is fixed at every φ.

	within-view class 5NN	cross-view class agreement	instance top-1
digits, $\rho = 1$			
$\varphi = 0.00$	92.7 ± 0.8	89.7 ± 1.0	3.1 ± 0.6
$\varphi = 0.25$	92.0 ± 0.8	89.3 ± 0.9	5.0 ± 1.1
$\varphi = 0.50$	91.4 ± 0.9	89.2 ± 1.0	7.1 ± 1.4
$\varphi = 0.75$	89.4 ± 1.1	86.5 ± 1.9	10.2 ± 2.0
$\varphi = 1.00$	82.5 ± 2.8	80.3 ± 3.6	20.7 ± 5.3
digits, $\rho = 0.254$			
$\varphi = 0.00$	91.5 ± 0.8	89.3 ± 1.8	3.2 ± 0.8
$\varphi = 0.25$	91.0 ± 0.8	88.0 ± 1.4	8.3 ± 1.1
$\varphi = 0.50$	90.4 ± 0.5	88.2 ± 1.0	14.0 ± 1.1
$\varphi = 0.75$	91.2 ± 0.7	89.0 ± 1.3	24.9 ± 1.4
$\varphi = 1.00$	92.1 ± 0.8	92.6 ± 0.7	49.0 ± 2.0
Fashion-MNIST, $\rho = 1$			
$\varphi = 0.00$	71.7 ± 1.8	65.8 ± 1.5	3.8 ± 1.0
$\varphi = 0.25$	72.6 ± 1.7	68.5 ± 2.1	5.6 ± 1.6
$\varphi = 0.50$	74.0 ± 1.5	71.7 ± 2.2	11.0 ± 1.6
$\varphi = 0.75$	73.1 ± 1.3	74.2 ± 2.3	22.5 ± 2.8
$\varphi = 1.00$	69.5 ± 1.4	85.9 ± 2.1	60.0 ± 5.3
Fashion-MNIST, $\rho = 0.254$			
$\varphi = 0.00$	72.1 ± 1.2	59.5 ± 1.7	1.7 ± 0.7
$\varphi = 0.25$	72.8 ± 1.7	63.3 ± 2.1	6.8 ± 1.1
$\varphi = 0.50$	74.3 ± 1.6	71.3 ± 1.4	21.7 ± 1.8
$\varphi = 0.75$	76.1 ± 0.9	82.1 ± 1.7	44.0 ± 1.8
$\varphi = 1.00$	77.8 ± 1.1	96.7 ± 0.7	84.1 ± 1.5

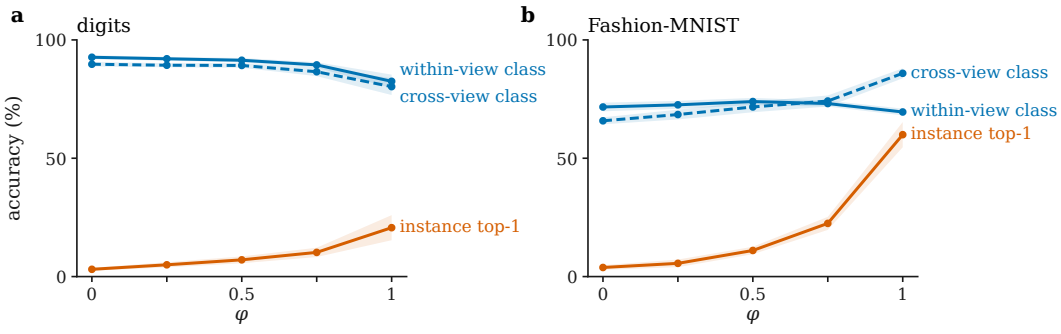

Figure A.1: The fixed-temperature sweep of Table A.1 at $\rho = 1$, mean and one standard deviation. Instance top-1 rises with φ on both datasets. Class structure falls on digits; on Fashion-MNIST cross-view class agreement rises while within-view class structure bends.

B Experimental Protocol

Every verdict in this paper was scored against predictions, thresholds and failure conditions written down before the runs. Eight failed: the class gain of a fixed class-mate on Fashion-MNIST (§3), the memorisation check at high temperatures (§4), the size of the free-temperature sharpness cost on Fashion-MNIST and the harm of one-hot on a correct target (§6), the sign flip under a linear readout (§7), the direction of the cross-view exchange rate in the fixed-temperature sweep (Appendix A), and one prediction each in Appendices H and I.

The one exception is the stopping rule of §5. The first grid put every optimum on the τ boundary; a first extension moved one of four cells from failing $J1'$ to clearing it; only then was the rule *extend until both compared cells are interior* written, and a second extension followed. The rule did not select the answer, since $J1'$ still fails in three of four settings, and each extension widened the search for the one-hot competitor too.

C Full Results of the Sharpness and Temperature Grid

Table C.1 gives every cell behind Figure 4 and Table 2. The comparison $J1'$ makes is between each panel's bold entry and the underlined entry in its $\rho = 1$ column, and both are interior to the grid in all four panels. The table stops at $\tau = 8$; a later predeclared stage ran $\tau \in \{16, 32, 64\}$ in the two leftmost columns of every panel. On digits without the anchor, Fashion-MNIST without it and Fashion-MNIST with it, the $\rho = 0.014$ column falls from its $\tau = 8$ value of 40.4, 85.5 and 58.6% to 38.5, 83.0 and 51.1% at $\tau = 16$, so every optimum in Figure 4 is interior, and no cell above $\tau = 8$ exceeds its column's maximum. The ridge is visible directly in the tables as the diagonal band of maxima: at each ρ the best τ moves one or two rows as ρ moves one column.

D Full Results for Target Correctness

Table D.1 is the surface behind §6. The starred cells are the best cross-view agreement in each row and are what the out-of-sample prediction was about: they move right as the target becomes more correct on digits ($\rho^\star = 0.055, 0.055, 0.75$ at $c = 0.42, 0.70, 1.00$), and Fashion-MNIST was chosen because a zero-training probe put its own inferred plan at $c = 0.806$ against digits' 0.424.

The row that does not fit is Fashion-MNIST at $c = 0.4238$, where the best ρ is 1 rather than the small value the digits ordering would suggest. On that dataset the two views begin below chance cross-view class agreement, so what a sharp target buys dominates what it destroys. The honest statement of the rule is therefore not that $\rho^\star$ increases with c. It is that sharpening buys cross-view correspondence and pays in within-view class structure at a price proportional to the target's error, and where the optimum falls depends on both sides of that trade. Correctness is one input to it, not the whole rule.

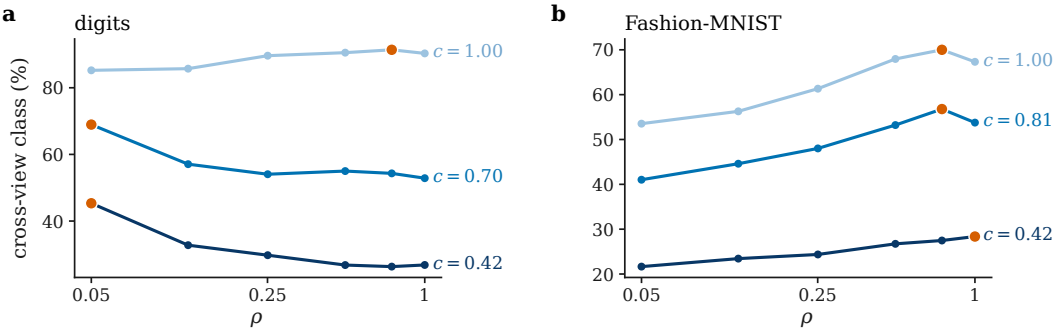

Figure D.1: Cross-view class agreement over the surface of Table D.1; darker is a more wrong target, and the orange point marks each curve's best ρ. On digits the best ρ moves from 0.055 to 0.75 as c rises. On Fashion-MNIST it is 0.75 or 1 for every c, including $c = 0.42$.

Table C.1: The whole sharpness by temperature grid: mean instance top-1 (%) in every cell. Bold is each panel's maximum and the underlined entry is the best cell of the $\rho = 1$ column, which is the one-hot competitor of Table 2.

digits, no anchor

$\tau\backslash\rho$	0.01367	0.05469	0.12717	0.25446	0.5	0.75	1
0.05	11.4	27.3	36.5	40.7	41.1	37.7	37.6
0.1	15.5	29.3	37.7	42.4	40.6	36.2	40.2
0.2	18.7	32.0	40.9	43.4	44.4	46.7	<u>47.2</u>
0.5	22.0	38.8	43.8	**48.1**	46.5	44.5	42.9
1	24.4	41.7	47.0	42.7	39.2	36.7	33.8
2	31.2	44.4	39.9	37.0	31.5	27.6	25.6
4	37.3	39.6	36.3	30.1	24.2	21.5	19.8
8	40.4	36.0	28.3	24.5	19.6	16.5	14.2

digits, reconstruction anchor

$\tau\backslash\rho$	0.01367	0.05469	0.12717	0.25446	0.5	0.75	1
0.05	10.7	29.2	36.5	42.9	42.7	39.0	39.4
0.1	14.1	31.1	40.8	44.1	41.4	38.1	41.4
0.2	17.7	34.2	41.4	45.0	44.5	47.1	<u>47.6</u>
0.5	17.0	35.8	47.3	49.0	47.0	44.9	43.3
1	15.9	46.1	51.5	47.3	43.6	40.2	37.0
2	14.8	**55.1**	51.5	47.3	40.8	36.7	33.9
4	14.9	54.6	51.9	47.7	40.4	36.6	33.1
8	14.3	46.7	49.9	44.9	40.0	33.6	29.7

Fashion-MNIST, no anchor

$\tau\backslash\rho$	0.01367	0.05469	0.12717	0.25446	0.5	0.75	1
0.05	11.1	51.2	71.7	79.9	78.2	74.1	80.8
0.1	18.1	57.6	75.3	80.9	79.9	79.0	81.6
0.2	28.0	65.1	79.7	84.8	84.8	86.9	<u>88.8</u>
0.5	42.8	77.9	87.6	88.4	**89.6**	88.7	87.9
1	53.2	84.6	89.1	88.5	85.4	84.0	82.8
2	64.5	88.3	86.3	83.7	80.9	79.5	77.4
4	80.6	84.9	81.2	78.7	76.6	73.1	71.0
8	85.5	81.0	77.9	75.8	70.2	67.0	63.1

Fashion-MNIST, reconstruction anchor

$\tau\backslash\rho$	0.01367	0.05469	0.12717	0.25446	0.5	0.75	1
0.05	8.8	38.8	71.4	80.8	80.5	76.7	81.3
0.1	15.0	45.9	74.8	83.4	81.0	79.6	82.2
0.2	20.0	53.5	80.8	84.8	85.5	87.8	87.8
0.5	24.9	62.7	87.8	89.5	89.7	88.6	<u>88.4</u>
1	28.2	80.3	**90.0**	88.9	87.4	85.1	83.5
2	31.4	87.9	88.3	87.1	84.8	82.8	82.3
4	43.4	85.1	86.8	86.7	85.0	83.2	82.4
8	58.6	84.2	88.0	87.9	86.6	85.0	83.8

E Full Results for All Evaluation Metrics

Table E.1 gives all four readouts for every arm of both datasets, from which the three numbers quoted in §7 are drawn. The pattern to read off it is that the raw and whitened columns agree closely wherever the target is class-correct and separate widely wherever it is not, and that the transfer column, which uses no nearest neighbour anywhere, selects a different ρ from the other three in most rows.

Table D.1: The authority by correctness surface, mean (%). Within is within-view class 5NN and cross is cross-view class agreement. The starred entry is each row's best cross-view cell, which is the quantity the out-of-sample prediction of §6 was about.

		ρ					
Dataset, c	readout	0.05469	0.12717	0.25446	0.5	0.75	1
digits, 0.4238	within	91.9	70.7	57.3	46.7	43.2	39.8
	cross	**45.3***	32.8	29.8	26.8	26.3	26.8
digits, 0.7002	within	92.0	77.6	71.9	68.7	66.7	64.7
	cross	**68.9***	57.1	54.0	55.0	54.3	52.9
digits, 1	within	92.9	90.9	92.0	92.6	93.2	92.7
	cross	85.2	85.7	89.6	90.5	**91.4***	90.3
Fashion-MNIST, 0.4238	within	71.1	60.7	60.4	55.0	47.7	41.4
	cross	21.7	23.4	24.4	26.7	27.5	**28.4***
Fashion-MNIST, 0.8057	within	72.5	65.8	67.8	66.4	66.6	62.7
	cross	41.0	44.6	48.0	53.2	**56.8***	53.8
Fashion-MNIST, 1	within	75.2	71.2	72.4	75.5	74.9	71.8
	cross	53.5	56.3	61.3	68.0	**70.0***	67.3

F Implementation Details

Encoders and optimisation. Both settings use two independent encoders, one per view, each an MLP of width $\to$ 128 $\to$ 64 $\to$ 16 with ReLU. The view width is 32 on digits (8×8 images split into even and odd pixel columns) and 392 on Fashion-MNIST (28×28, same split). On both datasets the data are 1024 training and 384 validation samples drawn from a fixed pool of 1797 images (all of digits, and a random subset of the Fashion-MNIST training set of the same size), with the y view permuted inside each split so that no index correspondence survives. Data splits are fixed and every random seed is set, so seeds vary initialisation and batch order but not the data; the p-values therefore understate variance across data draws, and they are not corrected for multiple comparisons. The verdicts rest on the predeclared tests, most of which are far below any correction threshold. Every arm is repeated with ten seeds (three for the CIFAR-100 stage of Appendix H, five for the three arms of its replication), and every mean, standard deviation and paired t-test in this paper is taken over those repetitions.

The fixed-temperature loss. Appendices A, D and I, the first half of §6, §7 and Figure 3 use the alignment loss of the release this project started from: a soft-target cross entropy on a Gibbs kernel built from squared Euclidean distances, scaled by a detached moving median, with a 16-dimensional latent followed by an affine LayerNorm and a separate decoder per view carrying a reconstruction term at weight 0.5. Optimisation is AdamW at learning rate 9×10^{-4} and weight decay 10^{-4} for 1000 updates, full batch over the 1024 training samples, with no convergence claim. Its temperature is fixed, which is the confound §5 exists to resolve.

The loss of §3 to §6. Sections 3, 4, 5 and the free-temperature half of §6 use a separate implementation written from scratch, not the trainer above: cosine similarity, an explicit temperature and a symmetric cross entropy against T of Eq. 2, in minibatches of 256 for 400 epochs under AdamW at learning rate 10^{-3} and weight decay 10^{-4}, with no pretraining and no reconstruction term unless the anchor arm is being run. Embeddings are L2-normalised at the encoder output. The sum against a uniform-floor target is evaluated in closed form,

$$\textstyle\sum_{ij} T_{ij} L_{ij} = \frac{1-\rho}{b^2} \sum_{ij} L_{ij} + \frac{\rho}{b} \sum_i L_{ii},$$

which avoids materialising T. The two implementation checks reported in §5 compare this against the textbook form at $\rho = 1$ and against a brute-force construction of T with an explicit double sum at $\rho < 1$.

Readouts. Within-view class 5NN retrieves the five nearest *training* embeddings of a validation point and votes on their labels. Cross-view class agreement and instance top-1 are nearest-neighbour retrievals between the two views' validation embeddings. Effective rank is the exponential of the

Table E.1: The readout audit in full. Four readouts applied to identical frozen embeddings, mean (%). Raw and whitened are within-view class 5NN before and after ZCA; probe is a within-view linear probe; transfer is a linear head fitted on one view and evaluated on the other, with no nearest neighbour anywhere.

Dataset	Arm	raw 5NN	probe	whitened 5NN	transfer
digits, $c = 0.4238$					
	$\rho = 0.05469$	91.9	88.7	91.2	28.3
	$\rho = 0.12717$	70.7	74.0	73.3	38.5
	$\rho = 0.25446$	57.3	69.1	69.6	36.6
	$\rho = 0.5$	46.7	69.4	73.5	38.4
	$\rho = 0.75$	43.2	67.0	75.8	39.4
	$\rho = 1$	39.8	65.6	76.1	40.1
digits, $c = 0.7$					
	$\rho = 0.05469$	92.0	89.5	91.5	49.7
	$\rho = 0.12717$	77.6	79.4	78.8	63.7
	$\rho = 0.25446$	71.9	76.6	78.0	62.1
	$\rho = 0.5$	68.7	77.3	80.5	65.2
	$\rho = 0.75$	66.7	76.5	81.6	67.0
	$\rho = 1$	64.7	74.8	82.3	66.8
digits, $c = 1$					
	$\rho = 0.05469$	92.9	91.9	92.8	76.4
	$\rho = 0.12717$	90.9	91.2	91.1	91.2
	$\rho = 0.25446$	92.0	91.9	92.1	91.8
	$\rho = 0.5$	92.6	92.9	92.3	92.8
	$\rho = 0.75$	93.2	93.1	92.1	93.2
	$\rho = 1$	92.7	92.5	91.2	92.5
Fashion-MNIST, $c = 0.4238$					
	$\rho = 0.05469$	71.1	70.9	66.5	23.5
	$\rho = 0.12717$	60.7	64.8	60.9	27.0
	$\rho = 0.25446$	60.4	64.9	61.3	28.9
	$\rho = 0.5$	55.0	62.5	60.5	31.3
	$\rho = 0.75$	47.7	61.7	61.0	33.6
	$\rho = 1$	41.4	60.0	61.7	35.1
Fashion-MNIST, $c = 0.8057$					
	$\rho = 0.05469$	72.5	73.2	69.7	32.6
	$\rho = 0.12717$	65.8	67.6	65.8	56.2
	$\rho = 0.25446$	67.8	68.6	68.3	58.5
	$\rho = 0.5$	66.4	69.1	68.4	62.1
	$\rho = 0.75$	66.6	70.8	69.5	65.0
	$\rho = 1$	62.7	69.2	67.3	63.5
Fashion-MNIST, $c = 1$					
	$\rho = 0.05469$	75.2	74.9	74.1	45.0
	$\rho = 0.12717$	71.2	71.4	71.3	70.8
	$\rho = 0.25446$	72.4	72.9	72.5	72.9
	$\rho = 0.5$	75.5	74.9	73.1	74.7
	$\rho = 0.75$	74.9	76.2	73.6	76.0
	$\rho = 1$	71.8	74.1	71.3	73.6

entropy of the normalised covariance spectrum. The linear probe, the whitened 5NN and the cross-view linear transfer of §7 are all applied to stored embeddings without retraining any encoder.

G The Specificity and Correctness Stages under a Free Temperature

These two stages answer whether the specificity and correctness results of the fixed-temperature loss survive a tuned temperature. Both were predeclared before any run and use the implementation of §5 unchanged, including its budget of 400 epochs in minibatches of 256. The only addition is a bijection m over the training set: a batch is $(x_i, y_{m(i)})$ for the sampled rows i, and T is the uniform-floor target of Eq. 2 on that batch. For the specificity stage, a fraction φ of rows keep $m(i) = i$ and

the rest are deranged in one cycle inside their class, drawn once, or redrawn before every update; because m maps each class onto itself, the class composition of each batch is identical across arms. For the correctness stage, a fraction c of rows keep $m(i) = i$ and the rest are sent to samples of other classes by a bijection among themselves. The grid is $\rho \in \{0.055, 0.254, 1\}$ against the eight temperatures of Figure 4, ten runs per cell, with and without the reconstruction anchor for specificity and without it for correctness. We saw one cell of the specificity stage before writing its predictions, the one-hot target at $\tau = 0.2$ for one run, and say so in the protocol.

Four predictions were scored for specificity, all with the temperature and sharpness chosen per arm and readout. The fixed class-mate beats $\varphi = 1$ within view and loses instance retrieval: passed on digits, **failed** on Fashion-MNIST. The redrawn class-mate beats the fixed one within view and has lower held-out spread: passed on both. The redrawn class-mate beats $\varphi = 1$ within view and loses instance retrieval: passed on both. Memorisation above 20% in every one-hot cell of the fixed arm and below 5% in the redrawn arm: the second half held everywhere, the first **failed** at $\tau \geq 4$ on both datasets. For correctness, the cost of sharpness at $c = 0.42$ exceeds that at $c = 1$ by more than 5 pp: passed on digits ($+10.0$ pp), **failed** on Fashion-MNIST ($+4.1$ pp, $p = 3{\times}10^{-5}$); the cost at $c = 1$ is below 3 pp: passed on both.

Table G.1: The specificity axis under a free temperature (§3). Each entry is the best cell of one arm for one readout over $\rho \in \{0.055, 0.254, 1\}$ and eight temperatures, mean over ten runs, with that cell's (ρ, τ). Class correctness is 1 and the class composition of every batch is identical across arms.

arm	within-view class 5NN	cross-view class	instance top-1
digits, no anchor			
$\varphi = 1$	91.3 (0.055, 0.2)	91.7 (0.254, 1)	48.4 (0.254, 0.5)
$\varphi = 0.5$	93.3 (0.254, 4)	91.7 (0.254, 4)	13.2 (0.254, 0.5)
$\varphi = 0$, fixed	93.2 (0.055, 8)	91.7 (0.055, 8)	3.7 (1, 1)
$\varphi = 0$, redrawn	94.7 (0.254, 0.1)	93.8 (0.254, 0.2)	3.9 (0.254, 1)
Fashion-MNIST, no anchor			
$\varphi = 1$	78.3 (0.254, 1)	98.1 (0.055, 2)	88.6 (0.254, 0.5)
$\varphi = 0.5$	78.0 (0.055, 8)	78.8 (0.055, 8)	28.4 (0.055, 2)
$\varphi = 0$, fixed	77.0 (0.055, 4)	70.4 (1, 1)	4.3 (1, 1)
$\varphi = 0$, redrawn	81.6 (0.254, 0.5)	78.1 (0.254, 0.2)	6.2 (0.254, 2)
digits, reconstruction anchor			
$\varphi = 1$	93.0 (0.055, 4)	94.5 (0.055, 2)	55.4 (0.055, 2)
$\varphi = 0.5$	93.8 (0.055, 4)	93.1 (0.055, 4)	26.7 (0.055, 4)
$\varphi = 0$, fixed	94.0 (0.055, 4)	92.5 (0.254, 2)	6.4 (0.055, 8)
$\varphi = 0$, redrawn	94.6 (0.055, 2)	94.4 (0.055, 0.5)	7.4 (0.055, 8)
Fashion-MNIST, reconstruction anchor			
$\varphi = 1$	79.7 (0.055, 2)	98.3 (0.254, 0.5)	89.6 (0.254, 0.5)
$\varphi = 0.5$	78.5 (0.055, 8)	83.5 (0.055, 8)	38.2 (0.055, 8)
$\varphi = 0$, fixed	77.8 (0.055, 8)	70.7 (1, 1)	4.4 (1, 8)
$\varphi = 0$, redrawn	81.4 (0.254, 1)	77.8 (0.254, 2)	8.9 (1, 8)

Table G.2: Sharpness against target class correctness under a free temperature (§6). Within-view class 5NN at the best temperature for each ρ, mean over ten runs; cost is the drop from $\rho = 0.055$ to $\rho = 1$. Rows with $c < 1$ point to the true partner or to a sample of another class; $c = 1$ is standard InfoNCE.

dataset	c	$\rho = 0.055$	$\rho = 0.254$	$\rho = 1$	cost
digits	0.42	71.8	60.5	61.1	$+10.7$
	0.70	82.5	76.1	72.3	$+10.2$
	1.00	91.3	91.1	90.6	$+0.7$
Fashion-MNIST	0.42	57.6	57.2	53.8	$+3.8$
	0.70	68.8	67.4	65.6	$+3.1$
	1.00	77.9	78.3	78.1	-0.2

H EXPERIMENTS ON CIFAR-100

This stage repeats §3 and §4 after changing four things at once: the encoder, the way the two views are built, the number of training samples, and the number of classes. It was predeclared before any run was launched, and the protocol was not edited afterwards.

Setting. CIFAR-100, $45{,}000$ training and $5{,}000$ validation images under a fixed split, test never opened. A CIFAR ResNet-18 with a 3×3 stem and no max pooling, followed by a projection head of width $512 \to 512 \to 128$ with L2 normalisation. The two views are two independent augmentations of one image, which is the construction §2 does not use. Optimisation is SGD with momentum 0.9, weight decay 10^{-4}, batch 512, learning rate 0.3 with one warmup epoch and then a cosine decay, 200 epochs, temperature fixed at 0.2.

Batches. Each batch is 16 coarse groups, 4 fine classes in each, and 8 images in each fine class. This structure is what makes a $\varphi = 0$ target constructible at all: the target must stay a permutation so that both its margins remain uniform, and a random batch of 512 over 100 classes holds about two images per class, which is too few. Every arm shares the structure, and an extra arm draws batches at random to show that the structure is not what produces the effect. It is not, as the first two rows of Table H.1 show.

Targets. Every target has $\rho = 1$, so every row places all of its mass on one column except the supervised arm. What moves is which column. The $\varphi = 0$ fine target rolls each block of eight by one position, so every row is answered by a different image of the same fine class. Because batches are redrawn at every step, the class-mate each image is paired with changes from step to step, whereas on digits and Fashion-MNIST the $\varphi = 0$ map is drawn once and kept for the whole of training. The $\varphi = 0$ coarse target moves each block of eight to the next fine class inside the same coarse group, so the answer shares the coarse label and not the fine one. The $\varphi = 0.5$ target sends half of each block around a four cycle and leaves the other half alone. The supervised arm spreads the row uniformly over all eight images of the fine class, the true partner included. All six were audited before training against forty three checks, covering the batch structure, both margins, the fact that each permutation is a permutation, the realised value of φ, and the class of the image each row is sent to.

Predictions and outcomes.

- J1 Instance retrieval decreases along $\varphi = 1$, $\varphi = 0.5$, $\varphi = 0$, with the first minus the last above 10 pp. **Passed**: 81.18, 30.72, 13.56.
- J2 The coarse demand costs more instance retrieval than the fine one, because a coarse class holds five times as many images. **Passed**: 71.97 pp against 67.62 pp, paired $t = 26.7$, $p = 7\times10^{-4}$. A later predeclared stage added two runs to the three arms this prediction compares, on full cards instead of slices, with nothing else changed; over five runs the harms are 72.22 and 67.93 pp ($t = 31.5$, $p = 3\times10^{-6}$), and both new runs go the same way.
- J3 The supervised target keeps the true partner as a positive, so it should lose less than half of what the $\varphi = 0$ target loses. **Failed**: 67.33 pp against a threshold of 33.81.
- J4 Effective rank decreases along $\varphi = 1$, $\varphi = 0$ fine, $\varphi = 0$ coarse. **Passed**: 75.78, 46.41, 17.63.
- J5 The arms ordered by class retrieval differ from the arms ordered by instance retrieval. **Passed**: $\varphi = 1$ is first of the six by instance retrieval and fifth of the six by fine class retrieval.

There is an exact reason for the coincidence. Because the images of a fine class appear in random order and batches are redrawn at every step, the $\varphi = 0$ fine target places, in expectation, $1/7$ of each row on each of the other seven images of its fine class in the batch; the class-mass target places $1/8$ on each of the eight, the true partner included. The two targets differ in expectation only by that one eighth. After the fact, instance retrieval follows the mass on the true partner ($1, 1/2, 1/8, 0$ give $81.18, 30.72, 13.85, 13.56$); this is a description, not a tested prediction.

Table H.1: The specificity axis on CIFAR-100, measured on the projection. Six targets over one shared batch structure, one shared set of negatives, identical initial weights. Values are mean±sd. Mass is what each target row places on the row's true augmented partner. The first arm draws batches at random and the rest draw them class balanced. Class labels enter the loss for every arm below the second, and the batch construction for every arm except the first. These are diagnostic controls and support no unsupervised claim.

Target	Mass	Instance	Fine 200NN	Coarse 200NN	Eff. rank	Probe
$\varphi = 1$, random batches	1	79.72±0.23	53.61±0.50	67.57±0.13	78.62±0.43	59.11±0.08
$\varphi = 1$	1	81.18±0.18	52.03±0.15	65.93±0.06	75.78±0.15	58.03±0.57
$\varphi = 0.5$, fine	1/2	30.72±0.56	71.34±0.36	81.76±0.13	53.45±0.39	71.35±0.06
Class mass, fine	1/8	13.85±0.05	71.43±0.04	81.70±0.39	58.73±0.28	71.89±0.45
$\varphi = 0$, fine	0	13.56±0.26	70.88±0.52	81.85±0.26	46.41±0.21	71.25±0.48
$\varphi = 0$, coarse	0	9.21±0.52	34.76±0.29	81.36±0.56	17.63±0.07	65.62±0.54

The coarse demand leaves *fine* class retrieval at 34.76 against 53.61 for an arm that never sees a label, so training on coarse labels makes a finer readout worse than training on no labels. That is what the account of §4 predicts, since the only content a coarse demand carries over is the coarse class, and it is also a warning about reading any single class readout as representation quality. The same effect is much weaker on the backbone, 59.61 against 53.15, so the collapse sits mostly in the projection head.

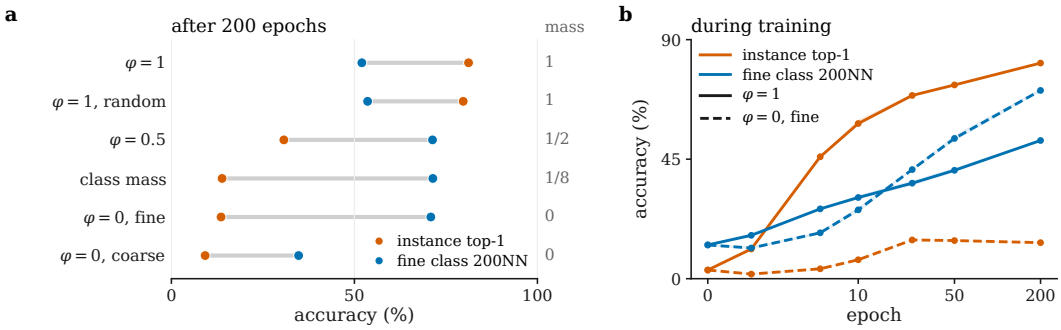

Figure H.1: CIFAR-100, projection head. (a) Every arm of Table H.1 after 200 epochs, one row each, with instance top-1 in orange and fine class 200NN in blue; "random" draws batches at random instead of class balanced. Rows are ordered by the mass each target row places on the true partner, given on the right. The class-mass row and the fine $\varphi = 0$ row nearly coincide. (b) Instance top-1 and fine class 200NN during training for $\varphi = 1$ (solid) and fine $\varphi = 0$ (dashed), with the range over runs shaded. Instance retrieval separates by epoch 5 (45.9 against 3.7). Fine class retrieval is at first higher under $\varphi = 1$ and is overtaken by the $\varphi = 0$ arm between epochs 10 and 25.

Scope. This stage does not test §5 or §6; the sharpness and temperature grid was not rerun at this scale. Its $\varphi < 1$ arms use class labels in the loss and every balanced arm uses them to build batches, so none of it is an unsupervised result. Twelve of the eighteen runs were stopped by a wall-clock limit and rerun with nothing else changed; which runs were stopped was decided by load on shared hardware and not by any result.

I Memorisation of the Assigned Class-Mate

This stage tests the account of §4. It was predeclared before any run, after a zero-training check on the stored $\varphi = 0$ encoders showed that they memorise the assigned class-mate on the training set.

Design. Three arms on both datasets under the trainer and budget of §4. The first is the $\varphi = 0$ arm of that section, with the within-class derangement σ drawn once. The second starts from the same σ and redraws it with the same construction before every update, so that class correctness, instance

correctness, the bijection and the same-class-negative count are unchanged at every step and only the possibility of memorising σ is removed. The third is $\varphi = 1$. The first and third arms reproduce the $\rho = 1$ rows of Table A.1.

Predictions and outcomes. The manipulation check required memorisation above 20% in the fixed arm and below 5% in the redrawn arm on every run; it **passed**. The main prediction was that held-out within-class spread is lower in the redrawn arm on both datasets (one-sided paired $p < 0.01$); it **passed** on both. A second prediction, that held-out instance top-1 would also fall, **failed** on both: it did not move. On reflection a memorised pairing has no bearing on held-out instances, so the prediction was not implied by the account. Not predicted, both class readouts rose when memorisation was removed.

Table I.1: Redrawing the class-mate at every update. Memorised is the share of training rows whose nearest view-y embedding is the assigned partner; the other columns are held-out readouts. Spread is within-class angular spread over the mean angle between class means; Within and Cross are within-view class 5NN and cross-view class agreement. Values are mean±sd; the last row of each block is the paired difference between the two $\varphi = 0$ arms and its two-sided p.

Target	Memorised	Spread	Instance	Within	Cross
digits					
$\varphi = 0$, fixed	39.9±5.0	0.149±0.008	3.07±0.61	92.67±0.82	89.73±0.96
$\varphi = 0$, redrawn	1.0±0.3	0.106±0.011	3.39±0.90	93.76±0.47	91.84±0.85
$\varphi = 1$	84.5±6.9	0.341±0.037	20.66±5.30	82.50±2.84	80.26±3.56
redrawn − fixed	−39.0 (1e-9)	−0.043 (1e-6)	+0.31 (0.39)	+1.09 (8e-4)	+2.11 (4e-4)
Fashion-MNIST					
$\varphi = 0$, fixed	51.7±11.8	0.270±0.026	3.84±0.97	71.65±1.84	65.82±1.49
$\varphi = 0$, redrawn	0.9±0.3	0.194±0.017	4.47±0.91	74.26±1.80	69.30±1.41
$\varphi = 1$	94.1±1.8	0.418±0.028	59.96±5.32	69.54±1.42	85.89±2.12
redrawn − fixed	−50.8 (3e-7)	−0.076 (6e-5)	+0.63 (0.22)	+2.60 (7e-4)	+3.48 (9e-4)

J The Failed Column of the Sharpness and Temperature Grid

At $\rho = 0.0137$ on digits with reconstruction, no temperature in Figure 4 exceeds 17.7% instance top-1, while the same column without reconstruction reaches 40.4% at $\tau = 8$. This stage asks whether the column fails because such a target cannot be learned or because of the setting. It was predeclared before any run.

Observation. At this ρ the target is close to uniform, and the contrastive loss moves by about 10^{-3} nats over training, from $\log 256$. The reconstruction term is larger than that at the start. With a single loss AdamW is insensitive to its scale, which is why the column learns without reconstruction; with both terms their relative size decides the update. Top-1 at $\tau = 8$ was still rising at the end of the fixed budget.

Design and outcome. Ten runs per arm on digits, $\rho = 0.0137$, everything else as in §5. Training ten times longer at reconstruction weight 0.5 and $\tau = 8$ raises top-1 from 14.3% to 32.4%; keeping the budget and weighting reconstruction ten times less raises it to 53.7% (paired $p = 3 \times 10^{-11}$ and 2×10^{-11}). Both cleared the predeclared threshold of 30%. At $\tau = 2$ ten times longer training gives only 20.7%, so the weak target also needs a high temperature. The same longer budget lowers the best cell of the panel, $\rho = 0.055$ and $\tau = 2$, from 55.1% to 50.0%, so a longer budget does not improve the grid as a whole and the rest of Figure 4 stands as measured.