PREDICTION DISTILLATION CAN BREAK EDITABILITY IN CONCEPT-EMBEDDING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept models expose human-named attributes for users to correct, but these edits help only if the classifier acts on them. We ask whether matching a teacher's predictions through knowledge distillation preserves this ability in concept-embedding models. Edits replace concept probabilities but retain image-dependent embeddings, allowing a student to preserve teacher distinctions absent from the annotated concepts. On CUB with frozen ResNet-18 features, distillation raises task accuracy from 59.3% to 64.9% but lowers accuracy after full concept correction from 99.9% to 88.0%. Predictions also show much weaker alignment with supplied alternative-class concepts. With frozen features, retaining distillation while shuffling teacher targets among images with identical concepts or replacing them with mean teacher logits largely restores correction benefits. These controls respectively break image–target pairing and remove image-specific variation within each group. The pattern extends across image encoders, to a second dataset, and to a nonlinear classifier. Under end-to-end training, the effect is weaker and mean-logit targets provide only partial recovery. Changes in correction accuracy are small on datasets with incomplete concepts. Prediction after an edit should therefore be evaluated alongside ordinary accuracy.

1 INTRODUCTION

A concept bottleneck model (CBM) predicts a class from human-named attributes such as a bird's wing pattern (Koh et al., 2020). Exposing these concepts lets an expert correct a mistaken attribute and ask the classifier to act on the correction. Concept-embedding models (CEMs) reduce the accuracy cost of a narrow bottleneck with image-dependent concept embeddings (Espinosa Zarlenga et al., 2022). The classifier receives these embeddings weighted by concept probabilities. An edit changes the probabilities but retains the original embeddings. Randomly corrected concepts during training encourage useful interventions, yet other objectives may reward information the edit cannot replace.

Knowledge distillation trains a student to match a teacher's predictions (Hinton et al., 2015). When those predictions differ between images with identical concepts, matching them rewards distinctions beyond the annotations. Image-dependent embeddings offer a route to retain these distinctions, potentially improving unedited accuracy while weakening corrections. Ordinary task and concept accuracy would miss this conflict.

We compare students with and without distillation under identical intervention-aware training. A linear decomposition identifies the terms fixed by an edit. We evaluate gold-concept corrections, alternative-class templates, and neutralizing probabilities to 0.5 to measure retained classification ability. Keeping distillation, we shuffle teacher targets among images with identical annotated concepts to break image–target pairing, or use the group's mean teacher logits to remove image-specific variation without random reassignment.

With frozen features, distillation improves unedited accuracy but weakens corrections and alignment with alternative-class concepts. The target controls largely restore correction benefits. The pattern extends across encoders, to a second dataset, and to a nonlinear classifier. Under end-to-end training, the effect is weaker and mean-logit targets provide partial recovery. Incomplete-concept datasets show small changes in correction accuracy, illustrating why retained image information is not always harmful. Distilled concept models should therefore be evaluated after edits, including simulated

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.

sparse corrections under limited editing budgets (Shin et al., 2023; Chauhan et al., 2023), alongside ordinary accuracy.

2 Related Work

Concept models as intervention interfaces. CBMs (Koh et al., 2020) make supervised concepts an intermediate state that a user can inspect and correct. Post-hoc (Yuksekgonul et al., 2023), label-free (Oikarinen et al., 2023), and probabilistic (Kim et al., 2023) variants address how this state is learned or represented; interactive models study its use during intervention (Chauhan et al., 2023). CEMs (Espinosa Zarlenga et al., 2022) address the loss of predictive capacity caused by a narrow bottleneck. Their RandInt training procedure exposes the model to corrected concepts so that the added embeddings remain useful for interventions. We keep this training active and ask whether subsequently adding a teacher's prediction targets changes the behavior of the resulting interface.

Leakage, bypasses, and residual information. Learned concept representations can contain information beyond their assigned labels (Mahinpei et al., 2021; Margeloiu et al., 2021), motivating methods that reduce or account for this leakage (Havasi et al., 2022). More recent work connects information-theoretic leakage measures to intervention behavior and analyzes several forms of leakage in CEMs (Parisini et al., 2026). These results establish that representation quality and intervention behavior are related. Our question concerns a specific training choice: does matching input-paired teacher predictions weaken concept edits, and does removing the within-concept pairing change that outcome?

Knowledge distillation. Prediction distillation trains a student to match a teacher's output distribution (Hinton et al., 2015). Feature transfer, contrastive objectives, and self-distillation extend the available training signals (Romero et al., 2015; Tian et al., 2020; Furlanello et al., 2018); analyses examine transfer efficacy and the roles of model capacity and training choices (Cho & Hariharan, 2019; Stanton et al., 2021; Beyer et al., 2022). Here the student's purpose adds a second criterion: a useful teacher signal should also leave the concept interface responsive. Comparing input-paired targets with targets shared within a concept stratum separates this concern from ordinary predictive performance.

Faithfulness and controllability. Concept-based explanation methods and their checks ask what a representation reveals about a prediction (Kim et al., 2018; Adebayo et al., 2018). Intervention studies ask what a user can change through that representation. Both matter when interpretability is intended to support decisions (Rudin, 2019). We therefore evaluate the prediction obtained after an edit directly, alongside concept accuracy; plausibly named, accurately predicted concepts alone do not establish that the classifier will use a correction effectively.

3 What a Concept Edit Can Change

The intended intervention. Let $p(x)$ denote the concept probabilities predicted from an image and c' a supplied concept vector. We study an interface in which the user replaces $p(x)$ by c' and asks the classifier to act on those values. For example, replacing an erroneous wing-pattern prediction with its annotation should help the species decision. This expectation is most informative when the supplied concepts distinguish the relevant classes.

We check this condition using the datasets' gold class templates: we identify their source, count classes with identical templates, and test whether the gold template uniquely identifies its class. When templates coincide, a concept edit cannot distinguish those classes. When concepts omit label-relevant information, retaining other image information may instead be useful. The primary experiments use distinct class templates; incomplete instance-level concepts provide the boundary comparison.

Concept-embedding prediction. A CEM represents each concept with two embeddings computed from the image: one for its presence and one for its absence. Its predicted probability chooses

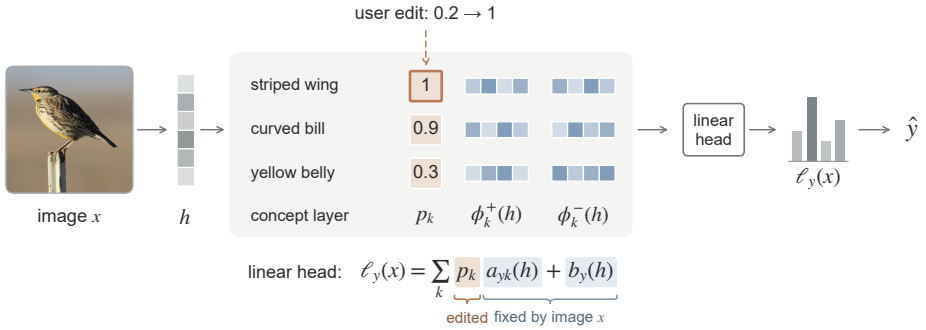


Figure 1: Editing changes p_k but leaves the image-dependent embeddings, $a_{yk}(h)$ and $b_y(h)$ fixed for a linear head. Values are illustrative. Photo: Krista Lundgren/USFWS (public domain).

their mixture. Writing the image features as h, the probability as $p_k = \sigma(s_k(h))$, and the two embeddings as $\phi_k^+(h)$ and $\phi_k^-(h)$, the representation supplied to the label head is

$$z_k = p_k \phi_k^+(h) + (1 - p_k)\phi_k^-(h).$$

To see what an edit leaves unchanged, consider a linear label head and collect the terms multiplying each p_k. Each class logit becomes

$$\begin{aligned} \ell_y(x) &= \sum_k p_k \, a_{yk}(h) + b_y(h), \\ a_{yk}(h) &= \left\langle W_{yk}, \phi_k^+(h) - \phi_k^-(h) \right\rangle, \end{aligned} \tag{1}$$

Here W_{yk} is the class-y, concept-k block of the label-head weights. The scalar $a_{yk}(h)$ tells us how changing concept k affects that class logit for this image, while

$$b_y(h) = b_{0y} + \sum_k \left\langle W_{yk}, \phi_k^-(h) \right\rangle$$

is the part independent of the edited probabilities. In $\ell_y(x \mid p \leftarrow c')$, we hold the original features h, embeddings $\phi_k^{\pm}(h)$, coefficients $a_{yk}(h)$, and offsets $b_y(h)$ fixed and replace only p_k by c'_k. Thus the same concept values can yield different logits for different original images.

Figure 1 makes the two sources of information visible. The edited probabilities still affect the logits through $a_{yk}(h)$; the decomposition alone does not imply that edits have little effect. It shows where image-specific information can remain after an edit. The experiments test whether distillation makes that information influential enough to impair correction. For a nonlinear head the additive identity no longer applies, but the same edited forward pass can still be evaluated (Section 5.3).

A margin test for limited edits. The linear decomposition also gives an image-specific condition under which a small editing budget cannot change a decision. For a current winner y and a competing class t, define $d_{yt,k} = a_{yk}(h) - a_{tk}(h)$ and the initial margin $M_{yt} = \ell_y(x) - \ell_t(x)$. An edit on coordinates S changes this margin exactly by

$$M'_{yt} - M_{yt} = \sum_{k \in S} (c'_k - p_k) d_{yt,k}. \tag{2}$$

Because probabilities remain in $[0, 1]$, its absolute change is at most $\sum_{k \in S} |d_{yt,k}|$. Let $|d_{yt}|_{(j)}$ denote these magnitudes in descending order. If at most B concepts can be edited, then

$$M_{yt} > \sum_{j=1}^{B} |d_{yt}|_{(j)} \quad \Longrightarrow \quad M'_{yt} > 0. \tag{3}$$

If this condition holds against every competitor, no admissible B-coordinate edit changes the winner. The bound follows directly from the triangle inequality; it is conservative because coherent class templates restrict which edits can occur together. Its converse need not hold. It also distinguishes annotation error from decision leverage: correcting the largest $|p_k - c_k|$ need not produce the largest class-margin change. Our correction curves retain their annotation-error ordering; this analytic test neither changes that protocol nor estimates how often the condition holds in the trained models.

Audit metrics. The first question is whether supplying correct concepts improves the original task. Corrective accuracy replaces all predicted probabilities with gold concepts c and measures classification accuracy (Koh et al., 2020):

$$\mathrm{Acc}\left(\arg\max_y \ell_y(x) \,\middle|\, p \leftarrow c\right).$$

This is the full-correction endpoint; later experiments also limit the number of corrected concepts.

Full correction may preserve the original class even when the model relies heavily on the image. A second test supplies concepts for a different class. For each test image i of class y_i, we sample one class $t_i \neq y_i$ uniformly and replace its probabilities with that class's template $c^{(t_i)}$; targets are resampled for each seed. We measure whether the resulting prediction $\hat{y}_i$ has a template closer in Hamming distance H to the target than to the source:

$$\frac{1}{n}\sum_i \mathbb{1}\left[H\left(c^{(\hat{y}_i)}, c^{(t_i)}\right) < H\left(c^{(\hat{y}_i)}, c^{(y_i)}\right)\right],$$

We call this fraction *responsiveness*; ties count as non-responsive. It measures alignment with the target after editing, rather than requiring the exact target class. It does not compare the prediction before and after the edit: an unchanged but initially wrong prediction can also satisfy the inequality. We therefore interpret it alongside gold-correction accuracy. All primary class templates are distinct, so every sampled target differs from its source.

The third question is how much classification ability remains without image-specific concept probabilities. We fix every probability to 0.5, keep the embeddings, and measure *concept-probability neutralized* (CPN) accuracy:

$$\mathcal{N}_{0.5} = \mathrm{Acc}\left(\arg\max_y \ell_y(x) \,\middle|\, p_k = 0.5\ \forall k\right).$$

At 0.5, each binary concept contributes an equal mixture of its two embeddings. For the linear head, the logits are $\sum_k 0.5\, a_{yk}(h) + b_y(h)$, so their image dependence comes from the retained embeddings. Chance accuracy is 0.005 on CUB and 0.020 on AwA2. High CPN together with weak performance on meaningful edits indicates reliance on information outside the editable probabilities. Because the all-0.5 vector need not occur in training, CPN is an ablation diagnostic; the correction and template-edit tests establish its relevance to the interface.

4 Experimental Design

Each primary comparison trains the same CEM with or without prediction distillation, then evaluates the same held-out images before and after concept edits. The split, architecture, concept supervision, RandInt training, and budget are fixed. The teacher-target controls then ask whether any loss of editability depends on pairing a particular image with its teacher prediction. Architecture extensions and incomplete-concept datasets test the reach of that explanation.

Datasets and concept regimes. CUB-200-2011 (Wah et al., 2011) and AwA2 (Xian et al., 2018) provide the primary bird and animal classification tasks. Their class templates make it possible to ask whether a classifier follows a coherent change in the supplied attributes. CelebA (Liu et al., 2015) and derm7pt (Kawahara et al., 2019) provide a contrasting regime with incomplete instance-level concepts. For CelebA, the three most balanced attributes define an eight-way label and the other 37 are concepts. For derm7pt, diagnoses are grouped into five classes and the seven criteria are encoded as 21 one-hot concepts. These two datasets test correction and neutralization, without a class-template responsiveness claim.

Class-template edits. CUB and AwA2 use class-level concept annotations. CUB has 200 distinct templates over 111 concepts, with minimum and mean pairwise Hamming distances 1 and 30.4; AwA2 has 50 distinct templates over 85 concepts, with distances 4 and 30.0. A nearest-template rule given the gold vector recovers the class with accuracy 1.000 on both datasets. This follows from the distinct class templates and verifies that the edited representation can specify a unique class; it is not an image-recognition result. No sampled source/target pair has an identical template (0% collisions), so none is excluded.

Students, teacher, and losses. The primary student has a linear label head and frozen image features. It learns both concept annotations and the class, while RandInt exposes it to corrected concepts during training. For example (x, c, y), an independent mask $m_k \sim \mathrm{Bernoulli}(0.25)$ chooses which of the K concept probabilities to replace:

$$\tilde{p}_k = m_k c_k + (1 - m_k) p_k.$$

The task loss uses this edited state, while concept supervision still compares the original predictions with the annotations:

$$\mathcal{L}_{\mathrm{base}} = \frac{\lambda_c}{K} \sum_k \mathrm{BCE}(p_k, c_k) + \lambda_y \mathrm{CE}(y, \ell(x \mid p \leftarrow \tilde{p})) .$$

The KD arm also matches the teacher's softened class distribution on that same edited student state. With teacher logits $q_T(x)$ for the original image and temperature τ, it adds

$$\mathcal{L}_{\mathrm{KD}} = \lambda_{\mathrm{KD}} \tau^2 \mathrm{KL}\Big(\mathrm{softmax}(q_T(x)/\tau) \,\Big\|\, \mathrm{softmax}(\ell(x \mid p \leftarrow \tilde{p})/\tau)\Big).$$

Both label losses therefore train the same randomly corrected prediction. The teacher target remains tied to the original image; it is not recomputed when the student's concept probabilities are substituted. The total loss is $\mathcal{L}_{\mathrm{base}} + \mathcal{L}_{\mathrm{KD}}$, without a separate RandInt loss.

KD-off sets $\lambda_{\mathrm{KD}} = 0$; the KD arm retains the same student settings and adds this loss. We use embedding width 16, $\lambda_c = 5$, unit task and distillation weights ($\lambda_y = \lambda_{\mathrm{KD}} = 1$ when KD is on), temperature $\tau = 3$, and RandInt probability 0.25. Concept and label heads train for 60 Adam epochs with learning rate 10^{-3}, weight decay 10^{-4}, and batch size 512. We use the final epoch in every arm, with no validation-based selection.

The teacher is a two-layer MLP with 512 hidden units, ReLU, and dropout 0.3, trained for 40 Adam epochs on the same frozen features and class labels. It reaches about 0.65 accuracy on CUB/ResNet-18, compared with the undistilled student's 0.59. It receives no concept annotations. Additional logistic and deep-MLP teachers test whether the result depends on this particular teacher.

Within-stratum teacher shuffle. To test the role of input-specific teacher information, we keep distillation but permute teacher targets among training images with identical gold concept signatures. These groups, or *strata*, number 200 on CUB and 50 on AwA2 because their class templates are distinct. The shuffle preserves the set of teacher distributions within each group while breaking which image receives which target; concept labels, class labels, loss weights, and the training schedule stay fixed.

If the image–target pairing matters, this change should reduce the loss of editability even though a distillation term remains. Shuffling also adds target noise, however, so recovery would not by itself distinguish the removal of pairing from that noise.

Concept-stratum mean control. The second control assigns every training image the mean teacher logit vector within its gold-concept stratum. It retains the stratum's average logits while removing image-specific variation without random reassignment. This averages logits before softmax, not output probabilities. The follow-up compares no distillation, input-paired targets, stratum-mean targets, and shuffled targets using the same student initialization and budget within each seed.

Matched comparisons. The main battery uses frozen ResNet-18, ResNet-50 (He et al., 2016), and ViT-B/16 (Dosovitskiy et al., 2021) features with five seeds per core comparison. Component controls use three seeds. The later CUB follow-up is a separate set of runs: five seeds for frozen linear and MLP heads, and three for end-to-end training. Its linear baseline and correction curves therefore have slightly different means from the earlier core battery.

Nonlinear, end-to-end, and local edits. To test whether the result depends on the linear head, the frozen-feature follow-up replaces it with an MLP of 256 hidden units and ReLU. To test whether

Table 1: CUB/ResNet-18, original five-seed comparison. Distillation improves task accuracy while weakening full gold correction and target-template responsiveness. Shuffling targets within concept strata largely restores the intervention outcomes. All entries are accuracies or fractions; $\pm$ gives the archived seed half-width (see text). Responsiveness is $0.998 \pm 0.002 / 0.109 \pm 0.005 / 0.986 \pm 0.003$ for KD-off/KD/shuffle. The other accuracy half-widths are below 0.02; CUB chance accuracy is 0.005.

Model	Task acc.	Concept acc.	Correct. acc.	Resp.	CPN acc.
KD-off	0.593	0.894	0.999	1.00	0.014 ± 0.016
KD	0.649	0.895	0.880	0.11	0.533 ± 0.033
Shuffle	0.614	0.890	0.991	0.99	0.004 ± 0.005

freezing features creates the problem, a separate experiment updates all parameters of an ImageNet-initialized ResNet-18 teacher and CEM student on the official CUB split. The teacher trains for 10 epochs and students for 15 with AdamW. Within each seed, conditions share initialization, augmentation, RandInt probability, and batch order.

For limited-budget correction, we start from predicted probabilities and replace the 1, 3, 5, 10, or 20 largest errors $|p_k - c_k|$ with gold values. For nearby-class edits, we instead start from the gold source template, choose the distinct class with smallest Hamming distance, and change coordinates toward its template, prioritizing those where the templates differ. We also replace the complete template. Sparse corrections are scored by task accuracy; nearby-class edits use the target-alignment criterion. These annotation-based simulations separate correcting a model error from asking it to follow a different class description.

5 Results

We first ask whether better prediction comes at a cost to concept edits. We then test target pairing, backbone choice, and the two architecture extensions. Correction curves show the cost at limited editing budgets, while incomplete-concept datasets locate the boundary.

Reading the results. Table 1 reports the original CUB/ResNet-18 comparison; Table 2 extends it across backbones and to AwA2. Table 3 and Figure 2 report the later follow-up. Accuracies and alignment scores are fractions, averaged across seeds; concept accuracy counts individual concept decisions. We retain the archived seed uncertainty half-widths: the original five-seed summaries use $2.571 s/\sqrt{5}$, and paired follow-up differences use $1.96 s_\Delta/\sqrt{n}$. Here s and s_Δ are sample standard deviations of seed values and paired differences. These are descriptive uncertainty summaries, not calibrated small-sample 95% intervals.

5.1 Accuracy Hides the Intervention Failure

On CUB/ResNet-18, distillation improves task accuracy from 0.593 to 0.649, while concept accuracy stays close (0.894 versus 0.895; Table 1). Yet supplying all correct concepts gives lower accuracy: 0.880 instead of 0.999. Target-template responsiveness falls from 1.00 to 0.11, and accuracy with every concept probability fixed to 0.5 rises from 0.014 to 0.533. The extra predictive performance thus accompanies weaker use of the editable values.

Shuffling teacher targets within concept strata largely reverses the intervention pattern: corrective accuracy is 0.991, responsiveness is 0.99, and CPN accuracy is 0.004. Its task accuracy is 0.614, between the undistilled and ordinarily distilled students. Thus the control recovers much of the editability while sacrificing part of the original accuracy gain; it does not make all metrics equal to the KD-off model.

5.2 The Pattern Appears Across Backbones

The same direction of change appears with other frozen features (Table 2). On CUB/ResNet-50, responsiveness falls from 1.00 to 0.11 and CPN rises from 0.016 to 0.501. ViT-B/16 already has

Table 2: Backbone and dataset comparisons (five-seed means). Paired cells give KD-off/KD; $\pm$ gives archived seed half-widths, with the convention stated in the text. Responsiveness half-widths are at most 0.01 on the CNN rows and 0.032 on ViT. The shuffle column reports CPN only; Table 1 gives its complete CUB/ResNet-18 profile. Chance CPN accuracy is 0.005 on CUB and 0.020 on AwA2.

Dataset	Backbone	Corrective acc. KD-off/KD	Responsiveness KD-off/KD	KD-off CPN acc.	KD CPN acc.	Shuffle CPN only
CUB	ResNet-18	$0.999/0.880$	$1.00/0.11$	0.014 ± 0.016	0.533 ± 0.033	0.004 ± 0.005
CUB	ResNet-50	$0.999/0.904$	$1.00/0.11$	0.016 ± 0.002	0.501 ± 0.031	0.004 ± 0.003
CUB	ViT-B/16	$0.990/0.872$	$0.53/0.05$	0.314 ± 0.026	0.689 ± 0.018	0.169 ± 0.025
AwA2	ResNet-18	$1.000/0.967$	$0.99/0.02$	0.074 ± 0.029	0.894 ± 0.007	0.001 ± 0.001

Table 3: CUB follow-up, separate from the original core battery. Entries are seed means: five for frozen features and three for end-to-end training. "Stratum" uses mean teacher logits within each gold-concept signature. Near-edit follow uses the complete Hamming-nearest distinct class template and the target-alignment criterion. Concept accuracy differs by at most 0.003 within each architecture block.

Training	Label head	Teacher target	Task acc.	Corrective acc.	CPN acc.	Near-edit follow
Frozen	Linear	None	0.593	0.999	0.014	0.936
		Input-paired	0.651	0.880	0.535	0.244
		Stratum	0.646	0.993	0.007	0.819
Frozen	MLP	None	0.556	1.000	0.007	0.963
		Input-paired	0.644	0.923	0.475	0.296
		Stratum	0.626	0.997	0.004	0.903
End-to-end	Linear	None	0.700	0.966	0.395	0.377
		Input-paired	0.732	0.957	0.472	0.280
		Stratum	0.731	0.970	0.390	0.322

weaker target alignment without distillation: responsiveness is 0.53 and CPN is 0.314. Distillation lowers the former to 0.05 and raises the latter to 0.689. AwA2 makes the distinction between correction and a change of class especially clear. Gold-correction accuracy remains 0.967 after distillation, but responsiveness to a different class template drops from 0.99 to 0.02.

The shuffle column reports only the neutralization diagnostic; it should be read with the full CUB/ResNet-18 profile in Table 1. On AwA2, shuffling also restores responsiveness to 1.000. On ViT, CPN remains 0.169, so this control does not eliminate all classification ability outside the editable probabilities.

5.3 Nonlinear and End-to-End Extensions

Does the failure require the exact linear structure in Equation 1? The MLP rows in Table 3 show a similar pattern without it. Distillation raises CPN from 0.007 to 0.475 and lowers alignment with the nearest class's full template from 0.963 to 0.296. The paired changes are 0.467 ± 0.035 and -0.667 ± 0.014 over five seeds. Mean-logit targets return the two diagnostics to 0.004 and 0.903, with task accuracy 0.626 versus 0.644 for ordinary distillation and 0.556 without it.

Allowing the representation to learn changes the starting point. The end-to-end model already has CPN accuracy 0.395 without distillation. Ordinary distillation raises it to 0.472 and lowers nearest-class alignment from 0.377 to 0.280; paired changes over three seeds are 0.077 ± 0.074 and -0.097 ± 0.022. Mean-logit targets return CPN to 0.390 but restore alignment only to 0.322. The editability loss therefore extends to this end-to-end setting, with a weaker contrast and partial recovery rather than the near-complete recovery of the frozen linear model.

The difference also appears at small editing budgets. Correcting the three largest concept errors gives linear-head accuracy 0.919 without distillation and 0.901 with mean-logit targets, compared with 0.737 under ordinary distillation. Starting from gold concepts and changing ten coordinates toward the nearest class gives target-alignment rates 0.752, 0.600, and 0.209, respectively; at twenty

coordinates they are 0.936, 0.818, and 0.244. The two simulations show a cost both when repairing predicted concepts and when modifying an already correct concept description.

5.4 Which Training Choices Matter?

The three-seed CUB/ResNet-18 component comparisons test alternatives to the embedding account. The full distilled CEM has corrective accuracy 0.879 and CPN 0.539. A distilled soft-CBM without image-dependent concept embeddings has CPN 0.005, although its corrective accuracy is only 0.915; changing the architecture does not by itself give perfect correction. Removing distillation from the CEM raises corrective accuracy to 0.999 and lowers CPN to 0.018. Removing RandInt instead worsens correction to 0.770 while CPN remains 0.466. Intervention-aware training helps, but its presence does not prevent the loss caused by adding distillation in this comparison.

The follow-up mean-logit control addresses the noise introduced by shuffling. It reaches corrective accuracy 0.993 and CPN 0.007, versus 0.880 and 0.535 under input-paired distillation; the follow-up shuffle reaches 0.990 and 0.005. Recovery therefore does not require randomly assigned soft targets. Across a logistic teacher and two MLP teachers, ordinary distillation gives CPN 0.447–0.539 and responsiveness 0.108–0.157. Together these comparisons support the role of input-specific teacher variation stored in the retained embeddings, without making it the only possible cause of poor editability.

5.5 Neutralization Exposes Retained Decision Power

Additional diagnostics check whether CPN tracks the intervention problem. Injecting a label-coded feature as a positive control raises CPN to 0.468. Across twenty settings of distillation strength and embedding width, CPN is negatively correlated with corrective accuracy (Pearson -0.89). A soft-versus-hard concept comparison misses this pattern: its accuracy gap is 0.112 without KD and 0.000 with KD, while CPN changes from 0.018 to 0.539. These checks explain why neutralization is useful alongside meaningful concept edits: rounding the probabilities alone can leave the predictive embeddings intact.

5.6 Correction Curves Show the Downstream Cost

Figure 2 shows how the difference accumulates as more predicted concepts are corrected. The order uses the known annotation error $|p_k - c_k|$, so this is an oracle correction curve: it measures the model's response once the concepts to repair are known.

Without distillation, five corrections raise accuracy from 0.593 to 0.968, and full correction reaches 0.999. With input-paired distillation, the same budgets yield 0.773 and 0.880. Mean-logit targets recover much of this benefit, reaching 0.948 and 0.993. A practitioner choosing only by the unedited accuracy would prefer the ordinarily distilled model, yet obtain less benefit from each of these correction budgets.

5.7 When Retained Image Information Is Useful

The CUB and AwA2 templates can specify a class on their own. CelebA and derm7pt have incomplete instance-level concepts, so their classifiers can reasonably use additional image information. Their undistilled CPN accuracies are already 0.164 and 0.578, and distillation increases them by 0.050 and 0.059. Gold-correction accuracy changes from 0.450 to 0.463 on CelebA and from 0.758 to 0.748 on derm7pt. These small, differently signed changes contrast with the CUB loss; they also show why high CPN alone is insufficient to declare an intervention failure.

6 Discussion

What the comparisons establish. The main result is a mismatch between two forms of usefulness: better predictions before intervention and worse outcomes after changing concepts. The linear identity explains how information can remain tied to the image; the teacher-target controls test how training makes use of that capacity. Their agreement across the frozen linear and nonlinear heads

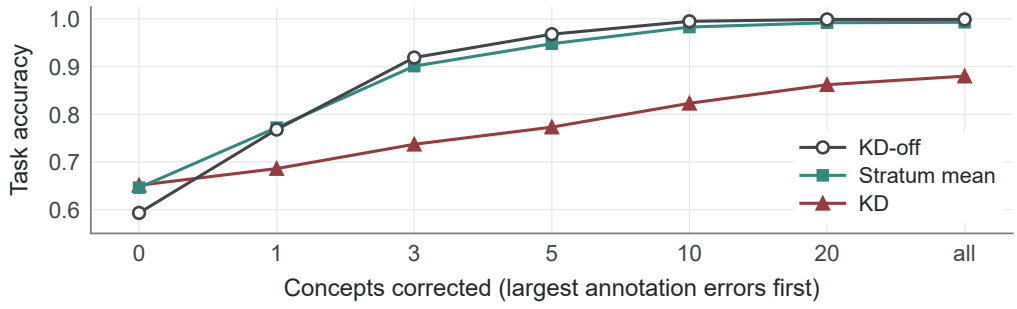


Figure 2: Oracle correction curves on CUB with frozen ResNet-18 features (five-seed follow-up). Equally spaced ticks denote evaluated budgets, including full correction. Stratum-mean targets recover much of the correction benefit lost under input-paired distillation.

supports the proposed account. The high undistilled CPN and partial recovery under end-to-end training show that distillation is only one contributor to reliance on the retained embeddings.

Operational implication. An audit of a distilled concept model should include the corrections its users expect to make. Ordinary task and concept accuracy can miss the change. CPN adds a simple diagnostic of what the editable probabilities leave behind, while correction curves and coherent template edits reveal whether that information matters for the intended interface. A mean-logit target is a useful comparison when gold concept strata are available; its accuracy/editability trade-off should be measured rather than assumed.

What neutralization leaves measurable. For the linear head, two images with identical edited probabilities can still have different margins through both their offsets and their concept slopes. Consequently, a high CPN score cannot attribute the residual prediction to $b_y(h)$ alone: at the neutral state, the slopes contribute $\frac{1}{2}\sum_k a_{yk}(h)$ as well. Separating these terms would require a different intervention. The present combination of neutralization and coherent edits evaluates the whole retained-embedding route, matching the interface available to a user who can change probabilities but not the embeddings.

7 Limitations

The nonlinear and end-to-end extensions use CUB with ResNet-18, and the end-to-end comparison has only three seeds. Their effect sizes therefore do not establish those of other backbones or training regimes. The class-template interventions and oracle correction ordering use annotations rather than observed human edits. CPN uses binary probability 0.5; other neutral values could change its scale. Finally, concept-stratum targets require gold annotations and may discard useful teacher distinctions when concepts are incomplete. A training objective that retains those distinctions while supporting the intended corrections remains open.

8 Conclusion

Prediction distillation can improve a concept-embedding model's accuracy while reducing the benefit of concept edits. The experiments connect this trade-off to image-specific teacher targets and embeddings that an edit leaves fixed. Averaging targets within gold-concept strata largely restores editability with frozen features and partly restores it under end-to-end training. Evaluating the edited prediction, alongside ordinary accuracy, makes this distinction visible.

Statement on the Use of Large Language Models

Generative AI tools assisted with manuscript refinement, language clarity and readability, and LaTeX formatting. The authors are responsible for the final manuscript.

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 (NeurIPS)*, 2018.

Lucas Beyer, Xiaohua Zhai, Amelie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Kushal Chauhan, Rishabh Tiwari, Jan Freyberg, Pradeep Shenoy, and Krishnamurthy Dvijotham. Interactive concept bottleneck models. In *AAAI Conference on Artificial Intelligence*, 2023.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lio, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born-again neural networks. In *International Conference on Machine Learning (ICML)*, 2018.

Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 2019.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning (ICML)*, 2018.

Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models. In *International Conference on Machine Learning (ICML)*, 2023.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning (ICML)*, pp. 5338–5348, 2020.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and pitfalls of black-box concept learning models. In *ICML Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI*, 2021.

Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? In *ICLR Workshop on Responsible AI*, 2021.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *International Conference on Learning Representations (ICLR)*, 2023.

Enrico Parisini, Tapabrata Chakraborti, Chris Harbron, Ben D. MacArthur, and Christopher R.S. Banerji. Leakage and interpretability in concept-based models. *Journal of Machine Learning Research*, 27(174):1–40, 2026. URL http://jmlr.org/papers/v27/25-1121.html.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations (ICLR)*, 2015.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A closer look at the intervention procedure of concept bottleneck models. In *International Conference on Machine Learning (ICML)*, 2023.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A. Alemi, and Andrew Gordon Wilson. Does knowledge distillation really work? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations (ICLR)*, 2020.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *International Conference on Learning Representations (ICLR)*, 2023.