FISHERCBM: CONCEPT BOTTLENECKS WITH CLOSED-FORM CONCEPT TRANSPORT AND FISHER-BOUNDED RESIDUAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept bottleneck models route a prediction through a layer of annotated concepts, but in practice the layer transmits information the concepts do not encode, and nothing determines which part of the input a concept is read from. Both problems arise because each concept is predicted from the entire input. We present FisherCBM, in which each concept is computed from an explicit set of input tokens. A concept transports the token features of the input, such as the patch tokens of a Vision Transformer, toward each of two learnable prototypes through an unbalanced optimal transport problem with a closed-form solution: the plan toward each state is the concept's saliency map and its mass is the evidence for each state. The head receives a one-bit state and a residual gated by an evidential belief and bounded by a fixed constant. The residual's sensitivity to a token is proportional to that token's share of the plan, its token Jacobian is explicit, and the resulting Fisher penalty on off-support tokens costs a forward pass. The bounded residual gives every intervention a guaranteed direction and a lower bound on its magnitude, and reduces whether the prediction follows to a sign condition on the head. On CUB and AwA2, FisherCBM matches the concept embedding model (CEM) in task accuracy with one fifth of its trainable parameters and rises monotonically to 100% accuracy under correct interventions, while wrong interventions drive it to chance as they do a hard CBM; it attains the highest concept accuracy on CUB and CelebA, and on CelebA, where two label attributes are not concepts, the bounded residual passes only part of the missing information.

1 INTRODUCTION

Concept bottleneck models (CBMs) (Koh et al., 2020) make a predictor interpretable by routing its output through a layer of human-defined concepts. An encoder predicts a vector of annotated concepts for an image of a bird, with attributes such as *bill shape: hooked* or *wing color: blue*, and a linear head predicts the target from these concepts alone. This design offers two guarantees that a black-box predictor does not: the prediction can be explained in terms of the concepts, and a user can intervene on a mispredicted concept and observe the corrected prediction.

Both guarantees rest on the assumption that the concept bottleneck of standard CBMs carries only the concepts. Two facts suggest that it does not. First, when the concept layer passes probabilities rather than binary states, it encodes label information beyond the concepts, and the head learns to exploit it (Mahinpei et al., 2021; Margeloiu et al., 2021). This concept leakage undermines both guarantees at once: the explanation omits what the head reads from the leak, and an intervention replaces the concept but not the leak. Binarizing the concept layer removes the leak at a substantial cost in accuracy, which is why concept embedding models (CEMs) (Zarlenga et al., 2022) and side-channel models (Havasi et al., 2022) widen the concept bottleneck instead, reintroducing the leak. Second, a CBM offers no account of how a concept is predicted from the input. In vision, concept predictors respond to regions unrelated to the concept (Margeloiu et al., 2021), and change under perturbations far from it (Raman et al., 2025). We argue that these two failures are one: a concept predicted from the entire input has no region to be localized to, and whatever it draws from outside the concept's own region is exactly the surplus that leaks. Existing remedies address the symptom rather than the cause. CEMs accept the wider channel and recover accuracy, and

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.



(a) Input (b) *belly color: white*, present (c) *belly color: buff*, absent

Figure 1: Transport plans on a CUB test image. Each panel shows the allocation $\pi^{c_i^{\text{hard}}}_{\cdot,i}$ of the predicted state, upsampled by Eq. (14); dimmed pixels are off-support, the cross marks the argmax token, and circles mark annotated part keypoints. For a present concept (b) and an absent one of the same family (c), the plan concentrates on the underparts; in (c), a small part of the mass falls on the tail and on a background patch at the right edge.

information-bottleneck variants (Lockhart et al., 2022; Galliamov et al., 2026) estimate how much the channel carries and penalize the estimate. Neither constrains where the concept prediction comes from, and neither guarantees what an intervention does to the prediction.

To address this, we propose FisherCBM. Each concept has two learnable prototypes, one per state, and obtains its feature by transporting the encoder's tokens toward them through an unbalanced optimal transport problem whose plan has a closed form (Theorem 1): a softmax over tokens that serves as the concept's saliency map, scaled by a mass that measures the evidence. The head receives the identity of the selected prototype, which carries at most one bit (Theorem 2), plus a residual gated by an evidential belief (Sensoy et al., 2018) and bounded by a constant $\eta < 1/2$; any information beyond one bit per concept must pass through this residual, whose Fisher information on off-support tokens we penalize at the cost of a forward pass (Proposition 2, Theorem 3). Because the residual is bounded, an intervention moves the coordinate in the corrective direction by at least $1-2\eta$ (Theorem 4), and whether the prediction follows reduces to a sign condition on the head (Theorem 5) toward which RandInt training moves it (Proposition 4). Our contributions are as follows:

- Independent unbalanced concept transport, a per-concept optimal transport with a closed-form plan that serves as the concept's saliency map.
- An evidential dual-channel bottleneck: a Beta readout from the two transports, an identity channel of at most one bit per concept, and a bounded, belief-gated residual.
- An off-support Fisher penalty with an explicit token Jacobian, evaluated in the forward pass, that defines and measures the residual's token-level confinement and, weighted to its scale, tightens it at no cost in accuracy.
- Intervention guarantees: a lower bound of $1 - 2\eta$ on every corrective change and a sign condition on the head under which corrections never lower accuracy.

2 Related Work

Concept Bottleneck Models and Concept Leakage The CBM of Koh et al. (2020) passes predicted concept probabilities to the label predictor. Mahinpei et al. (2021) show that this soft bottleneck carries label information beyond the concepts, and Margeloiu et al. (2021) show that the concept predictors respond to regions unrelated to the concept. Hard, binarized bottlenecks stop the leak at a cost in accuracy. Havasi et al. (2022) recover part of it with a side channel, GlanceNets (Marconato et al., 2022) use disentanglement and open-set rejection, and concept embedding models (Zarlenga et al., 2022) widen the bottleneck to an embedding per concept. Later work adds a residual regularized to be independent of the concepts (Zabounidis et al., 2023; Sheth & Kahou, 2023), bounds the mutual information between the two (Lockhart et al., 2022), or places an information bottleneck on the concept layer, penalizing a variational estimate of $I(X;C)$ (Galliamov

et al., 2026). These remedies either close the channel or estimate how much passes through it and penalize the estimate. FisherCBM keeps the channel open, fixes the capacity of its identity channel and the amplitude of its residual by construction, and penalizes its sensitivity to tokens outside the concept's own region. Intervention effects depend on how interventions are realized and trained for (Shin et al., 2022); we train with RandInt rather than a learned policy (Zarlenga et al., 2023) and intervene on the hard bit.

Input-Gradient and Fisher Penalties Fisher information measures how well an output distinguishes nearby inputs (Achille & Soatto, 2018; Hannun et al., 2021). For a Gaussian location family, it is the squared input gradient, so penalizing it is an input-gradient penalty in the line of double backpropagation (Drucker & Le Cun, 1992) and its uses for robustness and interpretability (Ross & Doshi-Velez, 2018; Hoffman et al., 2019). Closest to us, Ross et al. (2017) penalize gradients on annotated irrelevant regions. Our penalty reads the irrelevant region off the concept's own transport plan, and its token Jacobian is explicit, so it needs neither annotation nor a double backward pass.

Explaining Concept Predictions Raman et al. (2025) show that CBM concept predictions change under perturbations far outside the concept's region, and post-hoc attribution is an unreliable instrument for diagnosing this (Adebayo et al., 2018). Models in which the explanation is part of the computation include part-prototype networks (Chen et al., 2019; Nauta et al., 2021; 2023) and concept transformers (Rigotti et al., 2022), whose patch-to-concept attention is the explanation. Optimal transport (OT) offers an alternative to attention for assigning tokens to references (Cuturi, 2013; Mialon et al., 2021; Sander et al., 2022), at the cost of Sinkhorn iterations and a balanced marginal under which references compete for tokens. Unbalanced OT (Frogner et al., 2015; Chizat et al., 2018; Liero et al., 2018; Séjourné et al., 2019) relaxes the marginals to KL penalties; we relax both and give each concept its own copy of the token measure, which makes the plan closed-form and decoupled. The plan is both the explanation and the support outside which the residual is penalized.

3 Methodology

3.1 Background and Design Principles

Each input $\mathbf{x} \in \mathcal{X}$ is annotated with a label $y \in \mathcal{Y}$ and with K binary concepts $\mathbf{c} \in \{0,1\}^K$. The concepts are not mutually exclusive, and a single image region may provide evidence for several of them. An encoder Φ maps $\mathbf{x}$ to $N = hw$ token features $Z = \{\mathbf{z}_n\}_{n=1}^N \subset \mathbb{R}^D$, one per patch. We apply a parameter-free LayerNorm to every token, so $B_z := \|\mathbf{z}_n\|_2 = \sqrt{D}$. We employ a ViT (Dosovitskiy et al., 2021), whose tokens carry a fixed spatial index, but after self-attention, mix information from other patches; statements about tokens are therefore statements about the encoder's representation, not about disjoint image regions (Section 3.4 and Appendix A.1.8 make the gap precise). A CBM computes $\mathbf{r} \in \mathbb{R}^K$ from Z and predicts $\hat{y} = f(\mathbf{r})$ with a linear head f. An intervention on concept i replaces r_i with a value determined by the ground truth c_i.

The interpretability of a CBM rests on the head using r_i solely as an indicator of concept i. A scalar or embedding-valued r_i (Koh et al., 2020; Zarlenga et al., 2022) has far more capacity than the one-bit annotation, and the excess, computed from the entire image, is what leaks and what cannot be localized; binarizing r_i removes it but discards the fine-grained evidence on which accuracy depends. FisherCBM keeps a continuous channel but changes what computes it and what it may contain: the feature is a transport of the tokens toward the concept, so the plan records where the evidence came from; r_i is one bit plus a residual bounded by $\eta < 1/2$; and the residual is penalized for depending on tokens outside the plan's support. Proofs are given in Appendix A.1.

3.2 Independent Unbalanced Concept Transport

Each concept i has two learnable prototypes $\mathbf{p}_i^1, \mathbf{p}_i^0 \in \mathbb{R}^D$ in token space, one per state, normalized to $\|\mathbf{p}_i^c\| = \sqrt{D}$. The feature for concept i is an aggregate of the tokens under an assignment to that concept, and the assignment is the solution of an OT problem rather than the output of an attention module. An attention map has parameters of its own and its weights sum to one whether or not the concept is present, so it cannot express how much evidence there is; IUCT adds a mass that does.

Problem Formulation. Each concept transports the tokens twice, once toward each of its states. Transporting token n to state c of concept i costs the relative squared distance $\|\mathbf{z}_n - \mathbf{p}_i^c\|^2 - \|\mathbf{z}_n - \mathbf{p}_i^{1-c}\|^2$, whose offset $\|\mathbf{p}_i^c\|^2 - \|\mathbf{p}_i^{1-c}\|^2$ vanishes under the prototype normalization,

$$C_{n,i}^c = -2\,\langle \mathbf{z}_n, \Delta\mathbf{p}_i^c \rangle, \qquad \Delta\mathbf{p}_i^c := \mathbf{p}_i^c - \mathbf{p}_i^{1-c}, \qquad C_{n,i}^0 = -C_{n,i}^1 \qquad |C_{n,i}^c| \le 4D. \tag{1}$$

For simplicity, we will drop the state index when a statement holds for either transport. For state c, concept i has a transport vector $T_{\cdot,i}^\star \in \mathbb{R}_+^N$ and, for $\epsilon, \tau > 0$ and a reference mass $\rho = 1$, solves

$$\min_{T_{\cdot,i} \ge 0} \mathcal{J}_i(T_{\cdot,i}) = \min_{T_{\cdot,i} \ge 0} \sum_n T_{n,i} C_{n,i} + \epsilon\, D_{\mathrm{KL}}\big(T_{\cdot,i} \,\big\|\, \tfrac{1}{N}\mathbf{1}\big) + \tau\, D_{\mathrm{KL}}\Big(\sum_n T_{n,i} \,\Big\|\, \rho\Big), \tag{2}$$

where $D_{\mathrm{KL}}(\boldsymbol{x}\|\boldsymbol{y}) = \sum_j [x_j \log(x_j/y_j) - x_j + y_j]$ denotes the generalized KL divergence, applied also to scalars. We refer to this problem as independent unbalanced concept transport (IUCT). It is *independent* in that the K concepts solve K separate problems of the form Eq. (2), each on its own copy of the uniform token measure, with no constraint coupling the $T_{\cdot,i}^\star$ across concepts, as formalized in Definition 1. It is *unbalanced* in the sense of unbalanced optimal transport (Chizat et al., 2018; Liero et al., 2018): neither the token marginal nor the total mass is fixed; both are penalized by KL divergences, and the mass $\sum_n T_{n,i}$ is consequently determined by the data.

Definition 1. *Concept-wise independence: the mechanism imposes no competition between the visual supports of distinct concepts; it says nothing about the statistical dependence of the labels.*

Closed-Form Solution. Theorem 1 establishes that Eq. (2) is well-posed and its solution is closed-form: with both marginals relaxed, the stationarity conditions reduce to a Gibbs factor per token and a single scalar equation for the mass, so IUCT needs no inner loop. The closed form exhibits the factorization on which the remainder of the method relies: a direction $\pi_{\cdot,i}$, a softmax depending only on ϵ that serves as the saliency map, and a mass $m_i^\star$ depending on τ, ρ and the aggregate match Q_i. Lemma 1 is the formal counterpart of Definition 1. The share of a token allocated to concept i does not depend on how much of that token any other concept receives, so the concepts do not compete for tokens and the plan of each concept can be read as an explanation of that concept alone. The two transports of one concept are likewise separate problems. Since $C^0 = -C^1$, the second transport costs one additional softmax and no parameters.

Theorem 1. *For finite costs and $\epsilon, \tau, \rho > 0$, Eq. (2) has a unique minimizer with all entries positive. With $q_{n,i} = \frac{1}{N} e^{-C_{n,i}/\epsilon}$ and $Q_i = \sum_n q_{n,i}$,*

$$T_{n,i}^\star = m_i^\star\, \pi_{n,i}, \qquad m_i^\star = \rho^{\tau/(\epsilon+\tau)}\, Q_i^{\epsilon/(\epsilon+\tau)}, \qquad \pi_{n,i} = \frac{q_{n,i}}{Q_i} = \frac{T_{n,i}^\star}{m_i^\star}. \tag{3}$$

Lemma 1. *For $i \neq j$, $\partial T_{\cdot,i}/\partial C_{\cdot,j} = 0$.*

Concept feature. We aggregate the tokens with the unnormalized plan:

$$\mathbf{v}_i^c = \sum_n T_{n,i}^c\, \mathbf{z}_n = m_i^c\, \bar{\mathbf{v}}_i^c, \qquad \bar{\mathbf{v}}_i^c := \sum_n \pi_{n,i}^c \mathbf{z}_n, \tag{4}$$

where T^c, m_i^c and π^c are the quantities of Theorem 1 for the cost C^c. The direction $\bar{\mathbf{v}}_i^c$ records which tokens support state c, and the mass m_i^c records how many. With $\bar{\mathbf{v}}_i^c$ alone, τ and ρ would not appear, and IUCT would reduce to a per-concept softmax (ablated in Appendix C.2).

3.3 Evidential Dual-Channel Bottleneck

Concept prediction from prototype distances The feature transported toward state c is compared with the two prototypes. Its margin in favour of c is $\delta_i^c := \|\mathbf{v}_i^c - \mathbf{p}_i^{1-c}\|^2 - \|\mathbf{v}_i^c - \mathbf{p}_i^c\|^2$. Under prototype normalization, the prototype norms cancel as in Eq. (1):

$$\delta_i^c = 2\langle \mathbf{v}_i^c, \Delta\mathbf{p}_i^c \rangle = 2 m_i^c \langle \bar{\mathbf{v}}_i^c, \Delta\mathbf{p}_i^c \rangle. \tag{5}$$

The margin is the alignment of the direction transported toward a state, scaled by the mass transported toward it. The direction alone cannot tell an input in which one token matches $\mathbf{p}_i^c$ from one in which many do, but the mass can. The simplest readout is a logistic function of the difference; for reasons given below, we follow evidential deep learning (EDL) (Sensoy et al., 2018) and place a

$\mathrm{Beta}(\alpha_i^1, \alpha_i^0)$ distribution on the concept probability, with $\alpha_i^c = E_i^c + 1$ and non-negative evidence E_i^c read from the prototype geometry instead of a dedicated head. With a fixed temperature $\theta > 0$,

$$E_i^c = \mathrm{softplus}\big(\tfrac{\delta_i^c}{\theta}\big), \quad S_i = \alpha_i^1 + \alpha_i^0, \quad b_i^c = \frac{E_i^c}{S_i}, \quad u_i = \frac{2}{S_i}, \quad b_i^1 + b_i^0 + u_i = 1, \quad \hat{p}_i = \frac{\alpha_i^1}{S_i}. \tag{6}$$

Each state's evidence comes from the transport aimed at it. Since $\mathrm{softplus} > 0$, the evidence is never exactly zero ($E_i^c = \log 2$ at zero margin), so the uniform prior is approached but not attained. The predicted probability $\hat{p}_i$ is the mean of the Beta, and b_i^c and u_i are the belief masses and the uncertainty mass of subjective logic. The predicted state is $c_i^{hard} = \mathbf{1}[b_i^1 \geq b_i^0]$, and we write $\mathbf{p}_i^{c_i^{hard}}$, $\mathbf{v}_i^{c_i^{hard}}$, $\pi_{\cdot,i}^{c_i^{hard}}$ for the prototype, feature, and plan of the predicted state. The norm $\|\mathbf{v}_i\| = m_i^\star \|\bar{\mathbf{v}}_i\|$ scales the margin, so the transport mass enters the prediction as the amount of evidence.

Proposition 1. *For any $\theta > 0$: (i) $b_i^1 > b_i^0 \iff \hat{p}_i > \frac{1}{2} \iff \delta_i^1 > \delta_i^0$, with equality throughout iff $\delta_i^1 = \delta_i^0$; (ii) u_i is strictly decreasing in each of δ_i^1 and δ_i^0.*

Proposition 1 (i) states that the predicted state is the one whose own transport gathers the larger margin, and $\hat{p}_i$, the belief masses, and c_i^{hard} are mutually consistent. By (ii), the uncertainty is large when both margins are small, which happens when few tokens resemble either prototype. Reading both evidences from a single transport is insufficient for the reason given in Remark 1.

Learning the Beta distribution With the digamma function ψ and evidence-removed parameters $\tilde{\alpha}_i^{c_i} = 1$, $\tilde{\alpha}_i^{1-c_i} = \alpha_i^{1-c_i}$, which reset the annotated state's parameter to its prior value and leave the other unchanged, the concept loss is the standard EDL loss (Sensoy et al., 2018),

$$\mathcal{L}_{\mathrm{c}} = \frac{1}{K} \sum_{i=1}^{K} \Big[\psi(S_i) - \psi(\alpha_i^{c_i}) + \lambda_{\mathrm{KL}}(t)\, D_{\mathrm{KL}}\big(\mathrm{Beta}(\tilde{\alpha}_i^1, \tilde{\alpha}_i^0) \,\|\, \mathrm{Beta}(1,1)\big) \Big]. \tag{7}$$

The first term is the expected negative log-likelihood of the annotation under the Beta and is minimized by accumulating evidence for the annotated state, while the second penalizes only evidence for the wrong state, with a weight $\lambda_{\mathrm{KL}}(t)$ annealed from zero. What differs from EDL is the gradient path. The evidence has no parameters of its own: it is a function of the margin δ_i^c, so the concept loss reaches the prototypes through $\Delta\mathbf{p}_i^c$, and the two transport plans and the encoder through $\mathbf{v}_i^c$.

Dual-Channel Mechanism. Concept i contributes the coordinate

$$r_i = (c_i^{\mathrm{hard}} - \frac{1}{2}) + \eta\, g_i\, s_i, \qquad s_i = \cos\angle(\mathbf{p}_i^{c_i^{\mathrm{hard}}}, \mathbf{v}_i^{c_i^{\mathrm{hard}}}), \qquad g_i = \big(b_i^{c_i^{\mathrm{hard}}}\big)^\gamma, \tag{8}$$

where $\eta \in (0, \frac{1}{2})$ and $\gamma > 0$ are fixed. The identity term takes the values $\pm\frac{1}{2}$, so the sign of r_i is the predicted state whenever $|\eta g_i s_i| < \frac{1}{2}$, and for a linear head, the centering is absorbed by the bias. The bit is propagated with a straight-through gradient $c_i^{\mathrm{hard}} + \hat{p}_i - sg[\hat{p}_i]$, so that the task loss also trains $\hat{p}_i$. The only information about the identity of the concept transmitted to the head is which prototype was selected. The score s_i is the cosine between the selected prototype and the feature transported toward it, so it records how closely the tokens allocated to the selected state resemble that state's prototype, independently of how many there are. Since $|s_i| \leq 1$ and $g_i \leq 1$, the residual satisfies $|\eta g_i s_i| \leq \eta < \frac{1}{2}$ independently of D, of m_i^c and of the learned parameters. The cosine separates the roles of the two transport quantities: the mass enters through the margin and determines belief, gate, and state; the residual depends only on the direction $\bar{\mathbf{v}}_i^{c_i^{\mathrm{hard}}}$. The gate should pass the residual in proportion to the model's confidence in the selected state, and a probability cannot do this: the logistic readout gives $\hat{p}_i = 1/2$ both for an uninformative input, where no token resembles either prototype and the residual is noise, and for a conflicting input, where the tokens carry strong evidence for both states, and the residual is concept information. The evidential readout tells them apart through u_i (Proposition 1 (ii)): both margins near zero make u_i large and close the gate; both margins large make u_i small and open it.

Theorem 2. *Let X_i be any quantity from which the selection for concept i is computed, and $P_i \in \{\mathbf{p}_i^1, \mathbf{p}_i^0\}$ the selected prototype. Then $I(X_i; P_i) \leq 1$ bit and $I(X_{1:K}; P_{1:K}) \leq K$ bits.*

Theorem 2 states in what sense the identity channel cannot leak. Whatever the encoder computes, the head can learn at most one bit per concept from this channel, which is the information content of

the annotation c_i itself. The theorem bounds the capacity of the channel, not its content: which bit is emitted is learned, and the concept loss is what ties it to the annotation. Any information beyond one bit per concept must travel via the residual, and that is where the remaining constraints are placed.

3.4 Off-Support Fisher Regularization

The gate bounds the amplitude of the residual but not its content. Since s_i depends on every token through $\pi_{\cdot,i}^{c_i^{\text{hard}}}$, the residual of a confidently predicted concept may still encode information about unrelated regions, in which case an intervention on c_i does not control r_i. Throughout this section $\pi_{\cdot,i}$, $\bar{\mathbf{v}}_i$ and $\Delta\mathbf{p}_i$ denote $\pi_{\cdot,i}^{c_i^{\text{hard}}}$, $\bar{\mathbf{v}}_i^{c_i^{\text{hard}}}$ and $\Delta\mathbf{p}_i^{c_i^{\text{hard}}}$ for the selected state, and an intervention changes c_i^{hard} and with it all three. The support of concept i is the set of tokens whose allocation exceeds the uniform share, and the remaining tokens are weighted by their deficit:

$$\mathcal{S}_i = \{n : \pi_{n,i} > 1/N\}, \qquad w_{n,i} = \big(1 - N\pi_{n,i}\big)_+ \in [0, 1). \tag{9}$$

The threshold is placed on π rather than on $T = m_i\pi$ because the mass is under the optimizer's control through the prototypes, and a threshold on T could be met for every token by inflating m_i.

Fisher Penalty. We measure the dependence of the residual on off-support tokens by the Fisher information of the auxiliary family $R_i \mid Z \sim \mathcal{N}(c_i^{\text{hard}} + \eta g_i s_i(Z), \sigma_r^2)$ with respect to Z, holding c_i^{hard}, $\mathbf{p}_i^c$ and g_i fixed; the Gaussian is an analytical device that gives the sensitivity a KL interpretation. Proposition 2 identifies the penalized quantity as a weighted sum of squared token gradients of the score and bounds how distinguishable the residual's distribution is under perturbations confined to off-support tokens. $\mathcal{F}_i^{\text{off}} = 0$ is a training target, not a structural identity, and is compatible with the strict positivity of T. The gate is held fixed: it is a second path from off-support tokens to r_i, but it is bounded by η and small for uncertain concepts, so penalizing its sensitivity would penalize confident predictions.

Proposition 2. *The w-weighted trace of the Fisher information of $R_i \mid Z$ with respect to the tokens is $\mathcal{F}_i^{\text{off}} = \frac{\eta^2 g_i^2}{\sigma_r^2} \sum_n w_{n,i} \|\nabla_{\mathbf{z}_n} s_i\|_2^2$, and for any ΔZ supported on $\{n : w_{n,i} > 0\}$,*

$$D_{\text{KL}}\big(p(R_i \mid Z) \,\big\|\, p(R_i \mid Z + \Delta Z)\big) \le \tfrac{1}{2}\, \mathcal{F}_i^{\text{off}} \sum_n w_{n,i}^{-1} \|\Delta Z_n\|^2 + o(\|\Delta Z\|^2). \tag{10}$$

If $\mathcal{F}_i^{\text{off}} = 0$, the residual is first-order invariant to perturbations of the off-support tokens.

Closed-Form Gradient. A gradient penalty is ordinarily paid for with a second round of differentiation for each of the K concepts. Since s_i is the cosine of a closed-form aggregate, its token gradient is explicit. Theorem 3 gives it in terms of quantities the forward pass already produces.

Theorem 3. *Fix concept i, hold $\mathbf{p}_i^1, \mathbf{p}_i^0, c_i^{\text{hard}}$ fixed, let $\bar{\mathbf{v}}_i \neq 0$, and write $\mathbf{p}_i := \mathbf{p}_i^{c_i^{\text{hard}}}$. With $\bar{\mathbf{a}}_i := \partial s_i / \partial \bar{\mathbf{v}}_i = \mathbf{p}_i / (\|\mathbf{p}_i\| \|\bar{\mathbf{v}}_i\|) - s_i \bar{\mathbf{v}}_i / \|\bar{\mathbf{v}}_i\|^2$ and $\xi_{n,i} := \langle \mathbf{z}_n, \bar{\mathbf{a}}_i \rangle$,*

$$\nabla_{\mathbf{z}_n} s_i = \pi_{n,i} \Big[\bar{\mathbf{a}}_i + \frac{2}{\epsilon}\, \xi_{n,i}\, \Delta\mathbf{p}_i \Big], \tag{11}$$

$$\|\nabla_{\mathbf{z}_n} s_i\|^2 = \pi_{n,i}^2 \Big[\frac{1 - s_i^2}{\|\bar{\mathbf{v}}_i\|^2} + \frac{4}{\epsilon}\, \xi_{n,i} \langle \bar{\mathbf{a}}_i, \Delta\mathbf{p}_i \rangle + \frac{4}{\epsilon^2}\, \xi_{n,i}^2 \|\Delta\mathbf{p}_i\|^2 \Big]. \tag{12}$$

The gradient depends on $(Z, \mathbf{p}_i^1, \mathbf{p}_i^0)$ only through $\pi_{\cdot,i}$, $\bar{\mathbf{v}}_i$ and s_i, and not on m_i, τ or ρ.

Besides making the penalty computable, Theorem 3 has a structural consequence. The term that arises from m_i is proportional to $\langle \bar{\mathbf{a}}_i, \bar{\mathbf{v}}_i \rangle$, which vanishes because the cosine is scale-invariant. Therefore, the mass does not enter the residual. Since the support is mass-invariant, increasing the mass cannot reduce the penalty. Averaging over concepts and absorbing η^2/σ_r^2 into the loss weight:

$$\mathcal{L}_{\text{fisher}}^{\text{off}} = \frac{1}{K} \sum_{i=1}^{K} \text{sg}[g_i]^2 \sum_{n=1}^{N} \text{sg}[w_{n,i}]\, \|\nabla_{\mathbf{z}_n} s_i\|^2, \tag{13}$$

which needs only the inner products $\xi_{n,i}$ beyond the quantities of Eq. (4), at cost $O(NKD)$ per input. The token weight $(1 - N\pi)_+ \pi^2$ vanishes at $\pi = 0$ and for $\pi \ge \frac{1}{N}$ and peaks at $\pi = \frac{2}{3N}$,

so the penalty acts on tokens allocated marginally below the uniform share, where leakage through the softmax weights is possible. The token weight is at most $\frac{4}{27N^2}$ (Appendix A.1.7), so the penalty is small for any plan. The gate and the penalty control the residual at two granularities. The gate is coarse: one scalar per concept bounds how much of the residual reaches the head. The penalty is fine: it acts token by token on where the residual is read from, a quantity the gate cannot see. By Theorem 3, the transport already scales this sensitivity by $\pi_{n,i}$, and on our datasets that structural prior suffices: off-support tokens carry only 2×10^{-4} to 10^{-3} of the residual's token sensitivity at $\beta = 0$, and $\beta = 1$ changes neither this share nor accuracy (Appendix C.2, Table 8). The penalty therefore serves here as the measure of fine-grained confinement rather than its cause; at $\beta = 1$, it does not bind even when the gate is removed, whereas at $\beta = 10^5$, the order of the bound $27N^2/4$, it lowers the residual's off-support share to 0.3-0.4 of its value on CUB and CelebA and halves the support at unchanged or nearly unchanged accuracy (0 and 0.3 points), and at $\beta = 10^6$ to 0.13 at a cost of 0.3–0.8 points (Table 8). The residual is thus Fisher-bounded in a structural sense—its off-support Fisher information is at most $4/(27N^2)$ times the bracket of Eq. (12) for any plan (Appendix A.1.7)—and the penalty is a control that binds once its weight matches that scale.

Token-level penalty and scalability. Fine-grained concepts occupy small image regions, which suggests a pixel-level penalty. Its cost would grow with the number of concepts, the input resolution, and the depth of the encoder, whereas that of Eq. (13) is independent of all three. The guarantee of Proposition 2 is stated in token space and Proposition 3 relates it to the pixels. Write $Z = \Phi(x)$, and let U denote bilinear upsampling from the $h \times w$ token grid to the pixel grid.

Proposition 3. *Let Φ be differentiable with $\|J_\Phi(x)\|_{\mathrm{op}} \le L_\Phi$. (i) The Fisher information of $R_i \mid x$ with respect to the pixels is the pullback of the token-level Fisher information of Proposition 2, $\mathcal{F}_x = J_\Phi(x)^\top \mathcal{F}_Z J_\Phi(x)$; in particular, $\|\nabla_x s_i\| \le L_\Phi \|\nabla_Z s_i\|$. (ii) U is linear with convex weights, preserves nonnegativity and the ordering of cells away from cell boundaries, and acts per concept.*

By Proposition 3 (i), lowering the token-level Fisher information lowers the pixel-level one up to the factor L_Φ, an existence hypothesis that is never computed. Since Eq. (13) acts only on the off-support tokens, one may ask for the finer statement that it bounds a pixel-level off-support penalty. This holds when each token depends only on its own patch and acquires a spill term when receptive fields overlap, as they do in a ViT. Both versions are given in Appendix A.1.8. We reshape $\pi_{\cdot,i}$ to the $h \times w$ grid Π_i and upsample it,

$$A_i^{\mathrm{px}} = U(\Pi_i), \qquad \widetilde{W}_i^{\mathrm{px}} = \mathbf{1}\big[U(\Pi_i) \le 1/N\big], \tag{14}$$

which yields the per-concept saliency map and the pixel-level off-support set. By Proposition 3 (ii), the argmax token maps to the argmax pixel region and the display mask coincides with the off-support tokens except at patch boundaries. Upsampling is not used in training.

3.5 Intervention Guarantees and Training

An intervention on concept i replaces the predicted state c_i^{hard} with the ground truth c_i, leaves (b_i^1, b_i^0, u_i) unchanged, and re-evaluates the prototype, feature, and the gate at c_i, since all are indexed by c_i^{hard} in Eq. (8). If the model was confidently wrong, the re-evaluated gate is small, so the intervention corrects the identity without asserting confidence in evidence the model did not find.

Theorem 4. *For a state c_i, write s_i^c, g_i^c for score and gate evaluated with prototype p_i^c and feature v_i^c and suppose $c_i^{\mathrm{hard}} \neq c_i$. Then $\Delta r_i := r_i^{\mathrm{int}} - r_i^{\mathrm{free}} = \Delta c + \eta\, \Delta_i$ with $\Delta c = c_i - c_i^{\mathrm{hard}} \in \{\pm 1\}$, $\Delta_i = g_i^{c_i} s_i^{c_i} - g_i^{c_i^{\mathrm{hard}}} s_i^{c_i^{\mathrm{hard}}}$ and $|\Delta_i| \le 2$. Hence, $\mathrm{sign}(\Delta r_i) = \mathrm{sign}(\Delta c)$ and $|\Delta r_i| \ge 1 - 2\eta$; if $\Delta c \cdot \Delta_i \ge 0$ then $|\Delta r_i| \ge 1$. Moreover, $g_i^{c_i} \le g_i^{c_i^{\mathrm{hard}}}$, and if $b_i^{c_i} \le \underline{b} < \bar{b} \le b_i^{c_i^{\mathrm{hard}}}$ then $g_i^{c_i} / g_i^{c_i^{\mathrm{hard}}} \le (\underline{b}/\bar{b})^\gamma$.*

Theorem 4 decomposes the change into the identity flip of unit magnitude and a residual change of magnitude at most $2\eta < 1$, using only $|s_i| \le 1$ and $g_i \le 1$; this is the motivation for the cosine score and the bounded gate. Its last statement says that the gate contracts on a wrong prediction: the more confidently wrong the prediction, the more of the residual the intervention removes, and γ sets how sharply. Whether a corrective change in r_i also improves the prediction depends on the head.

Table 1: Accuracy (Mean ± Std) on CUB, AwA2, and CelebA datasets.

	CUB		AwA2		CelebA	
Model	Task	Concept	Task	Concept	Task	Concept
End-to-end	82.1 ± 0.3	–	95.7 ± 0.1	–	38.6 ± 0.2	–
Hard CBM (Koh et al., 2020)	76.1 ± 0.2	97.2 ± 0.0	96.0 ± 0.1	99.3 ± 0.0	30.7 ± 0.1	88.2 ± 0.0
Soft CBM (Koh et al., 2020)	80.7 ± 0.1	96.1 ± 0.3	96.1 ± 0.0	99.2 ± 0.0	31.5 ± 0.5	88.0 ± 0.0
Logit CBM (Mahinpei et al., 2021)	80.4 ± 0.1	96.4 ± 0.0	96.0 ± 0.0	99.2 ± 0.0	38.4 ± 0.2	88.3 ± 0.2
Hybrid CBM (Mahinpei et al., 2021)	81.0 ± 0.4	96.6 ± 0.3	96.1 ± 0.0	99.2 ± 0.0	39.6 ± 0.2	88.6 ± 0.1
CEM (Zarlenga et al., 2022)	83.6 ± 0.2	97.6 ± 0.0	95.9 ± 0.1	99.2 ± 0.0	39.6 ± 0.4	88.7 ± 0.1
ProbCBM (Kim et al., 2023)	76.7 ± 0.5	88.4 ± 0.2	95.1 ± 0.1	93.8 ± 0.0	35.3 ± 0.1	84.8 ± 0.3
CIBM-B (Galliamov et al., 2026)	79.0 ± 0.4	94.7 ± 0.6	95.8 ± 0.1	99.1 ± 0.0	32.5 ± 0.3	88.6 ± 0.0
CIBM-E (Galliamov et al., 2026)	80.8 ± 0.4	96.5 ± 0.1	96.0 ± 0.0	99.3 ± 0.0	32.4 ± 0.4	88.5 ± 0.1
FisherCBM (Ours)	82.5 ± 0.2	98.0 ± 0.1	96.1 ± 0.1	99.2 ± 0.0	33.2 ± 0.2	89.1 ± 0.1

Theorem 5. *Let $M_{y,\ell}(\mathbf{r}) = f_y(\mathbf{r}) - f_\ell(\mathbf{r})$ denote the margin of the true class over class ℓ. Consider $\mathbf{r} \mapsto \mathbf{r} + \Delta r_j \mathbf{e}_j$. If along the segment joining the two points,*

$$\Delta r_j \, \partial_{r_j} M_{y,\ell} \;\geq\; 0 \qquad \text{for all } \ell \neq y, \tag{15}$$

then no true-class margin decreases, and accuracy cannot decrease along nested sets of correct interventions. For a linear head, the condition is $\Delta r_j (w_{y,j} - w_{\ell,j}) \geq 0$.

Theorem 5 separates the two ingredients of a successful intervention: Theorem 4 guarantees the corrective change in r_j, and Eq. (15) is the property of the head that turns it into a non-decreasing margin. For a linear head, Eq. (15) is a sign agreement between the correction and the learned weights that can be measured directly. The architecture does not enforce it, and the training procedure moves the head toward it.

Training. The loss is $\mathcal{L} = \mathcal{L}_{\mathrm{t}} + \lambda_c \mathcal{L}_{\mathrm{c}} + \beta \mathcal{L}^{\mathrm{off}}_{\mathrm{fisher}}$, with $\mathcal{L}_t$ the task loss. Following RandInt, we intervene on a random subset $\mathcal{I} \subseteq [K]$ at every training step, drawing the intervention rate per sample from $\mathcal{U}(0, 1)$, so that the head sees intervened bottlenecks of every size; Proposition 4 shows that each such step also moves a linear head toward Eq. (15).

Proposition 4. *Let a linear head be trained with cross-entropy on the corrected bottleneck $\mathbf{r} + \sum_{j \in \mathcal{I}} \Delta r_j \mathrm{e}_j$. For every $\ell \neq y$ and $j \in I$ whose free prediction was wrong, one gradient step changes $w_{y,j} - w_{\ell,j}$ by a nonnegative multiple of $\mathrm{sign}(\Delta r_j)$, and so increases $\Delta r_j (w_{y,j} - w_{\ell,j})$ whenever it is negative.*

4 Experiments

We evaluate FisherCBM on CUB-200-2011, AwA2 and the incomplete-concept CelebA task against an end-to-end classifier and eight concept-based baselines (Table 1), all on the same frozen ViT-B/16 backbone, over three seeds. Datasets, baselines, hyperparameters, and protocols are in Appendix B; results on AwA2 and CelebA beyond Table 1 and the ablation study are in Appendix C.

4.1 Task Performance

On CUB (Table 1), FisherCBM reaches 82.5% task accuracy, within 1.1 points of CEM and on par with the end-to-end classifier, with the highest concept accuracy of all models (98.0%). Hard CBM, the only baseline whose bottleneck also carries one bit per concept, is 6.4 points lower, and CIBM, which matches FisherCBM's parameter count (Table 2) and leakage level (Table 3), is 1.7–3.5 points lower. On AwA2, all models saturate near 96%. CelebA is the informative case: its label depends on eight attributes, of which six are concepts. Hard and Soft CBMs obtain 31%. CEM, Hybrid CBM, and Logit CBM reach the 39% of the end-to-end classifier, so their bottlenecks pass the two hidden attributes. FisherCBM obtains 33.2%, recovering a third of the gap through its bounded residual and giving up the rest by design. Removing the transport mass ($m_i \equiv 1$) costs 3.2 points on CUB and none on CelebA or AwA2 (Appendix C.2).

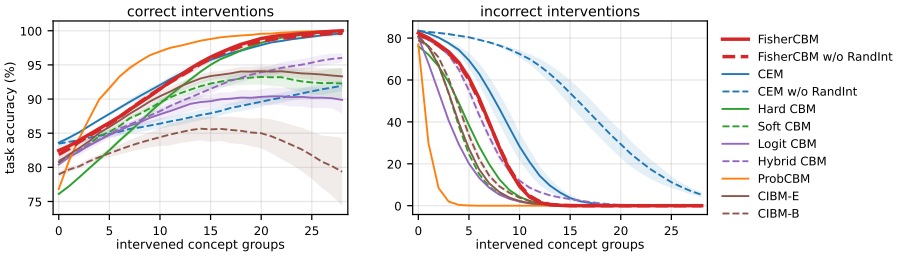


Figure 2: Test-time interventions on CUB (per-group protocol, mean and standard deviation over three seeds). Left: selected concept groups are set to their annotation. Right: they are set to the wrong value; a head that reads the concepts should lose accuracy, so lower is better.

4.2 Intervention

Figure 2 shows task accuracy on CUB against the number of intervened concept groups, with the groups set to their annotation (correct mode) or to the wrong value (incorrect mode, Havasi et al., 2022); Table 3 gives the summary statistics. Under correct interventions, FisherCBM rises monotonically from 82.5 to 100.0% with no drop at any step, consistent with Theorems 4 and 5 (the 100% endpoint follows from the class-level concept labels, Appendix B.1, and is shared by the hard CBM; the claim is the absence of drops), with an area of 93.9% equal to CEM's; the soft and logit CBMs plateau near 90%. Under incorrect interventions, it falls below 1% after 13 of 28 groups, with an area of 18.3%, against 24.0 for CEM, 18.2 for the hybrid CBM and 12.9 for the hard CBM; the soft, logit and CIBM bottlenecks lie lower still (8.8–12.5), because their clean accuracy is 1.7–3.5 points lower and their descent begins earlier, so the area is partly a measure of how much accuracy there was to lose. With half of the groups wrong, FisherCBM retains 0.1%, like the hard, soft, logit, and CIBM bottlenecks (0.1–0.4%), whereas the hybrid CBM retains 3.9% and CEM 6.3% (Table 3): the areas differ in where the descent begins, not in where it ends. The gap to the hard CBM is the price of the residual, while the gap to CEM is the information its embedding lets the head fall back on.

4.3 Interpretation

The concept-level reading of a CBM is unchanged (concept i contributes $w_{y,i}r_i$ to class y); the plan adds where in the input the concept was read from. Figure 1 shows a Downy Woodpecker on a feeder, classified correctly with all 112 concepts correct. For *belly color: white* (present), the plan toward $\mathbf{p}_i^1$ covers the white belly and flank with 22 of 196 tokens on the support; the argmax token lies two tokens from the annotated belly keypoint, closer than the median of 3.5 tokens over the test set (Appendix C.3); the feeder and the wire mesh are off-support. For *belly color: buff* (absent) the plan toward $\mathbf{p}_i^0$ covers the same underparts with 24 tokens: the evidence that the belly is not buff is read mainly from the underparts; one off-object token at the right edge remains on the support. The figure is typical: over 1,000 random CUB test images and all 112 concepts, the median support has 15 tokens and $91 \pm 5\%$ of the allocated mass (three seeds) lies inside the bird's bounding box, which covers 53% of the tokens; a uniform or random plan places 53% there and ProbCBM's per-concept attention 59%, with a median support of 93 tokens (Table 9). More examples are in Appendix C.3.

5 Conclusion

FisherCBM reads each concept through a closed-form transport toward two prototypes and passes the head one bit plus a bounded, belief-gated residual, so the plan shows where a concept was read from, and every correction moves its coordinate by a guaranteed amount. It matches CEM in task accuracy on CUB and AwA2 with one fifth of the parameters, has the highest concept accuracy on CUB and CelebA, and its interventions are monotone to 100% with or without RandInt, while wrong ones drive it to chance as for a hard CBM; on CelebA the bounded residual passes only part of the hidden information. Two limitations remain: the Fisher penalty binds only when weighted to its scale ($\beta \approx 10^5$, Table 8), and then sharpens plans without localizing better; and the plans locate objects, not parts (median 3.5 tokens from the annotated keypoint, Appendix C.3). Multi-valued concepts and non-image inputs are natural next steps.

AI USE STATEMENT

The method, its theoretical analysis and proofs, the implementation of FisherCBM and of all baselines, and all training runs and main experiments are the authors' own work, without generative AI assistance. We used a generative AI assistant (Claude, Anthropic) for two tasks that require disclosure: it provided feedback on the experimental design, on the basis of which the authors added the ablations and post-hoc diagnostics reported in the appendix, and it assisted in interpreting the outcomes of those ablations. Generating synthetic data, developing theoretical models, formulating or proving mathematical claims, implementing methods, translation, dataset cleaning, and qualitative analysis were not performed with generative AI. Additionally, we used the assistant to write scripts for the post-hoc evaluation of trained checkpoints and to edit the manuscript for length and readability. All AI-assisted work was reviewed by the authors: the evaluation scripts were checked and their accuracies verified against the metrics logged during training, and all text was revised and checked against the underlying results. We take responsibility for the final content of this work.

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–9, 2018. doi: 10.1109/ITA.2018.8503149.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 9525–9536, Red Hook, NY, USA, 2018. Curran Associates Inc.

Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: Deep learning for interpretable image recognition, 2019. URL https://arxiv.org/abs/1806.10574.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of computation*, 87(314):2563–2609, 2018.

Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 2292–2300, Red Hook, NY, USA, 2013. Curran Associates Inc.

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 *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

H. Drucker and Y. Le Cun. Improving generalization performance using double backpropagation. *IEEE Transactions on Neural Networks*, 3(6):991–997, 1992. doi: 10.1109/72.165600.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso Poggio. Learning with a wasserstein loss. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pp. 2053–2061, Cambridge, MA, USA, 2015. MIT Press.

Karim Galliamov, Syed Muhammad Ahsan Raza Kazmi, Adil Khan, and Adín Ramírez Rivera. Concepts'information bottleneck models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 82405–82430, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/8555cf308efef02adacb0f9e09bd5f10-Paper-Conference.pdf.

Awni Hannun, Chuan Guo, and Laurens van der Maaten. Measuring data leakage in machine-learning models with fisher information. In *Uncertainty in Artificial Intelligence*, pp. 760–770. PMLR, 2021.

Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Judy Hoffman, Daniel A. Roberts, and Sho Yaida. Robust learning with jacobian regularization, 2019. URL https://arxiv.org/abs/1908.02729.

Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211 (3):969–1117, 2018.

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

Joshua Lockhart, Nicolas Marchesotti, Daniele Magazzeni, and Manuela Veloso. Towards learning to explain with concept bottleneck models: mitigating information leakage. *ArXiv*, abs/2211.03656, 2022. URL https://api.semanticscholar.org/CorpusID:253384003.

Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and pitfalls of black-box concept learning models. *ArXiv*, abs/2106.13314, 2021. URL https://api.semanticscholar.org/CorpusID:235652059.

Emanuele Marconato, Andrea Passerini, and Stefano Teso. Glancenets: interpretable, leak-proof concept-based models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? *ArXiv*, abs/2105.04289, 2021. URL https://api.semanticscholar.org/CorpusID:234339919.

Grégoire Mialon, Dexiong Chen, Alexandre d'Aspremont, and Julien Mairal. A trainable optimal transport embedding for feature aggregation and its relationship to attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=ZK6vTvb84s.

Meike Nauta, Ron van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14928–14938, 2021. doi: 10.1109/CVPR46437.2021.01469.

Meike Nauta, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. Pip-net: Patch-based intuitive prototypes for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2744–2753, June 2023.

Naveen Janaki Raman, Mateo Espinosa Zarlenga, Juyeon Heo, and Mateja Jamnik. Do concept bottleneck models respect localities? *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=4mCkRbUXOf.

Mattia Rigotti, Christoph Miksovic, Ioana Giurgiu, Thomas Gschwind, and Paolo Scotton. Attention-based interpretability with concept transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=kAa9eDS0RdO.

Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32, 04 2018. doi: 10.1609/aaai.v32i1.11504.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pp. 2662–2670. AAAI Press, 2017. ISBN 9780999241103.

Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pp. 3515–3530. PMLR, 2022.

Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouv'e, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport. *ArXiv*, abs/1910.12958, 2019. URL https://api.semanticscholar.org/CorpusID:204949797.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 3183–3193, Red Hook, NY, USA, 2018. Curran Associates Inc.

Ivaxi Sheth and Samira Ebrahimi Kahou. Auxiliary losses for learning generalizable concept-based models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A closer look at the intervention procedure of concept bottleneck models. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*, 2022. URL https://openreview.net/forum?id=PUspzfGsgY.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 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*, 41(9):2251–2265, 2019. doi: 10.1109/TPAMI.2018.2857768.

Renos Zabounidis, Ini Oguntola, Konghao Zhao, Joseph Campbell, Simon Stepputtis, and Katia Sycara. Benchmarking and enhancing disentanglement in concept-residual models, 2023. URL https://arxiv.org/abs/2312.00192.

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 *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Mateo Espinosa Zarlenga, Katherine M. Collins, Krishnamurthy (Dj) Dvijotham, Adrian Weller, Zohreh Shams, and Mateja Jamnik. Learning to receive help: intervention-aware concept embedding models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

A Appendix

A.1 Proofs of Section 3

Throughout $\sigma(u) = 1/(1+e^{-u})$ denotes the logistic function, ψ the digamma function, and $\mathbf{a} = (a_1, \ldots, a_N)$ with $a_n = 1/N$ the uniform token measure.

A.1.1 PROOF OF THEOREM 1

Proof. Both KL terms in Eq. (2) use the generalized KL divergence:

$$D_{\mathrm{KL}}(\boldsymbol{x}\|\boldsymbol{y}) = \sum_j \Big[x_j \log \frac{x_j}{y_j} - x_j + y_j\Big], \qquad x_j \geq 0,\ y_j > 0, \tag{16}$$

with the convention $0 \log 0 = 0$. In the proof, we denote $\mathbf{t}_i = T_{\cdot,i}$. We record the two derivatives used below:

(i) Vector term $\epsilon\, D_{\mathrm{KL}}(\mathbf{t}_i\|\mathbf{a})$: Only the n-th summand depends on $T_{n,i}$, and $\frac{d}{dt}\big(t \log \frac{t}{a} - t + a\big) = \log \frac{t}{a} + t \cdot \frac{1}{t} - 1 = \log \frac{t}{a}$. Hence

$$\frac{\partial}{\partial T_{n,i}}\, \epsilon\, D_{\mathrm{KL}}(\mathbf{t}_i\|\mathbf{a}) = \epsilon \log \frac{T_{n,i}}{a_n}. \tag{17}$$

(ii) Scalar mass term $\tau\, D_{\mathrm{KL}}(m_i\|\rho)$: Here $m_i = \sum_{n'} T_{n',i}$ is a scalar, $D_{\mathrm{KL}}(m_i\|\rho) = m_i \log \frac{m_i}{\rho} - m_i + \rho$, and $\frac{d}{dm}\big(m \log \frac{m}{\rho} - m + \rho\big) = \log \frac{m}{\rho}$. Since $\partial m_i / \partial T_{n,i} = 1$ for every n, the chain rule gives

$$\frac{\partial}{\partial T_{n,i}}\, \tau\, D_{\mathrm{KL}}(m_i\|\rho) = \tau \log \frac{m_i}{\rho} \cdot \frac{\partial m_i}{\partial T_{n,i}} = \tau \log \frac{m_i}{\rho}. \tag{18}$$

Write $m_i = m_i(\mathbf{t}_i) = \sum_{n=1}^N T_{n,i}$ and recall

$$\mathcal{J}_i(\mathbf{t}_i) = \underbrace{\sum_n T_{n,i} C_{n,i}}_{\text{(linear)}} + \underbrace{\epsilon\, D_{\mathrm{KL}}(\mathbf{t}_i\|\mathbf{a})}_{\text{(entropy)}} + \underbrace{\tau\, D_{\mathrm{KL}}(m_i\|\rho)}_{\text{(mass)}}. \tag{19}$$

The argument proceeds in five steps. We first record the gradient (Step 0). We establish strict convexity (Step 1) so a minimizer, if it exists, is unique; establish coercivity (Step 2) so that a global minimizer exists. We show that the minimizer must be interior (Step 3) so that it satisfies the first-order condition. Only then do we solve that condition in closed form (Step 4). We do not treat the first-order equation as a solution before existence, uniqueness, and interiority are secured.

Step 0 (Gradient): Differentiating term by term with Eq. (17) and Eq. (18), and $\partial(\sum_{n'} T_{n',i} C_{n',i})/\partial T_{n,i} = C_{n,i}$,

$$\frac{\partial \mathcal{J}_i}{\partial T_{n,i}} = C_{n,i} + \epsilon \log \frac{T_{n,i}}{a_n} + \tau \log \frac{m_i}{\rho}, \qquad n = 1, \ldots, N. \tag{20}$$

Step 1 (Strict convexity on $\mathbb{R}^N_{>0}$): Differentiate Eq. (20) again. For the diagonal ($n = n'$),

$$\frac{\partial}{\partial T_{n,i}}\big[\epsilon \log(T_{n,i}/a_n)\big] = \frac{\epsilon}{T_{n,i}}, \tag{21}$$

$$\frac{\partial}{\partial T_{n,i}}\big[\tau \log(m_i/\rho)\big] = \tau \cdot \tfrac{1}{m_i} \cdot \tfrac{\partial m_i}{\partial T_{n,i}} = \frac{\tau}{m_i}; \tag{22}$$

for the off-diagonal ($n \neq n'$) the entropy term contributes 0 (it depends only on $T_{n,i}$), and the mass term contributes $\partial/\partial T_{n',i}\big[\tau \log(m_i/\rho)\big] = \tau/m_i$. Hence the Hessian is

$$\big(\nabla^2 \mathcal{J}_i\big)_{n,n'} = \epsilon\, \frac{\delta_{n,n'}}{T_{n,i}} + \frac{\tau}{m_i}, \qquad \text{i.e.} \qquad \nabla^2 \mathcal{J}_i = \epsilon\, \mathrm{diag}\Big(\tfrac{1}{T_{1,i}}, \ldots, \tfrac{1}{T_{N,i}}\Big) + \frac{\tau}{m_i} \mathbf{1}\mathbf{1}^\top, \tag{23}$$

with $\delta_{n,n'}$ the Kronecker delta and $\mathbf{1} = (1, \ldots, 1)^\top$. For any $\mathbf{x} \in \mathbb{R}^N$, $\mathbf{x} \neq 0$,

$$\mathbf{x}^\top \nabla^2 \mathcal{J}_i\, \mathbf{x} = \epsilon \sum_{n=1}^N \frac{x_n^2}{T_{n,i}} + \frac{\tau}{m_i} \big(\mathbf{1}^\top \mathbf{x}\big)^2 = \epsilon \sum_{n=1}^N \frac{x_n^2}{T_{n,i}} + \frac{\tau}{m_i} \Big(\sum_{n=1}^N x_n\Big)^2. \tag{24}$$

On $\mathbb{R}^N_{>0}$ we have $T_{n,i} > 0$ and $m_i > 0$. The first term is > 0 whenever $\boldsymbol{x} \neq \mathbf{0}$ (a sum of nonnegative terms with at least one $x_n^2/T_{n,i} > 0$) and the second is ≥ 0. Hence $\boldsymbol{x}^\top \nabla^2 \mathcal{J}_i \boldsymbol{x} > 0$, so $\nabla^2 \mathcal{J}_i \succ 0$ and $\mathcal{J}_i$ is strictly convex on the convex set $\mathbb{R}^N_{>0}$. A strictly convex function has at most one stationary point, which, if it exists, is the unique global minimizer over $\mathbb{R}^N_{>0}$.

Step 2 (Coercivity ⇒ A minimizer exists): We show $\mathcal{J}_i(\mathbf{t}_i) \to +\infty$ as $\|\mathbf{t}_i\| \to \infty$ within the closed orthant $\mathbb{R}^N_{\geq 0}$. Since $|T_{n,i}C_{n,i}| \leq T_{n,i} \max_n |C_{n,i}|$ and $T_{n,i} \geq 0$, the linear term is bounded by

$$\sum_n T_{n,i}C_{n,i} \geq -\Big(\max_n |C_{n,i}|\Big) \sum_n T_{n,i} = -m_i \max_n |C_{n,i}|, \tag{25}$$

which is linear in m_i. The entropy term $\epsilon D_{\mathrm{KL}}(\mathbf{t}_i \| \mathbf{a})$ grows like $\epsilon \sum_n T_{n,i} \log T_{n,i}$, that is, superlinearly in each coordinate, and the mass term is nonnegative. A superlinear positive growth dominates a linear negative term, so $\mathcal{J}_i \to +\infty$. Every sublevel set $\{\mathbf{t}_i \in \mathbb{R}^N_{\geq 0} : \mathcal{J}_i(\mathbf{t}_i) \leq c\}$ is therefore bounded. It is also closed, because $\mathcal{J}_i$ is continuous on $\mathbb{R}^N_{\geq 0}$ under the convention $0 \log 0 = 0$. Sublevel sets are compact, and a continuous function on a nonempty compact set attains its minimum, so a global minimizer $\mathbf{t}_i^\star \in \mathbb{R}^N_{\geq 0}$ exists.

Step 3 (the minimizer is interior): Suppose a minimizer has $T_{n,i} = 0$ for some n. Consider the one-sided derivative as $T_{n,i} \to 0^+$ with the remaining coordinates fixed. From Eq. (20), as $T_{n,i} \to 0^+$ we have $\log(T_{n,i}/a_n) \to -\infty$ while $C_{n,i}$ and $\tau \log(m_i/\rho)$ stay finite, so

$$\frac{\partial \mathcal{J}_i}{\partial T_{n,i}} = C_{n,i} + \epsilon \log \frac{T_{n,i}}{a_n} + \tau \log \frac{m_i}{\rho} \;\longrightarrow\; -\infty. \tag{26}$$

A strictly negative partial derivative means increasing $T_{n,i}$ slightly strictly decreases $\mathcal{J}_i$, so no point with a zero coordinate can be a minimizer. The global minimizer of Step 2 therefore lies in the open orthant $\mathbb{R}^N_{>0}$, where by Step 1, it is the unique stationary point, and it satisfies $\partial \mathcal{J}_i / \partial T_{n,i} = 0$ for every n.

Step 4 (solve the first-order condition for the closed form): Setting Eq. (20) to zero and solving for $T_{n,i}$:

$$\epsilon \log \frac{T_{n,i}}{a_n} = -C_{n,i} - \tau \log \frac{m_i}{\rho} \implies \log \frac{T_{n,i}}{a_n} = -\frac{C_{n,i}}{\epsilon} - \frac{\tau}{\epsilon} \log \frac{m_i}{\rho}, \tag{27}$$

and exponentiating both sides, with $q_{n,i} := a_n e^{-C_{n,i}/\epsilon}$ and $e^{-(\tau/\epsilon)\log u} = u^{-\tau/\epsilon}$,

$$T_{n,i} = a_n \exp\Big(-\frac{C_{n,i}}{\epsilon}\Big) \exp\Big(-\frac{\tau}{\epsilon} \log \frac{m_i}{\rho}\Big) = q_{n,i} \Big(\frac{m_i}{\rho}\Big)^{-\tau/\epsilon}, \tag{28}$$

Summing Eq. (28) over n and letting $Q_i := \sum_n q_{n,i}$, the factor $(m_i/\rho)^{-\tau/\epsilon}$ is independent of n and can be extracted:

$$m_i = \sum_n T_{n,i} = \Big(\sum_n q_{n,i}\Big) \Big(\frac{m_i}{\rho}\Big)^{-\tau/\epsilon} = Q_i \Big(\frac{m_i}{\rho}\Big)^{-\tau/\epsilon}. \tag{29}$$

Multiply both sides by $(m_i/\rho)^{\tau/\epsilon}$:

$$m_i \Big(\frac{m_i}{\rho}\Big)^{\tau/\epsilon} = Q_i \implies \frac{m_i^{1+\tau/\epsilon}}{\rho^{\,\tau/\epsilon}} = Q_i \implies m_i^{1+\tau/\epsilon} = Q_i\, \rho^{\,\tau/\epsilon}. \tag{30}$$

Raising to the power $\frac{1}{1+\tau/\epsilon} = \frac{\epsilon}{\epsilon+\tau}$,

$$m_i^\star = \Big(Q_i\, \rho^{\,\tau/\epsilon}\Big)^{\epsilon/(\epsilon+\tau)} = Q_i^{\epsilon/(\epsilon+\tau)}\, \rho^{\,(\tau/\epsilon)\cdot(\epsilon/(\epsilon+\tau))} = \rho^{\,\tau/(\epsilon+\tau)} Q_i^{\epsilon/(\epsilon+\tau)}, \tag{31}$$

which is the mass in Eq. (3). Back-substituting Eq. (31) into Eq. (28) and using $\frac{\tau}{\epsilon+\tau} - 1 = -\frac{\epsilon}{\epsilon+\tau}$ and $(\cdot)^{-\tau/\epsilon}$:

$$\begin{aligned}
\Big(\frac{m_i^\star}{\rho}\Big)^{-\tau/\epsilon} &= \Big(\rho^{\,\tau/(\epsilon+\tau)-1} Q_i^{\epsilon/(\epsilon+\tau)}\Big)^{-\tau/\epsilon} \\
&= \Big(\rho^{-\epsilon/(\epsilon+\tau)} Q_i^{\epsilon/(\epsilon+\tau)}\Big)^{-\tau/\epsilon} \\
&= \rho^{\,\tau/(\epsilon+\tau)} Q_i^{-\tau/(\epsilon+\tau)},
\end{aligned} \tag{32}$$

Therefore

$$T^{\star}_{n,i} = q_{n,i}\Big(\frac{\rho}{Q_i}\Big)^{\tau/(\epsilon+\tau)} = \rho^{\,\tau/(\epsilon+\tau)} Q_i^{\epsilon/(\epsilon+\tau)} \cdot \frac{q_{n,i}}{Q_i} = m_i^{\star}\,\pi_{n,i}, \tag{33}$$

$$\pi_{n,i} = \frac{q_{n,i}}{Q_i}, \tag{34}$$

which is Eq. (3). Since the minimizer exists (Step 2), is unique (Step 1), and is the interior stationary point (Step 3), the closed form is the unique global minimizer. It is a purely algebraic expression in $C_{\cdot,i}, \mathbf{a}, \rho, \epsilon, \tau$ and no Sinkhorn-type marginal iteration is involved. □

A.1.2 Proof of Lemma 1

Proof. We denote $\mathbf{t}_i = T_{\cdot,i}$. The joint objective over all concepts is $\mathcal{J}(T) = \sum_{i=1}^{K} \mathcal{J}_i(\mathbf{t}_i)$, and each summand $\mathcal{J}_i$ depends only on its own column $\mathbf{t}_i$ and its own cost column $C_{\cdot,i}$. There is no term coupling $\mathbf{t}_i$ with $\mathbf{t}_j$ for $i \neq j$, and IUCT imposes no shared token constraint of the form $\sum_i T_{n,i} = a_n$. The minimization therefore separates,

$$\min_{T\geq 0} \mathcal{J}(T) = \sum_{i=1}^{K} \min_{\mathbf{t}_i \geq 0} \mathcal{J}_i(\mathbf{t}_i), \tag{35}$$

and the minimizer of block i is obtained from block i alone. By Theorem 1, it is the explicit function $\mathbf{t}_i = \mathbf{t}(C_{\cdot,i}, \mathbf{a}, \rho, \epsilon, \tau)$ of Eq. 3, in which $C_{\cdot,j}$ for $j \neq i$ does not appear. Differentiating a function of $C_{\cdot,i}$ alone with respect to an argument $C_{\cdot,j}$ on which it does not depend gives $\partial \mathbf{t}_i / \partial C_{\cdot,j} = 0$.

□

A.1.3 Proof of Proposition 1

Proof. Recall $\delta_i^c = 2m_i^c \langle \bar{\mathbf{v}}_i^c, \Delta \mathbf{p}_i^c \rangle$ for $c \in \{0,1\}$, $E_i^c = \mathrm{softplus}(\delta_i^c/\theta)$, and $b_i^c = E_i^c / S_i$ with $S_i = E_i^1 + E_i^0 + 2 > 0$.

(*i*) The function $\mathrm{softplus}(u) = \log(1+e^u)$ has derivative $\mathrm{softplus}'(u) = e^u/(1+e^u) = \sigma(u) \in (0,1)$, so softplus is strictly increasing. With $\theta > 0$, $u \mapsto u/\theta$ is strictly increasing, so $E_i^c = \mathrm{softplus}(\delta_i^c/\theta)$ is a strictly increasing function of δ_i^c. Therefore,

$$b_i^1 > b_i^0 \iff \frac{E_i^1}{S_i} > \frac{E_i^0}{S_i} \tag{36}$$

$$\iff E_i^1 > E_i^0 \tag{37}$$

$$\iff \mathrm{softplus}(\tfrac{\delta_i^1}{\theta}) > \mathrm{softplus}\Big(\tfrac{\delta_i^0}{\theta}\Big) \tag{38}$$

$$\iff \delta_i^1 > \delta_i^0, \tag{39}$$

where the second equivalence divides through by the common positive S_i, and the last uses strict monotonicity of softplus. The equality case $b_i^1 = b_i^0 \iff E_i^1 = E_i^0 \iff \delta_i^1 = \delta_i^0$ follows identically. For the probability, $\hat{p}_i = \frac{\alpha_i^1}{S_i}$ and $1 - \hat{p}_i = \frac{\alpha_i^0}{S_i}$, so $\hat{p}_i > \frac{1}{2} \iff E_i^1 > E_i^0$, and the same chain applies

(*ii*) $S_i = 2 + \mathrm{softplus}(\frac{\delta_i^1}{\theta}) + \mathrm{softplus}(\frac{\delta_i^0}{\theta})$ is a sum of a constant and two strictly increasing functions of δ_i^1 and δ_i^0 respectively, so it is strictly increasing in each argument with the other held fixed. Since $u_i = 2/S_i$ and $s \mapsto 2/s$ is strictly decreasing on $(0,\infty)$, u_i is strictly decreasing in each of δ_i^1 and δ_i^0.

□

Remark 1. *With a single transport toward $\mathbf{p}_i^1$, the evidence for absence would have to be read from the negative side of the same margin, $E_i^0 = \mathrm{softplus}(-\delta_i^1/\theta)$. This margin is bounded below by a constant, so the belief in absence would be capped: the mass of a transport shrinks faster than the alignment grows when the tokens favour the other state. We make this precise. Fix a state and drop*

its index. Let $\mathbf{c}_n = \langle \mathbf{z}_n, \Delta\mathbf{p} \rangle$*, so that* $C_n = -2\mathbf{c}_n$*,* $\pi_n = e^{2\mathbf{c}_n/\epsilon}/(NQ)$ *with* $Q = \frac{1}{N}\sum_n e^{2\mathbf{c}_n/\epsilon}$*, and* $A := \langle \bar{\mathbf{v}}, \Delta\mathbf{p} \rangle = \sum_n \pi_n \mathbf{c}_n$*. The KL divergence of* π *from the uniform distribution is*

$$\sum_n \pi_n \log(N\pi_n) = \sum_n \pi_n \Big(\frac{2\mathbf{c}_n}{\epsilon} - \log Q \Big) = \frac{2A}{\epsilon} - \log Q \;\geq\; 0, \tag{40}$$

so $\log Q \leq 2A/\epsilon$ *and, by Theorem 1,* $m = \rho^{\tau/(\epsilon+\tau)} Q^{\epsilon/(\epsilon+\tau)} \leq \rho^{\tau/(\epsilon+\tau)} e^{2A/(\epsilon+\tau)}$*. If* $A \geq 0$ *then* $\delta = 2mA \geq 0$ *and there is nothing to prove. If* $A < 0$ *and write* $x = -A > 0$*, then*

$$\delta = -2mx \;\geq\; -2\rho^{\tau/(\epsilon+\tau)}\, x\, e^{-2x/(\epsilon+\tau)}. \tag{41}$$

The function $x \mapsto x e^{-2x/(\epsilon+\tau)}$ *on* $x > 0$ *is maximized at* $x = (\epsilon+\tau)/2$ *with value* $(\epsilon+\tau)/(2e)$*, so* $\delta \geq -\rho^{\tau/(\epsilon+\tau)}(\epsilon+\tau)/e$*, with equality when* π *is uniform and* $A = -(\epsilon+\tau)/2$*. With* $\epsilon = \tau = \sqrt{D}$ *and* $\theta = 0.1D$ *as in our experiments, this gives* $|\delta|/\theta \leq 20/(\mathrm{e}\sqrt{D})$*, about* 0.27 *at* $D = 768$*, so a single transport could never assign more than* $b_i^0 \approx 0.25$ *to absence. Evidence against a state is therefore read from the transport aimed at the other state, whose margin is unbounded above.*

A.1.4 Proof of Bayes Risk in Eq. (7)

Proof. Model the concept state as $c_i \sim \mathrm{Bernoulli}(p_i)$ with p_i the probability of the present state, so that the likelihood of the target is $p_i^{(c_i)}$ with $p_i^{(1)} = p_i$ and $p_i^{(0)} = 1 - p_i$. The Bayes risk is the expected negative log-likelihood under the second-order distribution,

$$\mathrm{BR}_i = \mathbb{E}_{p_i \sim \mathrm{Beta}(\alpha_i^1, \alpha_i^0)} \Big[-\log p_i^{(c_i)} \Big] = -\mathbb{E}\big[\log p_i^{(c_i)} \big]. \tag{42}$$

For $p \sim \mathrm{Beta}(\alpha^1, \alpha^0)$ with $S = \alpha^1 + \alpha^0$, the identities $\mathbb{E}[\log p] = \psi(\alpha^1) - \psi(S)$ and $\mathbb{E}[\log(1-p)] = \psi(\alpha^0) - \psi(S)$ follow from differentiating the logarithm of the normalizing constant $\log \mathrm{Beta}(\alpha^1, \alpha^0) = \log\Gamma(\alpha^1) + \log\Gamma(\alpha^0) - \log\Gamma(S)$ with respect to α^1 and α^0 respectively, since $\partial_{\alpha^1} \log B = \mathbb{E}[\log p]$ for an exponential family in natural parameterization. Substituting,

$$\mathrm{BR}_i = -\big(\psi(\alpha_i^{c_i}) - \psi(S_i) \big) = \psi(S_i) - \psi(\alpha_i^{c_i}), \tag{43}$$

which is the first term of Eq. (7).

□

A.1.5 Proof of Theorem 2

Proof. Fix all learned parameters, so the maps below are deterministic. The selected prototype $P_i = \mathbf{p}_i^{c_i^{\mathrm{hard}}}$ is a deterministic function of the bit $c_i^{\mathrm{hard}} = [b_i^1 \geq b_i^0] \in \{0, 1\}$, and c_i^{hard} is a deterministic function of X_i, so the chain $X_i \to c_i^{\mathrm{hard}} \to P_i$ is Markov.

Single concept. By the data-processing inequality applied to $X_i \to c_i^{\mathrm{hard}} \to P_i$, $I(X_i; P_i) \leq I(X_i; c_i^{\mathrm{hard}})$. Because mutual information is bounded by the entropy of either argument, $I(X_i; P_i) \leq H(c_i^{\mathrm{hard}})$. Since c_i^{hard} takes at most two values, $H(c_i^{\mathrm{hard}}) \leq \log_2 2 = 1$ bit. Chaining the three inequalities, $I(X_i; P_i) \leq 1$ bit.

All concepts jointly. Let $X_{1:K}$, $c_{1:K}^{\mathrm{hard}}$ and $P_{1:K}$ collect the corresponding quantities. Each P_i is a deterministic function of c^{hard}, so $P_{1:K}$ is a deterministic function of $c_{1:K}^{\mathrm{hard}}$, and $c_{1:K}^{\mathrm{hard}}$ is a deterministic function of $X_{1:K}$; thus $X_{1:K} \to c_{1:K}^{\mathrm{hard}} \to P_{1:K}$ is Markov. By data processing and the entropy bound, $I(X_{1:K}; P_{1:K}) \leq I(X_{1:K}; c_{1:K}^{\mathrm{hard}}) \leq H(c_{1:K}^{\mathrm{hard}})$. By subadditivity of entropy, which holds for any joint distribution and requires no independence, $H(c_{1:K}^{\mathrm{hard}}) \leq \sum_{i=1}^K H(c_i^{\mathrm{hard}}) \leq K$ bits.

□

A.1.6 Proof of Proposition 2

Proof. Condition on fixed g_i, c_i^{hard} and $\mathbf{p}_i = \mathbf{p}_i^{c_i^{\mathrm{hard}}}$, so that the auxiliary mean is $\mu_i(Z) = c_i^{\mathrm{hard}} + \eta g_i s_i(Z)$ with $s_i(Z) = \cos\angle(\mathbf{p}_i, \mathbf{v}_i(Z))$ regarded as the full closed-form function of the tokens through $T(Z)$. By linearity, $\nabla_Z \mu_i = \eta g_i \nabla_Z s_i$.

Fisher matrix of the auxiliary family. For $R_i \mid Z \sim \mathcal{N}(\mu_i(Z), \sigma_r^2)$ the log-density is $\log p(R_i \mid Z) = -\frac{1}{2\sigma_r^2}(R_i - \mu_i(Z))^2 - \frac{1}{2}\log(2\pi\sigma_r^2)$, so with $\vartheta = \mathrm{vec}(Z) \in \mathbb{R}^{ND}$ the score is

$$\nabla_\vartheta \log p(R_i \mid Z) = \frac{R_i - \mu_i(Z)}{\sigma_r^2}\, \nabla_\vartheta \mu_i. \tag{44}$$

The Fisher information matrix is the expected outer product of the score,

$$\mathcal{F}_Z = \mathbb{E}_{R_i|Z}\big[\nabla_\vartheta \log p\, \nabla_\vartheta \log p^\top\big] = \frac{\mathbb{E}[(R_i - \mu_i)^2]}{\sigma_r^4}\, \nabla_\vartheta \mu_i\, \nabla_\vartheta \mu_i^\top = \frac{1}{\sigma_r^2}\, \nabla_\vartheta \mu_i\, \nabla_\vartheta \mu_i^\top, \tag{45}$$

where $\mathbb{E}[(R_i - \mu_i)^2] = \sigma_r^2$ is the variance of the location family. Substituting $\nabla_Z \mu_i = \eta g_i \nabla_Z s_i$,

$$\mathcal{F}_Z = \frac{\eta^2 g_i^2}{\sigma_r^2}\, \mathrm{vec}(\nabla_Z s_i)\, \mathrm{vec}(\nabla_Z s_i)^\top, \tag{46}$$

which is a rank-one matrix.

Weighted trace. Assign token n the weight $w_{n,i} \geq 0$, applied to all D of its coordinates. For the rank-one matrix $\frac{\eta^2 g_i^2}{\sigma_r^2} v v^\top$ with $v = \mathrm{vec}(\nabla_Z s_i)$, the diagonal entry at coordinate (n, c) equals $\frac{\eta^2 g_i^2}{\sigma_r^2}(\nabla_{\mathbf{z}_n} s_i)_c^2$, so the weighted trace is

$$\mathcal{F}_i^{\mathrm{off}} = \sum_n w_{n,i} \sum_{c=1}^{D} (\mathcal{F}_Z)_{(n,c),(n,c)} = \frac{\eta^2 g_i^2}{\sigma_r^2} \sum_n w_{n,i} \sum_{c=1}^{D} (\nabla_{\mathbf{z}_n} s_i)_c^2 = \frac{\eta^2 g_i^2}{\sigma_r^2} \sum_n w_{n,i}\, \|\nabla_{\mathbf{z}_n} s_i\|_2^2, \tag{47}$$

which is the expression in Proposition 2. Since $w_{n,i} = (1 - N\pi_{n,i})_+$ is the positive part of an affine function of the continuous quantity $\pi_{n,i}$, it is continuous in Z, and $\mathcal{F}_i^{\mathrm{off}}$ is well defined everywhere. Unlike a hard mask $\mathbf{1}[\pi_{n,i} \leq 1/N]$, it requires no locally-constant-mask or measure-zero assumption.

Local KL bound For a smooth parametric family $p(\cdot \mid \vartheta)$, the Taylor expansion of $\vartheta' \mapsto D_{\mathrm{KL}}(p(\cdot \mid \vartheta) \| p(\cdot \mid \vartheta'))$ around $\vartheta' = \vartheta$ has vanishing value and vanishing gradient, the latter because the gradient equals $-\mathbb{E}_{p(\cdot|\vartheta)}[\nabla_{\vartheta'} \log p(\cdot \mid \vartheta')]$ evaluated at $\vartheta' = \vartheta$, which is the expected score and is zero. Hence

$$D_{\mathrm{KL}}\big(p(\cdot \mid \vartheta) \,\|\, p(\cdot \mid \vartheta + \Delta\vartheta)\big) = \tfrac{1}{2}\, \Delta\vartheta^\top \mathcal{F}(\vartheta)\, \Delta\vartheta + o(\|\Delta\vartheta\|^2). \tag{48}$$

With $\vartheta = \mathrm{vec}(Z)$, $\Delta\vartheta = \mathrm{vec}(\Delta Z)$ and the rank-one $\mathcal{F}_Z$ above,

$$\tfrac{1}{2}\Delta\vartheta^\top \mathcal{F}_Z \Delta\vartheta = \frac{\eta^2 g_i^2}{2\sigma_r^2}\big(\langle \nabla_Z s_i, \Delta Z_F \rangle\big)^2 = \frac{\eta^2 g_i^2}{2\sigma_r^2}\Big(\sum_n \langle \nabla_{\mathbf{z}_n} s_i, \Delta Z_n \rangle\Big)^2, \tag{49}$$

splitting the Frobenius inner product token by token. For a perturbation supported on $\{n : w_{n,i} > 0\}$, write each summand as $(\sqrt{w_{n,i}}\, \nabla_{\mathbf{z}_n} s_i)^\top (w_{n,i}^{-1/2} \Delta Z_n)$ and apply the Cauchy–Schwarz inequality over the token index:

$$\Big(\sum_n \langle \nabla_{\mathbf{z}_n} s_i, \Delta Z_n \rangle\Big)^2 \leq \Big(\sum_n w_{n,i} \|\nabla_{\mathbf{z}_n} s_i\|^2\Big)\Big(\sum_n w_{n,i}^{-1} \|\Delta Z_n\|^2\Big). \tag{50}$$

Multiplying by $\frac{\eta^2 g_i^2}{2\sigma_r^2}$ and recognizing $\mathcal{F}_i^{\mathrm{off}}$ gives the stated bound. If $\mathcal{F}_i^{\mathrm{off}} = 0$, every summand $w_{n,i}\|\nabla_{\mathbf{z}_n} s_i\|^2$ vanishes, so $\nabla_{\mathbf{z}_n} s_i = 0$ wherever $w_{n,i} > 0$. The first-order directional derivative $\sum_n \langle \nabla_{\mathbf{z}_n} s_i, \Delta Z_n \rangle$ then vanishes for every ΔZ supported on $\{n : w_{n,i} > 0\}$. □

A.1.7 Proof of Theorem 3

Proof. We omit the concept index and write $\kappa := \tau/(\epsilon + \tau) \in (0, 1)$ and $Q = \frac{1}{N}\sum_n e^{-C_n/\epsilon}$, so that $\pi_n = e^{-C_n/\epsilon}/(NQ)$ and $m = \rho^\kappa Q^{1-\kappa}$. All quantities refer to the transport of the selected state $c = c_i^{\mathrm{hard}}$: π, $\bar{\mathbf{v}}$, $\mathbf{p}$ and $\Delta\mathbf{p}$ stand for π^c, $\bar{\mathbf{v}}^c$, $\mathbf{p}^c$ and $\Delta\mathbf{p}^c = \mathbf{p}^c - \mathbf{p}^{1-c}$, so that $C_n = -2\langle \mathbf{z}_n, \Delta\mathbf{p} \rangle$ for either state and the derivation below is the same for both. Throughout, both prototypes and the hard state are held fixed, and we differentiate only with respect to the tokens.

Step 1 (Jacobian of the transport plan). Since $C_k = -2\langle \mathbf{z}_k, \Delta\mathbf{p}\rangle$ depends on Z only through $\mathbf{z}_k$, $\partial C_k/\partial \mathbf{z}_{n'} = -2\,\delta_{kn'}\Delta\mathbf{p}^\top$. From the closed form,

$$\log T_n = -\log N - \frac{C_n}{\epsilon} + \kappa \log \rho - \kappa \log Q, \qquad \frac{\partial \log Q}{\partial C_{n'}} = -\frac{e^{-C_{n'}/\epsilon}}{\epsilon N Q} = -\frac{\pi_{n'}}{\epsilon}, \tag{51}$$

so $\partial \log T_n/\partial C_{n'} = -\frac{1}{\epsilon}(\delta_{nn'} - \kappa\pi_{n'})$. By the chain rule,

$$\frac{\partial T_n}{\partial \mathbf{z}_{n'}} = T_n^\star \sum_k \frac{\partial \log T_n^\star}{\partial C_k}\frac{\partial C_k}{\partial \mathbf{z}_{n'}} = \frac{2}{\epsilon}\, T_n\big(\delta_{nn'} - \kappa\pi_{n'}\big)\Delta\mathbf{p}^\top \in \mathbb{R}^{1\times D}. \tag{52}$$

Two consistency checks follow. Summing Eq. (52) over n gives $\partial m/\partial \mathbf{z}_{n'} = \frac{2}{\epsilon}(1-\kappa)T_{n'}\Delta\mathbf{p}^\top$, which agrees with direct differentiation of $m = \rho^\kappa Q^{1-\kappa}$. In the balanced limit $\tau \to \infty$, $\kappa \to 1$, the mass becomes constant, and Eq. (52) reduces to the standard softmax Jacobian $\frac{2}{\epsilon} m\pi_n(\delta_{nn'} - \pi_{n'})\Delta\mathbf{p}^\top$.

Step 2 (Jacobian of the concept feature). Differentiating $\mathbf{v} = \sum_{n'} T_{n'}\mathbf{z}_{n'}$ and inserting Eq. (52),

$$\frac{\partial \mathbf{v}}{\partial \mathbf{z}_n} = T_n I_D + \sum_{n'} \mathbf{z}_{n'}\frac{\partial T_{n'}}{\partial \mathbf{z}_n} = T_n I_D + \frac{2}{\epsilon}\Big(T_n\mathbf{z}_n - \kappa\pi_n \sum_{n'} T_{n'}\mathbf{z}_{n'}\Big)\Delta\mathbf{p}^\top \tag{53}$$

$$= T_n I_D + \frac{2}{\epsilon}\big(T_n\mathbf{z}_n - \kappa\pi_n\mathbf{v}\big)\Delta\mathbf{p}^\top. \tag{54}$$

Because $\pi_n\mathbf{v} = \pi_n m\bar{\mathbf{v}} = T_n\bar{\mathbf{v}}$, this is

$$\frac{\partial \mathbf{v}}{\partial \mathbf{z}_n} = T_n\Big[I_D + \frac{2}{\epsilon}\big(\mathbf{z}_n - \kappa\bar{\mathbf{v}}\big)\Delta\mathbf{p}^\top\Big] \in \mathbb{R}^{D\times D}. \tag{55}$$

Step 3 (Gradient of the cosine and orthogonality). For $s(\mathbf{v}) = \langle \mathbf{p}, \mathbf{v}\rangle/(\|\mathbf{p}\|\|\mathbf{v}\|)$, direct differentiation gives

$$\mathbf{h} := \frac{\partial s}{\partial \mathbf{v}} = \frac{\mathbf{p}}{\|\mathbf{p}\|\|\mathbf{v}\|} - \frac{s\,\mathbf{v}}{\|\mathbf{v}\|^2}, \qquad \langle \mathbf{h}, \mathbf{v}\rangle = \frac{\langle \mathbf{p}, \mathbf{v}\rangle}{\|\mathbf{p}\|\|\mathbf{v}\|} - \frac{s\|\mathbf{v}\|^2}{\|\mathbf{v}\|^2} = s - s = 0. \tag{56}$$

The cosine is positively 0-homogeneous in its second argument, so $s(\mathbf{v}) = s(\bar{\mathbf{v}})$ with $\mathbf{v} = m\bar{\mathbf{v}}$. Differentiating this identity with respect to $\bar{\mathbf{v}}$ gives $\bar{\mathbf{a}} = \partial s/\partial\bar{\mathbf{v}} = m\mathbf{h}$, and the same computation with $\bar{\mathbf{v}}$ in place of $\mathbf{v}$ yields the explicit formula for $\bar{\mathbf{a}}$ stated in the theorem together with $\langle \bar{\mathbf{a}}, \bar{\mathbf{v}}\rangle = 0$. Expanding the square,

$$\|\bar{\mathbf{a}}\|^2 = \frac{1}{\|\bar{\mathbf{v}}\|^2} - \frac{2s\langle \mathbf{p}, \bar{\mathbf{v}}\rangle}{\|\mathbf{p}\|\|\bar{\mathbf{v}}\|^3} + \frac{s^2}{\|\bar{\mathbf{v}}\|^2} = \frac{1 - 2s^2 + s^2}{\|\bar{\mathbf{v}}\|^2} = \frac{1-s^2}{\|\bar{\mathbf{v}}\|^2}. \tag{57}$$

This is the standard property that the gradient of a scale-invariant function is orthogonal to its argument. It is the reason the unbalanced-mass term of Eq. (55) vanishes in the next step.

Step 4 (The token gradient). By the chain rule and Eq. (55),

$$\nabla_{\mathbf{z}_n} s = \Big(\frac{\partial \mathbf{v}}{\partial \mathbf{z}_n}\Big)^\top \mathbf{h} = T_n\Big[\mathbf{h} + \frac{2}{\epsilon}\,\Delta\mathbf{p}\,(\mathbf{z}_n - \kappa\bar{\mathbf{v}})^\top\mathbf{h}\Big] = T_n\Big[\mathbf{h} + \frac{2}{\epsilon}\Big(\langle \mathbf{z}_n, \mathbf{h}\rangle - \kappa\langle\bar{\mathbf{v}}, \mathbf{h}\rangle\Big)\Delta\mathbf{p}\Big]. \tag{58}$$

By Step 3, $\langle\bar{\mathbf{v}}, \mathbf{h}\rangle = \langle \mathbf{v}, \mathbf{h}\rangle/m = 0$, so the term carrying κ, which is the only place where the adaptive mass and the softmax normalization enter the Jacobian, vanishes identically. This gives $\nabla_{\mathbf{z}_n} s = T_n[\mathbf{h} + \frac{2}{\epsilon}\langle \mathbf{z}_n, \mathbf{h}\rangle\Delta\mathbf{p}]$. Substituting $T_n = m\pi_n$ and $\mathbf{h} = \bar{\mathbf{a}}/m$ cancels $m^\star$ and yields the first identity of Eq. (11) with $\xi_n = \langle \mathbf{z}_n, \bar{\mathbf{a}}\rangle$.

Step 5 (Mass-freeness). In the second form, π_n is the softmax of $-C_\cdot/\epsilon$ and involves neither τ nor ρ; $\bar{\mathbf{v}} = \sum_n \pi_n\mathbf{z}_n$ and $s = \cos\angle(\mathbf{p}, \bar{\mathbf{v}})$ are functions of $\pi_\cdot$, Z and $\mathbf{p}$ only; and $\bar{\mathbf{a}}$ is a function of $\mathbf{p}$, $\bar{\mathbf{v}}$ and s. Hence $\nabla_{\mathbf{z}_n} s$ is independent of m, τ and ρ, and it coincides with the gradient of $\cos\angle(\mathbf{p}, \sum_n \pi_n\mathbf{z}_n)$ computed from the normalized allocation alone. This is the final statement of Theorem 3.

Step 6 (Squared norm). Taking the squared norm of the first identity and using $\|\bar{\mathbf{a}}\|^2 = (1 - s^2)/\|\bar{\mathbf{v}}\|^2$ from Step 3,

$$\|\nabla_{\mathbf{z}_n} s\|^2 = \pi_n^2 \Big[\|\bar{\mathbf{a}}\|^2 + \frac{4}{\epsilon}\xi_n \langle \bar{\mathbf{a}}, \Delta\mathbf{p}\rangle + \frac{4}{\epsilon^2}\xi_n^2 \|\Delta\mathbf{p}\|^2 \Big] = \pi_n^2 \Big[\frac{1 - s^2}{\|\bar{\mathbf{v}}\|^2} + \frac{4}{\epsilon}\xi_n \langle \bar{\mathbf{a}}, \Delta\mathbf{p}\rangle + \frac{4}{\epsilon^2}\xi_n^2 \|\Delta\mathbf{p}\|^2 \Big], \quad (59)$$

the second identity of Eq. (11). Substituting into the expression for $\mathcal{F}_i^{\text{off}}$ in Proposition 2 gives Eq. (13) in terms of forward-pass quantities. The only new computation is $\xi_{n,i} = \langle \mathbf{z}_n, \bar{\mathbf{a}}_i \rangle$ for all (n, i), which costs $O(BNKD)$. Because Eq. (11) is an exact identity and not an approximation, the analytic and automatic-differentiation expressions of the penalty are the same function of the encoder and prototype parameters whenever g_i and $w_{n,i}$ are treated as constants in both; their values and parameter gradients coincide up to floating-point round-off, and by Step 5 the parameter gradient of $\mathcal{L}_{\text{fisher}}^{\text{off}}$ does not pass through m_i, τ or h except through the gate, which is held fixed.

Per-summand bound. By the triangle inequality, the Cauchy–Schwarz bound $|\xi_n| \le \|\mathbf{z}_n\|\|\bar{\mathbf{a}}\| \le B_z\|\bar{\mathbf{a}}\|$, and $\|\Delta\mathbf{p}\| \le \|\mathbf{p}^1\| + \|\mathbf{p}^0\| = 2\sqrt{D}$, we obtain

$$\Big\| \bar{\mathbf{a}} + \frac{2}{\epsilon}\xi_n \Delta\mathbf{p} \Big\| \le \|\bar{\mathbf{a}}\| + \frac{2}{\epsilon} B_z \|\bar{\mathbf{a}}\| \cdot 2\sqrt{D} = \|\bar{\mathbf{a}}\| \Big(1 + \frac{4B_z\sqrt{D}}{\epsilon} \Big), \quad (60)$$

and squaring with $\|\bar{\mathbf{a}}\|^2 = (1 - s^2)/\|\bar{\mathbf{v}}\|^2$ bounds the bracket of Eq. (11) by

$$\frac{(1 - s^2)\,(1 + 4B_z\sqrt{D}/\epsilon)^2}{\|\bar{\mathbf{v}}\|^2}. \quad (61)$$

For the token weight, the function $\phi(\pi) := (1 - N\pi)\pi^2$ on $[0, 1/N]$ satisfies $\phi'(\pi) = 2\pi - 3N\pi^2 = 0$ at $\pi = 2/(3N)$, where $\phi = \frac{1}{3} \cdot \frac{4}{9N^2} = \frac{4}{27N^2}$, and $\phi(0) = \phi(1/N) = 0$; moreover $(1 - N\pi)_+ = 0$ for $\pi \ge 1/N$. Multiplying the two bounds gives the per-summand bound quoted in Section 3.4. The only unbounded factor is $1/\|\bar{\mathbf{v}}\|^2$, which diverges when the allocation-weighted token mean vanishes; this is the degenerate configuration in which the cosine itself is undefined. □

A.1.8 Proof of Proposition 3

Proof. Throughout, $g_n := \nabla_{\mathbf{z}_n} s_i \in \mathbb{R}^D$, $\mathcal{P}_n$ is the patch of token n (the patches partition the pixels), $\mathcal{R}_n \supseteq \mathcal{P}_n$ is its receptive field, the pixels on which $\mathbf{z}_n$ depends, and $J_{n,p} := \partial \mathbf{z}_n / \partial x_p$ denotes the $D \times 3$ Jacobian of token n with respect to pixel p, which vanishes unless $p \in \mathcal{R}_n$. By the chain rule,

$$\nabla_{x_p} s_i = \sum_{n:\, p \in \mathcal{R}_n} J_{n,p}^\top g_n. \quad (62)$$

Let $\Phi : \mathbb{R}^{H_{\text{px}} W_{\text{px}} \cdot 3} \to \mathbb{R}^{ND}$ be the encoder, $Z = \Phi(x)$, and s_i a scalar computed from Z.

(*i*) By the chain rule for the composition $x \mapsto Z \mapsto s_i$, in vectorized form,

$$\nabla_x s_i = \Big(\frac{\partial \operatorname{vec}(Z)}{\partial \operatorname{vec}(x)} \Big)^\top \nabla_Z s_i = J_\Phi(x)^\top \nabla_Z s_i. \quad (63)$$

Taking Euclidean norms and using the definition of the operator norm together with the invariance of the spectral norm under transposition,

$$\|\nabla_x s_i\| = \|J_\Phi(x)^\top \nabla_Z s_i\| \le \|J_\Phi(x)\|_{\text{op}} \|\nabla_Z s_i\| \le L_\Phi \|\nabla_Z s_i\|. \quad (64)$$

(*ii*) Bilinear interpolation writes each output pixel as $\sum_{k=1}^4 \omega_k\, \Pi_i[c_k]$ over the four neighbouring coarse cells c_k with $\omega_k \ge 0$ and $\sum_k \omega_k = 1$. Nonnegativity of the weights and of the inputs makes each output a nonnegative combination, so $U(\Pi_i) \ge 0$. On the interior of a coarse cell, the output is a convex combination of neighbouring cell values, which is monotone in those values, so the relative ordering of cell values is preserved away from boundaries, and the argmax cell maps to the argmax pixel region up to boundary interpolation. Finally, U acts on the spatial indices only, and the concept index i is a free batch dimension, so $U(\Pi_i)$ depends on no Π_j with $j \ne i$. With nearest-neighbour expansion, the weights ω_k degenerate to indicators, and the same argument applies.

□

Remark 2. *The main text states the pixel-level guarantee for the total Fisher information. A token-wise version for the off-support penalty needs a locality hypothesis. Let $\|\partial \mathbf{z}_n/\partial x_{\mathcal{R}_n}\|_{\mathrm{op}} \le L_{\mathrm{loc}}$ for every n, let ω be the largest number of receptive fields that contain a single pixel, set the pixel weights $\tilde{w}_p := w_{n,i}$ for $p \in \mathcal{P}_n$, and let $\mathcal{O}_i = \{p : \tilde{w}_p > 0\}$ be the off-support pixels, the patches of the off-support tokens. Then:*

(a) If the receptive fields are disjoint, $\mathcal{R}_n = \mathcal{P}_n$, then $\sum_p \tilde{w}_p \|\nabla_{x_p} s_i\|^2 \le L_{\mathrm{loc}}^2 \sum_n w_{n,i} \|\nabla_{\mathbf{z}_n} s_i\|^2$: with the same support, the token-level off-support penalty bounds the pixel-level one.

(b) In general,

$$\sum_{p\in\mathcal{O}_i} \tilde{w}_p \|\nabla_{x_p} s_i\|^2 \;\le\; 2\omega L_{\mathrm{loc}}^2 \Big[\sum_{n:\, w_{n,i}>0} \|\nabla_{\mathbf{z}_n} s_i\|^2 + \sum_{n:\, w_{n,i}=0,\ \mathcal{R}_n\cap\mathcal{O}_i\neq\emptyset} \|\nabla_{\mathbf{z}_n} s_i\|^2 \Big]. \tag{65}$$

The first sum is the off-support token sensitivity the penalty drives down; the second is a spill term, the influence of off-support pixels through on-support tokens whose receptive fields reach them, on which the penalty does not act. It is empty when the receptive fields are disjoint.

(c) Any pixel allocation that is constant on each $\mathcal{P}_n$ and sums to $\pi_{n,i}$ on it induces the same support as Eq. (9).

The constants $L_{\mathrm{loc}} \le L_\Phi$ and ω are existence hypotheses and are not computed.

Proof of Remark 2. Based on the condition, we have:

(a) With disjoint receptive fields, $\mathcal{R}_n = \mathcal{P}_n$, so each pixel p lies in exactly one receptive field and Eq. (62) has a single term, $\nabla_{x_p} s_i = J_{n,p}^\top g_n$ for $p \in \mathcal{P}_n$. Since the $\mathcal{P}_n$ partition the pixels,

$$\sum_p \tilde{w}_p \|\nabla_{x_p} s_i\|^2 = \sum_n w_{n,i} \sum_{p\in\mathcal{P}_n} \|J_{n,p}^\top g_n\|^2 = \sum_n w_{n,i} \|J_{n,\mathcal{P}_n}^\top g_n\|^2 \le L_{\mathrm{loc}}^2 \sum_n w_{n,i} \|g_n\|^2, \tag{66}$$

where $J_{n,\mathcal{P}_n}$ stacks the $J_{n,p}$ over $p \in \mathcal{P}_n$ and the last step is the operator-norm bound.

(b) Fix $p \in \mathcal{O}_i$ and split the sum in Eq. (62) into off-support tokens ($w_{n,i} > 0$) and on-support tokens ($w_{n,i} = 0$), $\nabla_{x_p} s_i = a_p + b_p$. By $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ and the Cauchy–Schwarz inequality over the at most ω terms of each sum,

$$\|\nabla_{x_p} s_i\|^2 \le 2\omega \sum_{\substack{n:\, w_{n,i}>0 \\ p\in\mathcal{R}_n}} \|J_{n,p}^\top g_n\|^2 + 2\omega \sum_{\substack{n:\, w_{n,i}=0 \\ p\in\mathcal{R}_n}} \|J_{n,p}^\top g_n\|^2. \tag{67}$$

Multiply by $\tilde{w}_p \le 1$, sum over $p \in \mathcal{O}_i$ and exchange the order of summation. For each token n, the pixels that contribute lie in $\mathcal{R}_n \cap \mathcal{O}_i \subseteq \mathcal{R}_n$, so $\sum_{p\in\mathcal{R}_n\cap\mathcal{O}_i} \|J_{n,p}^\top g_n\|^2 \le \|J_{n,\mathcal{R}_n}^\top g_n\|^2 \le L_{\mathrm{loc}}^2 \|g_n\|^2$. An on-support token contributes only if $\mathcal{R}_n \cap \mathcal{O}_i \neq \emptyset$. Collecting terms gives the stated bound. When the receptive fields are disjoint, an on-support token has $\mathcal{R}_n = \mathcal{P}_n$ disjoint from $\mathcal{O}_i$, so the second sum is empty. The spill term can carry all of the pixel sensitivity: for $z_1 = x_1 + x_2$, $z_2 = x_2$, $s = z_1$ with token 2 off-support, $g_2 = 0$ while $\partial s/\partial x_2 = 1$.

(c) Let M be the number of pixels and $|\mathcal{P}_n| = M/N$ for all n. If a pixel allocation $\pi_{\cdot,i}^{\mathrm{px}}$ is constant on each $\mathcal{P}_n$ with $\sum_{p\in\mathcal{P}_n} \pi_{p,i}^{\mathrm{px}} = \pi_{n,i}$, then $\pi_{p,i}^{\mathrm{px}} = \pi_{n,i} N/M$ for $p \in \mathcal{P}_n$, and the pixel-level uniform share is $1/M$, so $\pi_{p,i}^{\mathrm{px}} > 1/M \iff \pi_{n,i} N/M > 1/M \iff \pi_{n,i} > 1/N$. The pixel allocation therefore induces the same support as Eq. (9). □

A.1.9 Proof of Theorem 4

Proof. For a state c write $s_i^c = \cos\angle(\mathbf{p}_i^c, \mathbf{v}_i^c)$ and $g_i^c = (b_i^c)^\gamma$, each state scored against its own prototype and its own transported feature; the free side uses $c = c_i^{\mathrm{hard}}$, the intervened side $c = c_i$.

From the bottleneck definition Eq. (8), evaluated on the free side and on the intervened side,

$$r_i^{\text{free}} = \left(c_i^{\text{hard}} - \frac{1}{2}\right) + \eta\, g_i^{c_i^{\text{hard}}}\, s_i^{c_i^{\text{hard}}}, \qquad r_i^{\text{int}} = \left(c_i - \frac{1}{2}\right) + \eta\, g_i^{c_i}\, s_i^{c_i}. \tag{68}$$

Subtracting, the centering cancels and

$$\Delta r_i := r_i^{\text{int}} - r_i^{\text{free}} = (c_i - c_i^{\text{hard}}) + \eta\big(g_i^{c_i} s_i^{c_i} - g_i^{c_i^{\text{hard}}} s_i^{c_i^{\text{hard}}}\big) = \Delta c + \eta \Delta_i. \tag{69}$$

Because the free prediction is wrong, $c_i^{\text{hard}} \neq c_i$ with both in $\{0,1\}$, so $\Delta c \in \{-1,+1\}$ and $|\Delta c| = 1$.

Bound on Δ_i Both scores satisfy $|s_i^{c_i^{\text{hard}}}| \leq 1$ and $|s_i^{c_i}| \leq 1$ by the Cauchy–Schwarz inequality, and each gate lies in $[0,1]$, being a belief in $[0,1]$ raised to the power $\gamma > 0$. By the triangle inequality,

$$|\Delta_i| = \big|g_i^{c_i} s_i^{c_i} - g_i^{c_i^{\text{hard}}} s_i^{c_i^{\text{hard}}}\big| \leq \big(g_i^{c_i}|s_i^{c_i}| + g_i^{c_i^{\text{hard}}}|s_i^{c_i^{\text{hard}}}|\big) \leq 2, \qquad \text{so } \eta|\Delta_i| < 1, \tag{70}$$

where the last step holds because $\eta < \frac{1}{2}$. The bound uses only the structural $|s_i| \leq 1$ and the ranges of g_i and η, with no assumption on the learned geometry and no data-dependent constant.

Sign preservation and lower bound By the reverse triangle inequality, $|\Delta r_i| = |\Delta c + \eta\Delta_i| \geq |\Delta c| - \eta|\Delta_i| \geq 1 - 2\eta > 0$. Since $\eta|\Delta_i| < 1 = |\Delta c|$, a perturbation of magnitude less than one cannot reverse the sign of a term of magnitude one, so $\Delta c + \eta\Delta_i$ has the sign of Δc: the intervention moves r_i in the corrective direction $\text{sign}(c_i - c_i^{\text{hard}})$.

Unit bound under corrective alignment If $\Delta c \cdot \Delta_i \geq 0$ in addition, the two terms do not cancel, and $|\Delta r_i| = |\Delta c| + \eta|\Delta_i| \geq 1$.

Gate contraction The free hard state $c_i^{\text{hard}} = \mathbf{1}[b_i^1 \geq b_i^0]$ selects the state with the larger belief, so $b_i^{c_i^{\text{hard}}} = \max\{b_i^1, b_i^0\}$. On a wrong prediction, the ground-truth state is the other one, whose belief is the smaller, so $b_i^{c_i} = \min\{b_i^1, b_i^0\} \leq \max\{b_i^1, b_i^0\} = b_i^{c_i^{\text{hard}}}$. Since $t \mapsto t^\gamma$ is non-decreasing on $[0,1]$ for $\gamma > 0$, $g_i^{c_i} = (b_i^{c_i})^\gamma \leq (b_i^{c_i^{\text{hard}}})^\gamma = g_i^{c_i^{\text{hard}}}$. For the quantitative form, if $b_i^{c_i} \leq \underline{b} < \bar{b} \leq b_i^{c_i^{\text{hard}}}$, $b_i^{c_i}/b_i^{c_i^{\text{hard}}} \leq \underline{b}/\bar{b}$, and raising to the positive power γ preserves the inequality, $g_i^{c_i}/g_i^{c_i^{\text{hard}}} = (b_i^{c_i}/b_i^{c_i^{\text{hard}}})^\gamma \leq (\underline{b}/\bar{b})^\gamma < 1$, the last inequality because $\underline{b}/\bar{b} < 1$.

□

A.1.10 Proof of Theorem 5

Proof. For a competitor class $\ell \neq y$, a correct classification of r against ℓ means $M_{y,\ell}(\mathbf{r}) > 0$. Consider the scalar path $t \mapsto \mathbf{r} + t\Delta r_j \mathbf{e}_j$ for $t \in [0,1]$ and define $\phi(t) := M_{y,\ell}(\mathbf{r} + t\Delta r_j \mathbf{e}_j)$. Only the j-th argument varies along the path, at rate Δr_j, so by the chain rule $\phi'(t) = \Delta r_j\, \partial_{r_j} M_{y,\ell}(\mathbf{r} + t\Delta r_j \mathbf{e}_j)$. By the fundamental theorem of calculus,

$$M_{y,\ell}(\mathbf{r} + \Delta r_j \mathbf{e}_j) - M_{y,\ell}(\mathbf{r}) = \phi(1) - \phi(0) = \int_0^1 \Delta r_j\, \partial_{r_j} M_{y,\ell}\big(\mathbf{r} + t\Delta r_j \mathbf{e}_j\big)\, dt. \tag{71}$$

Under Eq. (15), the integrand is nonnegative for every $t \in [0,1]$ and every $\ell \neq y$, and the integral of a nonnegative function is nonnegative, so $M_{y,\ell}(\mathbf{r} + \Delta r_j \mathbf{e}_j) \geq M_{y,\ell}(\mathbf{r})$. Therefore, every pairwise margin is non-decreasing, and so is the minimum margin $\min_{\ell \neq y} M_{y,\ell}$. If $\mathbf{r}$ was correctly classified, it remains so, and accuracy cannot decrease along a nested sequence of correct interventions. Taking expectations over images and intervention sets preserves the inequality, so the expected intervention accuracy is monotone non-decreasing.

For a linear head $f_\ell(\mathbf{r}) = \mathbf{w}_\ell^\top \mathbf{r} + b_\ell$, $M_{y,\ell}(\mathbf{r}) = (\mathbf{w}_y - \mathbf{w}_\ell)^\top \mathbf{r} + (b_y - b_\ell)$ and $\partial_{r_j} M_{y,\ell} = w_{y,j} - w_{\ell,j}$, a constant that does not vary along the path. Condition Eq. (15) then reduces to $\Delta r_j (w_{y,j} - w_{\ell,j}) \geq 0$ for all $\ell \neq y$. □

A.1.11 Proof of Proposition 4

Proof. We write $\mathbf{r}^{\mathrm{int}}$ for the corrected bottleneck of the sample. For cross-entropy with softmax class probabilities $\hat{y}_\ell = \mathrm{softmax}_\ell(f(\mathbf{r}^{\mathrm{int}}))$, the gradient of the loss with respect to the logit f_ℓ is $\hat{y}_\ell - \mathbf{1}[\ell = y]$, so $\partial\,\mathrm{CE}/\partial \mathbf{w}_\ell = (\hat{y}_\ell - \mathbf{1}[\ell = y])\,\mathbf{r}^{\mathrm{int}}$. A gradient-descent step with rate $\nu > 0$ is $\Delta\mathbf{w}_\ell = -\nu(\hat{y}_\ell - \mathbf{1}[\ell = y])\,\mathbf{r}^{\mathrm{int}}$, and in the j-th coordinate $\Delta w_{\ell,j} = -\nu(\hat{y}_\ell - \mathbf{1}[\ell = y])\,r_j^{\mathrm{int}}$. Hence, for $\ell \neq y$,

$$\Delta(w_{y,j} - w_{\ell,j}) = -\nu\big[(\hat{y}_y - 1) - \hat{y}_\ell\big]\,r_j^{\mathrm{int}} = \nu\,(1 - \hat{y}_y + \hat{y}_\ell)\,r_j^{\mathrm{int}}, \tag{72}$$

and $1 - \hat{y}_y + \hat{y}_\ell \geq 0$ always, since $\hat{y}_y \leq 1$ and $\hat{y}_\ell \geq 0$. The bias receives $\Delta b_\ell = -\nu(\hat{y}_\ell - \mathbf{1}[\ell = y])$, which does not enter $w_{y,j} - w_{\ell,j}$.

It remains to show that $\mathrm{sign}(r_j^{\mathrm{int}}) = \mathrm{sign}(\Delta r_j)$ for a corrected j. After the corrective flip $c_j^{\mathrm{hard}} \leftarrow c_j$, by Eq. (8), the coordinate is $r_j^{\mathrm{int}} = (c_j - \frac{1}{2}) + \eta g_j s_j$, where s_j and g_j are the score and gate evaluated at the corrected state. Since $c_j \in \{0, 1\}$ gives $c_j - \frac{1}{2} = \pm\frac{1}{2}$ and $|\eta g_j s_j| \leq \eta < \frac{1}{2}$, the identity term strictly dominates the residual on both branches:

$$c_j = 1 :\ r_j^{\mathrm{int}} \geq \tfrac{1}{2} - \eta > 0, \qquad c_j = 0 :\ r_j^{\mathrm{int}} \leq -\tfrac{1}{2} + \eta < 0. \tag{73}$$

Therefore, $\mathrm{sign}(r_j^{\mathrm{int}}) = \mathrm{sign}(c_j - \frac{1}{2})$. Because the free prediction was wrong, $c_j^{\mathrm{hard}} = 1 - c_j$, so $\mathrm{sign}(c_j - \frac{1}{2}) = \mathrm{sign}(c_j - c_j^{\mathrm{hard}}) = \mathrm{sign}(\Delta c) = \mathrm{sign}(\Delta r_j)$, the last equality by Theorem 4.

Combining, $\Delta(w_{y,j} - w_{\ell,j})$ is a nonnegative multiple of $\mathrm{sign}(\Delta r_j)$, so the step increases $\Delta r_j(w_{y,j} - w_{\ell,j})$ whenever that product is negative and reduces any current violation of Eq. (15). The statement concerns the contribution of one sample to one gradient-descent step, and it does not bound the contributions of other samples in the same batch, and it does not extend to adaptive optimizers, whose per-coordinate rescaling can change the sign of the aggregate update.

□

B Experimental Details

Code is available at `https://anonymous.4open.science/r/FisherCBM-B9D1`.

B.1 Datasets

CUB-200-2011. We use the 11,788 images and 200 classes of CUB-200-2011 (Wah et al., 2011) with the official train/test split (5,994 / 5,794 images). Concepts are the 112 binary attributes selected by Koh et al. (2020), in their standard order, partitioned into the 28 attribute groups of the official CEM implementation (*bill shape*, *wing color*, ...). Following Koh et al. (2020) and the CEM, ProbCBM and CIBM code bases, the per-image attribute annotations, which are noisy and do not determine the class, are replaced by the class-level majority vote over the training images, so every image of a class carries the same concept vector. A stratified 10% of the training images is held out for model selection.

AwA2. We use the 37,322 images and 50 classes of Animals with Attributes 2 (Xian et al., 2019) with the 85 binary predicates of `predicate-matrix-binary.txt` as class-level concepts, in the predicate order of the official CEM loader. Images are split 60/20/20 into train, validation, and test at random with a fixed seed, following the CEM protocol; the split is generated once and cached. For group-wise interventions, the 85 predicates are partitioned into semantic groups that partition the predicate set.

CelebA. The incomplete-concept task of Zarlenga et al. (2022), reproduced from the official CEM loader. Over all 202,599 images of CelebA (Liu et al., 2015), the eight attributes whose positive rate is closest to 50% are selected. The six most balanced (*Attractive*, *High Cheekbones*, *Male*, *Mouth Slightly Open*, *Smiling*, *Wearing Lipstick*) are the visible concepts; the remaining two (*Heavy Makeup*, *Wavy Hair*) are hidden, so they shape the label but are never annotated. The label is the 8-bit vector of all eight attributes, remapped to the 230 combinations that occur. A random $1/12$ subsample of 16,883 images is drawn and split 70/10/20 into train, validation, and test, both with the

official seed. Interventions are per attribute, since the six concepts form no groups. Because the task is constructed so that the concepts are insufficient for the label, it is the setting in which the residual channel and the choice of β would matter most, and we compare $\beta = 0$ and $\beta = 1$ on it. Different from the official CelebA configuration, inputs are 224×224 into the shared frozen ViT-B/16 with the same transforms as CUB and AwA2, and the concept BCE is unweighted.

Preprocessing. Inputs are 224×224 and ImageNet-normalized. Training uses random resized crops (area scale 0.6–1.0 on CUB, 0.7–1.0 on AwA2) and horizontal flips; evaluation resizes the short side to $224/0.875 = 256$ pixels and center-crops, preserving aspect ratio.

B.2 MODELS

All models share a ViT-B/16 encoder (Dosovitskiy et al., 2021) pretrained on ImageNet, taken from timm, which yields $N = 14 \times 14 = 196$ patch tokens of width 768. The encoder is frozen in all main runs; only the adapter, the bottleneck parameters, and the heads are trained. FisherCBM does not require a frozen encoder, and Section 3.1 places no restriction on it, but freezing fixes the backbone capacity across all compared models and makes the $K = 112$ sweep affordable. The adapter is one 1×1 convolution–BatchNorm–GELU block operating on the token grid. For FisherCBM, the adapter output is followed by a parameter-free LayerNorm on every token, so that $\|\mathbf{z}_n\| = \sqrt{D}$ exactly and the bound $B_z = \sqrt{D}$ of Section 3.1 holds with a known constant.

FisherCBM. Tokens and prototypes pass through the same parameter-free normalization $x \mapsto (x - \bar{x})/\mathrm{std}(x)$, computed without the usual ε so that $\|\mathbf{z}_n\| = \|\mathbf{p}_i^c\| = \sqrt{D}$ holds exactly and the offset in Eq. (1) vanishes identically rather than approximately. Prototypes are initialized from a standard Gaussian; since the normalization removes their scale, the initialization only sets the effective learning rate. Because the natural scale of the cost is $O(\sqrt{D})$ and that of the margin δ_i is $O(D)$, the transport and evidence hyperparameters are specified in D-free units and scaled internally: $\epsilon = \tau = 1 \cdot \sqrt{D}$, $\rho = 1$, and $\theta = 0.1 \cdot D$. The gate exponent is $\gamma = 1$, and the residual bound $\eta = 0.4$. The concept loss is the EDL loss Eq. (7) with $\lambda_{D_{\mathrm{KL}}}(t)$ ramped linearly from 0 to 1 over the first 1,000 steps. The concept feature $\mathbf{v}_i$ is fully differentiable through the transport; the hard state is the only detached quantity. The identity bit is propagated with a straight-through estimator, $c_i^{\mathrm{hard}} + \hat{p}_i - \mathrm{sg}[\hat{p}_i]$, on un-intervened concepts, while intervened concepts inject the exact bit with no gradient; the hard-cut ablation detaches the bit for all concepts, so the identity is trained by the EDL loss alone. RandInt draws a rate $\varrho \sim \mathcal{U}(0, 1)$ per sample, and a Bernoulli(ϱ) mask per concept, and with probability 0.1 sets the mask to all K concepts so that the endpoint of the intervention curve is seen in training. The Fisher penalty Eq. (13) is computed from the closed form of Theorem 3 with the gate g_i and the off-support weights $w_{n,i}$ detached, and the constant η^2/σ_r^2 absorbed into $\beta = 1$. The $m_i \equiv 1$ ablation sets $\tau = 10^6 \cdot \sqrt{D}$, for which m_i equals 1 to within 10^{-4}; its θ-matched variant divides θ by the median training mass of the full model (220 on CUB, 100 on CelebA, 4300 on AwA2). The θ sweep of Appendix C.2 changes only the penalty weight.

Baselines. All baselines are ports of the official CEM repository (Zarlenga et al., 2022), which also contains the CBM variants of Koh et al. (2020). *End-to-end*: the adapted, pooled feature is fed to a linear head with no concept layer. *Hard CBM*: concepts are thresholded to $\{0, 1\}$ before the head. *Fuzzy CBM*: sigmoid probabilities are passed to the head. *Logit CBM*: pre-sigmoid logits are passed. *HybridCBM* (Mahinpei et al., 2021): a fuzzy CBM with 50 additional unsupervised side-channel dimensions. *CEM*: one 16-dimensional embedding per concept (32 on AwA2), mixed by the predicted probability, trained with RandInt at $p = 0.25$; we also report CEM without RandInt. *ProbCBM* (Kim et al., 2023): Gaussian concept embeddings of dimension 16 with a 128-dimensional class embedding, 8 Monte-Carlo samples, replacement probability 0.25, no weight decay, following the official configuration. *CIBM* (Galliamov et al., 2026): a fuzzy CBM with a variational concept head and an information-bottleneck term, in both the estimator variant (IB_E, $\beta = 0.5$, marginal approximated over 64 batch samples) and the bounded variant (IB_B, $\beta = 0.5$). All models, FisherCBM included, use the same concept-loss weight $\lambda_c = 5$ on every dataset, rather than the per-model, per-dataset weights of the CEM repository, so that the comparison isolates the bottleneck design.

Parameter counts. Table 2 lists the trainable parameters of every model. FisherCBM adds to the hard CBM only the $2K$ prototypes ($2KD$ parameters, 172K on CUB) and has no concept head; its readout cost matches CIBM's, whose mutual-information estimator is comparable in size. CEM and ProbCBM carry 16 embedding dimensions per concept into the head, which multiplies both the head and the per-concept generators (3.7–5.2M trainable parameters on CUB and AwA2).

Table 2: Trainable parameters (thousands) and bottleneck width of every model on the three datasets. All models share the frozen ViT-B/16 encoder (85.8M parameters, not counted); the trainable parameters are the adapter, the concept layer and the task head. $K = 112$, 85 and 6 concepts on CUB, AwA2 and CelebA. The end-to-end classifier has no bottleneck; its adapter output is pooled directly into the head.

	CUB		AwA2		CelebA	
Model	params	width	params	width	params	width
End-to-end	894	–	825	–	797	–
Hard CBM	700	112	661	85	598	6
Soft CBM	700	112	661	85	598	6
Logit CBM	700	112	661	85	598	6
Hybrid CBM	749	162	702	135	648	56
CEM	3,706	1792	4,911	2720	761	96
ProbCBM	3,692	1792	5,200	2720	786	96
CIBM-B	786	112	726	85	602	6
CIBM-E	786	112	726	85	602	6
FisherCBM	786	112	726	85	602	6

B.3 Training

All models use AdamW with learning rate 10^{-3}, weight decay 10^{-4} (0 for ProbCBM), batch size 128, gradient-norm clipping at 1.0, and a cosine learning-rate schedule with 200 linear warm-up steps, stepped per optimizer step. Training runs for up to 300 epochs, with the checkpoint of highest validation task accuracy retained for test evaluation. Mixed precision is disabled. Runs use a single GPU; one CUB epoch takes about 17 s with the frozen encoder. We report mean and standard deviation over three seeds (0, 1, 42).

B.4 Intervention protocols

Interventions are evaluated on the test set at fractions $\{0, 0.1, \ldots, 1.0\}$ of the concepts. Two selection protocols are used. Under *per-concept* selection, a random subset of $\mathrm{round}(f \cdot K)$ individual concepts is drawn per image. Under *per-group* selection, a random subset of whole attribute groups is drawn (28 groups on CUB), and every concept in a selected group is intervened on; this is the standard CUB protocol, since the attributes within a group are mutually exclusive and only a whole-group correction is meaningful. On CelebA, each of the six attributes is its own group, so the two protocols coincide. In the *correct* mode, selected concepts are set to their ground-truth value; in the *incorrect* mode of Havasi et al. (2022) they are set to $1 - c_i$. The intervention value is encoded per model: the bit for FisherCBM (the belief and gate are re-evaluated at the new state as in Section 3.5), $\{0, 1\}$ for the hard, fuzzy, and hybrid CBMs and CIBM, the corresponding endpoint embedding for CEM and ProbCBM, and the logit extremes for the logit CBM.

Summary statistics. Let $I(0), I(1), \ldots, I(n)$ be the test accuracies of an intervention sweep, $I(0)$ without intervention and $I(k)$ with k of the n groups (or fractions) intervened. Following Galliamov et al. (2026) we report the area under the curve, $\mathrm{AUC} = \frac{1}{n}\sum_{k=1}^{n} I(k)$, that is, the mean accuracy over the non-trivial intervention levels, expressed in percent like the accuracies it averages; the accuracy $I(n)$ with every group intervened (*full* in the tables), and the accuracy $I(\lfloor n/2 \rfloor)$ with half of the groups intervened (*half*); in the incorrect mode the latter separates models that have already reached chance from those that still read the label from the residual channel. In the correct mode, a higher AUC means that fewer corrections are needed to reach a given accuracy. In the incorrect mode, the AUC measures how much accuracy the head retains while the concepts it is

Table 3: Intervention summary (%, mean ± std over three seeds; per-group protocol). AUC, *half*, *full* and the two modes are defined in Appendix B.4; *half* is the accuracy with half of the groups intervened (14 of 28 on CUB). In the correct mode, higher is better; in the incorrect mode, lower is better.

		correct		incorrect		
Dataset	Model	AUC	full	AUC	half	full
CUB	FisherCBM	93.9 ± 0.1	100.0 ± 0.0	18.3 ± 0.1	0.1 ± 0.0	0.0 ± 0.0
	CEM	94.0 ± 0.3	99.6 ± 0.1	24.0 ± 2.1	6.3 ± 2.5	0.0 ± 0.0
	CEM w/o RandInt	87.9 ± 0.7	92.0 ± 0.8	50.2 ± 3.0	58.6 ± 6.4	5.2 ± 1.2
	Hard CBM	92.0 ± 0.2	100.0 ± 0.0	12.9 ± 0.1	0.2 ± 0.1	0.0 ± 0.0
	Soft CBM	90.0 ± 0.9	92.3 ± 2.2	10.6 ± 0.1	0.2 ± 0.1	0.0 ± 0.0
	Logit CBM	88.1 ± 1.0	89.9 ± 2.0	8.8 ± 0.2	0.1 ± 0.0	0.0 ± 0.0
	Hybrid CBM	90.3 ± 0.6	96.0 ± 0.7	18.2 ± 0.6	3.9 ± 0.3	0.0 ± 0.0
	ProbCBM	96.3 ± 0.1	100.0 ± 0.0	1.5 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	CIBM-E	90.9 ± 0.6	93.3 ± 1.3	10.9 ± 0.2	0.1 ± 0.0	0.0 ± 0.0
	CIBM-B	83.3 ± 1.7	79.4 ± 5.0	12.5 ± 0.4	0.4 ± 0.1	0.0 ± 0.0
AwA2	FisherCBM	98.6 ± 0.0	100.0 ± 0.0	29.6 ± 0.2	4.8 ± 0.3	0.0 ± 0.0
	CEM	98.4 ± 0.0	99.8 ± 0.0	48.0 ± 0.8	47.7 ± 2.8	4.5 ± 1.1
	CEM w/o RandInt	96.8 ± 0.1	97.6 ± 0.2	75.7 ± 4.0	89.4 ± 2.6	25.9 ± 10.6
	Hard CBM	98.4 ± 0.0	100.0 ± 0.0	25.4 ± 0.3	3.1 ± 0.1	0.0 ± 0.0
	Soft CBM	98.7 ± 0.0	100.0 ± 0.0	26.2 ± 0.1	3.3 ± 0.3	0.0 ± 0.0
	Logit CBM	98.7 ± 0.0	100.0 ± 0.0	28.3 ± 0.1	10.6 ± 0.1	0.0 ± 0.0
	Hybrid CBM	98.1 ± 0.0	99.7 ± 0.1	41.7 ± 0.4	35.4 ± 1.3	0.0 ± 0.0
	ProbCBM	99.5 ± 0.0	100.0 ± 0.0	1.9 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	CIBM-E	98.7 ± 0.0	100.0 ± 0.0	26.3 ± 0.1	3.1 ± 0.1	0.0 ± 0.0
	CIBM-B	98.8 ± 0.0	100.0 ± 0.0	25.9 ± 0.1	2.9 ± 0.3	0.0 ± 0.0
CelebA	FisherCBM	46.9 ± 2.2	58.4 ± 4.2	1.8 ± 0.9	0.0 ± 0.0	0.0 ± 0.0
	CEM	57.9 ± 0.1	73.8 ± 0.4	2.9 ± 0.1	0.4 ± 0.1	0.0 ± 0.0
	CEM w/o RandInt	52.9 ± 2.1	64.1 ± 4.0	10.7 ± 1.8	10.2 ± 3.2	0.4 ± 0.3
	Hard CBM	45.8 ± 1.3	59.0 ± 2.5	1.6 ± 0.4	0.0 ± 0.0	0.0 ± 0.0
	Soft CBM	47.7 ± 0.5	61.4 ± 0.3	0.9 ± 0.2	0.0 ± 0.0	0.0 ± 0.0
	Logit CBM	48.6 ± 1.0	59.0 ± 2.4	0.4 ± 0.1	0.0 ± 0.0	0.0 ± 0.0
	Hybrid CBM	43.8 ± 0.1	46.9 ± 0.2	34.0 ± 0.2	34.6 ± 0.2	30.2 ± 0.5
	ProbCBM	52.1 ± 0.1	63.7 ± 0.6	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	CIBM-E	47.5 ± 0.4	60.1 ± 1.3	1.1 ± 0.1	0.0 ± 0.0	0.0 ± 0.0
	CIBM-B	46.3 ± 1.0	57.5 ± 1.8	1.5 ± 0.3	0.0 ± 0.0	0.0 ± 0.0

given are wrong, so a lower value means that the prediction rests more exclusively on the concepts; the accuracy $I(n)$ in this mode is the leakage summary of Section 4.2, and a head that reads only the concepts should fall to chance.

B.5 Post-hoc Evaluation Protocols

Table 8 evaluates, post hoc, how much of the residual's token sensitivity lies off the support, and whether the penalty changed it. For every trained checkpoint (`best.pt` of the full model and of the no-gate ablation $g_i \equiv 1$, $\eta = 1$, each with $\beta \in \{0, 1\}$, Table 5, Table 6, Table 7) we run the test split once without interventions and compute, per image and concept, the token gradients $\nabla_{\mathbf{z}_n} s_i$ from the closed form of Theorem 3, with prototypes and hard state fixed as in the penalty. From these, we report, averaged over images and concepts:

- $\mathcal{L}^{\text{off}}_{\text{fisher}}$: the value of Eq (13), i.e. $\frac{1}{K}\sum_i g_i^2 \sum_n w_{n,i} \|\nabla_{\mathbf{z}_n} s_i\|^2$, the quantity the penalty minimizes; it is the same function that was logged during training (Figure 5), evaluated on the test set instead of on RandInt-intervened training batches.
- *off share*: $S_i^{\text{off}}/(S_i^{\text{off}} + S_i^{\text{on}})$ with $S_i^{\text{off}} = \sum_n w_{n,i} \|\nabla_{\mathbf{z}_n} s_i\|^2$ the gate-free off-support sensitivity of Proposition 2 and $S_i^{\text{on}} = \sum_{n \in \mathcal{S}_i} \|\nabla_{\mathbf{z}_n} s_i\|^2$ the sensitivity on the support; it measures which fraction of the residual's sensitivity the penalty can act on at all, independently of the gate.

- *off mass*: $\sum_{n \notin \mathcal{S}_i} \pi_{n,i}$, the share of the plan that lies off the support.
- *support*: the median of $|\mathcal{S}_i|$ over images and concepts.

The ratio column divides $\mathcal{L}^{\text{off}}_{\text{fisher}}$ of the $\beta = 1$ checkpoint by that of the $\beta = 0$ checkpoint trained with the same seed; the table gives the mean and range of the three ratios. All quantities are forward-pass expressions, and no model is modified.

In-box mass (Section 4.3). On the 1,000-image CUB test subset of Appendix C.3, we compute, for every image and concept, the share of the map's mass that falls on tokens whose centre lies inside the annotated bird bounding box, mapped to the 14×14 token grid under the evaluation crop. For FisherCBM, the map is the plan $\pi^{c_i^{\text{hard}}}_{\cdot,i}$ of the predicted state. For ProbCBM, the only baseline that weights tokens per concept, it is its attention $\mathrm{softmax}(\langle \mathbf{z}_n, \mathbf{q}_i \rangle / \sqrt{C})$ of its probabilistic embedding module, where $\mathbf{q}_i \in \mathbb{R}^C$ is the learned query of concept i (Kim et al., 2023). Two references are the uniform plan, whose in-box mass equals the box's share of the tokens, and a random plan supported on as many tokens as FisherCBM's plan for the same image and concept. For CEM and ProbCBM, the probe is fitted on the flattened embedding that their heads read (the mixture for CEM, the sample mean for ProbCBM).

Part-level pointing (Appendix C.3). For the 94 CUB concepts whose attribute family refers to an annotated part (bill, eye, forehead, crown, nape, throat, breast, belly, back, wing, tail, leg), we measure the distance in tokens from the argmax token of the map to the nearest visible keypoint of that part, over the same 1,000 images, and the fraction of cases within one and two tokens. Chance is the expectation over a uniformly random token; centre is the fixed central token of the grid.

Hidden-attribute probes (Appendix C.1). On CelebA, a logistic-regression probe (scikit-learn, $C = 1$) is fitted on the training split to predict each hidden attribute from a stated feature set and evaluated by accuracy and ROC-AUC on the test split.

C Experimental Results

C.1 Additional Results

Table 3 lists the areas under the intervention curves and the accuracies with every concept intervened for all models and datasets, and Figure 3 shows the curves on AwA2 and CelebA. Figure 4 repeats the RandInt comparison of Section 4.2 on all three datasets, with the same outcome. On AwA2 correct interventions bring every model to 100% except CEM (99.8) and the hybrid CBM (99.7); under incorrect interventions, the CBM family and CIBM lie between 25.4 (hard CBM) and 28.3 (logit CBM), FisherCBM just above them at 29.6, and the models with a wider channel well above: the hybrid CBM at 41.7, CEM at 48.0 and CEM without RandInt at 75.7; ProbCBM (1.9) collapses under its anchor-replacement intervention, as on CUB. On CelebA, the concepts do not determine the class, so no model approaches 100% under correct interventions; the models that reach the highest accuracy there (CEM, 73.8) are those whose bottleneck also carries the two hidden attributes (Section 4.1). The hybrid CBM is the only model that retains accuracy (30.2%) with all concepts set to the wrong value, because its unconstrained side channel carries most of the class information.

What the residual carries on CelebA. Table 4 probes each part of the bottleneck for the two hidden attributes (protocol in Appendix B.5). Heavy Makeup is largely predictable from the concepts themselves (AUC 0.945 from the six ground-truth bits), so every channel scores 0.93–0.97. Wavy Hair is not (0.712), and here the channels separate: FisherCBM's bits give 0.702, its residual 0.726, and the full bottleneck 0.730, comparable to a probability channel (0.720), whereas ProbCBM's Gaussian embedding gives 0.791 and CEM's mixture embedding 0.888. The bounded residual thus passes a little hidden-attribute information beyond the concept bits; the unbounded embeddings pass progressively more, in the order of their task accuracy (33.2%, 35.3%, 39.6%), and this is the sense in which the residual's leakage is bounded. The bounded residual thus passes a little hidden-attribute information beyond the concept bits; CEM's unbounded embedding passes most of it. This is the difference between the 33.2% task accuracy of FisherCBM and the 39.6% of CEM, and the sense in which the residual's leakage is bounded.

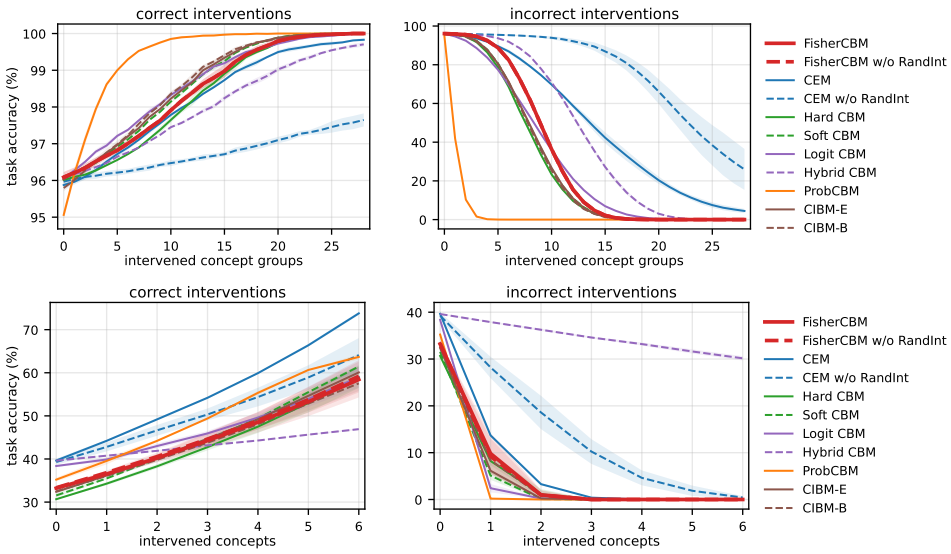


Figure 3: Test-time interventions on AwA2 (top; per-group protocol with the predicate groups of Zarlenga et al. (2022)) and CelebA (bottom; each of the six concepts is its own group), mean and standard deviation over three seeds, in the same layout as Figure 2.

Table 4: Hidden-attribute information in the bottleneck on CelebA: test ROC-AUC (and accuracy, %) of a logistic-regression probe fitted on the training split to predict each hidden attribute from the stated features (mean over seeds $\{0, 1, 42\}$; std $\leq$ 0.007 AUC throughout; protocol in Appendix B.5). Majority-class accuracy is 62.5% for Heavy Makeup and 69.6% for Wavy Hair.

		Heavy Makeup		Wavy Hair	
Model	features	AUC	acc	AUC	acc
–	ground-truth concepts (6 bits)	0.945	89.7	0.712	68.1
FisherCBM	hard bits c_i^{hard}	0.925	87.0	0.702	67.4
	residual $\eta g_i s_i$	0.958	88.9	0.726	70.1
	bits + residual (what the head sees)	0.958	88.0	0.730	70.8
	concept probabilities $\hat{p}_i$	0.952	88.1	0.720	69.1
CEM	concept probabilities $\hat{p}_i$	0.960	89.1	0.719	68.9
	mixture embedding (what the head sees)	0.967	89.9	0.888	82.8
ProbCBM	concept probabilities $\hat{p}_i$	0.937	87.0	0.700	68.5
	Gaussian embedding (what the head sees)	0.945	87.9	0.791	78.2

C.2 Ablation Study

Table 5, Table 6, and Table 7 change one model element at a time on CUB, AwA2 and CelebA. Four findings hold on every dataset. Reading both evidences from a single transport lowers concept accuracy (98.0 to 97.1% on CUB, 89.1 to 88.8% on CelebA), as the bound on the one-sided belief in absence (Appendix A.1.3) predicts, and on CelebA it also costs 1.6 points of task accuracy. Replacing the unbalanced mass by $m_i \equiv 1(\tau \to \infty)$, which turns IUCT into per-concept softmax attention, costs 3.2 points of task and 0.6 of concept accuracy on CUB at the default temperature and 4.2 points with θ divided by the median mass, so the loss is not a temperature artefact: the mass varies across concepts and images by orders of magnitude, which no single temperature can restore. The mechanism is visible in the readout: without the mass, the belief in the predicted state falls (mean gate 0.79 against 0.97) and the gate passes less of the residual, consistent with the lower incorrect-mode area (16.0 against 18.3). On CelebA, where six coarse attributes are read from large regions and the median mass is 100, the same change costs nothing at the default temperature (+1.3 task, +0.8 concept); with θ divided by the median mass, it costs 2.0 points, which follows the dataset's preference for coarser temperatures ($\theta = D$ gains 1.2 points in Table 7, $\theta = 0.01D$ loses

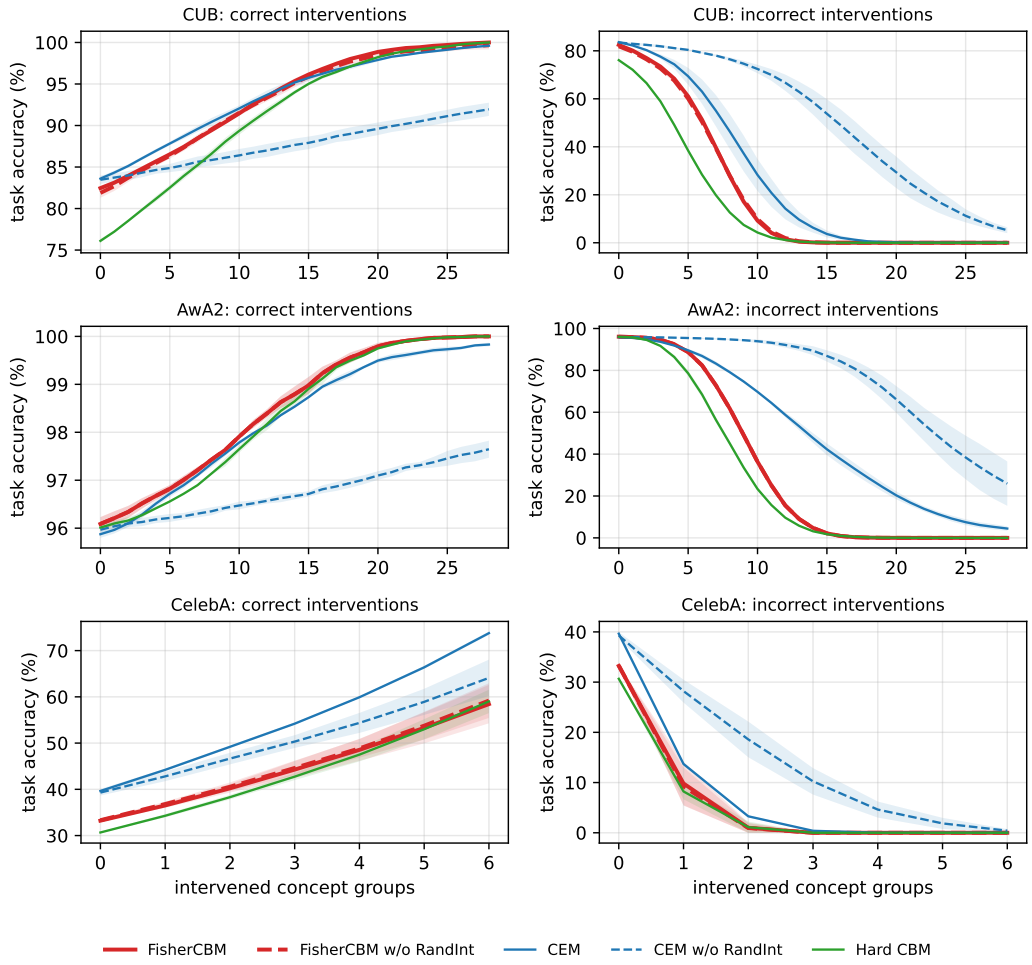


Figure 4: Effect of RandInt on the intervention curves (mean and standard deviation over three seeds). The curves of FisherCBM with and without RandInt coincide on every dataset and in both modes; CEM without RandInt responds less to correct interventions and retains accuracy under incorrect ones. The hard CBM is shown for reference.

0.2) rather than the absence of the mass. On AwA2, whose class-level predicates match most of the animal (median mass 4300), $m_i \equiv 1$ changes accuracy by at most 0.2 points with or without the matched temperature. The mass carries how many tokens support a state (Eq. (5)); this matters when concepts are fine-grained and numerous, and is redundant with the direction when they are few, coarse, or class-level. Interventions remain monotone in every case. Removing RandInt leaves accuracy and both intervention curves within seed variance (Figure 4). Raising η from 0.4 to 1 leaves accuracy and the correct-mode curves unchanged and raises the incorrect-mode area by 0.2 to 1.3 points. Removing the gate as well ($g_i \equiv 1$ at $\eta = 1$) costs 0.8 points of task and 0.4 of concept accuracy on CUB and nothing on AwA2 and CelebA, and raises the incorrect-mode area by a further 0.3 points on CUB; the gate's contribution to accuracy is small, and its main effect is on the plans (below). Finally, $\beta = 0$ and $\beta = 1$ are indistinguishable on every dataset and every metric. Three facts account for this. First, by Theorem 3, the residual's sensitivity to a token is proportional to that token's share of the plan, so off-support tokens, whose share is below $1/N$, contribute little to the residual whether or not they are penalized; the penalty acts only on the band $\pi \in (0, 1/N)$, and the trained plans, with a median support of 10–16 tokens, leave few tokens in it. Second, the gate bounds the amplitude of the residual by ηg_i and closes on uncertain concepts, and the incorrect-mode curves show that the head recovers nothing from the residual once the identities are wrong. Third, the token weight $(1 - N\pi)_+\pi^2$ is at most $4/(27N^2) \approx 4 \times 10^{-6}$, so at $\beta = 1$, the value of Eq. (13) is of order 10^{-7}, six orders below the task loss, and its gradient is negligible

next to the other losses. Table 8 confirms this directly: evaluated post hoc on the trained checkpoints with the closed form of Theorem 3, on the test sets without interventions, the value of Eq. (13) at $\beta = 1$ relative to $\beta = 0$ is 0.87 on CUB, 0.99 on AwA2 and 1.00 on CelebA; per seed the ratio ranges from 0.79 to 1.43, while the spread across seeds at fixed β is a factor of 4–5, so the penalty did not measurably reduce its own target. The same table shows why: already at $\beta = 0$, off-support tokens account for between 2×10^{-4} and 1×10^{-3} of the residual's total token sensitivity and carry 3–7% of the plan mass, so there is almost nothing left for the penalty to act on. The gate contributes to this confinement: removing it ($g_i \equiv 1$, $\eta = 1$) raises the off-support share at $\beta = 0$ by a factor of 1.7 to 3.4 and the off-support mass from 3–7% to 6–10%, because a concept whose residual barely reaches the head receives no task gradient that would spread its plan. The no-gate model is thus the regime in which the penalty has something to act on, yet even there $\beta = 1$ leaves the off-support share unchanged (paired ratios 0.93–1.00): checkpoints trained with and without the penalty from the same seed agree to three digits, so at $\beta = 1$ the penalty's gradient lies below the numerical noise of training. Making it bind requires a weight of the order of the bound $27N^2/4$: at $\beta = 10^5$ the off-support share falls to 0.36 of its $\beta = 1$ value on CUB and 0.29 on CelebA and the median support from 15 and 12 tokens to 7 and 4, with task accuracy unchanged on CUB (82.5) and -0.3 on CelebA; at $\beta = 10^6$ it falls to 0.14 and 0.13 with supports of 3 tokens at a cost of 0.3 and 0.8 points (Tables 5, 7 and 8). On CelebA, the penalty also removes most of the seed-to-seed spread of the off-support share. The sharper plans do not localize better, however: on CUB the in-box mass falls from 91% to 84%, and the part-level hit rate stays at 22–24%, because the plans of absent concepts, once supported on a few tokens, more often leave the bird. The penalty is thus an effective control on where the residual is read from, whose default weight was too small to act; we keep $\beta = 1$ in the main experiments because the transport and the gate already confine the residual there.

In detail: $\theta = D$ costs 1.1 points of task accuracy on CUB and is neutral on AwA2 (+0.1) and beneficial on CelebA (+1.2 task, +3.7 correct-mode area), so the softplus saturation it causes hurts only where many fine-grained concepts must be separated; one-sided evidence lowers concept accuracy by 0.9 points on CUB, 0.1 on AwA2 and 0.3 on CelebA, and task accuracy by 1.6 points on CelebA; $\beta = 0$ and the removal of RandInt are within seed variance of the full model on every dataset; $\eta = 1$ raises the incorrect-mode area by 0.2 to 1.3 points. The $\eta = 1$ row lies outside the hypothesis of Theorem 4, which needs $\eta < \frac{1}{2}$ so that the residual cannot overturn the identity for any score $|s_i| \leq 1$; empirically $|s_i|$ does not exceed 0.28 on CUB ($D = 768$, cosine between a token aggregate and a prototype), so the residual stays below 0.28 in magnitude at $\eta = 1$ and the guarantee holds with margin. The slightly larger incorrect-mode area is consistent with a residual that carries a little more information when its amplitude is allowed to grow. Figure 5 shows the logged value of Eq. (13) over training with $\beta = 1$ (mean and standard deviation over three seeds). After a rise during the first epochs, when the prototypes move and the planes are still broad, it decreases monotonically by a factor of 2.5 on CUB, 10 on AwA2, and 3.5 on CelebA, and stays between 5×10^{-8} and 1.3×10^{-6} throughout. The $\beta = 0$ runs did not log this quantity during training, but the post-hoc comparison in Table 8 shows that they end at the same value, so the decrease is due to the concept loss sharpening the plans, not to the penalty.

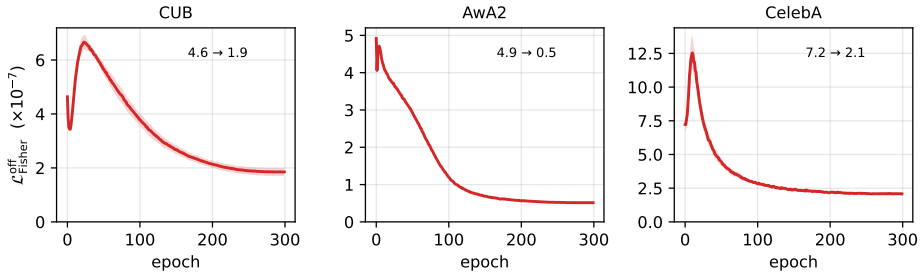


Figure 5: Value of the off-support Fisher penalty Eq. (13) during training with $\beta = 1$, in units of 10^{-7} (mean and standard deviation over three seeds). The annotation gives the first and last values.

Table 5: Ablation on CUB (%, mean $\pm$ std over three seeds). Each row changes one element of the full model in the first row. Intervention AUCs are defined as in Table 3. Here, full means $\theta = 0.1 \cdot D$, $\eta = 0.4$, $\beta = 1$, two-sided evidence, RandInt.

Variant	task	concept	correct AUC	incorrect AUC
FisherCBM (full)	82.5 ± 0.2	98.0 ± 0.1	93.9 ± 0.1	18.3 ± 0.1
$\beta = 0$ (no Fisher penalty)	82.2 ± 0.2	97.9 ± 0.1	93.9 ± 0.0	18.2 ± 0.1
$\beta = 10^4$	81.4 ± 0.8	97.6 ± 0.3	93.9 ± 0.2	17.8 ± 0.3
$\beta = 10^5$	82.5 ± 0.4	98.0 ± 0.1	94.0 ± 0.1	18.3 ± 0.1
$\beta = 10^6$	82.2 ± 0.1	97.8 ± 0.1	94.0 ± 0.1	17.9 ± 0.2
$\theta = 0.01 \cdot D$	82.1 ± 0.4	97.9 ± 0.1	94.0 ± 0.2	17.9 ± 0.1
$\theta = 1 \cdot D$	81.4 ± 0.4	97.7 ± 0.1	93.7 ± 0.1	17.7 ± 0.1
$\eta = 1$ (outside the range of Theorem 4)	82.2 ± 0.0	98.0 ± 0.0	93.8 ± 0.0	19.1 ± 0.2
one-sided evidence (single transport)	82.2 ± 0.5	97.1 ± 0.1	94.1 ± 0.1	17.0 ± 0.1
$m_i \equiv 1$ ($\tau \to \infty$, per-concept softmax)	79.3 ± 1.0	97.4 ± 0.2	93.5 ± 0.2	16.0 ± 0.5
$m_i \equiv 1$, θ matched to the median mass	78.3 ± 0.8	97.1 ± 0.1	93.0 ± 0.2	15.9 ± 0.3
no RandInt	81.9 ± 0.6	97.6 ± 0.4	93.7 ± 0.4	18.2 ± 0.1
no gate ($g_i \equiv 1$, $\eta = 1$)	81.4 ± 0.1	97.6 ± 0.0	93.8 ± 0.1	19.4 ± 0.1

Table 6: Ablation on AwA2 (%, mean $\pm$ std over three seeds), same variants as Table 5. Here, full means $\theta = 0.1 \cdot D$, $\eta = 0.4$, $\beta = 1$, two-sided evidence, RandInt.

Variant	task	concept	correct AUC	incorrect AUC
FisherCBM (full)	96.1 ± 0.1	99.2 ± 0.0	98.6 ± 0.0	29.6 ± 0.2
$\beta = 0$ (no Fisher penalty)	96.0 ± 0.1	99.2 ± 0.0	98.6 ± 0.0	29.5 ± 0.2
$\theta = 0.01 \cdot D$	96.1 ± 0.1	99.2 ± 0.0	98.6 ± 0.0	29.6 ± 0.1
$\theta = 1 \cdot D$	96.2 ± 0.0	99.3 ± 0.0	98.6 ± 0.0	29.4 ± 0.0
$\eta = 1$ (outside the range of Theorem 4)	96.1 ± 0.1	99.2 ± 0.0	98.6 ± 0.1	30.9 ± 0.3
one-sided evidence (single transport)	96.1 ± 0.1	99.1 ± 0.1	98.7 ± 0.1	28.9 ± 0.5
$m_i \equiv 1$ ($\tau \to \infty$, per-concept softmax)	96.0 ± 0.1	99.2 ± 0.0	98.7 ± 0.0	28.9 ± 0.1
$m_i \equiv 1$, θ matched to the median mass	95.9 ± 0.1	99.2 ± 0.0	98.6 ± 0.1	29.1 ± 0.1
no RandInt	96.1 ± 0.1	99.2 ± 0.0	98.6 ± 0.0	29.6 ± 0.2
no gate ($g_i \equiv 1$, $\eta = 1$)	96.1 ± 0.1	99.2 ± 0.0	98.6 ± 0.0	30.8 ± 0.1

Table 7: Ablation on CelebA (%, mean $\pm$ std over three seeds), same variants as Table 5. Here, full means $\theta = 0.1 \cdot D$, $\eta = 0.4$, $\beta = 1$, two-sided evidence, RandInt.

Variant	task	concept	correct AUC	incorrect AUC
FisherCBM (full)	33.2 ± 0.2	89.1 ± 0.1	46.9 ± 2.2	1.8 ± 0.9
$\beta = 0$ (no Fisher penalty)	33.3 ± 0.2	89.1 ± 0.0	46.8 ± 2.1	1.8 ± 0.9
$\beta = 10^4$	33.2 ± 0.5	89.1 ± 0.2	48.8 ± 1.6	1.1 ± 0.5
$\beta = 10^5$	32.9 ± 0.1	89.2 ± 0.1	48.4 ± 1.4	1.1 ± 0.6
$\beta = 10^6$	32.4 ± 0.2	89.1 ± 0.1	48.8 ± 0.2	0.8 ± 0.1
$\theta = 0.01 \cdot D$	33.0 ± 0.2	88.9 ± 0.1	49.1 ± 0.3	0.9 ± 0.1
$\theta = 1 \cdot D$	34.4 ± 1.4	89.3 ± 0.1	50.6 ± 1.6	0.8 ± 0.0
$\eta = 1$ (outside the range of Theorem 4)	33.5 ± 0.5	89.2 ± 0.1	47.8 ± 0.8	2.0 ± 0.4
one-sided evidence (single transport)	31.6 ± 0.2	88.8 ± 0.1	44.9 ± 1.9	2.3 ± 0.7
$m_i \equiv 1$ ($\tau \to \infty$, per-concept softmax)	34.5 ± 1.3	89.9 ± 0.1	48.3 ± 2.7	1.7 ± 0.9
$m_i \equiv 1$, θ matched to the median mass	31.2 ± 0.2	88.2 ± 0.2	46.9 ± 1.1	1.2 ± 0.5
no RandInt	33.3 ± 0.3	89.1 ± 0.1	47.3 ± 2.0	1.7 ± 0.8
no gate ($g_i \equiv 1$, $\eta = 1$)	33.3 ± 0.5	89.2 ± 0.1	47.5 ± 1.5	2.3 ± 0.6

Table 8: Off-support sensitivity of the residual on the test sets, evaluated post hoc on the trained checkpoints with the closed form of Theorem 3 and no interventions (mean $\pm$ std over seeds $\{0, 1, 42\}$; quantities defined in Appendix B.5). *gate* rows are the full model ($\eta = 0.4$); *no gate* rows set $g_i \equiv 1$ and $\eta = 1$. The last column is $\mathcal{L}^{\text{off}}_{\text{fisher}}$ at $\beta = 1$ divided by its value at $\beta = 0$, or, for $\beta \geq 10^5$, at $\beta = 1$, for the same seed and gate setting.

Dataset	gate	β	$\mathcal{L}^{\text{off}}_{\text{fisher}}$ ($\times 10^{-7}$)	off share ($\times 10^{-4}$)	off mass (%)	support (tokens)	ratio
CUB	gate	0	3.3 ± 1.4	5.7 ± 2.7	5.6 ± 2.2	16	–
		1	2.8 ± 0.9	4.9 ± 1.6	5.0 ± 1.3	15	0.90 (0.79-1.00)
		10^5	1.1 ± 0.2	1.7 ± 0.4	3.0 ± 0.6	7	0.40 (0.35–0.48)
		10^6	0.4 ± 0.2	0.6 ± 0.2	1.8 ± 0.4	3	0.16 (0.07–0.22)
	no gate	0	14.3 ± 1.6	19.6 ± 1.9	9.5 ± 0.8	21	–
		1	13.9 ± 1.7	18.6 ± 2.1	9.2 ± 0.5	20	0.98 (0.93–1.00)
AwA2	gate	0	1.3 ± 1.1	2.4 ± 2.4	2.8 ± 3.0	11	–
		1	1.3 ± 0.8	2.2 ± 1.6	2.7 ± 2.0	10	1.13 (0.84–1.43)
	no gate	0	3.6 ± 0.5	6.1 ± 1.1	5.6 ± 0.7	19	–
		1	3.5 ± 0.5	5.9 ± 1.0	5.5 ± 0.5	19	0.99 (0.96–1.00)
CelebA	gate	0	6.6 ± 3.9	10.1 ± 7.3	7.0 ± 3.3	11	–
		1	6.6 ± 3.9	10.0 ± 7.2	7.0 ± 3.3	11	1.00 (1.00–1.01)
		10^5	1.1 ± 0.6	2.2 ± 1.4	4.2 ± 2.0	4	0.29 (0.15–0.50)
		10^6	0.4 ± 0.1	0.8 ± 0.2	2.9 ± 0.6	3	0.13 (0.04–0.23)
	no gate	0	44.2 ± 15.3	17.2 ± 7.4	8.0 ± 2.6	12	–
		1	44.2 ± 15.3	17.2 ± 7.4	8.0 ± 2.6	12	1.00 (1.00–1.00)

Table 9: Object-level localization on CUB: share of the map's mass inside the bird's bounding box, over 1,000 test images and all 112 concepts (protocol in Appendix B.5). *present* / *absent* split the concepts by the model's prediction; *argmax in box* is the fraction of maps whose largest token lies inside the box; *support* is the median number of tokens with weight above $1/N$. FisherCBM: mean $\pm$ std over seeds $\{0, 1, 42\}$; ProbCBM: seed 42. The box covers 52.7% of the tokens.

Map	in-box mass (%)	present (%)	absent (%)	argmax in box (%)	support
FisherCBM plan $\pi^{c_i^{\text{hard}}}_{\cdot,i}$	91.0 ± 4.5	82.0 ± 1.8	92.6 ± 5.1	92.5 ± 3.8	15
ProbCBM concept attention	59.0	59.0	59.1	63.9	93
Random plan, matched support	52.6	–	–	–	15
Uniform plan	52.7	–	–	–	196

C.3 Additional Transport-Plan Visualizations

This section shows transport plans on all three datasets. Panels follow the convention of the main text: each shows the allocation $\pi^{c_i^{\text{hard}}}_{\cdot,i}$ of the predicted state, upsampled by Eq. (14); dimmed pixels are off-support and the cross marks the argmax token. All images shown are classified correctly, and every concept shown is predicted correctly with belief above 0.95. CUB is the only dataset with part annotations, so the quantitative localization analysis of Section 4.3 is restricted to it. On CelebA and AwA2, the plans are discussed qualitatively, and the two datasets illustrate two kinds of concept supervision. Support sizes quoted below are medians in tokens out of $N = 196$, over the selected CUB examples and over 80 evenly spaced test images on CelebA and AwA2.

C.3.1 CUB

Figure 6, Figure 7, Figure 8, Figure 9 show further CUB test images together with the transport plans of one present concept and one absent concept of the same family; circles mark the annotated part keypoints of that family. Table 10 lists, for each panel, the belief in the predicted state, the number of support tokens ($\pi_{n,i} > 1/N$out of $N = 196$), and the distance in tokens from the argmax token to the nearest annotated part; the panels were selected to show part-level agreement. Over the

1,000-image subset and the 94 concepts whose attribute family refers to an annotated part (protocol in Appendix B.5), the argmax token lies a median 3.5 ± 0.3 tokens from the nearest keypoint of that part and within one token of it in 9.8 ± 1.2% of the cases ($26 \pm 3\%$ within two), against 6.0% (20.7%) for the fixed image-centre token, 2.1% (9.1%) for ProbCBM's attention and 1.7% (6.6%) for a random token. The plans therefore locate the bird reliably (Table 9) but the specific part only weakly; in the selected panels the support is a small region on the bird for present and absent concepts alike, and where a background token enters the support, as in Figure 1 (c), it carries a small share of the mass.

C.3.2 CELEBA

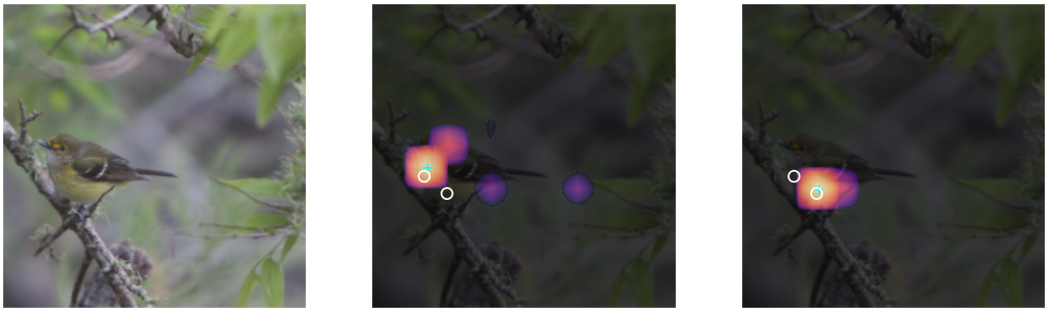
The task uses six attributes as concepts (Attractive, Male, High Cheekbones, Mouth Slightly Open, Smiling, Wearing Lipstick), each annotated per image. Figure 10, Figure 11 show eight test images with the plans of two present concepts and one absent concept each. Attributes that name a facial region are read off that region: the support of *Mouth Slightly Open* (median size 2 tokens) sits on the mouth, that of *Smiling* (8 tokens) covers the mouth corners and nasolabial folds, that of *High Cheekbones* (6 tokens) splits into two blobs on the cheekbones, and that of *Wearing Lipstick* (11 tokens) covers the lips. *Male*, which has no single location, is read from the beard and mouth region on men (9 tokens) and, when absent, from the eyes and brows (12 tokens); *Attractive*, a holistic attribute, has the most diffuse support (32 tokens). The plans of absent concepts are about as compact as those of present ones and lie on the region where the attribute would appear or on a correlated cue (absent *Wearing Lipstick* is read from the eyes and forehead, where make-up co-occurs), not on the background. The readout therefore does not impose locality; the support is compact exactly when the concept is local.

C.3.3 AWA2

The 85 predicates of AwA2 are annotated per class, not per image: every image of a class carries the same attribute vector, whether or not the attribute is visible (a dolphin's tail, a giraffe's hooves). The supervision thus ties a concept to a species rather than to an image region, and any cue that identifies the species is a valid source of evidence. Figure 12, Figure 13 show eight test images with three present predicates each; the selection is restricted to predicates that name a visible property (colours, textures, body parts, body plan) and excludes behavioural and habitat predicates. Two patterns appear. Predicates with a distinctive appearance in the image localize: *horns* on the cow and the ox, *patches* on the giraffe, *tusks* on the rhinoceros, *hooves* and *quadrapedal* on the legs of the deer and the horse. Predicates whose referent is not visible are read from the species context instead: the dolphin's *tail* is read from the water, and *hairless* and *toughskin* from the animal's body as a whole. Supports remain small (median 6 tokens for present predicates), but the transported mass is an order of magnitude larger than on CUB (median $m_i \approx 4300$ against ≈ 220 on CUB and ≈ 100 on CelebA), which indicates that many tokens match the prototype: with class-level labels the prototypes converge to species-level appearance shared across the body rather than to a part. The plan reports where the model's evidence is; under class-level supervision that evidence is legitimately contextual, and the visualization exposes it rather than hiding it. The intervention guarantees of Section 3.5 do not depend on grounding and hold on AwA2 as on the other datasets (Section 4.2), but the AwA2 concepts should not be read as detectors of the named body part.



Row 1: White throated Sparrow. *bill length*: shorter than head (present, middle) vs. about the same as head (absent, right).



Row 2: White eyed Vireo. *underparts color*: yellow (present, middle) vs. grey (absent, right).



Row 3: Sage Thrasher. *breast color*: white (present, middle) vs. black (absent, right).

Figure 6: Transport plans on CUB test images (1). Left: input; middle: plan of a present concept; right: plan of an absent concept of the same family.

Table 10: Statistics of the panels in Figure 6, Figure 7, Figure 8, Figure 9. The panels were selected for part-level agreement; population statistics are given in Appendix C.3.1. b is the belief in the predicted state (b_i^1 for present, b_i^0 for absent), $|\mathrm{supp}|$ the number of support tokens out of $N = 196$, and d the distance in tokens from the argmax token to the nearest annotated part.

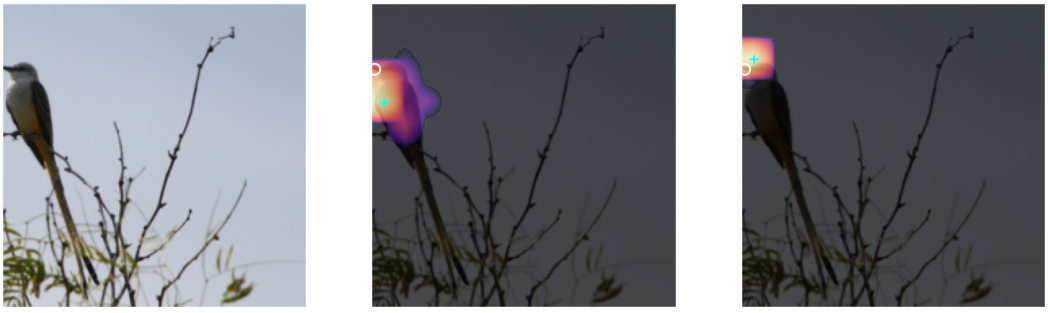
			present concept				absent concept			
#	Species	Family	value	b	$\lvert\mathrm{supp}\rvert$	d	value	b	$\lvert\mathrm{supp}\rvert$	d
1	White throated Sparrow	bill length	shorter than head	1.00	12	0.3	about the same as head	1.00	6	0.3
2	White eyed Vireo	underparts color	yellow	0.99	6	0.4	grey	1.00	2	0.2
3	Sage Thrasher	breast color	white	1.00	8	0.7	black	1.00	6	0.3
4	Tree Sparrow	wing color	brown	1.00	8	0.5	white	1.00	7	0.5
5	Swainson Warbler	breast pattern	solid	1.00	3	0.3	multi colored	0.99	10	1.2
6	Scissor tailed Flycatcher	bill color	black	1.00	8	1.6	buff	1.00	2	0.6
7	Blue winged Warbler	belly pattern	solid	1.00	7	0.5	multi colored	1.00	4	1.7
8	Heermann Gull	upperparts color	grey	1.00	11	0.8	buff	1.00	8	1.8
9	Brown Thrasher	underparts color	white	1.00	7	0.4	grey	1.00	6	0.4
10	Chipping Sparrow	wing color	brown	1.00	7	0.5	grey	1.00	5	0.8
11	Horned Puffin	breast color	white	0.98	6	1.2	brown	1.00	8	0.4
12	Prothonotary Warbler	bill color	black	1.00	10	1.4	buff	1.00	6	1.3



Row 1: Tree Sparrow. *wing color*: brown (present, middle) vs. white (absent, right).



Row 2: Swainson Warbler. *breast pattern*: solid (present, middle) vs. multi colored (absent, right).



Row 3: Scissor tailed Flycatcher. *bill color*: black (present, middle) vs. buff (absent, right).

Figure 7: Transport plans on CUB test images (2). Left: input; middle: plan of a present concept; right: plan of an absent concept of the same family.



Row 1: Blue winged Warbler. *belly pattern*: solid (present, middle) vs. multi colored (absent, right).



Row 2: Heermann Gull. *upperparts color*: grey (present, middle) vs. buff (absent, right).



Row 3: Brown Thrasher. *underparts color*: white (present, middle) vs. grey (absent, right).

Figure 8: Transport plans on CUB test images (3). Left: input; middle: plan of a present concept; right: plan of an absent concept of the same family.



Row 1: Chipping Sparrow. *wing color*: brown (present, middle) vs. grey (absent, right).



Row 2: Horned Puffin. *breast color*: white (present, middle) vs. brown (absent, right).



Row 3: Prothonotary Warbler. *bill color*: black (present, middle) vs. buff (absent, right).

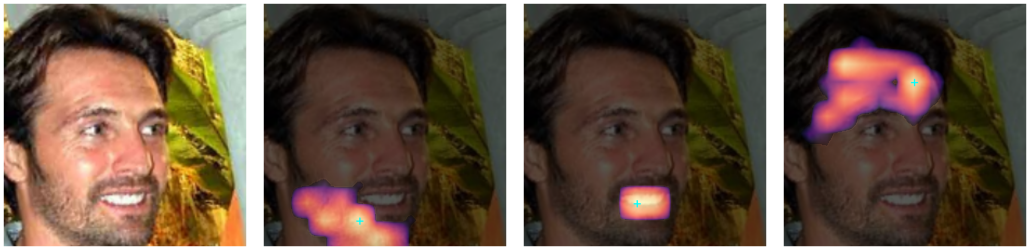
Figure 9: Transport plans on CUB test images (4). Left: input; middle: plan of a present concept; right: plan of an absent concept of the same family.



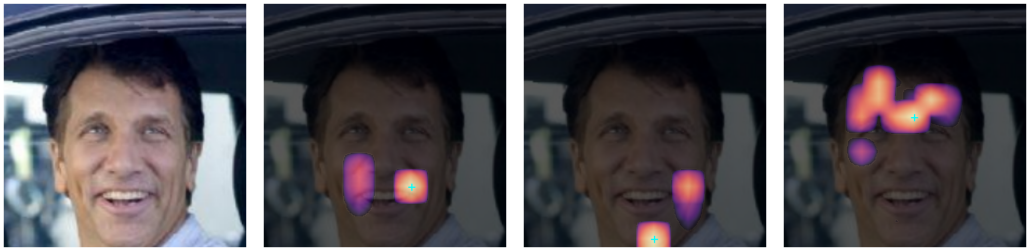
Row 1: *Smiling* (present), *Mouth Slightly Open* (present), *Male* (absent).



Row 2: *Mouth Slightly Open* (present), *Smiling* (present), *Male* (absent).



Row 3: *Male* (present), *Mouth Slightly Open* (present), *Wearing Lipstick* (absent).

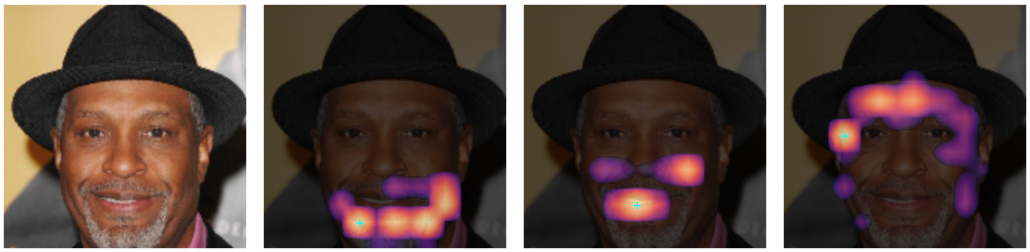


Row 4: *Smiling* (present), *Male* (present), *Wearing Lipstick* (absent).

Figure 10: Transport plans on CelebA test images (1). Left: input; then the plans of two present concepts and one absent concept, named under each row.



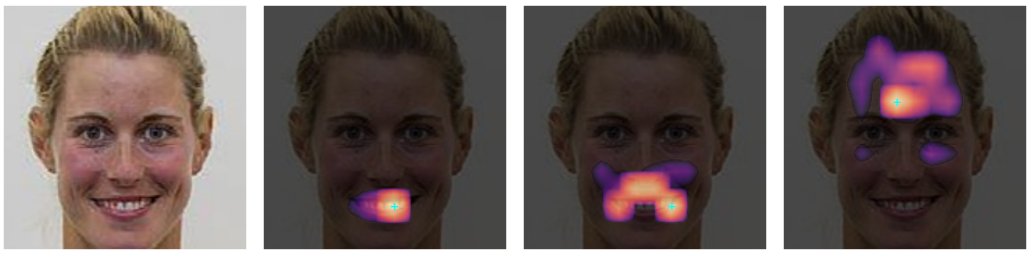
Row 1: *Mouth Slightly Open* (present), *Smiling* (present), *Male* (absent).



Row 2: *Male* (present), *Smiling* (present), *Wearing Lipstick* (absent).



Row 3: *Male* (present), *Mouth Slightly Open* (present), *Wearing Lipstick* (absent).



Row 4: *Mouth Slightly Open* (present), *Smiling* (present), *Wearing Lipstick* (absent).

Figure 11: Transport plans on CelebA test images (2). Left: input; then the plans of two present concepts and one absent concept, named under each row.



Row 1: cow. *quadrapedal*, *black*, *horns* (all present).



Row 2: ox. *quadrapedal*, *big*, *horns* (all present).



Row 3: giraffe. *horns*, *tail*, *patches* (all present).



Row 4: rhinoceros. *tusks*, *hairless*, *toughskin* (all present).

Figure 12: Transport plans on AwA2 test images (1). Left: input; then the plans of three present predicates, named under each row. Predicates are annotated per class, so a present predicate need not be visible in the image.



Row 1: deer. *quadrapedal*, *lean*, *hooves* (all present).



Row 2: horse. *furry*, *quadrapedal*, *gray* (all present).



Row 3: tiger. *quadrapedal*, *pads*, *paws* (all present).



Row 4: dolphin. *tail*, *toughskin*, *hairless* (all present).

Figure 13: Transport plans on AwA2 test images (2). Left: input; then the plans of three present predicates, named under each row. Predicates are annotated per class, so a present predicate need not be visible in the image.