# VLM-GAUGE: PRICING VISION-LANGUAGE SUPERVISION FOR CAUSAL REPRESENTATION LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Weakly supervised causal representation learning ties each latent variable to a labeled attribute and assumes those labels exist; because they come from human annotators, published experiments on real images supervise four to six attributes at most. We replace the human annotator with a vision-language model, which labels any attribute that can be named at negligible cost and so removes that limit, and ask what the substitution costs the representation. Calibrating the model on 500 human-labeled images per attribute yields its Youden index $J$, and we show, in theory and in experiments with DEAR on four to twelve CelebA attributes, that this index prices a free label before any training: a label of index $J$ carries roughly the signal of $J^2$ human labels, and annotator labels on 17,000 images outperform human labels on 4,000. The gap to human labels narrows from four attributes to twelve, as the human-label score falls and both annotators' scores hold, and persists for two annotators, for CausalVAE and SCM-VAE, and for a second attribute set. Three further findings follow. Label noise stays local: replacing one attribute's labels with the model's costs that attribute up to 31 points without significantly harming any other, and removing the weakest attributes' labels does not help the rest. Seven correction rules, among them forward correction, posterior relabeling, learned loss weights and human labels spent where the annotator is weakest, do not improve on the uncorrected labels. Errors from a real annotator cost less than simulated errors at matched rates. Finally, the scores the existing literature reports are underdetermined: for CausalVAE and SCM-VAE on Pendulum, Flow and CelebA, six scoring conventions that neither paper states move MIC on the same trained models by up to 71 points. We release the calibration protocol, the convention audit and the scores of every run.

## 1 INTRODUCTION

Causal representation learning seeks latent variables that correspond to the causal factors of the data (Schölkopf et al., 2021). Because unsupervised disentanglement is not identifiable without an inductive bias (Locatello et al., 2019), the methods that learn causal representations of real images supply one as labels, tying each supervised latent variable to a named attribute: CausalVAE (Yang et al., 2021), SCM-VAE (Komanduri et al., 2022), DEAR (Shen et al., 2022) and CFI-VAE (Li et al., 2024). Each method assumes these labels exist, and because they come from human annotators, published experiments on real images supervise four attributes, six for DEAR: annotation bounds how many attributes a causal representation can name.

A vision-language model can label any attribute that can be asked about in words, at the cost of one forward pass per image, and we call such labels *free*. On CelebA (Liu et al., 2015), such a model agrees with the human annotations 79.5% of the time at under 1% of their cost (Lu & Zhong, 2024), but its labels are not uniformly good. Measured on 500 human labels per attribute, the stronger of our two annotators reaches a Youden index (Youden, 1950), one minus its false positive and false negative rates, of 0.96 on MALE and comes within 0.10 of chance on four of twelve attributes (Figure 1a). To our knowledge, what labels of this kind cost a causal representation, and whether that cost can be known before training, has not been measured.

We measure it with four architectures on three datasets: DEAR on up to twelve CelebA attributes, with two annotators, seven correction rules and six metrics; CausalVAE and SCM-VAE, to test



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

whether the cost is specific to DEAR; Pendulum (Yang et al., 2021), whose causal graph is known, to test whether label noise follows it; and an audit of CausalVAE, SCM-VAE and ICM-VAE (Komanduri et al., 2024) on Pendulum, Flow and CelebA. Our contributions are the following.

- **Vision-language supervision of twelve attributes**, against four to six in published work, reaching 79% to 95% of the human-label score across six metrics, with a gap to human labels that narrows from four attributes to twelve (Sections 6.1 and 6.6).
- **A price for a label.** A label of index $J$ carries roughly the signal of $J^2$ human labels, and $J^2$, measured before training, orders what labels are worth (Sections 4.2 and 6.2).
- **Why corrections fail.** Label noise stays local, and seven correction rules from five families do not improve on the uncorrected labels (Sections 6.3 to 6.5).
- **An audit of evaluation.** Six scoring conventions that neither CausalVAE nor SCM-VAE states move MIC on the same trained models by up to 71.3 points; we release the protocol, the audit and every run's scores (Section 7).

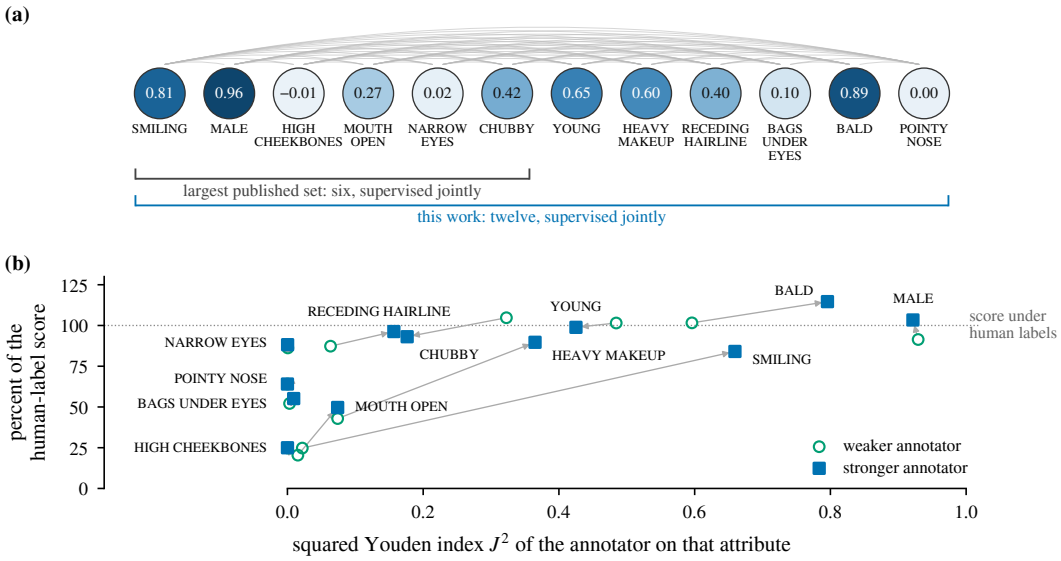


Figure 1: **What is supervised, and what a free label is worth.** (a) The twelve CelebA attributes in the assumed ordering, with the stronger annotator's Youden index $J$. (b) Each annotator's score as a percent of the human-label score at twelve attributes; arrows point to the stronger annotator.

## 2 Related Work

**Weakly supervised causal representation learning.** Without an inductive bias, disentanglement is not identifiable (Locatello et al., 2019), and this line of work supplies the bias as labels of the generative factors, following iVAE (Khemakhem et al., 2020). CausalVAE (Yang et al., 2021), SCM-VAE (Komanduri et al., 2022), DEAR (Shen et al., 2022), ICM-VAE (Komanduri et al., 2024) and CFI-VAE (Li et al., 2024), all weakly supervised in this sense, differ in what they assume about structure: CausalVAE learns the graph, DEAR is given a super-graph or an ordering, SCM-VAE and ICM-VAE the full graph, and CFI-VAE constructs one by hand. All but ICM-VAE use CelebA's own annotations, on four attributes, six for DEAR. They assume the labels exist and ask how much supervision is needed; we ask how good it has to be when a model supplies it, a question sharpened by how poorly existing methods fare on a real physical system (Gamella et al., 2025).

**Supervision signals other than labels.** Other work changes the signal rather than its source: paired observations (Locatello et al., 2020a), match and rank pairings (Shu et al., 2020), counterfactual pairs (Brehmer et al., 2022), sparse perturbations (Ahuja et al., 2022), data augmentations

(von Kügelgen et al., 2021), temporal sequences with intervened variables (Lippe et al., 2022), and interventional data (Squires et al., 2023; Ahuja et al., 2023; Buchholz et al., 2023; Zhang et al., 2023; Varici et al., 2024), unified by Yao et al. (2025). Locatello et al. (2020b) show that as few as 100 labels suffice and that 10% uniform label noise barely hurts, which Propositions 1 and 2 explain. None of these signals is measured, and none is available for an attribute named in words.

**Imperfect labels and models as annotators.** Methods for noisy labels, among them loss correction (Natarajan et al., 2013; Patrini et al., 2017), trusted subsets (Hendrycks et al., 2018), sample selection (Han et al., 2018), uncertainty weighting (Kendall et al., 2018) and instance-dependent models (Yao et al., 2021), surveyed by Song et al. (2023), target predictive accuracy, where the bias that noise introduces must be removed; Proposition 1 shows that this bias leaves rank-based representation criteria unchanged. Real label noise behaves unlike synthetic flips (Wei et al., 2022), as we find at matched rates, and Park et al. (2025) study synthetic concept noise in concept bottleneck models (Koh et al., 2020). Language models annotate text below crowd-worker cost (Gilardi et al., 2023; Wang et al., 2021), and Lu & Zhong (2024) compare a vision-language model with CelebA's annotations, themselves imperfect (Wu et al., 2023). To our knowledge no work supervises a causal representation with model-generated labels or prices label quality in human labels (Appendix B).

# 3 Problem Setup

## 3.1 Setting

We observe images $\boldsymbol{x}$ and want each of $k$ named attributes to have its own latent variable. The attributes are never observed; only binary labels $y_j \in \{0, 1\}$ impose the correspondence.

We train DEAR (Shen et al., 2022), which fits an encoder $E$ and a generator $G$ adversarially against a structured prior. Of its $d$ latent variables the first $k$ carry the attributes and follow a linear structural causal model,

$$z_{1:k} = f\big((I - A^\top)^{-1}\varepsilon\big), \qquad \varepsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $f$ is invertible and acts on one variable at a time and $A_{ij}$ is the weight of the edge from attribute $i$ to attribute $j$; the other $d - k$ are unstructured. Labels enter through DEAR's objective $\mathcal{L}(E, G) = \mathcal{L}_{\text{gen}}(E, G) + \lambda\mathcal{L}_{\text{sup}}(E)$, whose supervised term is

$$\mathcal{L}_{\text{sup}}(E) \;=\; \mathbb{E}_{\boldsymbol{x},y}\left[\frac{1}{k}\sum_{j=1}^{k}\text{BCE}(E(\boldsymbol{x})_j,\, y_j)\right]. \tag{2}$$

The support of $A$ is the only causal assumption, and we take the weakest DEAR admits: $A$ is strictly upper triangular under a fixed ordering of the attributes, which permits every earlier attribute to cause every later one. At $k = 12$ that leaves 66 learned edge weights, and no claim here rests on a specific graph, since they vary about as much across seeds as across conditions (mean Spearman 0.40 to 0.55 against 0.37 to 0.54; Appendix E).

## 3.2 A vision-language annotator

We replace the human labels with a vision-language model's answers. For each image and attribute the model is asked a question naming the attribute, and $\tilde{y}_j$ is its answer, the probability it assigns to "Yes" over "No" thresholded at $0.5$ (Appendix C). We use two annotators, the stronger Qwen3.5-9B (Qwen Team, 2026) and the weaker Qwen2.5-VL-7B (Qwen Team, 2025), so that what the substitution costs is not what one model happens to do. On attribute $j$ an annotator has a false positive rate (FPR) and a false negative rate (FNR) (Natarajan et al., 2013),

$$\rho_{0j} = P(\tilde{Y}_j = 1 \mid Y_j = 0), \qquad \rho_{1j} = P(\tilde{Y}_j = 0 \mid Y_j = 1), \tag{3}$$

summarized by the Youden index (Youden, 1950) $J_j = 1 - \rho_{0j} - \rho_{1j}$, which is 1 for a perfect annotator, 0 when its answers carry no information about the attribute, and negative when they are anticorrelated with it. Unlike accuracy, it does not depend on prevalence, here 2.4% to 77.9%.

Both rates are estimated per attribute on 500 human-labeled images, 250 positive and 250 negative, none of them among the 20,000 CelebA images we train and score on (Table 4 in Appendix C). On all 20,000 of those images, the stronger annotator's index is within 0.04 of this estimate on every attribute, so 500 images suffice. Stronger and weaker refer to the mean index, 0.43 against 0.32.

# 4 THEORETICAL ANALYSIS

We ask what label noise does to a rank-based criterion, what a noisy label is worth, what a label-side correction can do, and what changes when errors depend on the image; proofs are in Appendix A.

**Assumption 1** (Class-conditional channel)**.** $\tilde{Y}_j$ is independent of $\boldsymbol{x}$ given $Y_j$, so the rates $\rho_{0j}$ and $\rho_{1j}$ of Equation 3 do not vary from image to image.

The annotators violate Assumption 1 on some attributes: on HEAVY MAKEUP, the stronger annotator's false positive rate is 0.28 on images not labeled MALE and 0.01 on images labeled MALE (Appendix C). Section 4.4 drops the assumption.

## 4.1 INVARIANCE OF RANK-BASED CRITERIA

**Proposition 1** (Population equivalence, exact)**.** *Under Assumption 1 with $J_j \neq 0$, the minimizer $s^\star$ of the expected cross-entropy in Equation 2 against $\tilde{Y}_j$ satisfies*

$$\sigma(s^\star(\boldsymbol{x})) \;=\; \rho_{0j} + J_j\,\eta_j(\boldsymbol{x}), \qquad \eta_j(\boldsymbol{x}) = P(Y_j = 1 \mid \boldsymbol{x}), \tag{4}$$

*a strictly monotone transform of the minimizer $\eta_j$ obtained from clean labels, increasing when $J_j > 0$ and decreasing when $J_j < 0$.*

For $J_j > 0$, any criterion that depends on the score only through its order is therefore unchanged; for $J_j < 0$ the order reverses, which flips AUROC but not a criterion built on grid partitions, so the maximal information coefficient (MIC) (Reshef et al., 2011) and the total information coefficient (TIC) (Reshef et al., 2016), which CausalVAE and SCM-VAE report, are equal at both optima. The annotator's cost is therefore statistical rather than a bias in the population optimum, and a correction that removes such a bias cannot improve these criteria in the limit. The statement concerns the supervised term, not the generative objective the same latent variable also serves.

## 4.2 THE PRICE OF A LABEL

**Proposition 2** (Price of a label, first order)**.** *Fix a feature $\phi$ of the image. Under Assumption 1, $\mathrm{Cov}(\phi, \tilde{Y}_j) = J_j \,\mathrm{Cov}(\phi, Y_j)$ exactly. In the weak-signal regime, where $\phi$ is nearly independent of the label, and when the annotator answers positive as often as the attribute occurs, $n$ annotator labels carry the signal-to-noise ratio of $nJ_j^2$ human labels.*

The supervised term reaches a feature only through its covariance with the label, and attenuating that covariance by $J_j$ costs a factor $J_j^2$ in squared signal-to-noise ratio. When the annotator answers positive at a different rate than the attribute occurs, the ratio of the two label variances also enters (Appendix A), but the experiments are ordered better without it, Spearman 0.76 against 0.62 across attributes and Pearson 0.91 against 0.82 when only the annotator changes, and $J_j^2$ is therefore the price throughout the paper. The same argument prices a mixture: $n_1$ human labels pooled with $n_2$ annotator labels are worth $(n_1 + n_2 J_j)^2/(n_1 + n_2)$ human labels, and $n_1 + n_2 J_j^2$ if each annotator label is weighted by $J_j$, a prediction Section 6.4 tests.

## 4.3 LABEL-SIDE CORRECTIONS

A map from the binary answer to a supervision target has two degrees of freedom, so it is affine, $g(\tilde{y}) = g(0) + (g(1) - g(0))\,\tilde{y}$: it scales the supervised signal and its standard deviation alike, leaving the signal-to-noise ratio unchanged, and to first order acts as a change of $\lambda$ (Proposition 4, Appendix A). A loss weight, fixed or learned, does the same. Nothing that reads only the annotator's answer can add information, which covers posterior relabeling, learned per-attribute weights and the supervision weight, three of the rules of Section 6.4.

## 4.4 INSTANCE-DEPENDENT ERRORS

**Proposition 3** (Error structure, exact)**.** *Drop Assumption 1 and write $\rho(\boldsymbol{x})$ for the probability that the annotator errs at $\boldsymbol{x}$, so that $\mathbb{E}[\rho \mid Y_j = y]$ is the rate $\rho_{yj}$ of Equation 3 averaged over images of that class. With $\pi_y = P(Y_j = y)$ and $\bar{J}_j = 1 - \mathbb{E}[\rho \mid Y_j = 0] - \mathbb{E}[\rho \mid Y_j = 1]$,*

$$\mathrm{Cov}(\phi, \tilde{Y}_j) = \bar{J}_j \,\mathrm{Cov}(\phi, Y_j) + \pi_0 \,\mathrm{Cov}(\phi, \rho \mid Y_j = 0) - \pi_1 \,\mathrm{Cov}(\phi, \rho \mid Y_j = 1). \tag{5}$$

Under Assumption 1 the two covariance terms vanish and Proposition 2's attenuation is exact. When the annotator errs on images that resemble the other class, $\mathrm{Cov}(\phi, \rho \mid Y_j = 0)$ is positive and $\mathrm{Cov}(\phi, \rho \mid Y_j = 1)$ negative, so both terms raise the effective index $\mathrm{Cov}(\phi, \tilde{Y}_j)/\mathrm{Cov}(\phi, Y_j)$ above $\bar{J}_j$: errors placed this way cost less than the measured index predicts, the direction Section 6.5 finds.

# 5 Experimental Setup

## 5.1 Data and attributes

We use the first 20,000 aligned CelebA images, 17,000 for training and 3,000 held out, the same for every model and condition, and score every run on the held-out images against their human labels. The twelve attributes of Figure 1a are ordered as follows: positions 1 to 6 are DEAR's published smile group in DEAR's order, 7 to 10 complete its union with DEAR's age group, and 11 and 12 were drawn at random from the remaining 30. A run on $k \in \{4, 6, 8, 10, 12\}$ attributes supervises the first $k$. DEAR's age group, trained under DEAR's own super-graph for it, is a second attribute set. Pendulum (Yang et al., 2021) has four factors with a known graph; we corrupt 70% of the labels of one cause or of one effect and compare every factor with runs on clean labels (Appendix D).

## 5.2 Models and conditions

We train DEAR because its implementation accepts any number of supervised attributes and an ordering in place of a graph. CausalVAE and SCM-VAE build one decoder sub-network per attribute, written for four, so we train them on CelebA-Smile, the four-attribute subset of CelebA that SCM-VAE names (Komanduri et al., 2022); CausalVAE, whose released code generates Pendulum, also runs the Pendulum experiment. The conditions vary the label source (human labels; either annotator's labels; human labels on only 250 to 4,000 images; one attribute relabeled by the annotator; the seven attributes with $J_j < 0.5$ unlabeled; human labels flipped at the annotator's measured rates), the correction rule (Section 6.4) and $\lambda \in \{1, 5, 10\}$, with the stronger annotator unless stated. Every condition uses five seeds, three for the supervision weight; Appendix D gives every setting.

## 5.3 Metrics and statistics

All six metrics are higher-is-better and computed on the encoder means of the $k$ supervised latent variables. MIC and TIC are measured between latent variable $j$ and attribute $j$ with minepy (Albanese et al., 2013) at its defaults, divided by the label's entropy $H(Y_j)$, which bounds MIC (Proposition 5), and averaged over attributes; we call these normalized MIC and TIC. We also report the DCI disentanglement, completeness and informativeness scores (Eastwood & Williams, 2018) and the AUROC of a logistic-regression probe (Appendix D). On CelebA, Figures 1 to 4 use normalized MIC, and Table 1 reports all six. Differences between conditions are paired by seed and written as the mean $\pm$ half the width of a 95% Student-$t$ interval (Student, 1908); families of tests carry a Holm correction (Holm, 1979), and Table 6 in Appendix D lists every test.

# 6 Results

## 6.1 Scaling beyond four attributes

Annotator labels let supervision grow from four attributes to twelve, at a cost relative to human labels (Figure 2a). The human-label score falls at every step, from 48.1 at four attributes to 39.9 at twelve, while the stronger annotator's stays between 31.1 and 35.2 and the weaker's between 25.1 and 26.8. Human labels lead the stronger annotator at every count, by $10.6 \pm 2.3$ points pooled over counts and by 6.3 at twelve against 12.9 at four, and the weaker annotator by $17.6 \pm 2.0$. Free labels beat a substantial annotation budget: at twelve attributes, the stronger annotator's labels on all 17,000 images score $3.9 \pm 3.0$ points above human labels on 4,000 of them.

The same order holds on all six metrics (Table 1): human labels, then the stronger annotator, then the weaker, with the stronger reaching 79% to 95% of the human score. These are scores of our own runs under one convention, not published values, for the reason Section 7 gives. Figure 6 in Appendix E shows counterfactuals under each label source, and Figure 7 the curve by attribute.

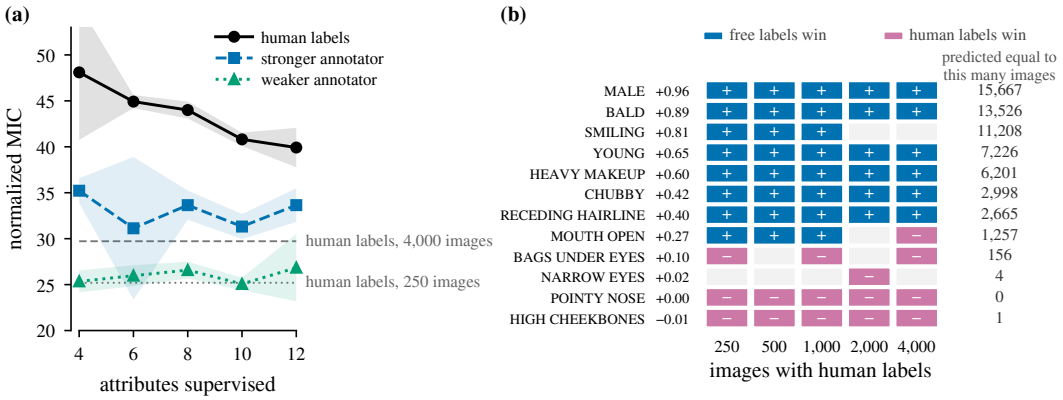


Figure 2: **What free labels buy.** (a) Normalized MIC against attributes supervised, bands one standard deviation: the human-label score falls as attributes are added, while both annotators' scores hold. (b) The stronger annotator's labels on each attribute against human labels on that many images, one-sided $t$-test at $p < 0.05$; in empty cells neither is significantly better. The last column is the price 17,000 $J^2$ of Section 4.2.

Table 1: Six metrics at twelve attributes, mean ± standard deviation across seeds; higher is better. Human labels are the reference; bold marks the better annotator. The last two rows give each annotator's score as a percent of the human-label score, both measured from chance level (0.5 for AUROC, 0 otherwise), ± half a 95% interval.

| Label source | MIC (%) | TIC (%) | DCI-D | DCI-C | DCI-I | Probe AUROC |
|---|---|---|---|---|---|---|
| Human labels (reference) | 39.9 ± 2.1 | 25.8 ± 2.2 | 0.23 ± 0.03 | 0.26 ± 0.03 | 0.24 ± 0.02 | 0.82 ± 0.01 |
| Weaker annotator | 26.8 ± 3.7 | 13.7 ± 3.6 | 0.10 ± 0.03 | 0.13 ± 0.05 | 0.09 ± 0.06 | 0.73 ± 0.05 |
| Stronger annotator | **33.6 ± 1.8** | **20.2 ± 1.8** | **0.18 ± 0.02** | **0.24 ± 0.04** | **0.20 ± 0.01** | **0.80 ± 0.01** |
| Weaker annotator, % of human | 67.0 ± 8.5 | 52.3 ± 14.4 | 41.7 ± 17.7 | 48.6 ± 21.9 | 35.5 ± 30.1 | 70.6 ± 17.4 |
| Stronger annotator, % of human | **84.4 ± 5.9** | **78.7 ± 9.8** | **80.1 ± 5.5** | **89.7 ± 7.9** | **85.8 ± 15.3** | **94.6 ± 5.2** |

### 6.2 The index prices a label

The index measured on 500 calibration images orders what each attribute's free labels are worth before any model is trained. Across the twelve attributes, the percent of the human score that the stronger annotator reaches rises with $J_j^2$ (Spearman 0.76). A sharper test changes only the annotator, holding the images, the attributes, their positions and the architecture fixed: the change in that percent then follows the change in $J_j^2$ (Pearson 0.91; Figure 1b), and the stronger annotator gains $7.0 \pm 1.7$ points over the weaker, pooled over attribute counts.

In annotation units (Figure 2b), six of the seven attributes with the highest index beat every budget up to 4,000 human-labeled images, three of the four near chance lose to as few as 250, and MOUTH OPEN, at $J_j = 0.27$, beats 1,000 and loses to 4,000, bracketing its price of 1,257. The price is an ordering and an order of magnitude rather than an exact count.

### 6.3 Label noise stays local

Noise on one attribute costs that attribute and leaves the others essentially unchanged (Figure 3a). Replacing the human labels of HIGH CHEEKBONES, which the stronger annotator answers at chance, with the annotator's costs it $31.3 \pm 4.2$ points, and none of the other eleven moves significantly. Relabeling BAGS UNDER EYES costs it $10.5 \pm 2.2$, and the one other attribute that moves significantly, YOUNG, improves. The reverse test removes the labels of the seven attributes with $J_j < 0.5$: if noisy labels harmed their neighbors, the five that keep labels would improve, and none does.

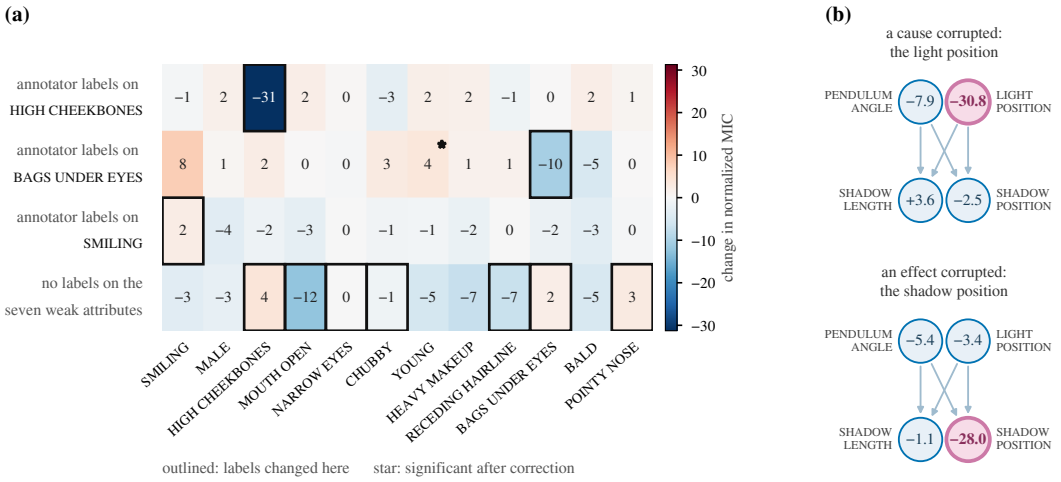


Figure 3: **Label noise stays where it is put.** (a) Change in normalized MIC per attribute. Rows 1 to 3 give one attribute the stronger annotator's labels, against human labels; row 4 removes the seven weakest attributes' labels, against the annotator's labels on all twelve. The star marks the only untouched cell significant after Holm correction. (b) CausalVAE on Pendulum: the change in MIC on each node when 70% of the ringed factor's labels are corrupted.

Where the graph is known, corruption does not travel along it (Figure 3b). On Pendulum, corrupting 70% of the light position's labels costs it 30.8 points, while its two effects move by $+3.6$ and $-2.5$ and the pendulum angle, which it does not cause, by $-7.9$.

### 6.4 Corrections do not help

No rule we tested improves on the uncorrected labels (Figure 4). The seven come from five families: forward correction (Patrini et al., 2017), learned per-attribute loss weights (Kendall et al., 2018), posterior relabeling, three ways of spending or withholding annotation, and the supervision weight. At twelve attributes the first six move the score by $-3.0$ to $+1.1$ points with no interval excluding zero, and a tenfold change in $\lambda$ leaves the gap where it was.

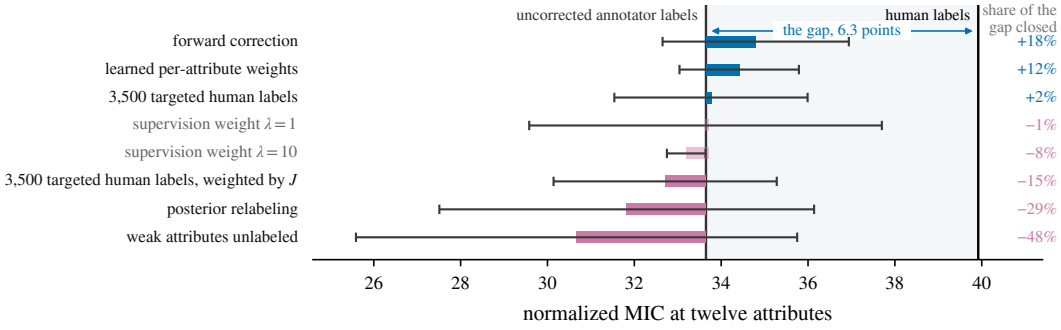


Figure 4: **No correction closes the gap to human labels.** Seven rules from five families, stronger annotator, twelve attributes. Each bar runs from the uncorrected labels to the normalized MIC the rule reaches, with its 95% interval; the band is the 6.3-point gap to human labels. Faded rows start from the default supervision weight and rest on three seeds.

The targeted rules are the strongest test: they spend human effort where the annotator is weakest. On the seven attributes they target, pooled over six and twelve attributes, they gain $0.2 \pm 1.8$ points unweighted and $0.4 \pm 1.9$ weighted, against a gap of $14.1 \pm 3.0$ to human labels, which excludes recovering more than about a sixth of it. Weighting by $J_j$, which Section 4.2 says should recover the

full value of the human labels, does not. Remedies that act on the optimizer, such as the sharpness-aware minimization Park et al. (2025) find helps concept bottleneck models, lie outside these five families.

### 6.5 Real versus simulated errors

At matched error rates, the annotator's own errors cost less than errors drawn at random: human labels flipped at each attribute's measured rates score 25.2 against the annotator's 33.6, $-8.4 \pm 7.5$ ($-4.2 \pm 2.4$ without two runs that failed to train), most where the annotator answers well ($-15.6$ on the five attributes it answers best against $-0.2$ on the four it answers near chance; Figure 8). Proposition 3 gives the direction but not the size: the annotator's effective index exceeds its measured index on all five well-answered attributes, while simulated labels sit at it (Appendix E).

### 6.6 Ablations and robustness

The gap between human and annotator labels survives every setting we fixed by choice. It holds for both annotators (Table 1) and, with every interval in Table 2 excluding zero, for CausalVAE and SCM-VAE, each measured against its own human-label baseline (Figure 9); for the second attribute set, trained under DEAR's own super-graph; and for $\lambda$ of 1, 5 and 10. The second set's gap, 4.3, is smaller than the first's, 13.8, in the direction its higher mean index predicts (0.52 against 0.41), but the difference is not significant, and the two sets also differ in assumed structure.

Table 2: The annotator's change from human labels under each variation, and its score as a percent of the human-label score, $\pm$ half a 95% interval; three seeds for the supervision weight. Human and annotator columns are seed means. DEAR's four attributes include HIGH CHEEKBONES where the other architectures have NARROW EYES. Every change's interval excludes zero.

| | Normalized MIC | | | |
|---|---|---|---|---|
| **What varies** | **Human** | **Annotator** | **Change** | **% of human** |
| *Architecture, four attributes* | | | | |
| DEAR | 48.1 | 35.2 | $-12.9 \pm 10.2$ | $75.2 \pm 20.7$ |
| CausalVAE | 64.7 | 51.7 | $-13.0 \pm 4.1$ | $80.0 \pm 5.3$ |
| SCM-VAE | 50.7 | 46.8 | $-4.0 \pm 0.6$ | $92.2 \pm 1.1$ |
| *Attribute set* | | | | |
| First published set (six) | 44.9 | 31.1 | $-13.8 \pm 10.5$ | $69.5 \pm 22.4$ |
| Second published set | 41.7 | 37.4 | $-4.3 \pm 1.7$ | $89.7 \pm 4.0$ |
| *Supervision weight* | | | | |
| $\lambda = 1$ | 41.1 | 33.6 | $-7.5 \pm 2.4$ | $81.8 \pm 5.4$ |
| $\lambda = 5$ (default) | 40.8 | 33.7 | $-7.1 \pm 1.6$ | $82.7 \pm 4.2$ |
| $\lambda = 10$ | 41.2 | 33.2 | $-8.0 \pm 6.9$ | $80.7 \pm 14.7$ |

## 7 Revisiting the Evaluation Protocol

Every comparison in Section 6 is between our own runs, scored one way, because in this line of work a published MIC is not determined by the model it describes. Its statistical properties have been contested (Kinney & Atwal, 2014); the problem here is simpler: computing it involves choices the papers do not report.

CausalVAE and SCM-VAE report MIC and TIC on four datasets: Pendulum, Flow (Yang et al., 2021), and CelebA-Smile and CelebA-Beard, two four-attribute subsets of CelebA. Both state the target: CausalVAE the labels, and SCM-VAE the labels transformed by the causal graph, $Au$ in its notation, which we compute as $(I + A)u$ so that no factor is identically zero. Neither states six other choices: the latent representation scored, the parameter $\alpha$ of the MIC estimator (Reshef et al., 2011), averaging or maximizing over a factor's dimensions, matching latent variables to factors by index or by assignment, dividing by the entropy ceiling, and averaging the factors or reporting the best.

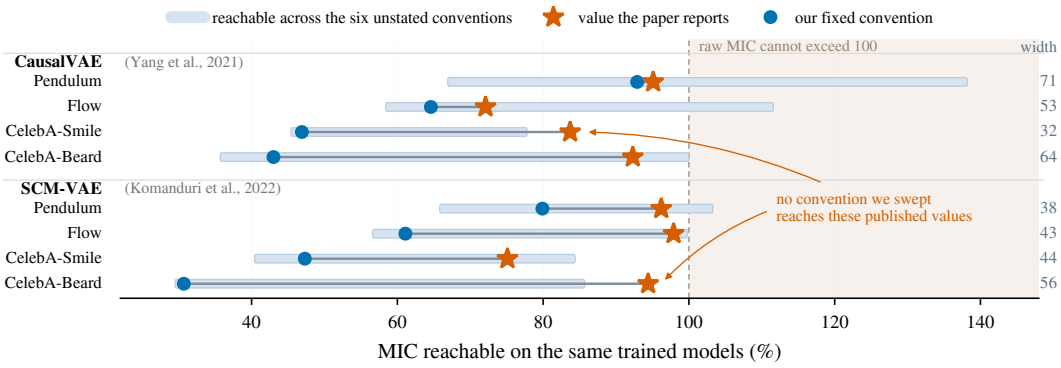


Figure 5: **Published scores are underdetermined.** The range of MIC reachable on the same trained models across the six scoring conventions neither paper states, with the target held at what each paper states. The right column gives each width in points.

We train each model on each dataset with five seeds, from its released code where that code runs the experiment (Appendix D), hold the target at what each paper states, and score every combination of the six choices, 96 for CausalVAE and 64 for SCM-VAE. The same trained models then reach MIC over ranges from 32.4 to 71.3 points wide (Figure 5). Two published values, CausalVAE's on CelebA-Smile and SCM-VAE's on CelebA-Beard, lie outside every range, and no single convention accounts for a paper's whole table: the closest misses one of CausalVAE's four values by 20.9 points and one of SCM-VAE's by 9.3. The papers also disagree with each other: SCM-VAE re-runs CausalVAE and reports values up to 24.7 points from CausalVAE's own, on Flow. Reporting the best factor rather than the mean moves MIC most, and matching least (Figure 11). Our scoring code comes within 2.2 points of CausalVAE's published Pendulum MIC and within 0.23 of ICM-VAE's (Komanduri et al., 2024) published DCI scores (Appendix E). A published MIC is therefore a property of a model and an unreported convention together.

A binary label carries at most one bit, and MIC cannot exceed it: for any score $s$, $\mathrm{MIC}(s, Y) \leq H(Y)$, the label's entropy in bits, with equality when a threshold on $s$ separates the classes (Proposition 5, proved in Appendix A). On CelebA-Beard, a mean of raw MIC over the four binary attributes cannot exceed 64.5 on CausalVAE's test labels or 62.8 on SCM-VAE's, because BALD is rare enough that its own ceiling is 16.0 and 14.0. The papers report 92.3 and 94.4, so these are not means of raw MIC, and each rests on a choice the paper does not state, such as dividing by the ceiling or reporting the best attribute. The same ceiling is why our convention normalizes: without it, a representation that separates BALD perfectly would score 16.0 rather than 100.

## 8 Discussion and Limitations

Vision-language annotators lift the assumption that labels exist, which has held published work on real images to four to six attributes: they label any attribute that can be named, what their labels cost can be estimated before training, and the gap to human labels narrows from four attributes to twelve, where our experiments stop but the annotator need not. On a dataset without attribute labels, human labels are then needed only to calibrate, 500 per attribute here, rather than for every training image. The results give a procedure: calibrate the annotator on 500 human labels per attribute and read the price from its index; keep the labels of attributes it answers poorly, since removing them does not help the rest; and improve the annotator rather than the loss, since the stronger annotator gains $7.0 \pm 1.7$ points over the weaker, pooled over attribute counts, while no correction rule improves on the uncorrected labels. A reported MIC or TIC should come with the convention that produced it.

The annotators labeled only CelebA, over several attribute sets of at most twelve; larger sets, other real-image datasets and other annotator families are untested. CelebA's own annotations contain errors (Wu et al., 2023), one index per attribute averages over subgroups (Appendix C), the ordering is assumed and was not permuted, and the attribute-count, pricing and correction experiments use DEAR, the only one of our architectures that accepts twelve attributes.

### AI USE STATEMENT

We used generative AI tools to polish the writing, extend our search of related work, help implement experiment and analysis code, and check the derivations of the propositions. The research questions, methodology, experimental design, choice of models and datasets, and claims are our own. We reviewed all AI-assisted work, including testing the code and checking every citation, and take full responsibility for the content of this paper.

### ETHICS STATEMENT

We use CelebA (Liu et al., 2015), images of public figures distributed for non-commercial research, under its terms, and show one held-out face in Figure 6, as existing literature does. Its human labels are CelebA's own; we collect no new human annotation. Some CelebA attributes are subjective, and MALE records a perceived presentation, not gender. The annotators' error rates differ between subgroups (Appendix C), so applications of this protocol to data about people should report them by subgroup. Our supplementary material contains no CelebA image or attribute label; the annotator labels it contains are released for non-commercial research under CelebA's terms.

### REPRODUCIBILITY STATEMENT

Sections 3 and 5 and Appendix D specify the models, data, conditions, training settings, metrics and statistics; Appendix C the calibration protocol and every question put to the annotators; and Appendix A the proofs. The anonymized supplementary material contains the code, including the changes to each baseline's released code, the annotator labels, their calibrated error rates, a script that builds every condition's labels from CelebA, the scores of every run, the convention grid of Section 7, and the scripts that recompute every number the paper reports from those scores. Every run uses a fixed seed, and Appendix D states which use deterministic kernels.

## REFERENCES

Kartik Ahuja, Jason S. Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 15516–15528, 2022.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pp. 372–407, 2023.

Davide Albanese, Michele Filosi, Roberto Visintainer, Samantha Riccadonna, Giuseppe Jurman, and Cesare Furlanello. minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers. *Bioinformatics*, 29(3):407–408, 2013.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco S. Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 38319–38331, 2022.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 45419–45462, 2023.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *International Conference on Machine Learning (ICML)*, 2024.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations (ICLR)*, 2018.

Juan L. Gamella, Simon Bing, and Jakob Runge. Sanity checking causal representation learning on a simple real-world system. In *International Conference on Machine Learning (ICML)*, 2025.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7482–7491, 2018.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2207–2217. PMLR, 2020.

Justin B. Kinney and Gurinder S. Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359, 2014. doi: 10.1073/pnas.1309933111.

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

Aneesh Komanduri, Yongkai Wu, Wen Huang, Feng Chen, and Xintao Wu. SCM-VAE: Learning identifiable causal representations via structural knowledge. In *IEEE International Conference on Big Data (Big Data)*, 2022.

Aneesh Komanduri, Yongkai Wu, Feng Chen, and Xintao Wu. Learning causally disentangled representations via the principle of independent causal mechanisms. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*, 2013.

Xiutian Li, Siqi Sun, and Rui Feng. Causal representation learning via counterfactual intervention. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 3234–3242, 2024.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Stratis Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning (ICML)*, pp. 13557–13603. PMLR, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2015.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning (ICML)*, pp. 4114–4124. PMLR, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning (ICML)*, pp. 6348–6359. PMLR, 2020a.

Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem. Disentangling factors of variation using few labels. In *International Conference on Learning Representations (ICLR)*, 2020b.

Haoming Lu and Feifei Zhong. Can vision-language models replace human annotators: A case study with CelebA dataset, 2024. NeurIPS 2024 Workshop on Evaluating Evaluations (EvalEval).

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep K. Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems (NIPS)*, volume 26, 2013.

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

Seonghwan Park, Jueun Mun, Donghyun Oh, and Namhoon Lee. An analysis of concept bottleneck models: Measuring, understanding, and mitigating the impact of noisy annotations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 39, 2025.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1944–1952, 2017.

Qwen Team. Qwen2.5-VL technical report, 2025. URL `https://arxiv.org/abs/2502.13923`.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, 2017. doi: 10.14778/3157794.3157797.

David N. Reshef, Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, 2011.

Yakir A. Reshef, David N. Reshef, Hilary K. Finucane, Pardis C. Sabeti, and Michael Mitzenmacher. Measuring dependence powerfully and equitably. *Journal of Machine Learning Research*, 17 (211):1–63, 2016.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23(241):1–55, 2022.

Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations (ICLR)*, 2020.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2023. doi: 10.1109/TNNLS.2022.3152527.

Chandler Squires, Anna Seigal, Salil S. Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pp. 32540–32560, 2023.

Divyansh Srivastava, Ge Yan, and Tsui-Wei Weng. VLG-CBM: Training concept bottleneck models with vision-language guidance. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 79057–79094, 2024.

Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 238 of *PMLR*, pp. 2314–2322, 2024.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 16451–16467, 2021.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 4195–4205, 2021.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations (ICLR)*, 2022.

Haiyu Wu, Grace Bezold, Manuel Günther, Terrance Boult, Michael C. King, and Kevin W. Bowyer. Consistency and accuracy of CelebA attribute values. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3258–3266, 2023.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves ImageNet classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2020.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Disentangled representation learning via neural structural causal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9588–9597. IEEE, 2021.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 53847–53890, 2025.

Yu Yao, Tongliang Liu, Mingming Gong, Bo Han, Gang Niu, and Kun Zhang. Instance-dependent label-noise learning under a structural causal model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 4409–4420, 2021.

W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 50254–50292, 2023.

# APPENDIX

The appendix is organized as follows. Appendix A proves the five propositions. Appendix B compares this work with prior work in more detail than Section 2. Appendix C describes the annotators, their calibration and every question they were asked. Appendix D gives the data, the construction of every condition, the metrics, the statistics with the table of every test, the training settings and the compute. Appendix E reports additional results, including counterfactuals.

## A PROOFS

Propositions 1, 2 and 3 are restated from Section 4; Propositions 4 and 5 are stated here and summarized in Sections 4.3 and 7. Throughout, $\phi$ is a fixed real-valued feature of the image, $\eta(\boldsymbol{x}) = P(Y = 1 \mid \boldsymbol{x})$, and the attribute index $j$ is dropped.

**Proposition 1** (restated)**.** *Under Assumption 1 with $J \neq 0$, the minimizer $s^\star$ of the expected cross-entropy against $\tilde{Y}$ satisfies $\sigma(s^\star(\boldsymbol{x})) = \rho_0 + J\,\eta(\boldsymbol{x})$, a strictly monotone transform of the clean minimizer $\eta$, increasing when $J > 0$ and decreasing when $J < 0$.*

*Proof.* Under Assumption 1,

$$P(\tilde{Y} = 1 \mid \boldsymbol{x}) = (1 - \rho_1)\,\eta(\boldsymbol{x}) + \rho_0\,(1 - \eta(\boldsymbol{x})) = \rho_0 + J\,\eta(\boldsymbol{x}).$$

Over all measurable score functions, the expected cross-entropy against a binary target is minimized pointwise by the target's conditional probability, so $\sigma(s^\star(\boldsymbol{x})) = \rho_0 + J\eta(\boldsymbol{x})$. For $J \neq 0$ the map $\eta \mapsto \rho_0 + J\eta$ is strictly monotone and $\sigma$ is strictly increasing, so $s^\star$ orders the inputs as $\eta$ does when $J > 0$ and in reverse when $J < 0$. A criterion that depends on a score only through its order therefore takes the same value at both minimizers when $J > 0$. A criterion that is also invariant to reversing the order takes the same value for either sign; MIC and TIC are, because a strictly monotone map, increasing or decreasing, sends every grid partition of the score to a grid partition with the same mutual information. □

**Proposition 2** (restated)**.** *Under Assumption 1, $\mathrm{Cov}(\phi, \tilde{Y}) = J\,\mathrm{Cov}(\phi, Y)$ exactly. In the weak-signal regime, and when the annotator answers positive as often as the attribute occurs, $n$ annotator labels carry the signal-to-noise ratio of $nJ^2$ human labels.*

*Proof.* Because $\phi$ is a function of $\boldsymbol{x}$, $\mathrm{Cov}(\phi, Y) = \mathrm{Cov}(\phi, \eta)$, and by the display above $\mathbb{E}[\tilde{Y} \mid \boldsymbol{x}] = \rho_0 + J\eta(\boldsymbol{x})$, so $\mathrm{Cov}(\phi, \tilde{Y}) = \mathrm{Cov}(\phi, \rho_0 + J\eta) = J\,\mathrm{Cov}(\phi, Y)$. The empirical covariance over $n$ independent draws has mean $\mathrm{Cov}(\phi, \tilde{Y})$ and, to leading order, variance $\mathrm{Var}\left((\phi - \mathbb{E}\phi)(\tilde{Y} - \mathbb{E}\tilde{Y})\right)/n$. In the weak-signal regime, where $\phi$ is nearly independent of the label, this variance is approximately $\mathrm{Var}(\phi)\,\mathrm{Var}(\tilde{Y})/n$, and likewise with $Y$. With $\kappa = \mathrm{Var}(Y)/\,\mathrm{Var}(\tilde{Y})$, the squared signal-to-noise ratio of $n$ annotator labels is therefore $J^2\kappa$ times that of $n$ human labels. When the annotator answers positive as often as the attribute occurs, $\tilde{Y}$ and $Y$ have the same variance, $\kappa = 1$, and $n$ annotator labels are worth $nJ^2$ human labels. □

**The general factor.** On the training labels, $\kappa$ lies between 0.6 and 2.0 for the stronger annotator on ten of the twelve attributes. Section 4.2 reports that the experiments are ordered better by $J^2$ than by $J^2\kappa$, which is why the paper uses $J^2$.

**Mixtures.** Pooling $n_1$ human labels with $n_2$ annotator labels, weighted 1 and $w$, gives a signal proportional to $n_1 + n_2 wJ$ and a variance proportional to $n_1 + n_2 w^2$ in the same regime, so the squared signal-to-noise ratio is $(n_1 + n_2 wJ)^2/(n_1 + n_2 w^2)$ in units of one human label. At $w = 1$ this is $(n_1 + n_2 J)^2/(n_1 + n_2)$; it is maximized at $w = J$, where it equals $n_1 + n_2 J^2$.

**Proposition 3** (restated)**.** *Without Assumption 1, with $\rho(\boldsymbol{x})$ the probability that the annotator errs at $\boldsymbol{x}$, $\pi_y = P(Y = y)$ and $\bar{J} = 1 - \mathbb{E}[\rho \mid Y = 0] - \mathbb{E}[\rho \mid Y = 1]$,*

$$\mathrm{Cov}(\phi, \tilde{Y}) = \bar{J}\,\mathrm{Cov}(\phi, Y) + \pi_0\,\mathrm{Cov}(\phi, \rho \mid Y = 0) - \pi_1\,\mathrm{Cov}(\phi, \rho \mid Y = 1).$$

*Proof.* Write $m_y = \mathbb{E}[\phi \mid Y = y]$, $\bar{\rho}_y = \mathbb{E}[\rho \mid Y = y]$ and $C_y = \mathrm{Cov}(\phi, \rho \mid Y = y)$. By the definition of $\rho$, $\mathbb{E}[\tilde{Y} \mid \boldsymbol{x}, Y = 1] = 1 - \rho(\boldsymbol{x})$ and $\mathbb{E}[\tilde{Y} \mid \boldsymbol{x}, Y = 0] = \rho(\boldsymbol{x})$. Conditioning on $Y$ and using $\mathbb{E}[\phi\rho \mid Y = y] = C_y + m_y\bar{\rho}_y$,

$$\mathbb{E}[\phi\tilde{Y}] = \pi_1 m_1(1 - \bar{\rho}_1) - \pi_1 C_1 + \pi_0 m_0 \bar{\rho}_0 + \pi_0 C_0,$$
$$\mathbb{E}[\phi]\,\mathbb{E}[\tilde{Y}] = (\pi_0 m_0 + \pi_1 m_1)\,(\pi_1(1 - \bar{\rho}_1) + \pi_0\bar{\rho}_0)\,.$$

Subtracting and using $\pi_0 + \pi_1 = 1$, the terms without $C_y$ collect to $\pi_0\pi_1(m_1 - m_0)\bar{J}$, so

$$\mathrm{Cov}(\phi, \tilde{Y}) = \pi_0\pi_1(m_1 - m_0)\,\bar{J} + \pi_0 C_0 - \pi_1 C_1.$$

Since $\mathrm{Cov}(\phi, Y) = \pi_0\pi_1(m_1 - m_0)$, this is the stated identity. Under Assumption 1, $\rho$ depends on $\boldsymbol{x}$ only through $Y$, so $C_0 = C_1 = 0$ and $\bar{J} = J$, recovering Proposition 2. □

**Proposition 4** (Label-side maps, first order)**.** *Any map $g$ from a binary answer to a supervision target satisfies $g(\tilde{y}) = g(0) + (g(1) - g(0))\,\tilde{y}$. It multiplies the supervised signal and its standard deviation by the same constant, leaving the signal-to-noise ratio unchanged, and to first order acts on that attribute as a change of $\lambda$.*

*Proof.* A function on $\{0, 1\}$ is determined by its two values, so $g(\tilde{y}) = g(0) + b\,\tilde{y}$ with $b = g(1) - g(0)$. Then $\mathrm{Cov}(\phi, g(\tilde{Y})) = b\,\mathrm{Cov}(\phi, \tilde{Y})$ and $\mathrm{Var}(g(\tilde{Y})) = b^2\,\mathrm{Var}(\tilde{Y})$, so the signal-to-noise ratio is unchanged. In the supervised term the gradient of the cross-entropy with respect to the score is $\sigma(s) - g(\tilde{y})$, whose label-dependent part is scaled by $b$, while the constant $g(0)$ shifts a bias that the model absorbs; to first order this rescales that attribute's term, as a change of $\lambda$ would. □

**Proposition 5** (Ceiling, exact)**.** *Let $Y$ be a binary label with $\pi_1 = P(Y = 1)$ and entropy $H(Y) = -\pi_1 \log_2 \pi_1 - (1 - \pi_1)\log_2(1 - \pi_1)$ bits. For any score $s$, $\mathrm{MIC}(s, Y) \le H(Y)$, with equality when a threshold on $s$ separates the two classes.*

*Proof.* MIC is the largest value, over grids with $a \ge 2$ columns on $s$ and $b \ge 2$ rows on $Y$ within the sample-size limit, of $I_G(s; Y)/\log_2 \min(a, b)$, where $I_G$ is the mutual information of the discretized pair (Reshef et al., 2011). Because $Y$ takes two values, the row variable of any grid is a function of $Y$, so $I_G(s; Y)$ is at most its entropy, which is at most $H(Y)$, and the denominator is at least $\log_2 2 = 1$. Every grid's ratio, and hence MIC, is therefore at most $H(Y)$. A two-by-two grid whose column boundary is a threshold separating the classes has $I_G = H(Y)$ and denominator 1. The bound holds for every entry of the characteristic matrix, and so for minepy's estimate. □

## B Extended Comparison with Prior Work

Table 3 places this work beside the methods and studies of Section 2. The attribute count is the number of CelebA attributes whose labels each work uses.

**Chosen against measured noise.** Locatello et al. (2020b) make labels imprecise in three ways: binning them, replacing 10% of them with random values, and labeling only a subset of the factors. Each rate is chosen rather than measured, and the first two are uniform across factors, whereas the annotator studied here errs at a rate that differs by attribute and comes within 0.10 of chance on four of the twelve. Propositions 1 and 2 explain why uniform noise is the cheap case: it preserves the order of the population optimum and, at 10%, keeps $J$ at 0.8 or above.

**Where a mitigation acts.** Park et al. (2025) test two mitigations for noisy concept annotations, sharpness-aware minimization and correcting high-entropy concepts at inference, and report that both help. Both act on optimization or inference, whereas the seven rules of Section 6.4 act on the labels, the loss, the loss weights, the annotation budget or the supervision weight. Neither remedy is tested here, so our negative result does not contradict theirs.

Table 3: Where each line of work gets its labels, how many CelebA attributes it uses, and whether it studies the quality of those labels. "Measured" means the paper estimates the error rate of its own labels against a reference; "synthetic" means errors are injected at a chosen rate.

| Work | Label source | CelebA attrs. | Noise | Causal latents |
|---|---|---|---|---|
| CausalVAE (Yang et al., 2021) | dataset labels | 4 | not studied | yes |
| SCM-VAE (Komanduri et al., 2022) | dataset labels | 4 | not studied | yes |
| DEAR (Shen et al., 2022) | dataset labels | 6 | not studied | yes |
| ICM-VAE (Komanduri et al., 2024) | synthetic truth | n/a | not studied | yes |
| CFI-VAE (Li et al., 2024) | dataset labels | 4 | not studied | yes |
| Locatello et al. (2020b) | dataset labels | n/a | synthetic | no |
| Park et al. (2025) | dataset labels | n/a | synthetic | no |
| Lu & Zhong (2024) | vision-language model | 40 | measured | no |
| Wu et al. (2023) | dataset labels | 40 | measured | no |
| This work | vision-language model | 12 | measured | yes |

**Machine labels.** Training on labels a model produced predates vision-language annotators: pseudo-labels (Lee, 2013), the labeling functions of Snorkel (Ratner et al., 2017) and self-training with a noisy student (Xie et al., 2020) all do it, and a strong model trained on a weak model's labels can exceed its supervisor (Burns et al., 2024). Snorkel estimates the accuracies of several labeling functions from their agreement; here a single labeler's error rates are measured directly against held-out human labels, which is what makes a price in human labels available before training. Concept bottleneck models have also been built from concepts that a language model proposes (Oikarinen et al., 2023) or that a vision-language model grounds (Srivastava et al., 2024). None of these trains a causal representation.

**Measuring label quality without training on it.** Lu & Zhong (2024) compare a vision-language model with CelebA's annotations on all 40 attributes, and Wu et al. (2023) audit those annotations; neither trains a representation on the result. The representation-learning methods in Table 3 train on their labels and none measures them. This work does both, which is what turns label quality into a number of human labels.

## C Annotators and Calibration

This appendix describes how the annotators are queried, how their error rates are measured, and where Assumption 1 fails.

**Querying.** Each annotator is asked one question per image and attribute; its answer is the probability it assigns to "Yes" over "No" as the next token, thresholded at $0.5$. Table 5 lists every question. Wording 1 labels the training set, and all three wordings are asked in calibration. Agreement with the human labels varies with wording by a standard deviation of 0.2 to 13.4 percentage points across attributes for the stronger annotator and 0.6 to 10.5 for the weaker.

**Calibration.** The rates of Equation 3 are estimated per attribute on 500 human-labeled images, 250 positive and 250 negative, so that both rates rest on the same number of images. Sampling by class rather than at random keeps the rare attributes estimable: BALD occurs in 2.4% of the data, and a random 500 would leave about a dozen positives. Prevalence is therefore taken from the training images, not from this set. No calibration image is among the 20,000 we train and score on. Measured instead on all 20,000 of those images, the stronger annotator's index differs from its calibration value by at most 0.04 on every attribute.

**Error rates.** Table 4 gives both annotators' rates and indices. A low index can come from either direction: on HIGH CHEEKBONES the stronger annotator answers yes to 94% of the negatives, and on POINTY NOSE it answers no to all but one of the 250 positives.

Table 4: The twelve CelebA attributes in the assumed causal ordering, grouped by the supervised count at which each enters; prevalence is over the 17,000 training images. $J = 1 - \mathrm{FPR} - \mathrm{FNR}$, with 0 chance and 1 perfect; rates are rounded so that this holds as printed.

| | | Weaker annotator | | | Stronger annotator | | |
|---|---|---|---|---|---|---|---|
| **Attribute** | **Prevalence (%)** | **FPR** | **FNR** | $J$ | **FPR** | **FNR** | $J$ |
| *In the four-attribute set* | | | | | | | |
| SMILING | 48.0 | 0.02 | 0.83 | $+0.15$ | 0.09 | 0.10 | $+0.81$ |
| MALE | 42.2 | 0.01 | 0.03 | $+0.96$ | 0.01 | 0.03 | $+0.96$ |
| HIGH CHEEKBONES | 45.0 | 0.75 | 0.27 | $-0.02$ | 0.94 | 0.07 | $-0.01$ |
| MOUTH OPEN | 47.9 | 0.61 | 0.27 | $+0.12$ | 0.70 | 0.03 | $+0.27$ |
| *Added at six* | | | | | | | |
| NARROW EYES | 11.6 | 0.10 | 0.92 | $-0.02$ | 0.06 | 0.92 | $+0.02$ |
| CHUBBY | 5.7 | 0.08 | 0.35 | $+0.57$ | 0.05 | 0.53 | $+0.42$ |
| *Added at eight* | | | | | | | |
| YOUNG | 77.9 | 0.07 | 0.23 | $+0.70$ | 0.09 | 0.26 | $+0.65$ |
| HEAVY MAKEUP | 38.6 | 0.03 | 0.70 | $+0.27$ | 0.08 | 0.32 | $+0.60$ |
| *Added at ten* | | | | | | | |
| RECEDING HAIRLINE | 7.7 | 0.53 | 0.22 | $+0.25$ | 0.07 | 0.53 | $+0.40$ |
| BAGS UNDER EYES | 20.4 | 0.41 | 0.53 | $+0.06$ | 0.39 | 0.51 | $+0.10$ |
| *Added at twelve* | | | | | | | |
| BALD | 2.4 | 0.23 | 0.00 | $+0.77$ | 0.01 | 0.10 | $+0.89$ |
| POINTY NOSE | 27.5 | 0.01 | 1.00 | $-0.01$ | 0.00 | 1.00 | $+0.00$ |

**Errors that depend on the image.** Assumption 1 fails where the error rate differs between groups of images with the same label. Over the 20,000 images, the stronger annotator's false positive rate on HEAVY MAKEUP is 0.28 on images not labeled MALE and 0.01 on images labeled MALE, while on SMILING it is 0.10 in both groups. A single index per attribute averages over such groups, which is why Section 4.4 treats the error as a function of the image.

## D Experimental Details

This appendix gives what is needed to reproduce every experiment: the data, how each condition is built, the metrics and statistics, the training settings, the code, and the compute.

**Data.** Every CelebA model of Sections 5 and 6 trains on the first 17,000 of the first 20,000 aligned images and is scored on the other 3,000. DEAR center-crops to $128 \times 128$ and resizes to $64 \times 64$; CausalVAE and SCM-VAE use $128 \times 128$ images, as their released code does. Pendulum's continuous labels are divided into five equal-width bins. Its two corrupted settings reassign 70% of the light position's or of the shadow position's labels uniformly to the other bins.

**Conditions.** A missing label is left out of the supervised term.

- *Human labels on $m$ images* keep CelebA's labels on a nested set of $m$ training images and leave the rest unlabeled.
- *A relabeled attribute* takes the annotator's answer on every training image and keeps human labels on the others; relabeling BAGS UNDER EYES changes 7,114 of its 17,000 labels.
- *Forward correction* (Patrini et al., 2017) applies the measured error rates inside the loss. *Learned weights* (Kendall et al., 2018) learn one loss weight per attribute during training. *Posterior relabeling* replaces each answer with the probability that the attribute is present given it, by Bayes' rule from the measured rates and the prevalence in the 20,000 images.
- *The targeted rule* keeps the annotator's labels everywhere and overwrites them with human labels on 500 images for each of the seven attributes with $J_j < 0.5$, 3,500 human labels in all; *the weighted rule* does the same with each annotator label weighted by $J_j$.

Table 5: The questions put to the annotators, verbatim. Wording 1 labels the training set; all three are asked in calibration. Wording 3 asks each attribute in the opposite direction ($-$), and its answer is inverted, so a model inclined to answer "Yes" disagrees with itself.

| Attribute | Wording | Sign | Question |
|---|---|---|---|
| SMILING | 1 | $+$ | Is this person smiling? |
| | 2 | $+$ | Does this person have a smile on their face? |
| | 3 | $-$ | Is this person's expression neutral or unsmiling? |
| MALE | 1 | $+$ | Is the person in this photograph male? |
| | 2 | $+$ | Is this person a man? |
| | 3 | $-$ | Is this person a woman? |
| HIGH CHEEKBONES | 1 | $+$ | Does this person have high, prominent cheekbones? |
| | 2 | $+$ | Are this person's cheekbones clearly defined and set high on the face? |
| | 3 | $-$ | Is this person's midface flat, without prominent cheekbones? |
| MOUTH OPEN | 1 | $+$ | Are this person's lips parted, showing a gap between the upper and lower lip? |
| | 2 | $+$ | Is this person's mouth at least slightly open? |
| | 3 | $-$ | Are this person's lips fully closed together? |
| NARROW EYES | 1 | $+$ | Does this person have narrow eyes, with a small vertical opening between the eyelids? |
| | 2 | $+$ | Are this person's eyes squinted or narrowed? |
| | 3 | $-$ | Are this person's eyes wide open? |
| CHUBBY | 1 | $+$ | Does this person have a round, full or heavy face? |
| | 2 | $+$ | Would you describe this person's face as chubby? |
| | 3 | $-$ | Is this person's face thin or narrow? |
| YOUNG | 1 | $+$ | Is the person in this photograph under 40 years old? |
| | 2 | $+$ | Would you describe this person as young? |
| | 3 | $-$ | Is this person middle aged or older? |
| HEAVY MAKEUP | 1 | $+$ | Is this person wearing heavy makeup? |
| | 2 | $+$ | Does this person have noticeable makeup, such as strong lipstick, eyeshadow or foundation? |
| | 3 | $-$ | Is this person wearing little or no makeup? |
| RECEDING HAIRLINE | 1 | $+$ | Does this person have a receding hairline, with the hair pulled back at the temples? |
| | 2 | $+$ | Has this person's hairline moved back from the forehead? |
| | 3 | $-$ | Does this person have a full hairline low on the forehead? |
| BAGS UNDER EYES | 1 | $+$ | Does this person have bags or puffiness under the eyes? |
| | 2 | $+$ | Is there visible swelling or dark shadowing beneath this person's eyes? |
| | 3 | $-$ | Is the skin under this person's eyes smooth and unpuffy? |
| BALD | 1 | $+$ | Is this person bald or balding, with visibly little or no hair on the top of the head? |
| | 2 | $+$ | Does this person have significant hair loss on the top of the head? |
| | 3 | $-$ | Does this person have a full head of hair? |
| POINTY NOSE | 1 | $+$ | Does this person have a pointy nose, narrowing to a sharp tip? |
| | 2 | $+$ | Is the tip of this person's nose narrow and pointed? |
| | 3 | $-$ | Is this person's nose broad or rounded at the tip? |

- *Simulated labels* flip each human label with the annotator's measured FPR or FNR, which matches the annotator's positive rate on every attribute to within 0.03.

**Metrics.** MIC and TIC use minepy (Albanese et al., 2013) with $\alpha = 0.6$ and $c = 15$, latent variable $j$ scored against attribute $j$ and divided by $H(Y_j)$. The DCI scores fit a gradient-boosted regressor

Table 6: Every test the paper reports, grouped by section; $n$ counts seeds, or attributes for a correlation. Differences are in points of normalized MIC (MIC on Pendulum) unless marked, paired by seed, with $\pm$ half a 95% Student-$t$ interval where given. Holm $p$ is given for corrected families.

| Comparison | $n$ | Estimate | $p$ | Holm $p$ |
|---|---|---|---|---|
| *Scaling, all attribute counts pooled (Section 6.1)* | | | | |
| Human minus stronger annotator | 25 | $+10.6 \pm 2.3$ | $2 \times 10^{-9}$ | – |
| Human minus weaker annotator | 25 | $+17.6 \pm 2.0$ | $1 \times 10^{-15}$ | – |
| Stronger minus weaker annotator | 25 | $+7.0 \pm 1.7$ | $2 \times 10^{-8}$ | – |
| Human minus stronger, TIC | 25 | $+10.1 \pm 2.4$ | $5 \times 10^{-9}$ | – |
| Human minus stronger, DCI disentanglement | 25 | $+0.12 \pm 0.03$ | $1 \times 10^{-7}$ | – |
| Human minus stronger, DCI completeness | 25 | $+0.09 \pm 0.04$ | $2 \times 10^{-4}$ | – |
| Human minus stronger, DCI informativeness | 25 | $+0.06 \pm 0.03$ | $6 \times 10^{-4}$ | – |
| Human minus stronger, probe AUROC | 25 | $+0.04 \pm 0.02$ | $1 \times 10^{-4}$ | – |
| Stronger annotator minus human labels, 4,000 images | 5 | $+3.9 \pm 3.0$ | 0.023 | – |
| *The index as a price, twelve attributes (Section 6.2)* | | | | |
| % of human score against $J^2$, Spearman | 12 | 0.76 | 0.004 | – |
| Change in % of human against change in $J^2$, Pearson | 12 | 0.91 | $4 \times 10^{-5}$ | – |
| *Locality: the relabeled attribute's own change (Section 6.3)* | | | | |
| HIGH CHEEKBONES | 5 | $-31.3 \pm 4.2$ | $3 \times 10^{-5}$ | – |
| BAGS UNDER EYES | 5 | $-10.5 \pm 2.2$ | $2 \times 10^{-4}$ | – |
| SMILING | 5 | $+2.4 \pm 15.7$ | 0.70 | – |
| YOUNG after relabeling BAGS UNDER EYES | 5 | $+3.6$ | – | 0.04 |
| *Locality: weak attributes unlabeled, each kept attribute* | | | | |
| BALD | 5 | $-5.1 \pm 7.7$ | 0.14 | 0.56 |
| HEAVY MAKEUP | 5 | $-7.1 \pm 4.9$ | 0.016 | 0.08 |
| MALE | 5 | $-2.6 \pm 16.7$ | 0.69 | 1.00 |
| SMILING | 5 | $-3.5 \pm 22.4$ | 0.69 | 1.00 |
| YOUNG | 5 | $-5.5 \pm 10.9$ | 0.24 | 0.71 |
| *Locality: Pendulum, 70% of one factor's labels corrupted* | | | | |
| *light position corrupted* | | | | |
| light position (corrupted) | 5 | $-30.8$ | 0.002 | – |
| pendulum angle | 5 | $-7.9$ | 0.041 | – |
| shadow length | 5 | $+3.6$ | 0.051 | – |
| shadow position | 5 | $-2.5$ | 0.040 | – |
| *shadow position corrupted* | | | | |
| light position | 5 | $-3.4$ | 0.069 | – |
| pendulum angle | 5 | $-5.4$ | 0.014 | – |
| shadow length | 5 | $-1.1$ | 0.47 | – |
| shadow position (corrupted) | 5 | $-28.0$ | $3 \times 10^{-5}$ | – |
| *Corrections minus uncorrected labels, twelve attributes (Section 6.4)* | | | | |
| Forward correction | 5 | $+1.1 \pm 2.1$ | 0.21 | 1.00 |
| Learned per-attribute weights | 5 | $+0.8 \pm 1.4$ | 0.20 | 1.00 |
| Posterior relabeling | 5 | $-1.8 \pm 4.3$ | 0.31 | 1.00 |
| 3,500 targeted human labels | 5 | $+0.1 \pm 2.2$ | 0.89 | 1.00 |
| The same, weighted by $J$ | 5 | $-0.9 \pm 2.6$ | 0.37 | 1.00 |
| Weak attributes unlabeled | 5 | $-3.0 \pm 5.1$ | 0.18 | 1.00 |
| Supervision weight 1 against 5 | 3 | $-0.1 \pm 4.1$ | 0.95 | – |
| Supervision weight 10 against 5 | 3 | $-0.5 \pm 0.4$ | 0.038 | – |
| *Targeted rules on the seven targeted attributes, six and twelve pooled* | | | | |
| Human labels minus uncorrected | 10 | $+14.1 \pm 3.0$ | $2 \times 10^{-6}$ | – |
| Targeted minus uncorrected | 10 | $+0.2 \pm 1.8$ | 0.76 | – |
| Weighted targeted minus uncorrected | 10 | $+0.4 \pm 1.9$ | 0.68 | – |
| Weighted minus unweighted | 10 | $+0.1 \pm 2.2$ | 0.90 | – |
| *Real against simulated errors, twelve attributes (Section 6.5)* | | | | |
| Simulated minus annotator labels | 5 | $-8.4 \pm 7.5$ | 0.035 | – |
| The same, seeds whose simulated run trained | 3 | $-4.2 \pm 2.4$ | 0.018 | – |
| Difference against $J$ across attributes, Spearman | 12 | $-0.87$ | $3 \times 10^{-4}$ | – |
| *Attribute sets (Section 6.6)* | | | | |
| First set, human minus annotator | 5 | $+13.8 \pm 10.5$ | 0.022 | – |
| Second set, human minus annotator | 5 | $+4.3 \pm 1.7$ | 0.002 | – |
| Second-set gap minus first-set gap | 5 | $-9.5 \pm 11.2$ | 0.078 | – |

per attribute on 70% of the held-out images; disentanglement and completeness come from its feature importances, and informativeness is its $R^2$ on the other 30%. The probe is a logistic regression from the supervised latent variables to each attribute under 5-fold stratified cross-validation, with its AUROC averaged over attributes.

**Statistics.** Conditions share seeds 42 to 46, or 42 to 44 for the supervision weight, and are compared paired by seed with a two-sided Student-$t$ interval unless a one-sided test is stated. Holm's correction is applied across the eleven untouched attributes of each relabeling, across the five attributes that keep labels when the weak ones are removed, and across the six correction rules. Runs that under-trained or collapsed are kept. Table 6 lists every test the paper reports, with its $p$ value, and shows that none of the six correction rules and none of the attributes kept after the weak labels are removed is significant after Holm correction.

**Training.** Table 7 gives the settings of every model. DEAR uses its ResNet encoder and implicit generator, one discriminator update per generator update, and $\lambda = 5$, the value in DEAR's released scripts. CausalVAE on CelebA first learns its adjacency from the labels alone for 10 epochs. DEAR's Adam uses $\beta = (0, 0.999)$ and, for the discriminator, a learning rate of $10^{-4}$; its adjacency matrix alone uses $\beta = (0.9, 0.999)$, at a learning rate of $10^{-3}$. The other models use $\beta = (0.9, 0.999)$.

Table 7: Training settings. Where two values are given, the first is for CelebA and the second for Pendulum and Flow; DEAR supervises the first $k$ of its 100 latent variables.

| Setting | DEAR | CausalVAE | SCM-VAE | ICM-VAE |
|---|---|---|---|---|
| Datasets | CelebA | all three | all three | Pendulum, Flow |
| Input size | $64^2$ | $128^2$ / $96^2$ | $128^2$ / $96^2$ | $96^2$ |
| Supervised latents | $k$ of 100 | $4 \times 32$ / $4 \times 4$ | $4 \times 32$ / $4 \times 4$ | $4 \times 4$ |
| Optimizer | Adam | Adam | Adam | Adam |
| Learning rate | $5 \times 10^{-5}$ | $10^{-3}$ | $10^{-3}$ | $10^{-3}$ |
| Batch size | 128 | 64 | 64 | 64 |
| Epochs | 200 | 80 / 101 | 100 | 101 |

**Code.** CausalVAE's released code covers Pendulum and Flow; for CelebA we implement the encoder and decoder of Table 2 of its supplement and keep the released objective. SCM-VAE's released code does not run on CelebA: we wire in the convolutional encoder and decoder it ships and remove three widths hard-coded for Pendulum, introducing no new architecture. ICM-VAE runs from its released training scripts on Pendulum and Flow. The audit's CelebA runs keep each code path's own sample and split of 20,000 images.

**Numerics.** GPUs compute some matrix products in reduced precision by default (TF32), and some kernels are nondeterministic. Both change results slightly, so every DEAR run disables TF32 and requests deterministic algorithms, and every CausalVAE run requests deterministic algorithms; SCM-VAE and ICM-VAE keep their released settings.

**Runs and compute.** DEAR accounts for 188 runs, CausalVAE and SCM-VAE for 15 each on CelebA-Smile, and CausalVAE for 15 on Pendulum, three settings over five seeds. The audit of Section 7 adds 15 CausalVAE, 20 SCM-VAE and 10 ICM-VAE runs, 278 runs in all. Each DEAR run trains on one GPU, taking 645 GPU-hours in all, a median of 2.87 hours per run, on NVIDIA L40S (101), Quadro RTX 6000 (72) and Quadro RTX 8000 (15) GPUs.

**Assets.** The annotators Qwen3.5-9B and Qwen2.5-VL-7B and the released code of DEAR and CausalVAE are used under the Apache 2.0 license; SCM-VAE's and ICM-VAE's released code carries no license file. CelebA is used under its terms for non-commercial research.

# E ADDITIONAL RESULTS

This appendix gives the results that support Sections 6 and 7 in more detail: counterfactuals, per-attribute breakdowns, the two further architectures, the learned graphs and the scoring choices.

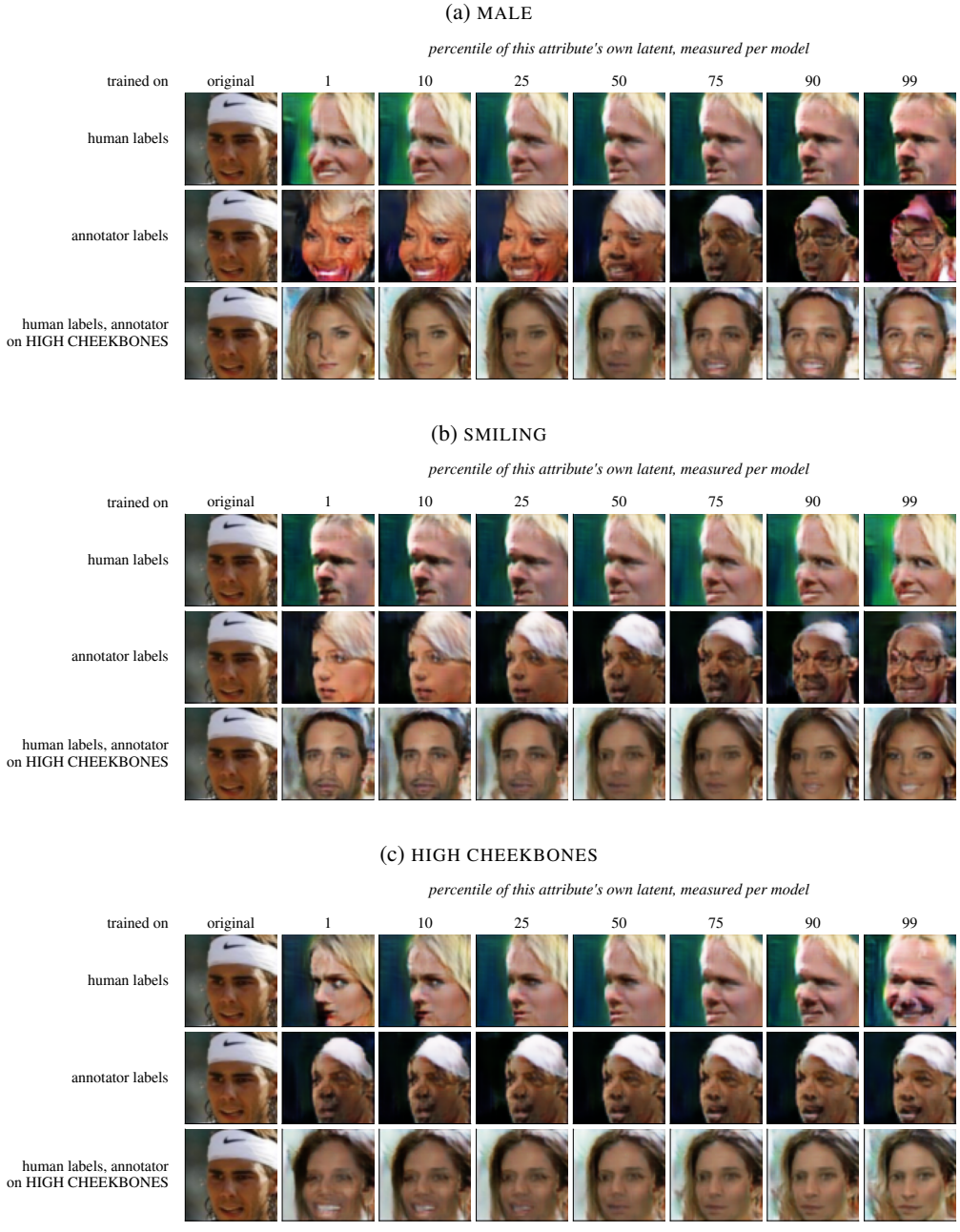


Figure 6: **Counterfactuals.** DEAR at twelve attributes, seed 42, on one held-out test face drawn at random. Each column sets the named attribute's latent variable to a percentile of its distribution over 200 held-out faces, measured per model, and recomputes its descendants through the learned causal model. Rows are trained on human labels, on the stronger annotator's labels, and on human labels with HIGH CHEEKBONES alone relabeled.

### E.1 COUNTERFACTUALS

Each baseline this work builds on shows counterfactuals or interventions for its own model. Figure 6 does the same for DEAR at twelve attributes under three label sources: it sets one attribute's latent variable to a percentile of its distribution and lets the attribute's descendants follow through the learned causal model. At DEAR's $64 \times 64$ resolution the generated faces do not preserve the identity of the input face, so the figure illustrates what each model generates, and no claim rests on it.

### E.2 PER-ATTRIBUTE RESULTS

Figure 7 breaks the scaling curve of Figure 2a down by attribute; an attribute appears from the count at which it enters the nested sets. The gap to human labels is not uniform. At twelve attributes, in seed means, human labels score 40.8 on HIGH CHEEKBONES against 10.2 for the stronger annotator, and 44.4 against 22.0 on MOUTH OPEN. On MALE and BALD, the two attributes the annotator answers best, its labels score above human labels, 66.1 against 64.0 and 85.5 against 74.6.

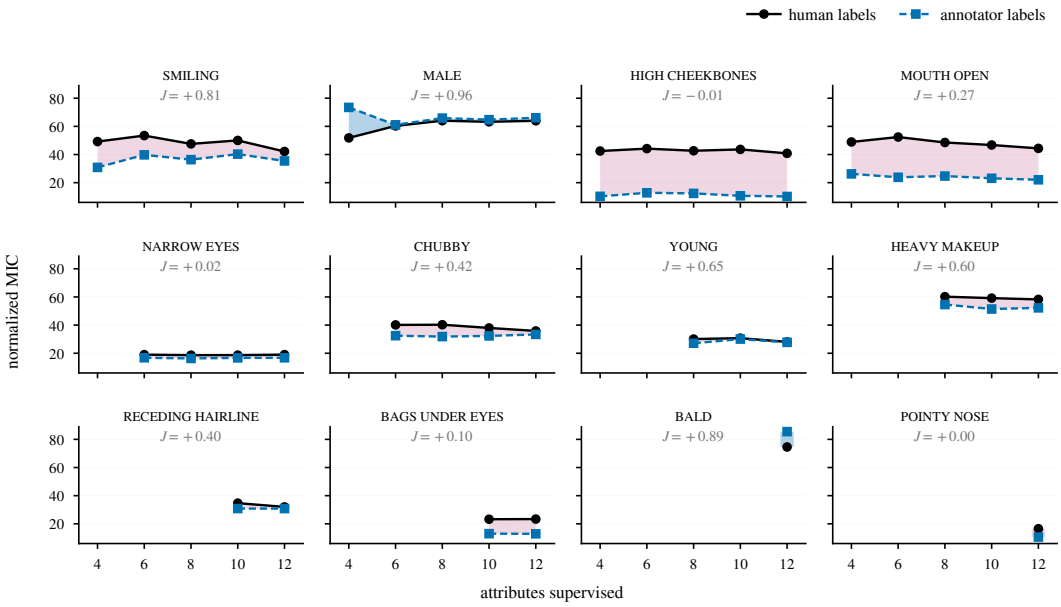


Figure 7: **The scaling curve by attribute.** Normalized MIC per attribute, seed means, from the count at which each attribute enters the nested sets. Rose shading marks where human labels lead the stronger annotator, blue where the annotator leads.

Figure 8 gives the comparison of Section 6.5 per attribute. Simulated errors cost most where the annotator is accurate: the difference averages $-15.6$ on the five attributes it answers well and $-0.2$ on the four it answers near chance, a Spearman correlation of $-0.87$ between index and difference across the twelve. Two of the five simulated runs did not train, with DCI informativeness of $-0.01$ and 0.02 against at least 0.19 for every run on the annotator's labels; over the other three seeds the difference is $-4.2 \pm 2.4$. With the human-trained representation as the feature, the annotator's effective index exceeds its measured index by 0.26 on average over the five well-answered attributes, while the simulated labels' falls short of it by 0.02, the direction Proposition 3 predicts. The size of that excess does not predict the size of the cost difference across attributes (Spearman 0.40).

### E.3 FURTHER ARCHITECTURES

Figure 9 repeats the four-attribute comparison of Table 2 and adds one relabeled attribute. With the annotator's labels on all four attributes, each architecture loses to its own human-label baseline, with every interval excluding zero. Relabeling NARROW EYES alone, which the stronger annotator answers near chance, gives intervals that include zero for both models: SCM-VAE moves by $-0.2$, and CausalVAE's mean of $-18.7$ is driven by the one seed the caption names.

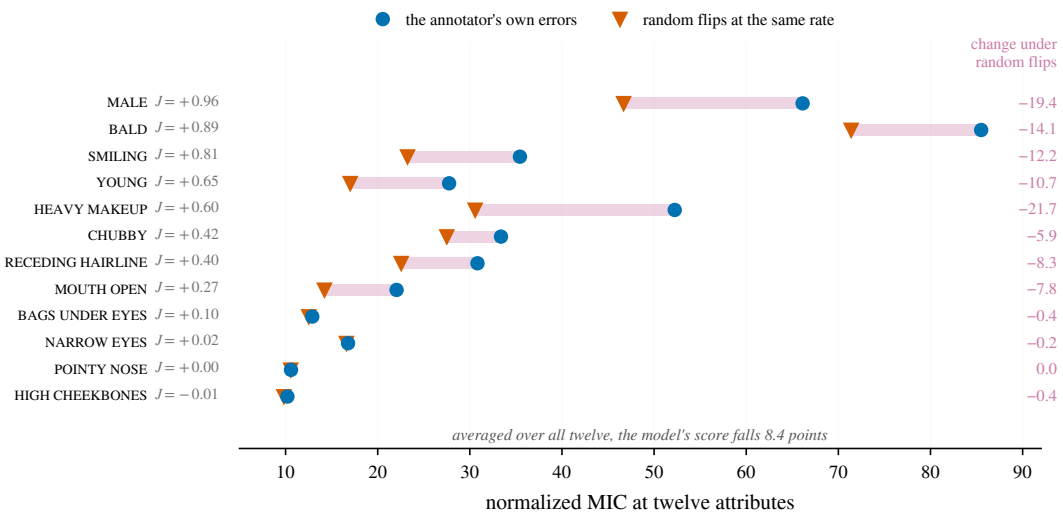


Figure 8: **Matching the error rate does not match the cost.** Normalized MIC per attribute for DEAR trained on the stronger annotator's labels and on human labels flipped at its per-attribute rates. Rows run from the best-answered attribute to the four near chance, where the two coincide.

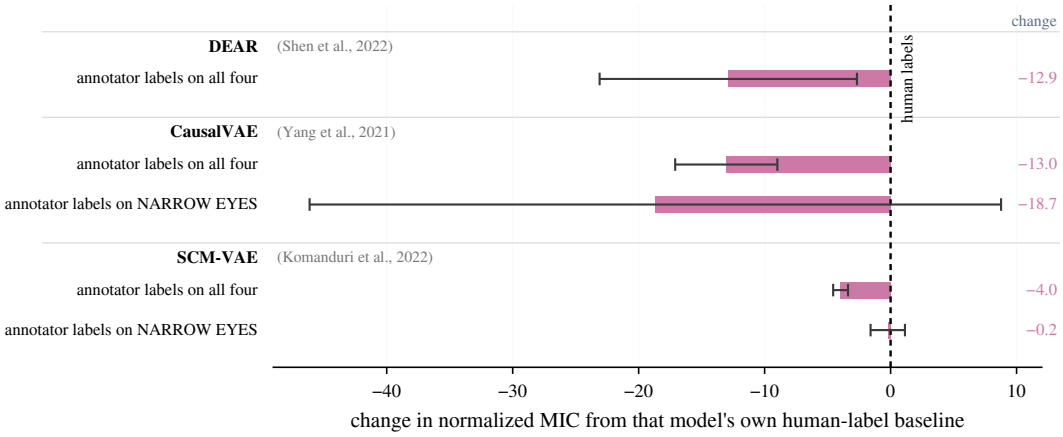


Figure 9: **Beyond DEAR.** Change in normalized MIC from each model's own human-label baseline at four attributes, with 95% intervals, for the stronger annotator's labels on all four attributes and on NARROW EYES alone. DEAR's four include HIGH CHEEKBONES where the other two have NARROW EYES. One CausalVAE seed scores 11.9, against 47.7 to 60.3 for the others, which widens its single-attribute interval.

### E.4 Learned graphs

The graphs DEAR learns support no structural claim, which is why none is made. Two seeds of one condition agree on the ranking of edge weights about as well as two conditions on one seed, mean Spearman 0.40 to 0.55 against 0.37 to 0.54 (Figure 10), so a difference between conditions' graphs cannot be told apart from seed-to-seed variation.

### E.5 Scoring choices and reproduction

Figure 11 isolates each choice of Section 7 by varying it alone with the others held at our convention. The reduction over factors moves MIC most, by a median of 24.0 points across the eight model and dataset pairs, and matching latent variables to factors by index or by assignment moves it least. Our scoring code reproduces published values where a check is possible: CausalVAE's Pendulum




Figure 10: **Learned graphs.** (a) Edge weights DEAR learns at twelve attributes under human labels, the stronger annotator's labels, and HIGH CHEEKBONES alone relabeled, averaged over seeds; the ordering fixes the lower triangle at zero. (b) Rank agreement between two learned adjacencies, for two seeds of one condition and for two conditions on one seed.

MIC to within 2.2 points and its TIC to within 0.3 under our convention, and ICM-VAE's DCI disentanglement and completeness on Flow to within 0.01; on Pendulum it falls short of ICM-VAE's values by 0.06 and 0.23. Across the 100 binary-label score matrices of the audit, raw MIC never exceeds its label's entropy ceiling, as Proposition 5 requires.

| *scoring choice* | reduction over factors | target | represen-tation | MIC parameter $\alpha$ | normal-ization | aggregation | matching | largest |
|---|---|---|---|---|---|---|---|---|
| **CausalVAE** | (Yang et al., 2021) | | | | | | | |
| Pendulum | 4.6 | 14.8 | 31.3 | 3.3 | 10.4 | 3.0 | 0.0 | 31.3 |
| Flow | 35.1 | 9.1 | 11.1 | 6.1 | 4.9 | 15.3 | 0.2 | 35.1 |
| CelebA-Smile | 18.8 | 1.6 | 1.5 | 8.0 | 3.1 | 2.5 | 0.0 | 18.8 |
| CelebA-Beard | 42.8 | 4.6 | 7.2 | 6.2 | 25.7 | 0.9 | 0.0 | 42.8 |
| **SCM-VAE** | (Komanduri et al., 2022) | | | | | | | |
| Pendulum | 19.8 | 9.7 | 16.2 | 3.5 | 4.8 | 0.7 | 0.0 | 19.8 |
| Flow | 21.6 | 0.8 | 4.5 | 7.2 | 3.5 | 5.5 | 0.0 | 21.6 |
| CelebA-Smile | 26.3 | 13.2 | 7.1 | 8.0 | 2.8 | 1.2 | 0.0 | 26.3 |
| CelebA-Beard | 43.3 | 9.7 | 1.5 | 7.2 | 18.7 | 1.2 | 1.6 | 43.3 |
| median effect | 24.0 | 9.4 | 7.2 | 6.7 | 4.8 | 1.9 | 0.0 | |

*points the score moves when this choice alone is varied*

Figure 11: **Which choice moves the score.** Points by which each scoring choice moves MIC on each model and dataset, with the others held at our convention; columns are ordered by median effect. The target, which both papers state, is shown for comparison. The right column gives each row's largest single effect.