arXiv:2512.18864v1 [cs.CV] 21 Dec 2025

# 跨模态反事实解释：揭示主观分类中的决策因素与数据集偏差

Alina Elena Baia、Andrea Cavallaro
洛桑联邦理工学院，洛桑，瑞士{alina.baia,
andrea.cavallaro}@epfl.ch

## 摘要

概念驱动的反事实解释通过语义变化改变模型预测，从而解释分类器的决策。本文提出一种新颖方法，利用跨模态分解性和图像特定概念，生成以自然语言表达的反事实场景。我们将这一可解释性框架命名为 *Decompose and Explain*（*DeX*），并应用于颇具挑战性的图像隐私决策领域，该领域的决策具有情境性和主观性。此应用能够量化关键场景元素对模型预测的差异化贡献。我们通过多准则选择机制识别相关决策因素，该机制同时考虑图像相似度以实现最小扰动，并考量决策置信度以优先选择影响显著的变化。该方法能够评估和比较多种解释，并考察解释属性之间的相互依赖和相互影响。借助图像特定概念，*DeX* 可生成基于图像且稀疏的解释，显著优于最先进水平。重要的是，*DeX* 是一种无需训练的框架，具有很高的灵活性。结果表明，*DeX* 不仅揭示了影响主观决策的主要因素，还识别出潜在的数据集偏差，从而能够制定有针对性的缓解策略，以提升公平性。

## 1. 引言

语言比图像更能有效地传达分享包含私人信息的照片所带来的风险 [10]。要生成有效的基于文本的解释，必须建立视觉模态与文本模态之间的映射，使模型能够以通俗易懂、便于人类阅读的方式表达其对图像的感知。

复合概念的上下文文本嵌入可以近似表示为其构成因素的向量表示的线性组合 [45]。大规模图像-文本多模态嵌入表现出类似于词嵌入 [11, 33, 36]的线性组合性 [23]。这种涌现的算术属性在联合表示模型 [14, 50]中跨模态延伸。这种对齐使得对嵌入进行简单的算术运算（例如，将文本向量与图像向量相加或相减）即可执行跨模态任务，而无需额外训练。

本文提出 DeX（分解与解释），这是一种免训练框架，利用跨模态组合性为图像隐私分类提供基于文本的反事实解释。DeX 将图像分解为以标签表示的关键语义元素，并通过潜在空间中的跨模态算术生成合理的替代场景[1]。DeX 评估这些场景构成变化对隐私预测的影响，并根据一组理想属性量化图像元素的相关性。我们的主要贡献如下：

•提出一种新颖的跨模态（解）组合策略，用于解释隐私分类器决策。

• 提出一种免训练框架，可生成满足关键属性（可行性、稀疏性和有效性）的基于文本的反事实解释。

• 对解释进行多标准评估，纳入了以往被忽视的方面，例如有效性、置信度、概念扎根性（可行性）和相似性。

• 应用 DeX 分析数据集内容，揭示了一个广泛使用的数据集中存在显著的主题偏差。

## 2. 相关工作

反事实解释（CEs） [47] 定义了能够翻转模型预测 [19] 的最小语义变化 [20–22, 42]。由于基于像素的解释（例如热力图）信息量不足，且非专业用户难以理解 [30, 32], ，近期研究开始聚焦于在生成模型的潜在空间中生成有意义的变化

[1]尽管扩散模型能够生成视觉解释，但它们需要引导式编辑和辅助技术（例如文本反演），从而带来与训练、优化和调优相关的重大负担 [22]。相比之下，DeX 利用多模态嵌入的算术性质直接生成解释，消除了这些额外开销。

模型 [19, 26, 31, 53], , 包括变分自编码器 [42] 和扩散模型 [3, 21, 22]。值得注意的是，ACE [21] 利用扩散模型将对抗噪声转换为语义扰动。TIME [22] 使用文本反演学习上下文词元和类别词元，从而引导基于扩散的反事实生成。StylEx [31] 通过训练特定于分类器的 StyleGAN2 [24],并操纵其 StyleSpace [51]，实现实例级解释。StylEx 的主要缺点是针对每个分类器训练 GAN 的计算成本高，以及依赖手动概念标注。其他方法 [34, 49]使用预训练生成器进行模型诊断。ZOOM [34]利用 CLIP 引导的文本提示差异来引导图像生成，而 UMO [49] 则寻找并标注具有影响力的编辑方向。一种无需训练的替代方法 DiffEx [25], 利用视觉语言模型提取分层的领域属性，随后引导现成的扩散模型生成反事实样本。然而，由于生成模型的局限性，此类方法面临多项挑战，包括计算成本、视觉伪影、生成失败（例如缺失物体）以及继承的偏差 [17, 32, 40]。 *DeX* 摒弃了生成式方法，转而操作多模态嵌入，以进行图像表征修改。这一选择消除了通过生成模型渲染图像所带来的计算开销和偏差。

使用自然语言潜在主题描述 [4] 或采用概念瓶颈模型（CBMs） [5]训练的模型可能会增强可解释性。然而，它们的表现往往不如端到端模型 [5]。相比之下，*DeX* 以事后方式运行，确保分类器的表现不受影响。这在隐私分类等高风险任务中是一项优势。

概念激活向量（CAVs）通过在模型空间中定义相应方向，量化概念的全局重要性；该方向由人工标注的正例和负例推导而来 [27]。CAV 与基于此类样本的激活训练得到的线性分类器的决策边界正交。然而，其一大缺点是依赖成本高昂的人工监督，且容易受到标注偏差的影响。无监督变体通过多分辨率图像分割和聚类生成概念，从而缓解了这一问题 [13]。*DeX* 则提供实例级解释，指出驱动每张图像分类决策的概念。

从图像集 [1,2] 或文本提示 [29]中提取的概念方向（概念激活向量）可用作无需生成的方法来生成反事实解释。概念反事实解释（CCEs） [1]方法学习分数，以沿概念方向对样本的嵌入进行最小幅度调整，从而改变预测结果。还可以识别出其纳入或移除会显著影响分类的最小关键概念子集 [2]。CounTEX [29] 推导出通过中性（即不包含任何概念）文本提示与特定概念文本提示的嵌入之间的向量差，在 CLIP空间中构建概念方向 [12, 28]。与 [1], 一样，通过概念激活向量（CAVs）的加权和对图像嵌入进行反事实扰动，以修改预测结果。尽管在结构化数据集（例如 AwA2 [52], CUB [48]）上取得了成功，CounTEX [29] 在复杂、非结构化的隐私数据集 [39, 54]上无法生成稀疏、可行且多样的解释。其质量取决于预定义的概念集和优化过程，从而导致解释坍塌，以及沿无关概念方向产生非稀疏扰动等问题。*DeX*区别于基于集合的概念方法 *[1, 32]* ，仅对输入图像进行操作。此外，为了扰动嵌入，*DeX*克服了*CounTEX* *[29]*的局限，采用自动提取的、特定于图像的概念，而非固定的通用概念集。这确保了解释以图像为依据，具有高度可行性，并提升了图像间的多样性。

## 3. 问题定义

设 $\mathcal{D} = \{(I,\ y)\}$ 为一组图像-标签对，其中标签 $y \in \mathcal{Y} = \{pr, pu\}$ 用于定义图像 $I$ 是私密还是公开。设 $E_I$ 为图像编码器， $E_T$ 为文本编码器，它们将各自模态嵌入到一个联合多模态空间中。设 $x = E_I(I) \in \mathbb{R}^d$ 为图像 $I$ 的嵌入，设 $f\colon \mathbb{R}^d \to \mathbb{R}$ 为一个隐私分类器，它将图像嵌入 $x$ 映射为预测 $\tilde{y} \in \mathcal{Y}$。

对于每张图像 $I$， $\mathcal{C}_I = \{c_1, \ldots, c_k\}$ 是一组用自然语言表达的图像特定概念。概念类型以及 $k$（即 $\mathcal{C}_I$ 的基数）取决于图像内容。我们的目标是通过探索（替代性）反事实情景，评估从 $I$ 中移除某个概念 $c_i \in \mathcal{C}_I$ 对解释隐私决策 $\tilde{y} = pr$ 的影响。我们旨在根据图像 $I$ 的概念 $c_i$ 修改其嵌入，并找到一个反事实解释 (CE) $\hat{}x$，使分类器的输出 $f(\hat{}x)$ 与原始预测 $\tilde{y}$: $f(\hat{}x) \neq f(x)$ 不同。如果移除 $c_i$ 会改变预测，那么它就构成一个反事实解释。

由于缺乏合适的基于概念的数据集，且必须平衡多个往往相互竞争的目标及其相关权衡 [16, 20–22, 38, 42, 43, 46, 47]，评估反事实解释颇具挑战。在本研究中，我们将可行性解释为依据性：若反事实解释以图像为依据，则称其可行。若反事实解释能改变分类结果，则称其有效 [21, 22, 29, 38, 42, 43, 47], ; 若其仅修改最少数量的属性或特征，则称其稀疏 [1, 20–22, 29, 38, 42, 43], ; 若图像（嵌入）[2] 与原图 [20–22, 29, 38, 42], 保持相似，则称其邻近 ; 若

[2]稀疏性和邻近性会尽量减少原始嵌入与其反事实之间的差异：稀疏的反事实（即只改变少数特征）通常会与原始实例相似。然而，如果改动幅度过大，反事实也可能稀疏但并不相似，或者很近处

解释彼此不同 [20, 21, 38, 42, 43]。此外，我们提出评估反事实的 置信度，它量化了分类器对预测的反事实类别的支持程度。

设 $\mathcal{E}_I^b = \{\hat{x}_1、\hat{x}_2、\ldots\ldots、\hat{x}_N\}$ 是通过多目标过程选出的最佳预测翻转反事实样本集合，其中每个解释 $\hat{x}_i$ 由一个反事实场景 $c_i \in \mathcal{C}_I^b$表示，其中 $\mathcal{C}_I^b \subseteq \mathcal{C}_I$ （详见第 4.2 节 ）。设 $\mathcal{D}_{pr}$ 是正确分类的私有图像集合，且 $\mathcal{D}_b = \{I \in \mathcal{D}_{pr} \mid \mathcal{E}_I^b \neq \emptyset\}$。

我们将相对于 $I$ 的依据性（可行性，$F$）定义为反事实解释中存在于 $I$ 的标签所占的平均比例：

$$F = \frac{1}{\sum_{I\in\mathcal{D}_b}|\mathcal{E}_I^b|}\sum_{I\in\mathcal{D}_b}\sum_{c\in\mathcal{C}_I^b}\frac{|\theta(I,l(c))|}{|l(c)|}, \tag{1}$$

其中， $l(\cdot)$ 返回构成 $c$ 的 $t_i$ 标签列表，而 $\theta(\cdot,\ \cdot)$ 是一种开放集图像标注器，可提供在 $I$ 中检测到的 $t_i$ 标签列表。

我们将有效性（*validity*， $V$）量化为：至少存在一个预测翻转反事实样本的正确分类私有图像所占的比例：

$$V = \frac{|\mathcal{D}_b|}{|\mathcal{D}_{pr}|}. \tag{2}$$

反事实样本引发的变化越大，$V$就越高，但这与稀疏性和邻近性的要求相冲突。我们将在 $F$ 和 $V$第5 节中以百分比的形式讨论结果。

我们将稀疏性（ $S$）度量为生成有效反事实样本所用标签数量的平均值：

$$S = \frac{1}{\sum_{I\in\mathcal{D}_b}|\mathcal{E}_I^b|}\sum_{I\in\mathcal{D}_b}\sum_{\hat{x}\in\mathcal{E}_I^b} g(\hat{x}), \tag{3}$$

其中， $g(\hat{x})$用于计算生成$\hat{x}$所用概念的数量。强制要求稀疏性可能会损害有效性。为了评估邻近性（ $P$），我们计算图像嵌入 $x$ 与其反事实样本$\hat{x}$ 的图像嵌入之间的平均余弦相似性cos(·)，涵盖 $\mathcal{E}_I^b$中的所有解释：

$$P = \frac{1}{\sum_{I\in\mathcal{D}_b}|\mathcal{E}_I^b|}\sum_{I\in\mathcal{D}_b}\sum_{\hat{x}\in\mathcal{E}_I^b}\cos(E_I(I),\hat{x}). \tag{4}$$

反事实类别中新预测的平均置信度（ $C$）为：

$$C = \frac{1}{\sum_{I\in\mathcal{D}_b}|\mathcal{E}_I^b|}\sum_{I\in\mathcal{D}_b}\sum_{\hat{x}\in\mathcal{E}_I^b} p(\hat{x}), \tag{5}$$

其中， $p(\hat{x})$ 是预测 $f(\hat{x})$ 的置信度。

在进行许多微小修改时，相似性会提高，但稀疏性则不会。因此，同时评估两者有助于深入了解这些解释方法。

我们通过计算数据集中针对同一 $I$ 的解释之间的成对余弦相似度的平均值，来衡量多样性，即 $D$:

$$D = \frac{1}{N(N-1)|\mathcal{D}_b|}\sum_{I\in\mathcal{D}_b}\sum_{\substack{c_i,c_j\in\mathcal{C}_I^b\\ i<j}}\big(1-\cos(E_T(c_i),E_T(c_j))\big), \tag{6}$$

其中 $N = |\mathcal{C}_I^b|$。 $D$ 值越高，表示解释之间的多样性越高。此外，为确保不同图像对应的解释不会重复，我们检测解释坍塌。对于所有图像，我们计算其解释的质心：

$$\bar{c}_I = \frac{1}{N}\sum_{c_i\in\mathcal{C}_I^b} E_T(c_i), \tag{7}$$

然后，我们计算它们的平均两两相似度：

$$R = \frac{1}{|\mathcal{D}_b|(|\mathcal{D}_b|-1)}\sum_{I\in\mathcal{D}_b}\sum_{\substack{J\neq I\\ J\in\mathcal{D}_b}} 1-\cos(\bar{c}_I,\bar{c}_J). \tag{8}$$

$R$ 值越高，表示图像之间的解释多样性越大，对应的解释也越贴合特定图像且与语境相关。

## 4. 分解与解释（DeX）

### 4.1. 跨模态算术

我们通过观察概念以算术方式相加时嵌入如何变化，来研究嵌入空间在多大程度上支持语义组合。与 [6], 类似，我们进行了线性测试，以检验将两个输入拼接后得到的嵌入是否与各自嵌入之和相似。我们考察了文本模态：给定概念 $c_a$ 和 $c_b$，将它们拼接成 $c_{ab} =$ "$c_a, c_b$"，使用 OmniBind [50]($E$)分别对 $c_a$、 $c_b$和 $c_{ab}$进行嵌入，并计算 $E(c_{ab})$与 $E(c_a)+E(c_b)$之间的余弦相似度（$CS$）。在随机选取的 1000 对 $c_a$、 $c_b$ 上进行 3 次实验，均值（标准差）$CS$ 为 0.61（0.13）。作为参照，在线同义词词典中"汽车"的同义词之间的平均两两相似度为 0.35。我们将分析扩展到概念三元组，计算拼接后的三元组嵌入与各个单独嵌入之和之间的 $CS$ 。结果显示，平均（标准差） $CS$ 为 0.53（0.11）。这些结果表明，O mniBind 嵌入具有线性组合性。

接下来，我们通过对标题嵌入进行算术调整，将其与一个概念嵌入相结合，从而生成修改后的标题嵌入；然后将后续的图像检索结果与原始结果进行比较，以确认预期的语义变化。例如，从绿色田野中的女性 中移除一名女性，检索到的便是空旷田野的图像；从一辆中移除汽车

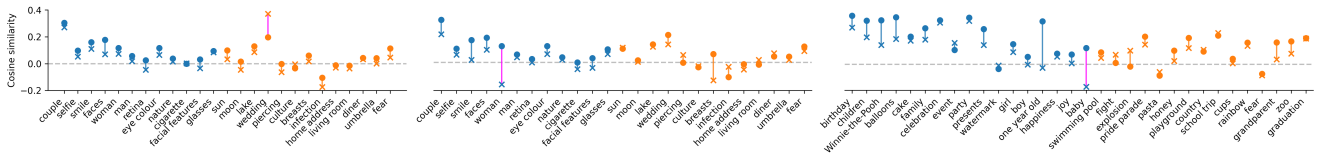


图1。跨模态算术示例。概念添加（左）选择性地提高目标相似度，同时保持与其他概念的整体相似度。概念移除（中、右）会导致目标概念及语义相关概念的相似度局部降低。关键：- ●/×: 算术运算前后的余弦相似度值；/： 相关/不相关概念；： 编辑后的概念。

红色汽车 会检索到没有汽车的红色物体；从医生进行手术 中减去进行手术，则会返回缺少手术情境的日常场景。这证实了嵌入空间支持概念的语义组合。

我们还通过研究嵌入算术，考察当向图像中添加新概念或移除已有概念时，图像与概念之间的相似性如何变化，从而探究跨模态组合（见图1）。我们随机抽取图像，并为每幅图像定义一组相关概念和无关概念。首先，我们计算原始图像与这两组概念之间的余弦相似度，以建立基线。我们假设图像与相关概念之间具有较高相似度，而与无关概念之间具有较低相似度。随后，我们重新计算概念相似度：添加一个无关概念应提高与该新增概念的相似度，同时保持与原有相关概念的相似度；移除一个相关概念应降低与该概念的相似度，同时维持与其他概念的相似度。当将新概念的文本嵌入添加到图像嵌入中时，与新增概念的余弦相似度显著提高，同时保持了与已有相关概念的相似性。例如，向一幅最初与婚礼无关的图像中添加婚礼，会使其与婚礼的相似度提高，而与其他原有概念的相似度变化很小。总体而言，添加新概念会提高与该概念的相似度，同时通常保持与已有相关概念的相似度。此外，当移除一个概念（例如，婴儿）时，修改后图像表示与该概念之间的相似度会显著降低，与其相关概念（例如，一岁儿童）之间的相似度也会降低。与无关概念之间的相似度总体保持稳定，表明该操作影响了被移除的概念及其语义相关概念，而不会影响其他概念。

## 4.2. 反事实解释

基于第 3 节的问题定义以及第4.1节对各项属性的验证，我们提出 DeX：一种基于跨模态组合性、无需训练的框架，用于提供基于文本的反事实解释。DeX 首先生成一组具有概念依据的反事实场景，然后通过嵌入运算，将模型的预测与操作后得到的嵌入表示进行对比（见图2）。

为了获得候选场景集合 $\mathcal{C}_I$，DeX 生

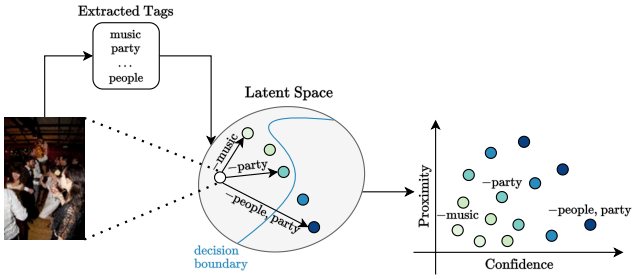


图2。DeX 通过一个三步流程，为给定图像 $I$ 生成基于概念的解释：提取概念并创建反事实场景；通过跨模态分解，在潜在空间中操纵图像表示；以及通过多准则选择，识别与隐私相关的场景（即解释）。

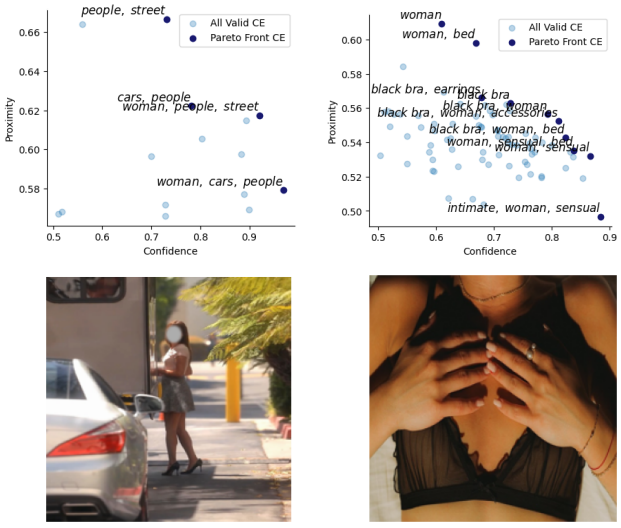


图3。反事实解释（上方）与原始图像（下方）的置信度–邻近度帕累托权衡：该图展示了相互竞争的准则之间的相互作用，以及不同概念如何影响模型决策。

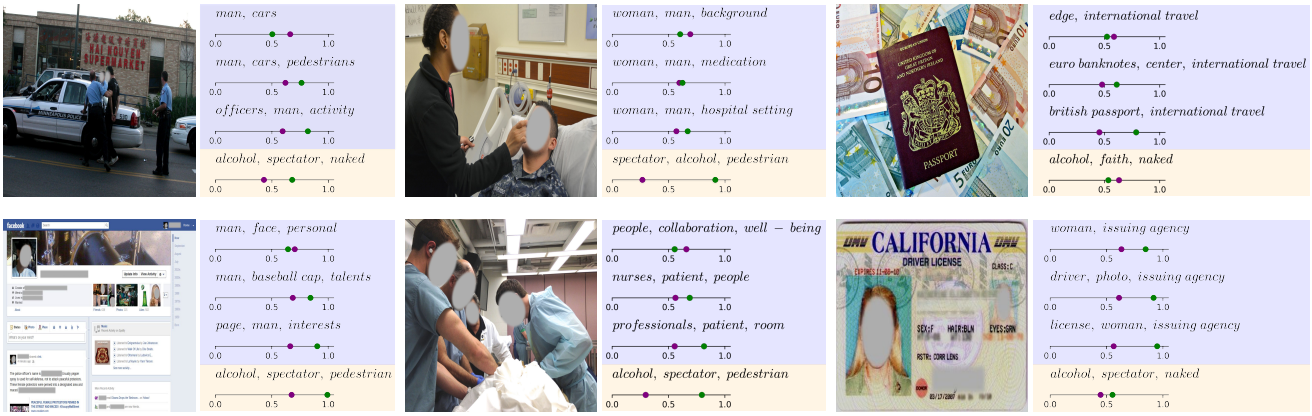


图 4. DeX（）和 CounTEX（）生成的解释示例。评估标准置信度（$C$、●）以及通过余弦相似度衡量的邻近性（$P$、●）。请注意，CounTEX 生成的解释 [29]（）内容重复，且以图像为依据。

成图像描述 $d$，描述 $I$ 通过一条指令 $p$（例如，尽可能详细地描述这张图像）由一个指令跟随型大型视觉语言模型（LVLM）生成，该模型旨在根据输入图像生成文本（InstructBLIP [9]）。描述涵盖物体及其属性，以及图像上下文（例如，图像描绘了一对男人和女性之间的浪漫时刻）。随后，DeX 通过大语言模型（Vicuna [8]）生成 $d$的 $\mathcal{T}_I = \{t_1, \cdots, t_n\}$基于标签的摘要（例如，男人、女性、浪漫时刻）。

我们通过生成 $\mathcal{T}_I$ 中所有元素长度不超过指定长度 $s$ 的唯一组合，从 $\mathcal{T}_I$ 推导出 $\mathcal{C}_I$：

$$\mathcal{C}_I = \bigcup_{i=1}^{s} \binom{\mathcal{T}_I}{i} = \bigcup_{i=1}^{s} \{S \subseteq \mathcal{T}_I, \text{ s.t. } |S| = i\}, \quad (9)$$

其中 $i \in \mathbb{N}$（例如，长度为 2 的场景是女性，浪漫时刻）。为确保稀疏性，对于 DeX，我们将 $s = 3$设为此值，因为先前的研究 [37]表明，人们更喜欢简洁的解释。

给定一对分类正确的样本（$I$，$pr$），我们通过从其表示中移除 $c_j \in \mathcal{C}_I$，为 $I$ 计算一个候选反事实解释 $\hat{}x_j$。具体而言，遵循 [29],，对于每个概念$c_j$，我们将其在潜在空间中的方向$e_{c_j}$ 定义为特定概念文本提示（$t_{\text{trg}}$）与概念中性锚点提示（$t_{\text{src}}$，例如一张物体的照片）之差：$e_{c_j} = E_T(t_{\text{trg}}) - E_T(t_{\text{src}})$。随后，将所得向量归一化为单位长度。候选反事实解释 $\hat{}x_j$ 定义为 $\hat{}x_j = x - e_{c_j}$，而 $I$ 的所有候选反事实解释构成的集合为$\mathcal{E}_I = \{\hat{}x_j | j = 1, \ldots\ldots, k\}$。

接下来，我们评估每个 $\hat{}x_j$的分类器预测。我们将 $I$ 的所有有效基于图像的反事实解释定义为集合$\mathcal{E}_I^v \subseteq \mathcal{E}_I$，满足 $\hat{}x_j \in \mathcal{E}_I^v \iff f(\hat{}x_j) \neq f(x)$；相应的 $c_j$ 概念集合记为 $\mathcal{C}_I^v$。由于 $c_j$ 用于生成 $\hat{}x_j$，我们将 $\hat{}x_j$（基于图像）和 $c_j$（文本）统称为反事实解释。图3展示了置信度与邻近性之间的权衡——有效反事实样本的权衡关系，并突出显示了具有帕累托最优解释的样本。

DeX 生成一组图像反事实 $\mathcal{E}_I^v$。选择最优的反事实解释集时，需要考虑多个相互冲突的标准，例如预测置信度和邻近性。我们采用多目标优化方法来确定帕累托前沿，即在所有标准上均不受支配的一组解；换言之，不存在其他解能同时在所有标准上表现得更好。

令每个 $c \in \mathcal{C}_I^v$ 都关联一个向量值目标函数$\mathbf{o}(c) = \big(o_1(c)\ o_2(c) \ldots,\ o_m(c)\big)$，其中 $o_i$表示一个目标函数（即预测置信度、邻近性），$i = 1, \ldots m$，且 $m \geq 2$。我们从最大化的角度定义支配关系：若反事实解 $c_j$ 支配解 $c_z$（记作 $c_j \prec c_z$），则 $o_i(c_j) \geq o_i(c_z)\ \forall i = 1, \ldots\ldots, m$ 且$\exists i = 1, \ldots\ldots, m$，使得 $o_i(c_j) > o_i(c_z)$。由此得到帕累托前沿：

$$\mathcal{P} = \{c \in \mathcal{C}_I^v | \nexists c' \in \mathcal{C}_I^v \text{ s.t. } c' \prec c\}, \quad (10)$$

其中可能包含许多最优解。然后，我们从 $\mathcal{P}$ 中选择 $\mathcal{C}_I^b$ 个 $q$ 解的子集，使解之间的相似性最小：

$$\mathcal{C}_I^b = \underset{|S|=q, S \subseteq \mathcal{P}}{\arg\min} \sum_{c_i, c_j \in S, i \neq j} \cos(E_T(c_i), E_T(c_j)), \quad (11)$$

这一选择过程确保 $\mathcal{C}_I^b$ 不仅满足所有非支配准则，还能最大限度地提高所选反事实解释之间的多样性。$\mathcal{E}_I^b$ 是与 $\mathcal{C}_I^b$对应的基于图像的反事实解释集合。我们设定 $q = 3$。

通过采用这种多目标优化策略，DeX 能够生成高质量的反事实解释，并在各项期望属性之间实现不同的权衡（见图4）。例如，可以通过以下概念解释描绘驾驶执照的图像所涉及的隐私判定：执照、女性、签发机构或驾驶员，

表 1. 使用 OmniBind 嵌入训练的线性分类器的分类性能 [50]。结果基于 Pri- vacyAlert [54] 和 VISPR [39] 测试集报告。说明： $ACC$ (%)：准确率， $F1$-m (%)：F1 宏平均值。

| 数据集 | $ACC$ | $F1$-m |
|---|---|---|
| PrivacyAlert | 87.17 | 83.34 |
| VISPR | 91.19 | 90.92 |

照片、签发机构。第一个反事实解释可对应于 VISPR [39]真实标注中的驾驶执照、性别和住址部分信息。第二个反事实解释可与真实标注中的完整面部 相关联，因为它涉及驾驶员的照片。对于展示一群医疗专业人员和一名患者身处病房的图像，其判定可以通过人们、协作、福祉、专业人员、患者、房间或护士、患者、人们来解释；这些内容可对应于真实标注中的专业关系 和职业。

请注意，可行性在设计上得到了增强，因为解释是基于一组图像衍生概念构建的，因此解释内容有图像内容作为依据。

## 5. 结果

实验设置。我们将 DeX 与 Coun- TEX [29]，进行比较，后者通过文本驱动概念操纵图像嵌入来生成文本反事实解释，是与我们工作直接相关的参照方法。此外，我们通过主题建模分析解释，并讨论驱动隐私决策的主要因素及潜在偏差。为进行公平比较，DeX 和 CounTEX [29] 均采用基于 OmniBind 嵌入 [50],OB+ 训练的线性分类器。表 1 展示了在两个隐私数据集上的分类准确率和宏平均 F1 分数。我们使用 Adam优化器训练 OB+线性分类器 100 个周期（学习率为 $10^{-3}$，批量大小为 64），并采用标准交叉熵损失。由于 CounTEX 需要预先定义概念列表，我们采用 [54]的隐私分类体系，并通过添加描述公众图像簇的概念，手动扩展其中的公众概念 [5]。利用这一预定义概念库及其 $L$概念关键词 $\mathcal{C} = \{c_1, \ldots, c_L\}$，CounTEX 在生成反事实 $\hat{x}$时，会为每个概念 $c$ 分配权重 $w_c(\hat{x})$。我们选取负值 $w_c$ 最高的排名前3的概念（即解释私密类别的概念），以展示解释示例，并评估其可行性和解释坍塌，因为该方法通常具有较高的稀疏性（即会使用许多概念来修改图像）。计算稀疏性时，我们将权重大于 0.1 的概念纳入考量，即$g(\hat{x}) = |\{c \in \mathcal{C} | w_c(\hat{x}) > 0.1\}|$。

由于代码未公开，我们根据

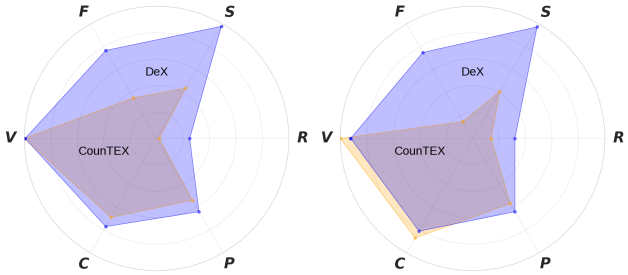


图5. DeX 与 CounTEX 在 PrivacyAlert [54]（左侧）和 VISPR [39]（右侧）上的比较，评估指标包括有效性、 $V$、可行性、 $F$、稀疏性、 $S$、解释坍塌、 $R$、通过原始图像嵌入与其反事实之间的余弦相似度衡量的邻近性、 $P$以及置信度、 $C$。为便于可视化， $S$ 缩放至 [0,1] 并取反（即数值越高越好）。缩放时，将 $S$ 的最大值设为100。对于 CounTEX， $F$ 和 $R$ 是针对排名前3的概念报告的。

论文中的说明和设置实现该方法。我们使用 SGD优化器优化概念分数 $w$ （学习率 $10^{-2}$，最多迭代 100 次），并在预测变化时触发提前停止。我们使用Xavier初始化初始化 $w$，并使用交叉熵损失；使用带有 0.1 正则化参数的识别损失，以及正则化参数为 0.1 的 $L_1$ 和 $L_2$ ，来优化$w$。

为评估解释坍塌，我们使用 Recognize Anything Model (RAM) [18], 作为图像标注器。该模型能够为图像分配多个语义标签，并泛化到未见类别。为获得帕累托前沿，我们使用 Pymoo [7]Python库中的实现。我们使用 SBERT [41] 对解释进行编码并计算相似性。

我们使用两个公开可用的数据集：Priva- cyAlert [54]和 VISPR [39]。PrivacyAlert 包含更多显式内容，而 VISPR 则包含文档类内容（护照、身份证件、电子邮件）。这两个数据集互为补充，对相似内容（例如包含汽车或儿童的图像）有不同的标注。

**讨论**。图5 展示了定量结果。对于我们的方法，我们报告了关于以下集合的结果： $\mathcal{E}_I^b$通过两个标准选出的最佳反事实样本：预测置信度（$C$）和邻近性（$P$），后者通过余弦相似度计算。此外，我们还报告了使用随机扰动改变图像的结果，记为 $rand$ （图7）。我们的方法在PrivacyAlert数据集上的有效性（$V$）达到99.43%，在VISPR数据集上达到92%，同时还具有较高的稀疏性。相比之下，CounTEX在两个数据集上都能达到100%$V$，但代价是稀疏性较低，导致图像编辑会影响许多概念，通常超过50个。此外，当要求预测具有高置信度时，尽管使用了大量6的概念，CounTEX在PrivacyAlert数据集上的有效性仍显著下降（图

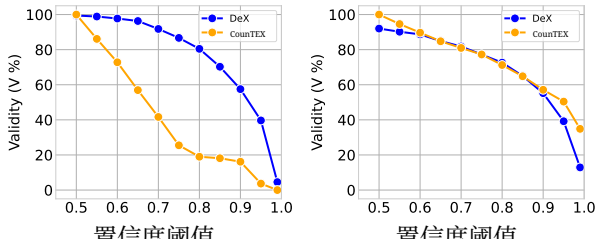


图 6. 不同置信度阈值下的有效性，Priva- cyAlert [54]（左侧）和 VISPR [39]（右侧）。DeX 生成稀疏性较高的高置信度解释。CounTEX [29] 为 VISPR 生成了同样具有高置信度但稀疏性较低的解释，而未能为 PrivacyAlert 生成高置信度解释。

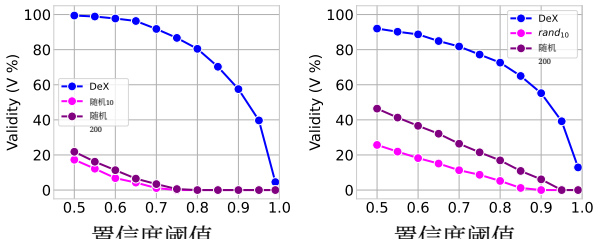


图 7. 对扰动的鲁棒性，PrivacyAlert [54]（左）和 VISPR [39]（右）。在选择最佳解释时，DeX 会过滤掉由扰动引起的预测翻转。

用于编辑图像。DeX能够以高稀疏性实现高置信度的预测变化。在置信度（$C$）方面，DeX在PrivacyAlert上优于CounTEX，在VISPR上取得了可比的结果。图像扰动后更加可靠的公众预测表明，DeX具备更强的移除隐私相关概念的能力。DeX的邻近性（$P$）更高，约为0.64，而CounTEX的得分介于0.54至0.57之间。更高的 $P$ 得分可归因于概念编辑过程的根本差异。我们的方法在设计上强制保持稀疏性，并限制用于修改图像的概念数量。这不同于CounTEX，后者会针对许多概念操纵图像，从而导致邻近性降低。

由于 CounTEX 对每张图像仅生成一个解释，因此我们比较的是各方法在整个数据集中的变化情况（$R$），而非关注图像内的多样性。结果表明，DeX 在不同样本间生成的解释具有更高的多样性，而 CounTEX 则出现了解释坍塌：对大量图像生成相同或高度相似的解释。这可能是因为它依赖于固定的预定义概念集，而这些概念可能较为通用，或与数据集内容不够契合。此外，DeX 在 PrivacyAlert 和 VISPR 数据集上的多样性（$D$）得分分别为 0.29 和 0.22。这些数值表明图像内的解释具有多样性，体现了 DeX 捕捉不同相关概念的能力。

DeX 在 $F$ 得分方面也表现出色，在两个数据集上的得分稳定在 75% 至 77% 之间，表明这些解释有充分依据，且与输入图像相关。相比之下，CounTEX 的 $F$ 得分较低，在概念数量不同的各种配置下为 12% 至 35%。具体而言，仅使用排名前 3 的概念时，$F$ 比使用排名前 10 的概念时更高。对于 PrivacyAlert，使用排名前 3 的概念时，$F$ 为 35.41%；而使用排名前 10 的概念时，该值为 18.16%。A

VISPR 也观察到类似趋势，排名前三的概念达到 $F$ 14.76%，优于排名前十的概念集合，后者达到 $F$ 12.06%。这表明，只有一小部分概念有依据，导致 Coun-TEX 的解释缺乏上下文相关性。

**鲁棒性**。我们通过向每个图像嵌入添加均值为零的高斯噪声，评估模型对非语义扰动的预测可靠性。我们考虑两种设置：10 个和 200 个随机向量，分别记作 $rand_{10}$ 和$rand_{200}$。图7 显示，随机扰动通常会导致预测变化，但大多数翻转的置信度较低。在置信度阈值为 0.5 时，PrivacyAlert 和 VISPR 在 $rand_{200}$下的 $V$ 分别为 22% 和 46%；当置信度阈值为 0.7 时，则分别降至 3% 和 26%。

相比之下，基于概念的操纵会导致置信度更高的预测变化，即使将置信度阈值设为 0.7，$V$ 仍然很高。我们通过过滤掉置信度较低的预测，确保解释具有意义。

**失败案例**。DeX 受标注模型局限性的制约，包括视觉语言模型的幻觉。若描述过于笼统或不准确，或者大语言模型未能提取相关关键词，DeX 就可能出现失败案例（图 8）。例如，一些描绘人们场景的图像，其描述准确无误，但大语言模型未能提取“人们”或“人”这两个词。



**描述**：图像描绘了六个人围坐在餐桌旁，场景像是在教室或办公室。他们正在交谈，可能是在讨论与工作或学习相关的事情.....

**提取的标签**：餐桌、椅子、杯子、瓶子、背包、对话、研究。



**描述**：图像中有三位女性正摆在一起合影。其中两位女性穿着婚纱，第三位女性则穿着绿色连衣裙。她们都面带笑容，看起来.....似乎正享受彼此的陪伴.....

**提取的标签**：婚纱、绿色连衣裙。

图 8. 当大语言模型未能识别相关关键词（例如，人们）时，DeX 可能会出现失败案例。这类遗漏凸显了数据集中的一种潜在偏差：仅凭图像中存在个人，就足以将其判定为私密图像。




图 9. 从 PrivacyAlert 上生成的解释中发现的主题列表 [54]（左）和 VISPR [39]（右）。在 PrivacyAlert 中，我们观察到主要主题是裸露/色情内容。在 VISPR 中，多种不同类型的内容被视为隐私内容，例如个人物品（如智能手机）、汽车和文档。请注意，主题名称由主题建模算法自动生成 [15]。

**识别数据集偏差**。DeX 不仅适用于可解释性场景之外，也可用于分析数据集及其偏差。基于 DeX 生成的解释，我们进行主题建模，以识别影响模型预测的主要因素，以及数据集中的潜在偏差。我们使用 BERTopic [15]。BERTopic 使用 HDBSCAN [35]，根据数据的嵌入表示对其进行聚类。在聚类之前，我们使用 UMAP [44]，将数据嵌入的维度降至 5，因为 HDBSCAN 在低维数据上的表现更好。为了进行可视化，我们使用 UMAP 将嵌入降至 2 维。寻找最佳主题建模配置，以及处理主题名称中的同义词和单复数形式的超参数搜索，均不在本研究的范围之内。请注意，不同的设置可能会产生不同的结果。

我们发现 VISPR 与 PrivacyAlert 之间存在显著差异：前者数据集中的隐私内容和场景具有较高的多样性（即人们出现在不同活动中、可见车牌的汽车、文档，以及智能手机等个人物品），而后者数据集中的隐私内容则主要与裸露/色情内容和亲密时刻有关。主题列表见图9。

## 6. 结论

我们提出了 DeX，这是一种基于概念的反事实跨模态分解方法，可生成基于文本的解释。DeX 利用图像特定概念修改图像嵌入，并识别影响决策的最少概念。通过结合关键图像元素并评估多个反事实场景，DeX 能够为每张图像生成多种解释，从而更全面地理解模型的决策。DeX 通过设计实现稀疏性，使该方法能够使用固定数量的概念生成解释，以便于理解。我们将 DeX 应用于图像隐私这一具有挑战性的案例，并展示了它能够识别决定分类结果的概念。此外，DeX 还揭示了数据集主题偏差，有助于提升样本多样性。未来的工作将聚焦于在不同隐私场景中应用 DeX，例如通过反事实概念操控来预测隐私评分并解释评分敏感性，以及将其用于公平性分析等其他任务。

**致谢**。本文所述的大部分研究是在第一作者任职于 Idiap 研究所期间开展的。

## 参考文献

[1] Abubakar Abid、Mert Yuksekgonul 和 James Zou。利用基于概念的反事实解释，有效调试模型错误。收录于 机器学习国际会议，页码 66 88 2022 2 – 。P [2] MLR，。Arjun Akula、Shuai Wang 和 Song-Chun Zhu。CoCoX：通过故障线生成基于概念的解释和反事实解释。*AAAI*人工智能会议论文集，34(03):2594–2601，2020年。2 [3] Maximilian Augustin、Valentyn Boreiko、Francesco Croce 和 Matthias Hein。扩散视觉反事实解释。神经信息处理系统进展，35:364–377，2022年。2.[4] Gonul Ayci、Arzucan Ozg¨ur、Murat Sensoy 和 Pinar Yolum。向我解释：探索隐私决策的理解。arXiv:2301.02079 [cs.AI], 2023。2 [5] Alina Elena Baia 和 Andrea Cavallaro。用于可解释隐私分类的图像引导主题建模。收录于欧洲计算机视觉会议研讨会，页码 200–217，2024年。2，6[6] Usha Bhalla、Alex Oesterling、Suraj Srinivas、Flavio P. Calmon 和 Himabindu Lakkaraju。利用

稀疏线性概念嵌入（SpLiCE）。收录于神经信息处理系统进展，页码 84298–84328, 2024. 3。

[7] Julian Blank 和 Kalyanmoy Deb。Pymoo：Python 中的多目标优化。IEEE Access，8:89497–89509，2020年。6[8] Wei-Lin Chiang、Zhuohan Li、Zi Lin、Ying Sheng、Zhang-hao Wu、Hao Zhang、Lianmin Zheng、Siyuan Zhuang、Yong-hao Zhuang、Joseph E. Gonzalez、Ion Stoica 和 Eric P. Xing。Vicuna：一款开源聊天机器人，以 90%* 的 ChatGPT 水平展现出色表现，2023年。5[9] Wenliang Dai、Junnan Li、Dongxu Li、Anthony Meng Huat Tiong、Junqi Zhao、Weisheng Wang、Boyang Li、Pascale Fung 和 Steven Hoi。InstructBLIP：面向通用视觉语言模型的指令微调。收录于神经信息处理系统进展，2023年。5[10] Laura Ferrarello、Andrea Cavallaro、Rute Fiadero 和 Riccardo Mazzon。通过系统思维设计重塑隐私叙事。收录于 DRS Conference 论文集，2022年。1[11]Louis Fournier、Emmanuel Dupoux 和 Ewan Dunbar。类比减去类比测试：衡量词嵌入中的规律性。收录于第24届计算自然语言学习会议论文集，页码 365–375。计算语言学协会，2020年。1[12] Rinon Gal、Or Patashnik、Haggai Maron、Amit H Bermano、Gal Chechik 和 Daniel Cohen-Or。StyleGAN-NADA：由 CLIP 引导的图像生成器领域自适应。ACM 图形学汇刊（TOG），41(4):1–13，2022年。2[13] Amirata Ghorbani、James Wexler、James Y Zou 和 Been Kim。迈向自动化的基于概念的解释。神经信息处理系统进展，32，2019年。2[14] Rohit Girdhar、Alaaeldin El-Nouby、Zhuang Liu、Mannat Singh、Kalyan Vasudev Alwala、Armand Joulin 和 Ishan Misra。ImageBind：统一万物的嵌入空间。收录于 IEEE 计算机视觉与模式识别会议，2023年。1[15] Maarten Grootendorst。BERTopic：采用基于类别的 TF-IDF 程序进行神经主题建模。arXiv: 2203 05794 。cs.CL。[ ], 2022 8

[16] Riccardo Guidotti。反事实解释及其寻找方法：文献综述与基准测试。《数据挖掘与知识发现》，38(5):2770–2824，2024年。2[17] Alec Helbling、Shruti Palaskar、Kundan Krishna、Polo Chau、Leon Gatys 和 Joseph Yitan Cheng。SafetyPairs：利用反事实图像生成分离安全关键图像特征。arXiv 预印本 arXiv:2510.21120，2025年。2[18] Xinyu Huang、Yi-Jie Huang、Youcai Zhang、Weiwei Tian、Rui Feng、Yuejie Zhang、Yanchun Xie、Yaqian Li 和 Lei Zhang。通过多粒度文本监督进行开放集图像标注。收录于第33届 ACM 多媒体国际会议论文集，页码 4117–4126，2025年。6[19] Paul Jacob、Eloi Zablocki、Hedi Ben-Younes、Micka¨el Chen、Patrick P´erez 和 Matthieu Cord。STEEX：利用语义引导反事实解释。收录于欧洲计算机视觉会议，页码 387–403, 2022. 1, 2。

[20] Guillaume Jeanneret、Loïc Simon 和 Fr´ed´eric Jurie。用于反事实解释的扩散模型。收录于亚洲计算机视觉会议论文集，页码 858–876，2022年。1、2、3

[21] Guillaume Jeanneret、Lo¨ıc Simon 和 Fr´ed´eric Jurie。对抗性反事实视觉解释。载于 IEEE 计算机视觉与模式识别会议，第 16425–16435 页，2023年。2, 3[22] Guillaume Jeanneret、Lo¨ıc Simon 和 Fr´ed´eric Jurie。用于反事实解释的文本到图像模型：一种黑箱方法。载于 IEEE/CVF 冬季计算机视觉应用会议论文集，第 4757–4767 页，2024年。1, 2[23] Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc Le、Yun-Hsuan Sung、Zhen Li 和 To m Duerig。通过噪声文本监督扩展视觉与视觉语言表征学习。载于国际机器学习会议，第 4904–4916 页。PMLR，2021年。1[24] Tero Karras、Samuli Laine、Miika Aittala、Janne Hellsten、Jaakko Lehtinen 和 Timo Aila。分析并提升 StyleGAN 的图像质量。载于 IEEE 计算机视觉与模式识别会议，第 8110–8119, 2020. 2 – 页。

[25] Tahira Kazimi、Ritika Allada 和 Pinar Yanardag。扩散中的解释：以扩散语义解释分类器。载于 IEEE 计算机视觉与模式识别会议，第 14799–14809 页，2025年。2[26] Saeed Khorram 和 Li Fuxin。通过潜在变换实现循环一致的反事实样本。载于 IEEE 计算机视觉与模式识别会议，第 10203–10212 页，2022年。2[27] Been Kim、Martin Wattenberg、Justin Gilmer、Carrie Cai、James Wexler、Fernanda Viegas 等。超越特征归因的可解释性：使用概念激活向量（TCAV）进行定量测试。载于国际机器学习会议，第 2668–2677 页。PMLR，2018年。2[28] Gwanghyun Kim、Taesung Kwon 和 Jong Chul Ye。Diffusionclip：用于稳健图像编辑的文本引导扩散模型。载于 IEEE 计算机视觉与模式识别会议，第 2426–2435 页，2022年。2[29] Siwon Kim、Jinoh Oh、Sungjin Lee、Seunghak Yu、Jaeyoung Do 和 Tara Taghavi。将图像分类器的反事实解释锚定到文本概念空间。载于 IEEE 计算机视觉与模式识别会议，第 10942–10950 页，2023年。2, 5, 6, 7[30] Sunnie S. Y. Kim、Elizabeth Anne Watkins、Olga Russakovsky、Ruth Fong 和 Andr´es Monroy-Hern´andez。“ 帮我帮助 AI”：理解可解释性如何支持人机交互。载于 2023年 CHI 计算系统中人为因素会议论文集，美国纽约州纽约市，2023年。美国计算机协会。1[31] Oran Lang、Yossi Gandelsman、Michal Yarom、Yoav Wald、Gal Elidan、Avinatan Hassidim、William T Freeman、Phillip Isola、Amir Globerson、Michal Irani 等。以风格解释：训练生成对抗网络在 StyleSpace 中解释分类器。载于国际计算机视觉会议，第 693–702, 2021. 2 – 页。

[32] Tiep Le、Vasudev Lal 和 Phillip Howard。Coco-counterfactuals：为图文对自动构建反事实示例。《神经信息处理系统进展》，36：71195–71221，2023年。1, 2[33] Lajanugen Logeswaran 和 Honglak Lee。一种高效的句子表征学习框架。载于国际学习表征会议，2018年。1

[34] Jinqi Luo、Zhaoning Wang、Chen Henry Wu、Dong Huang 和 Fernando De la Torre。零样本模型诊断。载于 IEEE 计算机视觉与模式识别会议，第 11631–11640 页，2023年。2[35] Leland McInnes、John Healy 和 Steve Astels。hdbscan：基于层次密度的聚类。《开放源代码软件期刊》，2017 8。Tomas [36] Mikolov、Kai Chen、Greg Corrado 和 Jeffrey Dean。在向量空间中高效估计词表示。arXiv 预印本 arXiv:1301.3781，2013年。1[37] Tim Miller。人工智能中的解释：来自社会科学的启示。《人工智能》，267 38，2019 1 5：–。[38] Ramaravind K. Mothilal、Amit Sharma 和 Chenhao Tan。通过多样化反事实解释来解释机器学习分类器。载于 2020年公平性、问责性与透明度会议论文集，第 607–617 页。美国计算机协会，2020年。2, 3[39] Tribhuvanesh Orekondy、Bernt Schiele 和 Mario Fritz。迈向视觉隐私顾问：理解并预测图像中的隐私风险。载于国际计算机视觉会议，2017年。2, 6, 7, 8[40] Viraj Prabhu、Sriram Yenamandra、Prithvijit Chattopadhyay 和 Judy Hoffman。LANCE：通过生成语言引导的反事实图像对视觉模型进行压力测试。载于《神经信息处理系统进展》，2023 2。Nils Reimers 和 [41] Iryna Gurevych。Sentence-BERT：使用孪生 BERT 网络的句子嵌入。载于自然语言处理实证方法会议，2019年。6
[42] Pau Rodriguez、Massimo Caccia、Alexandre Lacoste、Lee Zamparo、Issam Laradji、Laurent Charlin 和 David Vazquez。超越平凡反事实解释：多样且有价值的解释。载于国际计算机视觉会议，第 1056–1065 页，2021年。1, 2, 3[43] Chris Russell。高效搜索多样且连贯的解释。载于公平性、问责性与透明度会议论文集，第 20–28 页，2019年。2, 3[44] Tim Sainburg、Leland McInnes 和 Timothy Q Gentner。用于表征学习与半监督学习的参数化 UMAP 嵌入。《神经计算》，2021年。8[45] Matthew Trager、Pramuditha Perera、Luca Zancato、Alessan-dro Achille、Parminder Bhatia 和 Stefano Soatto。意义的线性空间：视觉语言模型中的组合结构。载于国际计算机视觉会议，第 15395–15404 页，2023年。1
[46] Sahil Verma、Varich Boonsanong、Minh Hoang、Keegan Hines、John Dickerson 和 Chirag Shah。机器学习中的反事实解释与算法补救：综述。《ACM计算综述》，56(12)：1–42，2024年。2[47] Sandra Wachter、Brent Mittelstadt 和 Chris Russell。无需打开黑箱的反事实解释：自动化决策与《通用数据保护条例》。《哈佛法律与技术期刊》，31：841，2017年。1, 2[48] Catherine Wah、Steve Branson、Peter Welinder、Pietro Perona 和 Serge Belongie。Caltech-UCSD Birds-200-2011 数据集。2011年。2
[49] Yinong Oliver Wang、Eileen Li、Jinqi Luo、Zhaoning Wang 和 Fernando De la Torre。无监督模型诊断。arXiv 预印本 arXiv:2410.06243，2024 年。2[50] Zehan Wang 、 Ziang Zhang、Minjie Hong、Hang Zhang、Lup- ing Liu、Rongjie Huang、Xize Cheng、Shengpeng Ji、Tao Jin、Hengshuang Zhao 和 Zhou Zhao。OmniBind：通过绑定空间实现大规模全方位多模态表征。载于《国际学习表征会议》，2025年。1、3、6[51] Zongze Wu、Dani Lischinski 和 Eli Shechtman。StyleSpace分析：用于 Style- GAN 图像生成的解耦控制。载于《IEEE计算机视觉与模式识别会议》，第12858–12867页，2021年。2[52] Yongqin Xian、Christoph H. Lampert、Bernt Schiele 和 Zeynep Akata。零样本学习——对优劣各面的全面评估。《IEEE模式分析与机器智能汇刊》，41(9):2251–2265，2019年。2[53] Mehdi Zemni、Micka¨el Chen、Eloi Zablocki、H´edi Ben-Younes、Patrick P´erez 和 Matthieu Cord。OCTET：面向对象的反事实解释。载于《IEEE计算机视觉与模式识别会议》，第15062–15071页，2023年。2[54] Chenye Zhao、Jasmine Mangat、Sujay Koujalgi、Anna Squic- ciarini 和 Cornelia Caragea。PrivacyAlert：用于图像隐私预测的数据集。载于《国际AAAI网络与社交媒体会议论文集》，2022年。2、6、7、8