

基于扩散的迭代反事实解释方法在胎儿超声图像质量评估中的应用

Paraskevas Pegios^{1,4}✉, Manxi Lin¹, Nina Weng¹, Morten Bo Søndergaard Svendsen², Zahra Bashir³, Siavash Bigdeli¹, Anders Nymark Christensen¹, Martin Tolsgaard², and Aasa Feragen^{1,4}

¹ Technical University of Denmark, Kongens Lyngby, Denmark
{ppar,afhar}@dtu.dk

² Region Hovedstaden Hospital, Copenhagen, Denmark

³ Slagelse Hospital, Copenhagen, Denmark

⁴ Pioneer Centre for AI, Copenhagen, Denmark

摘要 产科超声图像质量对于准确诊断和监测胎儿健康至关重要。然而，获取高质量的标准平面困难重重，这受到操作员的专业水平以及母体 BMI 或胎儿动态等因素的影响。在这项工作中，我们探索了基于扩散的反事实可解释 AI 技术，以从低质量非标准图像生成逼真的、高质量的标准平面。通过定量和定性评估，我们展示了我们的方法在生成更高质量反事实结果方面的有效性。这表明未来有可能通过提供视觉反馈来提高临床医生的培训水平，并可能改善下游诊断和监测所需的标准平面的质量和获取。

Keywords: 可解释的人工智能 · 扩散模型 · 胎儿超声图像

1 介绍

产科超声筛查图像的质量对于临床下游任务至关重要，包括胎儿生长估计 [31]、早产预测 [23] 以及异常检测 [21]。如果捕获的解剖平面不够精确或待测量的解剖区域显示不佳，则测量结果和估算结果将会不准确（图 1）。为了标准化图像质量，国际超声学会制定了定义胎儿超声质量的精确标准 [24]。然而，获得高质量图像受到了临床医生培训水平以及孕妇 BMI 或胎儿位置等物理特征的影响。因此，获取 标准平面 足够高质量的图像可能非常具有挑战性。

作为解决这个问题的一步，我们研究了反事实视觉解释用于胎儿超声图像质量评估。给定一个低质量的非标准平面（NSP）作为输入，我们估

算同一解剖结构的高质量标准平面（SP）。尽管我们的方法植根于可解释的人工智能，但它在模型可解释性之外还有潜在的应用，通过启用具有影响力的生成式人工智能应用如合成数据生成 [29]、个性化临床培训 [17] 以及计算机辅助平面获取 [18]，最终目标是支持在难以获得标准平面（SP）的情况下进行可靠的诊断。

我们的方法基于最先进的扩散式反事实解释方法 [9, 10, 27, 30]。然而，这些方法通常设计为进行最小的、高度局部化的更改。这种假设与胎儿超声质量评估不太相符，在后者中，提高质量往往需要更广泛的修改，因为纠正一个 NSP 可能需要更多的全局调整 [24]。为此，我们贡献了 1) 一种迭代扩散式的反事实生成方法，该方法引导产生更高的置信度的反事实结果，实现对输入的大范围更改；以及 2) 一项广泛的评估，证明我们可以为胎儿超声生成可信且高质量的反事实路径。

2 相关工作

深度学习方法已被广泛用于超声图像质量评估 [12, 31, 33]。生成模型通过改进细粒度分类 [15] 和为超声技师创建培训材料进一步支持了这一任务 [11, 17]。非常近期，提出了 ScanAhead [18]，这是一种具有领域适应的基于 GAN 的方法，能够从扫描视频中预测高质量的标准平面，并结合了三维探头运动信息。去噪扩散概率模型（DDPMs）的成功 [4, 7] 使创建高度逼真的胎儿超声图像 [8, 29] 和增强分布外检测 [20, 21] 成为可能。在本文中，我们仅依赖于图像数据和现有的超声数据库来探索一种引导扩散 [4] 反事实迭代框架，该框架旨在预测从低质量到高质量图像的路径。

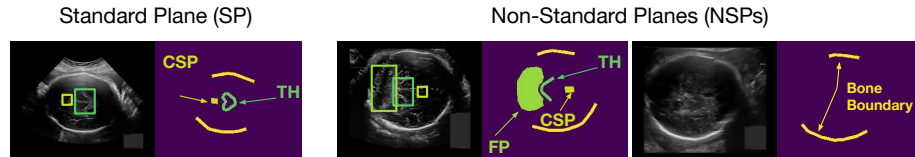


图 1: 标准平面（SP）被定义为穿过身体的特定解剖平面；这里我们展示了高质量的 SP（左）和低质量的非标准平面（NSP）（右）的例子，用于胎儿头部。一个胎儿头部的 SP 应该显示丘脑（TH）、透明隔腔（CSP），但不包括后窝（FP）。骨边界应支持卡尺的正确放置。

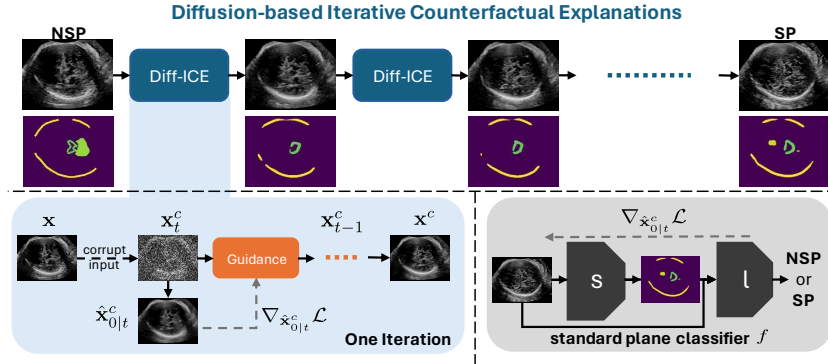


图 2: **顶部**: 每次迭代使用之前的输出来增强反事实置信度。**左下角**: 在每个时间步骤 t 对一次迭代进行高效的梯度估计。**右下角**: 标准平面分类器和引导梯度流。

反事实解释试图回答这样一个问题: 如果一个给定的分类器将图像分类为不同的类别, 那么这个图像会是什么样子? 与对抗样本不同, 反事实应该具有现实性, 即接近数据流形, 这通常用生成模型 [22] 来近似。在计算机视觉领域, 已经提出了各种方法, 包括基于 GAN 的方法 [26] 和基于扩散的方法 [9, 10, 25, 27, 30]。然而, 这些方法应用于变化高度局部化的任务中。在我们的工作中, 我们利用了扩散引导的最新进展 [6, 30] 来应用一种计算可行的迭代方法, 而不是稀疏性 [9]、修复术 [10, 30] 或区域约束 [27] 来细化基于扩散的反事实。这使我们能够实现更全局性的反事实变化。

3 方法

3.1 扩散模型的预备知识

DDPMs [7] 是由两个过程定义的: 前向和反向。在前者中, 输入图像 x_0 通过在每个时间步 t 添加高斯噪声而逐渐被破坏, 而在后者中则逐渐去除这些噪声以生成清洁图像。形式上, 前向过程通过从原始干净图像 x_0 迭代生成时间 t 的图像 $x_t \sim \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t I)$ 来定义, 其中 $\{\beta_t\}_{t=1:T}$ 控制每一步添加噪声的方差。时间 t 的图像 x_t 可以直接使用 x_0 和重新参数化技巧进行采样。

$$x_t = \sqrt{\alpha_t}x_0 + \epsilon\sqrt{1-\alpha_t}, \epsilon \sim \mathcal{N}(0, 1), \quad (1)$$

其中 $\alpha_t = 1 - \beta_t$ 和 $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ 。逆过程也包含由神经网络预测其均值和协方差的高斯转换: $\mathbf{x}_{t-1} \sim \mathcal{N}(\mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t))$, 其中 $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 。在实践中, 去噪器 $\epsilon_\theta(\mathbf{x}_t, t)$ 预测来自方程 (1) 的噪声而不是直接预测均值 $\mu_\theta(\mathbf{x}_t, t)$, 给出 $\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{1-\beta_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right)$ 。

3.2 Diff-ICE: 基于扩散的迭代反事实解释

我们使用分类器 f 来量化图像质量, 该分类器被训练以预测胎儿超声图像是否为标准或非标准 (SP 或 NSP) 平面。由于超声图像质量 [13] 和结果预测 [23] 都受益于将图像与分割相结合, 分类器 f 由一个分割模型 s 和一个按顺序训练的预测器 l 组成。分类器将图像和分割预测 $f(\mathbf{x}) = l(s(\mathbf{x}), \mathbf{x})$ 作为输入。这为分类器增加了可解释性, 因为可以将分割可视化为部分解释。

在 [9, 30] 之后, 我们对输入 $\mathbf{x}$ 进行扰动, 直至达到有限的噪声水平 τ , 使用 $1 \leq \tau \leq T$, 并利用公式 (1) 初始化输入的噪声版本, 并使用引导扩散 [4] 将 $\mathbf{x}_\tau^c$ 引导至期望的反事实类别 y 。为此, 我们最小化损失函数 $\mathcal{L}$ 并通过其梯度 g 移动平均样本,

$$\mathbf{x}_{t-1}^c \sim \mathcal{N}(\mu_\theta(\mathbf{x}_t^c, t) - \Sigma_\theta(\mathbf{x}_t^c, t) g, \Sigma_\theta(\mathbf{x}_t^c, t)). \quad (2)$$

由于分类器 f 在干净图像上进行训练, 我们使用学习到的去噪器获得一步去噪预测, 并通过求解方程 (1) 得到 $\mathbf{x}_0$, $\hat{\mathbf{x}}_{0|t}^c = \frac{\mathbf{x}_t^c - \sqrt{1-\alpha_t} \epsilon_\theta(\mathbf{x}_t^c, t)}{\sqrt{\alpha_t}}$. 将其传递给 f 。关于 $\mathbf{x}_t^c$ 的梯度计算 (如 [1, 32] 所示), 即 $g = \nabla_{\mathbf{x}_t^c} \mathcal{L}$, 需要反向传播通过去噪器。由于计算成本高昂, 我们采用了一种高效的梯度估计方法 [6, 30], 该方法避免了过多的反向传播并加快了采样速度。因此, 对于每个 t , 我们根据 $\hat{\mathbf{x}}_{0|t}^c$ 计算梯度, 即在方程 (2) 中使用反事实损失计算 $g = \nabla_{\hat{\mathbf{x}}_{0|t}^c} \mathcal{L}$ 。

$$\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}_{0|t}^c, y) = \lambda_c \mathcal{L}_c(f(\hat{\mathbf{x}}_{0|t}^c), y) + \lambda_p \mathcal{L}_p(\hat{\mathbf{x}}_{0|t}^c, \mathbf{x}), \quad (3)$$

其中 $\mathcal{L}_c$ 是分类损失, 引导至期望的标签 y , $\mathcal{L}_p$ 是基于 l2 的感知损失, 从接近度方面指导过程, 而 λ_c 和 λ_p 是控制引导强度的超参数。典型应用关注局部反事实变化, 并使用基于像素的 l1-范数进行 $\mathcal{L}_p$ [10, 30] 计算, 或者将其作为额外项添加到噪声图像中 [9]。我们的损失函数优先考虑广泛的改变同时保持解剖学保真度。

然而, 实现全局变化颇具挑战。设置 τ 在有限噪声水平下可以保持一步去噪预测中的语义信息, 但在大多数细化阶段允许指导 [32]。简单增加分类器引导强度 λ_c 可能会导致无意义的生成或对抗样本 [1]。因此, 我们提出了

一种基于差异融合-的 **I** 迭代 **C** 反事实 **E** 解释方法 (Diff-ICE)，以增强信心并实现全局更改。通过 L 次反事实引导的逆过程迭代，每次将前一次的输出作为输入，我们在保持接近原始输入 $\mathbf{x}$ 保真度约束的同时优化输出，在公式 (3) 中。我们的方法总结在图 2 中。

4 实验与结果

数据和基础实现。 我们使用了从一个国家胎儿超声筛查数据库（已匿名处理）中提取的两个数据集。增长，用于训练无条件扩散模型和在引导标准平面分类器 f 中使用的分割模型，包含来自生长扫描的 4363 张（训练/测试为 2842/1521）胎儿超声图像，包括头部、腹部、股骨和宫颈图像。标题用于训练和测试完整的引导分类器 f ，包含 240 张高质量标准平面 (SP) 和 1339 张低质量非标准平面 (NSP) 的胎儿头部超声图像。作为指导标准平面分类器 f ，我们选择了一种结合了 DTU-Net [14] 分割模型 s 和 SonoNet-16 [2] 分类器 l 的鲁棒且可解释的架构，遵循 [23]。鲁棒性对于确保高质量的反事实推理很重要，而可解释性则使得不同经验水平的技术开发者 and 临床医生更容易监控这些反事实。分割模型 s 是在增长上开发的。因此，我们依次训练和评估分类器的预测器 l ，保持 g 的权重固定，并按照 121/26/93 SP 和 712/204/423 NSP 图像进行训练/验证/测试分割，这些图像是来自非重叠患者的，最终得到 78% 平衡测试准确率。一个无条件的 DDPM [7] 也在增长上使用了 1000 个扩散步骤进行训练，遵循 [9] 的架构，进行了 30 万次迭代，批量大小为 16，学习率为 10^{-4} ，权重衰减为 0.05，并且没有使用 dropout。对于所有模型，图像被调整为 224×288 大小，嵌入的文本和卡尺被移除 [19]，并且像素强度被归一化到 $[-1, 1]$ 。为了生成反事实视觉解释，我们经验性地设置了 $L = 5, \tau = 120$ 的 400 个重新分配的时间步长， $\lambda_p = 30$ 并搜索 $\lambda_c \in \{40, 60, 80\}$ 。感知损失使用在 RadImageNet [16] 上训练的 ResNet-50。

基线。 1) DiME [9] 使用每一步的昂贵嵌套循环反向引导过程 t 来获得干净的图像，并应用缩放以估计相对于噪声图像的梯度。2) 单次迭代的 Diff-ICE，表示为 Diff-ICE₁。3) 受 [1] 的启发，我们实现了 Diff-ICE_{1-x_t}，对噪声图像取梯度。为了公平比较，我们使用相同的损失函数（方程 (3)）和超参数。

评估指标。 遵循现有的评估协议 [9–11, 30]，我们使用 Fréchet Inception Distance (FID) 和 Fréchet SonoNet Distance [11] (FSD) 来评估现实主义原始

表 1: Diff-ICE 与基于基线扩散方法的比较。

	现实主义			有效性				效率		
Method	FID ↓	FSD ↓	SonoSim ↑	MQD ↑	BKL ↓	MAD ↑	FR ↑	Batch Time (s)	Total Time (h)	GPU M (GB)
DiME	41.5	0.396	0.854	0.291	0.391	0.231	0.966	3151.9 ± 730.4	37.65	9.6
Diff-ICE _{1-x_t}	39.9	0.403	0.855	0.301	0.413	0.208	0.936	231.4 ± 56.8	2.77	33.7
Diff-ICE ₁	39.0	0.355	0.856	0.253	0.387	0.234	0.982	115.6 ± 34.2	1.38	9.6
Diff-ICE	42.4	0.435	0.790	0.371	0.336	0.284	0.982	448.6 ± 22.9	5.27	9.6

表 2: 每次迭代的中间结果为 Diff-ICE 。

	现实主义			有效性			重新识别	效率 (运行时间)	
Iteration	FID ↓	FSD ↓	SonoSim ↑	MQD ↑	BKL ↓	MAD ↑	Accuracy (%) ↑	Batch (s)	Iteration (h)
1	38.96	0.355	0.856	0.253	0.387	0.234	99.76	115.6 ± 34.2	1.38
2	41.26	0.420	0.830	0.317	0.370	0.250	98.80	88.6 ± 22.3	1.06
3	42.25	0.464	0.818	0.314	0.362	0.258	97.36	85.8 ± 21.9	1.03
4	43.17	0.454	0.807	0.355	0.357	0.262	94.48	77.9 ± 18.7	0.96
5	42.42	0.435	0.790	0.371	0.336	0.284	91.83	78.9 ± 17.4	0.95

NSP 图像及其有效的 SP 反事实之间的差异。我们进一步引入了 SonoSim, 它使用与 FSD 中使用的相似的 SonoNet-64 [2] 特征计算余弦相似度。为了测量生成图像的有效性, 我们使用计算机视觉中的标准指标, 例如翻转比率 (FR), 即反事实分类为 SP 的频率、原始 NSP 与反事实 SP 之间置信度预测的平均绝对差 [30] (MAD) 以及 KL 散度的有界重映射 [9] (BKL), 该指标衡量了由 f 预测的反事实与 SP 独热反事实标签之间的相似性。有效性指标是针对被 f 分类为 NSP 的那些 NSP 测试图像进行计算的。除了验证反事实确实根据引导分类器 f 向 SP 类移动之外, 我们还开发了一个渐进概念瓶颈模型 [13] (PCBM) 作为使用预言机的增长, 并利用其置信度来衡量输入 $\mathbf{x}$ 和反事实 $\mathbf{x}^c$ 的整体质量分数 (QS_O)。为了模拟一个真实的评估场景并确保我们分析中的可靠预言预测, 我们包括了对原始 NSP 输入有自信预测的案例, 即原始输入 $\mathbf{x}$ 被分类为 NSP, $QS_O(\mathbf{x}) < 0.5$ 。作为一个预言的有效性指标, 我们引入定义为 $MQD = \frac{1}{N} \sum_{i=1}^N I(QS_O(\mathbf{x}_i) < 0.5) \cdot QD_O(\mathbf{x}_i^c, \mathbf{x}_i)$ 的整体质量差异平均值, 其中 I 是指示函数, $N = 423$ NSP 测试图像和 $QD_O(\mathbf{x}^c, \mathbf{x}) = QS_O(\mathbf{x}^c) - QS_O(\mathbf{x})$ 。为了评估效率, 我们使用批处理大小为

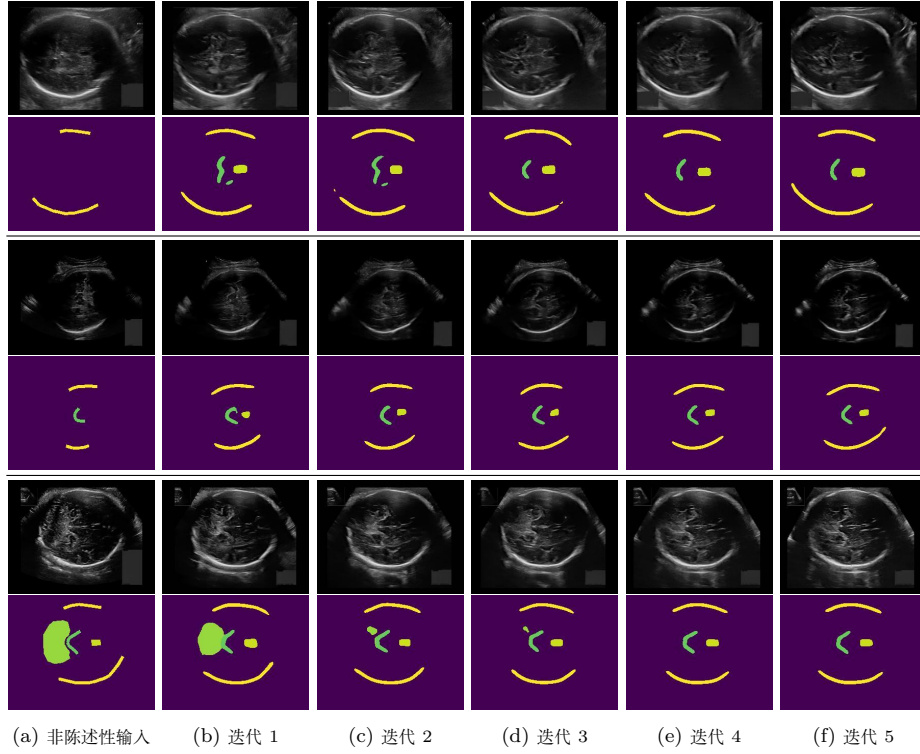


图 3: 从低质量的 NSP 到高质量 SP 的 Diff-ICE 迭代，使用预测分割进行可视化以增强可解释性（专家掩码注释用于 NSP 输入）。Diff-ICE 合成包含丘脑（TH）和透明隔腔（CSP）的平面（正确地来自 NSP 平面）。它还能够创建移除后窝（FP）的平面（正确地）。除了这些明确的更改，Diff-ICE 实现了广泛的变更和改进。

10 的 NVIDIA RTX A6000 计算批处理时间（秒）、总时间（小时）和 GPU 内存。

结果。我们为来自头部的所有 423 张 NSP 测试图像生成了反事实。表 1 和表 2 分别列出了所有方法和 Diff-ICE 每个中间迭代的结果。表 2 说明了预期的权衡，随着迭代次数的增加，现实性降低而有效性提高。为了评估身份保持 [5]，我们使用 SonoSim 并报告了来自 308 个独特胎儿的 417 个有效反事实情况中的排名-3（重新识别）准确性，验证 SP 反事实图像的胎儿身份是否在最相似的前 3 个输入 NSP 图像中。我们设置 $L = 5$ 是因为最后一次

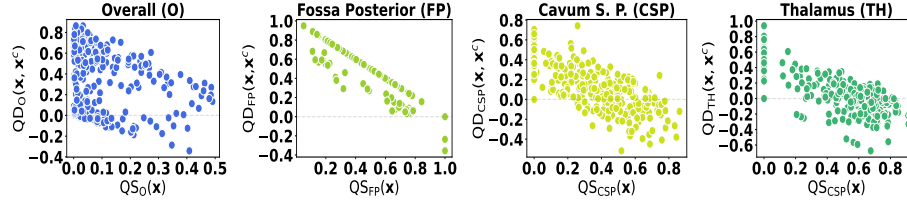


图 4: 质量差异 (QD) 作为 NSP 输入的质量分数 (QS) 的函数。

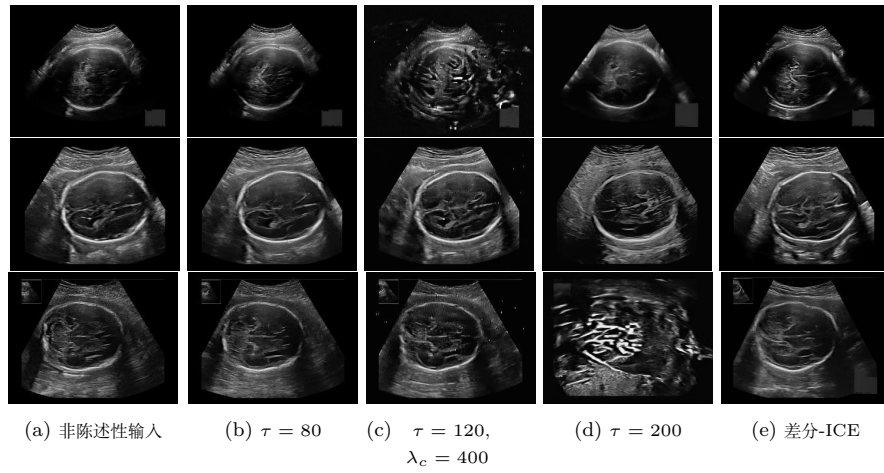


图 5: **迭代反事实**。Diff-ICE₁ 在列 (b)、(c) 和 (d) 中进行了说明。增加 λ_c (c) 可能会导致对抗性噪声。 τ (b) 的低值导致不显著的变化, 而高值 (d) 则可能导致完全改变解剖结构。

迭代之间的差异很小, 如图 3 所示, 这说明了从低质量 NSP 到更高质量 SP 的路径。除了图像级的 SP 置信分数外, oracle PCBM [13] 还预测个别质量得分 (QS_{FP} 、 QS_{CSP} 和 QS_{TH} 分别对应后凹 (FP)、透明隔腔 (CSP) 和丘脑 (TH))。作为对 Diff-ICE 的细粒度评估, 我们将反事实输入与 NSP 输入的质量分数差异 (QD_{FP} , QD_{CSP} , QD_{TH}) 作为输入质量的函数显示在图 4 中。

关于迭代反事实的必要性。我们评估了参数 τ (初始化噪声水平) 和 λ_c (引导强度), 在我们的单次迭代基线 Diff-ICE₁ 相对于迭代方法 Diff-ICE 的影响。如图 5 所示, 我们观察到通过简单增加引导强度来实现全局变化会导致

对抗样本的产生，而初始噪声水平过高则可能导致解剖结构的变化或坍塌为对抗样本。相比之下，Diff-ICE 有效提高了质量同时保持了图像保真度。

定性验证。具有 10 年产科成像经验的胎儿医学顾问被要求从真实 NSP 图像及其相关的 Diff-ICE 反事实图像对中选择最佳质量的图像。包含 $QD_O > 0$ 的图像对是均匀采样的，并以随机位置和顺序显示。专家在 50 对图像中的 41 对选择了 Diff-ICE 反事实情况，这表明了 Diff-ICE 确实能够提升输入的质量。专家还定性地指出差异是全球性的，包括改进了骨骼轮廓的呈现、**TH** 和 **约束满足问题** 以及整体图像质量的提升。

5 讨论与限制

我们的结果显示，迭代方法生成的反事实具有更高的置信度和更广泛的改变，同时在很大程度上保持了图像保真度和胎儿身份。此外，我们通过采用高效的梯度估计方案来维持计算可行性。图 4 显示，当初始质量较差时改进最大，这是直观的，因为对于接近 SP 的高质量输入的变化应该是最小的。然而，对于所有结构，质量一般都有所提高。Diff-ICE 在去除 **FP** 方面特别有效，而在 **TH** 和 **约束满足问题** 较为突出的情况下，它则在整体质量和个别结构质量之间取得平衡。

尽管我们的研究结果在图像保真度和胎儿超声特定指标方面颇具前景，但缺乏包含同一胎儿 NSP 和 SP 数据集的限制了我们进行直接比较、评估幻觉以及对生物测量影响的评价能力——所有这些都是临床验证的关键。然而，能够可视化通往 SP 的合理路径为个性化教育提供了强大的潜力，支持超声技师培训，并使非专家能够标注数据集 [3]。未来的工作可以整合视频扫描中的胎儿特定先验知识（如果可用），或直接适应反事实视频生成方法 [28]，并专注于临床评估，以支持下游诊断和监测的 SP 获取。

致谢。本工作得到了人工智能先驱中心（DNRF 资助编号 P1）、DIREC 项目 EXPLAIN-ME（9142-00001B）以及诺和诺德基金会通过生命科学中的基本机器学习研究中心（NNF20OC0062606）的支持。P.Pegios 想要感谢 Thanos Delatolas 的见解深刻的讨论。

参考文献

1. Bansal, A., Chu, H.M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., Goldstein, T.: Universal guidance for diffusion models. In: CVPR. pp. 843–852 (2023)
2. Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D.: Sononet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. *IEEE transactions on medical imaging* **36**(11), 2204–2215 (2017)
3. Chiquier, M., Avrech, O., Gandelsman, Y., Feng, B., Bouman, K., Vondrick, C.: Teaching humans subtle differences with diffusion. *arXiv preprint arXiv:2504.08046* (2025)
4. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* **34**, 8780–8794 (2021)
5. Gabler, E., Nissen, M., Altstidl, T.R., Titzmann, A., Packhäuser, K., Maier, A., Fasching, P.A., Eskofier, B.M., Leutheuser, H.: Fetal re-identification in multiple pregnancy ultrasound images using deep learning. In: EMBC. pp. 1–4. IEEE (2023)
6. He, Y., Murata, N., Lai, C.H., Takida, Y., Uesaka, T., Kim, D., Liao, W.H., Mitsufuji, Y., Kolter, J.Z., Salakhutdinov, R., et al.: Manifold preserving guided diffusion. In: The Twelfth International Conference on Learning Representations (2023)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
8. Iskandar, M., Mannering, H., Sun, Z., Matthew, J., Kerdegari, H., Peralta, L., Xochicale, M.: Towards realistic ultrasound fetal brain imaging synthesis. In: Medical Imaging with Deep Learning, short paper track (2023)
9. Jeanneret, G., Simon, L., Jurie, F.: Diffusion models for counterfactual explanations. In: Asian Conference on Computer Vision. pp. 858–876 (2022)
10. Jeanneret, G., Simon, L., Jurie, F.: Adversarial counterfactual visual explanations. In: CVPR (2023)
11. Lee, L.H., Noble, J.A.: Generating controllable ultrasound images of the fetal head. In: 17th International Symposium on Biomedical Imaging (ISBI). pp. 1761–1764. IEEE (2020)
12. Lin, M., Ambsdorf, J., Sejer, E.P.F., Bashir, Z., Wong, C.K., Pegios, P., Raheli, A., Svendsen, M.B.S., Nielsen, M., Tolsgaard, M.G., et al.: Learning semantic image quality for fetal ultrasound from noisy ranking annotation. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2024)

13. Lin, M., Feragen, A., Bashir, Z., Tolsgaard, M.G., Christensen, A.N.: I saw, i conceived, i concluded: Progressive concepts as bottlenecks. *arXiv:2211.10630* (2022)
14. Lin, M., Zepf, K., Christensen, A.N., Bashir, Z., Svendsen, M.B.S., Tolsgaard, M., Feragen, A.: Dtu-net: Learning topological similarity for curvilinear structure segmentation. In: *International Conference on Information Processing in Medical Imaging*. pp. 654–666. Springer (2023)
15. Maack, L., Holstein, L., Schlaefer, A.: Gans for generation of synthetic ultrasound images from small datasets. *Current Directions in Biomedical Engineering* **8**(1), 17–20 (2022)
16. Mei, X., et al.: Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence* **4**(5), e210315 (2022)
17. Men, Q., Zhao, H., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Towards standard plane prediction of fetal head ultrasound with domain adaption. In: *20th International Symposium on Biomedical Imaging (ISBI)*. pp. 1–5. IEEE (2023)
18. Men, Q., Zhao, H., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Scanahead: Simplifying standard plane acquisition of fetal head ultrasound. *Medical Image Analysis* p. 103614 (2025)
19. Mikolaj, K., Lin, M., Bashir, Z., Svendsen, M.B.S., Tolsgaard, M., Nymark, A., Feragen, A.: Removing confounding information from fetal ultrasound images. *arXiv:2303.13918* (2023)
20. Mishra, D., Zhao, H., Saha, P., Papageorghiou, A.T., Noble, J.A.: Dual conditioned diffusion models for out-of-distribution detection: Application to fetal ultrasound videos. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 216–226. Springer (2023)
21. Olsen, M.D.S., Ambsdorf, J., Lin, M., Taksøe-Vester, C., Svendsen, M.B.S., Christensen, A.N., Nielsen, M., Tolsgaard, M.G., Feragen, A., Pegios, P.: Unsupervised detection of fetal brain anomalies using denoising diffusion models. In: *International Workshop on Advances in Simplifying Medical Ultrasound*. pp. 209–219. Springer (2024)
22. Pegios, P., Feragen, A., Hansen, A.A., Arvanitidis, G.: Counterfactual explanations via riemannian latent space traversal. *arXiv preprint arXiv:2411.02259* (2024)
23. Pegios, P., Sejer, E.P.F., Lin, M., Bashir, Z., Svendsen, M.B.S., Nielsen, M., Petersen, E., Christensen, A.N., Tolsgaard, M., Feragen, A.: Leveraging shape and spatial information for spontaneous preterm birth prediction. In: *International Workshop on Advances in Simplifying Medical Ultrasound*. pp. 57–67. Springer (2023)

24. Salomon, L., et al.: Isuog practice guidelines: ultrasound assessment of fetal biometry and growth. *Ultrasound in obstetrics & gynecology* **53**(6), 715–723 (2019)
25. Sanchez, P., Kascenas, A., Liu, X., O’ Neil, A.Q., Tsaftaris, S.A.: What is healthy? generative counterfactual diffusion for lesion localization. In: *MICCAI Workshop on Deep Generative Models*. pp. 34–44. Springer (2022)
26. Singla, S., Eslami, M., Pollack, B., Wallace, S., Batmanghelich, K.: Explaining the black-box smoothly—a counterfactual approach. *Medical Image Analysis* **84**, 102721 (2023)
27. Sobieski, B., Grzywaczewski, J., Sadlej, B., Tivnan, M., Biecek, P.: Rethinking visual counterfactual explanations through region constraint. In: *The Thirteenth International Conference on Learning Representations* (2024)
28. Spyrou, N., Vlontzos, A., Pegios, P., Melistas, T., Gkouti, N., Panagakis, Y., Papanastasiou, G., Tsaftaris, S.A.: Causally steered diffusion for automated video counterfactual generation. *arXiv preprint arXiv:2506.14404* (2025)
29. Wang, F., Whelan, K., Silvestre, G., Curran, K.M.: Generative diffusion model bootstraps zero-shot classification of fetal ultrasound images in underrepresented african populations. In: *International Workshop on Preterm, Perinatal and Paediatric Image Analysis*. pp. 143–154. Springer (2024)
30. Weng, N., Pegios, P., Petersen, E., Feragen, A., Bigdeli, S.: Fast diffusion-based counterfactuals for shortcut removal and generation. In: *European Conference on Computer Vision*. pp. 338–357. Springer (2024)
31. Wu, L., Cheng, J.Z., Li, S., Lei, B., Wang, T., Ni, D.: Fuiqa: fetal ultrasound image quality assessment with deep convolutional networks. *IEEE transactions on cybernetics* **47**(5), 1336–1349 (2017)
32. Yu, J., Wang, Y., Zhao, C., Ghanem, B., Zhang, J.: Freedom: Training-free energy-guided conditional diffusion model. In: *CVPR*. pp. 23174–23184 (2023)
33. Zhao, H., Zheng, Q., Teng, C., Yasrab, R., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Memory-based unsupervised video clinical quality assessment with multi-modality data in fetal ultrasound. *Medical Image Analysis* **90**, 102977 (2023)