arXiv:2303.01555v1 [cs.CV] 2 Mar 2023

生成式评估的反事实编辑

Maria Lymperaiou[1], *, GiorgosFilandrianos[1], *, KonstantinosThomas[1], *
和Giorgos Stamou[1]

[1]*AILS*实验室，电气与计算机工程学院，雅典国立理工大学

摘要

尽管生成架构迅速发展，生成模型的评估仍是一个未得到充分研究的领域。大多数最新模型仍采用相当过时的指标进行评估，这些指标存在稳健性问题，同时无法评估视觉质量的更多方面，例如组合性和合成逻辑。与此同时，生成模型的可解释性仍是一个受限但重要的研究方向；当前的一些尝试要求能够访问生成模型的内部功能。不同于以往研究，我们将生成模型视为黑箱，并提出了一个基于概念而非像素的框架，用于评估和解释合成结果。我们的框架利用基于知识的反事实编辑，指出应从生成图像中插入、移除或替换哪些物体或属性，以使其接近真实条件。此外，通过累积局部编辑而生成的全局解释，还可以揭示模型总体上无法生成哪些概念。我们将该框架应用于为故事可视化和场景合成等具有挑战性的任务而设计的各种模型，验证了该方法在模型无关设置中的有效性。

关键词

图像生成、反事实解释、扩散模型、故事可视化、生成式评估、可解释人工智能

1. 引言

图像生成一直是最受欢迎的深度学习任务之一，催生了许多令人印象深刻的最先进的应用 [1, 2, 3, 4, 5, 6, 7]。自生成对抗网络（GANs） [8],问世以来——这标志着该领域最早的重大突破之一——大多数应用都专注于根据人类感知提升图像质量。同时，由于没有可用于对比的真实标注数据，生成样本的自动评估仍然是一个长期存在的问题。到目前为止，此类生成任务的评估依赖于像素级指标，例如 Inception分数（IS） [9], 、Fréchet Inception距离（FID） [10], 和学习感知图像块相似度（LPIPS） [

11], 为生成样本提供质量度量指标。因此，基于这些基准指标进行评估的文献数量众多；然而，人们已指出，这些指标的脆弱性 [12]正导致

收录于 *A. Martin*、*K. Hinkelmann*、*H.-G.Fill*、*A.Gerber*、*D. Lenat*、*R. Stolle*、*F. van Harmelen*（编）主编的《美国人工智能协会*2023*年春季研讨会论文集：需要结合机器学习与知识工程的挑战》（*AAAI-MAKE 2023*），凯悦酒店，旧金山机场，加利福尼亚州，美国，*2023*年*3*月*27*日至*29*日。*通讯作者。“marialymp@islab.ntua.gr （M. Lymperaiou）；geofila@islab.ntua.gr （G. Filandrianos）；kthomas@islab.ntua.gr （K. Thomas）；gstam@cs.ntua.gr （G. Stamou） 0000-0001-9442-4186 （M. Lymperaiou）； 0000-0002-7489-7776 （K. Thomas）； 0000-0003-1210-9874 （G. Stamou）©2023年本文版权归根据知识共享署名4.0国际许可协议（CC BY 4.0），允许使用。CEUR研讨会论文集（CEUR-WS.org）

结果不准确。尽管 Clean-FID 等近期指标 [12] 能够解决一些有关视觉伪影的问题，但仍无法处理图像复杂性、组合性、逻辑和生成公平性等重大问题 [13]。此外，对于条件生成，我们还需要衡量条件输入中提及的物体和属性是否成功呈现在生成样本中。目前，条件合成评估的尝试仍然有限 [14, 15] ，同时也面临其所依托的无条件对应方法的不足之处。

生成模型的可解释性是另一个新兴领域，但目前相较于判别模型 [16, 17]受到的关注少得多。在生成对抗网络（GAN）中引入可解释反馈已展现出颇具前景的研究方向 [18], , 而其他研究则专注于解读GAN内部结构 [19]。通过确定促使样本被判定为伪造或真实的图像区域，可以解释判别器的决策，从而应对生成对抗网络中的过拟合问题 [20]。尽管其他人工智能领域（例如自然语言处理）中的相关尝试已取得进展 [21, 22, 23]，有限的相关文献仍阻碍了生成模型可解释评估的发展。

我们认为，要解决生成式评估中的挑战，需要采用一种针对评估过程的概念化方法，而不是像素级路径。相较于像素，依赖概念能够提升评估过程的可解释性，并为生成模型的可解释评估铺平道路。识别能够或不能够被生成的概念（物体或属性），可以揭示相关模型的能力与偏差，从而推动潜在的架构修改。在本文中，我们提出了首个面向生成模型的可解释评估技术。具体而言，我们利用反事实解释，将条件生成评估表述为以下问题的答案：生成样本X需要改变哪些概念，才能达到其条件输入c？由外部知识源引导的概念编辑 [24] 已被证明能够有效指示到达条件概念的最短可能路径。此外，现有将可解释性与图像生成相结合的研究都针对特定模型[18, 19, 20] ，并要求访问其内部结构（白盒技术）；而我们的方法只需要生成的输出及其真实条件，同时仍将生成模型视为黑盒。因此，我们作出如下贡献：

1. 我们提出首个基于概念 而非基于像素的生成式评估框架[1]，，适用于场景生成（SG）和故事可视化（SV）等多种任务。2. 我们的指标在设计上具有可解释性 ，能够说明生成图像中需要插入、删除或替换哪些概念，才能使其接近真实条件。这些编辑操作在模型无关设置下执行，完全无需访问生成模型的内部机制。3. 全局解释能够自动揭示生成模型可能存在的盲点 ，即模型本质上无法生成的概念。

2. 相关工作

Ge生成对抗网络（GAN）[8]由两个神经网络组成：一个生成 -

[1]https://github.com/geofila/Counterfactual-Edits-for-Generative-Evaluation

器 $G(z;\theta_g)$和一个判别器 $D(x;\theta_d)$。 G 将从先验分布 $z \sim p_z$生成的随机噪声 z映射到数据空间。另一方面， D将来自同一数据空间的样本 x_i 映射为标量值 $p_i = D(x_i)$，该值表示该样本取自真实数据分布的概率。在条件生成对抗网络（cGAN）中， [25]不仅接收随机噪声 z，还接收额外的条件向量 y，这有助于引导生成来自目标分布特定子区域的样本。

若干图像生成条件生成对抗网络 [26, 27] 在生成具有鲜明纹理和颜色的图像方面表现出色。然而，由于卷积滤波器的局限性，它们往往难以生成连贯的整体物体结构以及捕捉其他长程依赖关系。自注意力生成对抗网络（SAGAN） [28] 旨在解决这一问题；它在 G 和 D中都采用了自注意力模块，并运用了权重谱归一化等现代稳定化技术 [29], ，同时利用双时间尺度更新规则 [30] 为 G 和 D设置不同的学习率。

扩散模型正在条件图像生成领域开辟新天地，并逐渐成为该领域的最先进技术 [2]。这些模型通过向图像添加噪声，再学习重建图像来运行。过去一年，基于扩散的图像合成领域取得了若干令人振奋的进展。稳定扩散 [31]将扩散过程应用于自编码器的潜在空间，而非图像空间中的像素级，从而即使在资源受限的情况下也能实现高质量图像合成。DALL-E2[4]在前代模型 [3]的成功基础上，将从 CLIP [32] 学习到的文本条件图像嵌入作为扩散模型的输入，并由该模型充当解码器。生成的图像具有照片级真实感，能够准确呈现输入文本，同时还支持通过语言引导操控源图像。Imagen [33] 的工作利用大型预训练语言模型（如 T5 [34], ）进行语言编码，并采用扩散过程进行图像合成。DreamBooth [35] 更进一步：只需提供对所需场景的文字描述，即可实现上下文感知图像合成。这使得模型能够生成各种视觉主体，同时保持较高的图像合成质量。

条件图像合成自文本条件图像生成 [36, 37]问世以来取得了长足进步。早期的尝试生成的图像缺乏细节，质量也不高。Stack- GAN [38] 是首个通过多阶段对抗训练显著提升生成图像质量的模型，随后出现的 StackGAN++ [39] 进一步改善了生成效果。AttnGAN [40] 运用注意力机制，根据输入文本中的单个词语生成图像的细粒度细节。SEGAN [41] 更进一步，只关注输入文本中的相关关键词。DM-GAN [42] 则通过处理模糊区域提升了生成图像的质量。

StoryGAN [43]是一种基于序列输入合成图像的生成模型（故事可视化），采用循环神经网络结构对输入文本进行编码，并向生成器提供上下文信息。生成器通过与两个判别器进行对抗训练：图像判别器用于评估图像质量和文本-图像相关性，故事判别器则确保在给定完整故事上下文的情况下图像之间的一致性。近期工作主要致力于改进基线 StoryGAN 模型 [44] 并探索替代方案

故事编码方法，例如使用 Transformer 架构 [45, 46, 47]。

生成式评估 尽管图像合成取得了快速进展，但由于主要出于基准测试目的而沿用过时的评估实践 [9, 10, 11, 48], ，生成式评估逐渐落后，并忽视了近期研究中指出的问题 [12, 13]。生成式建模中的可解释性能够带来有趣的洞见，但当前的研究要么仍然针对特定模型 [49, 18, 19, 20]，要么需要发现可解释的潜在方向[50, 51, 52, 53], ，而这是一项并不简单的任务。我们的方法在单一框架下同时服务于生成模型的评估与可解释性，并且能够适配任何生成模型——甚至是用于序列图像生成的模型 [43, 45, 46, 47]——因为它只关注输入和输出概念集。

反事实解释 当代人工智能研究正致力于解释神经网络的“思维”过程，因此，可解释人工智能（XAI）日益受到关注[54]。反事实解释基于最小的输入修改提供替代现实，从而揭示推理路径。在探索视觉替代方案时，生成模型是一种直接的方法 [55]。任何改动都应符合原始数据分布，这一要求为替代输入增加了限制 [56]。确定最小改动时，可以选择干预神经网络的黑箱特性[57], ，也可以不干预 [24]。我们选择黑盒路线，利用反事实解释来揭示生成模型推理过程的线索。

3. 作为反事实解释的概念编辑

我们整体的方法深受 Filandrianos 等人 [24],的启发。他们探讨了反事实推理的根本问题：“要让图像 I 被分类为 X 而不是 Y，需要发生最小 的变化是什么？”，其中 X 和 Y 是预定义图像分类器 F预测的类别。在我们的情形下， F 并非必要，因为我们默认将所有生成概念 s 置于一个集合 S中，并将所有真实概念 t 置于一个集合 T中。反事实解释能够回答上述问题，为所有 $s \in S$、 $t \in T$实现 $S \to T$ 转变所需的概念编辑数量提供最小值。

概念距离指明了连接两个特定概念的最短路径。概念层级用于以确定的方式定义概念之间的转换成本。我们探讨了两种选择：使用外部层级知识（例如 WordNet [58], 将提取的概念映射到同义词集），或者手动构建特定层级，以便对语义距离进行高度可控的定义。在这两种情况下，我们都将概念 s 和 t之间的距离记作 $d(s,\ t)$。有三种可用的概念**编辑操作**来实现转换：

- **替换（R）** $e_{s \to t}(S)$：将概念 $s \in S$ 替换为概念 $t \notin S$。
- **删除（D）** $e_{s-}(S)$：从 S中删除概念 $s \in S$ 。
- **插入（I）** $e_{s+}(S)$：在 S中插入概念 $s \in S$ 。

每个编辑操作都继承所选层级结构规定的概念距离。因此，**R** 操作会考虑 s 与 t 之间的路径，从而确保$min(d(s,t))$。正如

[24], d 还确保编辑具有可操作性，允许进行具有语义意义的转换（例如“食物”→“意大利面”），同时禁止无意义的转换（例如“食物”→“天空”）。**D** 和**I** 操作分别将层级结构的根节点 视为 t 和 s；对于 WordNet，entity.n.01 充当根节点。**概念集编辑距离 (CSED)** $D(S \to T)$通过汇总所有可能的最小成本编辑操作获得，从而最终实现 $S \to T$：

$$CSED = D(S \to T) = min \sum_{s \neq t}^{S,T} \sum^{R,D,I} d(s,t) \tag{1}$$

4. 方法

我们方法的核心是一个预训练黑盒生成模型 M，它接收语义描述 c（以自然语言或符号形式）作为条件输入，并生成图像 I，该图像与 c相对应。随后，我们使用目标检测、语义分割等现成的自动化方法，从生成图像中提取所有描绘的概念 I，并将它们添加到生成（或源）概念集中 S。同样，从 c 中提取的概念会添加到真实（或目标）概念集中 T。 c 的格式决定了所采用的概念提取技术：如果c 是文本句子，则采用语言学概念提取；如果 c 已经是集合格式，则只需进行简单的预处理。最终，我们希望回答以下问题：“从到 S， T究竟需要进行哪些最小改动？”我们的方法概述见图1。

4.1. 生成式评估

第3 节介绍的反事实骨干阐明了我们的生成式评估流程，该流程用于生成领域中两个难度较高的任务：**故事可视化（SV）**和**场景生成（SG）**。

故事可视化（SV）旨在按顺序生成图像 I_1， I_2，……， I_L，使其与给定的、总长度为 L 的顺序条件输入 c_1， c_2，……， c_L 一一对应。生成图像不仅需要忠实于其条件输入，还需保持序列一致性。因此，我们定义了故事可视化适用的两个目标：

- **忠实性**：对于任意故事帧 k，其中 $1 \leq k \leq L$， c_k 中提及的物体和属性也应出现在帧 I_k 中。
- **一致性**：对于任意故事帧 k，其中 $1 \leq k < L$，出现在帧 I_k 中的物体或属性不能在后续帧 I_{k+1}，……， I_L 中消失或发生变化。

由于定义明确的语义与反事实解释密切相关 [59], ，我们认为 CLEVR-SV [60] 是展示我们方法的理想数据集，因为它提供了一组概念 $\mathcal{C}$：形状 *(*立方体、球体、圆柱体*)*、尺寸 *(*小、大*)*、材质 *(*橡胶、金属*)* 以及八种颜色中的一种 *(*蓝色、青色、棕色、黄色、红色、绿色、紫色、灰色*)*。每个 CLEVR-SV 物体都包含 $|\mathcal{C}|$=4概念

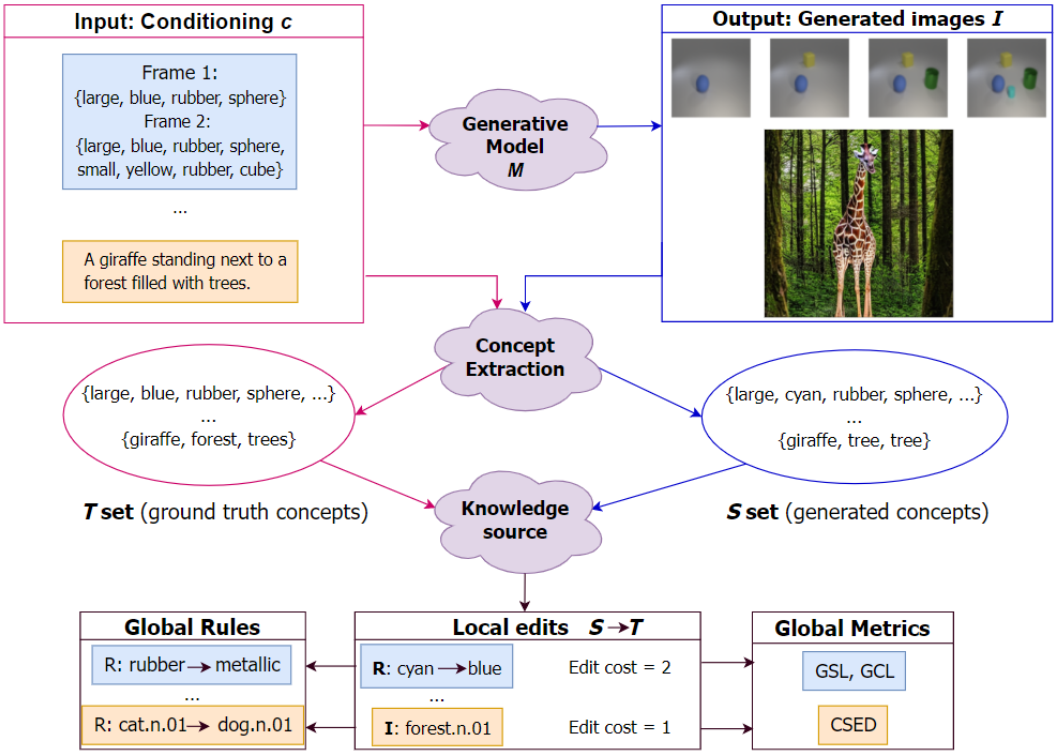


图1： 所提出框架的概述。

用于描述其形状、尺寸、材质和颜色。我们手工构建了一个简单的层级结构，将物体语义归入通用概念类别，以展示以下包含关系：

$$\text{(大，小)} \subset \text{尺寸（蓝色、黄色、棕色、灰色、绿色、紫色、青色、红色）} \subset \text{颜色（金属质、橡胶）} \subset \text{材质（球体、立方体、圆柱体）} \subset \text{形状} \tag{2}$$

CLEVR-SV包含长度为 L=4的故事，其中第 k帧严格包含 k 个物体。三种可用的编辑操作都可能适用于逐帧处理：当生成帧中的物体数量多于与之匹配的真实标注帧时，对某个概念进行**D**；在相反情况下，对某个概念进行**I**；**R** 相当于先进行**D**，再进行**I**，当帧中的物体数量正确但语义不同时，可应用此操作。在默认情况下，我们为所有语义以及**D** 和**I** 操作设定相同的成本1（**R** 的成本是**D** 和**I** 成本之和）。

为衡量故事**忠实性**，我们提出故事损失（***SL***）度量指标，该指标将故事中第 $k = 1$、2、.....、L 帧各自的概念集编辑距离（$CSED_k$）相加。第k帧生成的 CLEVR-SV 形状、尺寸、材质和颜色语义构成概念集 S_k，而条件输入的语义构成 T_k，其 $CSED_k$ 记为D（S_k，T_k）。因此，对应于将生成序列的语义转换为

$\{I_1, I_2, I_3, I_4\}$ 条件输入 $\{c_1, c_2, c_3, c_4\}$ 的语义所需的最低成本R D I 编辑的转移成本可表示为：

$$SL = \sum_{k=0}^{k=L} CSED_k = \sum_{k=0}^{k=L} D(S_k, T_k), \qquad L = 4 \tag{3}$$

通过扩大对包含 N 个故事的数据集计算故事损失（SL）的规模，我们得到了全局故事损失（***GSL***）度量指标：

$$GSL = \sum_{i=0}^{N} SL_i = \sum_{i=0}^{N} \sum_{k=1}^{k=L} D(S_k, T_k)_i \tag{4}$$

对于故事一**致性**，我们提出了一种名为一致性损失（***CL***）的度量指标：将帧 I_k 与生成序列中的 I_{k-1}， $k = 2$，..... L 帧进行比较，以捕捉语义变化。CL 的一个难点在于不存在真实概念集。不过，根据任务定义，我们知道第*k*帧包含*k* 个物体，并且数据集概念的基数是预先定义的（对于 CLEVR-SV，$|\mathcal{C}|$=4 ），因此可以假设，每个先前的帧都构成与概念集 T对应的“真实标注”。从*k*=1 帧开始，我们期望 T的基数等于 $|\mathcal{C}| \cdot k = |\mathcal{C}|$。任何差异都会导致对*k*=1的惩罚 $p_k = |T| - |\mathcal{C}| \cdot k = CL_k$。对于后续帧，我们将第*k*帧对应的概念集定义为 S，并将由*k*-1帧中的概念组成的“真实标注”集合定义为 T。数学上，CL 可写为：

$$CL = p_{k=1} + \sum_{k=2}^{k=L} D(S_k, T_k), \qquad T_k = S_{k-1}, L = 4 \tag{5}$$

在理想情况下，当第*k*帧包含具有 $\mathcal{C}$ 语义的*k* 个物体时，我们期望$p_{k=1} = 0$ 和 $CL_{k>1} = |\mathcal{C}| \cdot ($ $k-1$）。通过将一致性损失扩展到 N 个故事，全局一致性损失（***GSL***）从总体上评估生成模型 M 的一致性能力：

$$GCL = \sum_{i=0}^{N} CL_i = \sum_{i=0}^{N} \{p_{k=1,i} + \sum_{k=2}^{k=L} D(S_k, T_k)_i, \;\; T_{k,i} = S_{k-1,i}\} \tag{6}$$

可分别获得局部（**故事损失/一致性损失**）和全局（**GSL/GCL**）指标的平均值：

$$Avg\ SL = \frac{1}{k} SL, \qquad Avg\ GSL = \frac{1}{N}[Avg\ SL] = \frac{1}{N} GSL \tag{7}$$

对于一致性而言，与其导出 $\sum CL_k$上的平均值，不如统计 $p_{k=1} = 0$， $CL_{k>1} = |\mathcal{C}| \cdot$（$k-1$）要求未被遵守的次数，并对 $L = k$ 帧取平均，这样更有意义：

$$Avg\ CL = \frac{p_{k=1}}{k} + \frac{1}{k} \sum_{k=1}^{k=L} [CL_{k>1} \neq |\mathcal{C}| \cdot (k-1)], \quad Avg\ GCL = \frac{1}{N}[Avg\ CL] \tag{8}$$

SL 和**CL** 本质上具有可解释性，因为它们不仅提供质量衡量，还会揭示 $S_k \to T_k$ 编辑路径。这些路径可作为局部反事实解释，突出显示

该特定故事中错误生成的语义，具体表现为**忠实性**或**一致性**方面的问题。总体而言，较高的**SL/GSL** 和**CL/GCL** 值表示较低的概念生成质量。**GSL/GCL** 编辑路径对应于全局反事实解释：规则提取技术提供频繁模式，总结正在研究的 M 的行为。频繁的**GSL** 编辑路径实际上包含常见错误概念，即条件概念M 难以生成的概念。同样，**GCL** 编辑路径揭示了频繁的不一致模式，展示了在故事帧中任意发生变化的概念。因此，通过研究“为了从 S 转变为 T，必须最小程度地改变什么？”这一问题，我们最终回答了一个更为一般的问题：“ M无法生成或保留哪些概念？”

场景生成（SG）旨在根据条件输入 c合成视觉场景 I。合成的图像包含多个彼此交互的物体。场景中的物体还具有各自的属性。与为 SV 提供的条件输入相比，给定的条件输入 c 更为复杂，因为待生成的概念数量众多且并未预先定义；因此，概念集 $\mathcal{C}$ 的基数未知，但相对较大。

COCO 数据集 [61] 为评估场景生成中的生成式**忠实性**提供了理想环境，其中的文本描述 c 可用作条件输入。我们着重研究 Huggingface[2]，上的最先进开源扩散模型 [2]，具体包括 Stable Diffusionv1.4& v2 [62, 63] 和Protogen x3.4 & 5.8 [64, 65]（详情见附录）。这些模型能够生成逼真的图像——这是概念提取（目标检测）阶段的一个重要方面。由于视觉质量较差，且依赖场景图和布局来确保构图恰当，我们不再讨论较早的场景生成架构 [66, 67, 68, 69, 70, ，等等]。

在概念提取阶段，利用 YOLO-v8 [71] and YOLOS [72] 目标检测器构建生成概念集 S。由于 c 采用文本格式，因此使用 spaCy [73] 从构成目标概念集 T 的描述中提取真实概念。与 COCO 概念相关的概念距离在语义上较为复杂，因此需要采用丰富的知识体系，例如 WordNet。举例来说，如果 c 指的是“食物”或“动物”等概念，扩散模型可能会生成更细化的“食物”或“动物”实例，例如“意大利面”和“狗”。随后，目标检测器会返回这些细化类别，从而在转换过程中引入一些噪声。层级知识可以消除这类问题：尽管 $T = \{$食物，动物$\} \neq S = \{$意大利面、狗$\}$，但如果考虑到将 S 和 T 概念映射到 WordNet 同义词集后得到的层级关系**意大利面** –是 –食物以及狗 –是 –动物，这两组概念在**语义上等价**。在这种情况下，无需执行 $S \to T$ 转换。因此，使用外部知识可以实现概念上更准确的转换。此外，WordNet 还提供编辑操作所需的概念距离，能够准确反映概念之间的语义关系。随后，可应用 CSED 计算 $S \to T$ 转换的总成本。

[2]https://huggingface.co/models?pipeline_tag=文本生成图像&sort=downloads

5. 实验

5.1. 故事可视化

由于所有语义和**D, I** 编辑操作的成本相同，因此我们为所有语义以及**D,I**都赋予 d=1 。例如，删除一种颜色的编辑成本为 1。或者，将一种颜色替换为另一种颜色的编辑成本为 2，与先删除源颜色再插入目标颜色的成本相同。同样的逻辑也适用于物体的形状、尺寸和材质。

指标结果涵盖所选故事可视化模型的最佳变体 [43, 45, 46, 47] ，如表1所示。现有指标（FID、Clean-FID、LPIPS、SSIM）也一并提供，以供比较。

总体而言，我们观察到像素级指标与概念指标之间具有一致性。这在一定程度上是预期之中的，因为概念提取阶段依赖于像素级图像质量，生成的物体或语义质量越好，就越容易被识别。尽管如此，概念评估能够提供更多可解释的洞见：我们给出了每个概念（材质、尺寸、形状、颜色）的损失百分比，突出展示了所研究模型在不同语义上的优势与不足之处。例如，所有模型较高的形状损失（> 50%）表明，它们在大多数情况下合成的物体形状较为模糊。另一方面，相对较低的大小损失揭示了这些模型生成尺寸正确的物体的能力。

表1。 M中每个=4 故事的所有L个故事的平均评估指标（现有指标和所提指标以竖线分隔）。

M	FID ↓	干净-FID↓	LPIPS ↓	SSIM ↑	GCL ↓	GSL ↓	材质 ↓	尺寸 ↓	形状 ↓	颜色 ↓
[47]	41.54 ± 8.55	**115.46**	**0.21** ± 0.05	**0.71**	**4.97**	**7.01**	**20.83%**	**14.55%**	**56.62%**	**33.10%**
[43]	**41.45** ± 6.25	123.40	0.25 ± 0.03	0.65	11.44	15.33	30.89%	21.12%	62.34%	37.44%
[46]	41.96 ± 9.66	124.97	0.25 ± 0.08	0.67	10.95	8.06	21.45%	16.02%	56.78%	35.10%
[45]	41.80 ± 8.81	122.62	0.25 ± 0.05	0.68	8.32	11.51	25.34%	16.71%	56.83%	35.14%

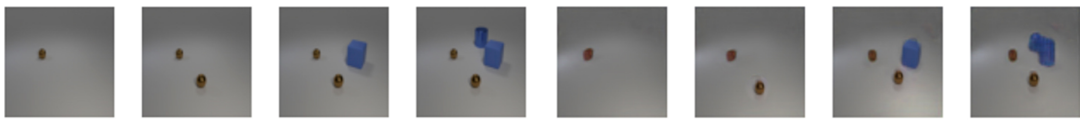
我们进一步聚焦于 Tsakas 等人表现最佳的 SV 模型，以深入探究我们的发现。 [47] 根据表1中报告的概念指标。具体而言，在表2 中，我们给出了逐帧**GSL**、**GCL** 以及各概念（材质、尺寸、形状、颜色）的损失。

表 2。 [47] 的逐帧平均概念评估指标。

帧	GSC↓	GSL↓	材质↓	尺寸↓	形状↓	颜色↓
1st	0.00	2.25	40.00%	6.20%	58.75%	7.50%
2nd	4.35	5.66	20.00%	11.88%	57.5%	32.50%
3rd	7.12	8.25	13.33%	16.67%	57.08%	43.33%
4th	8.42	11.49	10.00%	23.44%	53.13%	49.06%

局部解释 通过获取以下内容，验证了所提出的**故事损失**/**一致性损失** 指标的透明度局部解释针对 [47]。具体而言，我们考察图 2 中序列的编辑路径：

最左侧的 4 张图像（图 2a）对应真实序列，最右侧的 4 张图像（图 2b）则表示生成帧。因此， S 包含图 2b 的概念，而 T包含图 2a 的概念。如表 3 所示（详情见附录），所有帧中均观察到标准的 R 操作，这表明需将棕色小球的材质从“橡胶”转换为“金属质”，以匹配真实标注。在最后一帧中，还增加了一项 R 操作，这表明还需将新对象的形状从“球体”转换为“圆柱体”。每项 R 操作的成本为 2，相当于一步删除错误语义，再用一步添加正确语义。不过，如有需要，可以适当调整此成本权重。该故事的故事损失为 10，即逐帧汇总所有操作成本。通过观察一致性损失，我们发现每个后续帧中添加的物体数量都正确，因此维持了 $\mathrm{CL}_{k>1} = |\mathcal{C}| \cdot (k-1)$ $|\mathcal{C}| = 4$：从第 k=1 帧的 $\mathrm{CL}_1 = p_{k=1} = 0$ 开始，我们确认只添加了一个物体，符合帧编号应等于其中物体数量的要求。CL_2=4 是预期结果，因为第 k=2 帧中添加的物体包含 4 种语义。任何低于或高于该数量的情况都表明存在异常行为：$\mathrm{CL}_{k>1} < |\mathcal{C}| \cdot (k-1)$表示缺失了一个或多个物体，而 $\mathrm{CL}_{k>1} > |\mathcal{C}| \cdot (k-1)$则表示生成了一个或多个额外物体。第三帧和第四帧也呈现出预期的模式。通过这项分析，我们揭示了 [47] 在这张特定图像上的不足之处，并得出如下局部解释：需要进一步检查语义材质，因为在本例的所有故事帧中，生成的棕色小球的属性都是“橡胶”，而非“金属质”。若要深入了解模型合成离散语义的能力，则需要推导全局指标并给出解释。



(a) 真实故事帧　　(b) [47]生成的故事帧

图 2.: 使用 [47]为L=4生成的 CLEVR-SV 故事帧与真实故事帧的对比。

表 3.图2的可解释局部编辑。

帧	最小编辑路径	操作	编辑成本	语义	CL
1st	’橡胶’ →’金属质’	**R**	2	材质	0
2nd	’橡胶’ →’金属质’	**R**	2	材质	4
3rd	’橡胶’ →’金属质’	**R**	2	材质	8
4th	{’橡胶’, ’球体’} →{’金属质’, ’圆柱体’}	**R**, **R**	4	材质、形状	12

全局解释 为了全面评估我们模型的不足之处，我们测量了**故事损失**在 CLEVR-SV 的所有测试图像上的表现。因此，我们可以衡量模型捕捉某些离散语义的能力不足，无论是逐帧评估还是整体评估（表 1）。我们观察到，在较后的帧中，材质损失 有所降低，尽管我们原本预计，随着物体不断增加，问题会变得

表4。使用 [47] 在 CLEVR-SV 测试集图像上生成的可解释的全局编辑。

规则（编辑）	语义	支持度 %	前件支持度%	后件支持度%
’金属质’ →’橡胶’	材质	26.77	26.77	26.77
’橡胶’ →’金属质’	材质	22.05	22.05	22.83
’圆柱体’ →’立方体’	形状	18.11	33.07	31.50
“圆柱体” →“球体”	形状	14.96	33.07	18.90

越来越难，导致损失增大。这种预期的趋势出现在大小损失 和颜色损失中，而形状损失则没有呈现出明确的规律。较高的形状损失 表明，在所使用的生成对抗网络中有必要采用注意力机制 [28], ，以捕捉长程关系。我们还可以将尺寸 和颜色 损失的快速上升归因于故事序列中的一致性缺陷。

GSL 也能以规则的形式揭示整个测试集中的模式。我们利用 Apriori算法 [74] 提取频繁的语义组合和规则。表 4列出了最常见的 4 种语义编辑及各规则的频率（支持度）。此外，还列出了概念类别（根据方程 2得出）、前项支持度（源语义频率）和后项支持度（目标语义频率）。

我们观察到，材质 是最常见的概念误判，其中“橡胶”和“金属质”语义经常被混淆。形状 是第二常见的误判：与条件输入中的“圆柱体”相比，生成帧中出现“圆柱体”的频率更高；生成“圆柱体”时，则牺牲了“立方体”和“球体”形状。由于规则支持度并不特别高，最高值为 26.77%，我们可以放心地认为，Tsakas 等人的 SV 模型 [47] 并未严重偏向某些语义。尽管如此，我们仍发现模型有生成错误材质和形状的倾向，这一发现有助于改进模型架构。

5.2 场景生成

我们从 COCO 中选取前一万条样本，以减少使用 YOLO-v8 和 YOLOS 目标检测器提取视觉概念所需的推理时间。COCO 为每条样本提供 5 个描述性句子，这些句子彼此互为释义。因此，我们仅将 5 个句子中的第 1 个作为条件输入 c。对于场景生成，我们采用两种独立流程：基于 c 的实际生成 ，以及根据与 c 相似的说明进行图像说明配对检索 。

基于 COCO 图像说明的条件生成 对于生成实验，我们使用预训练的扩散模型，不进行任何进一步调优，如 4.1 中所述；所有模型均在相同的条件输入 c下进行测试。四个扩散模型中的每一个都需要约 15 小时，使用 2 个 T4 GPU 合成一万张图像，因此总计约需 60 小时。

检索与 COCO 相关的图像说明 为了在不花费时间和资源运行扩散模型数千次额外迭代的情况下，获取更多以与 COCO 相关的查询为条件的图像，我们使用了一个稳定扩散搜索引擎

（Lexica.art）[3]。我们采用的具体流程如下：使用 c 个 COCO 样本中前 10K 个样本的图像说明作为“查询”。对于这 10K 条图像说明，搜索引擎各返回 10 张图像；这些图像已由在线社区生成，且其输入查询与我们的图像说明最为接近。通过这种方法，我们获得了另外 100.000 张稳定扩散图像及其输入查询。随后，我们比较了网络检索图像与生成图像的结果。

目标检测 我们为检测选择 T_d =0.6 的默认阈值；将置信度为≥0.6的检测物体加入生成概念集 S。该阈值通过实验确定，以在假阳性和假阴性物体之间保持有效的权衡；事实上，由于不存在真实标注，即使定义假预测，在没有人工检查的情况下也是难以处理的。不过，我们的方法可以提供有关假检测概率的相关提示：较多的 **D** 操作可能意味着更高的假阳性率（如果 T_d 过低，则会检测到无关物体），而更多的 **I** 操作可能与更高的假阴性率相关（如果 T_d 过于严格，则无法检测到相关物体）。

指标结果 出于比较目的，我们在表 T_d =0.5、0.6、0.7中展示了5（YOLO-v8）和6（YOLOS）针对生成图像的结果，并在表7中展示了针对网络图像的结果，同时报告了两种目标检测器的目标提取结果。标为蓝色 的实例表示最低分数，这些分数更为理想；最高分数则以红色突出显示。我们展示编辑数量（**#I**、**# D**、**# R**），以及所有图像中每个 **I**、**D**、**R** 操作的总成本。平均 CSED 作为总体度量指标进行报告，不论执行频率更高的是哪种操作。

表 5.使用 YOLO-v8 对 COCO 查询生成的图像进行目标检测的指标结果。

T_d	M	# **I**	成本 **I**	# **D**	成本 **D**	# **R**	成本 **R**	平均 CSED
0.5	稳定扩散	37651	16762	1196	5655	126004	14323	35.75
	Stable Diffusion 2	36878	16067	1243	6301	129315	14839	36.32
	Protogen 基础模型	37072	16208	1233	5944	129290	14744	35.95
	protogen 5.8	38581	17715	1195	4702	117708	13411	34.66
0.6	稳定扩散	39070	18386	1157	4042	110260	12964	34.22
	Stable Diffusion 2	38678	17782	1200	4514	112499	13397	34.55
	Protogen 基础模型	38548	17794	1184	4270	114762	13427	34.35
	protogen 5.8	39766	19210	1134	3419	103579	12135	33.38
0.7	稳定扩散	40814	20391	1086	2681	93390	11337	32.96
	Stable Diffusion 2	40677	19806	1107	2938	95477	11756	33.08
	Protogen 基础模型	40397	19801	1101	2820	97314	11787	32.94
	protogen 5.8	41295	20944	1039	2308	89850	10726	32.39

关于选定的阈值 T_d，我们的初始假设得到了证实：阈值 T_d=0.7 越高，实现的**I**操作越多，这表明条件输入中的物体未被检测到；而在 T_d=0.5 时，执行的**I** 操作较少。同样，在最低阈值 T_d=0.5 下，**D**操作更多，因为虚假物体更容易被检测到。此外，

[3]https://lexica.art/

表6。使用 YOLOS 对 COCO 查询生成的生成图像进行目标检测的度量指标结果。

T_d	M	# **I**	成本 **I**	# **D**	成本 **D**	# **R**	成本 **R**	平均 CSED
0.5	稳定扩散	26302	9032	1382	44189	197623	21097	68.25
	Stable Diffusion 2	26684	8832	1403	43459	192198	21082	68.05
	Protogen 基础模型	26887	8966	1404	44406	193327	21035	68.81
	protogen 5.8	28880	10367	1373	34996	189677	19858	60.45
0.6	稳定扩散	27963	9920	1373	33891	188395	20286	60.10
	Stable Diffusion 2	28145	9662	1394	33933	182767	20322	60.36
	Protogen 基础模型	28499	9845	1394	34167	185217	20224	60.63
	protogen 5.8	30545	11330	1364	27218	179947	18963	54.13
0.7	稳定扩散	29998	10985	1357	24956	177213	19319	52.51
	Stable Diffusion 2	29831	10657	1347	25492	172860	19409	53.14
	Protogen 基础模型	29866	10790	1346	25255	175495	19350	52.98
	protogen 5.8	28880	10367	1373	34996	189677	19858	60.45

表 7。针对与 COCO 相似查询的从网络检索的 Stable Diffusion 图像的指标结果。

T_d	目标检测器	# **I**	成本 **I**	# **D**	成本 **D**	# **R**	成本 **R**	平均 CSED
0.5	YOLO-v8	186775	857448	1343	52247	1353479	224350	75.87
	YOLOS	163628	605321	1487	421525	2469635	473331	106.41
0.6	YOLO-v8	190047	891454	1317	37418	1174012	190928	73.74
	YOLOS	167576	646112	1467	308346	2303966	432851	98.06
0.7	YOLO-v8	193663	929183	1236	25388	982259	154063	71.81
	YOLOS	171778	688942	1449	214928	2115779	390304	90.56

更多**R** 操作需要较低的阈值，这也在意料之中，因为会提取更多物体并将其添加到 S 集合中。对于目标检测器，使用 YOLO-v8 得出的结果非常一致，这表明无论 T_d如何，所研究的模型都遵循相当可预测的行为。Protogen 5.8 始终得到最低的平均 CSED 分数，说明在所有阈值下，其转换成本都较低。在 T_d=0.7 阈值和 YOLOS 目标检测器（表 6）的情况下，这一观察略有变化；令人意外的是，protogen 5.8 产生的转换成本更高。比较表5 & 6可见，YOLOS 的平均 CSED 更高，**I**操作更少，**D** 操作的成本显著更高（尽管 **D** 操作的数量并未大幅增加），而且**R** 操作也更多、成本也更高。因此，我们可以有把握地认为，YOLOS 对检测更多物体相对更敏感，这可能会在检测过程中引入一些噪声。如果进一步探究评估所附带的解释，所有这些结果都会变得更容易理解。表7 中的发现也支持对生成图像的评估结果，验证了阈值假设以及 YOLOS 灵敏度较高这一结论。然而，网络检索图像似乎遗漏了查询中提到的物体，这一点可由大量的**I** 和**R** 操作得到证明。

局部解释根据针对特定生成图像执行的**I**、**D**和**R** 操作提供编辑路径。因此，我们采用图 3所示的场景。

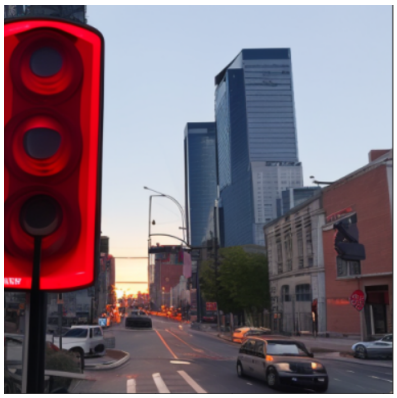


图 3： 由 Stable Diffusion 2 生成、用于提取局部解释的图像样本。

根据 YOLO-v8（默认阈值为 T_d=0.6），生成概念为 S={'car', 'car', 'traffic light', 'car', 'stop sign'}，真实概念为 T={'light', 'buildings'}。对于这项 $S \rightarrow T$ 变换，总成本最低为 59.00 的编辑操作如下：

I: { }**D:** {'汽车', '汽车', '汽车'}**R:** {'交通信号灯' →'灯', '停车标志'→'建筑物'}

使用 YOLOS 时，生成概念为 S={'car', 'traffic light', 'car', 'stop sign', 'traffic light', 'car', 'traffic light', 'traffic light', 'traffic light', 'traffic light', 'traffic light', 'traffic light', 'car', 'traffic light', 'traffic light', 'traffic light', 'traffic light', 'traffic light', 'car', 'traffic light', 'traffic light', 'traffic light', 'traffic light', 'car', 'car', 'traffic light', 'traffic light'}，真实概念为 T={'light', 'buildings'}。通过目视检查图像可以清楚地看出，YOLOS 高估了实际存在的物体数量，导致生成概念集 S 中出现噪声。尽管如此，我们的评估策略成功捕捉到了这种高估，并建议删除多个概念。具体而言，我们得到以下总成本为 104.04 的变换：

I: { }
D: {'汽车', '交通信号灯', '汽车', '交通信号灯', '汽车', '交通信号灯', '交通信号灯', '交通信号灯', ' 交通信号灯 ', ' 交通信号灯 ', ' 交通信号灯 ', ' 汽车 ', ' 交通信号灯 ', ' 交通信号灯 ', ' 交通信号灯 ', '交通信号灯', '汽车', '交通信号灯', '交通信号灯', '交通信号灯', '交通信号灯', '汽车', '汽车', '交通信号灯', '交通信号灯'}**R:** {'停车标志'→'灯', '交通信号灯'→'建筑物'}

所有图像的全局解释分别见表8 中的**插入**、**删除**编辑结果和表9 中的**替换**编辑结果。结果仅涉及 YOLO-v8 提取的概念，因为 YOLOS 会产生数量过多的检测实例。此处展示排名前三的结果，即频率最高的 3 个插入、删除和替换。插入和**删除**分别表示插入或删除的概念，而频率**插入**、**删除**表示某个特定概念在所有图像中被插入或删除的次数。**插入**、**删除**支持度表示某种特定编辑在所有**插入**、**删除**编辑中出现的频率。对于**替换**，支持度表示某种变换规则在所有生成的规则中出现的频率。

表 8。YOLO-v8 提取概念的全局解释（**I**和**D** 编辑）。

T_d	M	**I**	频率 **I**	**I** 支持度	**D**	频率 **D**	**D** 支持度
0.5	稳定扩散	街道	264	1.57%	人	2075	36.69%
		桌子	250	1.49%	绵羊	363	6.42%
		网球	247	1.47%	car	252	4.46%
	稳定扩散 2	网球	253	1.57%	人	2177	34.55%
		街道	242	1.51%	绵羊	466	7.40%
		桌子	237	1.48%	car	313	4.97%
	Protogen 基础模型	网球	247	1.52%	人	2281	38.37%
		街道	244	1.51%	绵羊	317	5.33%
		桌子	229	1.41%	car	311	5.23%
	protogen 5.8	桌子	270	1.52%	人	1564	33.26%
		网球	265	1.50%	car	261	5.55%
		街道	241	1.36%	雨伞	251	5.34%
0.6	稳定扩散	街道	290	1.58%	人	1572	38.89%
		桌子	281	1.53%	绵羊	311	7.69%
		网球	259	1.41%	car	158	3.91%
	稳定扩散 2	桌子	274	1.54%	人	1656	36.69%
		街道	269	1.51%	绵羊	376	8.33%
		网球	264	1.48%	car	203	4.50%
	Protogen 基础模型	街道	268	1.51%	人	1717	40.21%
		桌子	261	1.47%	绵羊	254	5.95%
		网球	255	1.43%	car	197	4.61%
	protogen 5.8	桌子	303	1.58%	人	1220	35.68%
		网球	278	1.45%	绵羊	198	5.79%
		街道	274	1.43%	雨伞	176	5.15%
0.7	稳定扩散	桌子	322	1.58%	人	1075	40.10%
		街道	316	1.55%	绵羊	254	9.47%
		网球	268	1.31%	甜甜圈	122	4.55%
	稳定扩散 2	桌子	313	1.58%	人	1134	38.60%
		街道	301	1.52%	绵羊	291	9.90%
		网球	267	1.35%	甜甜圈	111	3.78%
	Protogen 基础模型	街道	300	1.52%	人	1189	42.16%
		桌子	289	1.46%	绵羊	188	6.67%
		网球	262	1.32%	雨伞	143	5.07%
	protogen 5.8	桌子	330	1.58%	人	884	38.30%
		街道	299	1.43%	绵羊	152	6.59%
		网球	287	1.37%	雨伞	130	5.63%

我们可以观察到模型之间明显的一致性；**插入**编辑包含“街道”“网球”和“桌子”等概念。看来，所选的 M 无法有效生成**插入**概念，或者生成的概念视觉质量较低，因此无法通过 T_d=0进行检测。5、0.6、0.7。**删除**编辑主要包含“人”“绵羊”“汽车”“雨伞”“甜甜圈”等概念，这表明模型在生成这些概念类别的虚假实例时存在一定偏差。最后，**替换**编辑

指将 'person' 转换为 'people'、'man' 或 'woman'。由于 'person' 是一个包含两种性别的 YOLO 类别，因此进行此类转换在某种程度上是可以预期的。

表 9。YOLO-v8 提取概念的全局解释（**R** 编辑）。

T_d	M	**R**	Freq **R**	**R** 上标-port	M	**R**	Freq **R**	**R** 上标-port
0.5	稳定扩散	人 →男性	1090	7.61%	稳定扩散 2	人 → 男性	1115	7.51%
		人 →人们	520	3.63%		人 →人们	551	3.71%
		人 →女性	499	3.48%		人 →女性	511	3.44%
	Protogen base	人 →男性	1101	7.47%	Protogen 5.8	人 → 男性	1061	7.91%
		人 →人们	507	3.44%		人 →女性	476	3.55%
		人 →女性	500	3.39%		人 →人们	441	3.29%
0.6	稳定扩散	人 →男性	1065	8.22%	稳定扩散 2	人 → 男性	1087	8.11%
		人 →人们	503	3.88%		人 →人们	536	4.00%
		人 →女性	481	3.71%		人 →女性	482	3.60%
	Protogen base	人 →男性	1080	8.04%	Protogen 5.8	人 → 男性	1035	8.53%
		人 →人们	494	3.68%		人 →女性	449	3.70%
		人 →女性	485	3.61%		人 →人们	431	3.55%
0.7	稳定扩散	人 →男性	1022	9.01%	稳定扩散 2	人 → 男性	1033	8.79%
		人 →人们	473	4.17%		人 →人们	508	4.32%
		人 →女性	458	4.04%		人 →女性	441	3.75%
	Protogen base	人 →男性	1054	8.94%	Protogen 5.8	人 → 男性	989	9.22%
		人 →女性	461	3.91%		人 →女性	419	3.91%
		人 →人们	446	3.78%		人 →人们	408	3.80%

6. 结论

生成式评估中的概念性方法是一个尚未得到充分探索的领域，可为模型质量和结果可解释性提供一些新颖见解。在我们的工作中，我们提出了一种知识驱动的可解释评估框架，用于指出应添加、移除或替换哪些概念，以使生成图像更接近其条件输入。在故事可视化和场景生成等具有竞争力的任务上的结果展示了这种方法的优势，揭示了模型无法生成的概念，或模型倾向于生成过多特定概念类别的偏差。作为未来工作，我们计划将该方法扩展到其他模型和任务，并纳入其他知识来源，以考察生成的编辑路径在概念上与当前路径的偏离情况。

致谢

本研究工作获得希腊研究与创新基金会（HFRI）第三轮博士奖学金计划（奖学金编号5537）的资助。

参考文献

[1] J. Ho, A. Jain, P. Abbeel, 去噪扩散概率模型，2020年。网址：https://arxiv.org/abs/2006.11239。doi:10.48550/ARXIV.2006.11239.[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, 基于潜在扩散模型的高分辨率图像合成，发表于：IEEE/CVF计算机视觉与模式识别会议（计算机视觉与模式识别会议），2022年，第10684–10695页。[3] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, 零样本文本到图像生成，2021年。 arXiv:2102.12092.[4] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, 基于 CLIP 潜变量的分层文本条件图像生成，2022年。网址：https://arxiv.org/abs/2204.06125。doi:10.48550/ARXIV.2204.06125.[5] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, M. Norouzi, 具备深度语言理解能力的照片级真实感文本生成图像扩散模型，2022年。 arXiv:2205.11487.[6] A. Blattmann, R. Rombach, K. Oktay, J. Müller, B. Ommer, 半参数神经图像合成，2022年。网址：https://arxiv.org/abs/2204.11824。doi:10.48550/ARXIV.2204.11824。[7] J. Kim, M. Lee, 类别连续条件生成式神经辐射场，2023年。网址：https://arxiv.org/abs/2301.00950。doi:10.48550/ARXIV.2301.00950.[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, 生成对抗网络，发表于：Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Q. Weinberger（编），神经信息处理系统进展，第27卷，Curran Associates, Inc.，2014年。网址：https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf。[9] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, X. Chen, 改进的 GAN 训练技术，发表于：D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett（编），神经信息处理系统进展，第29卷，Curran Associates, Inc.，2016年。网址：https://proceedings.neurips.cc/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf。[10] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, 通过双时间尺度更新规则训练的生成对抗网络收敛到局部纳什均衡，发表于：I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett（编），神经信息处理系统进展，第30卷，Curran Associates, Inc.，2017年。网址：https://proceedings.neurips.cc/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf。[11] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, 深层特征作为感知度量的惊人效力，发表于：计算机视觉与模式识别会议，2018年。[12] G. Parmar, R. Zhang, J.-Y. Zhu, 混叠缩放及生成对抗网络评估中的意外细微之处，发表于：计算机视觉与模式识别会议，2022年。[13] A. Borji, 生成对抗网络评估度量的优缺点：最新进展，《计算机视觉与图像理解》215（2022年）103329。网址：https://www.sciencedirect.com/science/article/pii/S1077314221001685。doi:https://doi.org/10.1016/j.cviu.2021.103329。[14] M. Soloveitchik, T. Diskin, E. Morin, A. Wiesel, 条件 Fréchet Inception距离，

2021年。URL：https://arxiv.org/abs/2103.11521。doi:10.48550/ARXIV.2103.11521.[15] Y. Benny、T. Galanti、S. Benaim、L. Wolf，《条件图像生成的评估指标》，《国际计算机视觉杂志》129（2021年）1712–1731。URL：https://doi.org/10.1007%2Fs11263-020-01424-w。doi:10.1007/s11263-020-01424-w.[16] K. Abhishek、D. Kamath，《计算机视觉中的基于归因的可解释人工智能方法：综述》，2022年。URL：https://arxiv.org/abs/2211.14736。doi:10.48550/ARXIV.2211.14736.[17] V. Buhrmester、D. MÃ¼nch、M. Arens，《计算机视觉黑箱深度神经网络解释器分析：综述》，《机器学习与知识提取》3（2021年）966–989。URL：https://www.mdpi.com/2504-4990/3/4/48。doi:10.3390/make3040048.[18] V. Nagisetty、L. Graves、J. Scott、V. Ganesh，《XAI-GAN：通过可解释人工智能系统增强生成对抗网络》，2020年。URL：https://arxiv.org/abs/2002.10438。doi:10.48550/ARXIV.2002.10438.[19] A. Genovese、V. Piuri、F. Scotti，《利用生成对抗网络实现可解释的人脸老化》，载于：2019年IEEE国际图像处理会议（ICIP），2019年，第3806–3810页。doi:10.1109/ICIP.2019.8803616.[20] J. Kim、H. Park，《用于解决过拟合问题的可解释人工智能模型有限判别器生成对抗网络》，《ICT Express》（2022年）。URL：https://www.sciencedirect.com/science/article/pii/ S240595952100179X。doi:https://doi.org/10.1016/j.icte.2021.12.014.[21] C. Leiter、P. Lertvittayakumjorn、M. Fomicheva、W. Zhao、Y. Gao、S. Eger，《迈向自然语言生成的可解释评估指标》，2022年。URL：https://arxiv.org/abs/ 2203.11131。doi:10.48550/ARXIV.2203.11131.[22] J. Opitz、A. Frank，《迈向一种用于可解释评估AMR文本生成的可分解指标》，载于：第16届欧洲计算语言学协会分会会议论文集：主会场，计算语言学协会，线上，2021年，第1504–1518页。URL：https://aclanthology.org/2021.eacl-main.129。doi:10.18653/v1/2021.eacl-main.129.[23] M. Lymperaiou、G. Manoliadis、O. Menis Mastromichalakis、E. G. Dervakos、G. Stamou，《迈向基于视觉概念语义相似度的语言模型可解释评估》，载于：第29届国际计算语言学会议论文集，国际计算语言学委员会，韩国庆州，2022年，第3639–3658页。URL：https://aclanthology.org/2022.coling-1.321。[24] G. Filandrianos、K. Thomas、E. Dervakos、G. Stamou，《作为反事实解释的概念编辑》，载于：AAAI 2022年混合智能机器学习与知识工程春季研讨会论文集（AAAI-MAKE 2022），斯坦福大学，美国加利福尼亚州帕洛阿尔托，2022年。[25] M. Mirza、S. Osindero，《条件生成对抗网络》，CoRR abs/1411.1784（2014年）。URL：http://arxiv.org/abs/1411.1784。arXiv:1411.1784.[26] A. Odena、C. Olah、J. Shlens，《利用辅助分类器生成对抗网络进行条件图像合成》，2017年。 arXiv:1610.09585.[27] T. Miyato、M. Koyama，《采用投影判别器的条件生成对抗网络》，2018年。 arXiv:1802.05637.[28] H. Zhang、I. Goodfellow、D. Metaxas、A. Odena，《自注意力生成对抗网络》，2019年。arXiv:1805.08318.[29] T. Miyato、T. Kataoka、M. Koyama、Y. Yoshida，《生成对抗网络的谱归一化》，2018年。 arXiv:1802.05957。

[30] M. Heusel、H. Ramsauer、T. Unterthiner、B. Nessler、S. Hochreiter，《通过双时间尺度更新规则训练的生成对抗网络收敛到局部纳什均衡》，2018年。arXiv:1706.08500.[31] R. Rombach、A. Blattmann、D. Lorenz、P. Esser、B. Ommer，《基于潜在扩散模型的高分辨率图像合成》，2021年。arXiv:2112.10752.[32] A. Radford、J. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark、G. Krueger、I. Sutskever，《从自然语言监督中学习可迁移视觉模型》，2021年。[33] C. Saharia、W. Chan、S. Saxena、L. Li、J. Whang、E. Denton、S. K. S. Ghasemipour、B. K. Ayan、S. S. Mahdavi、R. G. Lopes、T. Salimans、J. Ho、D. J. Fleet、M. Norouzi，《具备深度语言理解能力的逼真文本到图像扩散模型》，2022年。URL：https://arxiv. org/abs/2205.11487。doi:10.48550/ARXIV.2205.11487.[34] C. Raffel、N. Shazeer、A. Roberts、K. Lee、S. Narang、M. Matena、Y. Zhou、W. Li、P. J. Liu，《利用统一的文本到文本转换器探索迁移学习的局限》，《机器学习研究杂志》21（2020年）1–67。URL：http://jmlr.org/papers/v21/20-074.html。[35] N. Ruiz、Y. Li、V. Jampani、Y. Pritch、M. Rubinstein、K. Aberman，《DreamBooth：微调文本到图像扩散模型以实现主体驱动生成》，2022年。URL：https://arxiv.org/ abs/2208.12242。doi:10.48550/ARXIV.2208.12242.[36] S. E. Reed、Z. Akata、X. Yan、L. Logeswaran、B. Schiele、H. Lee，《生成对抗式文本到图像合成》，CoRR abs/1605.05396（2016年）。URL：http://arxiv.org/abs/1605.05396。arXiv:1605.05396.[37] S. E. Reed、Z. Akata、S. Mohan、S. Tenka、B. Schiele、H. Lee，《学习绘制什么以及在哪里绘制》，CoRR abs/1610.02454（2016年）。URL：http://arxiv.org/abs/1610.02454。arXiv:1610.02454.[38] H. Zhang、T. Xu、H. Li，《StackGAN：利用堆叠生成对抗网络进行文本到逼真图像合成》，2017年，第5908–5916页。doi:10.1109/ICCV.2017.629.[39] H. Zhang、T. Xu、H. Li、S. Zhang、X. Wang、X. Huang、D. Metaxas，《StackGAN++：利用堆叠生成对抗网络进行逼真图像合成》，2018年。arXiv:1710.10916.[40] T. Xu、P. Zhang、Q. Huang、H. Zhang、Z. Gan、X. Huang、X. He，《AttnGAN：利用注意力生成对抗网络进行细粒度文本到图像生成》，载于：计算机视觉与模式识别会议2018，2018年。[41] H. Tan、X. Liu、X. Li、Y. Zhang、B. Yin，《用于文本到图像合成的语义增强生成对抗网络》，载于：2019年IEEE/CVF国际计算机视觉大会（ICCV），2019年，第10500–10509页。doi:10.1109/ICCV.2019.01060.[42] M. Zhu、P. Pan、W. Chen、Y. Yang，《DM-GAN：用于文本到图像合成的动态记忆生成对抗网络》，CoRR abs/1904.01310（2019年）。URL：http://arxiv.org/ abs/1904.01310。arXiv:1904.01310.[43] Y. Li、Z. Gan、Y. Shen、J. Liu、Y. Cheng、Y. Wu、L. Carin、D. E. Carlson、J. Gao，《StoryGAN：用于故事可视化的序列条件生成对抗网络》，2019年IEEE/CVF计算机视觉与模式识别会议（CVPR）（2019年）6322–6331。[44] C. Li、L. Kong、Z. Zhou，《用于序列图像可视化的改进型StoryGAN》，《视觉通信与图像表示杂志》73（2020年）102956。URL：https: //www.sciencedirect.com/science/article/pii/S1047320320301826。doi:https://doi.org/10.1016/j.jvcir.2020.102956.[45] A. Maharana、D. Hannan、M. Bansal，《改进视觉故事的生成与评估》

通过语义一致性，ArXiv abs/2105.10026（2021年）。[46] A. Maharana、M. Bansal，将视觉空间、语言和常识结构融入故事可视化，ArXiv abs/2110.10834（2021年）。[47] N. Tsakas、M. Lymperaiou、G. Filandrianos、G. Stamou，用于故事可视化的公正 Transformer，2023年。网址：https://arxiv.org/abs/2301.03563。doi:10.48550/ARXIV.2301.03563。[48] A. Horé、D. Ziou，图像质量指标：峰值信噪比与 SSIM，载于：2010年第20届国际模式识别会议，2010年，第2366–2369页。doi:10.1109/ICPR.2010.579。[49] D. Bau、J.-Y. Zhu、J. Wulff、W. Peebles、H. Strobelt、B. Zhou、A. Torralba，看见生成对抗网络无法生成的内容，2019年。网址：https://arxiv.org/abs/1910.11626。doi:10.48550/ARXIV.1910.11626。[50] Y. Shen、J. Gu、X. Tang、B. Zhou，解读生成对抗网络的潜在空间以进行语义人脸编辑，2019年 IEEE/CVF计算机视觉与模式识别会议（CVPR）（2019年）9240–9249。[51] Y. Shen、C. Yang、X. Tang、B. Zhou，InterfaceGAN：解读生成对抗网络学得的解耦人脸表示，《IEEE模式分析与机器智能汇刊》44（2020年）2004–2018。[52] L. C hai、J. Wulff、P. Isola，利用潜在空间回归分析并挖掘生成对抗网络中的组合性，ArXiv abs/2103.10426（2021年）。[53] J. Wen、F. Benitez-Quiroz、Q. Feng、A. M. Martinez，粗粝中的璞玉：通过遍历生成对抗网络潜在空间提升图像真实感，ArXiv abs/2104.05518（2021年）。[54] M. Nauta、J. Trienes、S. Pathak、E. Nguyen、M. Peters、Y. Schmitt、J. Schlötterer、M. van Keulen、C. Seifert，从轶事证据到定量评估方法：关于可解释人工智能评估的系统综述，2022年。网址：https://arxiv.org/abs/2201.08164。doi:10.48550/ARXIV.2201.08164。[55] W. Zhao、S. Oyama、M. Kurihara，生成自然的反事实视觉解释，载于：C. Bessiere（编），第二十九届国际人工智能联合会议论文集，IJCAI-20，国际人工智能联合会议组织，2020年，第5204–5205页。网址：https://doi.org/10.24963/ijcai.2020/742。doi:10.24963/ijcai.2020/742，博士生研讨会。[56] R. Poyiadzi、K. Sokol、R. Santos-Rodríguez、T. D. Bie、P. A. Flach，Face：可行且可操作的反事实解释，《AAAI/ACM人工智能、伦理与社会会议论文集》（2020年）。[57] Y. Goyal、Z. Wu、J. Ernst、D. Batra、D. Parikh、S. Lee，反事实视觉解释，载于：K. Chaudhuri、R. Salakhutdinov（编），第36届国际机器学习会议论文集，机器学习研究论文集第97卷，PMLR，2019年，第2376–2384页。网址：https://proceedings.mlr.press/v97/goyal19a.html。[58] C. Fellbaum，WordNet：电子词汇数据库（1998年）。[59] K. Browne、B. Swift，语义与解释：为何反事实解释会在深度神经网络中产生对抗样本，ArXiv abs/2012.10076（2020年）。[60] J. Johnson、B. Hariharan、L. van der Maaten、L. Fei-Fei、C. L. Zitnick、R. B. Girshick，CLEVR：用于组合语言和基础视觉推理的诊断数据集，2017年 IEEE计算机视觉与模式识别会议（CVPR）（2017年）1988–1997。[61] T.-Y. Lin、M. Maire、S. Belongie、L. Bourdev、R. Girshick、J. Hays、P. Perona、D. Ramanan、C. L.

Zitnick、P. Dollár，Microsoft COCO：上下文中的常见物体，2015年。arXiv:1405.0312。[62]Stable Diffusion v1.4，Huggingface。网址：https://huggingface.co/CompVis/ stable-diffusion-v1-4。[63]Stable Diffusion 2 Base，Huggingface。网址：https://huggingface.co/stabilityai/ stable-diffusion-2-base。[64] Protogen x3.4，Huggingface。网址：https://huggingface.co/darkstorm2150/Protogen_x3。4_官方发布。_[65] Protogen x5.8，Huggingface。网址：https://huggingface.co/darkstorm2150/Protogen_x5。8_官方发布。_[66] T. Park、M.-Y. Liu、T.-C. Wang、J.-Y. Zhu，采用空间自适应归一化的语义图像合成，2019年。arXiv:1903.07291。[67] A. Zareian、Z. Wang、H. You、S.-F. Chang，学习视觉常识以实现稳健的场景图生成，2020年。arXiv:2006.09623。[68]K. Ma、B. Zhao、L. Sigal，基于布局的属性引导图像生成，2020年。arXiv:2008.11932。[69] Y. Li、Y. Cheng、Z. Gan、L. Yu、L. Wang、J. Liu，BachGAN：根据显著物体布局进行高分辨率图像合成，2020年。arXiv:2003.11690。[70] W. Sun、T. Wu，学习用于可控图像合成的布局与风格可重构生成对抗网络，2021年。arXiv:2003.11571。[71] G. Jocher、A. Chaurasia、J. Qiu，Ultralytics 的 YOLO，2023年。网址：https://github.com/ultralytics/ ultralytics。[72] Y. Fang、B. Liao、X. Wang、J. Fang、J. Qi、R. Wu、J. Niu、W. Liu，你只需看一个序列：通过目标检测重新思考视觉领域中的 Transformer，CoRR abs/2106.00666（2021年）。网址：https://arxiv.org/abs/2106.00666。arXiv:2106.00666。[73] 工业级自然语言处理，spaCy。网址：https://spacy.io/。[74] R. Agrawal、R. Srikant，用于挖掘关联规则的快速算法，载于：第20届国际VLDB会议论文集，1994年，第487–499页。

A. 在线资源

以下模型用于图像生成：

- Huggingface 文本到图像生成模型。
- Stable Diffusion v1.4
- Stable Diffusion 2
- Protogen x3.4
- Protogen x5.8

以下网页用于检索已生成的稳定扩散图像：

- https://lexica.art/

以下模型用于目标检测：

- Huggingface YOLOS-基础
- Huggingface YOLO-v8

B. SV 局部详细示例

本节将针对图2所示序列，详细分析 CSED 在故事可视化中的逐帧局部属性。根据 [47], 中的分析，这属于中等难度示例，因为在第 4 帧中，蓝色圆柱体与蓝色立方体发生了重叠。我们将比较该序列的真实语义（对应真实标注帧，见图 2a）与生成语义（对应生成帧，见图 2b）。因此，我们得到以下结果：

帧 k=1
真实语义: {[小、棕色、**金属质**、球体]}
生成语义: {[小、棕色、**橡胶**、球体]}
两个序列的区别在于第三个位置上高亮显示的语义：真实标注中的语义是 '金属质'，而生成结果中的语义是 '橡胶'，因此 $CSED_{k=1}$ 提出将 '橡胶' →'金属质' 的替换操作，编辑成本为 = 2 = $CSED_{k=1}$ ，以使生成序列与真实标注序列完全一致。此外，由于该变换涉及材质语义，因此材质损失计数器中会增加一次生成失败，从而为所有测试集帧中的语义合成失败提供全局解释。

对于第 1 帧，生成序列的一致性损失（CL）为 $CL_{k=1}$ = 0，因为总共有 $|\mathcal{C}|$=4 个语义（材质、尺寸、形状、颜色），且序列中 k=1 位置上的 1 个物体包含 T =4 个语义：
$CL_{k=1} = p_{k=1} = |T| - |\mathcal{C}| \cdot k$=4-4=0。

帧 k=2
真实语义：{[小、棕色、**金属质**、球体], [小、棕色、金属质、球体]}生成语义：{[小、棕色、**橡胶**、球体], [小、棕色、金属质、球体]}
第 3 个位置的语义存在差异，且以粗体突出显示：真实标注语义为 '金属质'，而生成语义为 '橡胶'，因此 $CSED_{k=2}$ 提出了将 '橡胶' →替换为 '金属质' 的操作，编辑成本为 = 2 = $CSED_{k=2}$。此外，由于该变换涉及材质语义，因此材质损失计数器中还要增加一次生成失败。

同时，仅通过添加一个包含$|\mathcal{C}|$ =4 语义的物体，CL 就不可避免地会增加。因此，对于 CLEVR-SV，添加一个物体时 CL 的最小增量可以是 4。除此之外，如果 $k = 1$ 与 $k = 2$生成帧之间存在更多不一致， $CL_{k=2}$ 就会增加。因此，我们比较 $k = 1$ 生成序列$T = S_{k-1}$={[小、棕色、橡胶、球体]} 与 $k = 2$ 生成序列 $S = S_k$={[小、棕色、橡胶、球体], [**小、棕色、金属质、球体**]}，其中未发现额外差异。应用方程5对于 k=2 我们得到：

$$CL_{k=2} = p_{k=1} + D(S_{k=2},\ T_{k=2}) = 0 + \textbf{我}\{small, brown, metallic, sphere\} = 0+4 = 4$$

帧 k=3
真实语义：{[小，棕色，**金属质**，球体], [小，棕色，金属质，球体],[大，蓝色，橡胶，立方体]}

生成语义：{[小，棕色，**橡胶**，球体], [小，棕色，金属质，球体], [大，蓝色，橡胶，立方体]}第三个位置的语义差异仍然存在，因此 $CSED_{k=3}$提出了将“橡胶” →替换为“金属质”的替换操作，编辑成本为 = 2 = $CSED_{k=3}$。此外，由于该转换涉及材质语义，因此向材质损失计数器中增加了一次生成失败。

一致性损失将考虑 $T = 2$ 生成序列 [小、棕色、橡胶、球体，小、棕色、金属质、球体] 与 $k = 3$生成序列 {[小、棕色、橡胶、球体], [小、棕色、金属质、球体], [**大、蓝色、橡胶、立方体**]}之间的比较。两者仅相差在第三帧添加了大、蓝色、橡胶、立方体，因此得到：

$CL_{k=3} = p_{k=1} + D\left(S_{k=2},\ T_{k=2}\right) + D\left(S_{k=3},\ T_{k=3}\right) = 0 + 4 +$ **我** $\{large, blue, rubber, cube\}$= 0+4+4 = 8

帧 k=4

真实语义：{[小、棕色、**金属质**、球体], [小、棕色、金属质、球体],[大、蓝色、橡胶、立方体], [大、蓝色、金属质、**圆柱体**]}

生成语义：{[小、棕色、**橡胶**、球体], [小、棕色、金属质、球体], [大、蓝色、橡胶、立方体], [大、蓝色、金属质、**球体**]}

除了第 $CSED_{k=4}$ 三处语义不同（ $CSED_{k=4}$ 提出了将“橡胶”替换为“金属质” →的替换操作，编辑成本为 = 2）之外，最后一处语义也有一处差异，表示将“球体”转换为“圆柱体” →，编辑成本为= 2。将这两种转换合并后，我们得到$k = 4$的总转换：{’rubber’, ’ sphere’} →{’metallic’, ’cylinder’}，编辑成本为 = 4 = $CSED_{k=4}$。材质损失 和形状损失 的计数器各增加 1。

对于一致性损失（CL），对应于 $k = 3$ 生成序列{[小、棕色、橡胶、球体], [小、棕色、金属质、球体], [大、蓝色、橡胶、立方体]}和 $k = 4$ 生成序列{[小、棕色、橡胶、球体], [小、棕色、金属质、球体], [大、蓝色、橡胶、立方体], [**大、蓝色、金属质、球体**]}，两者仅相差新增的大、蓝色、金属质、球体项。因此 $CL_{T=4}$= 4。

$CL_{k=4} = p_{k=1} + D(S_{k=2},\ T_{k=2}) + D(S_{k=3},\ T_{k=3}) + D(S_{k=4},\ T_{k=4}) = 0 + 4 + 4 +$**I** $\{large, blue, metallic, sphere\}$ = 0+4+4+4 = 12 通过汇总结果，故事损失（*SL*）作为逐帧 CSED成本之和将为：

$$SL = 2 + 2 + 2 + 4 = 10$$

并根据公式 $L = 4$7 对所有帧的 SL 取平均值：

$$Average\ SL = \frac{1}{k} SL = 10/4 = 2.5$$

为保持一致性，我们遵循公式8：

$$Average\ CL = \frac{p_{k=1}}{k} + \frac{1}{k} \sum_{k=1}^{k=L} [CL_{k>1} \neq |\mathcal{C}| \cdot (k-1)] = 0 + 0 = 0$$

图中生成故事 2 **完全一致**，因为其 平均 *CL* 等于 0，这是理想情况。因此，生成的

序列中没有插入、删除或更改任何语义。不过，有趣的是，CL 无法捕捉在第 4 帧插入新项所产生的**忠实性** 错误：虽然真实标注项是一个大、蓝色、金属质的圆柱体，但生成序列插入的是一个大、蓝色、金属质的球体，而 CL 并不会因最后一个位置的语义差异而进一步惩罚。相反，SL 会对此错误进行惩罚。当然，在另一个示例中也可能出现相反的情况，即 CL 能够指出错误，而 SL 无法捕捉到该错误。这一观察表明，两个指标都很重要：SL 侧重于真实标注与生成故事之间的忠实性，而 CL 侧重于连续生成帧之间的一致性。模型越好，在全局层面上，这两个指标都应越低。