/TemplateVersion (IJCAI.2025.0)

通过步骤选择在基于去噪模型中实现文本与图像的对齐

Paul Grimal, Hervé Le Borgne & Olivier Ferret
Université Paris-Saclay,
CEA, List,
F-91120, Palaiseau, France
{ paul.grimal, herve.le-borgne, olivier.ferret } @cea.fr

Abstract

视觉生成 AI 模型经常遇到与文本图像对齐和推理能力限制相关的挑战。本文提出了一种新颖的方法，用于在关键去噪步骤中选择性地增强信号，根据输入语义优化图像生成。我们的方法解决了早期信号修改的不足之处，证明在后期阶段进行调整可以产生更好的结果。我们进行了广泛的实验，以验证我们的方法在扩散和流匹配模型上生成语义对齐图像的有效性，达到了最新的性能。我们的结果强调了明智选择采样阶段以提高性能和整体图像对齐的重要性。这项工作是最初在 OpenReview 上发布的预印本的更新和扩展版本。代码可在 https://github.com/grimalPaul/gsn-factory 获取。

1 介绍

视觉生成式人工智能模型通常依赖于去噪过程，如扩散模型 [Ho *et al.*, 2020] 或流动匹配 [Lipman *et al.*, 2023]。它们可以通过文本提示进行条件化，以引导推断，生成视觉上令人愉悦的图像 [Rombach *et al.*, 2022; Podell *et al.*, 2023; Ramesh *et al.*, 2022; Saharia *et al.*, 2022]。尽管这些模型展示了令人印象深刻的语义和组合能力，但即使是最好的模型仍然存在文本-图像对齐和推理的局限性（例如 空间，计数）。一些工作通过改进训练数据集中噪声较大的字幕 [Chen *et al.*, 2023; Chen *et al.*, 2024a; Segalis *et al.*, 2023] 或改进架构 [Peebles and Xie, 2022] 来解决这些问题，而另一些工作则采用基于注意力的生成语义护理（GSN）方法 [Chefer *et al.*, 2023; Rassin *et al.*, 2023; Guo *et al.*, 2024]，该方法通过在推理时进行校正或添加条件来更好地引导生成，避免重新训练整个模型。

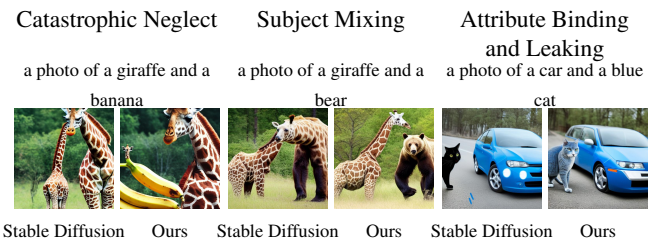

Figure 1: 由 Stable Diffusion 生成的样本与我们的样本对比。

早期的研究已经识别出几种文本-图像对齐问题 [Ramesh *et al.*, 2022; Saharia *et al.*, 2022; Chefer *et al.*, 2023; Feng *et al.*, 2023]。这些问题包括灾难性忽视，即提示中的一个或多个元素未能被生成；主体混合，即不同元素被不当组合；属性绑定，即属性（例如 颜色）被错误地分配给实体；以及属性泄漏，即属性正确地绑定到指定元素，但被错误地应用于场景中额外的、非预期的元素（Figure 1）。

为了改进生成，出现了无训练方法 [Chefer *et al.*, 2023; Rassin *et al.*, 2023; Li *et al.*, 2023b; Guo *et al.*, 2024; Agarwal *et al.*, 2023]。这些方法利用模型中扩散特征的文本-图像关系来优化扩散模型正在去噪的潜在图像以进行调整。然而，这些方法需要测试并仔细选择多个敏感的超参数（如选择各种扩散步来执行优化或设置不同的损失阈值以达到每个扩散步），这可能导致优化过程中的潜在失败。此外，尽管通常沿着扩散路径采用多个细化步骤，但对于其重复使用的必要性以及其位置的原因，主要是通过实验结果确定的，没有明确的解释。我们认为，更仔细地检查细化步骤的位置不仅可以改善性能，还可以更好地理解这些步骤的最佳位置。为减轻欠优化或过度优化的风险，InitNO [Guo *et al.*, 2024] 在第一次扩散步骤中仅优化多个初始潜在图像，此时潜在图像是纯高斯噪声。然而，扩散模型的逆向过程在图像生成过程中逐步重建信号，使得早期阶段的优化因信号较弱而效果不佳。随着信号在后来的扩散步骤中变得更强，提供了更多有用的信息来细化潜在图像。对信号退化动态的更深入理解可用于提高生成能力。在这项工作中，我们研究了基于语义内容选择最佳步骤以增强信号的影响，展示了仔细选择这些步骤带来了文本与图像对齐的显著改进。

本文的主要贡献包括：（1）将生成语义护理扩展到最新的流匹配模型 Stable Diffusion 3 的架构，以及（2）一种在扩散或流匹配过程中选择性增强关键信号的方法，以优化基于输入语义的图像生成。我们展示了早期阶段信号修改效果较差，并表明后期调整可以获得更好的结果。我们通过大量实验验证了我们的方法，展示了其在生成语义一致的图像和达到最新成果中的有效性，同时还研究了细化步骤的位置。

2 相关工作

新的控制 Li *et al.* [2023a] 和 Mou *et al.* [2023] 引入了可训练模块，以便向冻结模型添加新的条件。类似地，Zhang *et al.* [2023] 引入了可训练的模型副本，可以在各种控制输入（例如图画、边界框或深度图）的条件下使用。最近的研究重点是通过处理噪声潜在图像来调节模型。SDEdit [Meng *et al.*, 2022] 向图像添加不同级别的噪声，在原始图像的保真度和创造性变化之间取得平衡。

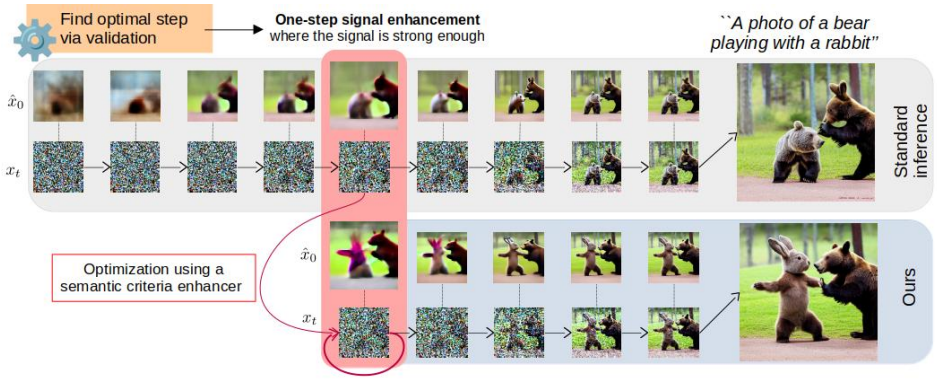

Figure 2: 扩散过程在一个关键步骤（通过验证子集确定）被暂停，以增强潜在图像中的信号。通过在这个关键点放大信号，我们确保模型能够正确构建图像的主要组成部分，从而获得更准确的最终结果。

Sun *et al.* [2024] 通过在背景上放置物体、添加噪声，然后在生成过程中去噪来创建伪引导图像，从而保持物体的放置。Choi *et al.* [2021] 在扩散过程中注入降采样的引导图像，以创建引导图像的变化。

FreeDoM [Yu *et al.*, 2023] 在各种采样步骤中对潜在图像进行多次更新，并依赖外部模型进行指导。与之相反，我们的方法使用模型的固有知识进行单步优化。这确保了生成过程中的更好对齐和连续性。外部模型在优化模糊图像时，可能无法像去噪模型那样理解信号，导致结果不够精确。例如，三个不同实体的模糊图像使得它们难以区分，而在模型的语义空间中，对应每个实体的信号部分可能更为明确。

高斯噪声 GSN 由 Attend & Excite [Chefer *et al.*, 2023] 引入，旨在在推理过程中优化潜在图像，以更好地考虑语义信息，而无需重新训练模型。在步骤 t 中，通过对模型在输入 x_t 提取的特征应用关于损失 $\mathcal{L}$ 的梯度下降步骤来修改潜在图像 $x_t : x_{t'} \leftarrow x_t - \alpha_t \cdot \nabla_{x_t}\mathcal{L}$（$\alpha_t$ 学习率）。因此，它调整潜在图像以实现由损失函数概念化的目标。Attend & Excite 考虑交叉注意特征，这建立了图像和文本特征之间的链接，以确保模型能够充分生成提示中的主体。在这种方法的基础上，Syngen [Rassin *et al.*, 2023]、Divide and Bind [Li *et al.*, 2023b]、InitNO [Guo *et al.*, 2024] 和 A-Star [Agarwal *et al.*, 2023] 设计了其他损失函数，以更好地增强提示的对齐，而其他工作则将布局信息与文本信息结合起来，以强制对象定位 [Chen *et al.*, 2024b; Xie *et al.*, 2023]。与我们的工作最接近的是 InitNO，它对初始潜在图像（初始噪声）执行一个预热的多轮优化。也就是说，他们试图调整初始潜在图像以达到期望的损失分数，旨在找到在生成过程中表现更好的初始噪声。“多轮”这一术语适用是因为如果未达到目标损失分数，这一过程最多可以进行五轮，每次都会重新采样和优化新的初始潜在图像。相比之下，我们认为在后期步骤优化潜在图像比在初始步骤更为有效。随着潜在图像的部分信息逐渐变得更准确，在较远的步骤中优化信息是有益的，此时潜在图像更容易与噪声区分开来，而扩散对潜在图像中的信号有更准确的理解。此外，我们的方法在不使用多轮优化的情况下更为高效。

扩散模型中的信号泄漏 Lin *et al.* [2024] 揭示出 Stable Diffusion 1.4 和其他一些扩散模型存在信号泄漏，这意味着即使在前向过程的最后步骤中信号也未完全消失。Everaert *et al.* [2024] 利用这种信号泄漏来控制生成的图像，使生成偏向于期望的风格，增强图像的多样性，并影响颜色和亮度。Grimal *et al.* [2024] 证明，某些在推理过程中使用的噪声在生成多个对象方面表现得更好。我们假设这种表现来自初始噪声中的某个信号，这种信号在使多个对象出现时更加一致。基于去噪过程中信号的构建，我们识别出可以改善信号并将其与文本对齐的扩散步骤。

3 方法论

3.1 初步：扩散与流动匹配

稳定扩散 1.4（标准差 1.4）[Rombach *et al.*, 2022] 和稳定扩散 3（SD 3）[Esser *et al.*, 2024] 分别基于扩散模型 [Ho *et al.*, 2020] 和流匹配 [Lipman *et al.*, 2023]。二者都依赖于一个逐渐使图像腐蚀的前向过程和一个逐步去除噪声的反向过程。尽管这些框架相近，但它们在核心机制上有所不同。两者都学习从一个分布 p_{source} 到 p_{target}，在我们的情况下，这就是图像的流形。

扩散模型 在扩散模型中，采样过程是随机的。它涉及不同时间步的噪声样本之间的联合分布 t 和 t'。在每一步 t'，模型估计前一步样本 $\mathbb{E}[x_{t'}|x_t]$ 的期望值，并从这个估计的分布中重新采样 $x_{t'}$。这个过程重复进行，直到达到最终目标分布。同样地，模型可以被训练来预测每一步添加的噪声，这通过以下损失函数来表达，其中 ϵ_θ 是模型：

$$\mathcal{L} = \mathop{\mathbb{E}}_{x,\epsilon\sim\mathcal{N}(0,I),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]. \tag{1}$$

流动匹配 采样过程是确定的。这个方法定义了从源分布到目标分布的路径。模型学习一个速度场 $v_t^{[\text{source,target}]}(x_t)$，该速度场将数据从 p_{source} 传输到 p_{target}。学习这个场涉及在每个步骤预测 $\mathbb{E}[v_t^{[\text{source,target}]}(x_t)|x_t]$。在训练期间，速度场是已知的；例如，我们可以选择线性流并让 $v_t^{[\text{source,target}]} = x_0 - \epsilon$ 和 $\epsilon \sim p_{\text{source}}$ 与 $x_0 \sim p_{\text{target}}$，线性地将 ϵ 传输到 x_0。通过最小化以下内容来学习模型 $v_\theta(x_t, t)$：

$$\mathcal{L} = \mathop{\mathbb{E}}_{x,\epsilon\sim\mathcal{N}(0,I),t}\left[\|v_t^{[\text{source,target}]}(x_t) - v_\theta(x_t, t)\|^2\right] \tag{2}$$

在此，我们考虑了 标准差 1.4 的扩散设置和 SD 3 的线性流匹配。在这两种情况下，初始分布是高斯分布，目

标分布是一组图像。在时间步 t，上述数据 x_t 是通过添加具有预定义计划的噪声得到的：

$$x_t = a_t x_0 + b_t \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{3}$$

，其中 a_t 和 b_t 由噪声调度器确定。这可以被看作是在原始信号 x_0 和噪声 ϵ 之间进行插值。增大 b_t 使得信号更难与噪声区分，从而控制退化水平。

为了用文本来调节生成过程，几位作者采用了交叉注意机制，该机制包括使用一个冻结的文本编码器（如 T5 或 CLIP）生成的提示 p 的嵌入。文本编码器生成 N 个标记的嵌入，模型在不同的交叉注意层中利用这些嵌入。在这些层中，对中间特征 Q 和文本嵌入 K 进行线性投影。然后计算注意图 $A = \text{softmax}(QK^T/\sqrt{d})$。这些注意图可以重塑为 $\mathbb{R}^{h \times w \times N}$，其中 h 和 w 表示交叉注意层中注意图的维度，N 表示提示嵌入的序列长度。如 [Hertz *et al.*, 2022; Tang *et al.*, 2023] 所示，交叉注意图揭示了空间布局和相应词语之间有意义的语义关系，可以用于可视化和控制。因此，模型被 $(x_t, t, \tau(p))$ 调节。

为降低扩散的计算成本，Rombach *et al.* [2022] 开发了一种在较小感知潜在空间中操作的潜在扩散模型（LDM）。该模型生成初始潜在噪声 z_T，并通过迭代去噪获得 z_0，然后将潜在图像投射到像素空间以生成最终图像 x_0。虽然我们的实验使用了 LDM，但我们的方法同样适用于像素空间。为清晰起见，我们将使用 x_t 描述该方法，即使我们的实验是在潜在空间中进行的。在推断过程中，我们可以通过使用一个将去噪过程离散化为较少步骤的采样调度器，在不遵循完整训练步骤的情况下生成图像。例如，使用带有 标准差 1.4 和 50 个采样步骤的 DDPM 调度器时，第一个采样步骤 0 对应于初始扩散过程的第 981 步，在保持生成质量的同时显著减少所需步骤的数量。

为了改进这个过程，最近的方法采用了两个可以结合但目的不同的过程。首先，他们采用了 GSN 引导（GSNg），使得步骤 t 处的潜在图像 x_t 通过应用关于损失 $\mathcal{L}$ 的（独特）梯度下降步骤进行偏移，以促进与提示的对齐，因此 $x_t : x_{t'} \leftarrow x_t - \alpha_t \cdot \nabla_{x_t} \mathcal{L}$，其中 α_t 是学习率。其次，该过程可以在一些预定义的采样步骤 $t_1 \dots t_k$ 处重复，直到 $\mathcal{L}$ 达到足够的阈值或已进行的偏移次数达到指定的最大数量。这个过程被称为迭代优化（迭代参考）步骤。我们认为仔细选择在哪个步骤执行 迭代参考 可以使我们只需执行一次，而无需与阈值进行比较，从而减少设置的超参数数量，同时获得更好的结果。

3.2 选择适合的 迭代参考 步骤来增强内容

我们的方法侧重于选择适当的去噪步骤，以在噪声中增强信号，从而生成更为真实的最终图像。先前的研究已经展示了扩散模型从粗到细的行为 [Park *et al.*, 2023]。在下文中，我们讨论了信号如何被退化然后通过扩散模型恢复。对于流匹配也可以得出相同的结论，因为信号退化遵循了 Equation 3 中描述的相同过程。

在反向过程过程中，模型首先重建图像的低频结构，然后逐步完善末端的细节。这种行为可以从 Equation 3 中的插值理解，在正向过程中，信号 x_0 逐渐减弱，而噪声 ϵ 增加。重要的是，在每一个扩散步骤，我们都可以估计最终图像，并获得底层信号的近似。

给定在特定扩散步骤中的任意 x_t，最终图像 x_0 可以估计为：

$$\hat{x}_0 = (x_t - b_t \epsilon_\theta(x_t))/a_t \tag{4}$$

在图 3 中，我们展示了扩散过程中的信号估计值以及 a_t 的值。随着过程的推进，信号变得更加明晰，使得最终图像的总体结构即便在早期阶段也能显现出来。信号 x_0 的降解和重构由噪声调度器控制。之前的研究 [Choi *et al.*, 2022; Chen, 2023] 已经强调过需仔细选择噪声调度，以为模型提供足够的时间构建图像的主要内容。这确保了模型有充分的机会准确地构建场景。在语义图像生成的背景下，这解释了为什么在图像的核心元素仍在形成的早期噪声水平时，对文本提示的关注较强 [Balaji *et al.*, 2023; Park *et al.*, 2023]。在后期阶段，文本输入的影响较小，模型专注于细节的完善，而总体空间结构保持不变。

我们的方法通过增强关键时间步的信号来利用这种理解：既不在信号较弱的情况下过早，也不在场景已经定义时过晚。这确保了信号在反向过程中始终保持足够强大，引导模型准确地语义构造最终图像。通过仔细选择步骤，我们可以放大潜在图像中的信号，从而更好地与文本提示进行语义对齐。为了自动选择表现最佳的步骤，我们提出了一种验证方法，通过评估指标（见 4.1）测试多个步骤。我们的方法总结在 Figure 2 中。

由于扩散模型的潜在空间本质上缺乏语义意义 [Kwon *et al.*, 2023; Park *et al.*, 2023]，导致其不适合直接操作来控制生成结果，我们依赖于模型对潜在表示的解释能力来赋予语义关联，并使用单一 迭代参考 步骤来增强信号，确保文本和图像之间的准确对齐。换句话说，我们修改模型解释的信号以提高其质量，确保模型接收到适当的信号进行正确的生成。此外，我们的仅需一次 迭代参考 步骤的方法具有多样性，可以与像 GSNg 这样的方法集成，以进一步改善图像生成。

3.3 根据文本到图像对齐任务增强信号

考虑一个包含主题标记列表 $\mathcal{S} = \{s_1, \dots, s_k\}$ 的提示 p，我们为每个主题提取注意力特征。对于 标准差 1.4，我们遵循 [Chefer *et al.*, 2023] 为每个主题 s 获得一个注意力图 A^s。对于 SD 3，架构集成了称为 MM-DiT 的转换器模块，其中潜在图像 x_t 被分块化并通过注意力机制与 T5 和 CLIP 嵌入一起处理。该机制可以被视为自注意力和交叉注意力的结合，使得提取有意义的注意力图变得具有挑战性。为了解决这个问题，我们隔离和优化对应于 CLIP 和 T5 的注意力图，在实体之间进行平均，并应用 GSN 标准进行对齐。更多细节见 补充材料。为了确保每个主题标记的注意力，我们考虑

$$\mathcal{L}_{CN} = \max_{s \in S}(1 - \max_{i,j}(A^s_{i,j})) \tag{5}$$

，由 Chefer *et al.* [2023] 提出，其中 $A^s_{i,j}$ 代表在位置 i, j 上的主题标记 s 的交叉注意力值。它鼓励激活最小的标记变得更活跃。此外，我们实现了已在 [Agarwal *et al.*, 2023] 中使用的交并比 (IoU) 损失，通过促进主题分离来缓解灾难性混合。对于所有主题标记对 $\mathcal{C}$ 的组合，损失定义为

$$\mathcal{L}_{\text{IoU}} = \frac{1}{|\mathcal{C}|} \sum_{\forall (m,n) \in \mathcal{C}} \left(\frac{\sum_{i,j} \min(A^m_{i,j}, A^n_{i,j})}{\sum_{i,j} (A^m_{i,j} + A^n_{i,j})} \right) \tag{6}$$

，其中 $A^s_{i,j}$ 表示主题标记 s 在位置 i, j 上的交叉注意力值。最后，我们的损失定义为 $\mathcal{L} = \mathcal{L}_{CN} + \mathcal{L}_{\text{IoU}}$，在潜在图像 x_t 的 50 次位移步骤中通过 Adam 优化器 [Kingma and Ba, 2017] 和学习率 1×10^{-2} 进行最小化。根据以前的研究，这些超参数被固定以进行公平比较。

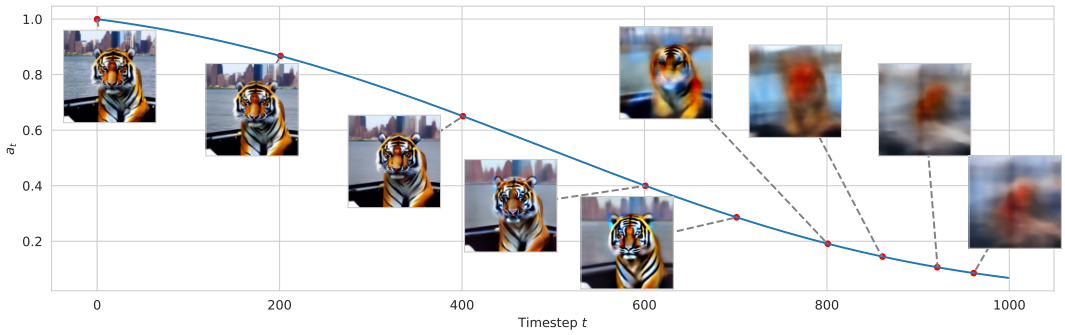

Figure 3: a_t 的值是时间步长 t 的函数（目标分布的 $t = 0$ 和高斯分布的 $t = 1000$ ）。在生成“老虎坐船抵达纽约的照片”的不同步骤中估计的 $\hat{x}\_0$ 被显示。观察到从粗略到细致的生成；随着去噪过程的进行，场景变得越来越容易辨认。使用 Stable Diffusion 1.4 生成。

Methods	迭代参考 Which Step	迭代参考 Reach Threshold	迭代参考 Max Shift	GSNg	Max Gradient Updates of x_t
Syngen	ø	ø	ø	25 first steps	25
Attend & Excite	0 10 20	✓	20	25 first steps	85
Divide & Bind	0 10 20	✓	50	25 first steps	175
InitNO	0	✓ up to 4 restart if it fails	50	ø	90
InitNO+	0 10 20	✓ up to 4 restart if it fails ✓	50 20	25 first steps	315
Ours	8	ø	50	ø	50
Ours+	2	ø	50	from 3 to 25	73

Table 1: 方法概述。步骤是根据采样调度器给出的。最大移位表示如果没有满足阈值或没有使用阈值时所应用的最大预定义移位。最大梯度更新指的是生成过程中更新潜在图像的最大次数。

4 实证分析与结果

4.1 实验设置

实现 我们使用 标准差 1.4 ，因为所有超参数方法都是基于这个模型。图像是在 Nvidia A100 80GB 中以 Float 32 精度使用 DDPM 调度器和 50 个推理步骤生成的，并具有 7.5 的无分类器指导 [Ho and Salimans, 2022] 。我们将我们的方法与其他仅依赖于模型内部知识的推理方法进行比较，包括标准的稳定扩散推理、Attend & Excite、Divide & Bind [Li *et al.*, 2023b] 、InitNO 和 Syngen。我们排除 A-Star，因为缺乏官方实现，并且 InitNO 报告了更优的结果。InitNO 的作者建议将他们的方法与 GSNg 和 迭代参考 步骤结合使用，我们称之为 InitNO+。我们的方法称为 Ours，其变体结合了 Syngen 的 GSNg 称为 Ours+，其中 GSNg 在迭代细化步骤之后应用。我们还比较了有无我们方法的情况下 SD 3 的结果。方法的总结见 Table 1 ，进一步细节见 补充材料 。

评估指标 继之前的 GSN 方法之后，我们采用了 [Chefer *et al.*, 2023] 提出的基于 CLIP 的指标，本文中称为相似度评分，包括完整提示相似度、最小对象相似度和文本-文本相似度。与之前的方法不同，我们还计算了 CLIP 评分 [Radford *et al.*, 2021] 以衡量文本和图像嵌入之间的平均对齐程度。然而，使用这些基于 CLIP 的指标时需要谨慎，因为它们通常难以理解关系，可能会错误地将对象与其属性关联，并表现出显著的顺序敏感性缺乏 [Yuksekgonul *et al.*, 2023] 。因此，我们报告了 TIAM 评分 [Grimal *et al.*, 2024] ，这是一种与人类判断密切相关、反映正确生成图像比例的指标。对于每个提示，我们生成了多张图像，并自动进行评估以确保请求实体正确呈现，并在适用时，包括颜色等属性。我们还报告了 LAION 的美学预测器 [Schuhmann *et al.*, 2022] ，评分范围为 1 到 10。最后，我们进行了用户研究以补充评估。关于指标的更多详情，请见 补充材料 。

数据集 根据 TIAM [Grimal *et al.*, 2024] 推荐的采样方法，我们利用 24 个 COCO 标签 [Lin *et al.*, 2014] 以及可能的颜色，为所有可能的两个和三个主题实体组合生成了提示。对于每个数据集，采样了 300 个提示，并使用相同的 16 个种子为每个提示生成 16 张图像，以创建测试集。此外，我们通过采样与 300 个测试提示不同的 10 个提示，创建了四个验证数据集，以确定最佳的 迭代参考 步骤。四个数据集包括两个实体、两个有色实体、三个实体和三个有色实体。

最优 迭代参考 步骤选择 我们评估了 即 中 50 个采样步骤中的 11 个采样步，每两个步骤间隔一个（0, 2, 4, . . . , 24）。我们重点关注前 25 步，因为之前的研究表明在这之后的益处有限 [Chefer *et al.*, 2023] 。对于每个验证数据集，我们使用相同的 16 个随机种子为每个提示生成 16 张图像，并计算 TIAM 得分。我们使用每个数据集的最小-最大标准化方法对得分进行标准化，并在 迭代参考 步中展示 Ours 和 Ours+ 的累积标准化 TIAM 得分 Figure 4 。根据得分，我们发现不使用 GSNg 时第 821 步（采样步骤 8）是最佳的，而在使用 GSNg 时，第 941 步（采样步骤 2）产生更好的结果。这种差异可以通过需要在没有 GSNg 的情况下在过程后期发生改变来解释，以确保调整后的信号足够强大，可以通过随机采样持续存在。相比之下，GSNg 使信号能够连续优化，即使在后期阶段也允许进行修正。此外，我们计算审美得分，观察到无论选择何种 迭代参考 步骤都没有退化，确认了选择（值可在 补充材料 中找到）。我们将在后续实验中使用这些选定的步骤。我们的验证方法在计算上是高效的，仅需 10 个提示和有限的样本数量即可确定最佳的迭代参考 步骤。我们采用相同的方法选择 SD 3 的最佳步骤（详情见 补充材料 ）。

4.2 定量结果

TIAM 我们在 Table 2 中展示了我们的方法以及其他方法的 TIAM、CLIP 和美学评分。凭借 标准差 1.4 ，我们的方法在所有配置中都在 TIAM 和 CLIP 评分上优于 InitNO，仅用一个 迭代参考 步，没有使用 GSNg 。这表

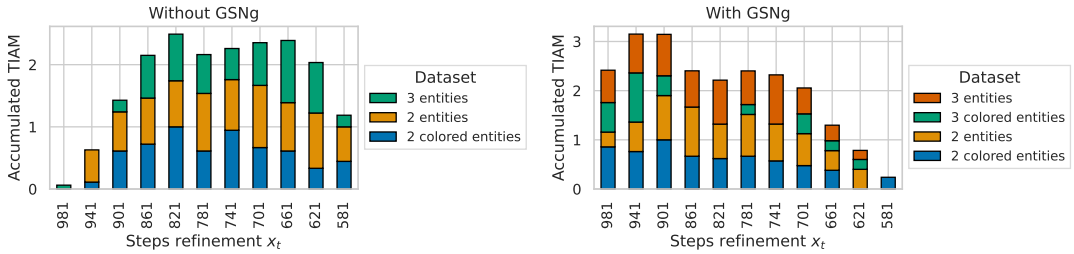

Figure 4: 累积的 TIAM 分数在没有（左）和有（右）GSNg 的情况下。左侧排除了具有三种颜色实体的数据集，因为其得分很低。步骤 821 和 941 被识别为最佳。

	迭代参考	GSNg	Methods	w/o colors 2 entities	w/o colors 3 entities	with colors 2 entities	with colors 3 entities
SD 1.4	0	✗	Stable Diffusion	$45.4_{32.2/5.5}$	$8.4_{33.5/5.5}$	$3.9_{34.6/5.4}$	$0.1_{34.5/5.4}$
	1	✗	InitNO	$62.1_{33.1/5.5}$	$14.2_{34.3/5.4}$	$7.2_{35.7/5.4}$	$0.2_{35.5/5.3}$
			Ours	$65.8_{33.7/5.5}$	$23.1_{35.4/5.5}$	$8.7_{36.4/5.4}$	$0.4_{36.3/5.4}$
	3	✓	Divide & Bind	$69.9_{33.7/5.5}$	$33.6_{35.9/5.4}$	$11.3_{36.1/5.4}$	$0.5_{36.1/5.3}$
			Attend & Excite	$71.4_{34.0/5.5}$	$32.0_{35.9/5.4}$	$10.5_{36.9/5.4}$	$0.6_{36.9/5.3}$
			InitNO+	$75.0_{34.1/5.5}$	$34.2_{36.0/5.4}$	$11.9_{37.1/5.4}$	$1.0_{37.3/5.3}$
	0	✓	Syngen	$\underline{78.5}_{34.1/5.4}$	$\underline{39.2}_{36.5/5.4}$	$\underline{20.4}_{37.1/5.3}$	$\underline{2.4}_{36.8/5.3}$
	1	✓	Syngen+	$75.8_{33.8/5.3}$	$36.2_{36.2/5.4}$	$20.1_{37.1/5.3}$	$1.9_{36.9/5.3}$
			Ours+	$81.1_{34.2/5.4}$	$45.8_{36.7/5.4}$	$20.5_{37.1/5.3}$	$2.8_{37.1/5.3}$
SD 3	0	✗	Stable Diffusion	$82.8_{34.8/5.5}$	$63.4_{37.9/5.5}$	$27.3_{38.2/5.4}$	$9.69_{39.4/5.3}$
	1	✗	Ours	$84.5_{34.9/5.6}$	$70.7_{38.2/5.6}$	$24.2_{38.1/5.4}$	$9.71_{39.6/5.4}$

Table 2: TIAM 在包含两个或三个实体的提示中的表现，有和没有颜色说明项。下标指的是 CLIP/美学分数 。最优值用粗体显示，对于 标准差 1.4 ，次优值用下划线标出。对于 SD 3 ，只有最优值用粗体显示。

	迭代参考	GSNg	Methods	Full Prompt	Minimum Object	Text-Text
SD 1.4	0	✗	Stable Diffusion	0.3313	0.2400	0.7682
	1	✗	InitNo	0.3411	0.2512	0.7901
			Ours	0.3470	0.2564	0.7979
	3	✓	Divide & Bind	0.3468	0.2597	0.8065
			Attend & Excite	0.3509	0.2634	0.8032
			InitNO+	<u>0.3520</u>	0.2638	0.8076
	0	✓	Syngen	0.3518	<u>0.2640</u>	<u>0.8122</u>
	1	✓	Ours+	0.3522	0.2643	0.8133
SD 3	0	✗	Stable Diffusion	0.3529	0.2616	0.8181
	1	✗	Ours	0.3535	0.2619	0.8190

Table 3: 基于[Chefer *et al.*, 2023] 的两个实体之间的相似度分数。最佳值用粗体显示，次佳值对于 标准差 1.4 用下划线表示。对于 SD 3 ，只有最佳值用粗体显示。

明，当信号强于初始扩散步骤时，一个单一的 迭代参考 步骤更有效，这与我们的方法预期一致。当与 GSNg 结合时，我们在 TIAM 评分方面超过了所有其他方法，表明 GSNg 在我们精心选择的 迭代参考 步骤中引导到更好的结果。在所有配置中，我们都实现了更高的 CLIP 评分，除非是三种颜色实体，在所有方法中 TIAM 对齐评分普遍非常低。为了公平比较，我们尝试为 Syngen 方法添加一个 迭代参考 步骤，称为 Syngen+，但获得了更低的分数。更多细节在 补充材料 中。通过 SD 3 ，我们的方法减轻了灾难性忽视，显示出对两种和三种实体的 TIAM 和 CLIP 评分的改善。我们注意到两个颜色实体的性能略有下降，但对于三个颜色实体，TIAM 评分几乎相同，而 CLIP 评分更好。

相似度分数 我们在 Table 3 中呈现了包含两个实体的数据集的评分结果。对于 标准差 1.4 ，没有 GSNg 的情况下，我们的方法始终优于 InitNO，这证实了谨慎选择扩散步骤以执行 迭代参考 步骤的重要性。使用 GSNg 时，我们超越了所有竞争方法。虽然在包含两个和三个实体的数据集上我们取得了稍好一点的表现，但对于包括颜色指定的数据集，Ours+ 略低。这可能是由于 CLIP-based 指标在捕捉精确的句法关系方面的限制导致的。Ours 在所有数据集上都优于 SD 3 ，只是在两个有颜色的实体的一个指标上略有下降。其他数据集的结果以及关于这个评分限制的进一步讨论在 补充材料 中。

用户研究 我们进行了一个主观用户研究，通过各个方法在 标准差 1.4 上的比较来评估人类偏好，包括 37 个候选项。每次比较中，我们展示了每个方法使用相同随机选择的提示和种子生成的图像，参与者被要求选择最佳匹配或者如果适用则选择 “无”。研究包括两个阶段。

	w/o GSNg		w GSNg		
	Ours	InitNO	Ours+	Syngen	InitNO+
%	43.1 %	36.9 %	57.4 %	51.9 %	43.3 %

Table 4: 用户研究的结果：左边（没有 GSNg ）和右边（有 GSNg ）。百分比表示每种方法被选择的频率。

在第一阶段，我们比较了 InitNO 与 Ours，随后在第二阶段我们评估了 Syngen、InitNO+ 和 Ours+。我们在 Table 4 中展示了结果。我们的方法在单步 迭代参考 设置中比 InitNO 显示出显著的改进，进一步验证了我们方法的有效性。此外，通过引导，参与者选择 Ours+ 的频率高于其他选项，表明其与文本提示具有更好的一致性。更详细的信息可以在 补充材料 中找到。

4.3 定性比较

我们针对没有在 Figure 5 和 Figure 6 中应用 GSNg 的方法，使用相同的两个种子和不同的提示词对图像生成进行了定性比较。我们的方法更好地缓解了灾难性忽视问题，例如 InitNO 在提示词中 “长椅” 和 “大象” 的照片难以清晰地代表这两个实体。即使在包含三个实体的复杂提示下，我们的方法仍能产生优越的结果，因为后面的迭代参考 步骤有助于更有效地区分实体。在 Figure 7 中，我们展示了使用 GSNg 的方法的结果。我们的方法显著增强了三个对象的分离。例如，Syngen 和 InitNO+ 有时无法生成某些实体（例如 Syngen：在第一个提示中未生成车，在第三个提示中未生成鸟；InitNO+：在第一个提示中未生成冰箱，在最后一个提示中未生成胡萝卜）。此外，我们的方法更好地区分了这些实体（例如 Syngen：在第二个提示中羊不能区分，而 InitNO+ 在第二个提示

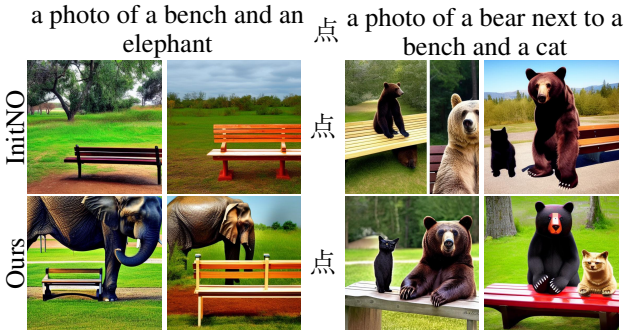

Figure 5: 不带 GSNg（ 标准差 1.4 ; 相同的种子) 生成的图像

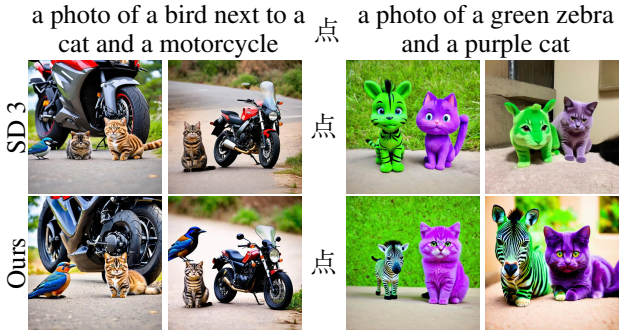

Figure 6: 使用 SD 3 生成的图像（两者使用相同的种子）

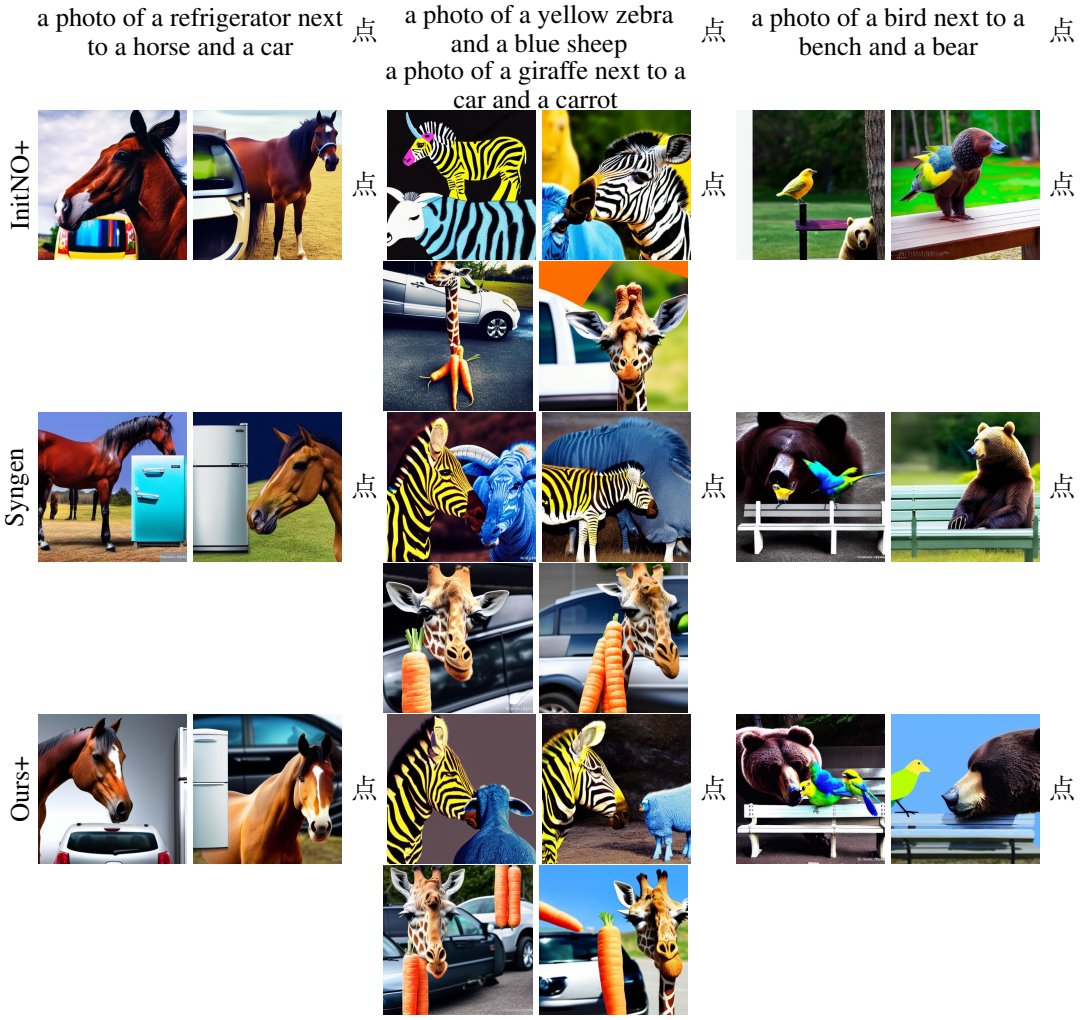

Figure 7: 使用 GSNg 生成的图像 (标准差 1.4 ; 相同的种子)

的第二张图像中将羊和斑马混在一起，并在第三个提示的第一张图像中将鸟和熊混在一起)。与现有方法相比，我们的方法在有效生成和区分实体方面表现出卓越的性能。我们为 标准差 1.4 /3 在 补充材料 中提供了更多示例。

4.4 迭代参考 布局研究

我们对进行 迭代参考 步（Figure 8 ）的最佳扩散步骤进行了详尽研究。subsection 4.1 中确定的候选步骤与结果高度一致，因为它们在所有数据集中始终表现出良好性能。这进一步证实了我们用于确定 迭代参考 步候选的验证方法的有效性：过早优化的效果不如在后期阶段进行调整。然而，如果过多地延迟修正也是有害的，这表

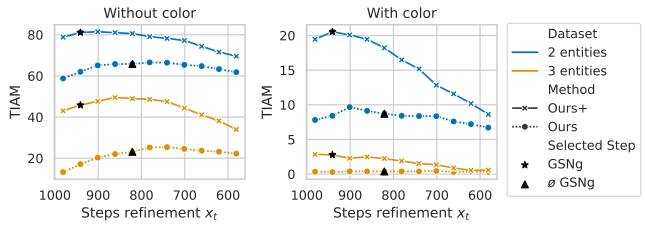

Figure 8: 根据 迭代参考 步骤评估有颜色（右）和无颜色（左）实体的 TIAM 分数。

明在修改信号时需要谨慎权衡时间点。我们注意到，使用 GSNg 遵循类似的趋势，但通过促进对信号的微小、持续调整，始终产生更好的结果。我们还发现，对于彩色数据集，通过设置不同的 迭代参考 步可以获得更好的结果。这一结论源于理解：颜色修改应在扩散过程的早期实施，因为颜色似乎在一开始就被确定。稍后进行调整可能会阻碍有效的集成。相比之下，在后期修改实体的信号更有利，这允许更精确地区分不同的实体。

5 局限性

尽管我们可以通过精心设计的 GSN 损失来整合外部信息，GSN 方法仍然受到模型固有知识的限制。这一限制影响了我们有效优化的能力，因为在稀有概念、对象混淆、推理、计数的情况下仍存在挑战。因此，由于模型的分布外行为，我们可能会遇到失败。我们的工作已证明，深入理解扩散过程中的信号构建，可以选择优化步骤来增强图像生成，同时限制超参数和 迭代参考 步骤的数量，例如根据扩散步骤设置优化阈值。然而，我们相信，尽管与测试众多阈值和超参数相关的挑战，这种利用精心设计的优化阈值的方法可以提高性能，特别是在考虑信号构建时。最后，和其他 GSN 方法一样，我们的方法需要通过模型进行反向传播，这在计算上是密集的。

6 结论

在本研究中，我们通过探索信号在去噪过程中的演变来改进 GSN 标准的应用。我们提出了一种识别和验证最佳优化步骤的方法。我们的研究结果表明，虽然早期的信号修改效果较差，但及时的调整可以显著提高性能，能够生成语义对齐的图像并实现最先进的成果，这在广泛的实验中得到了证明。此外，与一些最先进的方法 例如 InitNO 相比，该方法减少了超参数和 迭代参考 的数量，从而简化了模型设置并提高了整体效率。我们观察到，迭代参考 步骤的位置取决于我们希望纠正的具体元素。例如，颜色修改应在过程中较早进行，而实体的调整可以稍晚一点进行。GSN 方法的未来发展可以基于这些见解，选择针对特定需要调整方面的优化步骤。此外，结合一个提醒损失 [Agarwal *et al.*, 2023]，可以通过为模型提供跨采样步骤的信号记忆来进一步增强该方法。

7

致谢

本研究使用了 GENCI-IDRIS（资助号 2022-AD011014009）的高性能计算资源，并使用了由法兰西岛大区委员会资助的 FactoryIA 超级计算机。此项研究部分得到了 SHARP ANR 项目 ANR-23-PEIA-0008 的支持，该项目是在法国 2030 计划的背景下进行的。

References

[Agarwal *et al.*, 2023] Aishwarya Agarwal, Srikrishna Karanam, K J Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2283–2293, October 2023.

[Balaji *et al.*, 2023] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. *arXiv 2211.01324*, 2023.

[Chefer *et al.*, 2023] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans. Graph.*, 42(4), jul 2023.

[Chen *et al.*, 2023] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.

[Chen *et al.*, 2024a] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation, 2024.

[Chen *et al.*, 2024b] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5343–5353, January 2024.

[Chen, 2023] Ting Chen. On the importance of noise scheduling for diffusion models, 2023.

[Choi *et al.*, 2021] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14367–14376, October 2021.

[Choi *et al.*, 2022] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11472–11481, June 2022.

[Esser *et al.*, 2024] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024.

[Everaert *et al.*, 2024] Martin Nicolas Everaert, Athanasios Fitsios, Marco Bocchio, Sami Arpa, Sabine Süsstrunk, and Radhakrishna Achanta. Exploiting the Signal-Leak Bias in Diffusion Models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4025–4034, January 2024.

[Feng *et al.*, 2023] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis, 2023.

[Grimal *et al.*, 2024] Paul Grimal, Hervé Le Borgne, Olivier Ferret, and Julien Tourille. Tiam - a metric for evaluating alignment in text-to-image generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2890–2899, January 2024.

[Guo *et al.*, 2024] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization, 2024.

[Hertz *et al.*, 2022] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022.

[Ho and Salimans, 2022] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv 2207.12598*, 2022.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

[Kingma and Ba, 2017] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[Kwon *et al.*, 2023] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space, 2023.

[Li *et al.*, 2023a] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *CVPR*, 2023.

[Li *et al.*, 2023b] Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. Divide & bind your attention for improved generative semantic nursing. In *34th British Machine Vision Conference 2023, BMVC 2023*, 2023.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[Lin *et al.*, 2024] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5404–5411, January 2024.

[Lipman *et al.*, 2023] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[Meng *et al.*, 2022] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

[Mou *et al.*, 2023] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

[Park *et al.*, 2023] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[Peebles and Xie, 2022] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

[Podell *et al.*, 2023] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[Ramesh *et al.*, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv 2204.06125*, 2022.

[Rassin *et al.*, 2023] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[Rissanen *et al.*, 2023] Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation. In *International Conference on Learning Representations (ICLR)*, 2023.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[Saharia *et al.*, 2022] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton,

Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[Schuhmann *et al.*, 2022] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

[Segalis *et al.*, 2023] Eyal Segalis, Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. A picture is worth a thousand words: Principled recaptioning improves image generation, 2023.

[Sun *et al.*, 2024] Wenqiang Sun, Teng Li, Zehong Lin, and Jun Zhang. Spatial-aware latent initialization for controllable image generation, 2024.

[Tang *et al.*, 2023] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023.

[Xie *et al.*, 2023] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7452–7461, 2023.

[Yu *et al.*, 2023] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[Yuksekgonul *et al.*, 2023] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2023.

[Zhang *et al.*, 2023] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

A 附录

附录总结如下：

- 第 A.1 节：实施和方法的详细描述，
- 第 **??** 节：描述 GSN 在稳定扩散 3 上的适配以及在验证数据集上的实验结果，
- 第 **??** 节：TIAM 评估过程概述，
- 第 **??** 节：Attend & Excite 的评估框架总结及附加结果，
- 第 **??** 节：关于用户研究的详细信息，
- 第 **??** 节：额外的比较示例输出，
- 第 **??** 节：主文件中的图形值和补充结果，包括用于验证集的第 **??** 节和用于测试集的第 **??** 节。

A.1 文本到图像方法的设置以实现稳定扩散 1.4

在此我们提供一些关于稳定扩散 1.4 的实现细节和使用的方法。

稳定扩散版本 1.4 (SD 1.4) 我们使用托管在 HuggingFace 上的模型 [1]，与 DDPM Scheduler [2] 和 50 个采样步骤。所有方法都在 Classifier Free Guidance [Ho and Salimans, 2022] 为 7.5 的情况下进行。

参加 & 激励 我们使用了 Diffusers 库提供的实现 [3]。迭代优化发生在采样步骤 0、10 和 20 时，其中损失必须分别达到指定的阈值 0.05、0.5 和 0.8。最多执行 20 次迭代优化步骤。学习率随着每个采样步骤逐步降低，初始值为 20。在前 25 个采样步骤中，他们进行了 GSN 指导。

分而结合 我们使用官方实现 [4]。我们遵循作者的建议，对于无颜色的提示使用 tv loss，对于有颜色的提示使用 tv bind loss。迭代优化发生在采样步骤 0、10 和 20，在这些步骤中，损失必须分别达到 0.05、0.2 和 0.3 的指定阈值。最多进行 50 次迭代优化步骤。学习率随着每个采样步骤逐渐降低，起始值为 20。在前 25 个采样步骤中，他们进行 GSN 引导。

初始化 NO 我们利用了在代码库 [5] 中提供的官方实现。作者设计了一个包含三个成分的损失函数：自注意力损失、交叉注意力损失和 KL 散度损失。在多轮步骤中，会进行迭代优化。如果未达到交叉注意力和自注意力损失的定义阈值，则通过采样新的初始潜变量重复优化，最多可尝试五次。如果目标仍然难以达成，则使用实现相对于目标最佳得分的优化初始潜变量进行推断。KL 散度损失仅在提升步骤中应用，在对注意力损失进行每次反向传播后进行优化，以确保初始潜变量图像保持在适当的区间内。迭代优化步骤也在采样步骤 10 和 20 进行。对于提升步骤和迭代优化，损失必须满足指定的阈值：交叉注意力损失为 0.2，自注意力损失为 0.3。学习率随着每次采样步骤逐步下降，初始值为 20。此外，GSN 引导在前 25 步采样中被应用。

此外，我们在代码中发现实现包括一个干净的跨注意力损失，该损失在多轮步骤和 GSNg 期间使用大津阈值处理注意力图。代码还为 GSNg 加入了跨注意力对齐损失，似乎旨在鼓励扩散步骤中标记激活区的一致性。据我们所知，这些细节在主论文中没有提到。

Syngen 我们使用官方实现 [6]。他们仅在前 25 个采样步骤中应用 GSN 指导。他们使用的学习率为 20。

Syngen 被设计用于接受仅包含具有属性的实体的提示。例如，当提示为“猫和狗的照片”时，会利用与“照片”对应的交叉注意图。为了增强结果，我们移除了与初始标记相关的交叉注意图。这一调整在实验中使性能提高了大约 1。论文中报告的 Syngen 的得分反映了这些有益的修改。

我们尝试为 Syngen 引入一个细化步骤。具体来说，我们在首次采样步骤应用了一个细化步骤，类似于 InitNO，并使用 Syngen 的损失函数进行了 20 次和 50 次优化迭代。采用学习率为 1×10^{-2} 的 Adam 优化器。TIAM 分数在 **??** 中报告。然而，与没有细化步骤的配置相比，我们没有取得更好的结果。虽然可能存在改进空间，但需要进一步研究以确定最佳超参数。

我们的 根据 Attend & Excite 框架，我们对注意力图应用高斯平滑，使用核大小为 3 和标准差为 0.5。在迭代优化步骤中，我们进行了 50 次潜在图像偏移，而没有试图达到特定阈值。对于使用 GSN 指导的配置，在迭代优化步骤后，我们加入了 Syngen GSN 指导。

我们使用 Hugging Face [7] 中可用的、具有默认调度器 FlowMatchEulerDiscreteScheduler [8] 的实现，配置为 28 个采样步骤，使用无分类器引导 [Ho and Salimans, 2022] 为 7.0，以及 bfloat16 精度用于图像生成。对于 迭代参考 ，我们应用 Adam 优化器，学习率为 1×10^{-2} ，并进行 50 步的优化。

稳定扩散 3（SD 3）是一个流匹配模型，旨在在两个分布 p_0 和 p_1 之间构建一个概率路径，其中 p_0 是目标分布，$p_1 \sim \mathcal{N}(0, I)$ 。该模型学习将点从一个分布运输到另一个分布。潜在图像运输路径可以被解释为一个去噪过程，其中噪声以类似于图像破坏的方式逐步去除。具体而言，潜在图像 x_t 是使用重参数化技巧进行采样的，涉及图像和噪声的插值。正如 Rissanen *et al.* [2023] 所展示的，各向同性噪声抑制数据中功率谱密度比噪声方差更低的频率分量。因此，模型最初重建较低频率，然后进一步精细化较高频率，类似于扩散模型中观察到的过程。在去噪过程中，信号可以通过 GSN 方法精细化以确保与期望的输出对齐。此外，我们的方法可以用于选择去噪过程中的最佳步骤。虽然在稳定扩散模型 1.4 和 1.5 中的特征提取已被广泛记录，据我们所知，对使用基于 transformer 架构的稳定扩散 3 的研究还不深入。在这种架构中，T5 和 CLIP 作为两个不同的编码器提供指导。模型结合了两个独立的 transformer，每个在自己的模态空间（图像片段和文本）内工作，同时在处理注意力时考虑其他模态。我们首先描述如何处理和提取注意力图，然后说明如何选择潜在图像精细化的一个潜在优良步骤。

提取注意力图 稳定扩散 3 包含 24 个转换器块。潜在图像表示为 $x_t \in \mathbb{R}^{H \times W \times c}$ ，其中 c 是潜在空间中的通道数，H, W 为高度和宽度，被分块以生成一系列标记 $z \in \mathbb{R}^{hw \times d}$ ，其中 $hw = \frac{1}{2}H \times \frac{1}{2}W$ ，d 是标记嵌入维度。

文本嵌入 t 是通过将来自 CLIP 和 T5 的嵌入连接起来并将它们投射到相同的维度 d 来形成的。这产生了 $t \in \mathbb{R}^{(n_{\text{CLIP}}+n_{\text{T5}}) \times d}$ ，其中 n_{CLIP} 和 n_{T5} 分别代表来自 CLIP 和 T5 的标记数。

在处理注意力时，得到的注意力图 A 的大小为 $A \in \mathbb{R}^{(hw+n_{\text{CLIP}}+n_{\text{T5}})^2 \times n_{\text{head}}}$ ，其中 n_{head} 是注意力头的数量。我们提取注意力图，并专注于子集，其中图像块作为查询，文本嵌入作为键。这个子集至关重要，因为它捕捉了图像潜在特征与文本概念之间的关系，确保潜在图像内的信号与标记的语义含义一致。

为了简化注意力图，我们在注意力头和 transformer 块之间进行平均，得到 $A \in \mathbb{R}^{hw \times (n_{\text{CLIP}}+n_{\text{T5}})}$ 。我们进一步改进这些图，通过排除特殊标记（例如，起始标记和终止标记）来处理 CLIP 和 T5，因为这些标记往往主导注

[1] https://huggingface.co/CompVis/stable-diffusion-v1-4

[2] https://huggingface.co/docs/diffusers/api/schedulers/ddpm

[3] https://huggingface.co/docs/diffusers/api/pipelines/attend_and_excite

[4] https://github.com/boschresearch/Divide-and-Bind

[5] https://github.com/xiefan-guo/initno

[6] https://github.com/RoyiRa/Linguistic-Binding-in-Diffusion-Models

[7] https://huggingface.co/stabilityai/stable-diffusion-3-medium-diffusers

[8] https://huggingface.co/docs/diffusers/api/schedulers/flow_match_euler_discrete