Detail++: 无需训练的文本到图像扩散模型细节增强器

Lifeng Chen[1] Jiner Wang[1] Zihao Pan[1] Beier Zhu[1, 2] Xiaofeng Yang[1, 2] Chi Zhang[1†]

[1] AGI 实验室，西湖大学，[2] 南洋理工大学。

https://detail-plus-plus.github.io/

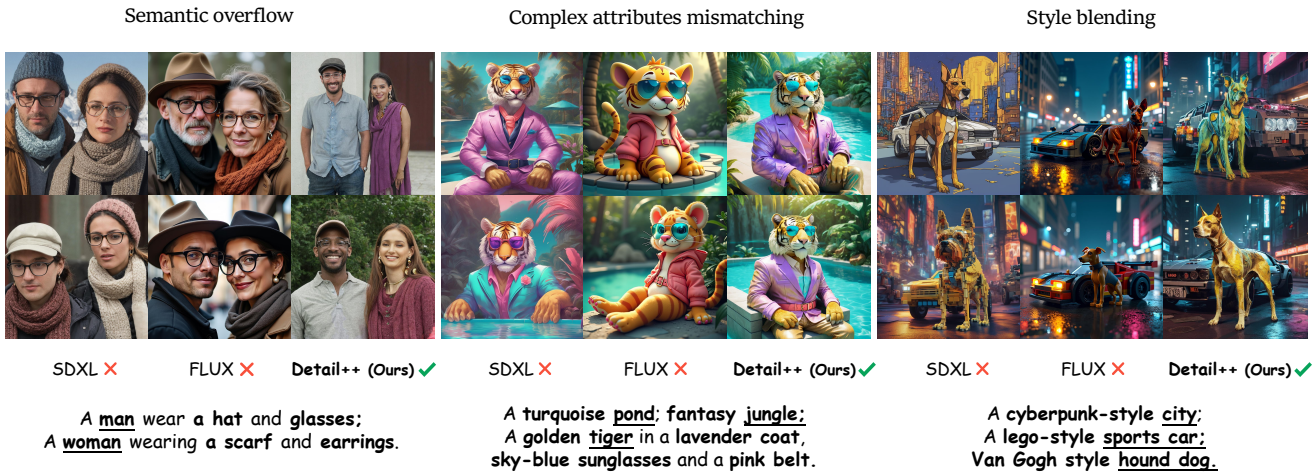

Figure 1. A comparison between our method and current state-of-the-art generative models. The mainstream models often suffer from issues such as semantic overflow, complex attribute mismatching, and style blending. Even Flux, the leading generative model under the DiT framework, struggles to overcome these challenges. In contrast, our method, Detail++, based on SDXL, achieves highly accurate semantic binding in a training-free way.

Abstract

近期在文本生成图像（T2I）方面的进展带来了令人印象深刻的视觉效果。然而，这些模型在处理复杂提示时仍面临显著挑战，尤其是涉及多个具有不同属性的主体时。受人类绘画过程的启发——先勾勒出构图，然后逐步增加细节，我们提出了 Detail++，一种无训练框架，引入了新颖的渐进式细节注入（PDI）策略来解决这一限制。具体来说，我们将复杂提示分解为一系列简化的子提示，引导生成过程分阶段进行。这个分阶段的生成利用自注意力固有的布局控制能力，首先确保整体构图，然后进行精确的细化。为了实现属性与相应主体之间的准确绑定，我们利用交叉注意力机制，并在测试时引入一个质心对齐损失，以减少绑定噪声并增强属性一致性。在 T2I-CompBench 和一个新构建的风格组合基准上的大量实验表明，Detail++ 显著优于现有方法，尤其是在涉及多个对象和复杂风格条件的场景中。

† Corresponding author.

1. 介绍

近年来，文本到图像（T2I）生成 [? ? ? ? ?] 技术取得了显著进步，能够根据文本描述创建高质量、细节丰富的图像，在艺术设计、广告、娱乐和教育方面展现出巨大的潜力 [? ? ?] 。尽管有所改进，目前的方法在处理涉及多个主体的复杂描述时仍然困难重重。当描述为多个主体分配不同属性集时，一个被称为“细节绑定”的著名挑战就会出现。这常常导致模型错误地将属性互换到不同主体上，造成语义泛滥、错误的属性匹配和不期望的风格混合，如图所示 1 。

针对上述挑战，我们观察到当前最先进的模型尝试同时渲染提示中的所有元素及其属性，导致细节绑定不完美。受人类艺术实践的启发——艺术家通常先绘制基本的空间布局和轮廓，然后逐步完善细节——我们逐步注入属性，以准确地将特定细节整合到其对应的主题区域，从而实现更语义一致的图像生成。为了实现我们的目标，我们提出了 Detail++，这是一种多分支的图像生成框架，通过不同的简化版本提示协同工作以生成准确的图像内容。该过程从使用原始提示的初始去噪分支开始，在此过程中，我们提取并共享自注意力图以保持一致的布局。第二个去噪分支使用最简

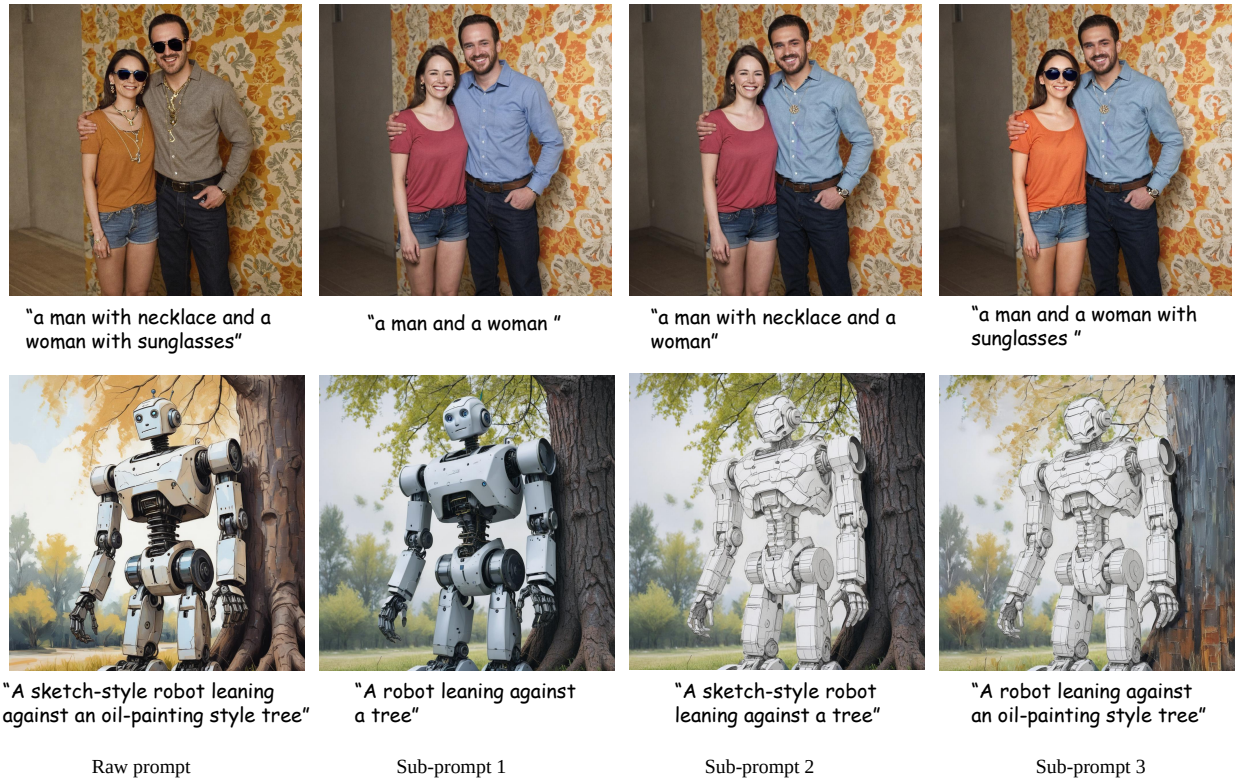

Figure 2. Detail++ 的基本过程。如第一列所示，在单个分支中生成复杂的提示常常导致不准确或混合的属性分配。例如，"sunglasses" 和"necklace" 这样的属性可能会错误地应用在错误的主体上。我们的方法通过逐步的方法来解决这一挑战：我们首先忽略所有复杂的修饰符以生成粗略的生成基础，然后系统地注入细节以确保每个属性被精确添加到其对应的主体区域。显示在每个图像下方的提示指示了用于该特定分支的输入，第二行展示了该方法在风格组合场景中的效果。注意，这里的所有四个分支都是并行生成的。

单的子提示——仅包含主题及其基本布局——以创建没有细节的基础。在随后的并行分支中，每个子提示中的特定属性逐步注入到图像中，所有分支保持相同的总体布局。为了确保细节正确绑定到对应的主体，我们引入了一种累积潜在修改策略，使用交叉注意力图生成二进制掩码，精确定位不同的主体区域以进行目标细节注入。此外，我们在测试时还引入了质心对齐损失，以优化主题特定的交叉注意力图，解决图不准确性并进一步提高细节注入的精度。整个过程如图 2 所示。

我们提出的 Detail++ 在广泛使用的 T2I-CompBench 基准测试 [?] 和我们自定义的风格组合基准测试上进行了评估。结果表明，Detail++ 始终优于现有方法，尤其是在涉及多重风格绑定的场景中。在定性评估中，Detail++ 无需额外微调 [? ? ? ?] 或预定义布局信息 [? ? ? ? ? ?] ，便能生成高质量的图像，突显了我们方法的实用性。本文的主要贡献如下：

- 我们创新性地提出了 Detail++，这是一种无需训练的多分支框架。它通过渐进属性注入提升了现有 T2I 模型在处理涉及多个主体的复杂提示时的准确性。
- 我们开发了一种自注意力图共享机制，利用 U-Net 中自注意力图的特性来控制图像布局，从而为准确的属性绑定创建一致的基础。
- 我们进一步引入了一种基于质心对齐损失的测试时优化策略，这进一步提高了属性绑定的准确性。
- 大量实验表明，Detail++ 相对于当前最佳方法的优势，随后我们将开源代码。

2. 相关工作

2.1. 文本到图像扩散模型

扩散模型已成为文本到图像（T2I）生成的主导方法 [? ? ? ? ? ? ?] 。类似 Stable Diffusion [?] 、DALL-E 2 [?] 和 Imagen [?] 的模型通过逐步去噪，生成高质量的图像。它们利用神经网络将文本映射到视觉图像，通常使用预训练的文本编码器，如 CLIP [?] ，以增强文本理解和指导生成。然而，即使是最先进的模型，如 Flux [?] ，仍然难以生成与文本提示密切对齐的图像。这导致了一系列研究旨在提高 T2I 的对齐性。

如前所述，T2I 基础模型难以准确遵循文本条件，常常面临属性溢出（属性溢出到未提及的主体上）、不匹配（属性被错误匹配）以及混合（来自不同主体的属性混合到一个主体上）等问题。为了解决这些挑战，已经提出了几种解决方案。其中，一些基于测试时优化的方

法旨在将属性的交叉注意力图与主体标记对齐，但它们通常较慢且难以高效地优化噪声。其他基于布局的方法依赖于用户或 LLM 预定义布局，通常涉及复杂的中间步骤。还有一些需要微调的方法，比如 ELLA，通过将 LLM 整合到 T2I 生成模型中来增强模型对文本的理解，但需要对基础模型进行显著修改，从而导致高昂的训练成本。虽然这些方法在解决属性不匹配问题上取得了显著进展，但仍然在溢出和混合方面存在困难。最近，一种称为 ToME 的方法通过合并 CLIP 文本编码器生成的标记嵌入，提供了解决溢出的新方法。然而，属性混合问题仍然在很大程度上未能解决。我们的方法 Detail++，通过逐步注入属性，有效地避免了这三种问题，从而实现了不同类型属性的精确绑定，包括对象、颜色、纹理和风格等。

2.2. 图像编辑

图像编辑在保持整体逼真的情况下操控特定区域。传统的文本驱动图像编辑方法 [? ? ?] 已通过结合 GANs [? ?] 和 CLIP 显示出前景。最近，大量基于扩散模型的新方法 [? ? ? ? ?] 已经涌现，旨在以无训练的方式解决图像编辑任务的各个方面，其中大多数 [? ? ? ? ?] 利用了注意力机制 [?] 。例如，Prompt-to-Prompt (P2P) [?] 调整交叉注意力图以实现各种编辑操作。MasaCtrl [?] 通过互相注意的方式实现了刚性编辑，同时保持整体纹理和身份。在这些工作的基础上，FPE [?] 区分了稳定扩散模型中自注意力和交叉注意力的不同功能。我们的方法从这些工作中汲取灵感，并进一步扩展到渐进编辑框架。

3. 绪论

本节介绍注意力机制的基础知识，它是我们提议的方法的关键组成部分。

U-Net 中的注意力机制。 U-Net 架构 [?] 已成为许多最先进扩散模型的基石。扩散模型中的 U-Net 集成了自注意力和交叉注意力机制，每种机制在图像生成中都有独特的作用。交叉注意力侧重于提取提示的语义元素，确保生成的图像与文本提示一致 [?] 。交叉注意力中的注意力矩阵可以定义为：

$$M_{\text{cross}} = \text{Softmax}\left(\frac{Q_{\text{img}}K_{\text{text}}^{\top}}{\sqrt{d_k}}\right), \tag{1}$$

其中，Q_{img} 是从 U-Net 的内部特征中得出的查询嵌入，K_{text} 是源自提示的文本嵌入的关键嵌入，d_k 是关键向量的维度。另一方面，自注意力主要与图像的空间布局和结构细节有关，通过计算图像特征空间内的关系 [?] 。这种注意力图可以表示为：

$$M_{\text{self}} = \text{Softmax}\left(\frac{Q_{\text{img}}K_{\text{img}}^{\top}}{\sqrt{d_k}}\right), \tag{2}$$

其中，Q_{img} 和 K_{img} 都是从 U-Net 内部特征中得出的查询和关键矩阵。自注意力图提供了一种内部关系结构，这对于保持空间一致性至关重要。为了更好地理解自注意力和交叉注意力之间的区别，我们在图 3 中分别可视化了它们的注意力图。我们提出的方法基于这些基础的注意力机制，特别是对它们进行了修改和扩展，以改善对复杂提示的处理。

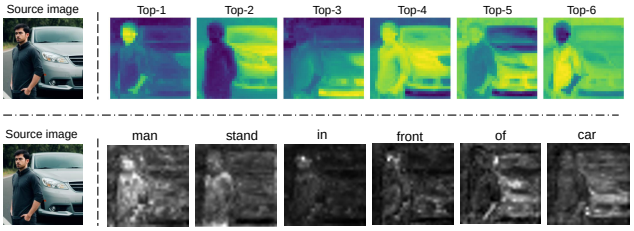

Figure 3. 注意力可视化。「男人站在汽车前面」提示的自注意力图和交叉注意力图的可视化。自注意力图通过显示奇异值分解后获得的前 6 个成分来实现可视化 [?] 。交叉注意力图的可视化对应于提示中的每个标记。

在本节中，我们提供了我们方法 Detail++ 的概述。我们的核心思想是首先忽略所有复杂、难以生成的细节以获得一个粗略的生成基础，然后逐步以类似绘画的方式添加更多细节。对于一个复杂的原始提示，我们自动生成一系列简化的提示，如 Sec. 3.1 中所述。为了确保从这些具有不同属性的提示生成的图像保持相同的布局，我们将原始提示生成过程中获得的自注意力图与所有子提示生成过程共享 (Sec. ??)。这个共享的自注意力图提供了一个稳定的结构基础。在我们的累积潜在修改策略（Sec. 3.1 ）的控制下，正确的细节逐步注入到相应的主题区域。在每个时间步中，每个主题关键词的交叉注意力图用于创建精确的掩码，以引导选择性的属性注入。此外，在 Sec. ?? 中，我们引入了一个中心对齐损失，以在测试阶段实现更准确和聚焦的交叉注意力图，确保添加的细节严格限制在它们预定的主题区域。整体工作流程如图 Fig. 4 所示。

3.1. 渐进细节注入

提示分解。 为了处理详细提示的复杂性，我们方法的第一步是简化输入提示。具体来说，我们使用一个语言模型，例如 spaCy [?] 或 ChatGPT [?] ，将原始复杂提示 p_0 分解为一系列逐步简化的子提示，记为 $\mathcal{P} = \{p_0, p_1, \ldots, p_{n-1}, p_n\}$ ，其中 p_1 代表通过移除所有修饰符后的最简化版本。此简化确保在 p_1 分支中消除所有属性，为逐步细节注入提供了一个清晰的基础。子提示 $p_2, p_3, \ldots, p_n$ 分别向 p_1 增加一个额外的修饰符。分解的选择在附录 ?? 中有更详细的说明。

此外，对于每一对提示 $p_{i+1}(i > 0)$ 和 p_1 ，我们也使用语言模型来识别新增属性在 p_{i+1} 中对应的主体 q_i 。这些主体用集合 $\mathcal{Q} = \{q_1, \ldots, q_{n-2}, q_{n-1}\}$ 表示。例如，给定提示 p_0 = “一个穿绿色运动服的红色泰迪熊”，集合 $\mathcal{P}$ 和主体集合 $\mathcal{Q}$ 如下所示：

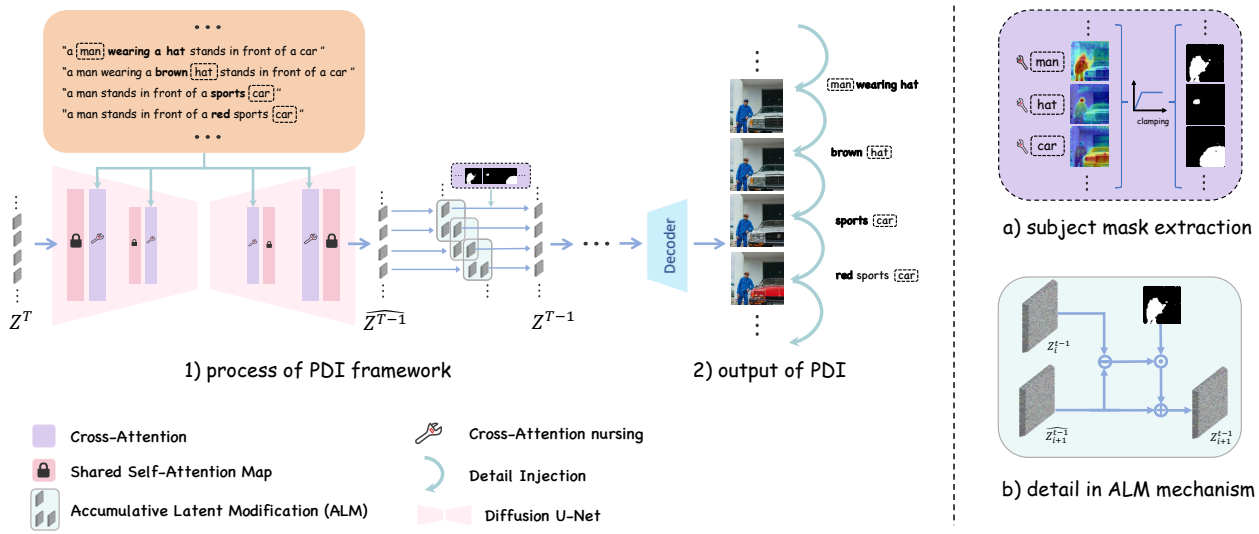

Figure 4. Detail++ 概述。我们的方法包含一个大的框架，称为渐进细节注入（PDI），以及基于重心对齐损失的测试时注意力培养。通过 spaCy 分解的子提示被一起传递通过 U-Net 进行并行推断。在每一步降噪中，生成的潜变量批次经历累积潜变量修改（ALM），仅修改当前添加属性对应的区域。请注意，在我们的框架中，所有分支的自注意力图都是统一的，这确保了一致的布局并避免了编辑效果之间的冲突。

$$\mathcal{P} = \left\{ \begin{array}{l} p_0 : \text{“}a\,red\,teddybear\,wearing\,a\,green\,tracksuit\text{”}, \\ p_1 : \text{“}a\,teddybear\,wearing\,a\,tracksuit\text{”}, \\ p_2 : \text{“}a\,red\,teddybear\,wearing\,a\,tracksuit\text{”}, \\ p_3 : \text{“}a\,teddybear\,wearing\,a\,green\,tracksuit\text{”} \end{array} \right\},$$

$$\mathcal{Q} = \left\{ \begin{array}{l} q_1 : \text{“}teddybear\text{”}, \\ q_2 : \text{“}tracksuit\text{”} \end{array} \right\}.$$

共享自注意力图。一旦我们得到了子提示集，下一步就是使用这些提示合作生成图像。由于先前的工作已经表明 U-Net 中的自注意力图存储了图像的布局信息，我们让网络基于每个子提示生成图像，同时通过共享自注意力图来实现一致的布局。具体来说，我们在第一个分支中缓存 U-Net 自注意力图，并在后续分支中重用它们。通过这种方式，我们使不同子提示生成的图像与原始提示的布局保持一致。观察到去噪的早期阶段对于构建图像的整体布局最为关键，因此我们仅在去噪过程的最初 S 步中使用这一策略。在剩下的 $T-S$ 步中，我们允许每个子提示的 U-Net 模型独立预测噪声，以确保图像的保真度，就像在标准扩散过程中那样。通过这种方式，我们可以让不同子提示生成的图像布局与原始提示保持一致。

累积潜在修改。 为了精确地将每个子提示中新增的属性绑定到它们对应的主体，我们进一步为每个属性注入开发了潜在级别的主体屏蔽。具体来说，我们基于每个主体 q_i (Sec. 3.1) 相应的交叉注意力图从一个子提示中提取一个二进制掩码，突出显示图像中每个主体的相关区域。这个过程涉及对交叉注意力图进行归一化，并应用阈值操作为每个主体 q_i 创建一个明确的二进制掩码。主体 q_i 的二进制掩码提取由以下公式给出：

$$B_i = \mathbf{1}\left[\frac{\overline{M_i} - \min(\overline{M_i})}{\max(\overline{M_i}) - \min(\overline{M_i})} > \tau\right], \tag{3}$$

，其中 B_i 是主体 q_i 的二值掩码，$\overline{M_i}$ 代表其跨层平均的注意力图。?? 。τ 是阈值参数。指示函数 $\mathbf{1}[\cdot]$ 将归一化后的注意力值转换为二值掩码，其中高于 τ 的值被设置为 1，其他值被设置为 0。$\min(\overline{M_i})$ 和 $\max(\overline{M_i})$ 分别表示平均注意力图的最小值和最大值。

二进制掩码可以将潜在修改限制在特定区域。具体而言，对于 i -th 分支，修改后的潜在表示计算如下：

$$z_{i+1}^{t-1} = z_i^{t-1} + B_i \odot (\widehat{z_{i+1}^{t-1}} - z_i^{t-1}), \tag{4}$$

，其中 $\widehat{z_{i+1}^{t-1}}$ 表示第 $i+1$ -th 分支中的去噪潜在，z_i^{t-1} 是来自前一分支的潜在，B_i 是当前主体的二进制语义掩码，z_{i+1}^{t-1} 是整合了正确属性的修改后潜在。运算符 $\odot$ 表示逐元素相乘。

例如，如果一个区域的掩码值 b 是 1，这表示该区域对应目标主体。在这种情况下，通过附加提示细节生成的新潜在 z_{i+1} 被应用于该区域，确保我们的框架将添加的细节准确融入指定区域。相反，如果某个区域的 b 为 0，表明该区域与当前属性无关，模型将重用来自先前分支的潜在 z_i ，从而剔除新属性对无关区域的影响。这种选择性应用能防止跨主体干扰，并保持每个主体特征在图像中的区别。因此，渐进注意力替换的整个过程在算法 1 中进行了概述。

在从交叉注意力图生成二元掩码时，我们观察到现有交叉注意力机制中的一个关键问题：注意力激活往

Algorithm 1 渐进细节注入

Require: A source prompt p_0 , and sub-prompts derived from it $p_1, ..., p_n$, DM represents the diffusion models to predict the noise of next step. T is the total denoising timesteps.

Ensure: Input $\mathcal{P} = \{p_0, p_1, \ldots, p_n\}$ to model in parallel.

$z_0^T, z_1^T, \ldots, z_n^T \leftarrow z^T \sim \mathcal{N}(0, \mathcal{I})$; ▷ Initialized with same latent.

$\mathcal{Z}^T \leftarrow \{z_0^T, z_1^T, \ldots, z_n^T\}$

for $t = T, \ldots, T - S + 1$ do

  $\mathcal{Z}^{t-1} \leftarrow \mathrm{DM}(\mathcal{Z}^t, \mathcal{P}, t, \overline{M_{self}^t})$; ▷ $\overline{M_{self}^t}$ is from first branch.

  for $i = 1, 2, \ldots, n-1$ do

    $\widehat{z_{i+1}^{t-1}} \leftarrow z_{i+1}^{t-1}$

    $z_{i+1}^{t-1} \leftarrow z_i^{t-1} + B_i \odot (\widehat{z_{i+1}^{t-1}} - z_i^{t-1})$; ▷ B_i is the bianry mask consistent with Sec 4.2.3.

  end for

end for

for $t = T - S, \ldots, 1$ do

  $\mathcal{Z}^{t-1} = \mathrm{DM}(\mathcal{Z}^t, \mathcal{P}, t)$;

end for

return $\mathcal{Z}^0$

往是分散且不集中的，导致一个主体的注意力扩散到其他主体区域，从而导致属性注入错误和生成质量下降。为了解决这个问题，我们提出了一种测试时优化方法，通过引入中心对齐损失来细化交叉注意力图。我们的关键见解是在理想情况下，一个主体的交叉注意力图应该在主体区域的中心形成集中的激活，而不是分布在整个图像空间中。基于此，我们设计了一种机制，鼓励每个主体的显著注意点（具有最大注意力值的点）与其注意力分布的中心对齐。

具体来说，我们首先计算每个主体 q_i 的注意力图的质心 $p_{\text{centroid}}(q_i)$ ，其中 q_i 表示在第 i 个分支中添加属性的主体词。这个质心作为对应于提示中 q_i 的标记的交叉注意力图的加权中心：

$$p_{\text{centroid}}(q_i) = \frac{1}{\sum_{h,w} \overline{M_i}(h,w)} \begin{bmatrix} \sum_{h,w} w \cdot \overline{M_i}(h,w) \\ \sum_{h,w} h \cdot \overline{M_i}(h,w) \end{bmatrix}, \tag{6}$$

，其中 h 和 w 分别代表高度和宽度的空间坐标，而 $\overline{M_i}(h,w)$ 是该位置的归一化注意力值。

接下来，我们通过最小化注意力中心与最亮点 $p_{\max}(q_i)$ 之间的欧几里得距离来定义重心对齐损失，从而鼓励更集中的注意力分布：

$$L_{\text{align}} = \sum_i \left\| p_{\text{centroid}}(q_i) - p_{\max}(q_i) \right\|^2. \tag{7}$$

我们将此损失项与 ToME [?] 中使用的熵损失相结合，以形成总损失函数 L_{total} 。熵损失定义为：

$$L_{\text{ent}} = -\sum\nolimits_{m \in \overline{M_i}} m \log(m), \tag{8}$$

其中 m 表示标准化交叉注意力图中主题 q_i 的注意力值 $\overline{M_i}$ 。我们的总损失函数变为 $L_{\text{total}} = L_{\text{align}} + \lambda L_{\text{ent}}$ ，其中 λ 是权衡超参数。在每个扩散步骤 t 期间，我们通过梯度下降更新潜在变量：

$$\boldsymbol{z}_t' \leftarrow \boldsymbol{z}_t - \alpha_t \cdot \nabla_{\boldsymbol{z}_t} L_{\text{total}}, \tag{9}$$

其中 α_t 是步长。

实验结果表明，引入重心对齐损失不仅可以生成更集中的交叉注意力图，还显著提高了属性注入的精度。优化后的注意力图生成更准确的二进制掩码，确保详细属性被正确注入到相应的主体区域，有效解决了复杂提示中的属性混淆问题。

4. 实验

4.1. 实验装置

评估基准和指标。我们在 T2I-CompBench [?] 上进行了广泛实验，这是一个专门为处理涉及复杂组合提示的文本到图像生成任务而设计的广泛使用的基准。具体来说，我们选择了 T2I-CompBench 中用于评估颜色、形状、纹理等属性的子集，因为这些直接反映了我们方法的语义对齐能力。为了进一步验证生成图像的主观质量并评估人类偏好得分，我们引入了 ImageReward [?] 模型，这是一种在图像生成任务中广泛接受的用于反映人类感知判断的学习指标。现有的基准通常缺乏有效的评估机制来评估与多种艺术风格组合相关的属性，这面临风格混合的问题。为了弥补这一缺陷并提供更全面的评估，我们提出了一种新颖的基准，称为风格组合基准 (SCB)。SCB 由 300 个精心设计的提示组成，涵盖广泛的独特艺术风格。更多细节可参见附录 8 。

我们对 SCB 的评估方法独立测量图像中每个元素与其相应的风格描述符之间的对齐程度。具体而言，它涉及解析每个生成图像的提示，以识别与不同语义组件相关的风格描述符。生成图像中的每个组件首先使用 Grounding DINO [?] 根据从提示解析出的组件词进行裁剪。然后，我们使用 CLIP-Score [? ?] 模型计算每个分割元素的组件风格对应关系。每个生成图像的最终风格分数被计算为所有分割语义元素跨风格匹配分数的平均值。因此，基于对参考图像的偏差，我们的新指标可以有效评估是否发生了风格混合。

实现细节。我们的方法基于 SDXL [?] 基线构建，使用 SpaCy [?] 自然语言处理工具包进行修饰符识别和自动提示管理。我们模型的不同分支是并行处理的，这大大减少了运行时间。自注意力图共享机制在总去噪步骤的 80 % 时激活，使用所有模块的 32×32 图大小进行注意力共享。在测试时优化中，我们设置 $\lambda = 1$ 以在这两种损失之间实现良好平衡。更多实验细节可以在附录中找到。

比较方法。为了全面评估我们提出的方法的性能，我们对多个方法进行了比较分析。除了与已建立的基线方

Table 1. 在多个指标上对细节绑定性能进行定量比较。我们提出的 Detail++ 在颜色、纹理、形状和风格绑定方面优于所有基线。最高得分在 blue 中突出显示，次高得分在 green 中。

Method	Train	BLIP-VQA ↑			Human-preference ↑			SCB ↑
		Color	Texture	Shape	Color	Texture	Shape	
SD1.5[?]	✓	0.4719	0.4334	0.3898	-0.360	-0.689	-0.483	0.2196
Composable Diffusion[?]	✗	0.4063	0.3645	0.3299	-0.209	-1.043	-0.813	0.2032
Structured Diffusion[?]	✗	0.4990	0.4900	0.4218	-0.025	-0.325	0.169	0.2207
GORS[?]	✓	0.6603	0.6287	0.4785	0.017	-0.417	1.115	-
ELLA v1.5[?]	✓	0.6911	0.6308	0.4938	-	-	-	-
SDXL[?]	✓	0.6369	0.5637	0.5408	0.733	0.124	1.184	0.2488
PixArt- α [?]	✓	0.6886	0.7044	0.5582	0.242	-0.788	0.165	0.2398
Attention Regulation[?]	✗	0.5860	0.5173	0.4672	0.268	-0.603	1.118	0.2248
Ranni xl[?]	✓	0.6893	0.6325	0.4934	-	-	-	-
ToMe[?]	✗	0.6583	0.6371	0.5517	-0.028	-0.851	0.454	0.2554
Detail++(Ours)	✗	0.7389	0.7241	0.5582	1.773	0.300	1.424	0.2687

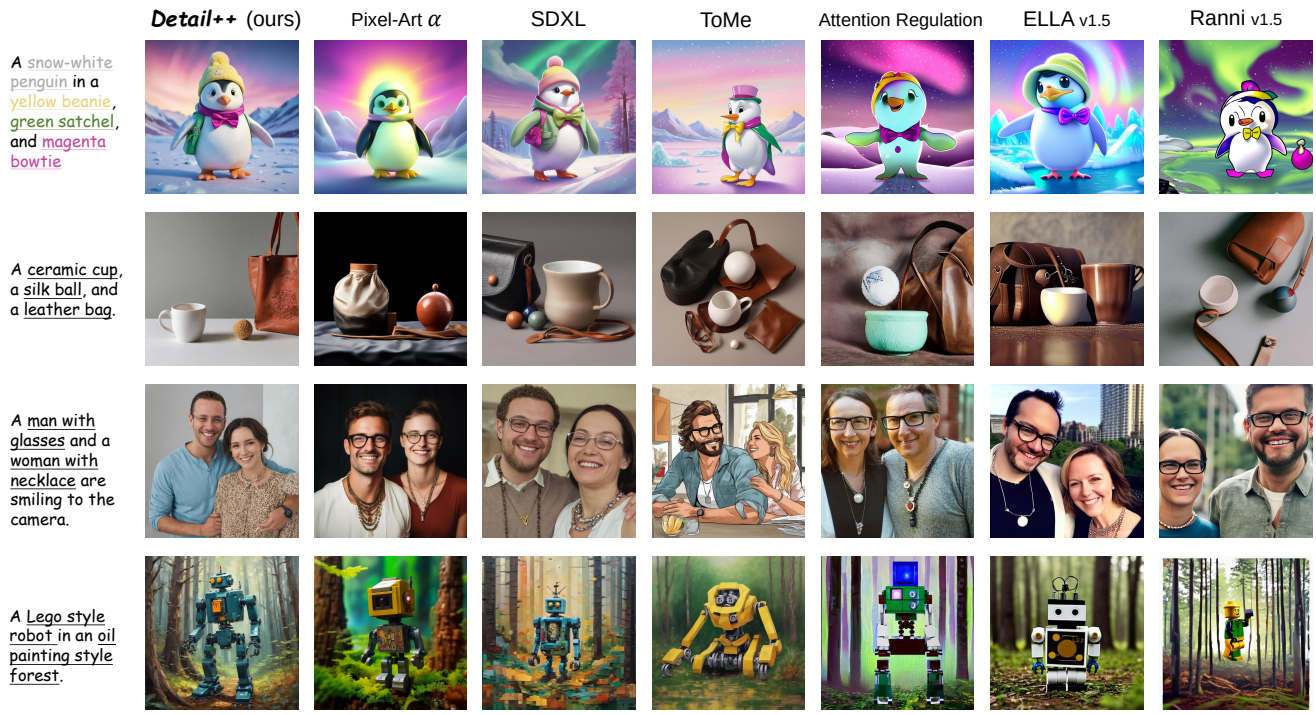

Figure 5. 对复杂提示（包括多种属性类型，如对象、颜色、纹理和风格）的各种方法进行定性比较。我们的方法有效地防止了属性溢出、复杂属性不匹配以及风格混合，同时保持了高视觉保真度。

法 SDXL [?] 进行比较外，我们还包括了多个专门为增强细节结合和构图性而设计的先进方法：Ranni [?]，这种方法引入了一个语义面板，作为文本提示与图像生成之间的中介；ELLA [?]，因利用大型语言模型增强文本提示与生成图像之间的对齐而闻名；注意力调节 [?]，这是一种专注于动态调节注意力权重以提高语义准确性并减少视觉模糊的复杂方法；ToME [?]，基于高效的令牌合并的最新方法，在构成清晰度和视觉保真度方面表现出显著改善。

4.2. 实验结果

定量比较。如表 1 所示，我们的方法在 T2I-CompBench [?] 上的颜色、纹理、形状绑定能力，以及在我们提出的 SCB 基准上的风格绑定能力方面，显示

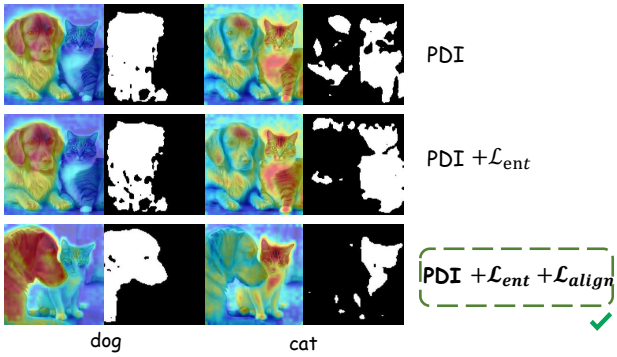

Figure 6. 关于二进制掩码提取中不同优化项的消融研究。结合使用 L_{ent} 和 L_{align} 比仅使用 L_{ent} 或没有优化能产生更准确且集中的主体掩码。

Figure 7. 对于提示“戴着墨镜的狗和戴着项链的猫”的不同优化策略的视觉结果。所有三行中的结果均基于相同的随机种子生成。结合使用 L_{ent} 和 L_{align} 能够实现更精确的细节注入，防止属性错误地溢入相邻的主体区域。

Table 2. T2I-CompBench 上不同优化项的消融研究。结合使用 L_{ent} 和 L_{align} 表现最好。

PDI	$\mathcal{L}_{\text{ent}}$	$\mathcal{L}_{\text{align}}$	BLIP-VQA Color	BLIP-VQA Texture	BLIP-VQA Shape
✗	✗	✗	0.6369	0.5637	0.5408
✓	✗	✗	0.6942	0.6836	0.5507
✓	✓	✗	0.7034	0.6995	0.5521
✓	✓	✓	0.7389	0.7241	0.5582

出优越或相当的性能。此外，使用 ImageReward [?] 分数作为人类偏好的代理指标，我们的方法（Detail++）显示出更强的人类偏好一致性。

定性比较。图 5 展示了一个定性比较，说明了不同方法如何处理四种不同的提示类型：复杂颜色绑定、复杂纹理绑定、对象绑定和风格构成。从第一和第二行可以看出，对于颜色和纹理属性，当主体的数量较多时，属性错误匹配变得更加突出。在第三行，对于对象类型属性，经常发生语义泄漏，导致未修改的主体受到其他主体修饰符的影响，例如项链修饰了男性的同时，女性也受到太阳镜的影响。在第四行，我们观察到现有主流方法难以处理应用于不同组件的多种风格修饰符，往往导致风格混合。然而，我们的方法是唯一能够成功生成乐高风格机器人，同时保持纯油画背景，没有两者之间的混合。视觉结果清楚地展示了我们的方法（Detail++）在防止属性溢出、复杂属性绑定和风格混合方面的优越性，这与表 1 中的定量结果一致。

消融研究。如表 2 所示，我们对不同组件和损失项进行了定量分析。可以观察到，仅使用 L_{ent} [?] 对实验结果稍有改善。然而，当加入所提出的 L_{align} 后，结果显示出显著提升。此外，我们在图 6 中可视化了生成提示词“戴墨镜的狗和戴项链的猫”时的生成二值化掩码。在中间图片集中，我们观察到添加熵损失有助于一定程度上使交叉注意力图收敛。但‘cat’令牌的高亮区域仍然包含了‘dog’的前额和颈部区域，这可能导致属性添加不精准。在图 7 的第二行中，这一点显而易见，其中给猫添加项链也无意中影响了狗的颈部区域。然而，我们提出的质心对齐损失进一步优化了初始分支生成，使主体的交叉注意力图更为集中和准确，大大减少了不同主体之间二值化掩码的重叠，这也可以在图 7 中生成的图像中看到。

为了全面评估我们提出的方法的有效性，我们在图 ?? 中进行了包含 145 名参与者的用户研究。该研究由每份问卷 12 个问题组成，涵盖三个关键方面：属性绑定、图像质量和风格绑定。为了确保评价的可靠性和无偏性，我们为每个对比模型构建了随机图像池（包括 SDXL [?] , ToMe [?] , 注意力调节 [?] , ELLA [?] , Ranni [?] 以及我们的方法），每个池中包含 120 幅图像，这些图像由 30 个不同的提示和 4 个不同的随机种子生成。调查中的每个问题，图像都是基于相应的提示从这些池中随机选择的。结果表明，我们的方法在所有三个指标上都稳定地优于所有基线方法。

在这项工作中，我们提出了 Detail++，这是一种新型的无需训练的方法，用于解决复杂提示中的文本到图像生成中精确细节绑定的挑战。我们的渐进细节注入框架结合了自注意力图共享和累计潜在修改，以确保适当的属性分配，同时保持一致的布局。引入的中心对齐损失通过优化交叉注意力图进一步提高绑定精度。对公共基准的大量实验表明，Detail++ 在防止属性溢出、错误匹配和风格混合方面明显优于现有方法，同时保持了高图像保真度。作为与当前扩散模型兼容的即插即用模块，Detail++ 代表了一项重要进展，使得无需额外训练即可实现更受控和语义上准确的文本到图像生成。

5. 局限性和讨论

提出的渐进细节注入（PDI）框架通过一种无训练的方法解决文本到图像生成中的细节绑定问题，但它仍然存在某些局限性。该方法非常依赖于初始自注意力图的质量和准确性——如果在早期阶段的布局生成不理想，随后的属性注入过程可能无法完全纠正这些问题，从而可能限制最终生成的质量。

6. 实现细节

方法细节。我们的方法是在 SDXL 上实现的。我们从 U-Net 中所有块的所有层中提取分辨率为 32×32 的交叉注意力图，因为我们发现这相对精细和准确。此外，我们所有的实验都在单个 Nvidia H-20 GPU 上进行。

基线方法的实现。在表 1 的定量比较中，我们使用 Stable diffusion 1.5 [?] 、Stable diffusion XL [?] 、Structured Diffusion Guidance [?] 、Composable diffusion [?] 、PixelArt- α [?] 、ELLA [?] 、ToME [?] 、Attention Regulation [?] 的官方实现。由于 Ranni [?] 的 SDXL 检查点尚未开源，因此我们直接参考他们论文中的得分。

在本节中，我们比较不同的提示管理粒度对生成质量的影响。例如，给定原始提示 $p_{\text{ori}} =$ “一个戴太阳镜的红色狗和一个戴项链的蓝色猫”，我们可以通过两种方式管理子提示：1）最复杂的优先，或 2）最简单的优先。关于粒度，我们可以通过两种方式管理它们：a）每个分支一个主题，或 b）每个分支一个属性。这导致总共有四种可能的配置组合。这些是：

配置 A (1 + a):

配置 B (1+b):

$$\mathcal{P} = \left\{ \begin{array}{l} p_0 : \text{“}a\,red\,dog\,with\,sunglasses\,and\,a\,blue\,cat\\ with\,a\,necklace.\text{”},\\ p_1 : \text{“}a\,dog\,and\,a\,cat.\text{”},\\ p_2 : \text{“}a\,red\,dog\,and\,a\,cat.\text{”},\\ p_3 : \text{“}a\,dog\,with\,sunglasses\,and\,a\,cat.\text{”},\\ p_4 : \text{“}a\,dog\,and\,a\,blue\,cat.\text{”},\\ p_5 : \text{“}a\,dog\,and\,a\,cat\,with\,a\,necklace.\text{”} \end{array} \right\},$$

$$\mathcal{Q} = \left\{ \begin{array}{l} q_1 : \text{“}dog\text{”},\\ q_2 : \text{“}dog\text{”},\\ q_3 : \text{“}cat\text{”},\\ q_4 : \text{“}cat\text{”} \end{array} \right\}.$$

配置 C (1 + a):

$$\mathcal{P} = \left\{ \begin{array}{l} p_1 : \text{“}a\,dog\,and\,a\,cat.\text{”},\\ p_2 : \text{“}a\,red\,dog\,with\,sunglasses\,and\,a\,cat.\text{”},\\ p_3 : \text{“}a\,dog\,and\,a\,blue\,cat\,with\,a\,necklace.\text{”} \end{array} \right\},$$

$$\mathcal{Q} = \left\{ \begin{array}{l} q_1 : \text{“}dog\text{”},\\ q_2 : \text{“}cat\text{”} \end{array} \right\}.$$

配置 D (2+b) :

因此，我们进行了大量实验来测试哪种方法能最好地反映我们方法的效率，如表格中所示。另一种分词方法是累积提示，其中先前添加的属性继续出现在下一个分支的提示中。这被称为累计提示：然而，如表格倒数第二行所示，量化结果并不理想。我们假设这是由于较长提示中的修饰符溢出，这降低了我们方法的效率。

7. 时间复杂度

时间复杂度。如表 ?? 所示，我们将时间成本与另外两种最新的无训练方法 [? ?] 进行了比较。由于并行推理的优势，我们的方法在保持与基线相似的运行时间的同时，交付了更好的结果。相对而言，其它方法通常需要使用 LLM [?] 进行局部推理或依赖于在去噪过程中的测试时优化，这往往导致时间消耗不理想。

8. SCB 详情

风格组合基准（SCB）构建。为了全面评估多风格组合场景下的文字到图像生成，我们采用统一的提示格式——“在一种 [type of style] 风格的 [背景] 中的 [type of style] 风格的 [主题]。” 如表 A1 所示，“[type of style]” 列列出了用于提示构建的常见风格类别。而 [subject] 和 [background] 则从表中对应的词池中随机选择。值得注意的是，为了更好地评估不同模型保持风格一致性的能力，我们确保同一提示中的 [主题] 和 [背景] 不具有相同的风格。

SCB 评估指标。我们提出了一种新的风格对齐评估方法，该方法采用目标检测将合成图像分割为主体和背景组件。每个部分均使用 CLIP 分数指标独立分析其预期风格描述。最终的平均分数反映了合成图像对所有元素的预期风格描述的整体保真度。如图 A1 A2 所示，我们的方法能够有效评估合成图像中的风格对齐。在图 A1 中，机器人融合了油画风格元素，而森林背景则展现了素描特征，导致较低的 CLIP 分数 (0.2344)，这表明由于风格污染而对齐度降低。相比之下，图 A2 显示出更清晰的风格区分——素描风格的机器人 (0.2818) 与油画背景的森林 (0.2493) ——产生了更高的平均对齐分数 (0.2656)。这种细致的方法使得能够精确测量不同图像元素的风格保真度。注意到我们选择裁剪方法而不是分割方法，是因为 Clip 无法准确识别部分空白的图像。

在本节中，我们提供了我们方法的额外生成结果，展示了该方法可以生成具有高审美视觉质量的图像，这可以在图 A3 中看到。

在本节中，我们介绍了作为我们提出方法基础的关键概念和技术。这些预备组成部分提供了对我们方法背后动机的更清晰理解，以及其设计中利用的技术机制。

Table A1. 风格组合基准

Type of style	Lego; Oil-painting; Cyberpunk; Sketch; Pixel-Art; Watercolor; Graffiti
Subject	Dog; Cat; Robot; Car; Unicorn
Background	forest; Space; Desert; City; Ruins

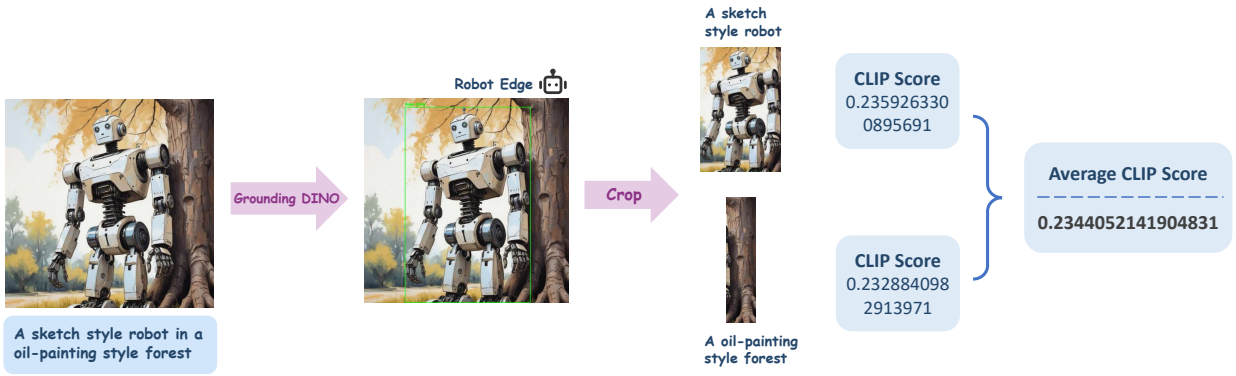

Figure A1. 在混合情况下测量风格准确性的流程。

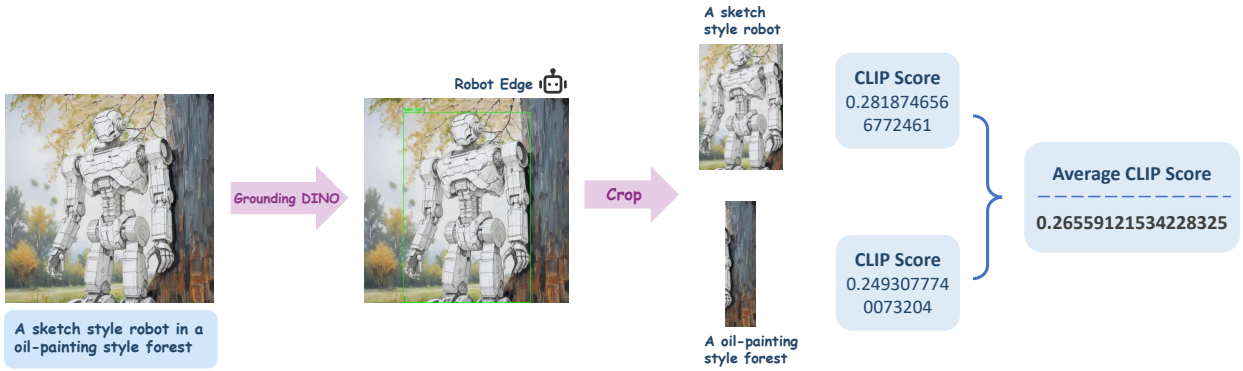

Figure A2. 我们的方法在不同环境下测量风格准确性的流程。

9. 为什么前三个 80 % 时间步自我扩展？

为什么选择 SA 映射？虽然许多近期的研究 [? ? ? ?] 表明自注意力映射携带结构信息，因此具备强大的布局保持能力，但基于交叉注意力映射的编辑方法 [?] 仍然具有显著影响。那么，为什么选择替换自注意力映射而不是交叉自注意力映射进行渐进编辑？在这里，我们可视化了在不同步骤替换自注意力映射（SA）和交叉注意力映射（CA）的差异。如图 A6 所示，基于自注意力映射的编辑往往提供更强的编辑能力，这与 FPE 中的发现 [?] 一致。

我们假设，基于交叉注意力进行编辑的相对较弱效果是因为它只替换或影响了被编辑对象的注意力图谱。相比之下，基于自注意力图谱的编辑更像是扩散模型布局的先验假设下的条件生成，这类似于 [?] 中描述的效果。通过这种方式，我们的框架 Detail++ 确保每个编辑步骤由精确的细节绑定关系提示引导，确保了精确的细节绑定，并防止出现溢出、不匹配和混合等问题。

为什么是 80%？在所有分支中扩展自注意力图的较高百分比可以更好地控制协调布局。然而，增加步骤数会削弱当前分支的原始生成能力，这可能导致图像的保真度较低。这个现象在图 A7 中有所说明。可以观察到，当 S 非常小时，每个分支输出的布局不一致，导致结构不一致。另一方面，当 S 非常大时，例如当 $S = T$ 时，图像质量下降，尤其是在主体的边缘，标记区域中会出现伪影。

在本节中，我们比较了不同层的选择用于生成二进制掩码。我们可视化了来自不同层的交叉注意力图（见图 A4 ），并在图 A5 中提供了额外的详细可视化。显然，从下、上和中间块提取 32×32 分辨率的交叉注意力图能生成最干净和最准确的目标掩码。

10. 与其他方法的结合

我们的方法还可以成功地与 AIGC 中常用的其他技术集成，如 Control-net [?] 。在本节中，我们展示了我们的方法与 Control-net 结合时的效率，如图 A8 所示。

A pale peach-colored fox mascot with sleek fur, a metallic silver jacket, a neon purple scarf, and a mint-green hat sits atop a dreamy turquoise log in the Forest of Dreams.

A magical deer, with iridescent fur of mint green, lavender and antlers adorned with glittering golden vines, stands in an enchanted forest of coral pink trees under a turquoise sky.

red teddy bear wearing a green tracksuit and a yellow hat is eating a blue cake.

A cream-colored bunny mascot with floppy ears, wearing a pastel teal scarf , rose gold sunglasses , and a lavender backpack , standing in a glowing coral-pink meadow under a twilight sky.

A **dog wearing sunglasses** and a **cat wearing necklace**.

Full body shot: a **man wearing necklace** and a **woman wearing sunglasses**

Figure A3. 我们的方法可以生成具有精确细节绑定和高保真度的图像。

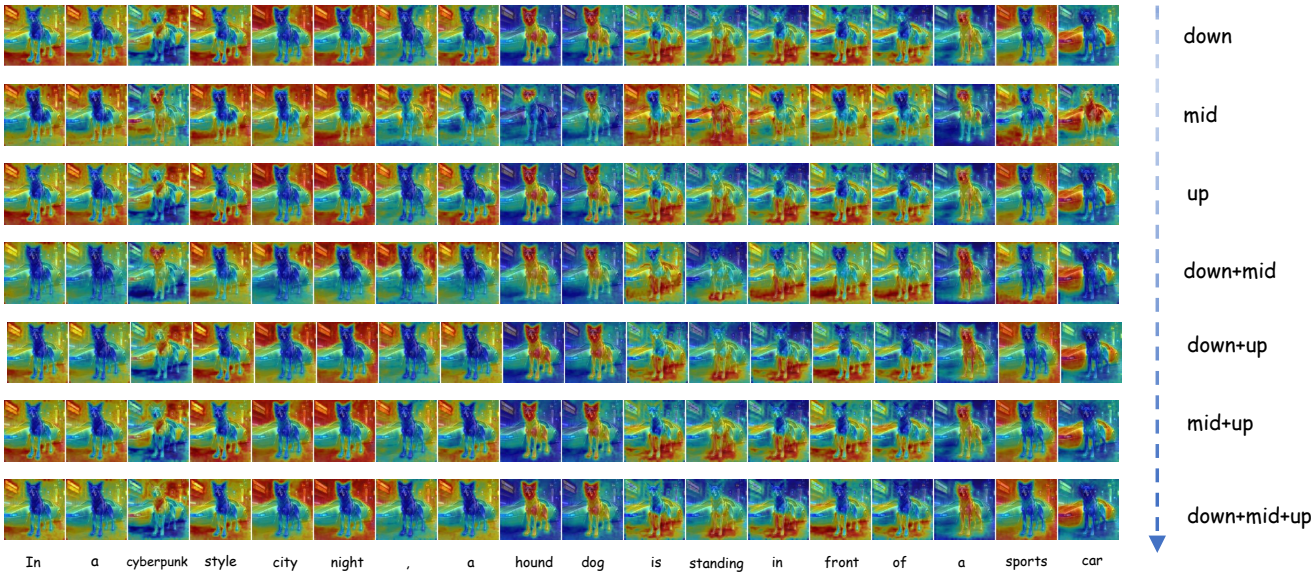

Figure A4. 不同层的交叉注意力图的可视化。“Down” 表示第一个下采样块中的交叉注意力层，而 “Up” 表示第二个上采样块中的交叉注意力层。由于具有 64 × 64 交叉注意力图的其他层需要显著更多的内存，因此在此省略以避免过多的 GPU 使用。

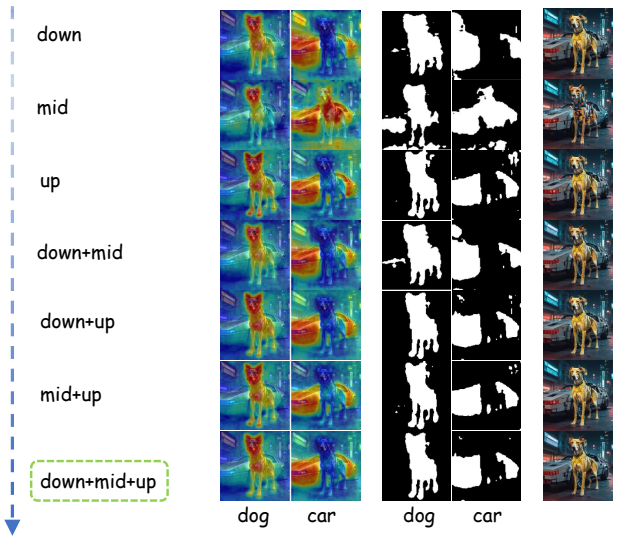

Figure A5. 比较在二进制掩膜提取中选择不同交叉注意力层的效果。可以观察到，从下游、中游和上游模块中提取 32 × 32 分辨率的交叉注意力图可以得到最干净且最准确的主体掩膜。

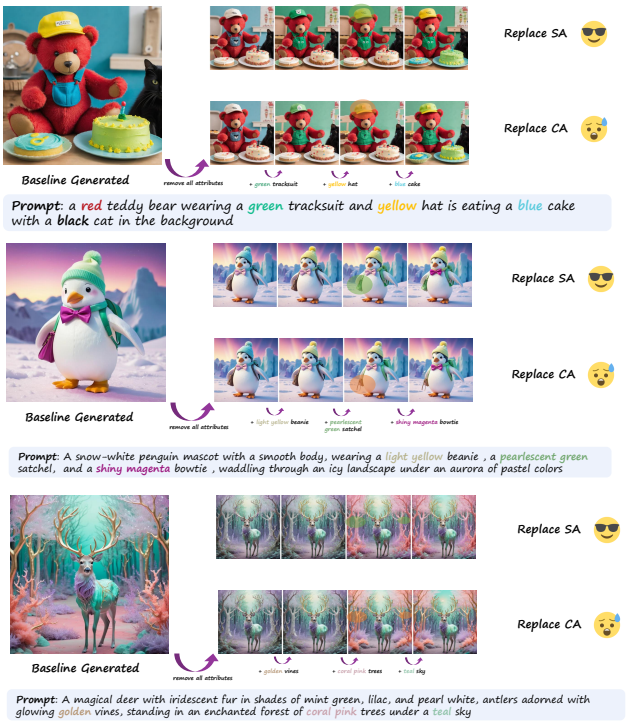

Figure A6. 图像生成中注意力图类型的比较分析。正如标记区域所示，替换交叉注意力图导致编辑效果较弱，使得在精确分配属性给主体时效果不佳。相比之下，替换自注意力图在同一场景中允许更精确的属性分配。

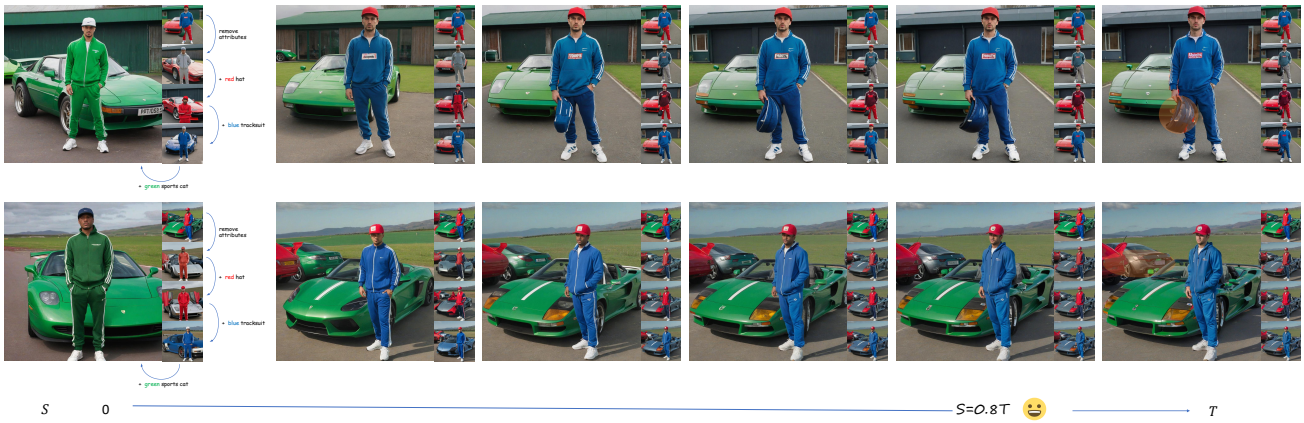

Figure A7. 不同注意力扩展时间步长的比较分析。两行图像均由提示“一个戴着红帽子和蓝色运动服的男人站在一辆绿色跑车前”生成。每个主图像右侧的子图代表渐进的生成过程，随着过程的推进，可以观察到布局的变化。

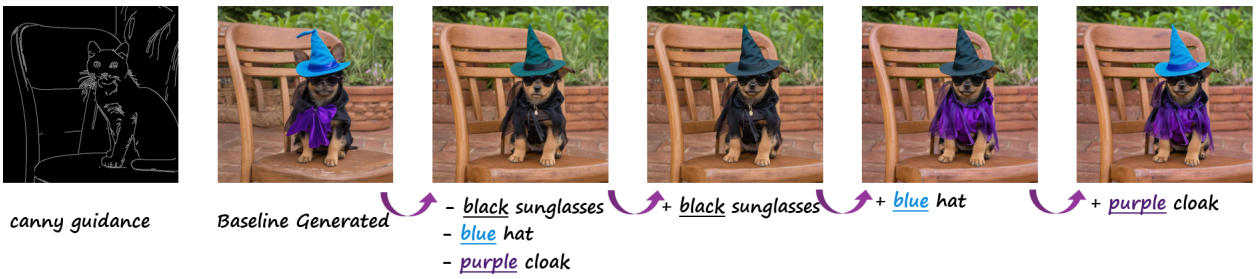

Figure A8. 我们的方法能够生成具有准确细节结合的图像。