arXiv:2305.01644v2 [cs.CV] 5 Jun 2024

面向文本到图像个性化的键锁定秩一编辑

YOAD TEWEL，英伟达、特拉维夫大学，以色列
RINON GAL，英伟达、特拉维夫大学，以色列
GAL CHECHIK，英伟达、巴伊兰大学，以色列
YUVAL ATZMON，英伟达，以色列



图 1。（左）*Perfusion* 支持生成个性化概念的图像，使其外观、姿势和背景发生显著变化，同时保持身份特征。每个概念仅需 100KB 的紧凑模型，且不牺牲保真度。（右）*Perfusion* 可在推理时组合已学习的概念，生成多个概念并列呈现的场景，甚至让它们彼此互动。

文本到图像模型（文生图）通过允许用户使用自然语言引导创作过程，提供了全新的灵活性。然而，要根据用户提供的视觉概念对这些模型进行个性化，使其与之相符，仍是一项具有挑战性的任务。文生图个性化面临多项艰巨挑战，例如在保持高视觉保真度的同时提供创作控制、在单张图像中组合多个个性化概念，以及控制模型大小。

我们提出了*Perfusion*，一种文生图个性化方法，通过对底层文生图模型进行动态秩一更新来应对这些挑战。*Per-fusion* 通过引入一种新机制，将新概念的交叉注意力键“锁定”到其上位类别，从而避免过拟合。此外，我们还开发了一种门控 秩一方法，使我们能够在推理时控制已学习概念的影响，并组合多个概念。借助单个经过训练、大小仅为 100KB 的模型，我们便能在运行时高效地平衡视觉保真度与文本对齐；该模型比当前最先进水平的模型小五个数量级。此外，无需额外训练，它就能覆盖帕累托前沿上的不同工作点。

最后，我们证明*Perfusion* 在定性和定量评估中均优于强基线方法。重要的是，与传统方法相比，键锁定能够带来新颖的结果，即使在单样本设置下，也能以前所未有的方式呈现个性化物体交互。

代码将在我们的项目页面[1]。

1 引言

文生图（T2I）个性化是指对基于扩散的文生图模型进行定制，使其能够处理全新的、由用户提供的视觉概念 [Cohen 等，2022；Gal 等人，2022 年；Kumari 等人，2022；Ruiz 等人， 2022]。用户首先提供少量该概念的图像示例；随后便可以使用自由文本，创作包含这些概念的新场景。这种工作流程可用于广泛的下游应用，从虚拟摄影、产品设计到生成个性化虚拟资产。

当前的个性化方法主要采用两种途径之一：要么通过文本编码器的输入端的词嵌入来表示一个概念 [Cohen 等，2022；Gal 等2022] ，要么微调整个基于扩散的去噪模块的权重 [Ruiz 等 2022]。遗憾的是，这些方法容易出现不同类型的过拟合。如下文所示，词嵌入方法难以泛化到未见过的文本提示词，这一点反映在其往往较低的文本对齐分数上。微调方法能够更好地泛化到新的文本提示词，但表达能力仍然不足；其文本和视觉对齐分数通常低于我们的方法，正体现了这一点。此外，微调方法通常需要大量存储空间，往往达到数百兆字节甚至数吉字节。最后，

[1]已被 SIGGRAPH 2023 接收

这两种方法都难以在单个提示词中组合分别训练的概念，例如泰迪熊*和茶壶*（图 1）。

在此，我们介绍“*Perfusion*”，一种旨在应对所有这些挑战的文生图个性化方法。它允许对概念进行富有表现力的变形，同时保持较高的概念保真度。它还支持在推理时组合概念，且模型大小很小——每个概念约为 100KB。为了实现这些目标，我们专注于基于扩散的文生图模型中的交叉注意力模块。

在典型的基于扩散的文生图中，输入的文本提示会通过文本编码器（如 T5 [Raffel 等人 2020] 或 CLIP [Radford 等人 2021]）转换为一系列编码。随后，作为交叉注意力模块的一部分，这些编码通过学习得到的投影矩阵映射为键和值。受 Balaji 等人 [2022]; Hertz 等人 [2022], 的启发，我们提出将这些投影的作用视为两条不同的路径：键（K）是“何处”路径，用于控制注意力图的布局，并由此控制生成图像的组合结构。值（V）是“何物”路径，用于控制图像组成部分的视觉外观。

我们的主要洞见是，现有技术在对 *Where* 路径过拟合时会失效（图 2、图 11），导致对新词的注意力超出概念本身的视觉范围。为了解决这一缺陷，我们提出了一种新颖的“键锁定”机制，即将某个概念的键固定到其所属超类别的键上。例如，表示某个特定泰迪熊的键可以被键锁定到泰迪熊这一超类别的键上，而不是锁定到该特定对象本身。直观而言，这使得新概念能够继承超类别的特征与创造能力（图 1）。随后，通过 *What*（V）路径实现个性化处理：我们将值投影视为扩展的潜在空间，并与输入词嵌入同步进行优化。

最后，我们将介绍如何通过对 K 和 V 投影矩阵的权重进行门控秩 1 更新，将这些组件直接整合到文生图模型中。这种更新的门控特性使我们能够在推理时有选择地仅对必要的编码应用秩 1 更新，从而组合多个概念。此外，同一门控机制还提供了一种调节所学概念强度的方法，这种强度体现在输出图像中。如此一来，无需为每个新的工作点专门训练模型，便可在运行时高效地权衡视觉保真度与文本对齐。实证结果表明，*Perfusion* 不仅能以一小部分模型大小实现更准确的个性化，还支持使用更复杂的提示词，并在推理时组合单独学习的概念。

总而言之，我们的贡献如下：首先，我们研究了当前个性化方法中观察到的过拟合现象，并提出了一种缓解该问题的“键锁定”机制。其次，我们提出了一种可控的网络秩一更新机制，仅需 100KB 的存储空间即可实现较高的物体保真度。第三，我们的方法能够高效地遍历单个训练模型的帕累托前沿，在运行时平衡视觉保真度与文本对齐，同时还能泛化到未见过的工作点。最后，我们证明 *Perfusion* 能够超越当前最先进的方法，并支持在推理时进行物体组合。

2 相关工作

基于扩散的文本引导合成。文生图生成的最新进展主要得益于预训练扩散模型 [Ho 等人 2020]，尤其是大规模模型 [Balaji 等人 2022; Nichol 等人 2021; Ramesh 等人 2022; Rombach 等人 2021]，这些模型在网络规模的数据上训练而成 [Schuhmann 等人 2021]。我们的方法扩展了这些预训练模型，使其能够描绘个性化概念。该方法应用于 Stable-Diffusion [Rombach 等人 2021],，但我们预计它也可以应用于任何采用类似交叉注意力机制的文生图生成器 [Saharia 等人 2022]。

文生图个性化 文生图个性化任务 [Gal 等人，2022; Ruiz 等人 2022] 旨在教会文生图生成模型合成特定目标概念的新图像，并通过自由文本进行引导。当前的个性化方法要么围绕固定嵌入微调去噪网络 [Ruiz 等人 2022]，要么优化一组词嵌入来描绘该概念 [Agrawal 等人 2021; Cohen 等, 2022; Daras 和 Dimakis 2022; Gal 等人，2022]2022]。[Kumari 等人 2022]提出了一项同期工作，该方法对去噪网络的 K 和 V 交叉注意力层进行微调，同时优化词嵌入。它采用闭式优化技术来组合概念。在 *Perfusion* 中，我们将 K 路径锁定到概念的超类别，并使用门控秩一编辑，而非微调及后续优化。由此获得了出色的新颖结果。我们独立于 [Kumari 等人 2022]，分别研究了 K 和 V 路径。

秩一模型编辑在自然语言处理领域，人们投入了大量精力来理解和定位大型语言模型的记忆机制。具体而言，研究发现，Transformer前馈层充当键值记忆存储器 [Geva 等，2022, 2020; Meng 等 2022]。近期，Bau 等 [2020]; Meng 等 [2022] 提出了 ROME，这是一种秩一编辑方法，通过更新这些关联记忆层来修改网络的事实知识。*Perfusion* 旨在将类似思路应用于文生图扩散模型。然而，正如 [Kumari 等 2022]所报告的那样，对扩散模型进行朴素的秩一编辑可能导致效果不佳。我们的方法解决了将这些方法应用于基于扩散模型的交叉注意力层时所面临的挑战。此外，*Perfusion* 能够通过动态门控机制组合多个秩一编辑，而非采用静态编辑。

基于文本的图像编辑。强大的多模态模型的出现带来了一系列基于文本的编辑方法 [Bau 等人 2021; Gal 等人 2021; Patashnik 等人 2021]。借助扩散模型，这些方法涵盖了单图像编辑方法 [Brooks 等人 2022; Kawar 等人 2022; Meng 等人 2021; Mokady 等人 2022; Valevski 等人 2022; Wu 等人 2022; Zhang 等人 2022]，以及图像修复任务 [Avrahami 等人 2022; Nichol 等人 2021; Ramesh 等人 2022; Yang 等人 2022]。与我们的工作最相关的是 Balaji 等人 [2022], 提出的文字绘画（PWW）方法和提示词到提示词（P2P） [Hertz 等人 2022]。PWW 在推理时将注意力图偏置到预定义的掩码。P2P 通过使用新提示词重新生成给定的生成图像，并在扩散过程中注入原始图像的注意力图，对其进行编辑。

过程。与这些方法不同，我们不编辑给定的图像，而是学习表示一种可在新提示词中调用的个性化概念。此外，我们不覆盖注意力图，而是约束新概念的交叉注意力键。这些键是影响注意力图的因素之一，但仍允许通过查询特征进行特定于概念的修改。

3 预备知识与符号说明

我们首先概述本文用于个性化的两种机制。第一种是文生图扩散模型中常见的交叉注意力机制 [Rombach 等人 2021]。第二种是近期提出的一种用于对大型语言模型进行秩一编辑的方法 [Meng 等人 2022]。

3.1 文本到图像模型中的交叉注意力

在当前基于扩散模型的文生图系统中（图3.A），输入的文本提示首先会转换为一系列词嵌入。随后，使用文本编码器（例如 CLIP [Radford 等人 2021]）将该序列转换为一系列编码。接着，每个编码都会通过两个交叉注意力矩阵进行线性投影：$\boldsymbol{W}_K$ 和 $\boldsymbol{W}_V$。这些投影的结果称为“键”和“值”。形式化地，令 M 为输入序列的长度，$\boldsymbol{w} \in \mathbb{R}^{M\times d_w}$ 为词嵌入序列，其中每个词嵌入的维度为 d_w，$\boldsymbol{e} \in \mathbb{R}^{M\times d_e}$ 为编码序列，其中每个编码的维度为 d_e。对于序列中的每个条目 $m \in M$，编码 $\boldsymbol{e}_m \in \mathbb{R}^{d_e}$ 会通过两个投影矩阵映射为一个“键”向量 $K_m = \boldsymbol{W}_K e_m \in \mathbb{R}^{d_k}$ 和一个“值”向量 $V_m = \boldsymbol{W}_V e_m$。与此同时，局部图像特征通过第三个矩阵 $\boldsymbol{W}_Q$ 进行投影，生成空间“查询”图。随后，这些查询会投影到键上，从而得到每个编码对应的注意力图：$A_m = \text{softmax}\left(QK_m^T/\sqrt{d_k}\right)$ [Rombach 等人 2021]。直观而言，该图会告知模型，m^{th} 词与图像各个空间区域的相关程度。最后，局部图像特征由“值”构成，并由这些图加权：$A \cdot V$（图3）。

3.2 秩-1 模型编辑

秩一模型编辑（ROME） [Meng 等人 2022] 最近提出了一种方法，用于编辑预训练语言模型（如 GPT）中的事实关联。ROME 会编辑网络中单个线性层 W 的权重，使该层在给定一个目标输入 $\boldsymbol{i}_*$ 时，输出一个目标输出 $\boldsymbol{o}_*$ [2]。为了编辑模型的事实知识，ROME 采用三个分别执行的步骤。(1) 查找目标输入 $\boldsymbol{i}_*$，即与第 W 层中经编辑词语（事实）相关联的输入。ROME 通过将不同提示词输入语言模型，并对词语所在索引处的激活值取平均，来确定经编辑词语的目标输入激活值。由此得到表征 $\boldsymbol{i}_*$。（2）查找目标输出 $\boldsymbol{o}_*$，通过针对特定目标优化该层的输出激活值来实现——例如，修改模型最终输出中呈现的事实。$\boldsymbol{o}_*$ 是在一个单一词语索引的输出上进行优化的。（3）更新该层 W，通过求解一个约束最小二乘问题，该问题具有闭式解

$$\hat{W} = W + \Lambda(C^{-1}\boldsymbol{i}_*)^T. \tag{1}$$

这里，$\Lambda = (\boldsymbol{o}_* - W\boldsymbol{i}_*)/[(\boldsymbol{i}_*^T(C^{-1})^T\boldsymbol{i}_*)]$，$C$ 是一个预先缓存的常数正定矩阵（附录C）。第 3 步中的更新是在找到 $\boldsymbol{o}_*$ 之后执行的，因此会影响所有序列编码，而不像第 2 步那样只影响单个被编辑的词。

通过将更新限制为局部的秩一变化，ROME 在不大幅改变模型知识的情况下，改变与单个事实相关的信息。我们的方法利用类似的机制来编辑文本到图像模型，并引入新的视觉概念。

4 方法

我们的目标是以富有表现力且高效的方式实现模型个性化。因此，一个自然的起点是研究该领域既有工作的局限性，尤其是文本反演 [Gal 等人 2022] 和 *DreamBooth* [Ruiz 等人 2022]的方法。我们注意到，这些方法，尤其是文本反演，容易出现过拟合：当包含已学习概念的提示词发生变化时，该概念会变得难以修改。在图 2 中，我们展示了这一问题源于注意力机制，因为新概念会将注意力吸引到超出其视觉范围的区域。图 A.3 提供了展示这一现象的其他示例。

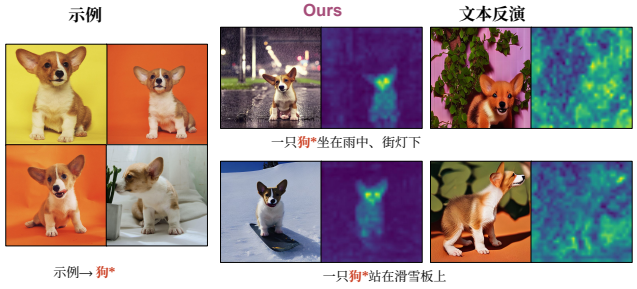

图 2. **注意力过拟合：** Textual-Inversion（文本反演，TI）中的典型过拟合现象，是由学习到的嵌入所对应的注意力接管整幅图像造成的。这里，我们将与“dog*”一词对应的注意力图可视化。文本反演注意力区域（右侧面板）分布在整幅图像中，而非聚焦于物体。这会导致生成过程忽略提示词的其余部分，只描绘“dog*”这一概念。

接下来，我们介绍 ***Perfusion***，这是一种通过秩一层编辑来解决该问题的方法。我们概述一种门控机制，它能在推理时提供更好的控制，并介绍如何利用它来组合独立学习的概念。

4.1 两个相互冲突的目标与一个朴素的解决方案

个性化文本到图像生成旨在实现两个目标：（1）避免对示例图像过拟合，使个性化概念能够以各种姿势、外观或情境生成；（2）保留生成图像中个性化概念的身份特征，即使其呈现为不同的姿势、外观或情境。这两个目标之间存在自然的权衡。对输入示例过拟合的方法往往能够保留身份特征，但无法匹配那些旨在将概念置于不同情境中的创意提示词。

[2]直观地说，目标输入 所扮演的角色类似于一把需要匹配的钥匙。为避免与变换器的 K 和 V 通路混淆，我们使用“目标输入”和“目标输出”，而不是 [Meng 等人 2022]所使用的“键”和“值”。

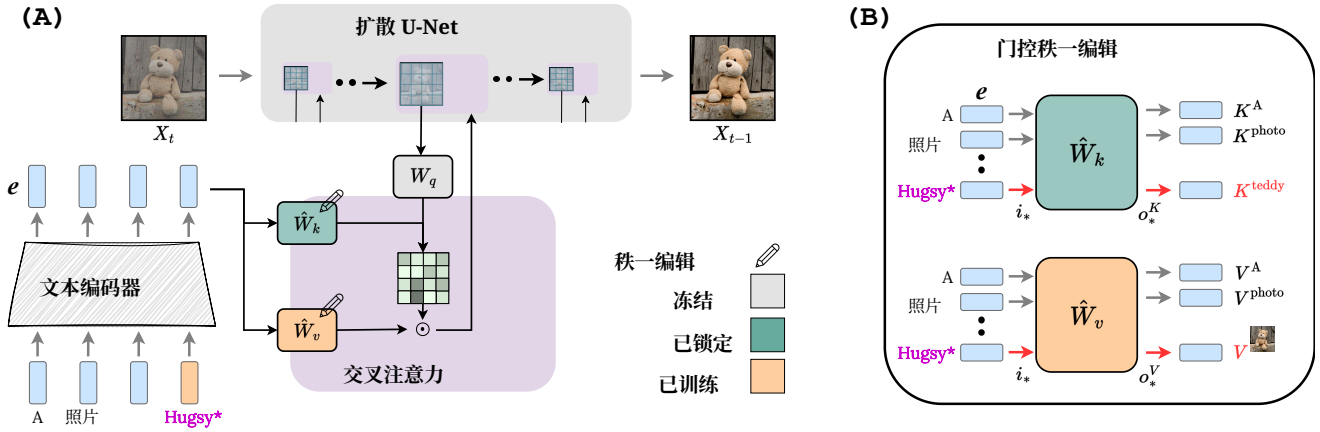

图 3。**架构概述（A）：** 提示词被转换为一系列编码。每个编码都会输入扩散 U-Net 去噪器的一组交叉注意力模块（紫色模块）。放大的紫色模块展示了键路径和值路径如何根据文本编码进行条件化。键驱动注意力图，随后由注意力图调制值路径。**门控秩一编辑（B）：** 顶部：K 路径被锁定，因此，任何到达 $\hat{}W_k$ 的 e_{Hugsy} 编码都会被映射到超类别 K^{teddy} 的键。底部：任何到达 $\hat{}W_v$ 的 e_{Hugsy} 编码都会被映射到 V^{Hugsy}，该值通过学习获得。此更新的门控特性使其能够仅应用于必要的编码，并提供调节所学习概念强度的手段，这一点体现在输出图像中。

“The”“*Where*”路径和“*What*”路径。为了同时改善这两个目标，我们的关键洞见是，模型需要将生成的内容与其生成位置解耦。为此，我们利用了第 3.1 节中所述的交叉注意力机制解释。 K 路径——与“键”相关的路径，与创建注意力图有关。因此，它充当控制物体在最终图像中所在位置的路径。相比之下， V 路径负责添加到每个区域的特征。从这个意义上说，它可以控制最终图像中出现的内容。因此，我们将 K 映射解释为“*Where*”路径，将 V 映射解释为“*What*”路径，而这一解释指导了我们提出的方法：

4.1.1 避免过拟合。在初步实验中，我们注意到，当使用数量有限的示例学习个性化概念时，*Where* 路径（$\boldsymbol{W}^K$）的模型权重容易对这些示例中的图像布局过拟合。图 2 展示了这一问题：个性化示例可能会“主导”整个注意力图，从而阻止其他词语影响合成图像。因此，我们旨在通过限制 *Where* 路径来避免这种基于注意力的过拟合。

4.1.2 保留身份特征。在 Image2StyleGAN 中，Abdal 等人 [2019]提出了一种分层潜在表示，以便更有效地捕捉身份特征。在该方法中，他们没有在生成器的输入空间预测单个潜在代码，而是在合成过程中为每种分辨率分别预测一个代码。鉴于 *What* (V) 路径激活值具有紧凑的特性，且底层 U-Net 去噪器具有多分辨率结构，我们提出将其视为类似的潜在空间。

4.1.3 朴素解决方案。 为同时实现这两个目标，可以考虑这一简单的解决方案：每当编码包含目标概念时，就确保其交叉注意力键 与其超类别的键相匹配，我们将此称为键锁定。此外，我们希望交叉注意力值 能够在多分辨率潜在空间中表示该概念。例如，如图 3（右上）所示，给定我们名为 Hugsy 的泰迪熊的图像示例，当编码包含“Hugsy”时，V 投影会输出概念特定代码，而 K 投影的目标是输出超类别 K泰迪熊的键。

实现这一想法的一种方法是简单地替换向量——只需交换分配给个性化概念索引处编码的键和值。然而，这种方法没有考虑文本编码器中的跨词信息共享。当编码到达去噪器的交叉注意力层时，其特征已经受到文本中其他词语特征的影响，同时也会反过来影响这些词语。我们希望确保实现能够考虑到这种影响，并正确修改所有受此影响词语的键和值激活值。

因此，一个自然的解决方案是使用 ROME 编辑交叉注意力层 $\boldsymbol{W}_V$和 $\boldsymbol{W}_K$的权重。具体而言，给定目标输入 iHugsy 时，我们强制 K 激活输出特定的目标输出 o^KHugsy = K泰迪熊。类似地，给定目标输入 iHugsy 时，我们强制 V激活输出学习得到的输出 o^VHugsy $= {}^V$Hugsy，见图3（右下）。现在，对于任意词语，如果其编码包含一个与 $\boldsymbol{i}_*$平行（对齐）的分量，那么它们的激活输出也会相应地改变。

遗憾的是，将 ROME 应用于这项任务会面临两个挑战：**挑战 1：** 使用 ROME 进行训练会导致训练与推理不匹配。这是因为在“步骤 2”的训练过程中，ROME 仅优化与提示词中某个特定条目 $\boldsymbol{o}_*$ 相关联的目标输出(m^{th}-索引）。然而，如上所述，在“步骤 3”中执行秩一矩阵更新时，预期该更新会影响提示词中其他词语的投影。事实上，我们观察到，这会导致训练测试不匹配，从而大幅降低重构概念的保真度。

挑战 2：类似的效应也阻碍了我们组合多个已学习的概念，因为它们对投影的影响并未充分解耦。此外，这些新概念对应多个目标输入 i_*，而这些输入本身也可能存在固有的纠缠（例如，概念之间存在相关语义）。这些因素共同导致在推理时尝试组合概念会产生视觉伪影。

为应对这些挑战，我们提出对齐 ROME 的训练与推理步骤，并引入一种新的门控机制。下面将介绍这两个部分。

4.2 用于个性化文生图的门控秩一模型编辑

通过端到端训练应对训练测试不匹配。为应对第一个挑战，即 ROME 的训练与推理不匹配，我们提出统一 ROME 的第二步和第三步。因此，目标输出优化与矩阵更新会在训练期间同步进行。网络会学习处理对其他提示词部分产生的任何影响，从而避免训练与推理不匹配。

为此，我们改写 ROME 的权重更新，以刻画层 $\hat{}W$ 在接收输入 e_m 时的输出 h。由此得到

$$h = We_m^{\perp} + o_* sim(i_*, e_m)/||i_*||^2_{C^{-1}}, \tag{2}$$

这里，$sim(i_*, e_m) := i_*^T (C^{-1})^T e_m$ 用于衡量 e_m 与 i_* 在由 C^{-1} [Atzmon 等人 2015]$||i_*||^2_{C^{-1}}$ 定义的度量空间中的相似度；$= sim(i_*, i_*)$ 用于衡量 i_* 在同一度量空间中的能量；$e_m^{\perp} := e_m - i_* \text{sim}(i_*, e_m)/||i_*||^2_{C^{-1}}$ 则是 e_m 在该度量空间中与 i_* 正交的分量。直观来说，公式（2）中的右侧加法项将词语编码（e_m）的 i_* 分量映射到 o_*。左侧项则消去词语编码中的 i_* 分量，并使用预训练矩阵 W 映射其余部分。更多细节见附录 A 和 B.1。

基于这一特性，我们通过使用公式（2）更新每一层，将每一层的前向传播替换为更新后的前向传播。这样可以确保训练和推理使用相同的更新表达式，从而消除二者之间的不匹配。结合对 i_*（第 4.4 节）的在线估计，这使得借助 ROME 进行端到端训练成为可能，而无需分别执行其 3 个步骤。

使用门控秩一更新来组合概念。在推理时组合单独学习的概念是一项艰巨的挑战。在初始实验中，我们测试了逐个添加概念的方法，即通过编辑 W，使用 $\hat{}W = W + \sum_{j=1..J} \Lambda^j (C^{-1} i_*^j)^T$ 作为公式（1）的一种变体。我们发现，即使提示词中只包含一个学习到的概念，这种方法也会引入视觉伪影。我们推测，这一问题源于单独学习的概念的不同i_*可能会相互干扰。例如，如果概念共享语义，它们可能具有相关的表征。理想情况下，仅当输入编码相对于该概念具有足够的能量时，才应进行概念更新，否则应减弱更新。这样，我们可以确保模型更新仅应用于相关概念，而不会影响其他概念。

为应对这一挑战，我们使用一种门控 机制，有选择地允许或削弱各个概念对层输出的影响。这里需要指出，式 (2) 的更新规则已包含一种线性 门控机制相似度$(e_m, i_*)/||i_*||^2_{C^{-1}}$，当 时其值接近 1。然而，较低的相似度值可能无法得到充分削弱。因此，我们提出通过 Sigmoid函数对相似度 进行变换，将相似度$(e_m, i_*)/||i_*||^2_{C^{-1}}$ 的值输入其中；该函数包含偏置和温度超参数。这样，权重更新便会集中作用于与个性化概念高度对应的输入。

因此，在单一概念的训练和推理时，每一层更新的前向传播为

$$h = We_m^{\perp} + o_* \sigma \left(\frac{sim(i_*, e_m)/||i_*||^2_{C^{-1}} - \beta}{\tau} \right) \tag{3}$$

其中，σ 是 Sigmoid激活函数，τ 是温度，β 是偏置项。此实现确保权重更新仅应用于与 i_*对齐（平行）的编码分量，即属于新概念或受其影响的分量。

因此，这种非线性门控机制带来**两项重要优势**：首先，它使我们能够在推理时更好地分离各个已学习概念的影响。其次，即使对于单一概念，它也能在推理时控制该概念的影响。通过调整 Sigmoid 超参数（偏置和温度）的值，我们可以在视觉保真度与文本对齐之间进行权衡，反之亦然。

下一节将进一步说明如何推广这一公式，以组合多个概念。

4.3 推理

单一概念：对于单个已训练概念的推理，我们只需将式（3）应用于每个经过编辑的交叉注意力层的前向传播。通过在推理时更改 Sigmoid 的 τ和 β的值，我们可以控制所描绘概念的强度。

组合多个概念：为了组合单独训练的概念，我们将公式 Eq. (3) 扩展为包含多个概念 $\{i_*^j,\ o_*^j\}_{j\in 1..J}$。为此，我们首先将 $e_m^{\perp}$泛化为与度量空间中所有 $\{i_*^j\}_{j=1..J}$ 张成的子空间正交，我们将其记为 $e_m^{\perp J}$。对于右侧项，我们只需对所有概念的门控响应求和。最终表达式为：

$$h = We_m^{\perp J} + \sum_{j\in 1..J} o_*^j \sigma \left(\frac{sim(i_*^j, e_m)/||i_*^j||^2 - \beta_j}{\tau} \right), \tag{4}$$

其中 $e_m^{\perp J} = e_m - \sum_{j\in 1..J} u_j$相似$(u_j,\ e_m)$，$u_j$与度量空间中的一个基向量相关联，该向量经逆柯 Cholesky 根 L^{T-1} 投影回文本编码器空间。$e_m^{\perp J}$ 的推导见附录B。

4.3.1 全局键锁定：键锁定可确保概念的键与其超概念正确对齐。然而，这并不能确保文本编码器以处理超概念及其与编码中其他词语之间相关性的相同方式处理该概念。我们还研究了一种推理时方法，用于将键锁定的概念与整个提示词对齐。我们将此变体称为全局 键锁定，并将我们的基础机制称为局部键锁定。详情见附录C。

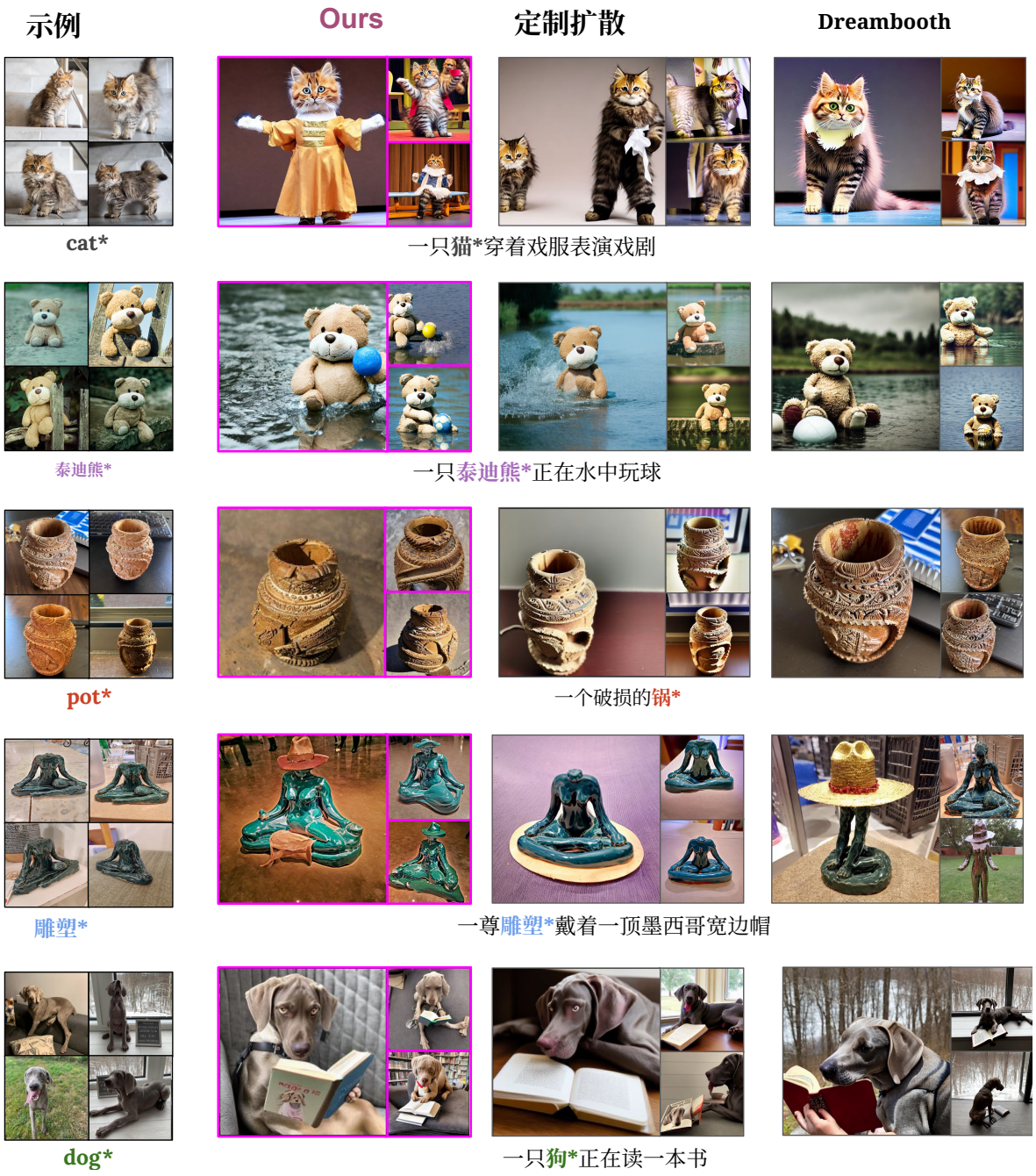

图 4. **单一概念示例的生成结果**。对于每个概念，我们展示了训练集中的示例，以及生成图像、对应的条件文本，并与 Custom-Diffusion（CD）和 Dreambooth（DB）基线方法进行比较。*Perfusion* 能够生成更具生命力的结果，提示词匹配效果更好，也较少受到原始图像背景特征的影响。特别值得注意的是，我们的猫（上方）的服装和戏剧化效果有所改善；当指示我们的狗读书时，其目光和姿势也更符合提示词（下方）。对于某些提示词，基线方法只是照搬训练集中的内容（例如罐子）。

4.4 实现细节

在线估计 i_*： 我们使用以下指数移动平均表达式，在训练期间估计 i_*：$i_* := 0.99 i_* +$ 0.01e概念，其中 e概念对应于文本编码器输出端概念词的编码。



图5. **多概念示例的更多生成结果**。我们展示了成对交互的概念，并与 CD 进行比较。除了 teddy* 提示词外，所有提示词均来自 CD 论文，并使用论文提供的图像。在 teddy* 示例中，*Perfusion* 将其描绘成戴着太阳镜*的样子，而 CD 则省略了太阳镜*。在水彩画中，*Perfusion* 更好地保留了椅子的形状。在 table* 示例中，*Perfusion* 更好地保留了桌子的颜色。

伪代码：附录D提供了*Perfusion*的秩一编辑模块的伪代码。

零样本加权损失：使用少量图像示例进行训练容易从图像背景中学习到虚假相关性。为了消除概念与其背景之间的相关性，我们使用通过零样本图像分割模型 [Lüddecke 和 Ecker 2021] 获得的软分割掩码，对标准条件扩散损失进行加权。掩码值通过其最大值进行归一化。

应用 *Perfusion* 于多个层： 与 [Gal 等人2022], 类似，对于每个概念，我们为一个超类别名称选择一个单词。我们使用该单词初始化其词嵌入，并将这些词嵌入视为可学习参数。我们将 *Perfusion* 编辑应用于 UNet 去噪器的所有交叉注意力层。对于每个 K 路径层（l），我们预先计算并冻结 $\boldsymbol{o}^K_{:,*^l}$，使其成为 $\boldsymbol{o}^K_{:,*^l} = \boldsymbol{W}^l_K \boldsymbol{e}$超类别，并使用提示词“A photo of a <超类别_word>”，同时随着训练的进行更新 $\boldsymbol{i}_*$。在每个 V 路径层中，我们将 $\boldsymbol{o}^V_{:,*^l}$视为可学习参数。

训练细节：我们以 0.03 的学习率训练 $\boldsymbol{o}_*$，嵌入的学习率设为 0.006。我们使用批次大小为 16，并采用 Flash-Attention [Dao 等人 2022；Lefaudeux 等人 2022]。我们只使用翻转增强 $p = 50\%$。不对不对称物体进行翻转。我们使用包含 8 个提示词的验证集，每 25 个训练步数采样一次，并选择使 CLIP 图像相似度分数与 CLIP 文本相似度分数的调和平均数最大化的模型步骤。我们在实验细节中更详细地介绍 CLIP 指标。为对生成进行条件控制，我们从 CLIP ImageNet 模板 [Radford 等人 2021] 中随机采样中性上下文文本。模板列表见补充材料。

我们的方法在单个 A100 GPU 上训练，平均训练 210 步，耗时 ~4 分钟；最多训练 400 步（ ~7 分钟）。与同期工作 [Kumari 等人 2022] 相比，这种训练所需的计算量减少了 ×2 – ×3，并使用 27GB 内存。

Sigmoid 超参数：在训练时，我们将 S 形函数偏置和温度设为 $b = 0.75$，$T = 0.1$。在推理时通常使用温度0。15，偏置设为 0.6 – 0.75 以进行局部键锁定，或设为 0.4 – 0.6 以进行全局键锁定。

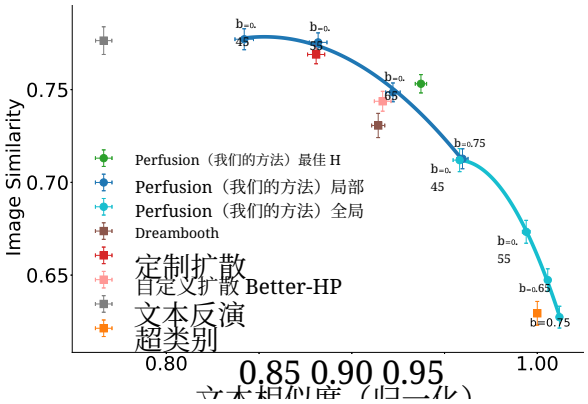

图 6. **视觉-文本相似度平面：**仅使用一个经过训练的 100KB 模型，并在运行时选择参数，*Perfusion*（蓝色和青色）即可覆盖帕累托前沿。误差线表示 95% 置信区间。

5 实验

我们证明了 *Perfusion* 优于强基线方法。我们进行了定性比较和定量评估，结果表明，尽管仅使用了部分参数量，它仍能覆盖视觉保真度与文本对齐的帕累托前沿，并且在面对更复杂的提示词时，能够取得更高保真度的结果。接下来，

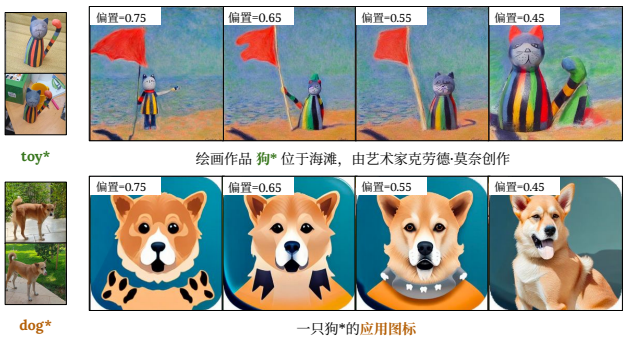

图 7. **平衡视觉保真度与文本对齐：** 控制 S 形函数的偏置阈值，可以平衡视觉保真度与文本对齐之间的权衡。较高的偏置值会减弱概念的影响，而较低的偏置值则会增强其影响。



图 8。**单样本训练结果：**使用单张图像进行训练时，我们的方法能够生成兼具高视觉保真度和文本对齐度的图像。量化结果见图 9(a)。

我们通过消融研究（第 5.3 节）研究 *Perfusion* 的属性。

对比方法：(1)*Perfusion*：如第 4 节所述的方法。我们为每个概念使用一个单独 训练的模型，但展示了采用不同 S形函数偏置以及局部 和全局 锁定的结果。这些参数调整均在运行时进行。**(2) *Perfusion* 最佳 H：**对于每个类别，我们会自动选择文本和视觉相似度调和平均数最高的运行时变体（见指标）。**(3) DreamBooth** [Ruiz 等人 2022]：一种最先进的方法，对去噪器 U-Net 的所有参数进行微调。我们采用 Patil 和 Cuenca [2022] 的实现。**(4) 文本反演（TI）** [Gal 等人 2022]：一种最先进的方法，仅优化词嵌入。我们采用官方仓库中的 Stable Diffusion 实现，并使用作者为 LDM 报告的参数 [Rombach 等人 2021]。**(5) Custom-Diffusion（CD）**：一项同期工作，训练 K 和 V 交叉注意力路径以及词嵌入。我们采用官方实现和超参数。**(6) Custom-Diffusion Better-HP：**采用 200 步训练的 CD，我们发现这能提高文本相似度得分。**(7) 超类别：** 一种纯文本基线；我们将概念词替换为其超类别。所有方法均应用于预训练的 Stable Diffusion 检查点 $v1.5$ [Rombach 等人 2021]。

数据集：概念：为进行公平且无偏的评估，我们采用了先前论文中的概念：来自 CD 的 6 个概念、来自 TI 的 2 个，以及来自 DB 或与 DB 类似的 3 个，共计 11 个个性化概念。这些概念分属四组：4 个玩具/小雕像和 3 只宠物（归为“动画”组）、2 个容器、1 件家具和 1 个可穿戴配饰。提示词：我们使用两类提示词。第一类是所有概念共用的 19 个提示词。这些提示词只改变场景，不会使概念变形。我们称之为“保持形状”提示词。第二类是每组提示词。这些提示词与概念所属的组别更为具体，并且往往会导致概念外观发生变形，例如“A broken pot*”或“A cat* is acting in a play wearing a costume”。总计有 86 个独特提示词，平均每个类别有 43 个提示词。在这 86 个独特提示词中，有 42 个是从 Custom-Diffusion 使用的提示词中随机选出的。

评估指标：与文本反演（TI）一样，我们在二维平面上报告结果，以展示视觉相似度与文本相似度之间的平衡。但与仅使用诸如“一张 S^*的照片”之类提示词来评估图像相似度的文本反演（TI）不同，我们使用所有保持形状不变的提示词，以同时衡量在新场景中的概念保真度。。

我们计算以下评估指标：**（1）图像相似度** 是概念图像与通过保持形状的提示词生成的生成图像之间 CLIP 余弦相似度的平均值。**（2）文本相似度** 是所有生成图像与其文本提示词之间 CLIP 相似度的平均值，其中我们省略占位符 S^*（即“ A 穿得像一位巫师”）。为了校准不同提示词之间的差异，我们将文本分数与使用超类别词而非学习词生成的图像分数进行比较，并据此对文本分数进行归一化。 所有相似度分数均按“类别”均衡 [Samuel 等人 2020]。具体而言，我们计算每个概念类别的平均分数，然后对所有类别分数取平均值。对于每个概念，我们为每个提示词采样 8 张图像，使用 50 个 DDIM 步数和 6 的引导尺度。

5.1 结果

在图 6 中，我们通过一个展示视觉相似度与文本相似度之间权衡关系的平面来说明结果。*Dreambooth* 成功地在视觉保真度和文本保真度之间取得了平衡。然而，与*Perfusion* 和 CD 相比，它在这两个指标上的得分都较低。*Textual-Inversion* 难以泛化到新的提示词，因此文本相似度得分较低。*Custom-Diffusion* 倾向于偏重视觉相似度，即使这会导致对目标的过拟合。我们观察到，只需对超参数进行小幅调整，*Custom-Diffusion Better-HP*便可以转而偏向文本相似度。*Per-fusion* 优于这些基线方法，并推动了帕累托前沿向前扩展。值得注意的是，我们可以通过推理时的参数修改覆盖该前沿，从而使用户能够根据所需的特性控制这种权衡。在实践中，这使我们能够轻松地为每个类别选择最佳工作点，从而获得更高的性能（*Perfusion Best H*）。请注意，*Perfusion* 仅需 100KB 的每个概念的参数即可实现这些提升，而 DB 需要数 GB，CD 则接近 100MB。**Importantly**，该比较表明，键锁定能够显著提升文本相似度，同时不会明显损害视觉相似度。

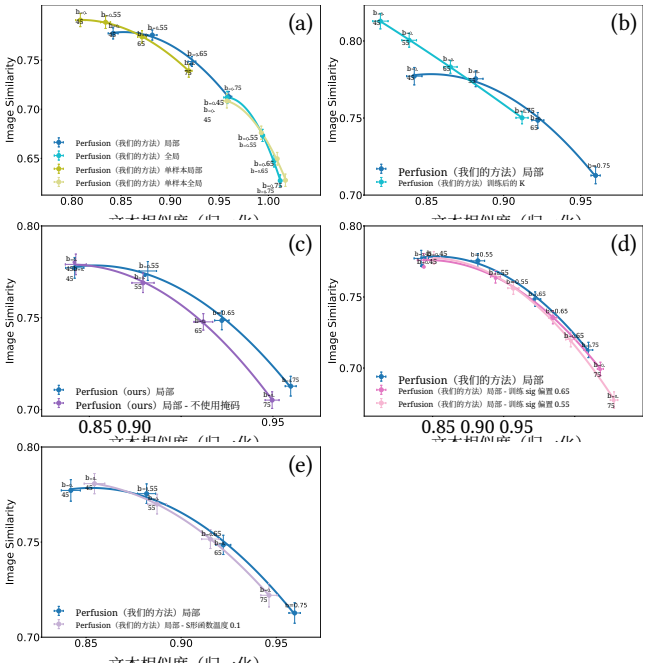

图 9. **消融研究：** 我们展示了不同消融条件下的帕累托前沿。**（a）单样本训练：** 我们的方法针对每个概念仅使用一张图像进行训练。**（b）训练后的 K：** 我们也训练键路径，而不是将其锁定。**（c）训练期间使用零样本掩码：** 我们比较使用和不使用零样本掩码时的 *Perfusion*。**（d）训练期间的 Sigmoid 偏置：** 我们比较使用不同 Sigmoid 偏置值进行训练的效果（默认值为 0.75）。**（e）推理期间的 Sigmoid 温度：** 我们比较 Sigmoid 温度为 0.1 的推理效果与默认温度 0.15 的推理效果。

5.2 用户研究

我们进一步通过两项使用亚马逊土耳其机器人开展的用户研究对模型进行评估。在第一项研究中，评分者会获得某个概念的两张图像和一个提示词，并被要求根据图像按照该提示词呈现概念的程度，对三种方法（*Perfusion*、CD、DB）生成的图像进行排序。我们使用了 11 个概念，每个概念包含 24 个提示词，每个提示词收集 8 份回答。*Perfusion* 以平均排名 2.18 ± 0.02 (SEM) 位居第一，CD 位居第二（2.06 ± 0.02），DB 位居最后（1.75 ± 0.02），这表明用户更偏好我们的方法。在第二项研究中，我们考察了 *Perfusion* 是否会损害生成先验。为此，我们将 *Perfusion* 与“原版”Stable-Diffusion（SD）进行了比较。评分者会看到一个提示词和两张生成图像，其中一张由 *Perfusion* 生成，另一张由 SD 生成。他们需要根据图像的真实感对其进行评分，分数范围为 1 至 3 分（3 分最佳）。我们收集了与第一项研究相同数量的概念、提示词和回答。*Perfusion* 的平均得分为 1.885 ± 0.017 (SEM)，而 SD 的得分为 1.894 ± 0.017。两者的结果在统计上无法区分，这表明 *Perfusion* 能够保留生成先验。有关这些实验的更多细节，请参见附录 I。

5.3 消融研究

在图 9 中，我们通过消融研究深入探讨 *Perfusion* 的特性。我们展示了以下条件下视觉相似度与文本相似度之间的权衡。

（a）**单样本：** 我们比较使用全部训练示例（每个概念平均 6.5 张图像）训练我们的方法，与每个概念仅使用一个示例进行训练的效果。我们观察到，仅使用一个示例进行训练会导致轻微过拟合。

(b) **键锁定：** 我们比较了采用键锁定的方法与采用训练后的键投影层的方法。显然，键锁定会使帕累托曲线向右移动，这意味着过拟合更少。此结果证实了我们的假设：锁定键投影层能够实现更好的文本对齐，并使学习到的概念能够发生复杂形变。

(c) **零样本掩码损失：** 我们比较了使用和不使用零样本掩码进行训练的效果。我们发现，使用零样本掩码往往能提升文本相似度，这意味着它有助于减少过拟合。(d) **Sigmoid 训练偏置**：我们比较了训练时使用的不同 Sigmoid 偏置值。我们发现，较高的偏置会带来更好的帕累托前沿。(e) **Sigmoid 推理温度** 我们比较了推理时使用的不同 Sigmoid 温度值。我们发现，推理时使用高于训练时的 Sigmoid 温度，能够带来更好的帕累托前沿。总体而言，0.15 的温度通常效果更好。

6 定性视觉比较

接下来，我们将通过定性比较展示我们方法的优势，并给出主要失败模式的示例。

单一概念文本引导合成： 图 1、图 4 和图 15 展示了我们使用 *Perfusion* 构建新颖场景的能力，并将其与强基线方法进行比较。对于每个概念，我们展示了训练集中的示例，以及生成图像和相应的条件文本。我们的方法能够改变概念的外观，同时保留其身份特征；与此同时，还能准确呈现概念和提示词的语义特质。例如，请注意，我们可以为图 1 中的泰迪熊*概念和图 4 中的猫*换上不同的服装，同时不损害它们的身份特征，并让它们与场景互动。我们还可以将茶壶*的材质改为纯金或透明玻璃，同时保留其鲜明特征。在图 4 中，只有 *Perfusion* 能够将锅*概念打碎。请注意，它还能改变狗*的姿势，使其看起来仿佛正在读一本书，目光落在文本上，爪子则握着书本。在图 A.2 中，我们提供了更多结果，其中包括文本反演。

多概念文本引导合成： 图 1 和图 5 展示了我们使用 *Perfusion* 合成包含多个概念的新颖场景的能力，并将其与 CD 通过“优化”结合单个概念的方法进行比较。在使用其论文中的提示词进行比较时，我们采用了他们提供的图像，否则



图 10. **比较锁定类型：**全局键锁定能够带来更多视觉变化，并能准确呈现物体或活动的细微差别，例如描绘以类人姿势阅读一本书或穿着厨师服的猫。局部键锁定也有成功之处，但效果不如全局键锁定。最后，训练式 K与训练图像的兼容性更好，但这是以牺牲其与文本的对齐为代价的。



图 11。**键锁定对训练动态的影响：**比较训练后的 K （未锁定）与局部键锁定 在保真度与文本对齐平面上的表现，其中“s”表示训练步骤。训练后的 K 会对训练图像过拟合，但会对文本对齐产生负面影响。相反，局部键锁定 训练表现出“过度泛化”，从而改善了文本对齐，但降低了视觉保真度。随着训练的进行，定性示例显示模型逐渐呈现出更多玩具特征。这表明，V 特征学习了超类别的潜在属性，以改善 Q 和 K 特征之间的对齐。最后，本例中的猫玩具展示了一个失败案例，因为 V 特征在训练过早阶段就捕捉到了玩具超类别。

使用“Better-HP”时。注意*Perfusion* 如何在不同场景中组合泰迪熊*和茶壶*，或者如何让茶壶*在帆船航行时托着泰迪熊*。与CD相比，我们观察到相似甚至更好的结果。例如，观察水彩画如何更好地保留椅子身份特征，或者泰迪熊*如何成功地戴上太阳镜*。

平衡视觉保真度与文本对齐：图 7 提供了定性示例，展示如何通过调整 Sigmoid 偏置阈值来平衡视觉保真度与文本对齐。较高的偏置值会减弱概念的影响，而较低的值则会使其在生成图像中更加突出。这是因为概念能量分散在文本编码器的多个编码中，而

不仅仅是与概念词对应的编码。降低偏置会增强概念对所有相关编码的影响。

接下来，我们展示键锁定机制的几个方面。首先，我们比较全局键锁定、局部键锁定和不进行键锁定的情况。然后，我们展示将概念锁定到不同超类别时会发生什么。最后，我们研究局部键锁定的训练动态，并展示一种“过度泛化”现象，它会使模型过度对齐到超类别。

6.1 键设计决策

接下来，我们展示并讨论键的设计决策。首先，我们比较全局键锁定、局部键锁定和不进行键锁定的情况。然后，我们展示使用原始 ROME、存在训练与推理不匹配，以及仅在优化步骤之后执行权重更新时会发生什么。

比较锁定类型： 图 10 比较了全局键锁定、局部键锁定 和训练式 K （不进行键锁定，K 路径的训练方式与 V 路径相同）。我们发现，与局部键锁定和训练式 K 相比，全局锁定能够生成丰富的场景，更好地表现物体属性或活动的细微差别，并且总体上能让概念呈现更多视觉变化。例如，请注意图中的猫以类人姿势读书，或穿着厨师服的样子。局部锁定也有一些成功案例，但效果较弱。最后，训练式 K 与训练图像中的姿势和外观更加一致，但牺牲了与文本的一致性。

使用原始 ROME 时的训练与推理不匹配： 图 A.1 展示了使用原始 ROME 进行编辑时，训练和推理过程中生成图像之间的不匹配。这会导致推理时生成的图像受损。

6.2 鲁棒性分析

接下来，我们将展示我们的方法在不同情况下的表现，并突出其优势与不足。

单样本学习： 我们将使用所有示例训练我们的方法（每个概念平均使用 6.5 张图像），与每个概念仅使用一个示例进行对比。图 8 和图 9(a) 显示，仅使用一个示例训练时会出现轻微过拟合。

向微调检查点进行零样本迁移： 使用原始扩散模型训练的 Perfusion 概念能够泛化到该模型的微调变体。图 12 展示了向 Stable-Diffusion 的两种热门变体迁移的能力：InkPunk-v2 [Envvi 2022] 以及 Protogen-v3.4 [darkstorm2150 2022]。

未经筛选的样本： 图 14 显示，批次大小为 8 通常足以确保获得几个优质样本。

锁定到非常规超类别 图 13 展示了锁定到非常规超类别时各个概念的呈现方式。我们观察到，这些概念会“继承”非常规超类别的定性轮廓。当罐子被锁定到鞋或时钟时，它会呈现出鞋或时钟的轮廓，但保留罐子的风格。当猫被锁定到灯时，它会变得明亮。

键锁定对训练动态的影响： 图 11 左侧面板比较了在不同训练步数（“s”）下的训练后的 K 与局部键锁定训练。随着训练的进行，训练后的 K （无锁定）会对训练图像过拟合，

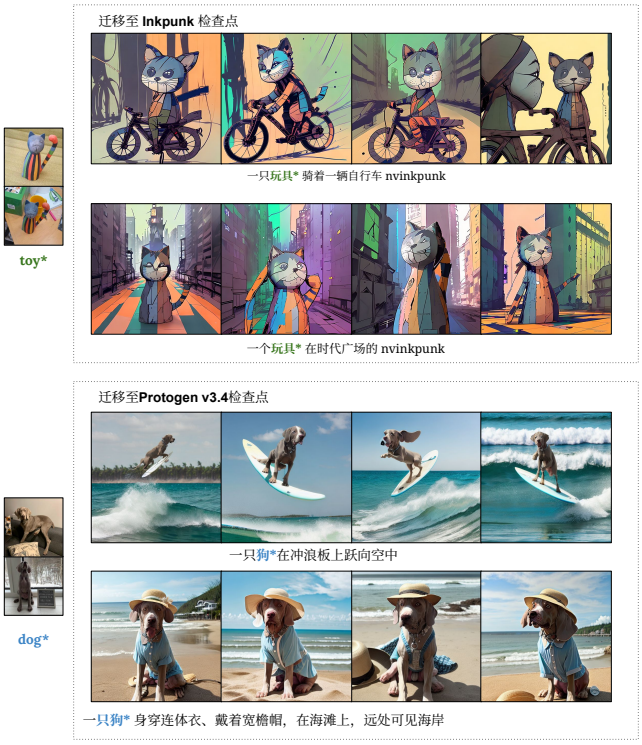

图 12. **向微调模型进行零样本迁移：** 使用标准扩散模型训练的 Perfusion 概念能够泛化到微调变体。



图 13. **锁定到非常规超类别：** 一个概念会“继承”其所锁定的超类别的定性轮廓。例如，将罐子锁定为鞋或时钟，而猫锁定为灯时则会发光。

尽管损害了文本对齐。有趣的是，局部键锁定训练揭示了一种“过度泛化”现象。随着训练的进行，概念会学习超类别（潜在）特征，从而改善文本对齐，但牺牲了视觉保真度。图 11 右侧面板展示了定性示例，其中，更长时间的训练会使“在外太空中的”这一概念呈现出越来越强的玩具化特征。值得注意的是，当使用重建提示词训练该概念时，其视觉保真度更高。这表明所学习的超类别特征是潜在的。这些发现是符合预期的，因为所学习的 V 特征会传播到 Q 特征中。

在下一个去噪步骤中，计算注意力图时，Q 特征应与锁定的超类别 K 特征对齐，因为它们之间存在内积。因此，我们预期 V 特征会学习编码超类别的潜在属性，以改善 Q-K 对齐。最后，猫-玩具示例展示了一种失败情况：在训练过程中，概念尚未学会其细粒度特征之前，V 特征就过早地捕捉了超类别特征。

7 结论与局限性

我们提出了 *Perfusion*，一种新颖的文生图个性化方法，将高视觉保真度与改进的文本对齐相结合。我们的方法采用门控秩一方法，能够控制已学习概念在推理时的影响，实现多个概念的组合，并使模型大小保持较小。重要的是，与传统方法相比，键锁定技术带来了新颖的定性结果。

局限性与未来工作。我们发现，选择要锁定的超类别词，有时会在新提示词中使用该概念时产生“过度泛化”效应。例如，如图 11 所示，将玩具猫这一概念设置为“玩具”，可能会促使模型在新的提示词中学习以幼稚风格生成它，同时牺牲其视觉保真度。此外，图 13 表明，将概念锁定到非典型超类别会导致这些概念呈现该非典型超类别的某些特征。第二个局限性是，组合概念需要大量的提示工程。有趣的是，我们发现，在由 CD 提出的提示词上更容易取得成功，而使用我们自行设计的提示词则较为困难。我们认为，这项任务仍有很大的改进空间。

致谢

我们感谢 SIGGRAPH 2023 的匿名审稿人提供宝贵意见，这些意见极大地提升了论文的最终版本。此外，我们还要感谢 Assaf Hallak、Xun Huang、Jim Fan 和 Eli Meirom 对本稿早期版本提出的深刻见解。本研究是第一作者攻读博士学位期间部分完成的工作。

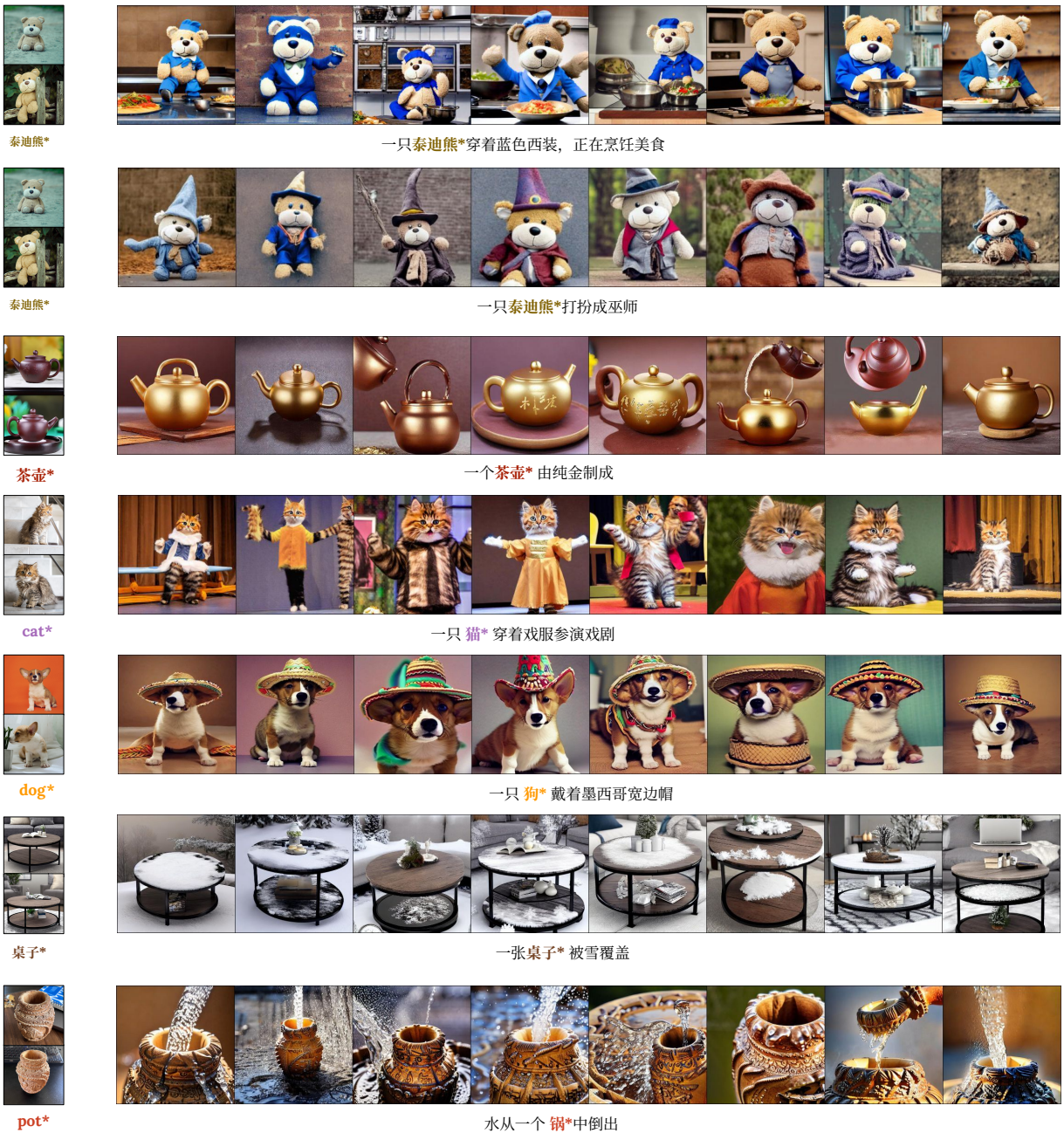

图 14. **使用文本引导提示词生成的图像变体的未经筛选的样本**。 我们发现，批次大小为 8 通常足以确保获得多个优质样本。

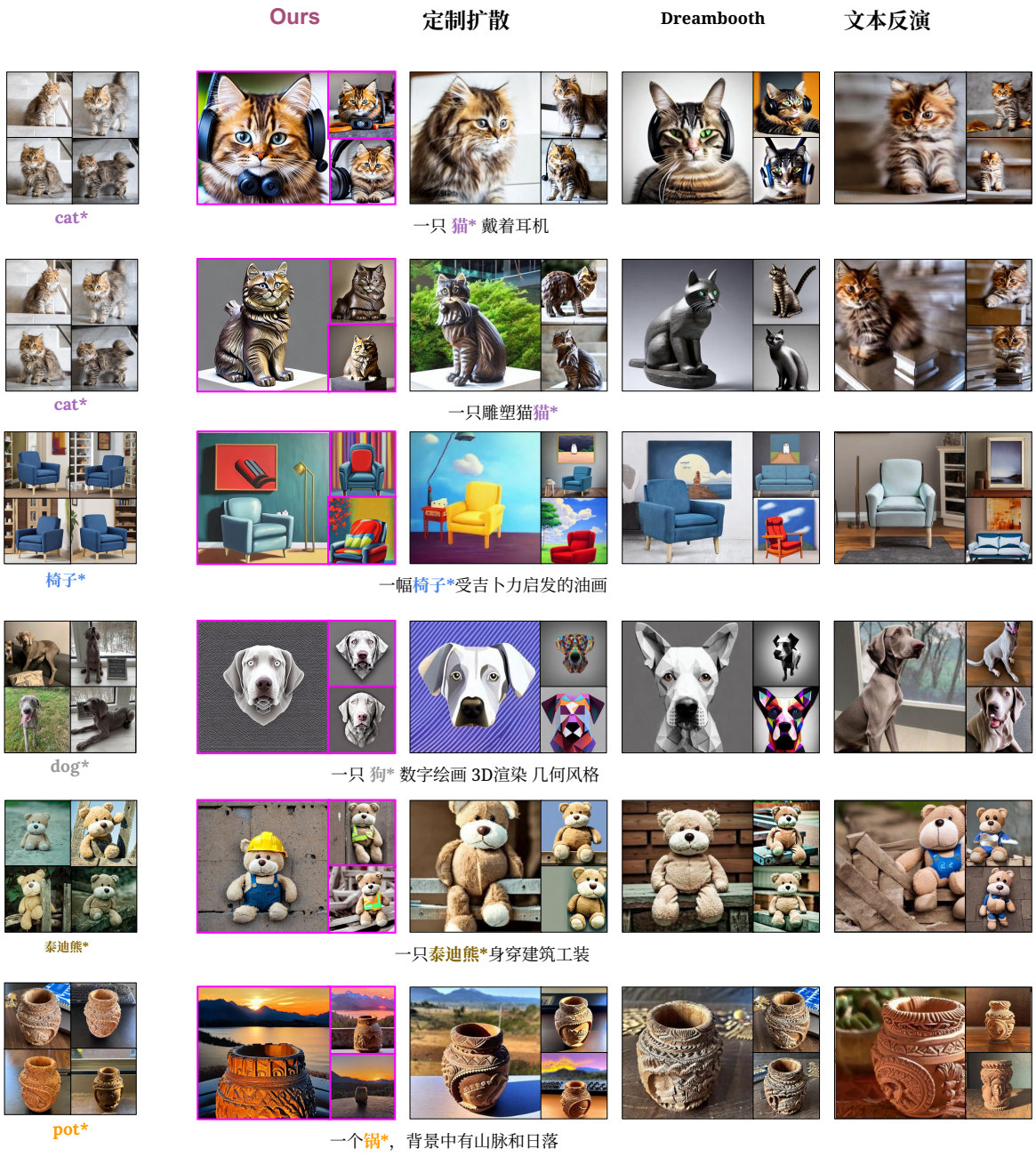

图 15. **单一概念示例的补充生成结果**。 对于每个概念，我们展示训练集中的示例图像，以及生成图像、其条件文本，并与 Custom-Diffusion（CD）、Dreambooth（DB）、 和文本反演（TI） 基线方法进行比较。所有提示词均取自 CD 论文，基线生成结果也同样如此。

参考文献

Rameen Abdal、Yipeng Qin 和 Peter Wonka。2019。Image2stylegan：如何将图像嵌入 StyleGAN 潜空间？载于*IEEE/CVF*国际计算机视觉会议论文集。4432–4441。

Harsh Agrawal、Eli A. Meirom、Yuval Atzmon、Shie Mannor 和 Gal Chechik。2021。已知的未知：使用排除式推理学习新颖概念。载于第三十七届人工智能不确定性会议论文集（机器学习研究会议录，第*161*卷），Cassio de Campos 和 Mar- loes H. Maathuis（编）。PMLR，504–514。

Yuval Atzmon、Uri Shalit 和 Gal Chechik。2015。逐个特征学习稀疏度量。载于特征提取：现代问题与挑战。PMLR，30–48。

Omri Avrahami、Dani Lischinski 和 Ohad Fried。2022。用于自然图像文本驱动编辑的混合扩散。载于*IEEE/CVF*计算机视觉与模式识别会议论文集。18208–18218。

Yogesh Balaji、Seungjun Nah、Xun Huang、Arash Vahdat、Jiaming Song、Karsten Kreis、Miika Aittala、Timo Aila、Samuli Laine、Bryan Catanzaro、Tero Karras 和 Ming- Yu Liu。2022。eDiff-I：采用专家去噪器集成的文本到图像扩散模型。*arXiv*预印本 *arXiv:2211.01324*（2022）。

David Bau、Alex Andonian、Audrey Cui、YeonHwan Park、Ali Jahanian、Aude Oliva 和 Antonio Torralba。2021。Paint by Word。arXiv:2103.10951 [cs.CV]

David Bau、Steven Liu、Tongzhou Wang、Jun-Yan Zhu 和 Antonio Torralba。2020。重写深度生成模型。载于欧洲计算机视觉会议（*ECCV*）论文集。

Tim Brooks、Aleksander Holynski 和 Alexei A. Efros。2022。InstructPix2Pix：学习遵循图像编辑指令。*ArXiv* abs/2211.09800（2022）。

Niv Cohen、Rinon Gal、Eli A. Meirom、Gal Chechik 和 Yuval Atzmon。2022。“这是我的独角兽 Fluffy”：个性化冻结视觉语言表示。载于欧洲计算机视觉会议（*ECCV*）。

Tri Dao、Daniel Y. Fu、Stefano Ermon、Atri Rudra 和 Christopher Ré。2022。FlashAttention：具有IO感知能力的快速且内存高效的精确注意力。载于神经信息处理系统进展。

Giannis Daras 和 Alexandros G. Dimakis。2022。多分辨率文本反演。*ArXiv* abs/2211.17115（2022）。

darkstorm2150。2022。*Protogen-v3.4*。https://huggingface.co/darkstorm2150/Protogen_ x3.4 官方发布/tree/main_ _

Envvi。2022。*InkPunk-v2*。https://huggingface.co/Envvi/Inkpunk-Diffusion

Rinon Gal、Yuval Alaluf、Yuval Atzmon、Or Patashnik、Amit H. Bermano、Gal Chechik 和 Daniel Cohen-Or。2022。一张图像胜过一个词：使用文本反演个性化文本到图像生成。https://doi.org/10.48550/ARXIV.2208. 01618

Rinon Gal、Or Patashnik、Haggai Maron、Gal Chechik 和 Daniel Cohen-Or。2021。Stylegan-nada：CLIP引导的图像生成器领域自适应。*arXiv*预印本 *arXiv:2108.00946*（2021）。

Mor Geva、Avi Caciularu、Kevin Ro Wang 和 Yoav Goldberg。2022。Transformer前馈层通过在词汇空间中提升概念来构建预测。*arXiv*预印本 *arXiv:2203.14680*（2022）。

Mor Geva、Roei Schuster、Jonathan Berant 和 Omer Levy。2020。Transformer前馈层是键值记忆。*arXiv*预印本 *arXiv:2012.14913*（2020）。

Amir Hertz、Ron Mokady、Jay Tenenbaum、Kfir Aberman、Yael Pritch 和 Daniel Cohen-Or。2022。使用交叉注意力控制进行提示到提示图像编辑。（2022）。

Jonathan Ho、Ajay Jain 和 Pieter Abbeel。2020。去噪扩散概率模型。神经信息处理系统进展 33（2020），6840–6851。

Bahjat Kawar、Shiran Zada、Oran Lang、Omer Tov、Huiwen Chang、Tali Dekel、In- bar Mosseri 和 Michal Irani。2022。Imagic：使用扩散模型进行基于文本的真实图像编辑。*arXiv*预印本 *arXiv:2210.09276*（2022）。

Nupur Kumari、Bingliang Zhang、Richard Zhang、Eli Shechtman 和 Jun-Yan Zhu。2022。文本到图像扩散的多概念定制。*arXiv*（2022）。

Benjamin Lefaudeux、Francisco Massa、Diana Liskovich、Wenhan Xiong、Vittorio Caggiano、Sean Naren、Min Xu、Jieru Hu、Marta Tintore、Susan Zhang、Patrick Labatut 和 Daniel Haziza。2022。xFormers：模块化且易于扩展的Transformer建模库。https://github.com/facebookresearch/xformers。

Timo Lüddecke 和 Alexander S Ecker。2021。基于提示词的多模态图像分割。*arXiv*预印本 *arXiv:2112.10003*（2021）。

Chenlin Meng、Yutong He、Yang Song、Jiaming Song、Jiajun Wu、Jun-Yan Zhu 和 Stefano Ermon。2021。SDEdit：使用随机微分方程进行引导式图像合成与编辑。载于国际学习表征会议。

Kevin Meng、David Bau、Alex Andonian 和 Yonatan Belinkov。2022。在GPT中定位并编辑事实关联。神经信息处理系统进展 36（2022）。

Ron Mokady、Amir Hertz、Kfir Aberman、Yael Pritch 和 Daniel Cohen-Or。2022。使用引导式扩散模型编辑真实图像的空文本反演。*arXiv*预印本 *arXiv:2211.09794*（2022）。

Alex Nichol、Prafulla Dhariwal、Aditya Ramesh、Pranav Shyam、Pamela Mishkin、Bob McGrew、Ilya Sutskever 和 Mark Chen。2021。GLIDE：探索使用文本引导扩散模型进行照片级图像生成与编辑。*arXiv*预印本 *arXiv:2112.10741*（2021）。或

Patashnik、Zongze Wu、Eli Shechtman、Daniel Cohen-Or 和 Dani Lischinski。2021。StyleCLIP：文本驱动的 StyleGAN 图像操控。*arXiv preprint arXiv:2103.17249*（2021）。

Suraj Patil 和 Pedro Cuenca。2022。HuggingFace DreamBooth 实现。https: // huggingface.co/docs/diffusers/training/dreambooth。

Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议。PMLR，8748–8763。

Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J Liu。2020。探索借助统一文本到文本转换器进行迁移学习的极限。《机器学习研究杂志》21, 1 (2020), 5485–5551。

Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。2022。利用 CLIP 潜变量进行层次化文本条件图像生成。*arXiv preprint arXiv:2204.06125*（2022）。

Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer。2021。使用潜扩散模型进行高分辨率图像合成。arXiv:2112.10752 [cs.CV]

Nataniel Ruiz、Yuanzhen Li、Varun Jampani、Yael Pritch、Michael Rubinstein 和 Kfir Aberman。2022。DreamBooth：针对主体驱动生成对文本到图像扩散模型进行微调。(2022)。

Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily Denton、Seyed Kamyar Seyed Ghasemipour、Burcu Karagol Ayan、S Sara Mahdavi、Rapha Gon-tijo Lopes 等。2022。具备深层语言理解能力的照片级文本到图像扩散模型。*arXiv preprint arXiv:2205.11487*(2022)。

Dvir Samuel、Yuval Atzmon 和 Gal Chechik。2020。从广义零样本学习到具有类别描述符的长尾学习。*2021 IEEE* 冬季计算机视觉应用会议 *(WACV)* (2020), 286–295。

Christoph Schuhmann、Richard Vencu、Romain Beaumont、Robert Kaczmarczyk、Clay-ton Mullis、Aarush Katta、Theo Coombes、Jenia Jitsev 和 Aran Komatsuzaki。2021。Laion-400m：包含 4 亿组经过 CLIP 筛选的图文对的开放数据集。*arXiv preprint arXiv:2111.02114*(2021)。

Dani Valevski、Matan Kalman、Yossi Matias 和 Yaniv Leviathan。2022。Unitune：通过在单张图像上微调图像生成模型来进行文本驱动的图像编辑。*arXiv preprint arXiv:2210.09477*(2022)。

Qiucheng Wu、Yujian Liu、Handong Zhao、Ajinkya Kale、Trung M. Bui、Tong Yu、Zhe Lin、Yang Zhang 和 Shiyu Chang。2022。揭示文本到图像扩散模型的解耦能力。*ArXiv* abs/2212.08698 (2022)。

Binxin Yang、Shuyang Gu、Bo Zhang、Ting Zhang、Xuejin Chen、Xiaoyan Sun、Dong Chen 和 Fang Wen。2022。Paint by Example：使用扩散模型进行基于示例的图像编辑。*ArXiv* abs/2211.13227 (2022)。

Zhixing Zhang、Ligong Han、Arna Ghosh、Dimitris N. Metaxas 和 Jian Ren。2022。SINE：使用文本到图像扩散模型进行单图像编辑。*ArXiv* abs/2212.04489 (2022)。

附录：用于文本到图像个性化的锁定键秩一编辑

A 引理：ROME 权重更新表达式的推导

本节给出一个引理，对受下方为方便起见而重述的方程 Eq. (1) 影响的层的输出进行重新组织和计算。

$$\hat{W} = W + \Lambda(C^{-1}\boldsymbol{i}_*)^T,$$

其中，$\Lambda = (\boldsymbol{o}_* - W\boldsymbol{i}_*)/[(\boldsymbol{i}_*^T(C^{-1})^T\boldsymbol{i}_*)]$, C 是一个预先缓存的常数正定矩阵。

引理： 给定输入 $\boldsymbol{e}_m$，编辑后层的输出 h 为：

$$h = W\boldsymbol{e}_m^{\perp} + \boldsymbol{o}_* sim(\boldsymbol{i}_*, \boldsymbol{e}_m)/||\boldsymbol{i}_*||^2_{C^{-1}}, \tag{5}$$

其中，(1) $sim(\boldsymbol{i}_*, \boldsymbol{e}_m) := \boldsymbol{i}_*^T(C^{-1})^T\boldsymbol{e}_m$ 用于衡量度量空间中 $\boldsymbol{e}_m$ 与 $\boldsymbol{i}_*$ 的相似度，该度量空间由 C^{-1} 定义；(2) $||\boldsymbol{i}_*||^2_{C^{-1}} := sim(\boldsymbol{i}_*, \boldsymbol{i}_*)$ 用于衡量同一度量空间中 $\boldsymbol{i}_*$ 的能量；(3) $\boldsymbol{e}_m^{\perp} := \left(\boldsymbol{e}_m - \boldsymbol{i}_* sim(\boldsymbol{i}_*, \boldsymbol{e}_m)/||\boldsymbol{i}_*||^2_{C^{-1}}\right)$ 是 $\boldsymbol{e}_m$ 在该度量空间中与 $\boldsymbol{i}_*$ 正交的分量。

证明：
现在我们继续证明该引理。

令 h 为输入 $\boldsymbol{e}_m$ 时编辑后层的输出。根据等式 Eq. (1)，我们有

$$h = \hat{W}\boldsymbol{e}_m = W\boldsymbol{e}_m + \Lambda(C^{-1}\boldsymbol{i}_*)^T\boldsymbol{e}_m.$$

注意，项 $(C^{-1}\boldsymbol{i}_*)^T\boldsymbol{e}_m$ 表示在由 C^{-1} 定义的度量空间中，$\boldsymbol{e}_m$ 与 $\boldsymbol{i}_*$ 的相似度，记为 $sim(\boldsymbol{i}_*, \boldsymbol{e}_m) =: \boldsymbol{i}_*^T(C^{-1})^T\boldsymbol{e}_m$。因此，我们可以将上面的等式改写为

$$h = W\boldsymbol{e}_m + \Lambda sim(\boldsymbol{i}_*, \boldsymbol{e}_m).$$

为简化上面的表达式，我们考察项 Λ。回想一下，$\Lambda = (\boldsymbol{o}_* - W\boldsymbol{i}_*)/[(\boldsymbol{i}_*^T(C^{-1})^T\boldsymbol{i}_*)]$。因此，我们可以代入并重新整理上面的等式，得到

$$\begin{aligned} h &= W\boldsymbol{e}_m + \frac{\boldsymbol{o}_* - W\boldsymbol{i}_*}{\boldsymbol{i}_*^T(C^{-1})^T\boldsymbol{i}_*} sim(\boldsymbol{i}_*, \boldsymbol{e}_m) \\ &= W\boldsymbol{e}_m + \frac{\boldsymbol{o}_*}{\boldsymbol{i}_*^T(C^{-1})^T\boldsymbol{i}_*} sim(\boldsymbol{i}_*, \boldsymbol{e}_m) - \frac{W\boldsymbol{i}_*}{\boldsymbol{i}_*^T(C^{-1})^T\boldsymbol{i}_*} sim(\boldsymbol{i}_*, \boldsymbol{e}_m). \end{aligned}$$

现在，为简洁起见，我们定义 $||\boldsymbol{i}_*||^2_{C^{-1}} = \boldsymbol{i}_*^T(C^{-1})^T\boldsymbol{i}_*$。据此，我们可以将上面的等式改写为

$$h = W\boldsymbol{e}_m + \frac{\boldsymbol{o}_*}{||\boldsymbol{i}_*||^2_{C^{-1}}} sim(\boldsymbol{i}_*, \boldsymbol{e}_m) - \frac{W}{||\boldsymbol{i}_*||^2_{C^{-1}}} \boldsymbol{i}_* sim(\boldsymbol{i}_*, \boldsymbol{e}_m).$$

为了进一步简化上述表达式，我们定义 $\boldsymbol{e}_m^{\perp} = \left(\boldsymbol{e}_m - \boldsymbol{i}_* sim(\boldsymbol{i}_*, \boldsymbol{e}_m)/||\boldsymbol{i}_*||^2_{C^{-1}}\right)$。这是 $\boldsymbol{e}_m$ 在由 C^{-1} 定义的度量空间中与 $\boldsymbol{i}_*$ 正交的分量。由此，我们可以将上述方程改写为

$$h = W\boldsymbol{e}_m^{\perp} + \frac{\boldsymbol{o}_*}{||\boldsymbol{i}_*||^2_{C^{-1}}} sim(\boldsymbol{i}_*, \boldsymbol{e}_m),$$

正如引理中所述。

也就是说，左项会从 $\boldsymbol{e}_m$ 中消除 $\boldsymbol{i}_*$ 的“能量”，并将其余部分（$\boldsymbol{e}_m^{\perp}$）通过原始矩阵 W。然后，右项会将该能量大致分配到 $\boldsymbol{o}_*$ 分量的方向上。

B 多个概念的正交分量 $\mathbf{e}_m^{\perp J}$ 的推导

为了计算多个概念的 $\boldsymbol{e}_m$ 正交分量，我们首先使用乔列斯基分解 $C^{-1} = LL^T$，将 $\boldsymbol{e}_m$ 和目标输入集合 $\{\boldsymbol{i}_*^j\}_{j=1..J}$ 投影到 C^{-1} 的度量空间中

$$\begin{aligned} \widetilde{\boldsymbol{e}_m} &= L^T\boldsymbol{e}_m \\ \widetilde{\boldsymbol{i}_*^j} &= L^T\boldsymbol{i}_*^j, \end{aligned} \tag{6}$$

其中，我们用“~”表示度量空间中的每个向量。

接下来，我们使用 QR分解，找到 $\{\boldsymbol{i}_*^j\}_{j=1..J}$ 的一组正交归一张成~基。我们用 $\{\widetilde{u^j}\}_{j=1..J}$ 表示这组基。

那么，$\widetilde{\boldsymbol{e}_m}$ 在该度量空间中的正交分量为

$$\widetilde{\boldsymbol{e}_m^{\perp J}} = \widetilde{\boldsymbol{e}_m} - \sum_{j\in 1..J} \widetilde{u_j} \langle \widetilde{u_j}, \widetilde{\boldsymbol{e}_m} \rangle, \tag{7}$$

其中，$\widetilde{u_j}, \widetilde{\boldsymbol{e}_m}$ 就是它们的点积 $\widetilde{u_j}^T \cdot \widetilde{\boldsymbol{e}_m}$。

将最后一个表达式通过乔列斯基平方根 $L^{T^{-1}}$ 的逆映射回文本编码器空间，从左侧乘以该逆矩阵可得

$$\boldsymbol{e}_m^{\perp J} = \boldsymbol{e}_m - \sum_{j\in 1..J} u_j \langle \widetilde{u_j}, \widetilde{\boldsymbol{e}_m} \rangle, \tag{8}$$

其中 $u_j = L^{T^{-1}}\widetilde{u_j}$。

最后，注意到 $\widetilde{u_j}, \widetilde{\boldsymbol{e}_m} = sim(u_j, \boldsymbol{e}_m)$，因为

$$\langle \widetilde{u_j}, \widetilde{\boldsymbol{e}_m} \rangle = \widetilde{u_j}^T \cdot \widetilde{\boldsymbol{e}_m} = u_j LL^T\boldsymbol{e}_m = sim(u_j, \boldsymbol{e}_m) \tag{9}$$

由此得到 $\boldsymbol{e}_m^{\perp J}$ 的最终表达式

$$\boldsymbol{e}_m^{\perp J} = \boldsymbol{e}_m - \sum_{j\in 1..J} u_j sim(u_j, \boldsymbol{e}_m) \tag{10}$$

B.1 单一概念的特殊情况

最后，我们证明，当仅存在一个概念（记为 $J = 1$）时，有 $\boldsymbol{e}_m^{\perp J} = \left(\boldsymbol{e}_m - \boldsymbol{i}_*\text{相似}(\boldsymbol{i}_*, \boldsymbol{e}_m)/||\boldsymbol{i}_*||^2_{C^{-1}}\right)$。

根据定义，$\widetilde{u_1} = L^T\boldsymbol{i}_*/||L^T\boldsymbol{i}_*||$，因此 $u_1 = \boldsymbol{i}_*/||L^T\boldsymbol{i}_*||$。将其代入公式 Eq. (10)，可得：

$$\boldsymbol{e}_m^{\perp J} = \boldsymbol{e}_m - \frac{\boldsymbol{i}_*}{||L^T\boldsymbol{i}_*||} sim(\frac{\boldsymbol{i}_*}{||L^T\boldsymbol{i}_*||}, \boldsymbol{e}_m) = \boldsymbol{e}_m - \frac{\boldsymbol{i}_* sim(\boldsymbol{i}_*, \boldsymbol{e}_m)}{||L^T\boldsymbol{i}_*||^2} \tag{11}$$

为完成证明，我们注意到，根据定义，$||L^T\boldsymbol{i}_*||^2 = \boldsymbol{i}_*^T C^{-1}\boldsymbol{i}_* = ||\boldsymbol{i}_*||^2_{C^{-1}}$。

至此，我们完成了证明：当只有一个单一概念时，$\boldsymbol{e}_m^{\perp J} = \left(\boldsymbol{e}_m - \boldsymbol{i}_* sim(\boldsymbol{i}_*, \boldsymbol{e}_m)/||\boldsymbol{i}_*||^2_{C^{-1}}\right)$，$J = 1$。

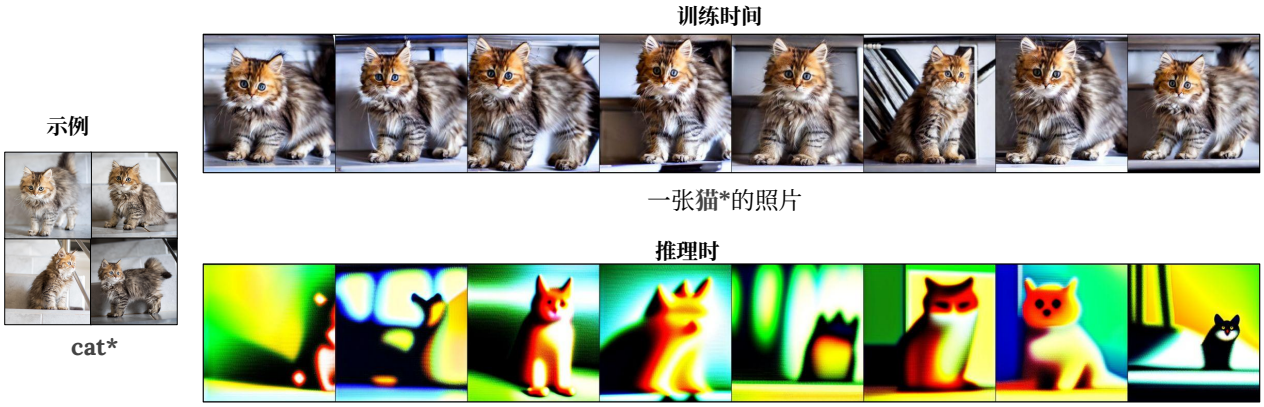

图 A.1. **训练与推理不匹配：** 如挑战 1 所述，使用 ROME 的朴素方案训练模型时，我们观察到推理时生成的图像出现损坏。我们表明，生成图像质量低下是推理与训练期间的优化目标不匹配所致。

C 附加技术细节

估计 C： C 是通过从 $100K$ 条 LAION 图像描述样本中估计 i 的非中心化协方差并预先缓存得到的常数矩阵，C^{-1} 是其逆矩阵，且二者均为正定矩阵（并且对称）。

全局键锁定： 本文介绍一种称为全局键锁定的推理时方法。

(1) 给定一个经过键锁定训练的概念以及一个包含该概念的提示词，我们首先通过文本编码器进行一次前向传播，并计算该提示词所引发的 K 和 V 序列激活，分别记为 K 和 V。(2) 我们将提示词中的概念词替换为其超类别词，并再次通过文本编码器进行前向传播。这一次，我们仅计算超类别提示词所引发的 K 序列激活，记为 $\widetilde{K}$。~(3) 我们用 K 覆盖 K，然后照常对图像进行采样。

我们将此变体称为全局键锁定，因为它锁定了整个提示词的 K 路径。

D 伪代码

算法 1 展示了秩一编辑神经模块的伪代码。该模块实现单层编辑，并且所有层都使用相同的实现。为方便起见，代码采用 PyTorch 模块风格。该模块在训练期间替换交叉注意力层中的键（K）和值（V）投影模块。它将式（3）与 i_* 的在线估计联合实现。当 o_* 时，唯一需要优化的变量是 $\boldsymbol{o}_*$。require_grad 为 *True*。对于单一概念的推理，我们只需禁用对 i_* 的在线估计。

其中，“@”运算符表示矩阵乘法，*init_input* 是提示词“A photo of a <superclass_word>”的编码，*input* 是整个提示词的编码，W 表示预训练层的权重，*concept_token_idx* 是概念词的索引

算法 1： 单概念秩一编辑模块（式 3）

初始化常量： W，C^{-1}，β，τ，
定义 初始化（初始化_输入，是否_键_锁定，概念_词元_索引）：

```
i* =初始化_输入[概念_令牌_索引]
o* = W @ 初始化_输入;

# 仅训练 o* 的 V 路径
o*.require_grad = !is_key_已锁定

def forward(输入, 概念_令牌_索引):
# i*的在线估计:
概念_编码 = 输入[概念_标记_索引]
 i* = 0.99 * i* + 0.01 * 概念_编码

# 公式 3 的组成部分:
i*_能量 = (C^-1@i*)^T@i*
相似度 = (输入@(C^-1@i*)^T)
S形函数_项 = σ((sim/i*_能量 - β)/τ)
# 计算中W  e_m^⊥:
W_em_正交_项 =
 W@input - (相似度 · (W @i*)/i*_能量)
# 式 3:
h = W_em_正交_项 +S形函数_项@o*
return h
```

word S^*，C^{-1} 是常数矩阵 C的逆矩阵，而 β，τ是 S形函数超参数。请注意，由于公式 (3) 对所有 m 都成立，因此该模块会同时处理整个提示词。

E 训练类别

- 猫玩具（来自 Textual-inversion）
- 无头雕塑（来自 Textual-Inversion）
- 猫（来自 Custom-Diffusion）
- 椅子（来自 Custom-Diffusion）
- 木盆（来自 Custom-Diffusion）

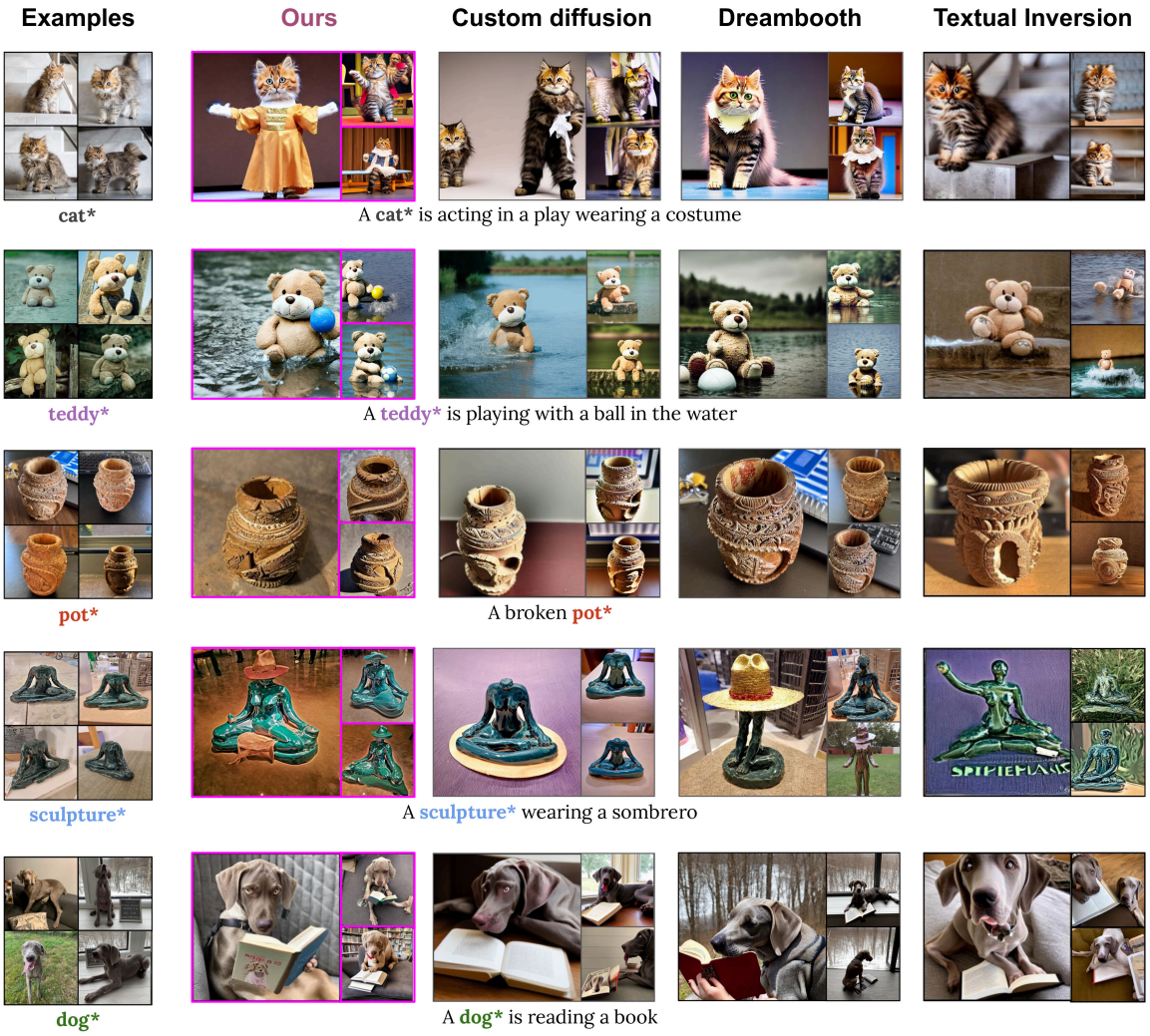

图 A.2. **包含文本反演的定性比较。**

- 狗（来自 Custom-Diffusion）
- 泰迪熊（来自 Custom-Diffusion）
- 乌龟毛绒玩具（来自 Custom-Diffusion）
- 小狗（来自 DreamBooth）
- 太阳镜（来自 DreamBooth）
- 茶壶（类似于 DreamBooth 茶壶）

F 全局锁定的成功与失败

全局锁定的图像相似度可能看起来不如局部锁定，但并非总是如此。我们发现，全局锁定难以处理形状独特的物体，但在宠物、泰迪熊和椅子等常见物体上效果良好。图 A.4 展示了全局锁定难以处理形状独特的物体，但在日常概念上效果良好。在全局锁定效果良好的情况下，我们发现它能够生成内容丰富的场景，并更好地呈现物体属性或活动的细微差别。

G 训练提示词

我们使用以下一组提示词进行训练：

- 一张 S^*的照片
- 一张 S^*的优质照片
- 一张 S^*的照片
- 一张优质 S^*照片
- 一张 S^*的图像
- S^*的图像
- 一张照片，内容是 S^*



图 A.3。 **文本反演中的注意力过拟合的更多结果**

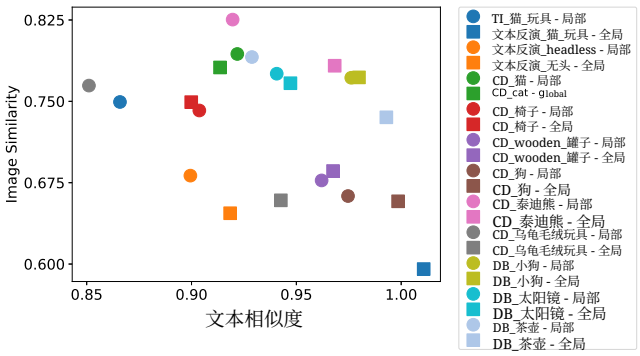

图 A.4。**全局锁定的成功与失败：**圆形——使用局部锁定。方形——使用全局锁定。全局锁定难以处理形状独特的物体类别，例如 CD_乌龟_毛绒玩具。不过，它在处理日常概念类别时很成功，例如 CD_狗。

- A S^* 显示在一张照片中，
- 一张 S^*的照片

H 定性图示参数选择

我们在定性图示中说明了为每个提示词选择的推理时参数：

图 1：

- "一个由{纯金、纱线、玻璃}制成的茶壶*" = 锁定：全局，训练偏置：0.75，推理偏置：0.45，推理温度：0.1
- "茶壶* 油画 吉卜力风格" = 锁定：全局，训练偏置：0.75，推理偏置：0.45，推理温度：0.1
- “一个身穿{武士，超级英雄，巫师}服装的泰迪熊*” =锁定：全局，训练偏置：0.75，推理偏置：0.45，推理温度：0.1
- “一个身穿蓝色西装的泰迪熊*正在烹制美食” = 锁定：局部，训练偏置：0.75，推理偏置：0.75，推理温度：0.1
- “一个泰迪熊*正在湖中的茶壶*里帆船航行” = 锁定：局部，训练偏置：0.75，推理偏置：0.55，推理温度：0.15
- "泰迪熊*在寒冷的夜晚坐在火边，旁边放着一把茶壶*" = 锁定：局部，训练偏置：0.75，推理偏置：0.55，推理温度：0.15

- "一幅泰迪熊*坐在野餐旁边的绘画" = 锁定：局部，训练偏置：0.75，推理偏置：0.55，推理温度：0.15
- “茶壶*上刻着泰迪熊*” = 锁定：局部，训练偏置：0.75，推理偏置：0.55，推理温度：0.15

图 4：

- “戴着墨西哥宽边帽的雕塑*” = 锁定：局部，训练偏置：0.55，推理偏置：0.65，推理温度：0.15
- “一只猫*穿着戏服在演戏剧” = 锁定：全局，训练偏置：0.75，推理偏置：0.65，推理温度：0.15
- “一只泰迪熊*正在水里玩球” = 锁定：局部，训练偏置：0 75，0 75，0 1 。推理偏置：。推理温度：。“一只破损的花盆*”锁定：全局，训练偏置：.75，推理偏置：0 45，0 15 。推理温度：“一只狗*正在读书” = 锁定：全局，训练偏置：0.75，推理偏置：0.45，推理温度：0.15

图 5：

- "猫*坐在椅子*上的水彩画" = 锁定：局部，训练偏置：0.6，推理偏置：0.6，推理温度：0.1
- "*桌子和椅子的照片" = 锁定：局部，训练偏置：0.7，推理偏置：0.75，推理温度：0.1
- "猫*在花园里玩耍，旁边有一个锅*" = 锁定：局部，训练偏置：0.75，推理偏置：0.4，推理温度：0.1
- "一只泰迪熊*坐在花园里，戴着太阳镜*" = 锁定：局部，训练偏置：0.75，推理偏置：0.55，推理温度：0.15

I 用户研究详情

我们通过亚马逊土耳其机器人开展了两项用户研究来评估这些模型。在第一项研究中，评分者需要对由三种方法（*Perfusion*、CD、DB）生成的一组图像进行排名。在第二项研究中，我们将*Perfusion* 与“原版” Stable-Diffusion（SD）进行比较，以研究*Perfusion* 是否会损害生成先验。

I.1 个性化方法比较

在第一项研究的每次试验中，评分者会看到某个概念的两张训练集图像和一段提示描述。随后，他们需要查看一组三张生成图像（每种方法各生成一张），并根据图像按照描述呈现该概念的效果，为每张图像分配一个从 1 到 3 的唯一排名（1 为最佳）。每次试验中的方法顺序均经过随机化。图 A.5 展示了试验所采用的实验框架。图 A.6 展示了用于帮助评分者理解说明的示例。

我们针对 11 个概念进行了 2104 次试验，每个概念平均有 24 个提示词，每个提示词进行 8 次试验。我们使用了*per-group* 提示词中挑战性子集的全部提示词。对于*Perfusion* 方法生成的图像，我们针对每个提示词选择文本和视觉相似度调和平均数最高的运行时变体。不过，两种方法所需的计算资源相同。报告的排名分数按“类别” [Samuel 等人 2020]进行平衡。具体而言，我们先计算每个概念类别的平均分数，再对所有类别的分数进行等权平均。

我们每次试验支付 0.12 美元。为确保查询质量，我们只选择了具备 AMT“大师”资格的评分者，他们在各类任务中都表现出较高的批准率，此外，我们还对预筛选评审员群体进行了资格测试，其中包含若干非常简单的精选试验。在每项资格测试中，我们都提供了一张根据提示词所描述的概念生成的优质图像、一张由原版稳定扩散生成的图像，以及一张概念训练示例中的图像。我们只让在 11 项试验中至少完成 5 项且得分满分的评分者通过资格测试。之所以没有让所有评分者完成全部试验，是因为 AMT 会随机分配试验。在我们的一项资格测试中（如图 A.5 所示），预筛选评审员群体中有相当一部分评分者表现不佳。因此，我们重新进行了这项任务，并确保所有通过资格测试的评分者都已成功完成该任务。

I.2 生成先验保持

对于第二项研究，在每次试验中，评分者都会获得一个提示词描述和两张生成图像，其中一张由 *Perfusion* 生成，另一张由 SD 生成。他们被要求根据文本描述，评价每张图像的逼真度。参与者需为每张图像赋予 1 到 3 分（最佳）的分数。与第一项研究不同，评分者可以为两张图像赋予相同的分数。每次试验中的方法顺序都是随机的。我们的试验、提示词和概念数量与第一项研究相同。此外，我们使用了为 *Perfusion*生成的相同图像。图 A.7 展示了试验中使用的实验框架。图 A.8 展示了用于帮助评分者理解说明的示例。

我们为每次试验支付 0.1 美元。为了保证查询的质量，我们只选择了具有 AMT“masters”资格的评分者。此外，我们还通过几项非常简单的精选试验进行了资格测试。我们只认可那些以满分完成所有试验且至少完成了 5 次试验的评分者。

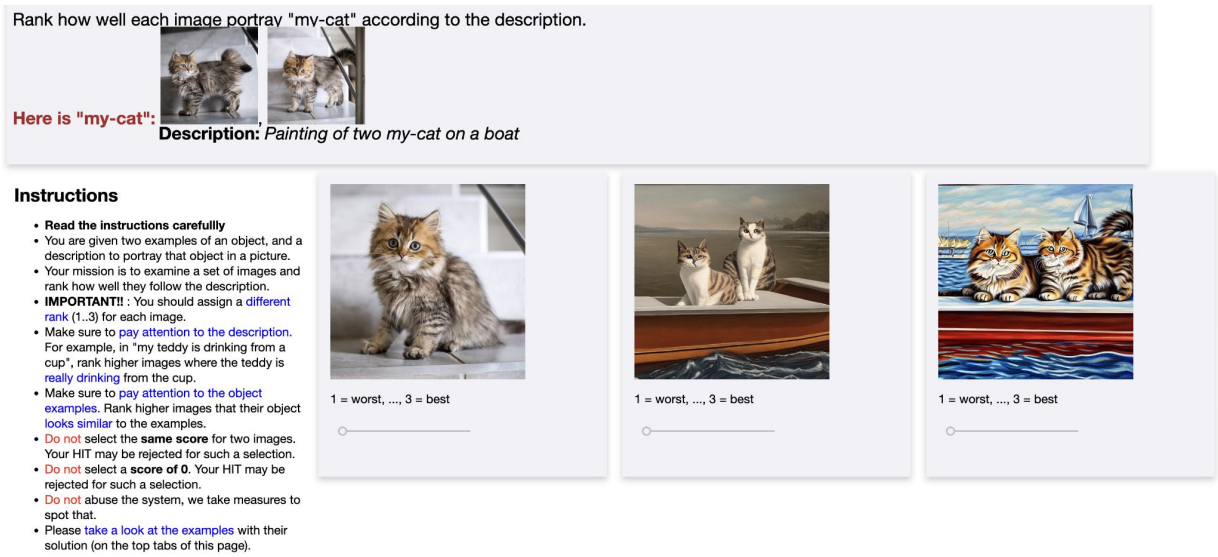

图 A.5. 方法比较用户研究中的一项资格测试。这是预筛选评审员群体中难度最高的一项，我们确保所有合格用户都参加了这项测试。

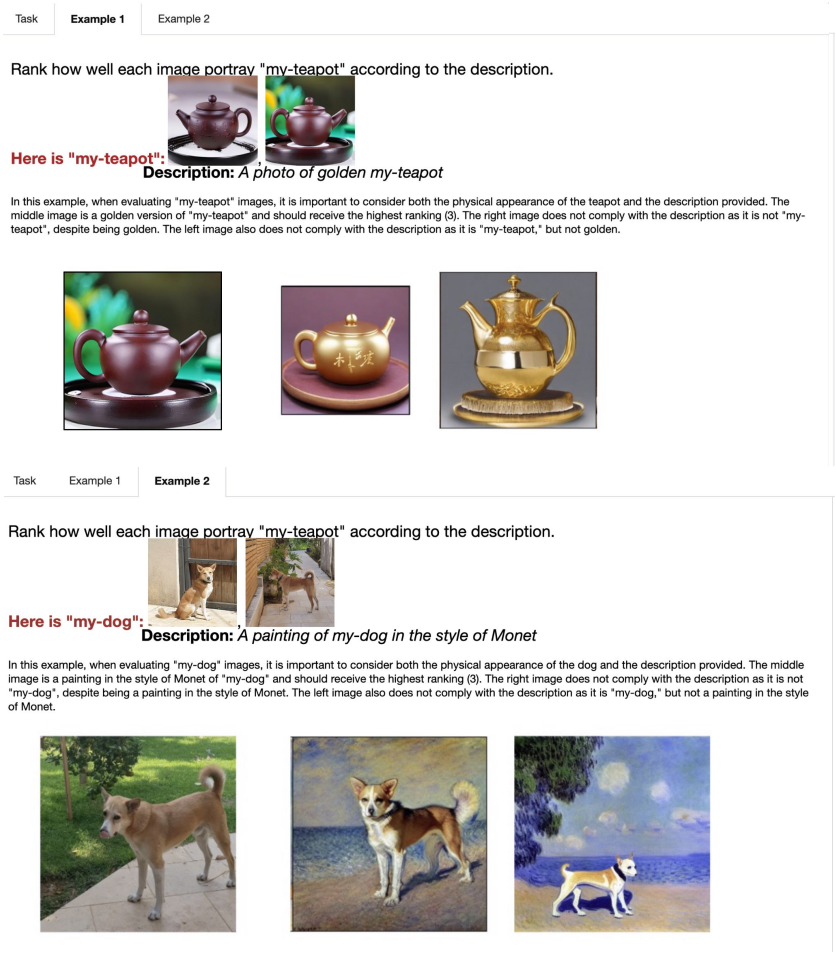

图 A.6. 方法比较用户研究中提供的示例。

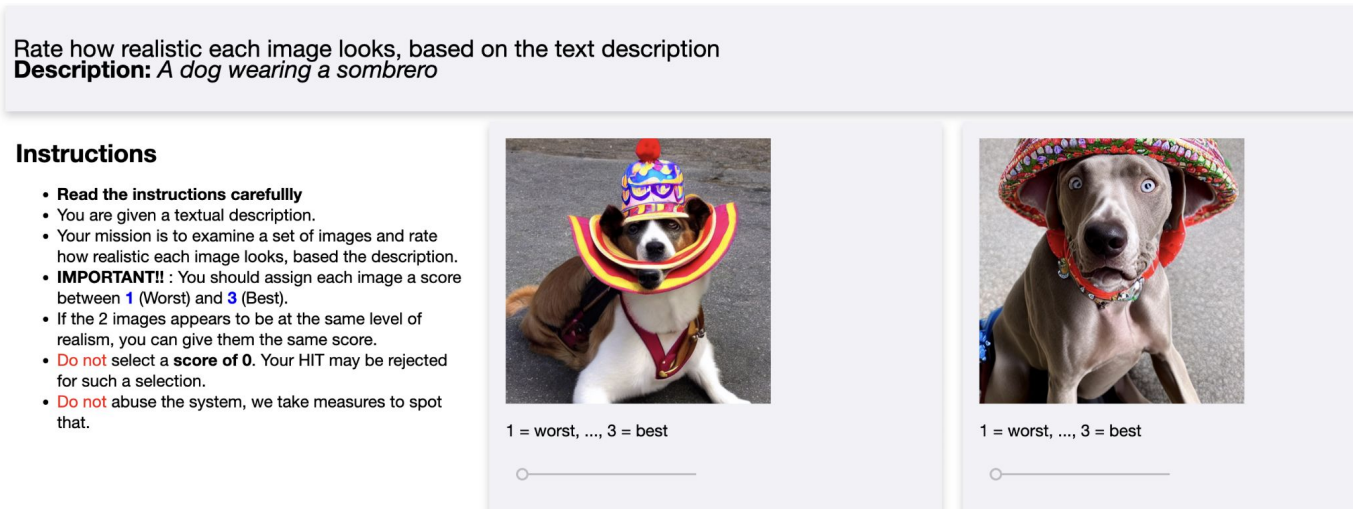

图 A.7. 先验保持用户研究中的一次试验。

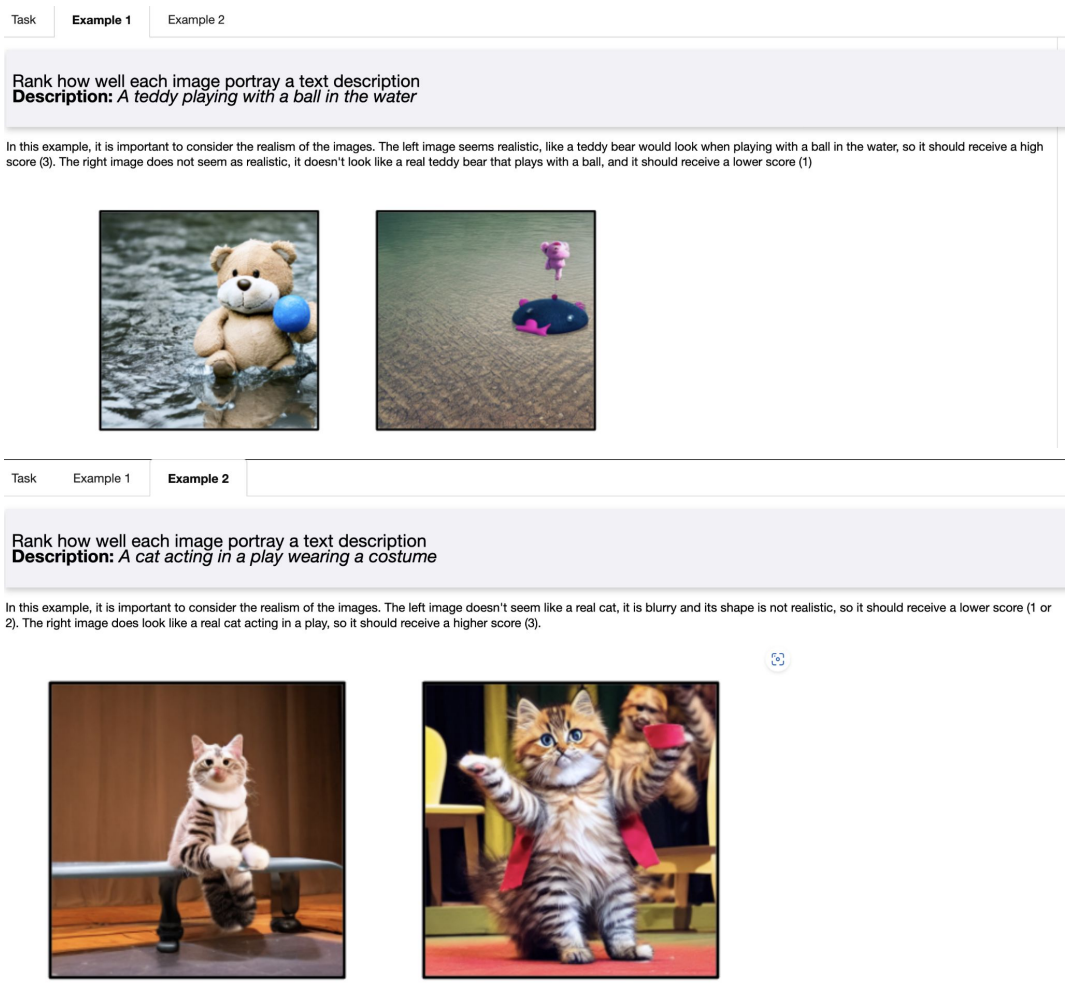

图 A.8. 先验保持用户研究中提供的示例。