VisualCloze: 一种基于视觉上下文学习的通用图像生成框架

Zhong-Yu Li[1,4*] Ruoyi Du[2,4*] Juncheng Yan[3,4] Le Zhuo[4]
Zhen Li[5†] Peng Gao[4] Zhanyu Ma[2] Ming-Ming Cheng[1†]
[1]VCIP, CS, Nankai University [2]Beijing University of Posts and Telecommunications
[3]Tsinghua University [4]Shanghai AI Laboratory [5]The Chinese University of Hong Kong
Project page: https://visualcloze.github.io

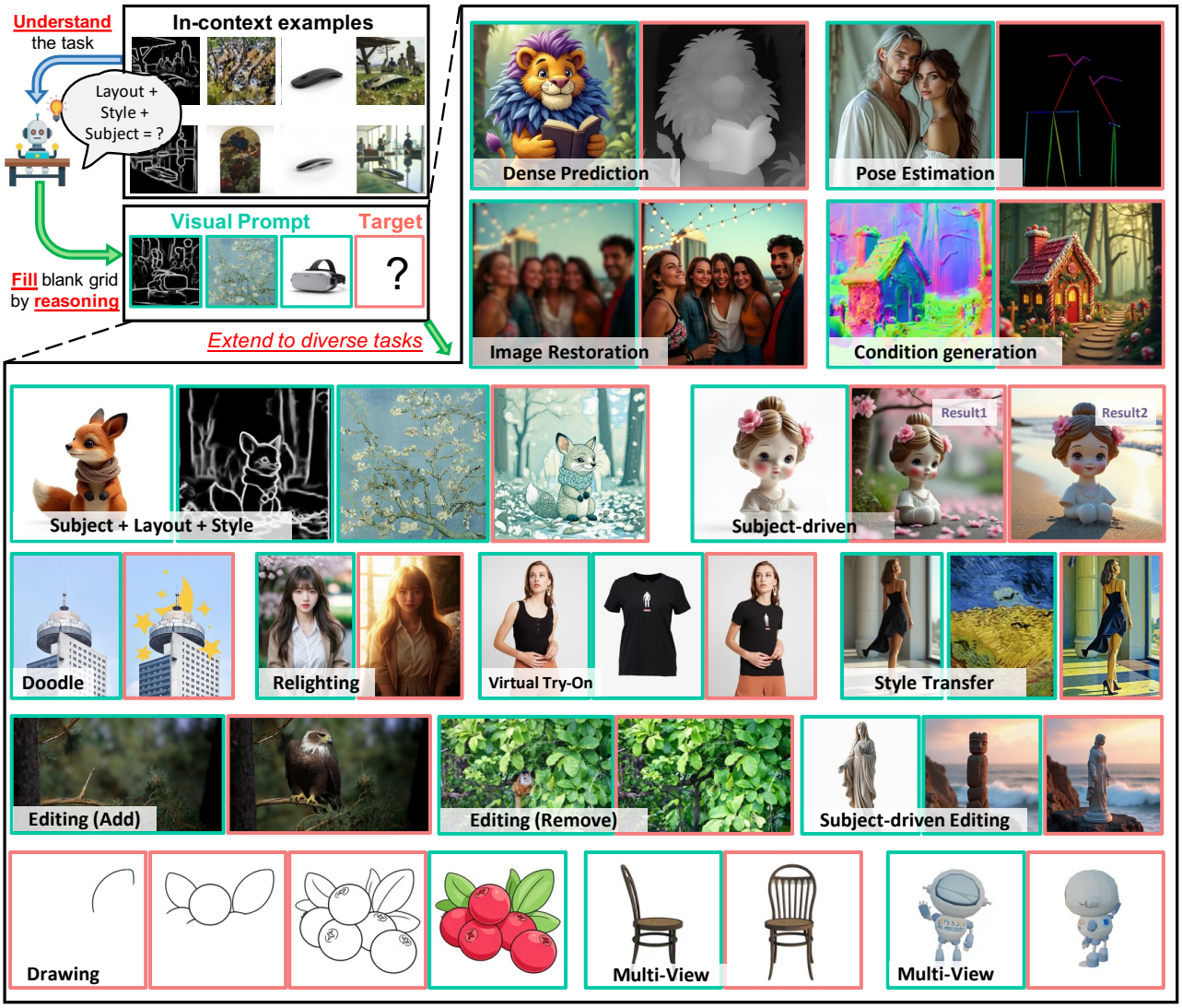
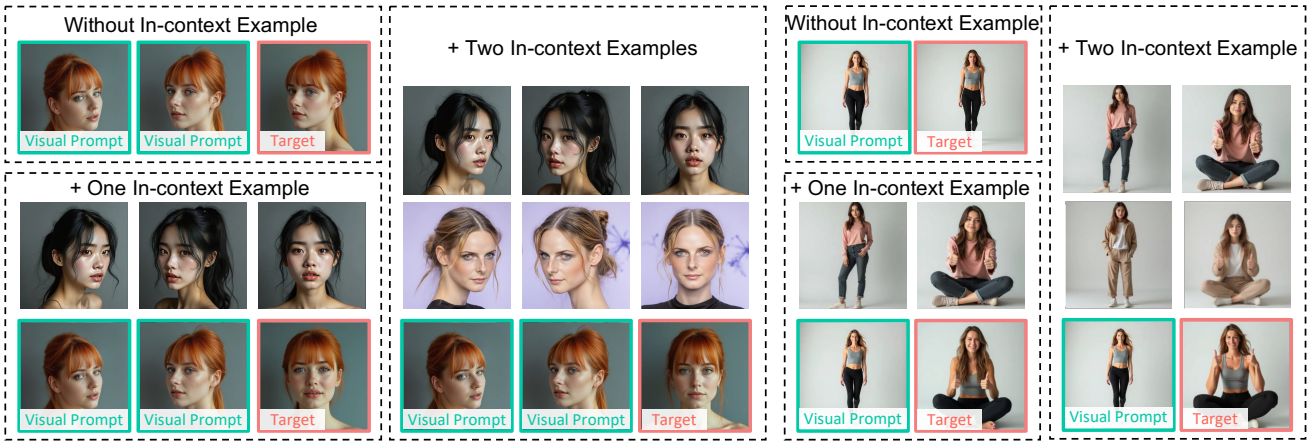


Figure 1. 左上角展示了我们基于视觉上下文学习的通用图像生成框架。给定一个特定任务的查询，生成模型通过观察少量作为示范呈现的上下文样例来学习该任务。对于每一个任务，生成结果用红色框标出。



Each row presents multi-view of a face, given a frontal face reconstruction task that leverages [IMAGE1] a left side of the face and [IMAGE2] a right side of the face, to generate [IMAGE3] that faces the center of the lens. The the last image in the final row is: the woman's frontal face that faces the center of the lens.

Each row shows a process to edit the image with the given editing instruction. The editing instruction in the last row is: making [IMAGE1] the standing woman [IMAGE2] sit down and give the thumbs up.

Figure 2. **未见任务** 👁: 通过上下文学习泛化到训练期间未见过的任务。提供更多的上下文示例将带来更准确的结果。

Abstract

扩散模型的最新进展显著推动了多种图像生成任务的发展。然而，当前主流的方法仍然专注于构建特定任务的模型，这在应对多样化需求时效率有限。尽管通用模型试图解决这一限制，但它们仍面临诸多关键挑战，包括可泛化的任务指令、合适的任务分布以及统一的架构设计。为了解决这些问题，我们提出了 VisualCloze，一个通用图像生成框架，能够支持多种**域内任务**、对**未知任务**的泛化、多任务的**统一处理**以及**逆向生成**。与现有依赖语言指令的任务驱动方式不同，这种方式常导致任务表述不明确和泛化能力弱，我们引入了视觉上下文学习，使模型能够通过视觉示范识别任务。同时，视觉任务分布本身的稀疏性也阻碍了跨任务的可迁移知识学习。为此，我们构建了 Graph200K，一个图结构的数据集，建立起多个相互关联的任务，从而提升任务密度并促进知识迁移。此外，我们发现，我们提出的统一图像生成表达方式与图像补全任务具有一致的优化目标，这使得我们可以在不修改网络结构的前提下，充分利用预训练补全模型的强大生成先验。

1. 引言

在扩散模型 [15, 33, 88] 的推动下，图像生成领域取得了显著进展，催生了包括图像编辑 [69]、风格迁移 [64, 81]、虚拟试衣 [11, 12]] 和个性化生成 [38, 54] 等在内的广泛应用。然而，这些任务通常依赖于专门的任务模型，这在实际应用中限制了其效率与可扩展性。近年来，学术界对通用生成模型 [27, 39, 44] 的兴趣日益增长，目标是在一个统一框架内处理多种图像生成任务，甚至包括未见任务。尽管相关研究已取得了一定进展，但仍存在关键问题尚未解决，包括：（1）可区分且具有泛化能力的任务指令；（2）训练阶段任务覆盖的全面性；（3）统一的模型架构。

理想的任务指令对于引导模型高效完成目标任务至关重要。现有方法主要依赖于语言指令 [27, 44] 或任务特定的 tokens [39] 来区分不同任务。但由于视觉任务本身的复杂性，以及视觉模态与语言模态之间的固有差异，使得模型难以准确理解仅通过语言描述的任务，从而引发任务混淆 [39] 并削弱对未知任务的泛化能力 [35, 71]。此外，预训练的任务 tokens 也限制了模型仅能处理已见任务。相比之下，大型语言模型已经成功实现了统一的多任务建模，这在一定程度上归功于上下文学习（ [5] 的发展，该方法允许模型通过少量示例快速适配新任务。因此，我们的目标是将上下文学习的理念迁移到纯视觉模态中，使模型能通过少量视觉示例来学习所需的任务指令，如 Fig. 1 左上所示。在这一设定下，视觉上下文学习展现出在通用图像生成任务中的巨大潜力。我们总结了四项关键发现：(1) 能够支持多种域内任务，并显著降低任务歧义 (Fig. 1)；(2) 能够对未见任务实现良好的泛化 (Fig. 2, Fig. 8)；(3) 作为一种未见的任务统一策略，能够将多个子任务整合为一步并生成中间结果 (Fig. 3)；(4) 能实现逆向生成，即从给定目标图像中反推出一组任务条件 (Fig. 9). 虽然已有研究 [1, 3, 4, 43, 66, 71, 82] 探索了视觉领域的上下文学习，但这些研究大多局限于特定领域 (如密集预测或风格迁移 [67, 87])，或仅涉及单一条件与目标图像的简化生成设置 [43, 60].

从任务分布的角度来看，视觉任务相比自然语言处理中的任务本质上更为稀疏，因为不同任务的专用数据 [71, 85] 之间几乎没有重叠 [19, 32, 79]。这种稀疏的任务学习导致各任务知识相互隔离，限制了模型跨任务学习共享特征的能力。此外，任务间的弱相关性也阻碍了知识迁移和对新任务的适应能力。然而，已有多任务学习的研究 [10, 16, 31, 53] 已验证相关任务间知识重叠的益处。为缓解视觉任务的稀疏性，我们引

* Equal contribution † Corresponding author

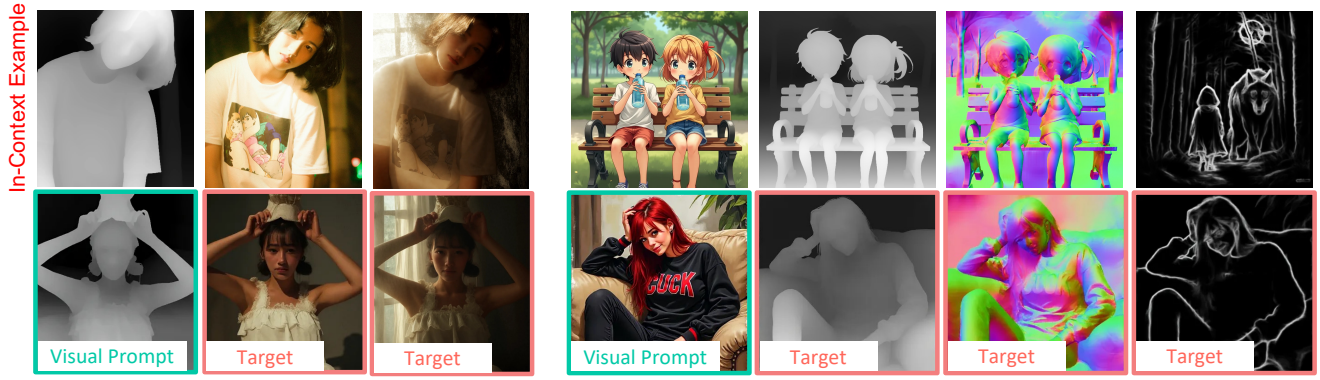


Figure 3. **未见任务**: 利用上下文学习将多个已见任务统一为一步式的未见任务。左图：将 [深度图转图像] 与 [重光照] 两个任务统一为一个新的 [不同光照条件下的深度图转图像] 任务。右图：将多个密集预测任务统一为一个联合预测任务。未提供视觉上下文的结果可见附录。

入了一个图结构数据集 Graph200K，中每张图像都关联了涵盖五个元任务的标注，条件生成 [80]，知识产权保护 [76]，风格迁移 [81]，图像编辑 [69]，和复原 [77]。通过组合不同条件，我们训练模型执行多个相互重叠的任务。依托这一高度重叠且紧凑的任务空间，我们的数据集显著提升了任务密度，使模型能更有效地学习共享和可迁移的知识。

在架构设计上，关键在于：1) 支持灵活的任务格式 [27, 35, 71]，确保上下文学习的无缝衔接，and 2) 兼容当前最先进模型 [33, 88] 充分利用其强大的生成先验。我们发现，现有最先进的图像补全模型 [33] 与我们基于上下文学习的通用生成框架目标一致。具体而言，我们将所有输入和输出图像拼接在一起，任务目标是填充输出区域。此目标的对齐使我们能在无需改动结构的前提下，基于先进的通用补全模型构建，实现以极低的数据和训练成本获得强大的通用生成能力。

基于此，我们提出了通用图像生成框架，VisualCloze，通过在 Graph200K 中采样相关任务对 FLUX.1-Filldev [33] 进行微调，学习可迁移知识并支持视觉上下文学习。随着上下文示例数量的增加，我们观察到性能提升和任务混淆减少，使模型能够支持包括条件生成、图像修复、编辑、风格迁移、知识产权保护及其组合在内的广泛域内任务。如 Fig. 2所示，在未见任务上，模型也展现出一定的泛化能力。综上，我们的主要贡献包括：

- 提出基于上下文学习的通用图像生成框架，支持多种域内任务并具备对未见任务的泛化能力
- 设计图结构数据集，Graph200K，构建紧凑任务空间，实现灵活的在线任务采样，促进模型学习跨任务共享与可迁移知识。
- 我们的统一图像生成表达形式与最先进的补全模型目标一致，无需修改结构即可通过最小调优实现卓越性能。

2. 相关工作

2.1. 图像生成

近年来，文本到图像生成技术取得了显著进展，这主要得益于自回归模型 [41, 58, 78] 和扩散模型 [2, 13, 15, 18, 24, 40, 42, 48, 51]。其中，修正流 transformers [15, 17, 33, 88] 表现出极高的训练效率和整体性能。在这些基础模型的推动下，涌现出多种应用，如条件生成 [80]，风格迁移 [64]，和个性化生成 [38]。近期，针对多任务的通用模型 [35, 44, 83] 也开始受到关注。例如，统一模型 OmniGen [71] 利用大型视觉语言模型将多任务整合入单一框架；类似地，UniReal [9] 将图像生成任务统一为不连续视频生成任务。然而，这些模型仍面临诸如过度依赖语言指令、视觉任务的孤立与稀疏性以及适应灵活任务格式的架构设计等问题。针对这些挑战，我们提出了一个通用图像生成框架，将多种生成任务统一为图像补全任务。通过视觉上下文学习以及我们设计的构建更密集任务空间以学习可迁移知识的 Graph200K 数据集，本方法有效缓解了任务歧义，支持多样的域内任务，并对训练中未见任务具备泛化能力。

2.2. 视觉上下文学习

随着大型语言模型（如 GPT-3 [5]）的兴起，上下文学习 [14] 已成为一种有效手段，使语言模型能够基于少量示例理解并完成复杂任务。视觉模态的早期工作 [21, 22] 提出利用图像类比自动创建图像滤镜。近年来，结合图像修复模型 [3, 4, 82]，掩码图像建模 [43, 66, 67]，或视觉语言模型 [1, 86]，视觉上下文学习被提出以处理更多任务。但这些工作主要聚焦于密集预测 [55, 59, 87] 或视觉理解 [63]。OmniGen [71] 也利用上下文学习实现对未见域的泛化，例如在训练中学习了分割任务后，能够分割未见概念。但其主要针对简单的密集预测任务，且训练域与未见域的差异较小。部分最新工作 [34, 43, 60, 68]] 尝试将视觉上下文学习扩展到图像生成，但仍局限于条件生成和密集预测等简单

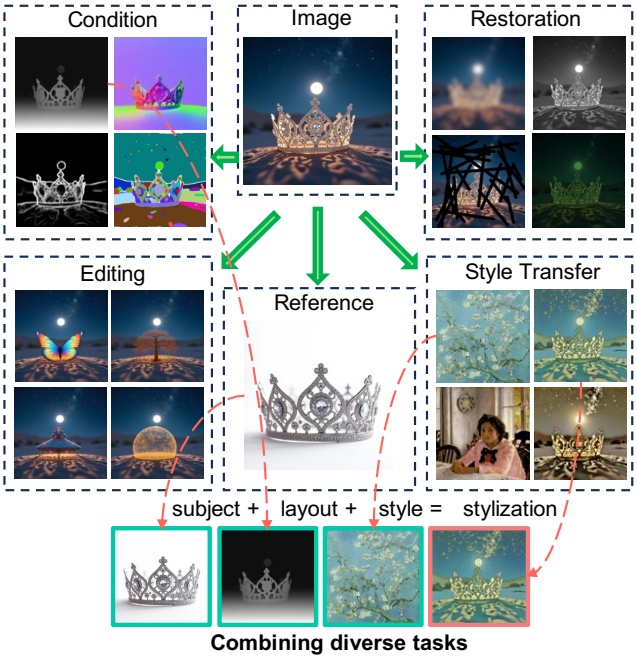


Figure 4. Graph200K 数据集示意图。每张图像都针对五类元任务进行了标注，分别为：条件生成、图像复原、图像编辑、知识产权保护以及风格迁移。基于这些任务，我们可以组合出种类丰富的复杂任务，如图中下方所示。

任务。此外，视觉任务的稀疏性使模型难以学习跨任务的可迁移与重叠知识，限制了上下文学习的生成能力。相较之下，我们引入了一个图结构数据集，支持多个相互关联任务，从而构建更密集的任务空间，促进模型学习共享且可迁移的知识，提升其适应能力。

3. 数据集

近期工作 [26, 44, 71] 在统一图像生成方面取得了显著进展，但其对未见任务的泛化能力仍然有限。我们认为这一问题部分源于视觉任务的稀疏性和孤立性，这阻碍了模型跨任务学习共享特征和处理未见任务的能力。此外，任务间的弱相关性进一步限制了知识迁移，降低了模型的适应性。因此，提高任务密度或加强任务间关联，有助于通过紧凑的任务分布提升模型的泛化能力。本文以 Subject200K [61] 数据集为基础，构建了 Graph200K 数据集，通过为每张图像增加涵盖五个元任务的 49 种标注，丰富了注释空间。该注释空间支持通过采样和组合不同元任务中的任意子集，灵活构造多种相关任务，如 Fig. 4所示。

3.1. 图结构多任务数据集

在自然语言处理领域，任务间有大量重叠，促进了强大的跨任务学习能力。相比之下，视觉任务本质上各异，使得视觉模型难以通过指令微调实现类似的泛化能力。为缓解此问题，我们引入了图结构多任务数据集。如 Fig. 4 (a) 所示，基于文本到图像的数据集，我们将每张图像视为图中的中心节点，围绕其构建多样的任务标注，包括不同空间条件、降质变换、图像编辑结果、知识产权保护用的参考图像，以及多种参考风格的风格迁移。每对任务的构建细节将在下一节介绍。

如 Fig. 4所示，每个任务标注与图像节点形成双向边，因此图是强连通的，即任意两节点间均存在双向路径。换言之，生成任务可以视为图中的一条路径，路径上的节点（除终点外）作为条件图像，相当于指令微调中的问题，而目标图像（终点）则相当于答案。具体而言，我们的 Graph200K, and 中包含 49 种节点类型，采样出多达 134 个高度重叠的任务，使模型能够学习更紧凑且共享的跨任务表征。此外，这也丰富了指令微调数据的多样性与灵活性。例如，路径 reference → editing → image 对应带参考图像的图像编辑任务，如 Fig. 4 底部所示。

3.2. 数据集构建

为了方便起见，我们继承了 Subjects200K 数据集 [61] 中的主体驱动数据。此外，对图像在线应用了 32 种不同的退化，以获得图像复原数据。本节总结了剩余三类任务的数据构建方法。

条件生成。每张图像配有 12 种由专门模型生成的不同条件，包括 Canny 边缘检测 [6], HED 边缘检测 [72], 霍夫线变换 [20], 语义分割图 [37], 深度图 [74], 形状法线图 [73], 和人体关键点 [7], 参考 ControlNet [80] 的做法。本工作通过引入 SAM2 [50] 掩码、前景分割和开放世界的边界框及掩码扩展了条件集合。前景分割基于 RMBG [84]，支持如图像修补和前景提取等多样任务。开放世界边界框由 Qwen2-VL [65] 的 grounding caption 功能生成，随后利用 SAM2 [50] 处理得到对应的掩码。

风格迁移。我们根据参考图像的风格进行迁移，涵盖语义不变和语义变异两种设置。语义不变迁移采用 InstantStyle [64] ，以保留语义内容；语义变异迁移则依赖 FLUX.1-Redux-dev [33], ，使用风格嵌入和深度作为条件。每张图像随机生成五个风格化版本。混合两种任务促使模型更好地跟随上下文示例，减少歧义。

图像编辑。设计了两类编辑任务：背景不变编辑和背景可变编辑。背景不变编辑首先定位主体，然后利用大型视觉语言模型 Qwen2-VL [65], 修改图像描述，将主体替换为新对象。将主体区域遮蔽后，使用图像修补模型 FLUX.1-Fill-dev [33] 将替换对象融合进遮罩区域。该过程重复五次以丰富数据集。背景可变编辑的区别在最后一步，使用 FLUX.1-Redux-dev [33] ，以深度作为条件，修改后的描述作为文本提示进行处理。

3.3. 其他数据

为了进一步扩展任务范围并增强模型的泛化能力，训练过程中我们引入了若干开源数据集，包括虚拟试衣

的 VITON-HD [11] 和艺术图像编辑的 PhotoDoodle [28]。图像编辑任务还结合了 OmniEdit [69]，其中两个子任务（对象添加和删除）用于训练，其余如属性修改和环境变化任务则作为未见任务，用于评估训练模型的泛化能力。此外，我们还利用了一部分高质量的内部数据，涵盖绘图过程 [62] 和多视角生成 [29] 任务。

4. 方法

本文识别出构建通用图像生成模型所面临的核心挑战，包括：明确且可泛化的任务定义、视觉任务的稀疏性，以及缺乏统一的多任务学习框架。在上一章节中，我们通过构建紧凑的 Graph200K 数据集解决了任务稀疏的问题。Sec. 4.1 提出将视觉上下文学习作为构建通用任务定义的理想范式。随后，Sec. 4.2 将图像补全模型视为统一的多任务框架，实现以最小成本获取强泛化能力。

4.1. 视觉上下文学习

当前主流做法常通过语言指令来指定任务目标，以便使用单一生成模型处理多个视觉生成任务。然而，由于视觉与语言模态之间存在语义鸿沟，图像生成模型对文本指令的理解能力仍然有限。这一问题导致现有通用生成模型在面对不同任务时常出现任务混淆 [39]，同时也难以对未见任务实现良好泛化。受到大型语言模型少样本学习成功实践的启发 [5]，我们认为：视觉上下文可作为更自然、直观的任务指令，尤其适用于具备强视觉理解能力的生成模型。

因此，本文重新提出视觉上下文学习，用于构建一个通用且可泛化的图像生成系统。为便于描述，，我们将任意条件生成任务的输入-输出表示为一个查询序列，包含 $L-1$ 个条件图像和一个空白目标 $\varnothing$，待模型完成填充。记作：$X = \text{concat}(\{x_1, \ldots, x_{L-1}, \varnothing\})$。在 Sec. 5.1，我们将展示，该方法可以推广到更一般的设置——模型可以在查询序列中的任意位置、任意数量地生成图像，而不仅仅是生成末尾的单张图像。在训练过程中，我们随机提供最多 C 个上下文示例，每个示例包含 L 张图像构成的查询序列。该策略保障了模型在不同上下文数量下的泛化能力。实验结果表明，引入上下文示例作为任务演示，不仅有助于缓解任务混淆、提升域内任务性能 [39]，还可显著增强模型在未见任务上的泛化能力。

4.2. 统一多任务框架

不同于以往视觉上下文学习方法主要聚焦于单个条件图像和单一上下文示例的简单场景 [43, 60]，本工作旨在构建一个统一的框架，能够处理不定数量的条件图像与上下文示例，从而灵活适应多种任务。为方便描述，我们先假设模型处理的所有图像尺寸一致，为 $W \times H$，并将在本节末扩展到不同宽高比的场景。在这种设定下，若给定 C 个上下文示例及一个查询示例，每个示例包含 L 张图像，则可将所有图像拼接为一个完整的网格图像，其尺寸为 $(L \times W, (C+1) \times H)$。

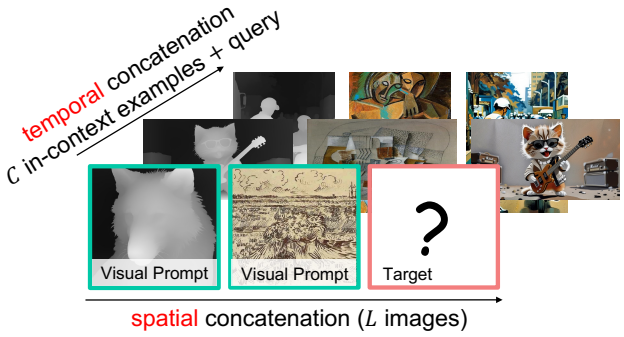


Figure 5. 在应用位置嵌入时进行图像拼接。对于 C 个上下文示例中的每个包含 L 张图像的集合以及查询图像，首先在水平方向上对图像进行拼接。然后，将这些拼接后的行在时间维度上进一步拼接，以处理长宽比不一致的问题。

模型可以通过在该网格中填补目标图像区域，实现任务完成，这类似于 视觉完形填空。因此，我们基于通用图像补全架构构建了我们的统一框架, VisualCloze，该架构本身具备处理多分辨率图像的能力。

与扩散模型的常见图像补全设计一致，我们将模型表达为：

$$\hat{X} = f(X \mid T, M), \tag{1}$$

其中 X 是拼接后的网格图像，最后一个网格留空；T 是语言指令；M 是掩码条件，指示哪些像素将被生成；并且 $\hat{X}$ 是最终补全后的图像。掩码 M 是一个二值矩阵，尺寸为 $(H \times (C+1), W \times L)$:

$$M(i,j) = \begin{cases} 1 & \text{if } i \in [H \times (C-1), H \times C) \\ & \text{and } j \in [W \times (L-1), W \times L), \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

当 $M(i,j) = 1$ 意味着该像素将被掩码并被修复网络生成。掩盖网格图像最后一行最后一列的区域，也就是目标图像区域。在训练过程中，我们还以 0.5 的概率随机遮盖前 $L-1$ 个网格中的一个，以促进反向生成能力（详见 Sec. 5.1）。在推理阶段，我们只需从 $\hat{X}$ 中裁剪出目标区域即可获得最终结果。 **对齐的优化目标**该设计的一大优势在于：VisualCloze 与通用图像补全模型在目标函数上高度一致，无需结构修改或显式条件输入。这样的设计使我们可以在无需改动结构的情况下，直接用构建的数据集对先进的图像补全模型进行微调，最大限度地利用其先验知识。相比之下，现有的任务专属模型常需要引入额外的可学习模块 [38, 69] 或适配额外条件输入 [61]，这可能削弱原模型的能力。

语言指令的作用值得注意的是，语言指令设计在 VisualCloze 同样必不可少，其作用包括：定义网格图像的布局；描述需要生成的图像内容；指明任务目标（在没有上下文示例时尤其重要）。在本统一框架中，语

言指令由三部分组成：(1) 布局指令：描述网格图像的排布，例如为 $(C+1)\times W$ ；(2) 任务指令：指定执行的任务类型；(3) 内容指令：描述目标图像的具体内容。这些指令的具体格式与示例详见附录 Appendix A。通过对公式 Equ. (1) 中的 X, T, 和 M 三个部分的重新组合，我们构建了一个基于图像补全的统一多任务框架，并自然支持上下文学习。

位置嵌入机制在上一部分中，所有图像被拼接成网格图像后，我们可以直接在该图像上使用位置嵌入 (如 RoPE [57]) 但一个潜在问题在于：上下文示例中图像尺寸的宽高比可能不同，不利于统一拼接。为了解决这个问题，我们在 Flux.1-Fill-dev 中利用了 3D-RoPE，将查询图像与上下文示例沿时间维度拼接，如 Fig. 5所示，有效克服了这一问题，同时没有带来显著的性能下降。

4.3. 实现细节

我们采用 FLUX.1-Fill-dev [33] 作为基础模型，因为它在开源图像补全模型中表现出色。在本研究中，我们选用 LoRA [25] 来微调模型，而非进行完整的参数微调，以降低训练成本并保留基础模型的能力。微调得到的 LoRA 还可以与社区中的其他 LoRA 模块融合，从而实现更广泛的应用。具体来说，我们将 LoRA 的秩（rank）设置为 256。模型在 8 张 A100 显卡上进行了 20,000 次迭代训练，累积批大小为 64。我们使用了 AdamW 优化器，学习率设为 $1e^{-4}$。遵循 FLUX.1-Fill-dev 的做法，我们采用了 lognorm 噪声策略，并加入了动态时间偏移。在训练期间，上下文示例的数量设置为最多 2 个 (在 Sec. 4.2中定义的 C)，而任务中涉及的图像数量 L, 在 Graph200K 数据集中取值为 2 到 4。在推理阶段，上下文示例的数量可以扩展为更大的数值。为了平衡计算效率，每张图像在拼接为网格布局前都会被调整为面积为 384×384 或者 512×512 。实际应用中，可以通过简单的后处理上采样技术 [45] 获得高分辨率的输出图像。

5. 实验

5.1. 上下文学习的定性分析

本节展示了一系列实验证明上下文学习在不同任务中的有效性，尤其是在训练过程中未见过的任务上。基于我们的大量实验，我们总结了四个关键发现，突显了上下文学习的作用。

上下文学习发现 1

上下文学习可以缓解已见任务中的任务混淆。

已见任务中的任务歧义。在已见任务中，模型有时会出现任务混淆，无法准确理解任务的目标，尤其是在密集预测任务上。上下文学习通过提供特定任务的示例，有效缓解了这一问题。例如，在 Fig. 6 (a) 和 (c) 中，在姿态估计和边缘检测任务中，若没有上下文示例，模型可能会产生较多噪声；而增加上下文示例数量则能提升性能与稳定性。在 Fig. 6 (b)，中的深度估计任务中，当模型原本预测不准确时，加入上下文示例也能显著提高远处区域的预测精度。此外，在某些任务中，例如条件生成任务，即使不使用上下文示例，模型也能稳定生成令人满意的结果，如 Fig. 6 (d) 所示。然而，如表 Tab. 1所示的量化结果表明，采用上下文学习仍然可以进一步提升任务完成的准确性。

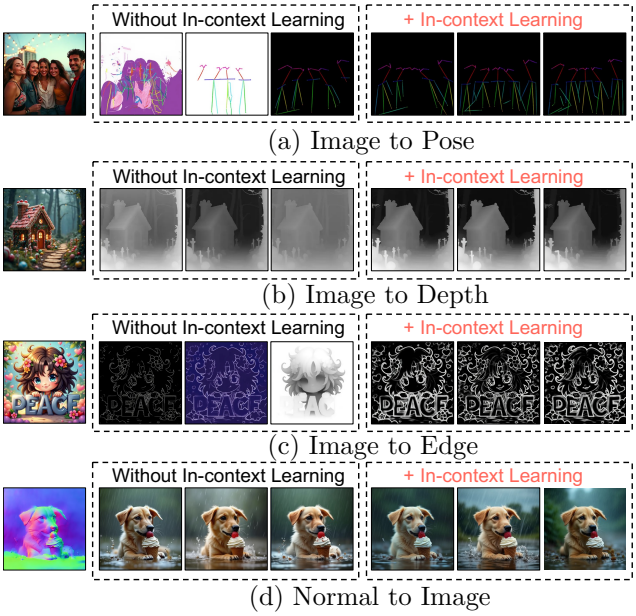


Figure 6. 上下文学习缓解了已见任务中的任务歧义性。我们展示了在不同初始噪声下的三组生成结果。

上下文学习发现 2

上下文学习支持对未见任务的泛化，提供更多的上下文样例会导致更加准确的生成结果。

对未见任务的泛化能力。除了缓解任务混淆问题，上下文学习还能使模型具备对训练过程中未见任务的泛化能力。Fig. 2 展示出模型能够通过上下文学习，从侧脸图像成功生成正脸图像，并传递编辑指令 [8] 尽管这些操作在训练中从未出现过。在这里，我们展示更多未见任务的示例。例如，虽然模型仅在包含物体添加和移除的图像编辑任务上进行了训练，但它仍能泛化到其他类型的编辑任务，如环境变化和属性修改，如 Fig. 7所示。此外，如 Fig. 8所示，，尽管模型仅在单人生成任务上接受训练，它仍然能够生成保留多个人物身份的图像。这些结果凸显了上下文学习作为一种有效的引导机制，可以在无需重新训练的情况下，使模型适应新颖任务。

Figure 7. **未见任务** : 尽管模型在训练中仅见过关于添加对象和移除对象的图像编辑任务，但它仍能通过上下文学习泛化到其他类型的编辑任务，例如环境修改（左图）和属性变换（右图）更多未见任务展示于 Fig. 2。

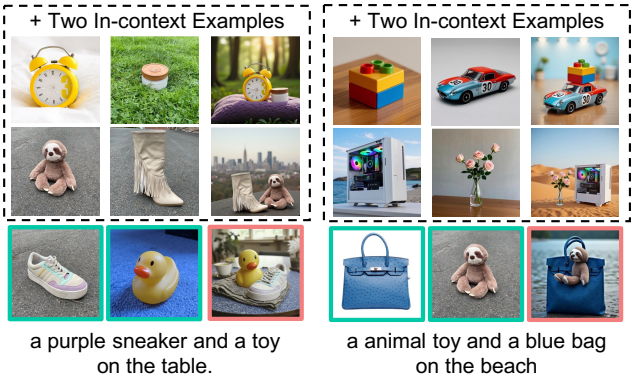


Figure 8. **未见任务** : VisualCloze 能够实现多主体驱动的图像生成 [70]，尽管模型在训练时仅接触过单主体驱动生成任务。建议放大观看效果最佳。

上下文学习发现 3

上下文学习实现了任务统一，一种未见的策略，可将多个子任务整合为单步执行，并生成中间结果。

多任务整合。我们还发现，通过上下文学习，可以将多个任务整合为一次性执行的单一步骤，这可以被视为一种未见的任务类型。Fig. 3 展示了两个示例：1）在左侧，我们将条件生成与光照重建任务合并；2）在右侧，我们同时执行深度估计、表面法线估计与边缘检测任务。类似地，Fig. 11 展示了如何在条件生成中组合多种条件，从而实现更精细的控制。例如，仅基于关键点生成肖像只能提供关于位置和身体姿态的粗略信息，在这种情况下，还可以加入轮廓条件来控制其他视觉元素的属性。

上下文学习发现 4

不同的上下文学习示例会产生不同效果，能够更好传达任务意图的示例将带来更好且更稳定的生成结果。

不同上下文示例的影响差异。参考先前关于提示选择的研究 [46, 52]，我们也发现不同的上下文示例会显著影响生成质量。具体而言，关键在于上下文示例是否能准确且有力地传达任务意图。例如，如 Fig. 10 (左) 所示，当侧脸朝向比 Fig. 10 (右)，更接近正面时，成功生成正脸图像的概率大幅下降。

上下文学习发现 5

上下文学习可以引导双向生成，即使是训练中未见的反向过程也能实现。

双向生成。除了根据给定条件生成目标图像外，我们的模型还展现了反向生成的能力，即从目标图像中推断出其背后的条件图像。尽管在训练阶段如 Sec. 4.2所述，我们是随机地将某一张条件图像作为目标进行训练，但模型在推理阶段仍能泛化到更具挑战性的、训练中未见的设置——即仅从目标图像推理出所有的条件图像。例如，如 Fig. 9 (左) 所示，模型可以根据一张风格化图像反推出原始图像和风格参考图像，展现了对内容与风格表征进行解耦的能力。类似地，如 Fig. 9 (右)，所示，模型能够根据边缘图生成对应的真实图像、深度估计图和表面法线估计图，这正是 Fig. 3 左）任务的反向过程。这种执行反向任务的能力，凸显了模型在理解不同图像表征之间复杂关系方面的灵活性和鲁棒性。

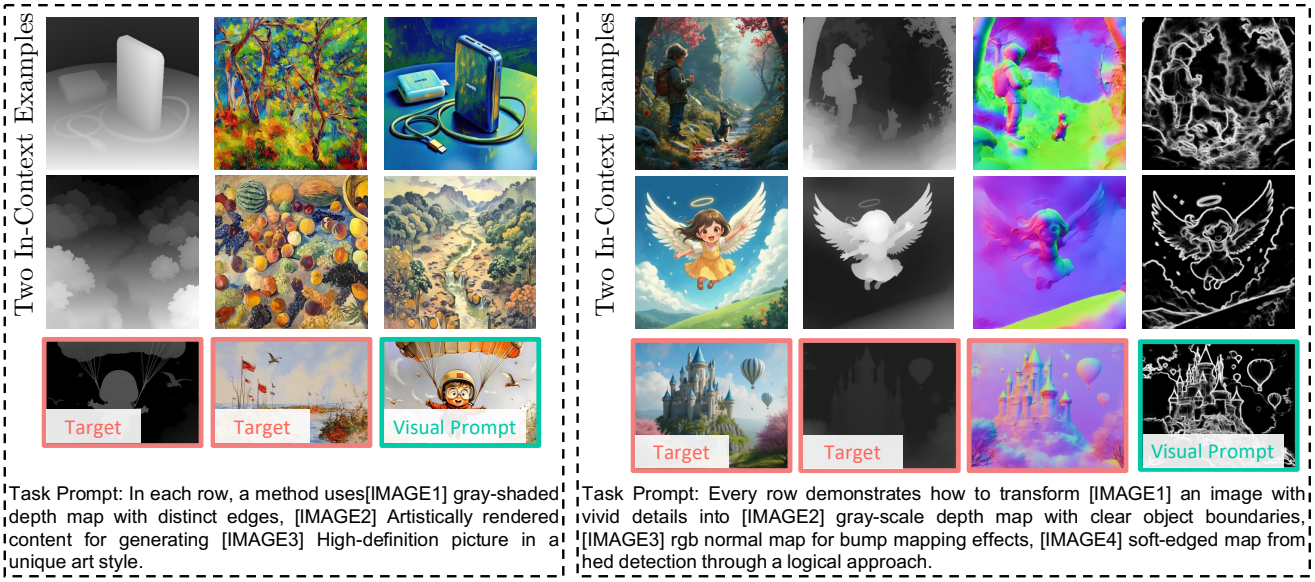


Figure 9. **未见任务** : 通过上下文学习，我们可以实现从目标到条件的逆向生成。例如，(a) 从风格化图像中分解出布局和风格；(b) 从边缘图同时推断图像、深度和表面法线，这对应于图 Fig. 3（左）的逆向任务。

Condition	Method	Context	Controllability		Quality				Text Consistency
			F1 ↑	RMSE ↓	FID [23] ↓	SSIM ↑	MAN-IQA [75] ↑	MUSIQ [30] ↑	CLIP-Score [49] ↑
Canny	ControlNet [80]		0.13	-	46.06	0.34	0.31	45.45	34.10
	OminiControl [61]		0.47	-	29.58	0.61	0.44	61.40	34.40
	OneDiffusion [35]		0.39	-	32.76	0.55	0.46	59.99	34.99
	OmniGen [71]		0.43	-	51.58	0.47	0.47	62.66	33.66
	Ours_{dev}	0	0.39	-	30.36	0.61	0.48	61.13	35.03
	$\text{Ours}_{\text{fill}}$	0	0.35	-	30.60	0.55	0.49	64.39	34.98
	$\text{Ours}_{\text{fill}}$	1	0.36	-	31.34	0.55	0.49	64.12	34.96
	$\text{Ours}_{\text{fill}}$	2	0.36	-	31.15	0.56	0.49	64.08	34.85
Depth	ControlNet [80]		-	23.70	36.83	0.41	0.44	60.17	34.49
	OminiControl [61]		-	21.44	36.23	0.52	0.44	60.18	34.08
	OneDiffusion [35]		-	10.35	39.03	0.49	0.49	60.49	34.71
	OmniGen [71]		-	15.07	86.08	0.26	0.49	64.90	29.72
	Ours_{dev}	0	-	25.06	42.14	0.53	0.46	58.95	34.80
	$\text{Ours}_{\text{fill}}$	0	-	10.31	33.88	0.54	0.48	64.85	35.10
	$\text{Ours}_{\text{fill}}$	1	-	9.91	34.44	0.54	0.49	64.32	34.95
	$\text{Ours}_{\text{fill}}$	2	-	9.68	34.88	0.54	0.48	64.29	34.89
Deblur	ControlNet [80]		-	37.82	53.28	0.49	0.45	61.92	33.80
	OminiControl [61]		-	19.70	26.17	0.85	0.45	60.70	34.53
	OneDiffusion [35]		-	-	-	-	-	-	-
	OmniGen [71]		-	-	-	-	-	-	-
	Ours_{dev}	0	-	25.03	56.76	0.74	0.38	46.68	33.52
	$\text{Ours}_{\text{fill}}$	0	-	26.53	40.59	0.74	0.46	59.62	34.56
	$\text{Ours}_{\text{fill}}$	1	-	25.87	36.93	0.76	0.48	61.58	34.82
	$\text{Ours}_{\text{fill}}$	2	-	25.57	36.28	0.76	0.48	61.77	34.82

Table 1. 条件生成与图像复原任务的定量比较。训练专用于每个任务的方法以灰色 标注。除这些方法外，性能最优的方法使用 **加粗**表示，次优方法使用 下划线 标注。

5.2. 主要结果

我们将所提出的方法与通用生成模型进行比较，包括 OmniGen [71] 和 OneDiffusion [35]，，以及一些专用模型，如 ControlNet [80] 和 Omini Control [61]。。评估指标的详细信息列于 Appendix C。此外，我们还在与 FLUX.1-Fill-dev 相同的设置下微调了 FLUX.1-dev [33] ，以便进行横向对比，并将微调后的模型分别命

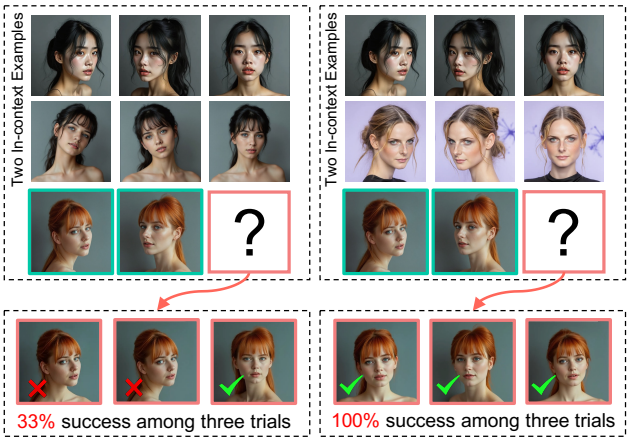


Figure 10. 不同上下文示例对上下文学习影响的示意图。在左侧第二个示例中，左右两张脸过于偏向正面，因此未能充分体现任务意图的核心目标。

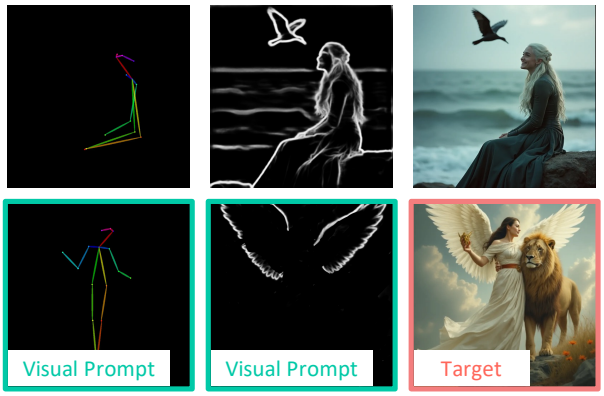


Figure 11. **未见任务** 👁: 多任务的未见组合。在条件生成中，我们整合多种条件以实现更精确的控制。更多示例见图 Fig. 3。

名为 Ours_{dev} and $\text{Ours}_{\text{fill}}$。关于 Ours_{dev} 的详细信息请见 Appendix B。

在条件生成与图像复原任务中，我们参考 Omini-Control [61] 的评估方法，从可控性、视觉质量和文本一致性三个方面对模型进行评估。如 Tab. 1所示，我们的框架在可控性方面与现有通用方法相当，同时在视觉质量和文本一致性方面表现更优。与专用方法相比，我们的模型在多个任务上表现不相上下，甚至在 depth-to-image（深度图转图像）任务中超过了这些方法。

在风格迁移任务中，我们使用 [49] 模型评估文本一致性与风格对齐情况。如 Tab. 3所示，，我们的方法在文本对齐和风格一致性上分别优于 OmniGen [71] 2% 和 3% 。即使与专用模型 InstantStyle-Plus [81], 我们

Method	Context	DINOv2	CLIP-I	CLIP-T
OminiControl [61]		73.17	87.70	33.53
OneDiffusion [35]		73.88	86.91	34.85
OmniGen [71]		67.73	83.43	34.53
Ours_{dev}	0	78.05	87.68	<u>35.06</u>
$\text{Ours}_{\text{fill}}$	0	80.41	89.63	35.16
$\text{Ours}_{\text{fill}}$	1	79.33	89.22	35.02
$\text{Ours}_{\text{fill}}$	2	<u>80.32</u>	<u>89.36</u>	35.01

Table 2. 主体驱动图像生成的定量比较。我们报告了文本一致性和风格一致性的 CLIP 得分。专用模型以灰色 标注。在剩余方法中，性能最优者以 **加粗**表示，次优者以 <u>下划线</u> 标注。

	text↑	image↑
InstantStyle [64]	0.27	0.60
OmniGen [71]	0.27	0.52
Ours_{dev}	0.30	<u>0.53</u>
$\text{Ours}_{\text{fill}}$	<u>0.29</u>	0.55

Table 3. 风格迁移任务的定量比较。我们报告了文本一致性与风格一致性的 CLIP 得分。专用模型以灰色 标注。在其余方法中，性能最优者以 **加粗**表示，次优者以 <u>下划线</u> 标注。

也在文本一致性上提升了 2% ，风格对齐上仅略有下降。

此外，我们还在基于主体驱动的图像生成任务中对模型进行了评估，并使用 DINOv2 [47],、CLIP-I [49] 和 CLIP-T [49] 指标报告语义对齐情况。如 Tab. 2所示，在所有指标上，我们的方法始终实现了性能提升。例如，与专用模型 OminiControl [61] 相比，我们在这三个指标上分别提升了 7.15%、1.66% 和 1.48%。

图像补全模型的优势。我们的方法 ($\text{Ours}_{\text{fill}}$) 构建于 FLUX.1-Fill-dev [33] 之上，其目标与我们的一体化图像生成框架保持一致。为验证其有效性，我们还在相同设置下微调了 Fill.1-dev [33] (称为 Ours_{dev}) 进行对比。与无需改动即可适配通用图像生成的 $\text{Ours}_{\text{fill}}$ 不同，Ours_{dev} 需要针对通用图像生成任务进行模型结构上的修改，详见 Appendix B。尽管如此，$\text{Ours}_{\text{fill}}$ 依然在多个任务中展现出更优的性能。

如 Tab. 1所示，在 canny-to-image 生成任务中，Ours_{dev} 的 F1 分数高于 $\text{Ours}_{\text{fill}}$ 。但在其他任务中，$\text{Ours}_{\text{fill}}$ 展现出显著优势。例如，在 depth-to-image 任务中，$\text{Ours}_{\text{fill}}$ 将 RMSE 从 25.06 降低至 10.31。在去模糊任务中，$\text{Ours}_{\text{fill}}$ 在保持更高 SSIM 的同时也降低了 RMSE，显示出更好的图像质量。而在主体驱动图像生成中，Tab. 2 展示出 $\text{Ours}_{\text{fill}}$ 始终优于 Ours_{dev}。此外，在语义不变的风格迁移任务中，如 Tab. 3所示，$\text{Ours}_{\text{fill}}$ 与 Ours_{dev} 表现相当。

Fig. 12展示了两者的可视化对比，$\text{Ours}_{\text{fill}}$ 在多个方面明显优于 Ours_{dev}。尤其是在 depth-to-image 任务

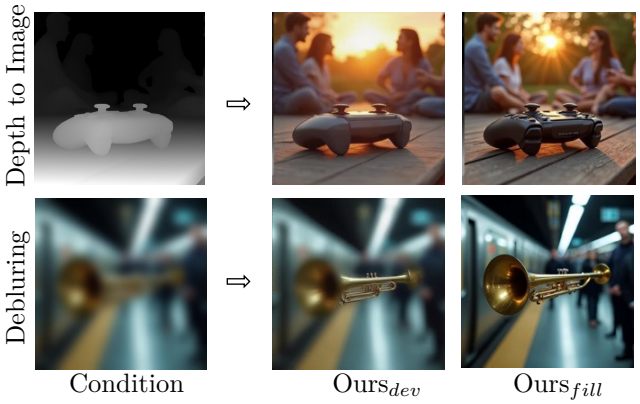


Figure 12. Flux.1-dev (Ours$_{dev}$) 和 Flux.1-Fill-dev (Ours$_{fill}$) 的对比。

中，Ours$_{\text{dev}}$ 生成的图像常出现对角线条纹伪影，严重影响了视觉保真度。综合性能、视觉质量和结构效率来看，Ours$_{\text{fill}}$ 是更具优势的模型。

上下文学习的量化对比。我们进一步分析了上下文学习在已见任务上的影响。Tab. 1 展示了在不同图像生成任务中，上下文学习的效果。在 canny 条件下，我们的方法在不使用上下文示例时的 FID 为 30.60，而在使用两个上下文示例后提升为 31.15。在 depth 条件下，RMSE 随上下文示例数量增加从 10.31 降低至 9.68，反映出结构一致性的增强。类似地，在去模糊任务中，RMSE 从 26.53 降低至 25.57，显示出更高的内容保真度。这些结果表明：上下文学习是一种有效的引导机制，使模型更好地契合任务意图。

6. 限制

尽管我们的模型在大多数同分布任务中表现出较强的稳定性，但在某些特定任务中仍表现出一定的不稳定性，例如目标移除。这一限制表明，模型性能对某些任务特征较为敏感。

此外，模型在未见任务上的稳定性仍然不足。除了任务本身的复杂性以及与已见任务之间的差异之外，模糊的上下文示例也可能导致结果的不稳定，这一点在 Sec. 5.1中已有讨论。

7. 结论

在本工作中，我们提出了 VisualCloze，一个通用图像生成框架，用于解决现有方法中存在的关键挑战，包括：可泛化的指令设计、合理的任务分布以及统一的架构设计。与仅依赖语言指令来表达任务意图的方法不同，我们重新提出了视觉上下文学习的概念，使模型能够通过少量示例学习任务。这一方法提升了模型在未见任务上的泛化能力，并减少了任务歧义。为了解决视觉任务分布稀疏性所带来的可迁移知识学习受限的问题，我们构建了 Graph200K，一个图结构的数据集，用于建立任务间的关联关系。在这一紧凑的任务空间中，模型得以学习可迁移的表示，并增强适应能力。同时，我们识别出图像补全任务与通用图像生成目标之间的一致性，从而能够无缝地将通用图像补全模型适配为通用图像生成框架，无需对原模型进行结构性更改。实验结果表明，我们的方法不仅能够通过上下文学习支持多样的同分布任务，还在未见任务上展现出强大的泛化能力。

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In NeurIPS, 2022. 2, 3

[2] Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In ICLR, 2023. 3

[3] Ivana Balazevic, David Steiner, Nikhil Parthasarathy, Relja Arandjelovic, and Olivier J Henaff. Towards in-context scene understanding. In NeurIPS, 2023. 2, 3

[4] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei A Efros. Visual prompting via image inpainting. In NeurIPS, 2022. 2, 3

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. NeurIPS, 2020. 2, 3, 5

[6] John Canny. A computational approach to edge detection. IEEE TPAMI, 1986. 4

[7] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE TPAMI, 2019. 4

[8] Lan Chen, Qi Mao, Yuchao Gu, and Mike Zheng Shou. Edit transfer: Learning image editing via vision in-context relations. arXiv preprint arXiv:2503.13327, 2025. 6

[9] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, Hui Ding, Zhe Lin, and Hengshuang. Unireal: Universal image generation and editing via learning real-world dynamics. arXiv preprint arXiv:2412.07774, 2024. 3

[10] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In ICML, 2018. 2

[11] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In CVPR, 2021. 2, 5

[12] Zheng Chong, Xiao Dong, Haoxiang Li, shiyue Zhang, Wenqing Zhang, Hanqing Zhao, xujie zhang, Dongmei Jiang, and Xiaodan Liang. CatVTON: Concatenation is all you need for virtual try-on with diffusion models. In ICLR, 2025. 2

[13] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In NeurIPS, 2021. 3

[14] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. arXiv preprint arXiv:2301.00234, 2024. 3

[15] 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. arXiv preprint arXiv:2403.03206, 2024. 2, 3

[16] Christopher Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. In NeurIPS, 2021. 2

[17] Peng Gao, Le Zhuo, Dongyang Liu, Ruoyi Du, Xu Luo, Longtian Qiu, Yuhang Zhang, Chen Lin, Rongjie Huang, Shijie Geng, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. arXiv preprint arXiv:2405.05945, 2024. 3

[18] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Mdtv2: Masked diffusion transformer is a strong image synthesizer. arXiv preprint arXiv:2303.14389, 2024. 3

[19] Golnaz Ghiasi, Barret Zoph, Ekin D. Cubuk, Quoc V. Le, and Tsung-Yi Lin. Multi-task self-training for learning general representations. In ICCV, 2021. 2

[20] Geonmo Gu, Byungsoo Ko, SeoungHyun Go, Sung-Hyun Lee, Jingeun Lee, and Minchul Shin. Towards light-weight and real-time line segment detection. In AAAI, 2022. 4

[21] Aaron Hertzmann. Algorithms for rendering in artistic styles. PhD thesis, New York University, Graduate School of Arts and Science, 2001. 3

[22] Aaron Hertzmann, Charles E. Jacobs, Nuria Oliver, Brian Curless, and David H. Salesin. Image analogies. In Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, 2001. 3

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30, 2017. 8, 14

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In NeurIPS, 2020. 3

[25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In ICLR, 2022. 6

[26] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Huanzhang Dou, Yupeng Shi, Yutong Feng, Chen Liang, Yu Liu, and Jingren Zhou. Group diffusion transformers are unsupervised multitask learners. arXiv preprint arxiv:2410.15027, 2024. 4

[27] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. arXiv preprint arxiv:2410.23775, 2024. 2, 3

[28] Shijie Huang, Yiren Song, Yuxuan Zhang, Hailong Guo, Xueyin Wang, Mike Zheng Shou, and Jiaming Liu. Photodoodle: Learning artistic image editing from few-shot pairwise data. arXiv preprint arXiv:2502.14397, 2025. 5

[29] Zehuan Huang, Yuanchen Guo, Haoran Wang, Ran Yi, Lizhuang Ma, Yan-Pei Cao, and Lu Sheng. Mv-adapter: Multi-view consistent image generation made easy. arXiv preprint arXiv:2412.03632, 2024. 5

[30] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In Proceedings of the IEEE/CVF international conference on computer vision, pages 5148–5157, 2021. 8, 14

[31] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In CVPR, 2018. 2

[32] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In CVPR, 2017. 2

[33] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2, 3, 4, 6, 8, 9, 14

[34] Bolin Lai, Felix Juefei-Xu, Miao Liu, Xiaoliang Dai, Nikhil Mehta, Chenguang Zhu, Zeyi Huang, James M Rehg, Sangmin Lee, Ning Zhang, et al. Unleashing in-context learning of autoregressive models for few-shot image manipulation. arXiv preprint arXiv:2412.01027, 2024. 3

[35] Duong H. Le, Tuan Pham, Sangho Lee, Christopher Clark, Aniruddha Kembhavi, Stephan Mandt, Ranjay Krishna, and Jiasen Lu. One diffusion to generate them all, 2024. 2, 3, 8, 9

[36] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In NeurIPS, 2023. 14

[37] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. IEEE TPAMI, 2023. 4

[38] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In CVPR, 2024. 2, 3, 5

[39] Weifeng Lin, Xinyu Wei, Renrui Zhang, Le Zhuo, Shitian Zhao, Siyuan Huang, Junlin Xie, Yu Qiao, Peng Gao, and Hongsheng Li. Pixwizard: Versatile image-to-image visual assistant with open-language instructions. arXiv preprint arXiv:2409.15278, 2024. 2, 5

[40] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In ICLR, 2023. 3

[41] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. arXiv preprint arXiv:2408.02657, 2024. 3

[42] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003, 2022. 3

[43] Yihao Liu, Xiangyu Chen, Xianzheng Ma, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Unifying image processing as visual prompting question answering. arXiv preprint arXiv:2310.10513, 2023. 2, 3, 5

[44] Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. Ace++: Instruction-based image creation and editing via context-aware content filling. arXiv preprint arXiv:2501.02487, 2025. 2, 3, 4

[45] 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. arXiv preprint arXiv:2108.01073, 2021. 6

[46] Noor Nashid, Mifta Sintaha, and Ali Mesbah. Retrieval-based prompt selection for code-related few-shot learning. In 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE), pages 2450–2462. IEEE, 2023. 7

[47] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023. 9, 14

[48] William Peebles and Saining Xie. Scalable diffusion models with transformers. In ICCV, 2023. 3, 14

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PmLR, 2021. 8, 9, 14

[50] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714, 2024. 4

[51] 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, 2022. 3

[52] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. arXiv preprint arXiv:2112.08633, 2021. 7

[53] Sebastian Ruder. An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098, 2017. 2

[54] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In CVPR, 2023. 2, 14

[55] Dianmo Sheng, Dongdong Chen, Zhentao Tan, Qiankun Liu, Qi Chu, Jianmin Bao, Tao Gong, Bin Liu, Shengwei Xu, and Nenghai Yu. Towards more unified in-context visual understanding. In CVPR, 2024. 3

[56] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. arXiv preprint arXiv:2306.00983, 2023. 14

[57] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. arXiv preprint arXiv:2104.09864, 2021. 6

[58] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525, 2024. 3

[59] Yanpeng Sun, Qiang Chen, Jian Wang, Jingdong Wang, and Zechao Li. Exploring effective factors for improving visual in-context learning. arXiv preprint arXiv:2304.04748, 2023. 3

[60] Yasheng SUN, Yifan Yang, Houwen Peng, Yifei Shen, Yuqing Yang, Han Hu, Lili Qiu, and Hideki Koike. Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation. In NeurIPS, 2023. 2, 3, 5

[61] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. arXiv preprint arXiv:2411.15098, 3, 2024. 4, 5, 8, 9

[62] Paints-Undo Team. Paints-undo github page, 2024. 5

[63] Alex Jinpeng Wang, Linjie Li, Yiqi Lin, Min Li, Lijuan Wang, and Mike Zheng Shou. Leveraging visual

tokens for extended text contexts in multi-modal learning. NeurIPS, 2024. 3

[64] Haofan Wang, Peng Xing, Renyuan Huang, Hao Ai, Qixun Wang, and Xu Bai. Instantstyle-plus: Style transfer with content-preserving in text-to-image generation. arXiv preprint arXiv:2407.00788, 2024. 2, 3, 4, 9

[65] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024. 4

[66] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In CVPR, 2023. 2, 3

[67] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Towards segmenting everything in context. In ICCV, 2023. 2, 3

[68] Zhendong Wang, Yifan Jiang, Yadong Lu, yelong shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. In-context learning unlocked for diffusion models. In NeurIPS, 2023. 3

[69] Cong Wei, Zheyang Xiong, Weiming Ren, Xinrun Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision. arXiv preprint arXiv:2411.07199, 2024. 2, 3, 5

[70] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. arXiv preprint arXiv:2504.02160, 2025. 7

[71] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. arXiv preprint arXiv:2409.11340, 2024. 2, 3, 4, 8, 9

[72] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In CVPR, 2015. 4

[73] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. IEEE TPAMI, 2023. 4

[74] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. arXiv:2406.09414, 2024. 4

[75] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1191–1200, 2022. 8, 14

[76] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721, 2023. 3

[77] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. arXiv preprint arXiv:2401.13627, 2024. 3

[78] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. Transactions on Machine Learning Research, 2022. 3

[79] Hayoung Yun and Hanjoo Cho. Achievement-based training progress balancing for multi-task learning. In ICCV, 2023. 2

[80] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In ICCV, 2023. 3, 4, 8

[81] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In CVPR, 2023. 2, 3, 9

[82] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? In NeurIPS, 2023. 2, 3

[83] Canyu Zhao, Mingyu Liu, Huanyi Zheng, Muzhi Zhu, Zhiyue Zhao, Hao Chen, Tong He, and Chunhua Shen. Diception: A generalist diffusion model for visual perceptual tasks. arXiv preprint arXiv:2502.17157, 2025. 3

[84] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. CAAI Artificial Intelligence Research, 2024. 4

[85] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In CVPR, 2017. 2

[86] Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. Visual in-context learning for large vision-language models. arXiv preprint arXiv:2402.11574, 2024. 3

[87] Muzhi Zhu, Yang Liu, Zekai Luo, Chenchen Jing, Hao Chen, Guangkai Xu, Xinlong Wang, and Chunhua Shen. Unleashing the potential of the diffusion model in few-shot semantic segmentation. In NeurIPS, 2024. 2, 3

[88] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Xiangyang Zhu, Fu-Yun Wang, Zhanyu Ma, Xu Luo, Zehan Wang, Kaipeng Zhang, Lirui Zhao, Si Liu, Xiangyu Yue, Wanli Ouyang, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-next : Making lumina-t2x stronger and faster with next-dit. In NeurIPS, 2024. 2, 3

Appendix A. **指令格式**

在我们的一体化框架中，指令由三个部分组成：(1) 布局指令：描述网格图像的布局；(2) 任务指令：指定任务类型；(3) 内容指令：描述目标图像的内容。Fig. 13 展示了两个示例指令：上图为风格、主体与布局的概念融合指令 (Fig. 13 顶部)，下图为带参考图的图像编辑任务指令。(Fig. 13 底部)。在某些条件信息已提供明确视觉线索的任务中（如风格迁移），可省略内容指令。

Appendix B. **微调** FLUX.1-dev **模型**

除了 FLUX.1-Fill-dev，我们还将所提方法适配于通用文本生成图像模型 FLUX.1-dev [33]，与在目标一致的补全模型中直接应用不同，对于 FLUX.1-dev，需要做定制化修改以处理干净的条件图像与带噪声的目标图像。具体地，我们仿照补全模型将图像拼接为网格布局后，在采样过程中始终保持条件区域对应的潜在表示为干净状态。这一策略要求对图像采样流程进行调整，因为原始 FLUX.1-Fill-dev 模型的输入是带噪声的潜在表示。此外，在 adaLN-Zero 模块 [48]，中，我们必须分别为干净的条件区域与噪声的目标区域计算均值和偏移量参数。实现方式是分别向 adaLN-Zero 输入时间步 $T = 0$ 与 $T = t$ 。t 表示当前采样步骤的时间步，采样过程中 t 会从 0 逐渐增长至 1。该策略符合 FLUX.1-dev 的预训练设定，即不同的噪声水平对应不同的均值与偏移量。如 Fig. 14所示，此方法能有效保证生成图像的视觉保真度。

Appendix C. **评估指标**

C.1. 条件生成与图像复原

我们从可控性、图像质量和文本一致性三个方面评估条件生成与图像复原任务的生成效果：

可控性。对于条件图像生成，我们通过比较输入条件与从生成图像中提取的条件之间的差异来衡量可控性。具体地，我们在边缘图转图像任务中计算 F1 分数，在深度图转图像任务中计算均方根误差（RMSE）。此外，对于去模糊任务，我们计算原始图像与复原图像之间的 RMSE。

生成质量。我们使用 FID [23], SSIM, MAN-IQA [75], and MAN-IQA [75] 来评估生成质量。FID [23] 衡量生成图像与真实图像在特征分布上的相似度；SSIM 通过比较图像的亮度、对比度与结构模式，来评估感知质量。它基于局部图像块计算统计量，并将结果组合成一个范围在 -1 到 1 之间的综合得分，数值越高表示结构保持越好。MANIQA [75] 和 MUSIQ [30] 利用神经网络来预测图像质量得分。

文本一致性。 借助 CLIP [49] 的强大多模态能力，我们还衡量生成图像与文本提示之间的语义一致性，从而反映模型对指令的遵循程度。

C.2. 主体驱动生成

参考 DreamBooth [54] 和 BLIP-Diffusion [36], 我们使用 DINOv2 [47], CLIP-I [49], 和 CLIP-T 分数来评估主体驱动图像生成的效果。DINOv2 [47] 和 CLIP-I 分数分别通过余弦相似度与 CLIP 得分来衡量参考主体与生成图像之间的对齐程度；CLIP-T 衡量生成图像与相应文本提示之间的对齐程度。

C.3. 风格迁移

参考 StyleDrop [56], 我们通过文本一致性与风格一致性来评估风格迁移的性能。对于文本一致性，我们计算生成图像与文本提示之间嵌入向量的余弦相似度，嵌入由 CLIP [49] 提取。对于风格一致性，我们计算生成图像与风格参考图之间嵌入向量的余弦相似度。需要注意的是，这两个指标应结合考虑：当模型崩溃时（即模型完全复制风格参考图，生成一幅复合图像，而忽略了文本指令），风格一致性可能达到 1.0。

Layout instruction:

12 images are organized into a grid of 3 rows and 4 columns, evenly spaced.

Task instruction:

Each row describes a process that begins with [IMAGE1] white edge lines on black from canny detection, [IMAGE2] Photo with a strong artistic theme, [IMAGE3] a reference image showcasing the dominant object and results in [IMAGE4] High-quality visual with distinct artistic touch.

Content instruction:

∅

Layout instruction:

A 3x3 grid containing 9 images, aligned in a clean and structured layout

Task instruction:

Every row provides a step-by-step guide to evolve [IMAGE1] a reference image with the main subject included, [IMAGE2] an image with flawless clarity into [IMAGE3] a high-quality image.

Content instruction:

The bottom-right corner image presents: A glossy gel nail polish bottle. At the edge of a bustling city park, this item rests on vibrant green grass, captured with a subtle bokeh effect as joggers and pets move in the background.

(a) Concatenated images (b) Language instructions

Figure 13. 包含关于拼接图像布局、任务意图及目标图像内容的语言指令示例。

(a) separate mean and shift (b) unified mean and shift

Figure 14. 在微调 FLUX.1-dev 时，分别使用均值和偏移的效果。