/TemplateVersion (2026.1)

纳米控制：一种用于扩散变压器中精确高效控制的轻量级框架

Shanyuan Liu[1 *], Jian Zhu[1 2 *], Junda Lu[1 3], Yue Gong[1 4], Liuzhuozheng Li[1],
Bo Cheng[1], Yuhang Ma[1], Liebucha Wu[1], Xiaoyu Wu[1], Dawei Leng[1], Yuhui Yin[1],

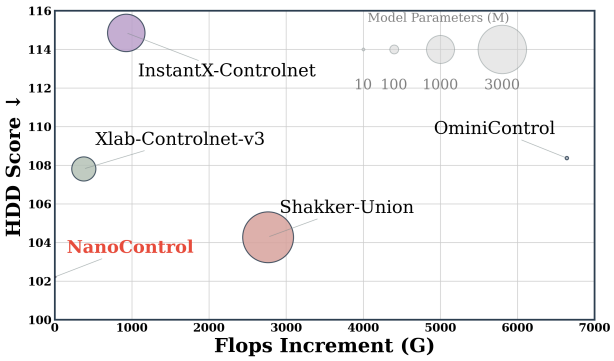

Figure 1: 不同模型中 HDD 分数与 FLOPs 增加的比较。圆圈大小表示模型参数数量（以百万计）。我们提出的模型 NanoControl 以最低的计算成本和适中的参数大小实现了有竞争力的性能。

Abstract

扩散变压器（DiTs）在文本到图像合成方面展示了卓越的能力。然而，在使用 DiTs 进行可控文本到图像生成的领域中，大多数现有方法仍然依赖于最初为基于 UNet 的扩散模型设计的 ControlNet 范式。该范式引入了显著的参数开销和增加的计算成本。为了解决这些挑战，我们提出了纳米控制扩散变压器（NanoControl），采用 Flux 作为骨干网络。我们的模型在只增加 0.024 % 的参数量和 0.029 % 的 GFLOPs 的情况下，实现了可控文本到图像生成性能的最新水平，从而实现了高效的可控生成。具体而言，我们设计了一种 LoRA 样式（低秩适应）的控制模块，不是复制 DiT 骨干进行控制，而是直接从原始条件输入中学习控制信号。此外，我们引入了一种 KV-Context 增强机制，以简单却高效的方式将条件特定的键值信息整合到骨干中，促进条件特征的深度融合。广泛的基准实验表明，与传统控制方法相比，NanoControl 显著降低了计算开销，同时保持了卓越的生成质量，并提升了可控性。

最近，扩散模型在文本到图像生成任务中取得了最先进的性能，代表性模型包括 Stable Diffusion、DALL · E 2、PixArt- 以及 FLUX.1 dev。这些模型通过利用文本信息和视觉信息之间的交互来生成高质量的图像。然而，仅依靠文本输入往往无法充分传达用户意图，因此难以生成与用户期望准确匹配的图像。为了解决这一局限性，研究人员提出了一系列基于图像的条件生成方法，这些方法结合了辅助输入——例如 Canny 边缘图和深度图，以更高的精度指导生成过程。

尽管当前的方法在可控图像生成方面取得了显著进展，但该领域仍然面临几个挑战，如控制精度有限、模型规模庞大以及计算成本高。普遍的方法是直接将最初为基于 UNet 的 (Ronneberger, Fischer, and Brox 2015) 扩散模型设计的控制机制（如 ControlNet (Zhang, Rao, and Agrawala 2023) ）适配到 DiT 框架 (Peebles and Xie 2023; Mao et al. 2025; Chen et al. 2024) 。ControlNet 的核心思想是在不影响预训练模型能力的情况下，通过复制骨干网络并注入控制条件，实现精确的图像控制。然而，这会导致模型参数和计算负载的显著增加，从而对用户设备施加更大的负担。另一种方法是利用 LoRA 实现可控模型的更高效训练，减少参数。例如，OminiControl (Tan et al. 2024) 采用轻量级的 LoRA 微调以及统一的序列处理策略，以避免昂贵的全参数更新。然而，这种方法将所有条件信息通过整个骨干网络传递，显著增加了推理时的计算成本（如图 1 所示）。

为了解决上述问题，我们提出了纳米控制扩散变压器（NanoControl），它能够生成最高质量和最强控制力的图像，同时仅引入极少的额外参数和计算成本。如图 1 所示，NanoControl 在达到最新的生成质量的同时，与现有方法相比，将额外参数数量和 FLOPs 减少了几个数量级。具体而言，我们在每个 MM-DiT 块的注意力层旁引入了两个 LoRA 风格的控制模块分支，直接处理输入的条件信号以生成条件特定的键和值表示。然后，在每个注意力层中，我们引入了键值-上下文增强机制，该机制将骨干中的原始键和值序列与条件特定的对应项结合，从而实现控制信息的直接和有效注入。

受益于我们的设计，控制相关的参数开销本质上是最小的——在实际应用中几乎可以忽略不计。在计算方面，与以前基于 LoRA 的控制模型将条件信息通过完整的骨干网络传递不同，我们的方法直接使用轻量级的 LoRA 风格控制模块进行计算，从而导致极低的额外计算成本。此外，这种设计能够更直接和高效地利用条件信号，从而显著增强生成图像的可控性。

总之，我们的主要贡献如下：

- 我们提出了 NanoControl (Nano Control Diffusion

*Equal contribution. [1]360 AI Research [2]Nanjing University of Science and Technology [3]University of Science and Technology Beijing [4]Beijing University of Aeronautics and Astronautics. Corresponding author: Dawei Leng (lengdawei@360.cn).

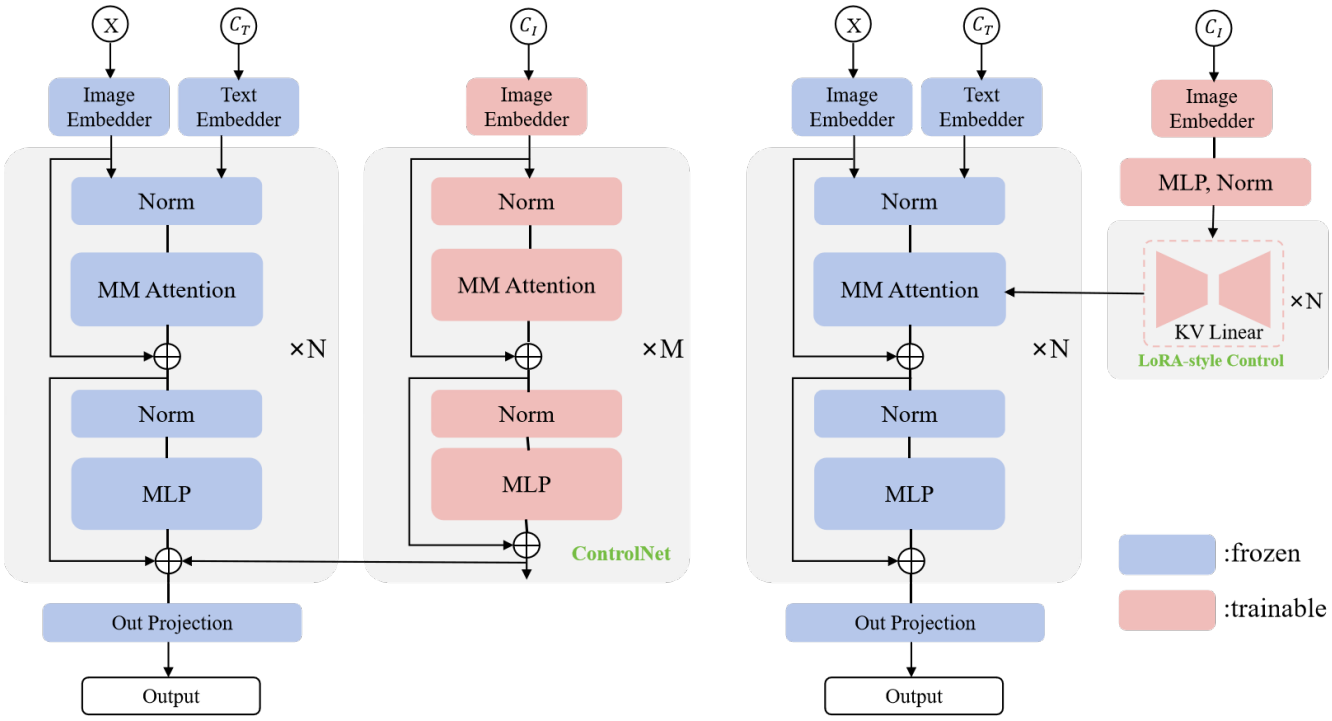

Figure 2: NanoControl 的整体框架与 ControlNet 的比较。

Transformer)，该模型只需增加 0.024 % 的参数和 0.029 % 的额外 FLOPs，即可实现高效的条件文本到图像生成，使其成为迄今为止最轻量级的控制机制之一。

- 我们设计了一个 LoRA 风格的独立控制分支来提取条件特征，并引入了 KV-上下文增强机制将其有效地注入主干中，从而提高控制保真度。
- 大量实验表明，我们的方法在保持极低参数和计算开销的同时，实现了最先进的可控性。

相关工作

扩散模型

扩散模型 (Ho, Jain, and Abbeel 2020; Song et al. 2020) 在近年来经历了快速发展，特别是在图像生成领域。文本到图像合成已取得特别显著的进展，尤其是在潜在扩散模型 (Rombach et al. 2022; Podell et al. 2023; Saharia et al. 2022) 引入之后。这些方法利用从大型预训练语言模型 (Radford et al. 2021; Raffel et al. 2020) 中提取的文本嵌入，并通过交叉注意力机制将其整合到潜在空间中，从而能够根据文本输入生成高保真度和多样化的图像。为了进一步提升生成性能，最近在扩散框架中采用了 Transformers 架构 (Vaswani et al. 2017) 。像 DiT (Peebles and Xie 2023) 这样的模型体现了这一趋势，通过去噪过程利用 Transformers 的力量来建模全球图像依赖性，从而获得更好的图像质量。在 DiT 架构的基础上，FLUX (Labs 2024) 整合了流匹配的概念，更有效地将生成轨迹与数据分布对齐，取得了更加引人注目和强大的结果。

尽管仅仅依赖文本条件在图像生成方面取得了显著的进展，但这往往不能完全捕捉用户的意图。模糊的描述、不完整的规格或缺乏精确的词汇可能导致生成的图像与期望明显偏离。为了克服这些局限，最近的方法已经开始探索结合额外的控制信号——特别是视觉输入——来增强可控性。一个突出的例子是 ControlNet，它引入了可学习的控制模块，将辅助输入（例如，边缘图、深度图）整合到预训练扩散模型中，从而实现对图像结构和内容的精确和细粒度控制。T2I-Adapter 使用轻量级适配器来支持多模态条件，而 IP-Adapter 采用了解耦的跨注意机制以促进基于图像的提示。然而，这些方法主要是为基于 UNet 的 Stable Diffusion 模型设计的，并不容易转移到像 DiT 这样的基于 transformer 的架构中。为了解决这一空缺，OminiControl 提出了一个统一的序列处理框架与 LoRA 微调，为 DiT 模型提供了一个简单但可推广的控制机制。尽管有这些发展，但对于能够同时实现高度可控性、计算效率以及优越图像质量的控制框架的需求依然强烈。

方法

预备知识

扩散变压器 目前，最先进的开源文本到图像模型，如 Flux (Labs 2024) 、Stable Diffusion 3 (Rombach et al. 2021) 和 PixArt- α (Chen et al. 2023) 都是基于 DIT 架构。DIT 在文本到视频生成方面也展示了出色的性能。它采用多层变压器块作为去噪网络，迭代地完善噪声图像表示，以生成高质量的输出。一个显著的例子是 FLUX 中的 MMDIT，它处理两种类型的标记：噪声图像标记 $X \in \mathbb{R}^{N \times d}$ 和文本条件标记 $C_T \in \mathbb{R}^{M \times d}$ ，其中 d 代表嵌入维度，N 和 M 分别代表图像和文本标记的数量。这两种模态的信息在网络中进行多模态融合，同时在整个过程中保持标记数量不变。

Task	Methods	Controllability	Image Quality		Consistency	
		HDD ↓ / MSE ↓	FID ↓	MUSIQ ↑	CLIP Text ↑	CLIP Image ↑
Canny	Flux-Contolnet	104.48	20.04	71.85	0.238	0.704
	InstantX-Controlnet	114.86	30.36	52.31	0.234	0.701
	Shakker-Union	104.28	41.82	63.10	0.231	0.699
	Xlab-Controlnet-v3	107.81	24.63	68.90	0.254	0.706
	OminiControl	108.37	16.74	70.91	0.252	0.790
	Ours	102.22	16.99	71.86	0.254	0.793
Depth	Flux-Contolnet	628.74	15.35	70.66	0.249	0.741
	InstantX-Controlnet	1144.19	36.25	61.61	0.228	0.678
	Shakker-Union	1010.42	24.95	67.41	0.241	0.712
	Xlab-Controlnet-v3	4205.97	23.69	70.44	0.259	0.714
	OminiControl	719.60	19.00	71.06	0.250	0.746
	Ours	567.95	19.48	71.92	0.253	0.759
Colorization	InstantX-Controlnet	662.21	17.95	57.51	0.248	0.873
	Shakker-Union	693.56	17.95	55.43	0.251	0.872
	OminiControl	126.31	10.17	67.63	0.250	0.911
	Ours	129.57	10.95	67.07	0.249	0.911
Hed	Xlab-Controlnet-v3	124.82	26.29	68.83	0.255	0.707
	Ours	120.57	16.07	70.60	0.252	0.819
Canny*	Flux-Contolnet	120.51	78.02	68.62	0.244	0.767
	InstantX-Controlnet	124.95	79.96	49.84	0.239	0.786
	Shakker-Union	114.96	81.41	55.47	0.239	0.782
	Xlab-Controlnet-v3	109.68	84.17	59.09	0.253	0.750
	OminiControl	135.86	77.15	63.84	0.249	0.783
	Ours	114.03	73.63	69.26	0.255	0.816

Table 1: 我们从可控性（使用 HDD 评估 Canny 和使用 MSE 评估其他方法）、图像质量（FID 和 MUSIQ），以及一致性（CLIP-Text 和 CLIP-Image）方面评估基线方法和 NanoControl。最佳结果以粗体显示，次佳结果为 underlined 。对于 Canny* 的定量评估是在 500 张分辨率为 1024×1024 的图像上进行的。

一种增强扩散模型的普遍策略是将空间控制条件加入到基础模型中，从而扩展其超越基于文本引导的能力。例如，在利用 UNet 架构的扩散模型中，ControlNet (Zhang, Rao, and Agrawala 2023) 使用一个并行的侧分支，该分支由主网络参数的一个副本初始化。在合并阶段引入了一个零卷积层，以确保稳定的训练初始化。该方法使模型能够集成 Canny 边缘、深度图和草图等空间条件，从而使用户能够生成更符合他们预期规格的图像。

$$\boldsymbol{y}_c = \mathcal{F}(\boldsymbol{x};\Theta) + \mathcal{Z}_2\left(\mathcal{F}c\left(\boldsymbol{x} + \mathcal{Z}1(\boldsymbol{c};\Theta z1);\Theta_c\right);\Theta z2\right) \tag{1}$$

在本文中，$\mathcal{F}(\boldsymbol{x};\Theta)$ 是冻结的原始模型（例如，Stable Diffusion 的 U-Net 块），其参数 Θ 保持不变。$\mathcal{F}_c(\cdot;\Theta_c)$ 是可训练的副本模型，其参数 Θ_c 是可学习的。$\mathcal{Z}_1$ 和 $\mathcal{Z}_2$ 是零卷积层（1×1 卷积，初始权重为零），其参数 Θ_{z1} 和 Θ_{z2} 在训练期间更新。$\boldsymbol{c}$ 表示条件输入（例如，边缘图、姿态图），通过 $\mathcal{Z}_1$ 转换到特征空间，然后添加到原始输入 $\boldsymbol{x}$ 。

纳米控制

为了在 DIT 架构中实现轻量化高效的控制方案，并利用 MMDIT 模块的多模态交互优势，我们提出了 Nano Control Diffusion Transformer（NanoControl）。我们使用当前最先进的开源 DIT 文本到图像模型 Flux.1 dev (Labs 2024) 进行实验。

首先，我们通过 Flux.1 dev 的 VAE (Lopez et al. 2020) 编码模块将条件控制图像映射到潜在空间。该方法提供了两个关键好处：1）VAE 映射几乎是无损的，在不牺牲信息的情况下减少了令牌长度；2）它重用了原有参数，避免了额外的参数增加并确保了兼容性。然后，通过图 2 所示的可学习图像嵌入模块，条件特征在特征维度上被对齐为与主骨干特征相同。经过非线性 MLP 和规范层，我们获得了原始条件信息 C_h 。

随后，我们设计了一条独立的侧分支，类似于 ControlNet (Zhang, Rao, and Agrawala 2023) 。这一侧分支独立处理条件信息并将其整合到主干网络中。此设计的优点是它不影响主干网络的参数，从而便于继与其他主干插件（例如 LoRA (Hu et al. 2021) ）的整合。然而，ControlNet 的侧分支过于冗余。如若干与 Unet 相关的论文（(Peng et al. 2024) , (Zhao et al. 2023) ）所述，已经提出了多种方法以减少侧分支的参数量。与 ControlNet 不同，为最小化控制模块的参数量，我们的独立侧分支仅包含匹配块层数量的线性 K 和 V 投影层。然而，由于网络中的特征维度已然很大——例如，Flux 使用一个 3072 的特征维度——因此我们独立侧

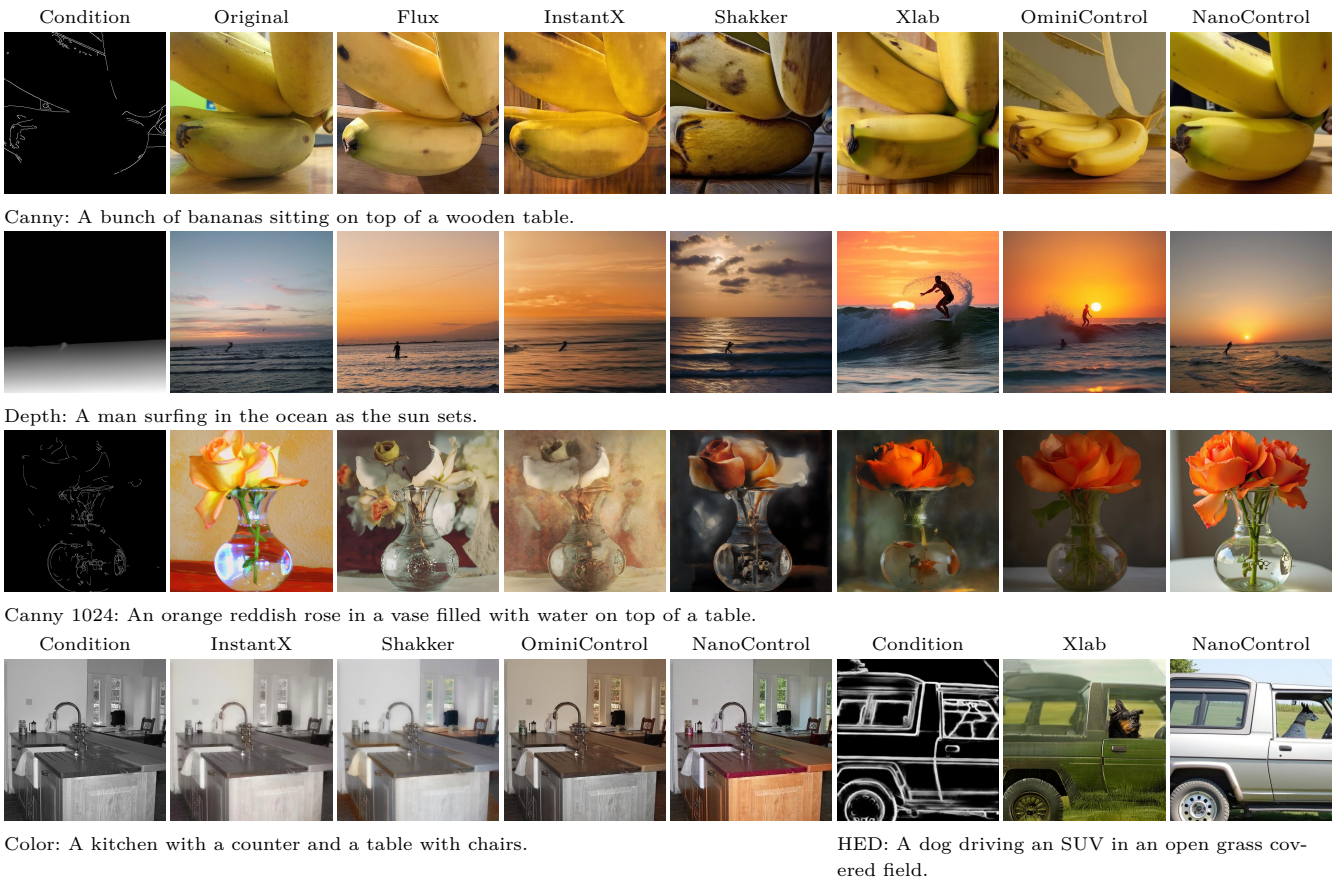

Figure 3: 在四个任务下与其他模型的定性比较

分支中的每个线性 K 和 V 层是一个 3072×3072 的矩阵，仍然过于庞大。因此，我们对 K 和 V 矩阵采用了类似 LoRA 的低秩适应，从而得到了一个超轻量化的 NanoControl 模型，其参数仅增加了 0.024 % ，而 Flux 中只增加了 0.029 % 的 GFLOPs。

然后，在每个模块中，这个相同的 C_h 被输入到独立侧支的 K 和 V 投影以进行计算。控制功能是通过与主干信息的 KV-上下文增强机制的交互来实现的。

KV-上下文增强机制 在 ControlNet (Zhang, Rao, and Agrawala 2023) 中，条件信息仅在第一个块中注入，随后的每个块都使用前一个块输出的条件特征。然后，处理后的条件特征通过逐元素加法与主干块对应的特征融合。OminiControl (Tan et al. 2024) 采用类似策略，仅在第一个块中注入原始条件信息。尽管残差连接可以缓解特征消失的问题，但原始信息仍容易被遗忘。

与上述方法不同，我们不仅仅在第一个模块中注入原始的条件特征，也不依赖于简单的特征融合相加。相反，我们提出了一种 KV-Context 增强机制。具体来说，我们将原始的条件信息 C_h 注入到每个模块中，这意味着每个模块都不依赖于前一个模块的条件特征结果。然后，在 MMDIT 部分，键（K）和值（V）与主干网络的文本符号、图像符号以及由侧分支处理的条件符号进行拼接，而查询（Q）仍然使用主干网络的文本符号和图像符号的拼接。这允许条件信息通过多模态注意力机制集成到主干网络中。

这是我们的 KV-上下文增强机制，它确保条件特征不会被遗忘，并通过注意力机制而不是简单的叠加来实现特征交互，充分利用了 DIT 架构的能力。

实验

实验设置

我们从多个角度使用一系列指标对图像生成质量进行全面评估。FID (Heusel et al. 2017) 和 MUSIQ (Ke et al. 2021) 用于评估生成图像的感知质量，而 CLIP-Text 和 CLIP-Image (Radford et al. 2021) 分数分别衡量与输入文本和参考图像的语义一致性。为了评估可控性，我们计算生成图像与条件输入之间的相似度。具体来说，对于以 Canny 或 HED 边缘图为条件的任务，我们使用 Hausdorff 距离 (HDD) (Huttenlocher, Klanderman, and Rucklidge 1993; Cao et al. 2025) 来量化对齐情况。对于深度条件的任务，我们采用均方误差 (MSE)。

实现细节。 我们在一台配备了 8 个 NVIDIA H100 GPU（每个 80GB）的机器上训练我们的模型，使用批大小为 1 并进行梯度累积 4 次，最终有效批大小为 32。训练使用 AdamW 优化器，固定学习率为 0.0001，权重衰减为 0.01。模型在 Text-to-Image-2M 数据集上训练一个 epoch。预处理期间，我们将每个图像较短的一边调整为 512 像素，并应用中心裁剪以获得 512×512 大小的图像。图像和文本的 dropout 率均设为 0.1。在

推理时，我们使用基于 Flow Matching Euler Discrete 方法 (Lipman et al. 2022) 的 24 步采样计划，指导比例为 3.5。随机种子固定为 42 以确保可重复性。

定量分析

我们通过与多种最新技术（SOTA）可控生成方法进行比较，对我们的方法进行了全面评估，包括 Flux-ControlNet（Labs 2024）、InstantX-ControlNet（InstantX 2024）、Shakker-Union（Shakker-Labs 2025）、XLab-ControlNet-v3（XLabs-AI 2023）和 OminiControl（Tan et al. 2024）。比较在四个条件任务上进行：边缘检测（Canny）、深度（Depth）、上色（Colorization）和加权边缘检测（HED）。我们从 COCO2017 验证集选择了 5000 张图像，将它们调整大小并中心裁剪到 512×512，并使用特定任务的预处理函数生成相应的条件输入。为了计算 1024 分辨率的指标，我们从 COCO2017 验证集中选择了 500 张图像，将它们处理成 1024 分辨率，并进一步处理成特定条件。

如表格 1 所示，我们从三个角度评估生成图像的质量：可控性、视觉质量和语义一致性。在可控性方面，我们的方法在 Canny、Depth 和 HED 任务中取得了最佳表现，在 Colorization 任务中与 OminiControl 相比仅表现出细微差异。在图像质量方面，我们的方法在所有四个任务中根据 FID 和 MUSIQ 指标取得了最佳或次佳的分数。在一致性方面，我们在 CLIP-Text Score 上与其他顶尖方法具有竞争力，并在 CLIP-Image Score 上显示出明显优势，这表明更好地与原始图像结构对齐。在表 1 的最后一行，我们展示了在 1024 分辨率下，不同模型的 canny 条件生成能力的对比结果。在大多数指标中，我们仍然取得了最佳表现。

值得注意的是，模型性能需要从控制指标和生成质量指标两个方面进行综合评估，因为需要考虑两种情况：首先，如果模型几乎无法学习任何信息，它可以生成高质量的图像（毕竟其骨干是 Flux），但其控制能力太弱，无法满足用户需求。其次，如果模型直接将条件信息融入生成的图像中，控制指标可以非常高（几乎直接将条件图粘贴到生成图像中并加一些颜色），但这会导致图像质量非常差，使得模型依然不可用。

定性分析

图 3 展示了我们的方法与现有方法的定性比较。如图所示，我们的模型在各种控制任务中始终生成高质量、语义对齐且视觉一致的图像。生成的输出不仅保留了条件输入提供的结构线索，还展示了与输入提示高度一致的细粒度细节和风格忠实度。（附录中提供了更多的定性结果以供进一步比较。）

这些结果结合定量指标证明了我们的方法在控制保真度、视觉逼真度和语义一致性之间达到了良好的平衡。这突出了我们方法在处理多样的控制模式时保持卓越生成质量的有效性和普适性。

我们进一步分析了 NanoControl 与现有最先进方法相比的模型效率。如表 1 所示，我们的模型相比传统的基于 ControlNet 的架构显著减少了参数数量，相较于 Flux 主干只增加了 0.024 % 的参数量。即使与轻量级的 OminiControl 相比，我们的模型也仅包含其参数的四分之一，明显表现出其紧凑性的优势。类似地，在计算成本方面，我们的方法所需的 FLOPs 大幅减少，相较于 Flux 主干仅增加了 0.029 % FLOPs，使其在效率上远超传统的 ControlNet 变体和 OminiControl。

Models	Parameters	FLOPs (512)	FLOPs (1024)
Flux-base	12 B	9,926 G	29,758 G
InstantX	+1,792 M	+926 G	+2,811 G
Shakker	+3,302 M	+2,764 G	+8,419 G
Xlab	+744 M	+375 G	+1,134 G
OminiControl	+14 M	+6,639 G	+26,539 G
Ours	+3 M	+3 G	+11 G

Table 2: 模型参数和计算量增加的比较。FLOPs(512) 是指在 512×512 大小图像上的结果，而 FLOPs(1024) 是指在 1024×1024 大小图像上的结果。

总而言之，我们的方法在生成质量和计算效率之间实现了良好的平衡。它不仅提供了优越的视觉结果和量化指标，同时也有效地减少了参数规模和推理开销，使其更适合于资源受限环境中的实际应用和部署。

我们对模型中两个关键组件的 LoRA 阶数设置进行了消融研究：LoRA 风格控制模块和图像投影 MLP 模块。表 3 (a) 展示了在不同 LoRA 阶数下 LoRA 风格控制模块的结果。随着阶数从 1 增加到 32，我们观察到在所有指标上的性能总体改善，尽管增益逐渐变得不显著。考虑到计算成本和生成质量之间的权衡，我们选择阶数 4 作为一个实用且高效的配置。表 3 (b) 报告了图像投影 MLP 模块的结果。在这种情况下，当 LoRA 阶数为 32 或更高时，模型保持相对稳定的性能。然而，当阶数降至 16 以下时，性能指标明显下降。因此，我们在这个组件中设置阶数为 32 以确保性能的稳定和可靠性。

条件策略 大多数基于 DiT 架构的现有 ControlNet 方法通过直接将条件信息添加到 hidden_states 来融合这些信息。为了更好地理解这种设计选择的影响，我们还评估了一种采用这种加法策略的方法变体。具体来说，这种方法不是将 cond_key 和 cond_value 与原始 key 和 value 连接在一起，而是仅对 query 应用交叉注意力，并通过简单的加法使用条件特征。如表 3 (c) 所示，相较于我们提出的方法，这种替代方法表现较差，证明了我们的 KV-Context 增广机制设计的有效性及结构化条件整合的重要性。

结论

NanoControl 作为一个极其简约的独立分支，被作为插件集成到 DIT 基本模型中。仅增加 0.024 % 参数量和 0.029 % Gflops，它就实现了业界领先的（SOTA）控制能力和条件生成效果。我们通过大量的定量和定性实验证明了 NanoControl 的有效性。然而，我们只在空间对齐的任务上验证了该方法，还未在空间不对齐的任务上进行实验，这将在我们后续的研究中解决。未来，我们将 NanoControl 扩展到文本到视频模型，为开源社区提供更强大的工具。

References

Cao, K.; Wang, J.; Ma, A.; Feng, J.; Zhang, Z.; He, X.; Liu, S.; Cheng, B.; Leng, D.; Yin, Y.; and Zhang, J. 2025. RelaCtrl: Relevance-Guided Efficient Control for Diffusion Transformers. ArXiv, abs/2502.14377.

Chen, J.; Wu, Y.; Luo, S.; Xie, E.; Paul, S.; Luo, P.; Zhao, H.; and Li, Z. 2024. Pixart-{\delta}: Fast and

rank	Hdd ↓	FID ↓	CLIP-I ↑
2	113.55	73.73	0.797
4	111.26	70.95	0.809
8	110.84	70.01	0.810
16	109.57	69.78	0.815
32	108.05	70.99	0.811

(a)
(a)

rank	Hdd ↓	FID ↓	CLIP-I ↑
4	113.30	72.37	0.793
8	110.83	71.85	0.803
16	110.48	71.58	0.807
32	111.26	70.95	0.809
64	111.32	71.63	0.805

(b)
(b)

method	Hdd ↓	FID ↓	CLIP-I ↑
KV-Context	111.26	70.95	0.809
add	113.49	72.99	0.800

(c)
(c)

Table 3: 消融研究。(a) LoRA 风格控制模块中的等级设置。(b) 图像投影 MLP 模块中的等级设置。(c) 不同的条件设置策略。

controllable image generation with latent consistency models. arXiv preprint arXiv:2401.05252.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J. T.; Luo, P.; Lu, H.; and Li, Z. 2023. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. ArXiv, abs/2310.00426.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In Neural Information Processing Systems.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33: 6840–6851.

Hu, J. E.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. ArXiv, abs/2106.09685.

Huttenlocher, D. P.; Klanderman, G. A.; and Rucklidge, W. 1993. Comparing Images Using the Hausdorff Distance. IEEE Trans. Pattern Anal. Mach. Intell., 15: 850–863.

InstantX. 2024. FLUX.1-dev-Controlnet-Union.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. MUSIQ: Multi-scale Image Quality Transformer. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 5128–5137.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2022. Flow matching for generative modeling. arXiv preprint arXiv:2210.02747.

Lopez, R.; Boyeau, P.; Yosef, N.; Jordan, M. I.; and Regier, J. 2020. AUTO-ENCODING VARIATIONAL BAYES.

Mao, C.; Zhang, J.; Pan, Y.; Jiang, Z.; Han, Z.; Liu, Y.; and Zhou, J. 2025. Ace++: Instruction-based image creation and editing via context-aware content filling. arXiv preprint arXiv:2501.02487.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF international conference on computer vision, 4195–4205.

Peng, B.; Wang, J.; Zhang, Y.; Li, W.; Yang, M.; and Jia, J. 2024. ControlNeXt: Powerful and Efficient Control for Image and Video Generation. ArXiv, abs/2408.06070.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In International Conference on Machine Learning.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140): 1–67.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10674–10685.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, 234–241. Springer.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems, 35: 36479–36494.

Shakker-Labs. 2025. FLUX.1-dev-ControlNet-Union-Pro.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456.

Tan, Z.; Liu, S.; Yang, X.; Xue, Q.; and Wang, X. 2024. OminiControl: Minimal and Universal Control for Diffusion Transformer. ArXiv, abs/2411.15098.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. Advances in neural information processing systems, 30.

XLabs-AI. 2023. flux-controlnet-collections.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 3813–3824.

Zhao, S.; Chen, D.; Chen, Y.-C.; Bao, J.; Hao, S.; Yuan, L.; and Wong, K.-Y. K. 2023. Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models. ArXiv, abs/2305.16322.

Models	Hdd ↓	FID ↓
Flux-base	110.28	108.04
InstantX	116.96	116.77
Shakker	119.15	126.65
Xlab	109.97	101.19
OminiControl	114.27	99.16
Ours	106.43	95.70

Table 4: 我们的方法与其他结合了 LoRA 的方法的量化结果。

使用 LoRA 集成进行评估

作为一个即插即用的模型，我们的方法表现出很强的扩展性。为了进一步评估其适应性，我们研究了将我们的方法与现有的 LoRA 模型结合的效果，并使用量化指标评估融合性能。具体而言，我们将一个 PixArt 风格的 LoRA 模型纳入我们的框架和所有其他方法中，以生成像素艺术风格的图像。这个 LoRA 模型是从一个开源社区网站获取的，需要在提示词中添加前缀 “Pixel Art ”。我们从 COCO 测试集中随机选择了 500 张图片，并在基于 Canny Conditioning 任务的基础上进行实验，使用 FID 和 HDD 作为评估指标。对于 FID 计算，参考图像集由通过在提示词中加上前缀 “Pixel Art ” 的原生 Flux 模型生成的输出组成。它用于评估各种结合了 LoRA 插件的 canny condition 模型保持 LoRA 风格的能力。较低的 FID 表明更好的风格保留能力和更强的与 LoRA 结合的能力。表 4 展示了比较结果，证明了我们的方法在可控性和保持 LoRA 风格方面均取得了卓越的性能。

条件注入的消融研究

在本节中，我们进一步研究了注入条件信息的不同策略。在我们提出的方法中，原始条件信号通过我们用于信息融合的 LoRA 风格控制模块直接嵌入到每个块的注意力层中。然而，现有条件控制模型中常用的方法是在网络入口处输入控制信息，处理后传递到下一个块，依此类推。我们将这种方法称为逐层传输。表 5 展示了这两种方法的比较。结果表明，我们的方法实现了更优越的性能，这表明更直接地注入条件信息有助于增强模型的可控性。

Methods	Hdd ↓	FID ↓	CLIP Image ↑
Layer-by-Layer	117.85	84.19	0.752
Ours	111.26	70.95	0.809

Table 5: 关于条件注入的消融研究。

在 1024×1024 分辨率下的附加定量结果

表格 6 展示了在分辨率为 1024×1024 的图像上，深度、着色和 HED 任务的相关指标结果，进一步证明了所提方法在可控性和相关方面的优越性。由于 1024×1024 图像中的信息量较大，LoRA 排名根据不同任务进行了调整。例如，在着色任务中使用了排名为 32 的 LoRA。

除了与基于 Flux 骨干的模型进行比较外，我们还对基于 SD 骨干构建的模型进行了定量评估。特别是，我们将我们的方法与基于 SD1.5 和 SDXL 骨干构建的 ControlNet 变体进行比较。结果如表 7 所示，按照报告的指标，我们的方法继续表现出强劲的性能。实验是在 512 的分辨率下进行的。

图 4 展示了我们方法在四个任务上的附加视觉结果，包括在 512 和 1024 分辨率下的输出。

Task	Methods	Controllability	Image Quality		Consistency	
		HDD ↓ / MSE ↓	FID ↓	MUSIQ ↑	CLIP Text ↑	CLIP Image ↑
Depth*	Flux-Contolnet	445.01	75.17	<u>63.91</u>	0.254	<u>0.777</u>
	InstantX-Controlnet	1455.66	116.41	50.92	0.226	0.702
	Shakker-Union	1267.40	96.28	52.35	0.241	0.734
	Xlab-Controlnet-v3	5690.33	109.89	61.20	<u>0.252</u>	0.689
	OminiControl	1640.95	101.92	52.29	0.240	0.705
	Ours	<u>488.60</u>	<u>84.75</u>	69.00	0.249	0.779
Coloring*	InstantX-Controlnet	36.36	41.51	<u>50.62</u>	0.248	0.908
	Shakker-Union	<u>27.65</u>	40.55	49.19	<u>0.251</u>	<u>0.916</u>
	OminiControl	67.98	42.11	50.11	<u>0.251</u>	0.912
	Ours	21.41	<u>40.94</u>	50.83	0.253	0.920
Hed*	Xlab-Controlnet-v3	114.32	90.47	64.96	0.258	0.741
	Ours	113.37	69.74	57.31	0.255	0.838

Table 6: 使用分辨率为 1024×1024 的图像进行深度、上色和 HED 任务的结果。最佳结果以粗体显示，第二好的结果为 <u>underlined</u> 。

Task	Methods	Controllability	Image Quality		Consistency	
		HDD ↓ / MSE ↓	FID ↓	MUSIQ ↑	CLIP Text ↑	CLIP Image ↑
Canny	SD1.5-Controlnet	<u>105.59</u>	15.99	<u>71.29</u>	0.250	0.790
	SD1.5-T2I-Adapter	109.04	17.90	68.39	<u>0.253</u>	0.763
	SD1.5-unicontrol	106.72	17.10	69.19	0.251	0.780
	SDXL-Controlnet-XS	110.58	18.70	63.34	0.248	<u>0.792</u>
	Ours	102.22	<u>16.99</u>	71.86	0.254	0.793
Depth	SD1.5-Controlnet	<u>776.35</u>	19.16	<u>70.39</u>	0.252	<u>0.735</u>
	SD1.5-T2I-Adapter	1516.98	21.36	69.04	<u>0.253</u>	0.730
	SD1.5-unicontrol	1019.84	23.75	65.85	0.259	0.726
	SDXL-Controlnet-XS	1076.37	28.82	50.59	0.251	0.723
	Ours	567.95	<u>19.48</u>	71.92	<u>0.253</u>	0.759

Table 7: 与 SD 系列模型进行比较的定量评估。最佳结果以粗体显示，次佳结果为 <u>underlined</u> 。

A wooden ladder with a hand-woven, multi-colored blanket hanging from it. The blanket is made of yarn and has a unique design, making it an eye-catching piece. The ladder is placed against a wall, and the blanket is draped over it, creating a cozy and inviting atmosphere. The combination of the wooden ladder and the hand-woven blanket adds a rustic touch to the scene.

A breathtaking view of a mountainous landscape with a clear blue sky above. The mountains are covered in snow, creating a picturesque scene. The sky is filled with clouds, adding depth and beauty to the landscape. The mountains are visible in the distance, with some appearing closer and others further away. The scene is serene and captivating, showcasing the beauty of nature.

A woman wearing a blue hoodie with the letters ”A” and ”U” on it, representing the University of Arizona. She is standing in front of a white background, posing for the camera. The hoodie is a sweatshirt, and the woman appears to be smiling, showcasing her pride in her university. The hoodie is a popular choice for college students, as it is comfortable and easy to wear.

A BMW car parked in a showroom, showcasing its sleek design and luxurious appearance. The car is positioned in the center of the scene, occupying a significant portion of the image. The showroom is well-lit, creating an inviting atmosphere for potential buyers. The car is displayed prominently, highlighting its features and attracting attention to the brand.

Figure 4: 关于四个任务的更多可视化结果，其中前两行显示的是 512 分辨率下的结果，后两行显示的是 1024 分辨率下的结果。