arXiv:2602.06850v1 [cs.CV] 6 Feb 2026

重新思考多条件DiT：通过位置对齐与关键词范围限定消除冗余注意力

Chao Zhou[1], Tianyi Wei[2*], Yiling Chen[1], WenboZhou[1], NenghaiYu[1]

[1]中国科学技术大学网络空间安全学院，中国安徽省合肥市复兴路100号，邮编230026。[2]南洋理工大学计算与数据科学学院，新加坡南洋大道50号，邮编639798，新加坡。

*通讯作者。电子邮箱：tianyi.wei@ntu.edu.sg；贡献作者：chaozhou@mail.ustc.edu.cn；ustccyleric@mail.ustc.edu.cn；welbeckz@ustc.edu.cn；ynh@ustc.edu.cn；

摘要

尽管现代文生图模型在基于提示词的生成方面表现出色，但它们往往缺乏满足特定用户需求所需的精细控制能力，例如空间布局或主体外观。多条件控制能够应对这一问题，但将其集成到扩散变换器（DiT）中，受限于传统的“拼接后注意力”策略；随着条件数量增加，该策略会带来二次级的计算和内存开销。我们的分析表明，这种跨模态交互中有很大一部分在空间或语义层面存在冗余。为此，我们提出位置对齐与关键词范围注意力（PKA），这是一种旨在消除这些冗余的高效框架。具体而言，位置对齐注意力（PAA）通过强制局部图块对齐，将空间控制线性化；关键词范围注意力（KSA）则通过语义感知掩码，剪除无关的主体驱动交互。为促进高效学习，我们进一步提出条件敏感性感知采样（CSAS）策略，对训练目标重新加权，使其侧重于关键去噪阶段，从而大幅加快收敛并提升条件保真度。实验结果表明，PKA实现了10.0× 倍推理加速和5.1× 倍显存节省，为高保真多条件生成提供了一种可扩展且节省资源的解决方案。

关键词：扩散变换器、多条件图像合成、高效注意力机制、推理加速

1 引言

经过多年快速发展，扩散变换器 (DiT)（Esser 等人，2024年；Peebles 与 Xie，2023）已成为图像生成领域的主流架构。尽管其性能十分出色，但现有的大多数扩散变换器主要由文本提示词引导。在许多现实场景中，用户往往需要更精细的控制，例如指定空间排列、布局或视觉参考。这就需要多条件扩散模型，能够灵活地结合文本提示词和视觉条件。

在基于 UNet 的扩散模型中（Podell 等人，2024年），这一挑战通常通过以下方式解决：

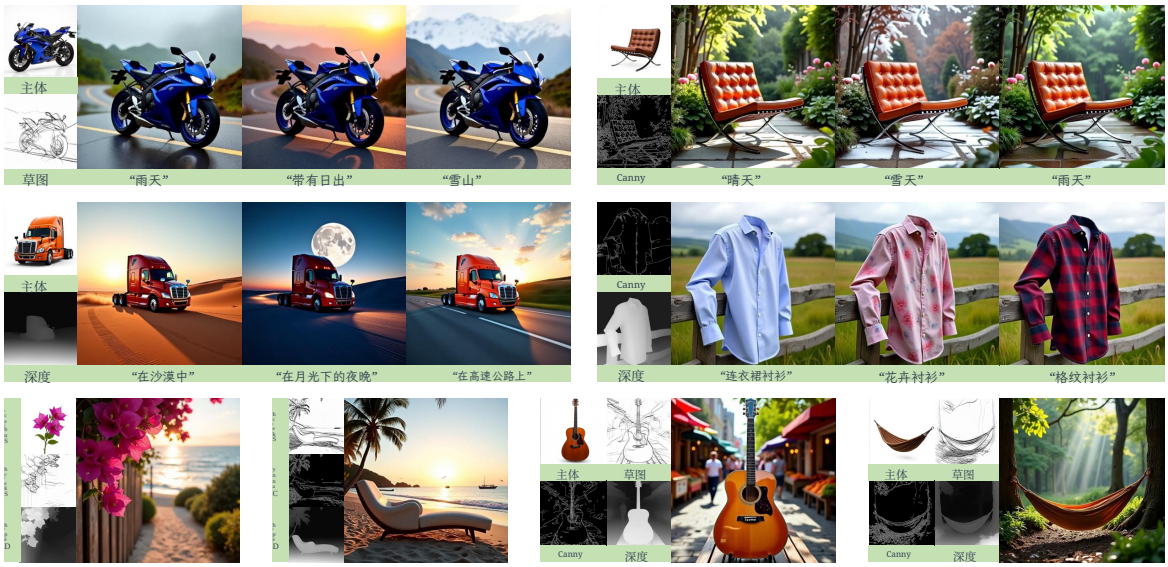

图1 我们提出的 PKA 在多条件生成中的视觉效果。我们提出的 PKA 以显著的效率实现了高质量的多条件生成。放大以获得更清晰的视觉效果。

特征级融合，以 ControlNet（L. Zhang、Rao 和 Agrawala，2023）等方法为例：不同的条件模态通过特征相加或调制注入 UNet 的各个层中（J. He 等人，2025；Q. He 等人，2024年）。由于在 Transformer 架构中进行特征融合并不那么直接，扩散变换器通常采用另一种范式：基于注意力的交互，将所有条件词元和含噪图像词元拼接起来并联合处理（M. Yang 和 Wang，2025）。

然而，这种“拼接并注意”策略的计算开销极高。假设 c个条件输入，每个条件包含 n 个词元，那么由于每个 Transformer 块中所有条件与噪声图像词元之间都要进行成对注意力计算，最终的注意力计算量将按 $\mathcal{O}(c^2n^2)$ 增长。随着条件数量增加（例如布局、参考图像和深度图），总序列长度会大幅增长。注意力机制的计算和内存需求呈二次增长，形成关键瓶颈，导致内存消耗过高、推理延迟增加。这自然引出了一个核心问题：有效的多条件控制是否真的需要如此庞大的注意力计算？

为回答这一问题，我们系统地分析了现有多条件DiT中的注意力模式（Tan、Liu、Yang、Xue 和 Wang，2025）。实证研究表明，“拼接并注意”机制存在明显的结构冗余，主要表现为两种不同模式。对于空间对齐条件（例如布局图），我们观察到注意力权重显著集中在注意力矩阵的对角线上（图2）。这表明，有意义的跨模态交互严格局限于空间位置一致或邻近的图像块。绝大多数非对角线元素——代表远距离、非对齐的交互——几乎不产生激活能量，却仍然需要付出相同的二次计算成本。对于主体驱动条件（例如参考物体），其交互呈现出稀疏激活特征。如图 3 所示，强烈的注意力响应仅限于关键词相关区域，其余图像区域基本不活跃。

受这些观察的启发，我们提出了位置对齐与关键词范围注意力（PKA），这是一种用于高效多条件控制的新机制。PKA 利用其固有的

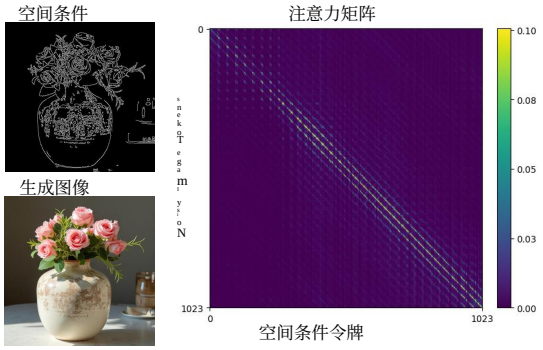

图 2 空间对齐生成中的注意力矩阵可视化。热力图呈现出显著的对角线占优模式，表明注意力激活主要受限于空间位置对应或邻近的图像块。



图 3主体驱动生成中的注意力图。提示词：“在海滩上，一位女士穿着这件衬衫，坐在海滩遮阳伞下。”X 是噪声图像。

注意力模式稀疏性，通过两个不同的条件感知组件来消除计算浪费。

对于空间对齐条件，位置对齐注意力（PAA）以噪声图像与位于相同空间坐标的条件词元之间直接的一对一 对应关系取代全注意力。PAA 仅计算这些对齐词元对之间的注意力，从而以极低的开销实现高度局部化的控制。

对于主体驱动条件，关键词范围注意力（KSA）旨在剪除无关计算。它首先通过以下方式运作：利用注意力图的朴素检测能力，识别与文本关键词和含噪图像词元之间最相关的图像区域（Zhou、Wei 和 Yu，2025）。随后利用该映射创建相关性范围掩码，将后续注意力计算限制在这些显著区域内，从而大幅减少查询—键交互的数量。关键的是，KSA 利用了预训练注意力图的涌现特性，无需额外设计或修改。

此外，我们认为，流匹配（Lipman、Chen、Ben-Hamu、Nickel 和 Le，2023）中常用的标准配置对数均匀时间步采样策略，旨在使训练偏向中等噪声水平，但对多条件微调而言仍未得到充分优化。尽管对数均匀采样能够通用地调整密度，却未考虑模特对异质视觉条件的具体时间敏感性。

为改进这一点，我们进行了系统的扰动分析，以刻画整个去噪轨迹中的条件影响概况。研究发现，模特对条件信号的依赖比对数均匀采样所考虑的更加集中于前期，其敏感度在最早期的结构形成阶段急剧达到峰值。因此，我们提出条件敏感性感知采样（CSAS），通过策略性地重塑采样分布，进一步优先采样这些高敏感度区间。通过使训练密度与分析揭示的内在条件动态相匹配，CSAS 能够更有效地利用梯度来处理复杂控制任务，与标准配置的对数均匀基线方法相比，可实现更快的收敛和更优的条件保真度。

通过整合这些进展，我们的实验验证了，在不影响模特生成性能的前提下，我们能够显著降低注意力机制的计算延迟和内存占用。定量而言，在条件数量较多的场景中，我们的方法实现了最高达 10.0×倍的加速1×，并将注意力模块的内存消耗降低了 5.倍。部分样例如图所示1。

总而言之，我们的贡献如下。

• 我们系统地刻画了多条件DiT中的空间局部性偏差和语义稀疏性，为消除标准全注意力机制固有的冗余提供了实证基础。

• 我们提出了位置对齐与关键词限定注意力框架。通过利用位置对齐注意力和关键词限定注意力，我们的方法实现了近乎线性的计算复杂度。

•我们提出了条件敏感性感知采样。通过将训练目标重新侧重于高敏感度去噪阶段，CSAS 显著加快了微调收敛，并提升了结构控制保真度。

•我们开展了全面的实验，结果表明，我们的方法实现了最先进的效率，最高可达 $10\times$ 倍加速，同时保持甚至提升了生成质量和可控性。

2 相关工作

2.1 可控扩散生成

多条件生成使用户能够通过空间布局或参考主体等多样化输入来引导合成过程。在基于 UNet 的架构中，这通常通过特征级融合实现。这种通用策略已成功应用于广泛的条件。典型示例包括 ControlNet（L. Zhang 等，2023）、UniControl（Qin 等，2023）和 UniControlnet（Zhao 等，2023），它们注入边缘图或姿态等空间引导信息（M. Li 等，2024；Lin、Mo、Klingher、Mu 和 Zhou，2024；W. Wang 等，2025；X. Zhang 等，2025；Zheng 等，2024）。遵循相同原理，IP-Adapter（Ye、Zhang、Liu、Han 和 Yang，2023）、StyleAdapter（Z. Wang 等，2025）、EZIGen（Duan 等，2024）和 InstantID（Q. Wang、Bai、Wang、Qin 和 Chen，2024）等方法利用特征融合，从参考图像中融入主体外观，从而确保身份一致性（Dong、Wei 和 Lin，2025；Jin、Shen、Zhao、Fu 和 Yang，2025；X. Li 等，2025）。尽管这些基于 UNet 的方法在经验上取得了成功，但它们往往依赖临时拼凑的特征融合策略。这增加了架构复杂性并限制了可扩展性，尤其是在组合多个异构条件时。

相比之下，基于 DiT 的模型通常通过基于注意力的交互实现多条件控制。OminiControl（Tan、Liu 等人，2025）和 UniCombine（H. Wang 等人，2025）等框架展示了这一范式的可行性：所有条件词元和潜在token均被拼接起来，通过全自注意力进行联合处理（H. Zhang 等人，2025；Y. Zhang、Yuan、Song、Wang 和 Liu，2025）。然而，这种“拼接后注意力”方法存在一个关键局限：计算成本会随着词元数量呈二次方增长。这会造成显著的内存和运行时开销，使这些方法难以应用于需要丰富多样条件输入的实际场景。

2.2 扩散变换器的高效机制

针对扩散变换器（DiT）固有的巨大计算开销，已有多种缓解策略。一项重要的研究方向聚焦于推理时优化，采用时间特征缓存或分解冗余词元等技术，以避免重复计算（Chen,Li, Jia, Ye, & Ma，2025；Hunter et al.，2025；Liu et al.，2025；Ma, Fang, Mi, &Wang，2024）。另一种相辅相成的方法是简化架构，通过识别并移除或简化对最终生成质量贡献甚微的层来提升效率（Fang, Li,Ma,&Wang，2025；X. Yang etal.，2025；Zhu et al.，2024）。对于专门的多条件生成，PixelPonder（Y. Panet al.，2025）和 OminiControl2（Tan, Xue, Yang, Liu, & Wang，2025）等近期框架通过动态词元剪枝和大幅输入下采样，引入了针对特定任务的加速方法。

尽管这些方法有效，但它们往往会损害复杂条件与生成内容之间的细粒度对齐。与之形成鲜明对比的是，我们的 PKA模块从第一性原理的视角重新评估多条件交互。我们不依赖事后词元

复用或盲目进行架构裁剪，而是利用领域特定结构先验，即布局的空间一致性和主体的语义局部性，从源头上消除冗余。

3 方法

3.1 预备知识

扩散变换器（扩散变换器），例如 FLUX.1（Labs, 2024年）和 Stable Diffusion 3（Esser 等人, 2024年），采用 Transformer 架构作为其去噪骨干。这些模型逐步细化含噪图像词元（$X \in \mathbb{R}^{N\times d}$），并由文本等各种条件词元引导（$C_T \in \mathbb{R}^{M\times d}$）。

在多条件框架（Tan、Liu 等人，2025）中，额外的视觉条件词元（$C_I \in \mathbb{R}^{N_I\times d}$）通过与文本词元和图像词元拼接而被纳入。随后，所有词元经由多模态注意力（MMA）机制联合处理：

$$\mathrm{MMA}([C_T; X; C_I]) = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \quad (1)$$

这种“拼接并注意力”范式的主要问题在于其计算成本。注意力矩阵 $QK^\top \in \mathbb{R}^{(M+N+N_I)\times(M+N+N_I)}$随序列长度呈二次方增长，随着更多条件的加入而变得代价高昂、难以承受。

在训练过程中，这些模型通常使用流匹配（Lipman 等人，2023）来学习去噪过程。通常，每个训练样本的时间步 t 都从标准 logit正态分布 Logit-$\mathcal{N}(0,1)$ 中抽取，以确保模型在生成轨迹的各个阶段都得到训练。

3.2 位置对齐与关键词范围注意力

基于 DiT 的文本生成图像模型，我们提出了位置对齐与关键词范围注意力（PKA），这是一种将标准全注意力分解为一系列轻量级、专用注意力的机制。我们的方法作用于由文本（**T**）、噪声图像（**X**）和空间条件（**SP**）以及主体条件（**SJ**）组成的词元序列。如图 4(b) 所示，我们从根本上重新设计了注意力结构，以降低计算开销。一个关键设计原则是，条件词元（**SP** 和 **SJ**）仅在各自的条件内部执行自注意力。这种结构选择实现了高效的条件缓存机制，如图4(a) 所示。所有条件词元的键和值投影仅在第一次去噪步骤中计算一次，随后进行缓存，并在后续所有步骤中重复使用。这消除了整个去噪轨迹中的冗余计算。含噪图像词元（**X**）通过我们提出的位置对齐注意力（PAA）和关键词范围注意力（KSA）模块，有选择地与条件进行交互，同时与文本（**T**）保持全注意力。

3.2.1 位置对齐注意力

对于空间条件，我们引入位置对齐注意力（PAA）。我们先前的分析表明，空间控制中的交互呈对角占优，这意味着非对齐图块之间的长程依赖是冗余的。因此，PAA 将全面的全局注意力重新构建为一系列局部化、逐位置的交互。

如图 4(c) 所示，不同于关注所有条件词元，每个图像令牌 $\mathbf{X}_i$ 在坐标 i 处仅与其对应的空间条件令牌 $\mathbf{SP}_i$交互。形式上，位置 i 处的 PAA 操作定义如公式 2所示。

$$\mathrm{PAA}(\mathbf{X}_i; \mathbf{SP}_i) = \mathrm{Softmax}\left(\frac{\mathbf{Q}_{X_i}\mathbf{K}_{SP_i}^\top}{\sqrt{d}}\right)\mathbf{V}_{SP_i} \quad (2)$$

通过强制执行这种一对一 位置映射，PAA 有效规避了标准“拼接后注意力”机制的 $\mathcal{O}(N^2)$ 复杂度。这种设计将计算开销线性化为 $\mathcal{O}(N)$，使模特能够纳入任意数量的空间条件，而不会出现标准扩散变换器中常见的、令人望而却步的内存增长。

3.2.2 关键词范围注意力

对于主体条件，我们提出关键词范围注意力（KSA），以利用语义

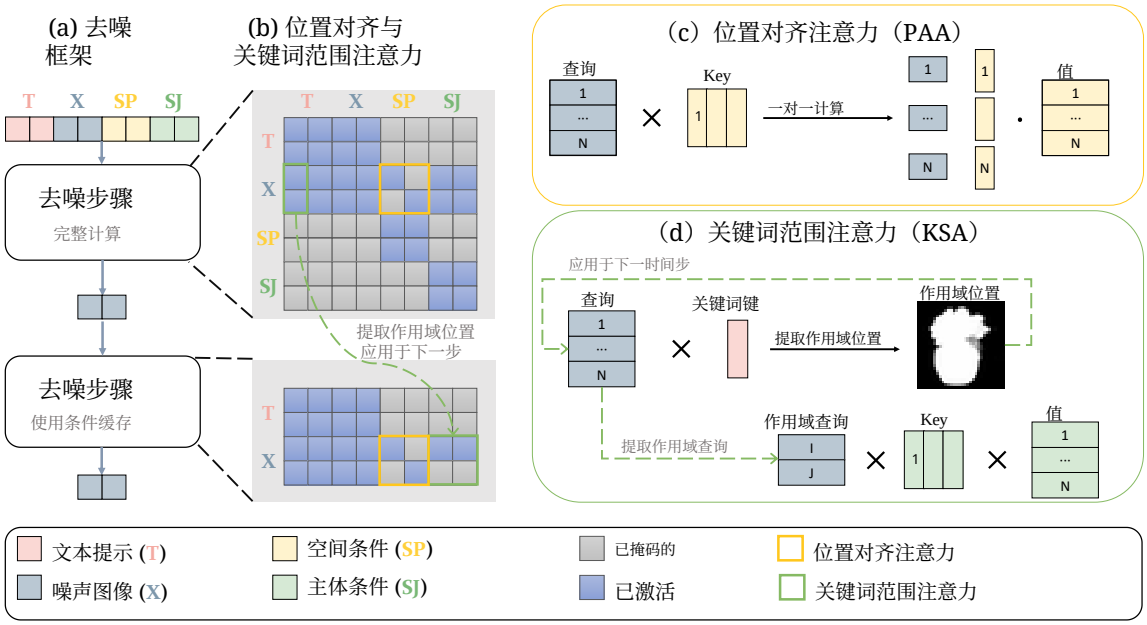

图 4 我们方法的概览。(a) 去噪框架。仅在第一步执行完整计算；随后缓存所有条件词元的键和值，以供后续步骤使用。(b) 位置对齐与关键词范围注意力。我们的分解式注意力机制中，条件仅执行自注意力（从而支持键值缓存）。随后，含噪图像词元 (X) 分别通过 PAA 和 KSA 与空间 (SP) 条件和主体 (SJ) 条件交互。(c) 位置对齐注意力 (PAA)。PAA 在图像 (X) 词元与空间条件 (SP) 词元的对齐位置之间执行高效的一对一注意力。(d) 关键词范围注意力 (KSA)。KSA 在一个步骤中根据文本关键词计算相关性掩码。随后在后续步骤中应用此掩码，将图像 (X) 与主体 (SJ) 之间的注意力计算限制在最相关的区域内。

稀疏性。我们的核心洞见是，特定主体的影响通常局限于其局部语义足迹，而非整个画布。因此，标准配置的全局注意力会因关注无关的背景区域而造成大量计算浪费。

为解决这一问题，KSA 提出了一种动态掩码策略，该策略以去噪轨迹的时间一致性为基础（Zhou 等人，2025）。如图4(d)所示，我们采用解耦式掩码生成机制。在时间步 t，根据公式3，我们首先通过捕捉图像查询$\mathbf{Q}^t$ 与一组稀疏关键词词元 $\mathbb{K}$之间的亲和度，推导出二值语义掩码$\mathbf{M}^t$，这些关键词词元锚定了主体的身份。

$$\mathbf{M}^t = \mathbb{I}\left[\text{Softmax}\left(\sum_{T_i \in \mathbb{K}} \frac{\mathbf{Q}_X^t \mathbf{K}_{T_i}^{t\top}}{\sqrt{d}}\right) \geq \epsilon\right] \quad (3)$$

其中， $\mathbb{I}[\cdot]$ 是指示函数， ϵ表示激活阈值。实际上， $\mathbb{K}$ 极为稀疏（通常只有 1–2 个词元），因此掩码生成的开销可以忽略不计。除非另有说明，实验中我们均使用 $\epsilon = 0.2$ 。

利用相邻时间步之间语义布局保持稳定这一观察，我们将 $\mathbf{M}_t$ 应用于后续步骤 $t+1$，以剪枝注意力流形。最终的 KSA 操作仅在已激活的图像词元 $\hat{\mathbf{Q}}_X^{t+1} = \mathbf{Q}_X^{t+1} \odot \mathbf{M}^t$上执行，从而有效地将主体的影响限制在其局部语义足迹内。

$$\text{KSA}(\mathbf{X}; \mathbf{SJ}) = \text{Softmax}\left(\frac{\hat{\mathbf{Q}}_X^{t+1} \mathbf{K}_{SJ}^{t+1\top}}{\sqrt{d}}\right) \mathbf{V}_{SJ}^{t+1} \quad (4)$$

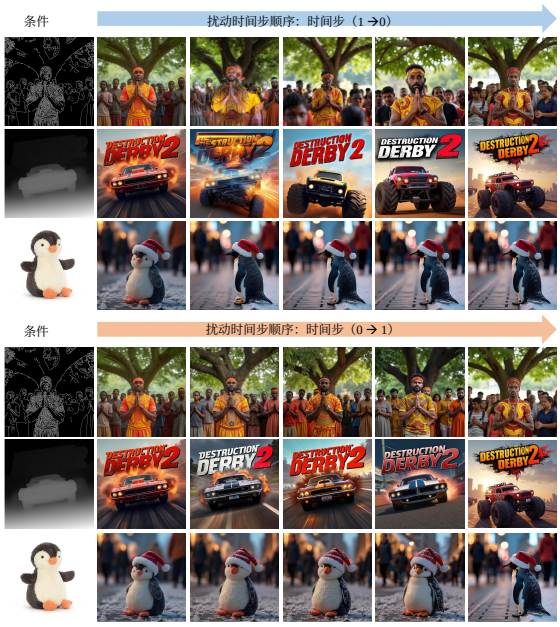

图 5 视觉条件扰动的定性结果。从左至右依次为：视觉条件、0（无扰动）、7、14、21 个扰动步骤，以及 28 个步骤（无视觉条件）。放大查看以获得更好的可视化效果。

3.3 条件敏感性感知采样

传统的流匹配框架（Lip- man 等人，2023）通常采用 logit-normal 时间步采样策略，其中 $t \sim \text{Logit-}\mathcal{N}(0, 1)$，以使训练偏向中等噪声水平。然而，我们的研究表明，这种通用分布对于多条件控制微调而言并非最优。如图5所示，模型（Tan、Liu 等人，2025）对条件引导的依赖在去噪轨迹中并不均匀；相反，这种依赖高度集中于噪声占主导的初始阶段（$t \to 1$）。

为了使优化目标与这种不对称敏感性相一致，我们提出了条件敏感性感知采样（CSAS）。CSAS 不采用均匀采样或标准的 logit-normal 采样，而是策略性地重塑采样密度，以优先关注高噪声下的结构形成阶段。形式上，我们从一个参数化分布中抽取时间步

平移对数几率正态分布：

$$t \sim \text{Logit-}\mathcal{N}(\mu, \sigma^2), \quad \text{with } \mu > 0, \sigma > 1 \tag{5}$$

通过将均值 μ 向 ODE 轨迹的起始处移动，CSAS 将梯度信号集中在条件对齐最为关键的时间段。这种有针对性的学习避免了训练能力消耗在后期的像素细化上，从而显著加快收敛并增强结构可控性。

4 实验

4.1 实验设置

训练详情。 我们从 Subject200K 数据集（Tan、Liu 等人，2025）中筛选出一个子集，确保每条图像描述都包含一个描述性关键词。随后将该子集划分为训练集和测试集。为确保公平比较，我们使用低秩适配（LoRA）微调 FLUX.1（Labs，2024年）模型（Hu 等人，2022），并使用 Prodigy（Mishchenko 和 Defazio，2024年）优化器进行 20,000 次迭代，批量大小为 1，梯度累积步数为 4。

评估详情。 我们遵循先前工作的设置，对主体-空间条件控制（主体-边缘图到图像和主体-深度图到图像）以及多空间条件（边缘图-深度图到图像）进行评估。对于主体-空间条件控制任务，我们将本方法与 OminiControl2（Tan、Xue 等人，2025）、UniCombine（H. Wang 等人，2025）和 EasyControl（Y. Zhang 等人，2025）进行比较。对于多空间条件任务，我们还将其与 PixelPonder（Y. Pan 等人，2025）和 UniControlnet（Zhao 等人，2023）进行比较。效率指标包括推理延迟和注意力显存占用，均在单张 NVIDIA RTX 6000 Ada GPU 上测量。

指标。 为评估主体一致性，我们计算生成图像与真实图像之间的 CLIP-I（Radford 等人，2021年）和 DINOv2（Oquab 等人，2024年）分数。为衡量可控性，我们计算从生成图像中提取的映射与原始条件输入之间的边缘条件 F1 分数以及深度条件的均方误差分数。

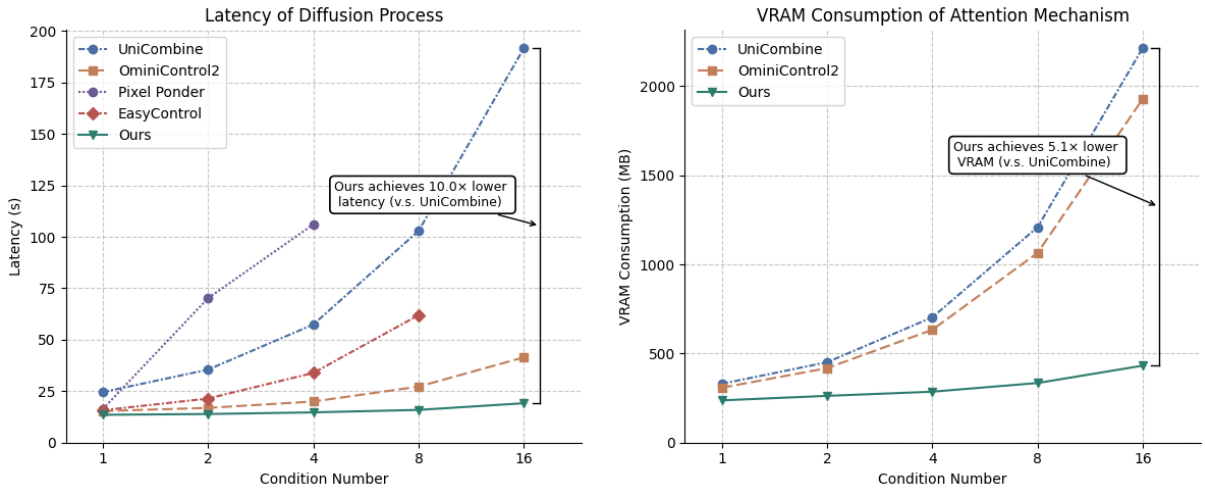

图6 不同空间条件数量下推理延迟（左）和显存占用（右）的可扩展性分析。

为评估生成质量，我们计算生成图像集与真实图像集之间的 FID（Heusel、Ramsauer、Unterthiner、Nessler 和 Hochreiter，2017年）和结构相似性指数（Z. Wang、Bovik、Sheikh 和 Simoncelli，2004年）。此外，我们采用 CLIP-T（Radford 等人，2021年）分数来估计生成图像与文本提示之间的文本一致性。

4.2 主要结果

4.2.1 效率与可扩展性

为评估我们框架的计算效率和可扩展性，我们针对位置对齐注意力（PAA）进行了受控评估。我们分析了空间条件数量从1增至16时的推理延迟和注意力相关显存占用，结果如图6所示。请注意，在这项可扩展性测试中，我们聚焦于PAA，以确保比较具有确定性，因为关键词范围注意力（KSA）的开销本质上取决于数据（随主体尺寸而变化）；相关详细讨论见消融实验。

就推理延迟而言（左图），我们的方法（绿色）的执行时间几乎保持恒定，基本不受条件数量增加的影响。与之形成鲜明对比的是，UniCombine（蓝色）、OminiControl2（黄色）、EasyControl（橙色）和 PixelPonder（紫色）等现有方法的延迟显著增长，且呈非线性增长。在 4 个条件时，我们的方法比 UniCombine 快约 $4\times$，比 PixelPonder 快 $7\times$。当条件数量达到 16 个时，效率差距扩大至 $10.0\times$，凸显了我们线性化注意力设计的优越性。值得注意的是，PixelPonder 最多只能扩展到 4 个条件，而 EasyControl 在 16 个条件时会遇到显存不足（OOM）错误。

在显存占用方面（右图），我们分析了关键瓶颈：注意力机制的开销。我们的方法在这方面同样表现出色，其内存消耗随条件数量近乎线性增长（$\mathcal{O}(N)$）。相比之下，UniCombine和OminiControl2都受制于“拼接后注意力”范式固有的二次增长（$\mathcal{O}(N^2)$），导致内存需求急剧攀升。在16个条件下，我们的方法比UniCombine少消耗 $5.1\times$显存。值得注意的是，我们未将PixelPonder纳入此次比较，因为其基于适配器的融合设计会带来难以承受的显存开销，使其在这一多条件场景中缺乏竞争力。

4.2.2 定性比较

我们在两个具有挑战性的场景下进行了定性比较：主体-空间（图7）和多空间条件（图8）控制。与其他方法相比，我们的方法对文本提示具有更高的遵循度

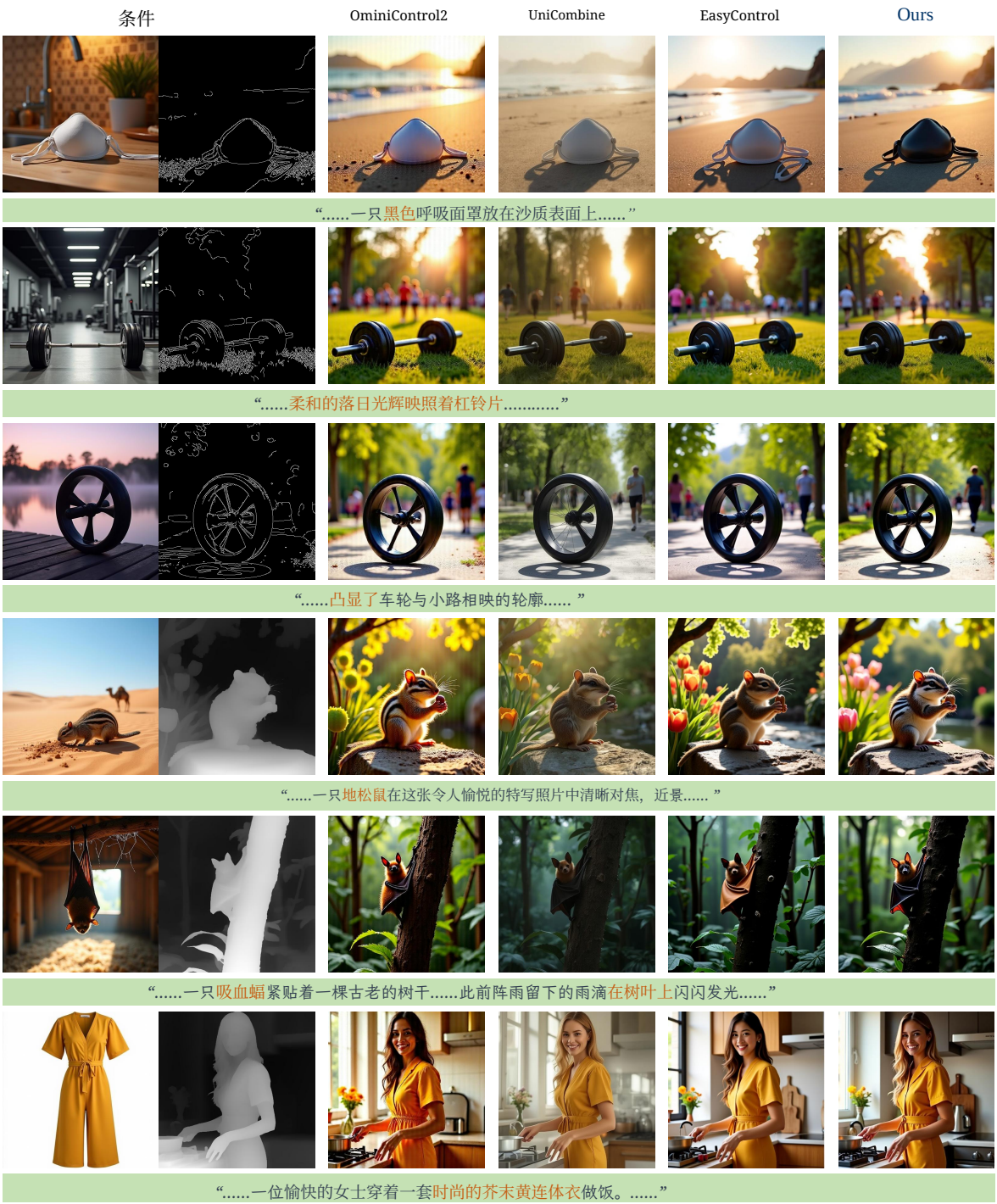

图7 主体-空间条件控制的定性比较。我们的方法具有更高的保真度和更好的图像质量。

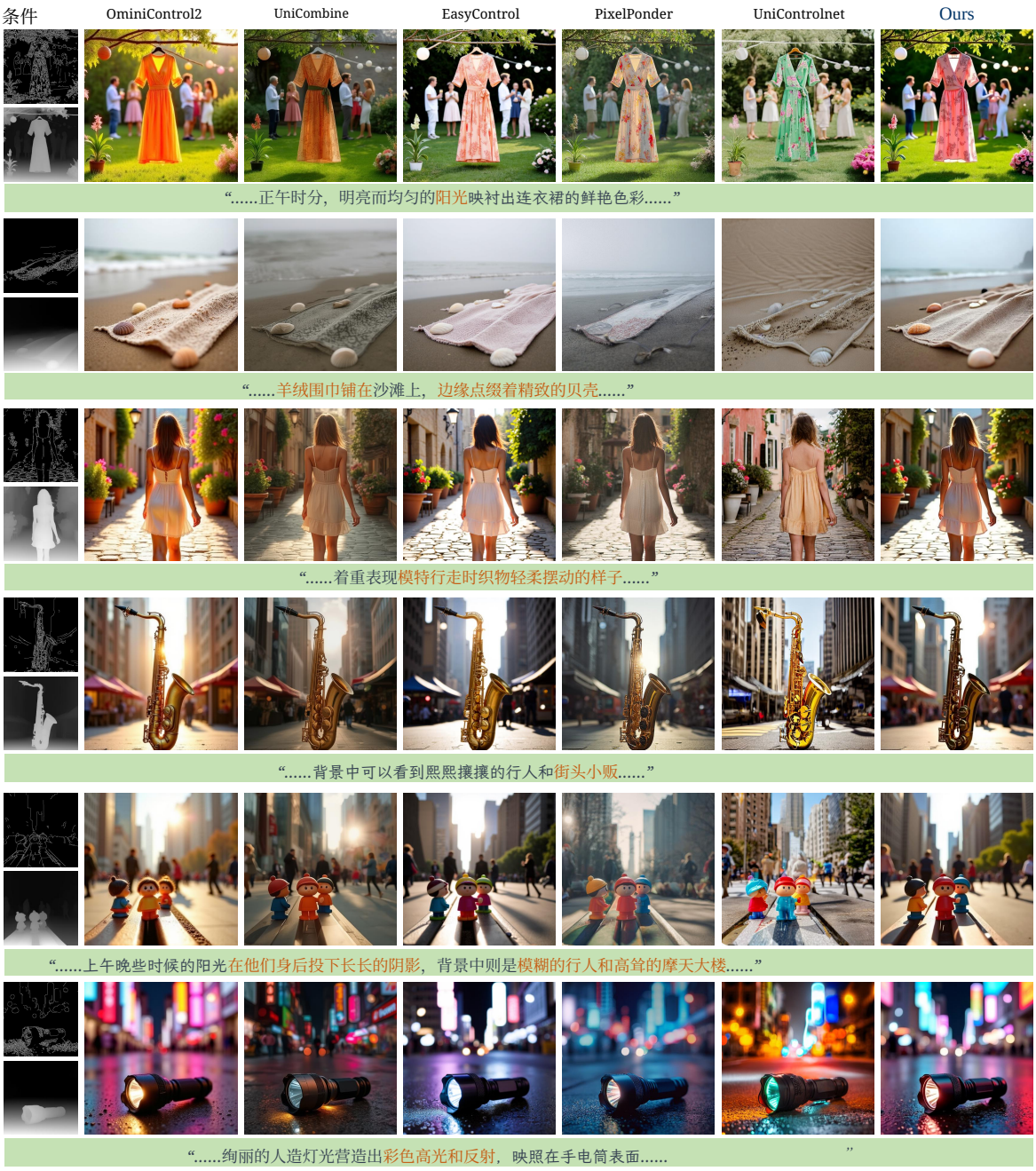

图8 多空间条件控制的定性比较。我们的方法展现出更高的保真度和更优的图像质量。

表 1 不同方法在各项任务和指标上的比较。粗体表示最优结果 lt.

Task	方法	质量 FID↓	结构相似性指数↑	可控性 F1↑	MSE↓	一致性 CLIP-I↑	DINOv2↑	保真度 CLIP-T↑
主体-Canny	OminiControl2	72.03	0.406	0.192	-	0.878	0.867	0.327
	UniCombine	61.03	0.493	**0.551**	-	0.912	0.901	**0.352**
	EasyControl	57.53	0.506	0.287	-	0.935	0.915	0.347
	Ours	**52.99**	**0.553**	0.414	-	**0.945**	**0.926**	0.349
主体-深度	OminiControl2	80.20	0.391	-	366	0.867	0.838	0.325
	UniCombine	70.22	0.454	-	312	0.911	0.879	**0.350**
	EasyControl	68.36	0.464	-	330	0.922	0.838	0.325
	Ours	**62.08**	**0.515**	-	**160**	**0.935**	**0.904**	0.348
多空间	OminiControl2	71.87	0.475	0.194	303	-	-	0.342
	UniCombine	67.40	0.508	0.369	250	-	-	**0.354**
	EasyControl	62.38	0.516	0.291	233	-	-	0.344
	PixelPonder	65.60	0.506	0.385	321	-	-	0.350
	UniControlnet	90.71	0.337	0.269	563	-	-	0.345
	Ours	**53.01**	**0.613**	**0.411**	**114**	-	-	0.353

以及视觉条件，持续生成视觉质量更丰富、更鲜明的图像。

如图7所示，我们的框架展现出卓越的能力，能够同时保留参考主体的身份特征并遵循空间布局。相比之下，其他方法在双重约束下往往会出现语义漂移或结构变形，而我们的模型对这两种模态都保持了较高的保真度。这一点在图7的第三行尤为明显：只有我们的方法能够忠实地重建车轮及其对应的阴影中复杂的 Canny 结构，而其他方法生成的几何形状则模糊不清或不完整。同样，在第四行中，我们的模型展现出出色的主体保真度，能够精准捕捉独特的高频纹理，例如地松鼠的条纹，而这些纹理在基线结果中往往会丢失或过度平滑。

在涉及多个空间条件的场景中，现有模型常会遭遇特征干扰，导致色彩组合显得沉闷、对比度降低。相比之下，我们的 PKA 框架通过将交互线性化，避免了密集全局注意力通常带来的噪声注入。如图 8所示，与基线方法相比，我们的模型生成的图像色调范围更丰富，色彩饱和度也更鲜明。不同条件之间的一致性也显著增强，证明了我们的轻量化设计能够有效平衡多个控制信号，同时不牺牲底层 DiT 的表达能力。

4.2.3 定量评估

表1 全面展示了我们的方法在各种多条件任务中的定量评估结果。我们的方法在生成质量和主体一致性方面始终优于当前最先进的基线方法，同时在可控性和文本保真度方面也保持了出色或极具竞争力的表现。

具体而言，在图像质量方面，我们的方法在所有类别中均取得了最低的 FID 分数，在主体-Canny 和多空间任务中分别达到 52.99 和 53.01。这表明，与 UniCombine 和 OminiControl2 相比，我们的方法在视觉真实感方面有显著提升。在主体一致性方面，我们的模型表现领先，CLIP-I 和 DINOv2 分数均为最高，这证明键词范围注意力（KSA）能够通过将去噪过程聚焦于语义相关区域，有效保留参考主体的身份特征。

在可控性方面，我们的方法在主体-深度和多空间场景中确立了新的最先进水平，深度控制的均方误差（114）显著低于 UniCombine（250）。尽管 UniCombine 在主体 -Canny 任务的 F1 分数上略占优势，但我们的方法在结构遵循度与视觉美感之间实现了更优的平衡

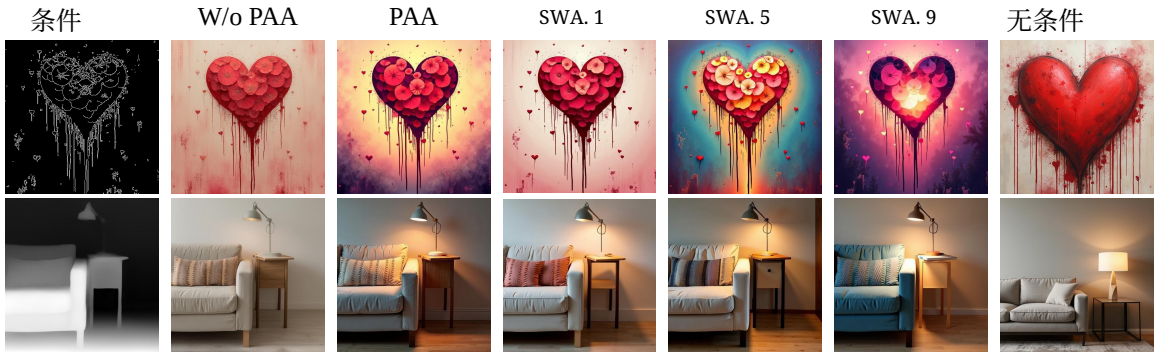

图 9 PAA 的定性消融实验。我们将位置对齐注意力（PAA）与不同窗口大小（$k \in \{1, 5, 9\}$）的滑动窗口注意力（SWA）进行比较。



图 10 效率与性能权衡分析。条形图展示了不同空间交互策略的推理延迟（上）和相关注意力的显存占用（下）。

，避免了强控制常伴随的过度饱和或伪影。此外，我们的模特在文本保真度（CLIP-T）方面达到了与领先基线方法相当的水平，数值上的微小差异在感知上无法区分。

4.3 消融实验

4.3.1 位置对齐注意力的影响

为评估位置对齐注意力（PAA）的有效性和必要性，我们针对两类基线方法进行了对比研究：标准DIT（记作“不使用 PAA”）和滑动窗口注意力（SWA）（X. Pan、Ye、Xia、Song 和 Huang, 2023），其感受野各不相同（$k \in \{1, 5, 9\}$）。

如图 9所示，从定性结果来看，所有变体——包括标准全注意力和多尺度滑动窗口注意力——都能生成高保真图像，并严格遵循给定的空间条件（例如心形布局和内部几何结构）。不同注意力范围下呈现出的这种视觉一致性印证了一项重要的经验观察：标准交叉注意力中的非对角交互对于空间定位而言大多是冗余的。图 10进一步凸显了我们设计在定量指标上的优势。尽管

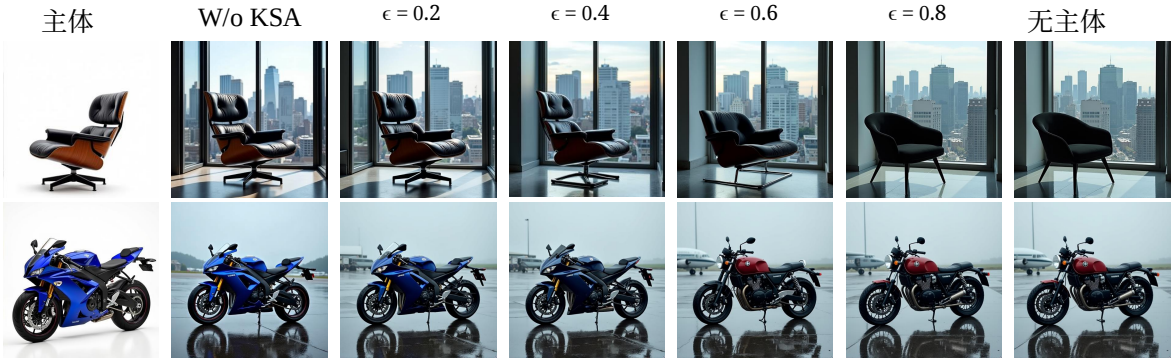

图11 KSA 掩码阈值 ϵ的可视化消融实验。我们评估激活阈值对主体保真度的影响。随着 ϵ 增大，语义掩码会变得更加严格，逐步滤除主体的外围特征。

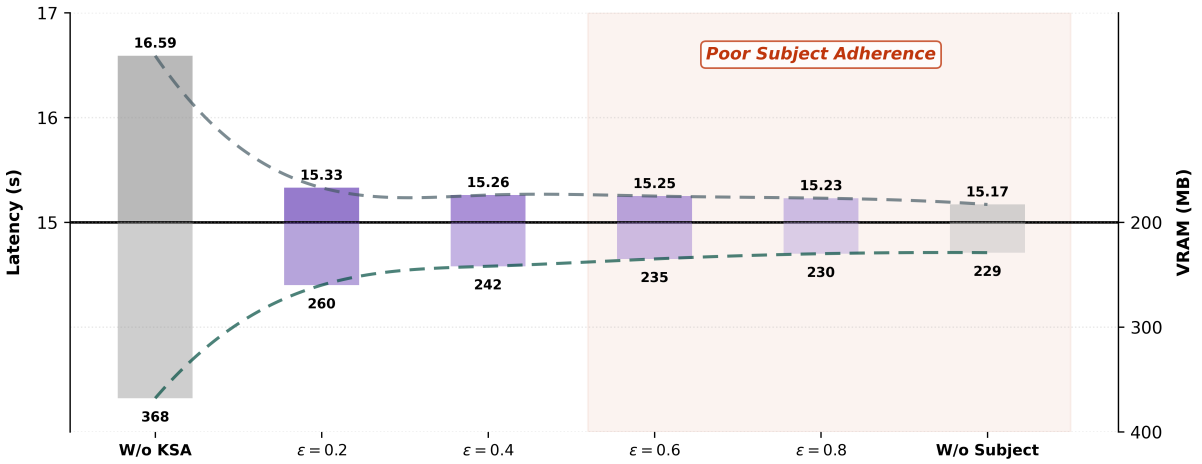

图12 效率衡取舍 以及不同 KSA 阈值下的显存占用可扩展性。镜像条形图展示了推理延迟（上）与显存占用（下）之间的权衡。

我们的 PAA 在视觉质量上可与更复杂的多尺度滑动窗口注意力基线方法相媲美，却实现了最低的推理延迟（13.63 秒）和最小的显存占用（237 MB）。

4.3.2 关键词范围注意力的效果

关键词范围注意力（KSA）模块提供了一种合理的机制，用于权衡计算效率与主体保真度。通过调节掩码阈值 ϵ，KSA 能够实现可调节的稀疏性先验，从而有选择地剪除跨模态交互。

如图 12所分析，基线配置（“无 KSA”，等同于 $\epsilon = 0$）带来了沉重的计算负担，延迟需 16.59 秒，显存占用达 368MB。启用 KSA 并设置 $\epsilon = 0.4$后，我们观察到显著的帕累托改进：显存占用降低了 34.2%（降至 242MB），延迟也缩短至 15.26 秒。值得注意的是，这一显著的效率提升并未导致参考身份灾难性崩溃。如图 11的定性结果所示，即使在较为激进的阈值设置下，生成图像仍与参考主体高度一致。从 $\epsilon = 0$ 到 $\epsilon = 0.4$ 的转变表现为非必要细节的平滑退化，例如精确的

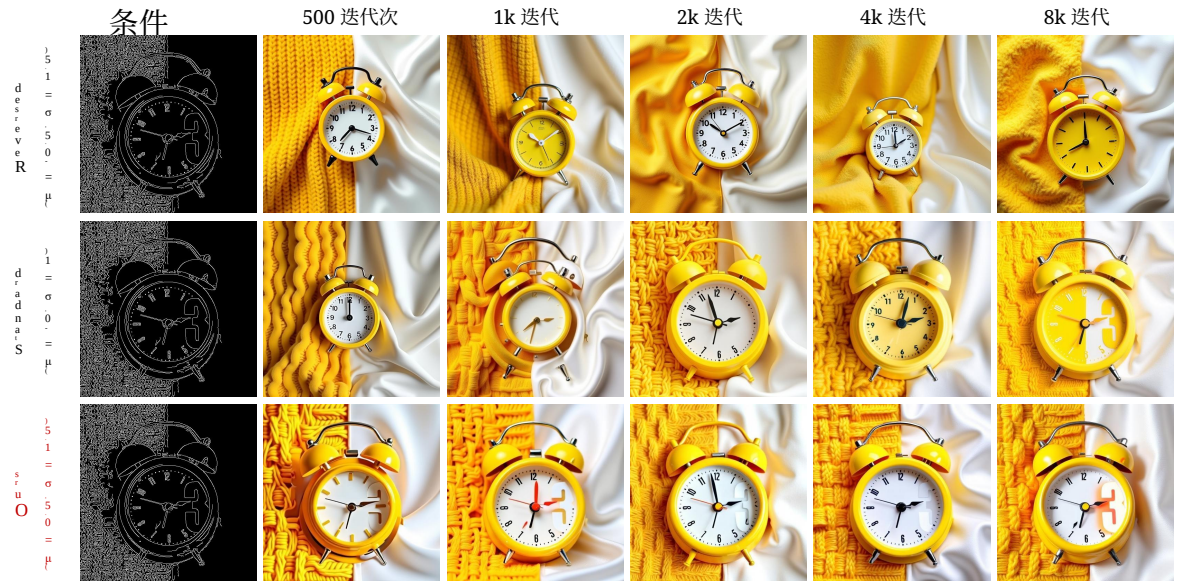

图13 CSAS采样策略在不同微调迭代阶段的定性消融。我们评估了不同时间步采样分布（$\mu$and σ）下的结构接地能力。

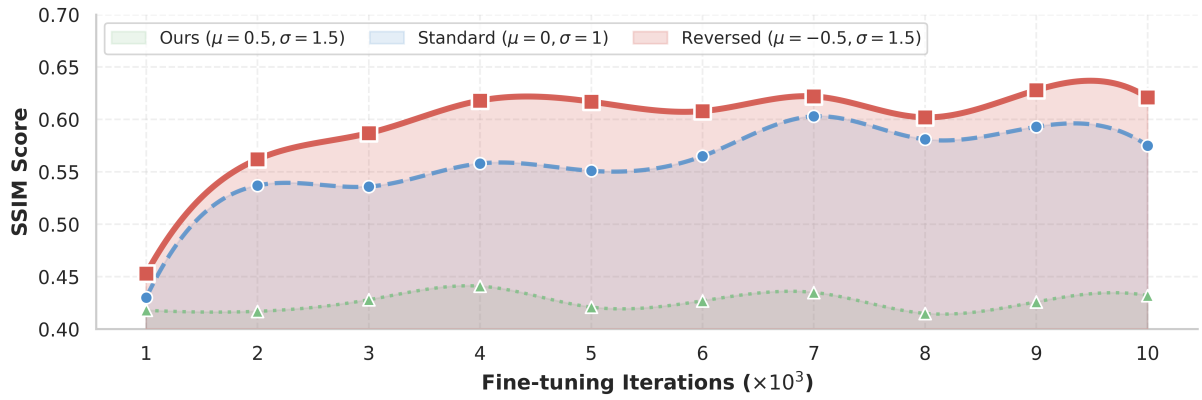

图14 通过SSIM分数对CSAS进行定量收敛分析。该图展示了结构遵循度随 10^3 微调迭代次数的变化。

椅子腿部的曲率，或摩托车挡风玻璃上的细微镜面反射，而不是核心语义特征的丢失。

这些结果凸显了KSA的超参数鲁棒性，ϵ 可以作为一个直观的控制旋钮，让用户能够有策略地分配计算资源。KSA成功识别并消除了注意力流形中的语义冗余，其所处的“保真度稳定”区域（其中 $\epsilon \in [0.2, 0.4]$）范围很广。

4.3.3 条件敏感性感知采样的效果

为了评估采样分布对微调效果的决定性影响，我们比较了 CSAS策略的三种代表性配置：我们的（$\mu = 0.5$，$\sigma = 1.5$，偏向早期高噪声阶段）、标准配置（μ =0，$\sigma = 1$，标准logit正态分布），以及

反向（$\mu = -0.5$，$\sigma = 1.5$，偏向后期低噪声阶段）。

图13和图14中的结果有力地证明了我们的方法在收敛速度和最终生成质量方面均具有明显优势。如图14中的收敛曲线所示，我们的方法（红色曲线）的斜率显著优于基线方法。我们的 SSIM 分数在仅迭代 2k 次后便达到 0.56，而标准配置需要两倍以上的训练时间（约 6k 次迭代）才能达到这一性能水平。这一“快速起步”优势也在图13中得到了直观印证：在极早期的 1k 次迭代阶段，只有我们的方法能够重建闹钟的几何轮廓。除了速度优势外，我们的方法还达到了显著更高的饱和性能，最终 SSIM 为 0.62，明显优于标准采样的 0.58。相比之下，反向策略遭遇了灾难性的学习失败，SSIM 分数停滞在 0.45 以下。

5 结论

本文针对多条件扩散变换器计算效率低下的问题，提出了位置对齐与关键词范围注意力（PKA）这一新型机制，将全注意力分解为两个高效模块：用于空间条件的位置对齐注意力（PAA），以及用于主体驱动条件的关键词范围注意力（KSA）。大量实验验证了这一方法，结果表明，与最先进的方法相比，该方法在保持甚至提升生成质量和可控性的同时，可使推理速度最高提升 10.0×，并将注意力模块的显存占用减少 5.1× 。展望未来，PKA 在多条件控制方面带来的显著效率提升，使其有望成为应对更复杂生成任务的基础。一个尤为令人期待的未来方向是将该框架扩展到视频生成领域，在这一领域中，PKA 的原理可用于以可控的计算成本确保各帧之间的时间一致性。总之，PKA 提供了一种可扩展且实用的解决方案，为新一代复杂且资源友好的人工智能应用铺平了道路。

声明

基金资助

本研究部分由国家自然科学基金项目（62372423、62121002 和 62072421）资助，另获中央高校基本科研业务费专项资金 WK2100250070 资助。

利益冲突

作者声明不存在任何利益冲突。

数据可用性

本文所使用的数据集收集自公开可用的在线来源。论文接收后，我们将以开源形式发布该数据集，以进一步惠及研究界。

代码可用性

本研究所使用的代码将在论文发表后发布，以支持可复现性和进一步研究。

作者贡献

Chao Zhou 和 Tianyi Wei 提出了研究构想，开发了所提出的方法，开展了实验，并参与了论文撰写。Yiling Chen、周文博和余能海提供了科学指导，监督了研究设计与实验验证，并协助修订了论文。所有作者均审阅并批准了论文的最终版本。

可复现性

我们通过明确区分结果与解读，并详细报告实现组件（如评估指标和计算资源），确保研究具有可复现性。

多条件图像生成中的伦理考量

多条件扩散模型的发展虽然带来了巨大的创作潜力，但也有必要对其

伦理影响进行严谨探讨。我们提出的 PKA 框架旨在实现轻量级的主体驱动生成和空间控制，理论上可能被滥用于未经授权地创作合成媒体或深度伪造。由于该技术能够精准控制角色身份和空间布局，因此存在被用于生成误导性内容，或在未经同意的情况下侵犯个人生物特征隐私和知识产权的风险。

为降低这些风险，我们强调，我们的模型旨在伦理边界内用于学术研究和创意辅助。我们倡导整合数字水印和主动内容审核协议，以确保可追溯性，并防止传播有害或未经同意的图像。此外，由于底层的扩散变换器（DiT）是在大规模公开数据集上训练的，模型可能会无意中继承或放大有关性别、种族或文化刻板印象的社会偏见。我们鼓励未来的研究人员采用去偏技术和多样化数据集，推动生成式人工智能技术朝着更加公平、负责任的方向发展。

参考文献

Chen, Z.、Li, K.、Jia, Y.、Ye, L.、Ma, Y.（2025）。通过结合通道感知奇异值分解的增量校准缓存加速扩散变换器。计算机视觉与模式识别会议论文集（第 18011–18020 页）。

Dong, Z., Wei, P., Lin, L.（2025）。Dreamartist：通过正负适配器实现可控的单次文本到图像生成：Z. dong 等。国际计算机视觉期刊，*133*（10），7037–7053,

Duan, Z., Ding, Y., Gou, C., Zhou, Z., Smith, E., Liu, L. (2024年). EZIGen：通过精确的主体编码和解耦引导增强零样本个性化图像生成。*arXiv* 预印本 *arXiv:2409.08091*, ,

Esser, P., Kulal,S.,Blattmann, A., Entezari, R.,M¨uller, J., Saini, H., . .. 等人（2024年）。扩展用于高分辨率图像合成的整流流 Transformer。第四十一届国际机器学习会议。

Fang, G., Li, K., Ma, X., Wang, X.（2025）。Tiny-fusion：学习浅层表示的扩散变换器。计算机视觉与模式识别会议论文集 （第 18144–18154 页）。

He, J., Tuo, Y., Chen, B., Zhong, C., Geng, Y., Bo, L.（2025）。*Anystory*：迈向文本到图像生成中单主体与多主体个性化的统一。获取地址：https://arxiv.org/abs/2501.09503

He, Q., Peng, J., Xu, P., Jiang, B., Hu, X., Luo, D., . .. Zhang, J.（2024）。Dynamiccontrol：通过自适应条件选择提升文本到图像生成效果。*arXiv* 预印本 *arXiv:2412.03255*, ,

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.（2017）。采用双时间尺度更新规则训练的生成对抗网络收敛到局部纳什均衡。神经信息处理系统进展，*30*, ,

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., .. . Chen, W.（2022）。LoRA：大型语言模型的低秩适配。国际学习表征会议。取自 https://openreview.net/forum?id=nZeVKeeFYf9

Hunter, R., Dudziak, ¨L., Abdelfattah, M.S., Mehrotra, A., Bhattacharya, S., Wen, H.（2025）。通过复用扩散模型中的注意力图实现快速采样。国际计算机视觉杂志，1–10,

Jin, J., Shen, Y., Zhao, X., Fu, Z., Yang, J. (2025). Unicanvas：通过定制化文本到图像生成实现具备可供性意识的统一真实图像编辑。国际期刊

计算机视觉，*133*(6)，3456–3480，

B.F. Labs (2024年). *Flux.* https://github.com/black-forest-labs/flux.

Li, M., Yang, T., Kuang, H., Wu, J., Wang, Z., Xiao, X., Chen, C. (2024年). Controlnet++：通过高效的一致性反馈改进条件控制：项目页面：liming- ai. github. io/controlnet plus plus. 欧洲计算机视觉会议(第129–147页)。

Li, X., Ren, Y., Jin, X., Lan, C., Wang, X., Zeng, W., . .. Chen, Z. (2025). 用于图像复原与增强的扩散模型：全面综述。国际计算机视觉期刊，*133*(11)，8078–8108，

Lin, K., Mo, S., Klingher, B., Mu, F., Zhou, B. (2024年). Ctrl-x：无需引导即可控制文生图生成的结构与外观。神经信息处理系统进展(第37卷，第128911–128939页)。

Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M. (2023). 用于生成建模的流匹配。第十一届国际学习表征会议。取自 https://openreview.net/forum?id=PqvMRDCJ

Liu, D., Zhang, J., Li, Y., Yu, Y., Lengerich, B., Wu, Y.N. (2025年). *Fastcache*：通过可学习的线性近似实现扩散变换器的快速缓存。取自 https://arxiv.org/abs/2505.20353

Ma, X.、Fang, G.、Mi, M.B.、Wang, X.（2024年）。Learning-to-cache：通过层缓存加速扩散变换器。第三十八届神经信息处理系统年会。 检索自 https://openreview.net/forum?id=ZupoMzMNr

Mishchenko, K. 和 Defazio, A.（2024年）。Prodigy：一种快速自适应的无参数学习器。第四十一届国际机器学习大会。 检索自 https://openreview.net/forum?id=JJpOssn0 O uP

Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., . .. Bojanowski, P. (2024年). DINOv2：无需监督地学习鲁棒的视觉特征。机器学习研究汇刊，，取自 https://openreview.net/forum?id=a68SUt6z Ft

(特色认证)

Pan, X., Ye, T., Xia, Z., Song, S., Huang, G.（2023）。Slide-transformer：结合局部自注意力的分层视觉Transformer。*IEEE/CVF*计算机视觉与模式识别会议论文集（第2082–2091页）。

Pan, Y., He, Q., Jiang, Z., Xu, P., Wang, C., Peng, J., ...其他作者（2025）。Pixelponder：用于增强多条件文本到图像生成的动态图像块自适应。*arXiv*预印本 *arXiv:2503.06684*，，

Peebles, W., & Xie, S.（2023）。基于Transformer的可扩展扩散模型。*IEEE/CVF*国际计算机视觉会议论文集（第4195–4205页）。 T9t

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., M¨uller, J., .. . Rombach, R.（2024）。SDXL：改进用于高分辨率图像合成的潜在扩散模型。第十二届国际学习表征会议。取自 https://openreview.net/forum?id=di52zR8xgf

Qin, C., Zhang, S., Yu, N., Feng, Y., Yang, X., Zhou, Y., . .. others（2023）。UniControl：一种用于开放环境中可控视觉生成的统一扩散模型。*arXiv* 预印本 *arXiv:2305.11147*，，

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., .. . others（2021）。通过自然语言监督学习可迁移视觉模型。国际机器学习会议 （第 8748– 8763 页）。

Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.（2025）。Ominicontrol：面向扩散变换器的最简通用控制。*IEEE/CVF* 计算机视觉国际会议论文集（第 14940–14950 页）。

Tan, Z., Xue, Q., Yang, X., Liu, S., Wang, X.（2025）。Ominicontrol2：面向扩散变换器的高效条件控制。*arXiv* 预印本 *arXiv:2503.08280*，

Wang, H., Peng, J., He, Q., Yang, H., Jin, Y., Wu, J., .. . 其他作者（2025）。Unicom- bine：基于扩散变换器的统一多条件组合。*arXiv*预印本 *arXiv:2503.09277*， ， Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A.（2024）。Instantid：数秒内实现零样本身份保持生成。*arXiv*预印本 *arXiv:2401.07519*， ， Wang, W., Zhao, C., Chen, H., Chen, Z., Zheng, K., Shen, C.（2025）。Autostory：以最少的人力生成多样化的叙事图像。国际计算机视觉期刊，*133*(6)，3083–3104， Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.（2004）。图像质量评估：从误差可见性到结构相似性。*IEEE*图像处理汇刊，*13*(4)，600–612， Wang, Z., Wang, X., Xie, L., Qi, Z., Shan, Y., Wang, W., Luo, P.（2025）。Styleadapter：统一的风格化图像生成模型。国际计算机视觉期刊，*133*(4)，1894–1911，

Yang, M., & Wang, Z.（2025）。有限数据条件下的图像合成：综述与分类体系。国际计算机视觉期刊，*133*(6)，3689–3726，

Yang, X., Lu, L., Fan, Q., Yang, C., Lin, J., Wang, Y., ... Gao, S.（2025）。Alter：面向高效扩散生成的全能层剪枝与时序专家路由。*arXiv* 预印本 *arXiv:2505.21817*，

Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.（2023）。IP-Adapter：用于文本到图像扩散模型的文本兼容图像提示适配器。

Zhang, H., Hong, D., Wang, Y., Shao, J., Wu, X., Wu, Z., Jiang, Y.-G. (2025). Creatilayout：用于创意布局到图像生成的孪生多模态扩散 Transformer。*IEEE/CVF*国际计算机视觉会议论文集（第 18487–18497 页）。

Zhang,L., Rao,A., Agrawala, M. (2023). 为文本到图像扩散模型添加条件控制。

Zhang, X., Li, Z., Li, W., Zhang, X., Wei, J., Lin, C., Tang, Y. (2025). Objctrl：用于条件文本到图像生成的基于对象的控制松弛。第*33*届 *ACM*国际多媒体会议论文集（第 10064–10073 页）。

Zhang, Y., Yuan, Y., Song, Y., Wang, H., Liu, J. (2025). Easycontrol：为扩散变换器添加高效灵活的控制。*IEEE/CVF*国际计算机视觉会议论文集（第 19513–19524 页）。

Zhao, S., Chen, D., Chen, Y.-C., Bao, J., Hao, S., Yuan, L., Wong, K.-Y.K.（2023）。Uni-ControlNet：面向文生图扩散模型的一体化控制。神经信息处理系统进展， ，

Zheng, J., Liu, D., Wang, C., Hu, M., Yang, Z., Ding, C., Tao, D.（2024年）。Mmot：用于组合式多模态条件图像合成的模态令牌混合变换器。国际计算机视觉期刊，*132*（9），3537–3565,

Zhou, C., Wei, T., Yu, N. (2025). 扩展你的指令：通过自适应注意力缩放增强统一图像生成模特对指令的遵循保真度。*IEEE/CVF*计算机视觉国际会议论文集（第 15171–15181 页）。

Zhu, H., Tang, D., Liu, J., Lu, M., Zheng, J., Peng, J., . .. 等人（2024年）。Dip-go：一种通过少步梯度优化实现的扩散剪枝器。神经信息处理系统进展，*37*，92581–92604,