FreeText：通过注意力定位与频谱字形注入实现扩散 Transformer 中的免训练文本渲染

Ruiqiang Zhang [*1] **Hengyi Wang** [*1] **Chang Liu** [1] **Guanjie Wang** [1] **Zehua Ma** [1] **Weiming Zhang** [1]

arXiv:2601.00535v1 [cs.CV] 2 Jan 2026

摘要

大规模文本到图像（T2I）扩散模型擅长开放域生成，但在精确文本渲染方面仍面临困难，尤其是在多行布局、密集排版以及中文等长尾文字系统的处理上。以往方法通常需要昂贵的重新训练或僵化的外部布局约束，这可能损害美学评分并限制灵活性。我们提出**FreeText**，这是一种无需训练、即插即用的框架，通过利用扩散*Transformer*（*DiT*）模型的内在机制来改进文本渲染。**FreeText** 将问题分解为在哪里书写和书写什么。对于在哪里书写，我们从内生的图像到文本注意力中读取逐令牌空间归因，并利用汇聚型词元作为稳定的空间锚点，通过拓扑感知细化生成高置信度掩码，从而定位书写区域。对于书写什么，我们提出频谱调制字形注入（SGMI），通过频域带通调制注入与噪声对齐的字形先验，以强化字形结构并抑制语义泄漏（即渲染概念而非文字）。在 Qwen-Image、FLUX.1-dev 和 SD3 变体上开展的大量实验表明，在 LongText-Benchmark、CVTG 以及我们提出的 CLT-Bench 上，该方法均能稳定提升文本可读性，同时基本保持语义对齐和美学质量，并仅带来适度的推理开销。

1. 引言

近年来，大规模文本到图像（T2I）扩散模型（例如 Stable Diffusion（Esser 等人，2024）、FLUX（Labs 等人，2025年）和 Qwen-Image（Wu 等人，2025年）已实现高质量的开放领域图像合成。然而，精确的文本渲染仍然颇具挑战，常见问题包括拼写错误、笔画缺失、失真以及“语义漂移”（渲染出概念而非文字）。在多行、文本密集、多语言和语义复杂的场景中，这些问题尤为突出。对于中文等表意文字系统，问题尤其严重：字符分布呈高度长尾特征，许多生僻字和低频组合在训练中代表不足；与此同时，许多字符在视觉上相似，且内部部件和笔画结构复杂（Chen 等人，2021）。因此，模型往往无法从有限的覆盖范围中学到可靠的字形先验，并容易出现细粒度混淆，导致生成的文本即使经过反复采样也常常无法使用。

从应用和研究的角度来看，文本渲染并非锦上添花，而是检验文本到图像模型细粒度可控性、复杂场景规划和跨模态对齐能力的关键压力测试。文本是一种高度结构化的视觉对象，其笔画、字形及排列方式对局部几何结构和全局布局约束提出了严格要求。此外，人类对文本错误极为敏感：在海报和用户界面设计等现实场景中，文本往往充当关键标识，拼写错误或畸形字形会严重降低可用性。因此，更出色的文本渲染对于实际应用至关重要，因为即使是轻微的拼写错误，也可能使原本出色的图像失去价值。

现有提升文本渲染效果的方法大多依赖两种手段：额外训练或微调（基于重新训练），以及明确的布局或位置条件（布局条件控制）。TextDiffuser（Chen 等人，2023年）和 AnyText（Tuo 等人，2023年）等方法利用框、掩码或字形监督训练布局预测器或控制分支，从而提升可控性和 OCR 准确率。这些方法需要付出高昂的数据和计算成本，而且往往会使生成分布和视觉风格偏离基础模型。在推理阶段，它们还会进一步注入边界框、掩码或字形作为硬条件，机械地将文本区域固定在预设位置。这类外部约束可能抑制模型内在的场景规划行为，使模型难以在复杂背景或含糊、相互冲突的提示词下兼顾多样性与自然度。

[1]中国科学技术大学数字安全安徽省重点实验室。通讯作者：Zehua Ma <mzh045@ustc.edu.cn>、Weiming Zhang<zhangwm@ustc.edu.cn>。

预印本。*2026*年*1*月*5*日。

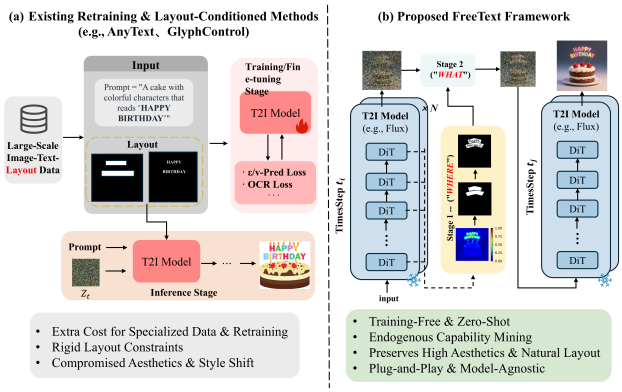

图 *1.* 系统概览。(a) 以往的文本渲染方法通常需要重新训练和/或严格的布局条件。(b) FreeText 将文本渲染分解为位置 和 内容：通过内生注意力图定位文本区域，然后以兼容模型的方式注入字形结构先验，从而在保留基础版模型美学风格的同时，实现无需训练的增强。

与此同时，经过充分预训练和广泛后训练的大型模型（例如 FLUX 和 Qwen-Image）已经展现出出色的美学质量；对它们施加僵化的布局约束，不仅难以微调，还可能明显损害其美学表现。相反，AnyText 等专用于文本的模型主要针对文本渲染进行训练，通常无法复现大型基础模型完整的预训练和后训练流程，因此往往需要在渲染准确度与美学表现之间取舍，而在复杂的开放域场景中，其美学表现尤其逊色。迄今为止，在不修改架构或参数的前提下，仅利用基础模型的内部机制，同时实现高渲染准确度与出色美学表现，仍进展有限。

受这些局限的启发，我们提出一种全新的视角：与其投入高昂成本重新训练模型，让其掌握通用的如何书写，不如将文本渲染分解为两个更基础、且基础版模型本身已有潜力支持的子问题：在哪里书写和书写什么。基于这一观点，我们提出 FreeText，这是一种免训练、即插即用的增强框架。其设计源于一项关键观察：让模型识别文本，远比让它精确渲染文本（生成像素级字形）容易。FreeText 利用易于获取的文本视觉先验和模型的内部结构来解决这两个子问题。

1. 写在哪里。 文本到图像模型未必缺乏布局规划；更准确地说，是我们尚未有效读出其为文本区域制定的内部规划。事实上，在生成过程中，采用 DiT 式架构的扩散模型会在图像到文本交叉注意力中，隐式编码不同文本令牌的空间归属关系（Peebles &Xie，2023年）。跨时间步和网络层深度的注意力图共同描述了模型的内生布局。

基于此，我们提出一种无监督定位策略：FreeText 不依赖脆弱的外部光学字符识别或视觉语言模型（VLM）检测器进行事后检测，而是选取最稳定的注意力层作为空间锚点，并在没有布局标注（零布局监督）的情况下，精准锁定目标文本词元的书写区域。

2. 要写什么。 如附录图 1 所示，模型可能会将词元“Car”渲染成汽车图像，而非文字本身。我们认为，这是由于嵌入空间中语义概念（高层含义）与字形结构（视觉形式）相互耦合所致。在生成早期，强烈而具体的语义先验可能占据主导地位并抑制字形信息，从而导致语义泄漏——也就是说，概念压过笔画，导致“文字变成图像”。为实现局部规则“字形 > 语义”，我们提出了频谱调制字形注入（SGMI）。SGMI 不会简单地混合潜变量，而是在频域中应用带通调制，增强承载字形结构的中高频分量，同时抑制背景和无关噪声的传播，从而引导精准的字形合成。

总而言之，我们的贡献如下：

- 一种无需训练且与基础模型无关的文本渲染增强框架。FreeText 作为推理时插件运行，可无缝集成至 StableDiffusion、FLUX、Qwen-Image 及其他文本到图像模型，无需修改任何参数，并显著提升双语（中文/英文）及具有挑战性的渲染场景中的文本渲染效果。

- 一种基于内生注意力的无监督文本区域定位方法。我们利用 DiT 风格的图像到文本注意力信号，以及类似注意力汇聚点的稳定性线索，在无需任何监督的情况下，实现通用且高精度的文本区域锁定。

- 一种频域字形先验注入方案。SGMI 使用带通频谱调制来突出承载结构信息的字形频率，同时抑制语义背景泄漏，从而提升渲染保真度。

- 中文长尾文本渲染基准。我们提出了 CLT-Bench，这是一项分级评估基准，面向长尾汉字（生僻且结构复杂），系统评估模型性能从常见到生僻、从简单到复杂设置下的下降情况。

2. 相关工作

2.1. 文本生成图像扩散基础模型

近年来，大规模 T2I扩散模型在分辨率、语义对齐和文本渲染方面不断取得提升（Wu et al.，2025年；Seedream et al.，2025年；Esser et al.，2024；Labs et al.，2025年）。以 Stable Diffusion 3、Qwen-Image 和 FLUX.1 为代表的系统，将这些提升归因于更强的 MMDiT/DiT 主干网络、流/修正流目标函数以及大规模专用数据处理流程，从而带来更优的整体视觉质量和排版效果。然而，这类改进通常需要成本高昂的预训练和后训练，并且与特定架构和数据配方紧密耦合，使得文本渲染能力难以低成本地迁移到不同基础模型上。相比之下，FreeText 保持基础模型不变，通过利用内生注意力和潜空间结构进行推理时控制，从而实现跨模型、细粒度的文本渲染增强。

2.2. 重新训练与依赖布局的文本渲染

大多数既有文本渲染方法都采用基于重新训练、依赖布局的范式。TextDiffuser 风格的（Chen 等人，2023年）方法在大型 OCR 标注语料库上学习布局预测模块，因此在生成时需要显式的布局模板或分割先验。AnyText（Tuo 等人，2023年）、GlyphDraw（Ma 等人，2023年）、GlyphControl（Yang 等人，2023年）和 UniGlyph（Yang 等人，2023年）等方法在 Stable Diffusion/DiT 上引入类似 ControlNet 的控制分支或专用条件分支，并利用额外输入（例如字形图像、文本掩码或分割图）进行重新训练，以提升光学字符识别准确率和字体可控性。尽管这些方法效果显著，但它们依赖额外标注和控制分支，将生成过程与外部布局/视觉条件紧密绑定，限制了提示词的自由度和图像多样性，也未能充分利用基础模型内生的场景规划能力。FreeText 无需训练，也不依赖布局：它从内生注意力中定位文本区域，并通过频谱调制注入字形先验，无需任何额外训练成本即可改善文本渲染。

2.3. 注意力汇聚点

注意力长期以来一直是解读 Trans- former 行为的重要视角。在大语言模型中，注意力汇聚现象已被广泛观察到：少数语义较弱的词元会吸收不成比例的注意力，通过缓冲全局上下文来稳定推理（Tigges 等人，2023年；Razzhigaev 等人，2025年；Chauhan 等人，2025年；Zhang 等人，2025年）。多模态模型中的相关分析则利用注意力模式研究跨模态对齐与幻觉（Kang 等人，2025年）。然而，注意力汇很少被用于空间生成，也尚未被系统地用于文本到图像扩散模型中的文本区域定位。FreeText 通过实证发现，基于DiT的文生图模型中的汇聚型词元在不同时间步和层之间会产生相对稳定的边界线索，并将其视为空间锚点，从内生的图像到文本注意力中提取文本区域，无需监督即可为后续的字形先验注入提供可靠的定位。

3. 方法

FreeText旨在增强复杂场景中的文本渲染效果，而无需修改基础文生图扩散模型的架构或参数。给定目标文本片段 s及其字形参考图像，FreeText分两个阶段处理，如图所示。2。

1. **注意力引导的内生文本区域定位**(第3.1节）：我们在采样过程中提取 DiT/MMDiT 模块中的图像到文本（I2T）交叉注意力，聚合并筛选信息丰富的时间步–层对，并应用拓扑感知后处理，以获得潜空间中的高置信度书写掩码$\mathbf{R}_s$ 。

2. **频谱调制字形注入**(第3.2节）：我们将字形参考图像编码到潜空间，将其与当前噪声水平对齐，构建基于 Log-Gabor 的频谱调制字形注入（SGMI）先验，并在一个较短的时间窗口内通过余弦退火将其注入$\mathbf{R}_s$ ，以强化字形结构并抑制语义泄漏（例如，渲染出概念而非对应的词语）。

3.1. 注意力引导的文本区域定位

为回答“写在哪里”，我们直接利用内生注意力定位书写区域（Tang 等人，2023年），无需外部布局预测器、光学字符识别或视觉语言模型检测器。我们从注意力图中读取逐令牌空间归因，然后进行时间步—层选择和拓扑感知细化，以生成高置信度的区域掩码。

3.1.1. 注意力提取

令 $\mathbf{A}^{(t,l)}$ 表示时间步 t和第 l个 DiT/MMDiT 模块处的头部平均I2T注意力：

$$\mathbf{A}^{(t,l)} \in \mathbb{R}^{H\times W\times N_{\text{text}}}, \tag{1}$$

其中， N_{text} 是文本令牌的数量。对于目标跨度s，我们首先定位其令牌子序列 $\mathcal{T}_s$，并将其与若干类汇聚特殊令牌合并；这些令牌在不同层和注意力头中均表现出稳定的高响应。我们将合并后的集合称为锚点˜令牌集 $\mathcal{T}_s$。

˜然后，我们对 $\mathcal{T}_s$ 上的注意力取平均，得到初始的局部

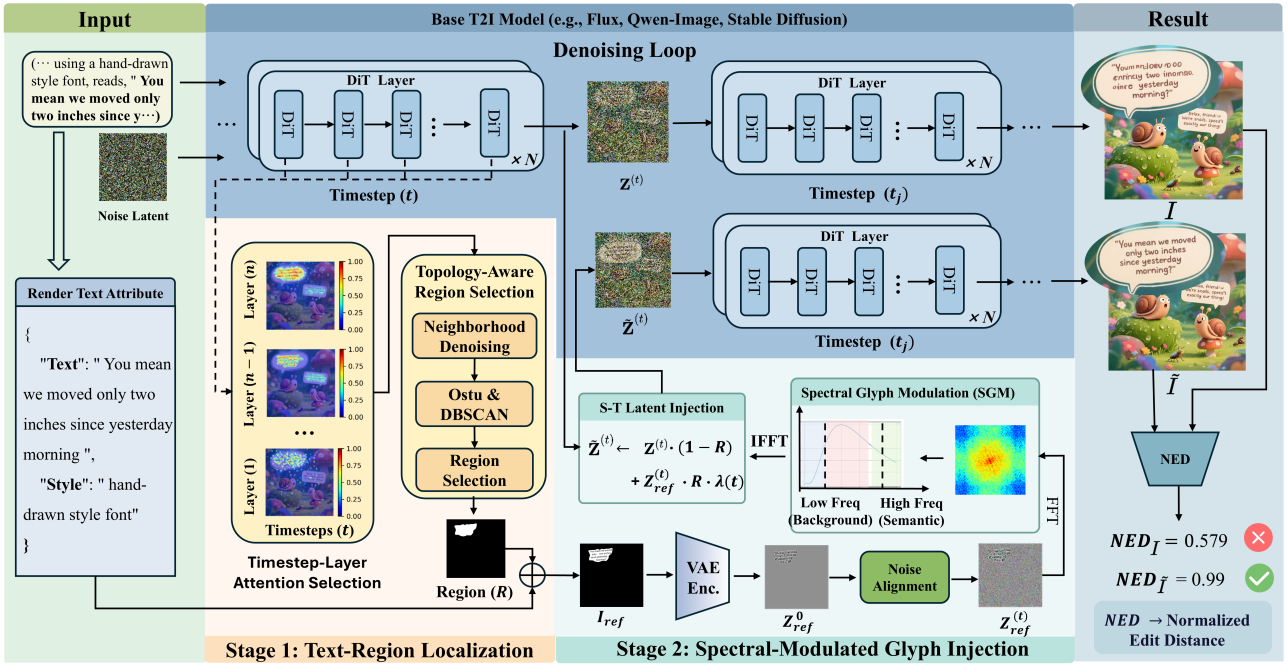

图 2. FreeText 概览。

化图：

$$\mathbf{M}^{(t,l)}(x,y) = \frac{1}{|\tilde{\mathcal{T}}_s|} \sum_{k \in \tilde{\mathcal{T}}_s} \mathbf{A}^{(t,l)}_{x,y,k}, \tag{2}$$

并将 $\mathbf{M}^{(t,l)}$ 线性归一化为 $[0,1]$。为清晰起见，下文省略下标 s。

3.1.2. 时间步-层选择

如图 3所示，直接汇总所有时间步和区块的注意力会引入大量噪声：早期步骤较为粗略，反映全局规划；中期步骤对于确定文字位置最具参考价值；后期步骤则因全局细化而变得分散（Chefer 等人，2023年；Darcet 等人，2023年）。此外，浅层区块侧重局部几何结构，而深层区块则整合全局语义。因此，我们在汇总之前先筛选出信息量较高的时间步—层组合。

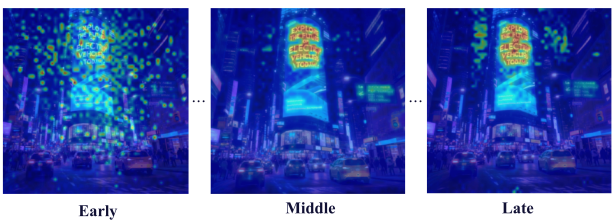

图 3. 跨时间步的典型图像到文本注意力模式：早期时间步较为粗略，中间时间步集中于目标区域，而后期时间步变得分散。

给定候选集合 $\mathcal{T}_{\text{cand}}$ 和 $\mathcal{L}_{\text{cand}}$，我们使用 软交并比 对每一对 $(t,\ l)$ 进行评分，比较 $\mathbf{M}^{(t,l)}$ 与参考掩码 $\mathbf{Y} \in [0,1]^{H \times W}$：

$$\text{IoU}(t,l) = \frac{\langle \mathbf{M}^{(t,l)}, \mathbf{Y} \rangle}{\|\mathbf{M}^{(t,l)}\|_1 + \|\mathbf{Y}\|_1 - \langle \mathbf{M}^{(t,l)}, \mathbf{Y} \rangle}. \tag{3}$$

我们选择排名前K 的配对来构成 $\mathcal{S}$ 并进行聚合：

$$\mathbf{M}(x,y) = \frac{1}{|\mathcal{S}|} \sum_{(t,l) \in \mathcal{S}} \mathbf{M}^{(t,l)}(x,y). \tag{4}$$

3.1.3. 拓扑感知区域选择

聚合后的映射$\mathbf{M}$ 仍可能包含孤立峰值和碎片化的簇。我们采用轻量级后处理流程，生成最终的书写掩码。

我们首先对$\mathbf{M}$ 进行局部邻域聚合，以抑制小型离群点并增强连通响应。接着，我们使用$\mathbf{M}$ 生成二值图$\mathbf{B} \in \{0,1\}^{H \times W}$，阈值采用通过最大化类间方差选取的自适应阈值（Otsu 等人，1975年）。随后，我们在前景像素上运行 DBSCAN（Ester 等人，1996），以获得候选连通区域 $\{\mathcal{C}_i\}$，同时丢弃稀疏噪声。

每个区域 $\mathcal{C}_i$ 均根据原始$\mathbf{M}$进行评分：

$$q_i = \frac{|\{(x,y) \in \mathcal{C}_i \mid \mathbf{M}(x,y) > \tau\}|}{|\mathcal{C}_i|}, \tag{5}$$

其中，τ被设置为候选区域并集内$\mathbf{M}$ 的高分位数。我们选择最佳区域，并将其调整为

潜变量分辨率，以获得二值书写掩码：

$$\mathbf{R} \in \{0,1\}^{H_{\text{lat}} \times W_{\text{lat}}}. \tag{6}$$

在第 3.2节中，$\mathbf{R}$ 会跨通道广播，以进行局部潜变量注入。

3.2. 频谱调制字形注入

为了解决“写什么”这一问题，我们增强字形结构，同时抑制语义泄漏。我们将字形参考图编码到潜空间，使其与当前噪声水平对齐，再应用基于 Log-Gabor 的 SGMI 来突出承载结构的频率，并在一个较短的时间窗口内将所得先验注入 $\mathbf{R}$。

3.2.1. 噪声对齐潜变量投影

我们将目标文本 s 栅格化为字形参考图像$\mathbf{I}_{\text{glyph}}$，并将其置于区域$\mathbf{R}$中，再使用与基础版模型相同的变分自编码器对其进行编码：

$$\mathbf{z}_{\text{ref}} = E_{\text{VAE}}(\mathbf{I}_{\text{glyph}}) \in \mathbb{R}^{C \times H_{\text{lat}} \times W_{\text{lat}}}. \tag{7}$$

在时间步 t，根据噪声调度（α_t，σ_t），我们通过前向扩散匹配噪声水平：

$$\mathbf{z}_{\text{ref}}^{(t)} = \alpha_t\, \mathbf{z}_{\text{ref}} + \sigma_t\, \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}). \tag{8}$$

3.2.2. LOG-GABOR 频谱调制

在 $\mathbf{z}^{(t)}_{\text{ref}}$上，我们应用 Log-Gabor 滤波器（Field, 1987年），以增强承载字形结构的中高频，同时抑制低频背景和超高频噪声。设 $G(\rho,\ \theta)$ 为二维频域中的 Log-Gabor 核。对于每个通道 c:

$$\widehat{\mathbf{z}}^{(t)}_{\text{ref},c} = \mathcal{F}\left(\mathbf{z}^{(t)}_{\text{ref},c}\right), \tag{9}$$

$$\widehat{\mathbf{z}}^{(t)}_{\text{sgmi},c}(\rho,\theta) = G(\rho,\theta) \cdot \widehat{\mathbf{z}}^{(t)}_{\text{ref},c}(\rho,\theta), \tag{10}$$

$$\mathbf{z}^{(t)}_{\text{sgmi},c} = \mathcal{F}^{-1}\left(\widehat{\mathbf{z}}^{(t)}_{\text{sgmi},c}\right), \tag{11}$$

其中，$\mathcal{F}$和 $\mathcal{F}^{-1}$ 分别表示二维快速傅里叶变换和逆快速傅里叶变换。所得$\mathbf{z}^{(t)}_{\text{sgmi}}$ 即为时间步t的 SGMI 增强的参考潜变量。

3.2.3. 退火时空注入

让采样轨迹从时间步 T演变至 0。我们仅在一个中前期窗口注入字形先验：

$$t_{\text{start}} = 0.8T, \quad t_{\text{end}} = 0.6T, \tag{12}$$

以避免干扰早期的全局规划或后期阶段的精细细节优化。对于 $t \in [t_{\text{start}}, t_{\text{end}}]$，我们定义一个余弦退火权重：

$$\lambda(t) = \frac{1}{2}\left(1 + \cos\left(\pi \cdot \frac{t - t_{\text{start}}}{t_{\text{end}} - t_{\text{start}}}\right)\right), \tag{13}$$

并通过掩码替换更新去噪潜变量 $\mathbf{z}^{(t)}$（Avrahami 等人, 2023年）：

$$\tilde{\mathbf{z}}^{(t)} = \left(\mathbf{I} - \lambda(t)\mathbf{R}\right) \odot \mathbf{z}^{(t)} \ + \ \lambda(t)\mathbf{R} \odot \mathbf{z}^{(t)}_{\text{sgmi}}. \tag{14}$$

对于 $t \notin [t_{\text{start}}, t_{\text{end}}]$，我们保持 $\mathbf{z}^{(t)}$ 不变。

3.3. CLT-Bench：中文长尾文本渲染

中文文本渲染颇具挑战，因为字符分布呈长尾特征，且类内视觉相似性很高。现有基准测试过度侧重常用字符和/或英语，因而掩盖了从常见/简单案例到生僻/复杂案例时的性能退化（Zhao 等人，2025年；Fang 等人，2025年；Du 等人，2025年）。我们推出 CLT-Bench，在生僻字和复杂布局设置下对文生图文本渲染进行压力测试。

我们为每个提示词分配一个复杂度分数，综合衡量字符难度和布局难度。对于字符 c，我们对笔画数 $s(c)$ 和频率排名 $r(c)$ 进行归一化：对于字符 c，我们对笔画数 $\kappa(c)$ 和频率排名 $r(c)$ 进行归一化：

$$K(c) = \frac{\kappa(c) - \kappa_{\min}}{\kappa_{\max} - \kappa_{\min}}, \quad R(c) = \frac{r(c) - r_{\min}}{r_{\max} - r_{\min}}. \tag{15}$$

并定义字符难度

$$D(c) = \frac{w_s K(c) + w_f R(c)}{w_s + w_f} \in [0,1]. \tag{16}$$

给定包含字符 $\{c_j\}_{j=1}^{N_{\text{chars}}}$ 的文本片段 $\{\text{txt}_i\}_{i=1}^{N_{\text{seg}}}$，我们计算

$$\begin{aligned} C_{\text{char}} &= \frac{1}{N_{\text{chars}}} \sum_j D(c_j), \\ C_{\text{len}} &= \min\left(\frac{N_{\text{chars}}}{N_{\max}}, 1\right), \\ C_{\text{seg}} &= \min\left(\frac{N_{\text{seg}} - 1}{M_{\max} - 1}, 1\right), \end{aligned} \tag{17}$$

其中，$N_{\max}$ 是提示中待渲染字符总数的预设上限，$M_{\max}$ 是待渲染文本片段（区域）数量的预设上限。提示分数随后为

$$\mathit{Score} = \frac{w_{\text{char}} C_{\text{char}} + w_{\text{len}} C_{\text{len}} + w_{\text{seg}} C_{\text{seg}}}{w_{\text{char}} + w_{\text{len}} + w_{\text{seg}}} \in [0,1]. \tag{18}$$

我们按分数对提示词进行分层，形成从常见/简单字符到罕见/复杂字符的子集，其中包含具有挑战性的多片段布局。

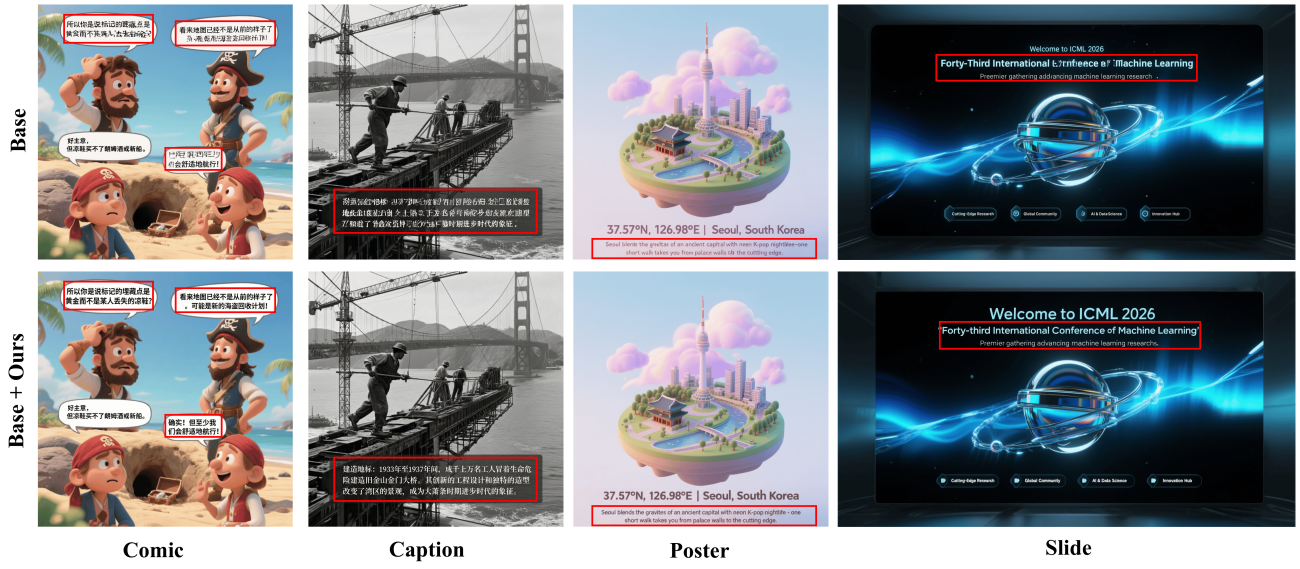

图 4. 四种文本渲染场景（漫画、说明文字、海报、幻灯片）中的基线对比（漫画、说明文字、海报、幻灯片）。上：基础版；下：基础版+FreeText。红框突出显示了目标文本区域，FreeText 可减少拼写错误和畸形字形，并提高可读性。

4. 实验

4.1. 实验设置

4.1.1. 基础模型

我们在四种具有代表性的文本到图像基础模型上评估 FreeText：(i) Qwen-Image（中文/英文提示词），(ii) FLUX.1-dev（仅英文），(iii) Stable Diffusion 3.5 Large（SD3.5-L；仅英文），以及 (iv) Stable Diffusion 3 Medium（SD3-M；仅英文）。所有实验均比较基础版与基础版 + *FreeText*。FreeText 作为推理时插件使用：它不会修改模型参数或架构，也不会引入可学习分支。

4.1.2. 基准测试与协议

我们使用三个基准测试，涵盖长文本、多区域渲染和中文长尾字符：(1)**longText-Benchmark** 包含 longText-en/zh，重点关注长提示词、段落级多行文本（Geng 等，2025年）；(2)**CVTG**，包含 2/3/4/5 个文本区域（2–5 个片段），通常使用短提示词（Du 等，2025年）；(3) **CLT-Bench**（第3.3节），旨在测试罕见且结构复杂的汉字。

语言对齐。Qwen-Image 和 FLUX.1-dev 在 longText-Benchmark 和 CVTG 上进行评估。SD3.5-L 和 SD3-M 仅在 CVTG 上进行评估，因为它们的文本编码器可能会截断长提示词。CLT-Bench 仅在 Qwen-Image 上进行评估（支持中文）。

推理设置。 除非另有说明，基础版和基础版 +FreeText 使用相同的分辨率、采样步数和采样器超参数。FreeText 使用默认的退火注入窗口（第 3.2节）；本节中我们将注入模块称为 **SGMI**。

4.1.3. 指标

我们同时衡量文本可读性和整体图像质量（除非另有说明，数值越高越好）：**归一化编辑距离**（Normalized Edit Distance，通过固定的光学字符识别引擎计算（Cui 等人，2025年）），**CLIP-Score**（文本-图像对齐），**AestheticScore**（LAION 美学预测器），以及**VQA 分数**（基于视觉语言模型的可用性/清晰度问答；模板见附录）。对于定位分析，我们报告**交并比**，用于衡量预测文本区域与参考文本区域之间的差异。

4.2. FreeText的有效性

4.2.1. QWEN-IMAGE 和 FLUX.1-DEV

表1报告了在 longText-Benchmark 和 CVTG 上的结果。FreeText 一贯提升归一化编辑距离和 VQA 分数（Fang 等人，2025年），表明文本可读性更高；与此同时，CLIP-Score 和 AestheticScore 基本保持稳定，说明其对语义对齐和美学的影响有限。

4.2.2. SD3-M 和 SD3.5-L

由于 SD3 变体对长提示词较为敏感，我们仅在 CVTG 上对其进行评估（表 2）。FreeText 提升了两个模型的归一化编辑距离和视觉问答分数，而 CLIPScore 和 AestheticScore 则基本持平，这表明局部 SGMI 注入不会造成明显的语义漂移或质量下降。

表 *1.* longText-Benchmark 和 CVTG 上的端到端结果。

模型	设置	子集	NED↑	CLIP↑	Aes↑	VQA↑
Qwen-Image	Base	长文本-英语	0.625	0.858	4.912	2.650
Qwen-Image	基础版 + FreeText	长文本-英语	**0.713**	**0.864**	**5.013**	**4.177**
FLUX.1-dev	Base	长文本-英语	0.598	0.863	**5.365**	2.563
FLUX.1-dev	基础版 + FreeText	长文本-英语	**0.690**	**0.868**	5.342	**4.211**
Qwen-Image	Base	长文本-中文	0.639	0.474	4.607	3.657
Qwen-Image	基础版 + FreeText	长文本-中文	**0.694**	**0.537**	**4.749**	**4.211**
Qwen-Image	Base	CVTG	0.574	0.781	4.386	2.756
Qwen-Image	基础版 + FreeText	CVTG	**0.619**	**0.794**	**4.391**	**3.469**
FLUX.1-dev	Base	CVTG	0.712	0.836	5.910	4.050
FLUX.1-dev	基础版 + FreeText	CVTG	**0.722**	**0.839**	**5.936**	**4.952**

表 *2*。 SD3 模型在 CVTG 上的端到端结果。每组模型对中表现最佳的结果以粗体显示。

模型	设置	归一化编辑距离↑	CLIP↑	美学评分↑	视觉问答↑
SD3.5-L	Base	0.848	**0.879**	**5.634**	3.849
SD3.5-L	基础版 + FreeText	**0.864**	0.871	5.608	**4.595**
SD3-M	Base	0.616	0.851	5.906	2.903
SD3-M	基础版 +FreeText	**0.669**	**0.852**	**5.917**	**3.674**

4.2.3. CLT-BENCH

在 CLT-Bench 上（仅评估 Qwen-Image），FreeText 提升了归一化编辑距离，但增幅较小（表3）。这表明，当基础模型已具备可用于目标字符的表征时，SGMI 的效果最为显著；它能强化字形结构，而非从零开始生成模型未见过的字符。

4.2.4 。完整注意力下的收益传播　　TION

我们观察到跨区域的效益传播：使用 FreeText 修正一个文本区域，可以改善其他未经显式处理的区域，这体现为更高的全局指标（例如 VQA分数）。我们将此归因于 DiT/MMDiT 中的全局自注意力：在每个去噪步骤中，图块令牌会全局混合信息，因此某一区域中的严重错误可能会扰乱其他位置的更新；一旦修正了关键错误，这种干扰便会减弱。

表*3.* 端到端归一化编辑距离（NED），基于 CLT-Bench。

模型	设置	NED↑
Qwen-Image	Base	0.458
Qwen-Image	基础版 +FreeText	**0.488**



(a) Base　(b) Inject prior (Text-2 only)

(c) Base + ours　(d) Attn while generating Text-1

图*5*。 跨区域收益传播与注意力证据（包含两行文本的示例；仅细化其中一行即可改善另一行）。

4.3. 定位策略

4.3.1. 令牌选择

我们针对每个目标跨度比较三种词元集合：**仅实体**（目标字符串的词元）、**仅汇聚令牌**（类似汇聚令牌的特殊词元），以及 **实体 + 汇聚令牌**。如表4和图 6所示，仅汇聚令牌在时间上更加稳定，但上限较低；而实体 + 汇聚令牌通过结合明确的语义归因与稳定的汇聚响应，实现了最佳交并比，从而为 SGMI 生成更可靠的掩码。

4.3.2. 与基于视觉语言模型的定位进行比较

表 5比较了我们的内生定位与多个闭源视觉语言模型基线的表现。在实际应用中，多行文本、杂乱背景和畸形字形可能会

表4. 不同词元集合的定位交并比。

设置	IoU↑
仅实体	0.495
仅汇聚令牌	0.479
实体 + 汇	**0.561**

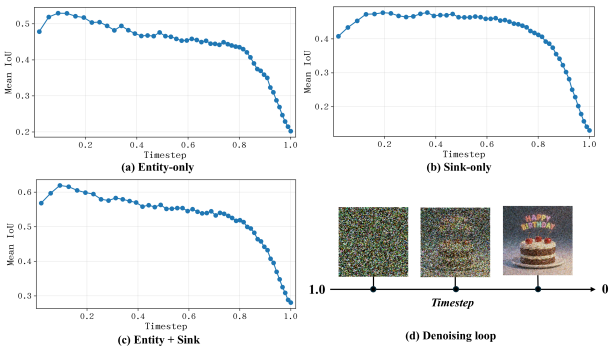

图6. 不同词元集合下交并比随时间步的变化。

破坏“先识别再定位”的流程；识别失败往往会进一步导致定位失败。FreeText 通过直接读取图像到文本注意力，避免了这一连锁问题，并提供更稳定的信号。

表5. 定位交并比比较。

方法	IoU↑
豆包-seed-1-6-251015	0.325
Gemini 2.5 Flash Lite	0.139
GPT-5.1	0.159
通义千问3-VL-Plus-2025-09-23	0.195
FreeText（我们的方法）	**0.561**

4.4. 消融研究

我们比较了三个变体：**B**（基础版），**+F**（基础版+FreeText），以及 **+F–SGMI**（基础版+FreeText，不含 SGMI，即移除频谱带通调制，同时保持其余部分不变）。如表 6所示，移除 SGMI 会降低归一化编辑距离和 VQA分数，而 CLIP 和美学评分基本保持不变，这表明 SGMI 主要有助于提升文本可读性。如图 8进一步所示，仅注入低频分量会丢失笔画级结构，而仅注入高频分量（此时语义占主导）则可能引发概念纹理侵入。相比之下，SGMI 的带通设计提供了最有利于字形结构、同时最能抑制语义泄漏的注入信号。这表明，频域调制的关键不在于注入更多信息，而在于注入恰当的频谱带。



图7 多行文本和退化字形情况下典型的视觉语言模型定位失败，与内生定位的对比。



图 8. 定性消融实验展示了不同频谱设置下的语义泄漏和笔画退化。

4.5. 效率

我们在 NVIDIA A6000 上使用 bfloat16、分辨率 928×928和 50 个采样步数测量推理开销。表7显示，FreeText 会带来适度的开销（主要来自第一阶段定位，该阶段会在注入前累积并选择图像到文本注意力），使端到端延迟增加约 12%–18%，峰值内存开销增加 < 1 GB。

5. 结论

我们提出了 FreeText，这是一种无需训练且与基础模型无关的框架，无需改变模型权重或架构即可提升 T2I扩散模型中的文本渲染效果。通过将文本渲染分解为在哪里书写和书写什么，FreeText (i) 通过基于汇点锚定、拓扑感知的选择，从内生注意力中定位书写区域；(ii) 通过 SGMI 提升字形保真度。SGMI 是一种噪声对齐的频域注入方法，可增强结构承载成分并缓解语义泄漏—

表*6*。 关于 SGMI 的消融实验。

模型	设置	归一化编辑距离↑	CLIP↑	Aes↑	VQA↑
Qwen-Image	B	0.625	0.858	4.912	2.650
	+F−SGMI	0.686	0.860	**5.027**	3.724
	+F	**0.713**	**0.864**	5.013	**4.177**
FLUX.1-dev	B	0.598	0.863	**5.365**	2.563
	+F− SGMI	0.671	0.865	5.361	3.816
	+F	**0.690**	**0.868**	5.342	**4.211**

表*7*。 推理效率。

模型	设置	时间（秒）↓	内存（GB）↓
Qwen-Image	Base	**37.64**	**53.76**
Qwen-Image	基础版 +FreeText	42.33	54.35
FLUX.1-dev	Base	**41.56**	**31.44**
FLUX.1-dev	基础版 +FreeText	47.17	32.17
SD3.5-L	Base	**35.03**	**26.11**
SD3.5-L	基础版 +FreeText	41.17	26.91
SD3-M	Base	**9.85**	**14.53**
SD3-M	基础版 +FreeText	11.47	14.97

漏。跨多个基础模型和基准测试，FreeText 持续改善可读性指标，同时保持 CLIPScore 和 AestheticScore，并且仅带来适度的运行时间和内存开销。未来的研究将重点通过将我们的方法适配到多种新兴基础模型，验证其普适性。

参考文献

Avrahami, O., Fried, O. 和 Lischinski, D. 混合潜空间扩散。*ACM* 图形学汇刊 *(TOG)*，42(4): 1–11，2023年。

Chauhan, S., Chaudhary, M., Choy, K., Nellessen, S. 和 Schoots, N. 语言模型中的标点符号与谓词。*arXiv* 预印本 *arXiv:2508.14067*，2025年。

Chefer, H.、Alaluf, Y.、Vinker, Y.、Wolf, L. 和 Cohen-Or, D. Attend-and-Excite：针对文生图扩散模型的基于注意力的语义引导。*ACM* 图形学汇刊 *(TOG)*，42(4)：1–10，2023年。

Chen, J.、Li, B. 和 Xue, X. 基于笔画级分解的零样本文字识别。*arXiv* 预印本 *arXiv:2106.11613*，2021年。

Chen, J.、Huang, Y.、Lv, T.、Cui, L.、Chen, Q. 和 Wei, F. TextDiffuser：作为文本绘制器的扩散模型。神经信息处理系统进展，36：9353–9387，2023年。

崔，C.，孙，T.，林，M.，高，T.，张，Y.，刘，J.，王，X., Zhang, Z., Zhou, C., Liu, H., 等。PaddleOCR 3.0 技术报告。*arXiv* 预印本 *arXiv:2507.05595*，2025年。Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. 视觉 Transformer 需要寄存器。*arXiv* 预印本 *arXiv:2309.16588*，2023年。Du, N., Chen, Z., Gao, S., Chen, Z., Chen, X., Jiang, Z., Yang, J., and Tai, Y. TextCrafter：在复杂视觉场景中准确渲染多段文本。*arXiv* 预印本 *arXiv:2503.23461*，2025年。

Esser, P.、Kulal, S.、Blattmann, A.、Entezari, R.、M¨uller, J.、Saini, H.、Levi, Y.、Lorenz, D.、Sauer, A.、Boesel, F. 等。用于高分辨率图像合成的整流流 Transformer 扩展。在 第四十一届国际机器学习大会，2024年。

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 等. 一种用于在含噪声的大型空间数据库中发现簇的基于密度的算法. 收录于*KDD*，第 96 卷，第 226–231 页，1996年。

Fang, R., Yu, A., Duan, C., Huang, L., Bai, S., Cai, Y., Wang, K., Liu, S., Liu, X., 和 Li, H. Flux-Reason-6M 与 Prism-Bench：百万级文生图推理数据集及综合基准测试。*arXiv* 预印本 *arXiv:2509.09680*，2025年。

Field, D. J. 自然图像的统计特性与皮层细胞响应特性之间的关系。美国光学学会会刊 *A*，4(12)：2379–2394，1987年。

Geng, Z., Wang, Y., Ma, Y., Li, C., Rao, Y., Gu, S., Zhong, Z., Lu, Q., Hu, H., Zhang, X., 等. X-omni：强化学习让离散自回归图像生成模型重焕生机。*arXiv* 预印本 *arXiv:2507.22058*，2025年。

Kang, S.、Kim, J.、Kim, J. 和 Hwang, S. J. 看见你所获告知的内容：大型多模态模型中的视觉注意力汇。*arXiv* 预印本 *arXiv:2503.03321*，2025年。

Labs, B. F.、Batifol, S.、Blattmann, A.、Boesel, F.、Consul, S.、Diagne, C.、Dockhorn, T.、English, J.、English, Z.、Esser, P. 等。Flux. 1 kontext：用于潜空间中上下文内图像生成与编辑的流匹配。*arXiv* 预印本 *arXiv:2506.15742*，2025年。

Ma, J.、Zhao, M.、Chen, C.、Wang, R.、Niu, D.、Lu, H. 和 Lin, X. Glyphdraw：在文生图中无缝呈现具有复杂空间结构的文本。*arXiv* 预印本 *arXiv:2303.17870*，2023年。

Otsu, N. 等人：《一种基于灰度直方图的阈值选择方法》。*Automatica*，11(285-296)：23–27，1975年。

Peebles, W. 和 Xie, S.：《基于变换器的可扩展扩散模型》。载于*IEEE/CVF* 国际计算机视觉会议论文集，第 4195–4205 页，2023年。

Razzhigaev, A., Mikhalchuk, M., Rahmatullaev, T., Gon- charova, E., Druzhinina, P., Oseledets, I., and Kuznetsov, A. LLM-Microscope：揭示标点符号在变换器上下文记忆中的隐藏作用。*arXiv* 预印本 *arXiv:2502.15007*，2025年。Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al. Seedream 4.0：迈向下一代多模态图像生成。*arXiv* 预印本 *arXiv:2509.20427*，2025年。Tang, R., Liu, L., Pandey, A., Jiang, Z., Yang, G., Kumar, K., Stenetorp, P., Lin, J., and T¨ure, F. DAAM：使用交叉注意力解读 Stable Diffusion。载于第*61*届计算语言学协会年会（第*1*卷：长篇论文），第5644–5659页，2023年。Tigges, C., Hollinsworth, O. J., Geiger, A., and Nanda, N. 大语言模型中的情感线性表征，2023年。网址 *https://arxiv. org/abs/2310.15154*，2023年。Tuo, Y., Xiang, W., He, J.-Y., Geng, Y., and Xie, X. AnyText：多语言视觉文本生成与编辑。*arXiv* 预印本 *arXiv:2311.03054*，2023年。Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.-m., Bai, S., Xu, X., Chen, Y., et al. Qwen-Image 技术报告。*arXiv* 预印本 *arXiv:2508.02324*，2025年。Yang, Y., Gui, D., Yuan, Y., Liang, W., Ding, H., Hu, H., and Chen, K. GlyphControl：用于视觉文本生成的字形条件控制。神经信息处理系统进展，36：44050–44066，2023年。Zhang, S., Khan, M., and Papyan, V. 注意力汇：一种针对嵌入的“捕获、标记、释放”机制。载于第*39*届神经信息处理系统年会，2025年。Zhao, S., Wu, Q., Li, X., Zhang, B., Li, M., Qin, Q., Liu, D., Zhang, K., Li, H., Qiao, Y., et al. Lex-Art：通过可扩展的高质量数据合成重新思考文本生成。*arXiv* 预印本 *arXiv:2503.21749*，2025年。