TeEFusion: 融合文本嵌入以提炼无分类器引导

Minghao Fu*[1,2,3] Guo-Hua Wang†[3] Xiaohao Chen[3] Qing-Guo Chen[3]
Zhao Xu[3] Weihua Luo[3] Kaifu Zhang[3]
[1] School of Artificial Intelligence, Nanjing University
[2] National Key Laboratory for Novel Software Technology, Nanjing University
[3] Alibaba International Digital Commerce Group
fumh@lamda.nju.edu.cn, { wangguohua, xiaohao.cxh, qingguo.cqg, changgong.xz, weihua.luowh, kaifu.zkf } @alibaba-inc.com

Abstract

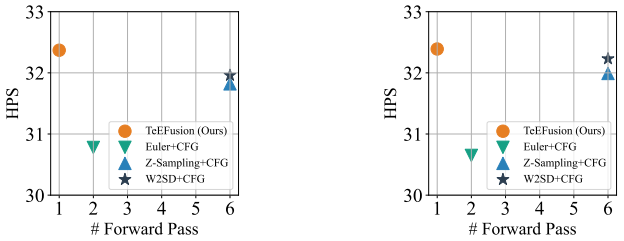
近年来，文本到图像合成的进步主要得益于复杂的采样策略和无分类器指导 (CFG)，以确保高质量生成。然而，CFG 依赖于两次前向传递，特别是当与复杂的采样算法结合时，导致极高的推理成本。为了解决这个问题，我们引入了 TeEFusion（文本嵌入融合），一种新颖且高效的蒸馏方法，它直接将指导幅度并入文本嵌入，并蒸馏教师模型的复杂采样策略。通过简单地使用线性操作融合有条件和无条件文本嵌入，TeEFusion 无需添加额外参数即可重建所需的指导，同时使学生模型能够从教师通过复杂采样方法产生的输出中学习。对诸如 SD3 等最先进模型的广泛实验表明，我们的方法使学生能够以更简单、更高效的采样策略逼真地模仿教师的表现。因此，学生模型的推理速度比教师模型提高最多 6 ×，同时保持图像质量与教师的复杂采样方法所获得的水平相当。代码在 github.com/AIDC-AI/TeEFusion 公开可用。

1. 引言

基于流匹配 [14, 16] 管道的模型，例如 SD3 [5] 和 FLUX [11]，已经成为通过文本提示直接合成结构化图像的领先生成框架。在文本到图像生成领域，它们的多功能性由广泛的应用范围证明，从图像编辑 [4] 到数字艺术创意概念生成 [21]。

无分类器引导 (CFG) [7] 是一种关键技术，通过利用条件和无条件预测来引导生成过程朝向具有更高文本条件密度的区域，从而确保高质量的图像合成。然而，使用 CFG 进行采样需要两个前向传递，一个基于提供的文本提示进行条件化，另一个基于空（或背景）提示进行条件化，这引入了大量的计算开销并减慢了推理速度。

为了解决这个问题，指导蒸馏 [19] 通常在最先进的模型中使用，例如像 FLUX.1-dev [11] 这样的变体。这种方法将指导作为输入参数进行整合，减少了所需前向传递的次数至一次。因此，指导蒸馏变得在最先进模型中既流行又重要，因为它允许模型扩展而不必担心成本，或可以采用更复杂的推理采样算法。如图 1 所示，诸如 Z-Sampling [1] +CFG 和 W2SD [2] +CFG 这样的高级采样策略即使能产生更高的图像质量（如 HPS [34] 所证实），也需要几乎传统欧拉 [16] +CFG 方法三倍的推理负担，从而突出显示了效率优化的巨大潜力。



(a) SD3 [5] (b) 内部 T2I

Figure 1. 对两种 DiT [22] 模型的不同采样策略进行比较。“# Forward Pass” 指定一次去噪步骤中所需的前向传递次数。

最近的一些研究 [13, 20, 24, 36–38, 41] 试图推进图像合成的蒸馏 [19]。然而，这些方法仅限于教师和学生模型使用相同采样算法的场景。换句话说，他们没有专注于从特别复杂的采样算法中蒸馏，同时使学生模型仅依赖于简单的采样策略。此外，这些方法过于复杂，给实施和适应带来了重大障碍。这一问题因现代最先进模型的大规模特性而进一步放大，在这种情况下，调整超参数变得越来越具有挑战性，最终阻碍了这些复杂方法的部署。

因此，这些方法目前缺乏稳健和大规模的验证，尤其在两个关键方面：1）是否包含通用基准，如 DPG-Bench [8] 或审美评分指标如 HPS [34]；以及 2）在与最先进模型如 SD3 [5] 和 FLUX [11] 中使用的大型 DiT [22] 相匹配的规模架构上进行测试。因此，尚不

*Work done during the internship at Alibaba International Digital Commerce Group.
†G. Wang is the corresponding author.

Figure 2. 不同提示策略的定性结果。最左边的三列显示由不同提示输入生成的图像，最后一列展示了通过加法融合保留和掩盖组件的语义信息的结果。

清楚这些蒸馏技术是否能有效应用于最新的、最先进的文本到图像生成模型。

受到这些挑战的启发，在本文中我们探讨了更有效和高效的 CFG 蒸馏方法。我们发现嵌入空间中对不同文本特征的线性操作可以有效地融合或抑制原文本的某些方面（参见图 2）。基于这一观察，我们提出了 TeEFusion（文本嵌入融合），这是一种新颖但非常简单的蒸馏方法，它直接利用 CFG 的基本采样公式并将引导强度整合到文本嵌入中。具体来说，我们的方法将条件模型和非条件模型输出的线性组合向前推进，直接应用于组合相应的文本嵌入。学生模型被优化以学习由复杂采样策略生成的教师模型的去噪输出，从而同时蒸馏指导信号和采样过程。

TeEFusion 提供了两个主要优势。首先，与现有的蒸馏技术相比，我们的方法非常简单，具有清晰且易于实现的算法原则。与使用前馈网络来架构指导的 [20] 不同，我们的方法不引入额外的架构——蒸馏后的模型结构与未蒸馏模型保持相同。此外，TeEFusion 不引入任何额外的超参数，确保蒸馏过程保持高效且易于集成到现有流程中。

其次，我们对 TeEFusion 在最先进的文本到图像模型上进行了全面评估，包括 SD3 [5] 和我们的内部开发模型 T2I，该模型显示出与 SD3 相当的性能。实验证明，TeEFusion 在美学评分 [34] 、对象组成 [8] 和提示追踪能力 [23] 方面相较于基线方法取得了显著的改进。此外，我们的方法能够无缝兼容教师模型使用的多样化采样器。我们希望我们的评估能够为评估最先进的大规模文本到图像模型中的蒸馏技术提供有价值的见解。

本文的贡献可以概括为：

- 我们通过实验证明，对于当前最先进的文本到图像生成模型，从特定的提示中增加或减去文本嵌入可以有效地在生成的图像中合并或缓解某些视觉概念。
- 我们提出了 TeEFusion，这是一种新颖且简单的蒸馏策略，将 CFG 的指导直接融入文本嵌入，不引入任何额外参数。我们的方法可以与教师模型使用的各种复杂采样策略无缝兼容。
- 我们在工业级大型文本到图像模型上评估 TeEFusion，包括诸如 SD3 [5] 这样的最新框架和我们内部的 T2I 模型，使用标准基准如 DPG-Bench [8] 、美学指标（HPS [34] ）以及提示跟随评估 [23] 。实证结果表明，TeEFusion 在美学质量、物体构图和提示遵循方面显著优于基准方法。

2. 相关工作

2.1. 测试时缩放

近年来，研究越来越关注扩散模型中推理的缩放行为 [7, 18] 。RePaint [18] 通过重新引入高斯噪声来恢复先前的噪声水平，迭代地改进潜在变量，建立了一种基于反射的采样范式，并在后续工作中被广泛采用 [1–3, 33] 。例如，Z-Sampling [1] 通过交替使用不同引导强度的前向和反向传递，以之字形方式改变去噪输出，从而将随机输入引导到语义有意义的区域。W2SD [2] 进一步扩展了这一方法，通过显式结合强模型和弱模型。

2.2. 扩散蒸馏

早期的研究通过利用先进的数值求解器来提高扩散模型的采样效率 [15, 28, 40] 。尽管有了这些改进，仍有

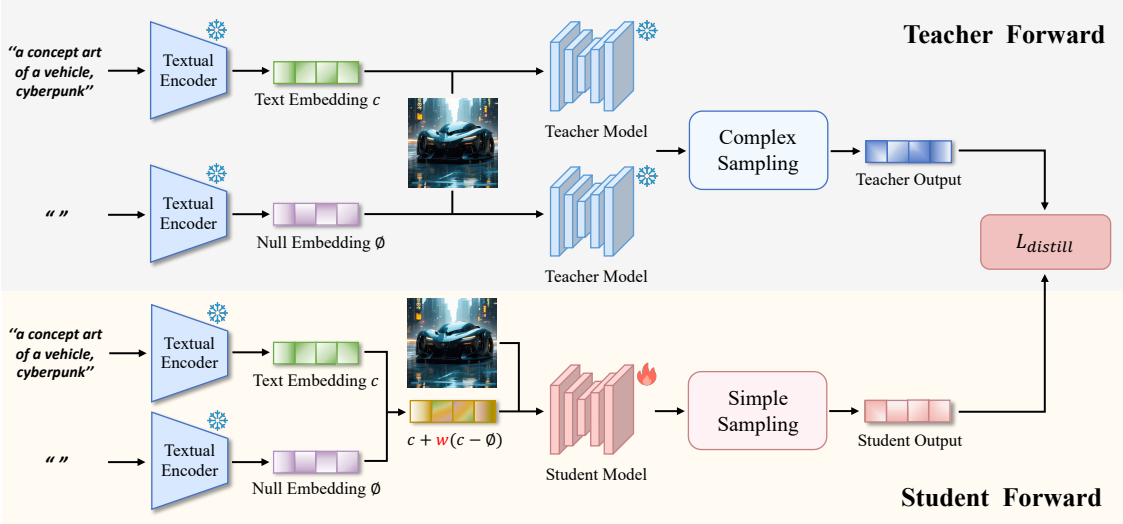


Figure 3. TeEFusion 的说明。我们的方法将 DiT [22] 模型输出的融合直接转移到文本嵌入空间中，如公式 7 所指定。此设计与教师模型采用的各种采样方法兼容，为测试时扩展提供了显著潜力。

很大的提升空间。扩散蒸馏 [19] 通过训练一个更简单的学生模型来复制更复杂的教师模型的行为，主要是通过步骤蒸馏 [9, 24–26, 29, 39] ，从而减少所需的推理步骤。某些方法 [25, 26] 通过引入一个额外的判别器来促进评分函数的泛化，从而将对抗训练结合进来。然而，这些方法固有的复杂性限制了它们的实际应用，尤其是在当代扩散模型的规模和复杂性面前。

在这些方法中，DistillCFG [20] 是最相关的，它使用额外的 MLP 来编码指导比例。DICE [42] 也添加了一个增强注意力的 MLP，但仅蒸馏一个固定比例，这在推理时限制了灵活性。更多变体，如 MG [30] 尚未在工业规模的数据集和模型上进行评价。此外，现有的蒸馏框架，如渐进蒸馏 [24] ，尚未明确解决涉及采用复杂测试时间采样策略的教师模型的场景，这种情况在最先进的扩散模型中日益流行。

为弥补这一关键差距，我们提出的框架 TeEFusion 通过将指导强度明确纳入条件和无条件文本嵌入的线性组合，推进了指导蒸馏。此外，TeEFusion 能够有效地泛化至教师模型采用的多种复杂采样策略，解决了扩散模型蒸馏中一个重要但很大程度上被忽视的方面。

3. 预备知识

3.1. 无分类器引导

无分类指导（CFG） [7] 通过调整采样分布，提高生成图像与文本描述之间的一致性：

$$\tilde{p}_\theta(x_t|c) \propto p_\theta(x_t|c)^{1+w} p_\theta(x_t)^{-w}, \tag{1}$$

其中 c 表示文本提示，$p_\theta(x_t|c)$ 是条件分布，$p_\theta(x_t)$ 表示无条件对应的分布。CFG 通过估计扩散评分 [1] 来优化生成过程：

$$\tilde{\epsilon}_\theta(x_t, c) = (1 + w)\epsilon_\theta(x_t, c) - w\epsilon_\theta(x_t), \tag{2}$$

标量 w 决定了引导的强度：较大的值更强烈地将合成引向文本提示，从而促进输出忠实反映输入文本，同时减少对未受约束预测的依赖。请注意，需要两次前向传播才能从 $\epsilon_\theta(x_t, c)$ 和 $\epsilon_\theta(x_t)$ 得到预测。

3.2. 带反射的采样。

一些方法 [1, 2] 采用超出标准的含反射的欧拉方法的采样策略，以更加丰富的语义信息来优化去噪过程的起始点。

弱到强扩散（W2SD）[2] 利用弱模型 $\mathcal{M}^w$ 和强模型 $\mathcal{M}^s$ 之间的差异来弥合生成分布与理想状态之间的差距。在 W2SD 中，反射操作更新了潜在变量 x_t ，具体如下：

$$\tilde{x}_t = \mathcal{M}^w_{\text{inv}}(\mathcal{M}^s(x_t, t), t) \tag{3}$$

$$x_{t-1} = \mathcal{M}^s(\tilde{x}_t, t), \tag{4}$$

从而引导采样轨迹朝向文本密度更高的区域。Z-采样 [1] 是 W2SD 的一个更简单的版本，通过在 $\mathcal{M}^s$ 和 $\mathcal{M}^k$ 间共享权重，但使用不同的引导信号。

然而，弱模型处理强模型的去噪输出的额外反演步骤在每次采样迭代中引入了两个额外的前向传递。因此，反射采样的时间成本是标准采样的 3 × （一个用于标准去噪，一个用于反演，另一个用于后续去噪）。当结合 CFG 时，总体时间成本上升到几乎 6 × 。

4. CFG 的深入分析

我们首先研究 CFG 的基本机制。如方程 2 所示，CFG 的输出可以表示为条件预测 $\epsilon_\theta(x_t|c)$ （系数为 $1 + w$ ）

[1] 为了符号简化，我们将扩散模型和流匹配模型的分数预测表示为 ϵ_θ 。

Algorithm 1 TeEFusion 的蒸馏流程。

Require: Teacher model $\epsilon_{\theta_\mathrm{T}}$, Dataset D
  $\epsilon_{\theta_\mathrm{S}} \leftarrow \epsilon_{\theta_\mathrm{T}}$ ▷ Initialize student using teacher weights
  while not converged do
    $x_0, c \sim \mathcal{D}$ ▷ Sample data
    $t \sim U[0,1]$ ▷ Sample time
    $w \sim U[w_{\min}, w_{\max}]$ ▷ Sample guidance
    $\epsilon \sim N(0,I)$ ▷ Sample noise
    $x_t = (1-t)x_0 + t\epsilon$ ▷ Add noise to data
    $\tilde{\epsilon}_{\theta_\mathrm{T}}(x_t, w, c) = (1+w)\epsilon_{\theta_\mathrm{T}}(x_t, t, c) - w\epsilon_{\theta_\mathrm{T}}(x_t, t, \varnothing)$
  ▷ Compute target using CFG
    if Reflection then
      $\tilde{\epsilon}_{\theta_\mathrm{T}}(x_t, w, c) \leftarrow$ Refine $\tilde{\epsilon}_{\theta_\mathrm{T}}(x_t, w, c)$ using Eq. 3 and 4
    end if
    Compute $\hat{z}_{t,c,\varnothing,w}$ using Eq. 7
    $L_{\text{distll}} = \|\epsilon_{\theta_\mathrm{S}}(x_t, \hat{z}_{t,c,\varnothing,w}) - \tilde{\epsilon}_{\theta_\mathrm{T}}(x_t, w, c)\|_2^2$ ▷ Loss
    $\epsilon_{\theta_\mathrm{S}} \leftarrow \epsilon_{\theta_\mathrm{S}} - \gamma \cdot \nabla_{\epsilon_{\theta_\mathrm{S}}} L_{\text{distill}}$ ▷ Optimization
  end while
  return $\epsilon_{\theta_\mathrm{S}}$ ▷ Output the trained student model

和无条件预测 $\epsilon_\theta(x_t)$（系数为 $-w$ ）的线性组合。这种公式激发了一个想法，即在处理管道中更早地进行这种线性组合，从而将两个独立的前向传播减少到一个。一个自然的方法是将这种组合直接集成到文本嵌入中，以有效地将指导强度纳入模型，如下所示：

$$\hat{\epsilon}_\theta(x_t, c) = \epsilon_\theta(x_t, \hat{c}), \tag{5}$$

，其中 $\hat{c} = (1+w)c - w\varnothing = c + w(c-\varnothing)$ 被标记为融合文本嵌入，与来自空提示的 $\varnothing$ 一起提供生成的背景。然而，这种方法的成功依赖于将输入条件线性组合是有意义的这一假设。

因此，我们进行了一项初步测试，以验证将条件和无条件文本嵌入合并为单个文本嵌入是否可以产生必要的语义表示。在该实验中，我们随机遮蔽了提示的某些组件，并用一个 [mask] 令牌替换它们。然后，我们在三种不同的条件下生成图像：

1. 仅使用提示中未被遮挡的部分。
2. 仅使用被掩盖的组件（即，被 [mask] 替换的内容）。
3. 使用原始的、未修改的提示。
4. 未遮掩和遮掩的组件都被转换为文本嵌入，然后通过元素级加法进行合并。合并后的嵌入随后用于图像生成。

这种设置使我们能够评估从保留和遮蔽的提示组件中融合语义信息的加法（或减法）是否产生了有意义的特征，这些特征可以通过线性组合（参见方程 5 ）在文本嵌入空间中指导高质量的图像合成。如图 2 所示，通过添加保留和遮蔽部分的文本嵌入生成的图像，其质量可以匹配（甚至超越）从原始、未修改的提示中生成的图像。这些发现表明，文本嵌入空间中的加法操作可以有效地合并不同的提示模式，从而突显了我们采样技术的潜力，即使用方程 5 作为替代来整合指导幅度，同时在采样期间保持高效率（定量结果见第 ?? 节）。

5. 利用 TeEFusion 蒸馏 CFG

第二节的分析表明，对不同提示的文本嵌入进行简单的加法或减法可以有效地重构或消除原始提示的特定语义成分，这表明嵌入空间中的这种操作可以产生具有明确语义的特征。基于这一观察，我们假设这些文本嵌入的任何线性组合——使用适当的中等系数——同样可以产生语义上有意义的表示。这一见解激发了使用适当配置的 w 进行指导蒸馏的方程 5 的应用。

受到这一假设的启发，我们提出了 TeEFusion（文本嵌入融合），这是一种简单的蒸馏框架，可以有效地同时蒸馏无分类器指导和复杂的基于反射的采样技术。在 TeEFusion 中，通过线性组合条件和无条件文本嵌入，将指导强度 w 集成到学生模型中，w 作为组合系数。随后，训练学生模型以模拟使用基于反射的采样的教师模型的去噪输出（见图 3 ）。

然而，直接将方程 5 应用于 TeEFusion 会带来挑战。随着引导尺度 w 的增加，融合提示 $\hat{c}$ 的方差以 $\mathcal{O}(w^2)$ 的顺序增长，当 w 变得过大时，会导致数值稳定性差。为了解决这个问题，我们使用正弦和余弦时间嵌入 [32] 将 w 投影到一个向量空间中。该方法不仅确保不同的 w 值可以区分，还缓解了由于方差过大而引起的数值不稳定性。

具体而言，文本到图像模型 [5, 11] $\epsilon(x_t, t, c)$ 中文本提示和时间步的联合嵌入被表述为：

$$z_{t,c} = \mathcal{G}(\psi(t)) + \mathcal{F}(c), \tag{6}$$

其中 $\mathcal{F}(\cdot)$ 和 $\mathcal{G}(\cdot)$ 表示两个不同的多层感知机（MLPs），而 $\psi(\cdot)$ 表示时间步的正弦编码。该公式有效地解耦了文本和时间信息的处理，从而实现了指导强度的更有结构的整合。

我们的 TeEFusion 通过将指导幅度 w 融合到 $z_{t,c}$ 中来提炼它，如下所示：

$$\hat{z}_{t,c,\varnothing,w} = \mathcal{G}(\psi(t)) + \mathcal{F}(c) + \underbrace{\mathcal{G}(\psi(w))\,\mathcal{F}(c-\varnothing)}_{\text{extra term}}, \tag{7}$$

，其中术语 $c-\varnothing$ 在投影到文本嵌入空间后计算。假设教师和学生模型分别表示为 $\epsilon_{\theta_\mathrm{T}}$ 和 $\epsilon_{\theta_\mathrm{S}}$ 。获得 $\hat{z}_{t,c,\varnothing,w}$ 后，用于计算学生模型的引导输出。蒸馏目标随后让学生模拟教师模型通过复杂采样得出的 CFG 衍生输出（参见算法 1 ）。对比方程 7 与 $\hat{c}$ 的定义，其中术语 $\mathcal{F}(c) + \mathcal{G}(\psi(w))\mathcal{F}(c-\varnothing)$ 正好对应于 $\hat{c}$ 中的术语 $c + w(c-\varnothing)$ ，但在嵌入空间中计算。

TeEFusion 提供了几个显著的优点。首先，它的 w -注入方法非常简单。与其他需要额外网络结构和参数的方法 [20, 25, 26] 不同，我们的方法非常容易实现。其次，它表现出强大的可扩展性。通过采用更强大的教师采样策略或更大的网络架构，我们的方法可以进一步增强学生模型。最后，结果如图 5b 所示，训练简单且收敛迅速。这些特性使得 TeEFusion 成为在最先进

Table 1. HPS 在美学基准上的结果（↑）。更高的分数表明生成的图像更符合人类的美学标准。”Cost” 指的是推理所需的时间，分别对每个模型进行测量和比较。”I.-T2I” 是”In-house T2I” 的缩写。学生模型的结果来自使用 W2SD+CFG 的教师模型。最佳结果以粗体显示。

Model	Method	Cost	Prompt Group			
			Anime	Concept-Art	Paintings	Photo
SD3	Teacher					
	CFG	2 ×	30.78	30.06	30.28	27.93
	W2SD+CFG	6 ×	31.96	30.65	30.67	29.76
	Student					
	DistillCFG	1 ×	31.14	29.52	30.03	29.04
	TeEFusion	1 ×	32.37	30.88	30.74	29.84
I.-T2I	Teacher					
	CFG	2 ×	30.65	29.27	28.94	27.99
	W2SD+CFG	6 ×	32.23	31.29	31.22	29.93
	Student					
	DistillCFG	1 ×	30.80	29.41	29.14	28.79
	TeEFusion	1 ×	32.39	31.34	31.27	29.96

的文本到图像生成系统中进行高效引导蒸馏的一个有吸引力且实用的解决方案。

6. 实验

在本节中，我们首先描述实验设置。接下来，我们展示主要结果，通过定量和定性分析证明 TeEFusion 的有效性。最后，我们进行综合消融研究，以调查我们方法中各个组件的贡献，并评估其在各种配置下的鲁棒性。

6.1. 实验设置

模型。我们采用了两种文本到图像生成模型，这两种模型都利用了 DiT [22] 架构和 Flow Matching [16] 框架：SD3 [5] 和我们的内部开发 T2I。SD3 [5] 是一个公开可用的开源模型，拥有 20 亿个参数，提供卓越的视觉质量和增强的组合一致性。内部开发的 T2I 针对电商场景中的逼真图像进行了优化，由十亿个参数组成。

训练细节。我们利用 LAION-5B [27] 的子集进行蒸馏。具体来说，选择分辨率高且美学评分 [27] 超过 6 的图像作为训练集。学生模型使用 AdamW [17] 优化，实际批处理大小为 128，默认学习率为 5×10^{-6} 。对于教师模型，我们采用 W2SD [2] +CFG [7] 作为采样策略，因为它在图 1 中表现出色。由于 W2SD 在采样过程中利用了两种不同优势的模型，这些模型是通过 CHATS [6] 获得的。我们实验中的所有蒸馏模型均在 W2SD+CFG 框架内训练，算法 1 中的 $w_{\min}$ 和 $w_{\max}$ 分别设置为 2 和 14。

基准。鉴于在大型文本到图像模型中直接应用指导蒸馏的研究有限，我们采用 DistillCFG [20] 作为我们的基准。DistillCFG 通过一个额外的 MLP 嵌入整合了指导尺度，如最近的 FLUX.1-dev [11] 。此外，我们将 TeEFusion 与教师模型进行比较，以进一步验证其有效性。

评估细节。HPS [34] 提供了一个广泛的美学评估基准，包含 3,200 个提示，平均分布在四种不同风格中：动漫、概念艺术、绘画和照片（每种风格 800 个提示）。这个广泛的提示集合支持稳健且一致的评估结果。DPG-Bench [8] 包含 1,065 个提示，每个提示描述了几个具有不同特征的物体，并解释这些物体之间的关系。它旨在测试文本到图像模型如何将这些元素组合成一个连贯的图像。用于美学评估，我们采用了三种最新的评估器：HPS [34] （及其 v2 版本）、ImageReward (IR) [35] 和 PickScore [10] 。我们还使用 CLIP [23] 分数来评估生成的图像与其对应输入提示的匹配程度。

6.2. 主要结果。

如表格 1 所示，我们的 TeEFusion 方法在美学评估中始终获得最佳分数，超越了 DistillCFG 基线，甚至在所有提示类别中都优于教师模型，包括 SD3 和内部 T2I 模型。表格 2 进一步强调了 TeEFusion 在 DPG-Bench 上的优秀表现，不仅在对象属性、空间关系和整体构图上表现出色，还获得了比 DistillCFG 更高的 CLIP 分数，突显了其忠实捕捉文本语义的能力。

这些结果清楚地表明，TeEFusion 通过线性融合在文本嵌入空间中有效地整合了引导强度，从而在显著提高采样效率的同时保留了语义结构和构图质量。值得注意的是，TeEFusion 实现了比教师模型快 6 × 的推理速度。总体而言，TeEFusion 成功地将复杂的采样策略简化为流畅的推理过程，以最少的折衷提供高质量的图像，使其成为现实应用中一种有前途的方法。

6.3. 消融研究

模块化消融。我们在内部 T2I 上使用动漫的提示进行模块化消融（见表 3 ）。通过用 $\mathcal{G}(\psi(w))$ 替换 DistillCFG 中使用的额外 MLP，我们观察到所有分数均有所提高。此外，完整的 TeEFusion 配置获得了最佳结果，始终优于所有其他变体。这些发现表明，TeEFusion 中的每个组件都对引导蒸馏的有效性有实质性贡献。值得注意的是，学生模型采用简单的欧拉采样而不使用 CFG。

切换教师采样。我们还研究了在不同教师模型采样策略下 TeEFusion 的有效性（参见表格 4 ）。结果表明，无论教师采样方法如何，TeEFusion 都能实现几乎无损的蒸馏。虽然我们观察到在从更简单的采样方法蒸馏到 Euler+CFG 和 Z-Sampling+CFG 时性能有轻微下降，但这些细微差别是可以预见的，因为与教师模型完美匹配是有挑战的。

重要的是，当从更健壮的采样方法（W2SD+CFG）进行蒸馏时，TeEFusion 不仅实现了近乎无损的蒸馏，而且在多个指标上略微超越了教师的表现。这表明 TeEFusion 在蒸馏过程中有效地保留了模型质量。此外，这些结果表明了 TeEFusion 的一个有前景的特性：随着更强的采样策略或更大的教师模型在未来可用，TeEFusion 的性能预计将进一步提高，使得我们的方法在测试时刻扩大时具备高度的可扩展性和适应性。

w -蒸馏成功吗？我们进一步考察了改变指导尺度 w 对 TeEFusion 性能的影响。如图 4 所示，TeEFusion

Figure 4. 在不同引导尺度 w 上各种方法的定性比较。提示词：1) 一个粉色头发和浓妆的电子女孩。2) 一个带有复古电子游戏氛围和电影灯光的舒适酒馆，具有卡通风格动画和细致的背景艺术。

Table 2. 关于 CLIP 得分和 DPG-Bench（%）的定量结果。更高的得分表明生成的图像在整体表现上更加优越，反映出增强的视觉质量、更精细的细节以及与输入提示的更好对齐。†： 使用动漫提示测量。

Model	Method	CLIP †	DPG-Bench					
			Global ↑	Entity ↑	Attribute ↑	Relation ↑	Other ↑	Overall ↑
SD3	Teacher							
	CFG	34.93	85.71	90.84	87.79	93.58	86.40	85.09
	W2SD+CFG	34.96	82.98	92.13	88.60	94.24	91.60	86.56
	Student							
	DistillCFG	34.53	81.46	90.47	87.85	93.42	84.40	84.13
	TeEFusion	34.63	81.76	90.60	88.11	93.15	86.80	84.56
In-house T2I	Teacher							
	CFG	34.25	83.89	88.08	88.07	91.57	76.40	81.88
	W2SD+CFG	34.62	82.37	91.41	89.08	93.73	78.80	85.20
	Student							
	DistillCFG	33.38	84.50	90.21	88.24	93.54	80.00	83.52
	TeEFusion	33.81	84.19	90.08	88.56	93.73	81.60	84.13

Table 3. 使用内部 T2I 结合来自动漫的提示对 TeEFusion 进行模块化消融。TeEFusion 的 “Config” 指定了方程 7 中 “额外项” 的配置。

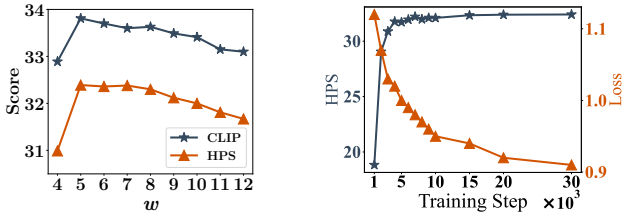
Config	Metrics			
	HPS ↑	IR ↑	PickScore ↑	CLIP ↑
DistillCFG	30.80	113.12	22.42	33.38
$\mathcal{G}(\psi(w))$	31.05	116.39	22.49	33.59
$\mathcal{G}(\psi(w))\,\mathcal{F}(c-\varnothing)$	32.39	128.50	22.68	33.81

在不同 w 设置下产生的图像既多样又在视觉上富有吸引力。此外，该模型在 $w = 5$ 时获得了最高的 HPS 和 CLIP 分数（见图 5a），这表明最佳的指导尺度能有效提升美学质量和提示依从性。在不同 w 值上的明显分数变化进一步证实了 TeEFusion 在蒸馏后仍然对指导尺度的变化保持敏感性。这些发现验证了我们方法在保留有意义的指导尺度特征方面的能力，从而使用户能够有效平衡文本-图像一致性和视觉美感。

TeEFusion 的收敛速度有多快？我们在图 5b 中进一步检查了 TeEFusion 的收敛速度。结果显示，在训练的早期阶段（低于 5000 步），HPS 值迅速增加，然后逐渐饱和。类似地，损失曲线开始快速下降，并在训

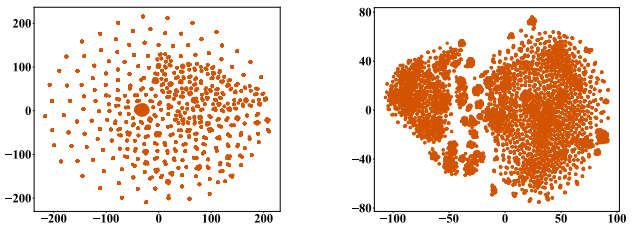
Table 4. 关于使用不同采样策略进行蒸馏的消融研究。在每个组中，第一行显示了教师的采样策略，第二行显示了使用该策略蒸馏的 TeEFusion。模型：内部 T2I。提示：动漫。

Sampling	Cost	Metrics			
		HPS ↑	IR ↑	PickScore ↑	CLIP ↑
Euler+CFG	2 ×	30.65	118.05	22.79	34.25
TeEFusion	1 ×	30.61	117.29	22.78	33.58
Z-Sampling+CFG	6 ×	31.99	125.42	22.61	34.47
TeEFusion	1 ×	32.01	125.07	22.65	33.62
W2SD+CFG	6 ×	32.23	127.56	22.49	34.62
TeEFusion	1 ×	32.39	128.50	22.68	33.81



(a) 指导刻度 w (b) 训练步骤

Figure 5. 关于变化 (a) 指导比例 w 和 (b) 内部 T2I 训练步骤数量的消融研究。



(a) DistillCFG (b) TeEFusion

Figure 6. 嵌入的 w 的插图，其值从范围 [2, 14] 中随机抽样，使用 t-SNE [31] 可视化。

练步骤超过 10000 后趋于平稳。这些发现清楚地表明 TeEFusion 收敛速度非常快。这个快速收敛可以归因于与引导尺度 w 相关的所有编码操作都利用了直接继承自预训练模型的模块，没有任何随机初始化的组件。事实上，对于 In-house T2I，当在 16 个 A100 GPU 上训练时，TeEFusion 在不到 4 小时内就收敛，展示了极为高效的训练过程。

定性分析。在图 4 中，我们展示了不同 w 下各种方法的定性结果。我们观察到，使用 W2SD+CFG 的教师模型生成的图像并不总能达到最佳的美学质量，特别是当 w 过高时。值得注意的是，在高 w 值时，W2SD+CFG 更容易崩溃。例如，左图显得不太吸引人，而右图则显示出明显的周期性伪影 [12]。相比之下，蒸馏模型并没有这些问题，这解释了为何 TeEFusion 由于其增强的稳定性，获得了比 W2SD+CFG 更高的美学评分。此外，我们一致观察到，TeEFusion 的表现优于 DistillCFG，尤其是在较低的 w 值时。

此外，在图 6 中，我们展示了嵌入的 w 的特征分布。一个有趣的观察是，由 TeEFusion 获得的嵌入 w 表现出更一致的分布，而 DistillCFG 的则显得明显更离散。这可以归因于 TeEFusion 额外利用了来自 $c - \varnothing$ 的信息，这导致了一个更平滑的指导信号，从而提高了生成质量。

7. 结论和局限性

在本文中，我们介绍了一种高效的蒸馏方法 TeEFusion，该方法将引导幅度融合到文本嵌入中，从而简化了文本到图像合成的推理过程。通过线性结合有条件和无条件嵌入，TeEFusion 能够复现教师模型的性能，达到了相似的图像质量，同时实现了最高可达 × 倍的推理速度。然而，TeEFusion 有时会产生语义不一致（例如属性不匹配），其输出不总是与教师模型完全一致（参见图 4 ）。未来的工作将致力于缓解这些问题。

References

[1] Lichen Bai, Shitong Shao, Zikai Zhou, Zipeng Qi, Zhiqiang Xu, Haoyi Xiong, and Zeke Xie. Zigzag diffusion sampling: Diffusion models can self-improve via self-reflection. In International Conference on Learning Representations, pages 1–14, 2025.

[2] Lichen Bai, Masashi Sugiyama, and Zeke Xie. Weak-to-strong diffusion with reflection. arXiv:2502.00473, 2025.

[3] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 843–852, 2023.

[4] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18392–18402, 2023.

[5] Patrick Esser, Sumith Kulal, Andreas Blattmann, et al. Scaling rectified flow Transformers for high-resolution image synthesis. In International Conference on Machine Learning, pages 12606–12633, 2024.

[6] Minghao Fu, Guo-Hua Wang, Liangfu Cao, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. CHATS: Combining human-aligned optimization and test-time sampling for text-to-image generation. arXiv:2502.12579, 2025.

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. arXiv:2207.12598, 2022.

[8] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. EllA: Equip diffusion models with LLM for enhanced semantic alignment. arXiv:2403.05135, 2024.

[9] Tero Karras, Miika Aittala, Samuli Laine, and Timo Aila. Elucidating the design space of diffusion-based

generative models. In Advances in Neural Information Processing Systems, pages 26565 – 26577, 2022.

[10] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In Advances in Neural Information Processing Systems, pages 36652–36663, 2023.

[11] Black Forest Labs. FLUX.1-dev. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024.

[12] Jianze Li, Jiezhang Cao, Yong Guo, Wenbo Li, and Yulun Zhang. One diffusion step to real-world super-resolution via flow trajectory distillation. arXiv:2502.01993, 2025.

[13] Yanyu Li, Huan Wang, Qing Jin, et al. SnapFusion: Text-to-image diffusion model on mobile devices within two seconds. In Advances in Neural Information Processing Systems, pages 20662–20678, 2023.

[14] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In International Conference on Learning Representations, pages 1–13, 2023.

[15] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In International Conference on Learning Representations, pages 1–11, 2022.

[16] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In International Conference on Learning Representations, pages 1–15, 2023.

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, pages 1–18, 2019.

[18] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11451–11461, 2022.

[19] Zhiyuan Ma, Yuzhu Zhang, Guoli Jia, et al. Efficient diffusion models: A comprehensive survey from principles to practices. arXiv:2410.11795, 2024.

[20] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14297–14306, 2023.

[21] OpenAI. DALL · E 2, 2022. https://openai.com/dall-e-2/.

[22] William Peebles and Saining Xie. Scalable diffusion models with Transformers. In IEEE/CVF International Conference on Computer Vision, pages 4172–4182, 2023.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, pages 8748–8763, 2021.

[24] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In International Conference on Learning Representations, pages 1–21, 2022.

[25] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In ACM SIGGRAPH Conference and Exhibition on Computer Graphics and Interactive Techniques in Asia, pages 1–11, 2024.

[26] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In European Conference on Computer Vision, pages 87–103, 2024.

[27] Christoph Schuhmann, Romain Beaumont, Richard Vencu, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In Advances in neural information processing systems, pages 25278–25294, 2022.

[28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In International Conference on Learning Representations, pages 1–12, 2022.

[29] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In International Conference on Machine Learning, pages 32211 – 32252, 2023.

[30] Zhicong Tang, Jianmin Bao, Dong Chen, and Baining Guo. Diffusion models without classifier-free guidance. arXiv:2502.12154, 2025.

[31] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. Journal of Machine Learning Research, 9(86):2579–2605, 2008.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In Advances in Neural Information Processing Systems, page 6000–6010, 2017.

[33] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, et al. Tune-A-Video: One-shot tuning of image diffusion models for text-to-video generation. In IEEE/CVF International Conference on Computer Vision, pages 7589–7599, 2023.

[34] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human Preference Score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv:2306.09341, 2023.

[35] Jiazheng Xu, Xiao Liu, Yuchen Wu, et al. ImageReward: Learning and evaluating human preferences for text-to-image generation. In Advances in Neural Information Processing Systems, pages 15903–15935, 2024.

[36] Yilun Xu, Weili Nie, and Arash Vahdat. One-step diffusion models with f-Divergence distribution matching. arXiv:2502.15681, 2025.

[37] Tianwei Yin, Michael Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6613–6623, 2024.

[38] Bowen Zheng and Tianming Yang. Diffusion models are innate one-step generators. arXiv:2405.20750, 2024.

[39] Hongkai Zheng, Weilie Nie, Arash Vahdat, Kamyar Azizzadenesheli, and Anima Anandkumar. Fast sampling of diffusion models via operator learning. In International Conference on Machine Learning, pages 42390 – 42402, 2023.

[40] Kaiwen Zheng, Cheng Lu, Jianfei Chen, and Jun Zhu. DPM-Solver-v3: Improved diffusion ode solver with empirical model statistics. In Advances in Neural Information Processing Systems, pages 55502 – 55542, 2024.

[41] Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity Distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In International Conference on Machine Learning, pages 62307–62331, 2024.

[42] Zhenyu Zhou, Defang Chen, Can Wang, Chun Chen, and Siwei Lyu. DICE: Distilling classifier-free guidance into text embeddings. arXiv preprint arXiv:2502.03726, 2025.

Figure 7. 失败案例的生成示例。提示：1）不是猫。2）液体玻璃。3）冷火。

A. 更多实验结果和分析

为了验证附加文本嵌入的有效性，我们在不同的文本到图像模型中进行了定量实验。原始嵌入和融合嵌入之间的余弦相似度（余弦相似度 $_{\text{txt}}$ ）及其对应生成图像（余弦相似度 $_{\text{img}}$ ）的结果总结在下表中：

Metric	SD3	In-house T2I	FLUX.1-dev
Cos Sim.$_{\text{txt}}$	0.8073	0.8192	0.8286
Cos Sim.$_{\text{img}}$	0.8732	0.9137	0.9318

这些结果证实，加性嵌入操作在文本空间中保留了超过 80 % 的余弦相似性，在图像空间中则超过 90 %，证明了它们有效融合多样语义模式的能力。

我们的融合机制 $\mathcal{G}(\psi(w))\,\mathcal{F}(c-\varnothing)$ 在编码器的线性机制中通过有界的正弦-余弦位置编码（$\|\mathcal{G}(\psi(w))\,\mathcal{F}(c-\varnothing)\|_2 \leq \delta$ ）进行操作。然而，失败的情况发生在：

- 语义向量表现出非正交性（例如，像 “冷火” 这样的矛盾短语）
- 上下文干扰发生在复合提示中（例如，“not a cat”）

这些限制在图 7 中展示。