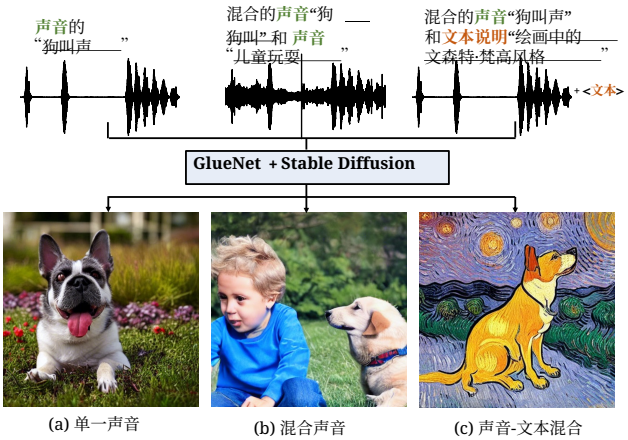
arXiv:2303.10056v2 [cs.CV] 2 Nov 2023

GlueGen：即插即用的 X到图像生成多模态编码器

Can Qin*†, Ning Yu†, Chen Xing†,ShuZhang†,Zeyuan Chen†,Stefano Ermon‡, YunFu*,CaimingXiong†,RanXu†*东北大学， †Salesforce AI Research， ‡斯坦福大学qin.ca@northeastern.edu, ermon@cs.stanford.edu, yunfu@ece.neu.edu,{ning.yu, cxing, shu.zhang, zeyuan.chen, cxiong, ran.xu}@salesforce.com

摘要

基于扩散过程的文本到图像（*T2I*）模型在利用用户提供的图像说明进行可控图像生成方面取得了显著成功。然而，当前 *T2I* 模型中的文本编码器与图像解码器紧密耦合，导致替换或升级它们颇具挑战。这类改动往往需要大量微调，甚至从头训练，代价高昂。为解决这一问题，我们提出 *GlueGen*，它采用新提出的 *GlueNet* 模型，将单模态或多模态编码器的特征与现有 *T2I* 模型的潜在空间对齐。该方法引入了一种新的训练目标，利用平行语料库对齐不同编码器的表示空间。实证结果表明，*GlueNet* 能够高效训练，并实现了超越以往最先进模型的多种能力：*1*）可将 *XLM-Roberta* 等多语言语言模型与现有 *T2I* 模型对齐，从而根据英文以外语言的图像说明生成高质量图像；*2*）*GlueNet* 可将 *AudioCLIP* 等多模态编码器与 *Stable Diffusion* 模型对齐，实现声音到图像生成；*3*）它还可以升级潜在扩散模型当前的文本编码器，以生成具有挑战性的案例。通过对齐各种特征表示，*GlueNet* 能够灵活高效地将新功能集成到现有 *T2I* 模型中，并为 *X*到图像（*X2I*）生成提供了新的思路。[1]



(a) 单一声音 (b) 混合声音 (c) 声音-文本混合

图 1. GlueGen 的设定。GlueNet 旨在为稳定扩散模型提供一个灵活的入口，以接收多模态数据，例如文本、音频，即（a）和（b），或文本-音频混合信号，即（c），用于 X到图像生成。

1. 引言

得益于算法进展，文本到图像（T2I）生成模型在过去几年中取得了巨大进展以及大规模配对训练数据集[43, 72, 46, 51, 50]的出现。尤其是基于扩散的文本到图像生成模型，在图像质量方面取得了显著成果 [24, 37, 3, 48, 36, 75]。尽管成果斐然，这些方法仍难以实现可控生成：生成图像往往无法忠实呈现图像说明，组合能力不足，而且通常需要进行提示工程才能获得理想结果 [38]。此外，大多数大规模模型仅使用英文文本描述进行训练，极大地限制了它们在世界各地的应用。

近期研究强调了文本编码器在提升文本到图像模型性能方面的关键作用，其理解和表征文本的能力被认为是图像生成的瓶颈 [48, 10]。然而，当前文本到图像模型的文本编码器通常是在简短的图像说明上训练的，这限制了它们处理复杂提示的表现，并对其特征提取能力提出了挑战

*本研究是在 Can Qin 于 Salesforce AI Research 实习期间完成的。主要联系人：qin.ca@northeastern.edu[1]代码将发布于：https://github.com/salesforce/GlueGen

。 [46]此外，文本到图像模型的能力仅限于根据文本生成图像，无法轻松纳入声音和音频等多模态条件。尽管如此，替换现有文本到图像模型中的文本编码器仍颇具挑战，因为文本编码器与图像生成器的表示空间紧密耦合 [46, 43]。新条件与现有模型之间严重的领域差距会妨碍图像生成的最终效果；而使用质量更高的图像-描述对从头训练整个文本到图像模型，成本将高得难以承受 [14]。[2]

如图1所示，我们提出了 GlueNet，以应对如何高效替换或升级现有基于扩散的文本到图像模型中的文本编码器这一挑战。借助 GlueNet，现成的预训练语言模型和多模态编码器可以轻松地与文本到图像模型的图像编码器对齐，以较低成本大幅扩展其功能。重要的是，这一过程无需从头训练，甚至无需微调，同时还能保持文本编码器与图像编码器之间的表示对齐。所提出的方法采用编码器-解码器结构。GlueNet 的编码器首先将新条件编码器的表示空间与文本到图像模型图像生成器的表示空间对齐，同时最小化逐元素差异和全局差异。随后，GlueNet 的解码器通过最小化重建损失，将对齐后的条件表示映射回新条件编码器的原始表示空间，从而在对齐训练过程中保留预训练模型所捕获的丰富语义。对现有模型进行对齐不可避免地会降低特征判别性 [7, 11], , 因此特征解码器必不可少。GlueNet 的整个训练过程只需要一个内容相同、但模态或语言不同的平行语料库。在推理阶段，只需在新的条件编码器上应用 GlueNet 的编码器，即可实现表示对齐。

为验证所提出框架的有效性，我们针对单模态和多模态编码器开展了三项主要实验。首先，我们使用更强大的语言模型 T5-3B [42]，升级了潜在扩散模型 [46] 中现有的文本编码器。与基线相比，我们的模型在 FID 分数和用户研究排名方面取得了有竞争力的提升，但要全面提升性能，仍需进行微调。其次，我们使用我们的方法对多语言语言模型 XLM-Roberta-L [9]，进行对齐，从而实现多语言文本到图像生成。与基于翻译的模型相比，该模型以显著更低的训练成本取得了有竞争力的结果。最后，我们展示了 GlueNet 将文本信号以外的新功能引入现有文本到图像模型的能力。对 AudioClip [20] 编码器进行对齐后，无需进行任何图像生成器的参数微调，即可实现声音到图像生成。这项新能力使现有的 Stable Diffusion 模型能够根据狗叫声、街头音乐等声音信号生成高质量图像。这项新能力超越了传统的文本生成图像，并为面向 X到图像（X2I）生成的多媒体内容创作开辟了新的可能性。

我们的贡献总结如下：

- 据我们所知，这是首项探讨如何高效对齐预训练音频模型与预训练文生图扩散模型，以实现声音到图像生成的工作。

- 在文生图生成基准上进行的大量实验表明，与基线潜在扩散模型方法相比，我们的模型在图像质量和语言可控性方面都更胜一筹。

- 我们的框架还支持使用英文提示词以外的语言进行文本到图像生成，且无需多语言图文对来重新训练。

2. 相关工作

2.1. 文本到图像模型

生成对抗网络（GAN） [18] 是文本到图像生成的主要框架之一。 [45] 中的方法是早期采用生成对抗网络训练文生图生成网络的一种方法。此后，其他基于生成对抗网络的方法，例如 Stack-GAN [74], 、Attn-GAN [68] 和 SD-GAN [71], , 都取得了令人瞩目的成果。此外，近期的研究进一步提升了生成质量。DM-GAN [77]通过引入动态记忆组件提升了文本到图像的表现。DF-GAN [58] 设计了一个融合模块，用于融合文本和图像特征。LAFITE [76] 利用 CLIP 模型构建伪图文对，并提出了一种生成对抗网络模型来学习图文对。

自回归 Transformer 是文本到图像生成的另一种主要框架。DALL-E [44] 和 CogView[13] 采用自回归 Transformer [65] 来学习视觉词元与文本词元之间的对应关系。Parti [72] 使用强大的视觉编码器 ViT-VQGAN[73], , 在 DALL-E 的基础上提升了效果。[16]中的方法与 CogView 和 DALL-E 相似，同时引入了额外的控制元素来改进词元化¨过程。NUWA [67] 和 NUWA-infinity [66] 训练了自回归视觉合成模型，以支持文本到图像生成和文生视频任务。

与此同时，扩散模型 [55, 56, 57] 成为了研究的主要焦点。在这类方法中，向图像添加噪声，并训练评分网络对其进行去噪，以恢复输入图像。GLIDE [37] 采用有引导的

[2]训练一个 Stable Diffusion 模型的成本约为 60 万美元。

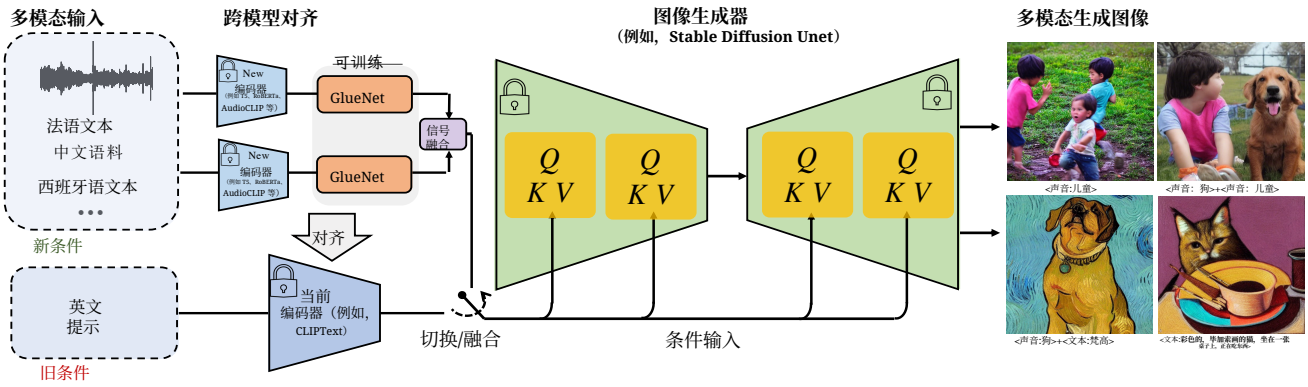


图 2. 我们所期望的 GlueGen 框架示意图。借助 GlueGen 框架中提出的 GlueNet 模型，预训练的图像生成器（即 UNet）可以与现成的单模态或多模态编码器连接，从而以有限的预算扩展其功能，即，实现多语言／声音到图像生成。GlueNet 在线下训练，无需对 UNet 进行反向传播，也无需使用图文对进行训练。因此，GlueGen 灵活且高效。

推理方法，在有无分类器网络的情况下生成高保真图像。DALL-E 2 [43] 和 Imagen [48]在文本到图像生成领域创下了新的最先进成果。在 DALL-E 2 中，模型学习了一个先验模型，用于根据文本生成 CLIP 图像嵌入表示，并使用扩散模型将图像嵌入表示解码为图像。在 Imagen 中，采用大规模冻结的文本编码器 T5-XXL[42] 生成嵌入表示，并使用一系列条件扩散模型将这些嵌入表示映射为分辨率逐渐提高的图像。潜在扩散模型和稳定扩散模型 [46] 是最先进的方法，它们像 [54, 62]一样在潜在嵌入空间中应用扩散模型。这些方法计算效率高，同时取得了令人瞩目的成果。为了融入更多条件以实现可控内容生成，[75, 36, 4]引入了额外参数和新数据，以便在当前预训练模型中注入外部知识。 [8] 尝试改造语言模型以提升其能力。然而，所有这些模型都是为适配文本和图像（或类似的二维网格）条件而设计的，并未尝试处理音频等序列信号。

2.2. 声音到图像生成

声音和图像是两种截然不同的模态。 [26] 尝试通过将声音特征与生成对抗网络模型对齐，来解决声音引导的语义图像操控问题。CoDI [6] 采用两阶段策略：首先，使用对齐的提示编码器对各个扩散模型进行预训练；随后，对齐潜在表示，将多模态特征投影到共享空间中。AAI [70] 将声音引导的图像生成、编辑和风格化整合到一个模型中。

3. 预备知识

3.1. 潜在扩散模型

潜在扩散模型（LDM） [46]及其扩展版本，即 Stable Diffusion[3], , 属于去噪扩散概率模型（DDPM） [23] 系列的变体。由于在效率和视觉质量之间取得了出色的平衡，我们选择它们作为基线模型。一般来说，LDM由两个阶段组成。首先，使用海量图像对自编码器（AE）进行预训练，并采用KL散度或向量量化 [63, 1]进行正则化。随后，使用编码器网络（即 E）进行潜在特征提取，得到 $z = E(x)$；解码器网络可将其映射回图像空间。输入图像 x与重建图像$\hat{}x$ 几乎完全相同 $x \approx \hat{}x$。

其次，LDM 在潜在空间中训练扩散模型 [54, 62]。它遵循标准的去噪扩散概率模型（DDPM） [23] ，采用去噪损失，并使用 U-Net [47] 作为图像解码器，如 [56]所示。为了实现生成可控性，LDM 引入了多种条件信号（y），例如文本、掩码或布局，这些信号经过单独编码后，在交叉注意力层的帮助下输入 U-Net。其形式化表达如下：

$$\mathcal{L}_{LDM} := \mathbb{E}_{z\sim E(x),\varepsilon\sim N(0,1),t,y}\left[\|\varepsilon - \varepsilon_\theta(z_t, t, c_\phi(y))\|\right], \tag{1}$$

其中， t 表示时间步， z_t 是时间步 t时噪声污染的潜在张量， $z_0 = E(x)$。 ε 是未缩放的高斯噪声， c_ϕ 是由 ϕ 参数化的条件网络，而 ε_θ 是类似 UNet 的去噪网络（亦称图像解码器）。条件网络和去噪网络的参数，即 θ, ϕ，均通过潜在扩散模型损失进行训练。在推理过程中，可以生成干净的图像——

[3]https://github.com/CompVis/stable-diffusion

通过无分类器引导 [24]如下：

$$\hat{\varepsilon}_\theta(z_t|y) = \varepsilon_\theta(z_t) + s \cdot (\varepsilon_\theta(z_t, c_\phi(y)) - \varepsilon_\theta(z_t)), \quad (2)$$

其中， s 是用于平衡文本可控性和图像保真度的引导权重。

基础文本到图像潜在扩散模型在 Laion- 400M [51] 数据集上训练。稳定扩散模型（SDM）经过了更多轮次的训练，并使用了更多训练数据，即，Laion2B-en 和 Laion-high-resolution [50]。

3.2. 条件编码器

潜在扩散模型使用类似 Bert 的 [12] 编码器作为文本编码器，并通过去噪扩散概率模型损失与图像解码器联合训练。SDM 使用预训练的 CLIP [41] 文本编码器，该编码器在训练期间保持冻结。研究发现，增大文本编码器的规模对提升性能非常有帮助 [48]。为了升级现有模型，我们的目标是能够高效地接入更先进的语言条件编码器，例如 T5-3B [42], AudioClip [20], 或 XLM-Roberta [12]。

4. GlueNet

文本编码器是文本到图像模型的关键组件之一，因为生成过程需要精确且细粒度的文本嵌入来提供引导。增大文本编码器的规模可以大幅提升整体性能 [48], ，然而，将现有文本模型升级为功能更强大的模型颇具挑战。现有文本到图像模型并非模块化设计，因此直接替换（或增添）另一个条件编码器并不可行。关键技术难点在于新条件编码器与旧图像解码器之间存在不对齐。此外，联合微调也无法解决问题，因为更新训练充分的参数时会发生灾难性遗忘。

将这两个不同的条件编码器视为源域和目标域后，我们使用神经网络学习对齐。如图2所示，本文提出了一个名为 GlueGen 的通用框架。在该框架中，GlueNet 作为一个附加模块，用于连接新的语言模型和旧的图像解码器。

4.1. 目标

给定两组平行语料库 $\mathcal{X}_{src} = \{x^s_{ij} | i = 1, ..., L^s, j = 1, ..., N^s\}$ 和 $\mathcal{X}_{tar} = \{x^t_{ij} | i = 1, ..., L^t, j = 1, ..., N^t\}$ L N，其词元长度为，总数量为，源模型 F^s 和目标编码器 F^t 已将原始数据映射为嵌入表示，记作 $\mathcal{S} = \{s_{ij} | i = 1, ..., L^s, j = 1, ..., N\}$，其中 $s_{ij} = F^s(x^s_{ij}) \in \mathbb{R}^{d_s}$，以及 $\mathcal{T} = \{t_{ij} | i = 1, .., L^t, j = 1, ..., N\}$，其中 $t_{ij} = F^t(x^t_{ij}) \in \mathbb{R}^{d_t}$。由于模型不同，源特征与目标特征之间存在严重的分布失配 $p(s) \neq p(t)$。GlueNet 是一种自编码器，其编码器 $M(\cdot, \Theta_M)$ 学习将来自源语言模型（即， $s \in \mathcal{S}$）的新特征映射为与当前目标编码器（即， $t \in \mathcal{T}$）对齐，其中 $p(M(s, \Theta_M)) \approx p(t)$。因此，这种转换（即， $F^s \xrightarrow{\Theta_M} F^t$）使图像生成器无需微调即可理解来自新条件模型的新特征。为实现这一目标，我们将其视为一个域适应 [17, 40, 32]问题，并同时采用逐元素对齐损失和分布级对齐损失，其中包括均方误差（MSE）损失的最小化（即， $\mathcal{L}_{mse}$），以及在判别器网络（即， $\mathcal{L}_{adv}$）上测得的对抗损失 $D(\cdot, \Theta_D)$:

$$\mathcal{L}_{mse}(\Theta_M) := \mathbb{E}_{x_t \sim X_t, x_s \sim X_s}[||F^t(x_t) - M(F^s(x_s), \Theta_M)||_2^2], \quad (3)$$

$$\mathcal{L}_{adv}(\Theta_D, \Theta_M) := \mathbb{E}_{x_t \sim X_t}[\log D(F^t(x_t), \Theta_D)] + \mathbb{E}_{x_s \sim X_s}[\log 1 - D(M(F^s(x_s), \Theta_M), \Theta_D)], \quad (4)$$

其中， X_s 和 X_t 分别表示源输入和目标输入，它们可以是相同的英文提示词（例如，LDM → T5-3B [42]），也可以是用于多语言文本到图像或音频-标签对的双语平行内容。

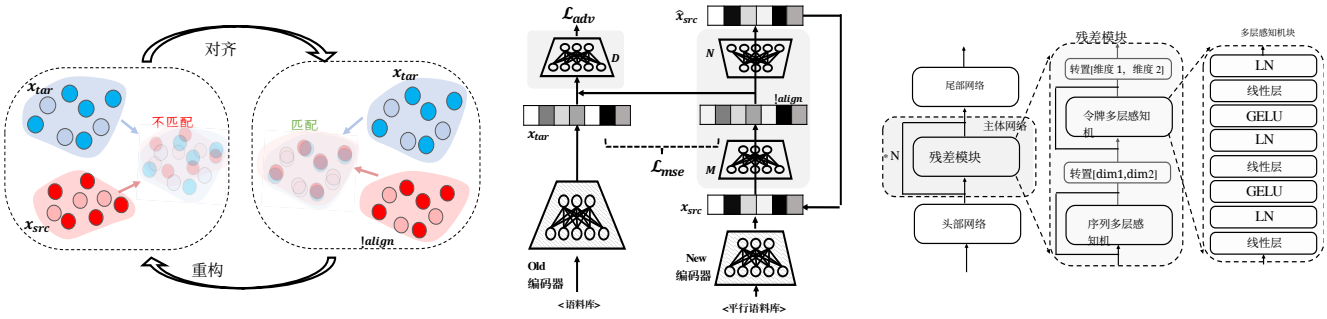
迁移学习中的一些深入研究表明，跨域对齐不可避免地会降低特征判别性 [7, 11]。为了投射模型升级所需的丰富语义，我们进一步使用解码器网络，即即， $N(\cdot, \Theta_N)$，进行特征重建:

$$\mathcal{L}_{rec}(\Theta_M, \Theta_N) := \mathbb{E}_{x_s \sim X_s}[||x_s - N(M(F^s(x_s), \Theta_M), \Theta_N)||_2^2]. \quad (5)$$

GlueNet 以端到端的方式训练，相关细节请参阅附录和第5节。[35] 指出，通用学习引擎应寻求具有良好结构保持特性的压缩式闭环转录。我们的跨模型映射也需要类似的功能。由于重构损失的约束，GlueNet 希望循环前后的输入和输出几乎保持等价，从而确保循环过程的稳定性；ˆ也就是说， $x \approx \hat{x} \approx \hat{x} \approx ...$。因此，将重构后的数据ˆ$x$ 输入 GlueNet 后，预期其结果与前一次循环保持一致。这有助于加速 GlueNet 在 T5 到 LDM 适配过程中的训练。

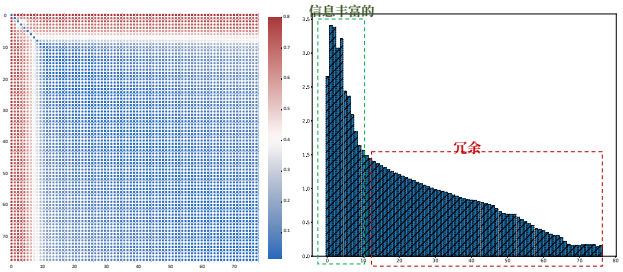
4.2. 模型架构

对齐的输入数据可表示为如下词元序列：$X_{i,*} = \{x_{i,1}, x_{i,2}, ..., x_{i,L}\}$ 其中$X_{i,*} \in \mathcal{X}$ 和$x_{i,j} \in \mathbb{R}^C$，序列长度为 L ，词元维度为 C。受图像超分辨率方法的启发， [28, 2] 为在不同输入输出维度下密集回归目标数据，翻译器网络依次堆叠了三个子网络，如图3 (c) 所示。头部网络仅用于简单的特征变换。主体



(a) (b) (c)

图3。（a）模型转换/对齐过程中各项特征变换的示意图。（b）我们提出的 GlueNet 的总体流程和学习目标。（c）GlueNet 编码器/解码器的详细架构。



(a) (b)

图4。对 CLIPText 编码器处理 77 个词元（用于 Stable Diffusion）时的信号分析，结果取自 10,000 条随机采样提示的平均值。（a）CLIPText 词元的平均差异度图（1−余弦相似度）。（b）各词元与最后一个词元之间的距离，即，$dist(s_{:,j}, s_{:,L})$，其中有 L 个词元。由此可知，只有前 K 个词元包含有效信息。

网络包含许多残差模块，用于细粒度表征学习。尾部网络附加在末端，主要用于维度转换，以匹配目标张量。每个残差模块均基于高效且具有强大表征学习能力的基于 MLP 的词元混合器实现 [61, 34] 。在 MLP 混合器模块中，它由令牌混合器和序列混合器组成，分别用于正交目的。令牌混合器通过共享的 MLP 学习每个词元的表征，而序列混合器旨在学习序列词元中的相同通道：

$$\mathrm{U}_{:,i} = \mathrm{X}_{:,i} + \mathrm{W}_2\sigma(\mathrm{W}_1 * \mathrm{LN}_{\gamma_1,\beta_1}(\mathrm{X})_{:,i}), i = 1, ..., C \quad (6)$$

$$\mathrm{X}_{j,:} = \mathrm{X}_{j,:} + \mathrm{W}_4\sigma(\mathrm{W}_3 * \mathrm{LN}_{\gamma_2,\beta_2}(\mathrm{U})_{j,:}), j = 1, ..., L \quad (7)$$

其中，$\mathrm{X}_{:,:}$ 表示特征，$\mathrm{LN}_{\gamma,\beta}$ 表示由 γ 和 β 参数化的层归一化，σ 表示非线性算子，例如 GELU [21]。$\mathrm{U}_{:,i}$ 表示序列多层感知机的输出，W 表示 MLP 混合器中的权重矩阵。

4.3. 跨模态对齐

GlueNet 提供了一种对齐文本编码器与 CLIPText [41]，之间跨模态表示的方法，以及将声音信号映射到嵌入空间的 AudioCLIP [20]模型。这种对齐有助于扩展 Stable Diffusion 等扩散模型的功能，这些模型仅接受文本信号作为输入。然而，实现这种对齐颇具挑战，因为每个模型中词元的长度不同。CLIPText 的嵌入表示需要序列中包含 77 个词元，而 AudioCLIP 会将声音信号映射为单个词元，序列长度为 1。在这种情况下，直接使用标准 GlueNet 会失败。

在探索过程中，我们发现，CLIPText 输出的全部 77 个词元中，条件信息的分布并不均匀。前 K 个词元包含了大部分信息。我们计算了 CLIPText 词元的平均差异图，并在图中进行了可视化。4 (a)，观察到第八或第九个词元附近存在明显差距，而之后的词元彼此高度相似。我们还发现，给最后一个词元嵌入添加一些随机噪声，并不会降低生成结果。因此可以得出结论，不同词元的贡献逐渐递减，并且可以通过与最后一个词元的特征距离来估算，$w_{:,j} = dist(s_{:,j}, s_{:,M})$，如图所示。4 (b)。我们采用 l2 距离作为 $dist(\cdot, \cdot)$，并使用该值对声音到图像生成中不同词元对应的 GlueNet 目标函数进行重新加权。

跨模态信号融合是另一项贡献。为此，我们从 GlueNet 输出的每种模态中选取前 K 个词元。然后，将两种模态特征中其余后续词元（不包括最后 K 个词元）的平均值与选出的前 K 个词元拼接起来。这是一种无需训练的非参数操作。K 的取值根据经验设为 4 到 8（或更高）。更多详情请参阅附录。

5. 实验

5.1. 实验设置

站在巨人的肩膀上，本文采用潜在扩散模型（LDM）和稳定扩散模型（SDM） [46] 作为基线，两者模型相似，但文本编码器和可扩展性有所不同，



图 5. XLM-Roberta 的多语言生成结果，分辨率为 512 × 512 + Glue-Net + SDM 解码器（sd-v1-4）使用相同的图注，即，“印象派画家绘制的午后花园油画”。借助不同的 Glue-Net 和多语言文本编码器，SDM 解码器能够支持多种语言，包括日语、意大利语、中文、法语和西班牙语。引导权重设为 7.5，PLMS [30] 采样步数为 50。

。LDM 使用类似 Bert 的 [12] 模型作为文本编码器，并基于去噪扩散概率模型损失与图像解码器联合更新；而 SDM 在训练期间使用冻结的预训练 CLIP 文本编码器。LDM 面向 256 × 256 中的图像，SDM 则以 512 × 512 图像为目标。我们基于 SDM 实现了多语言文本到图像和声音到图像生成。

实现。GlueNet 的实现如图 3所示，包含一个编码器 M和一个解码器N ，采用对称架构，由 ∼51M 个参数组成。为简化记法，本文将 GlueNet 编码器简称为 GlueNet。

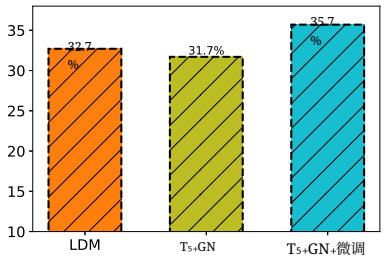


图 6. 使用 Drawbench [48] 和 Winoground [60] 中的提示评估 T5-to-LDM 文本可控性的用户投票百分比。

在以下实验中，其主体网络默认由 5 个残差模块组成。我们采用 AdamW [33]作为基于 PyTorch [39]的优化器。GlueNet 训练的学习率设为 1×10^{-4} ，LDM-T5-GlueNet 微调的学习率设为 1×10^{-6}。GlueNet 在 Nvidia-A100 上训练，根据不同用途需要 0.5 至 5 个 GPU 天。

条件编码器。我们使用 T5-3B [42]升级 LDM Bert 的文本编码器。T5-3B 的参数量是 LDM Bert 的 6 倍，并且使用规模庞大的文本语料库进行训练，其语料范围远超图像描述。此外，为了实现多语言文本到图像生成，我们选择 XLM-Roberta-L [12] 作为新的条件编码器。在声音到图像生成中，我们采用 AudioCLIP [20]音频编码器。

数据集。我们收集了用于训练 GlueNet 的文本数据。在单语种文本到图像（T2I）任务中，只需要英文文本。我们从 Laion-400M 中提取了 1800 万条图像说明作为训练语料库，更多数据将更有帮助。

多语言 GlueNet 网络需要不同语言的平行文本。这些数据通过 Wikimarix [52]和 Googletrans 收集，每种方法均包含 ∼2百万个句子。为了训练用于声音到图像生成的 GlueNet，我们使用了 Urban-sound8k 数据集 [49][4]，，其中包含从城市环境中收集的 8000 多个音频样本，涵盖十个不同类别，例如狗叫声、儿童玩耍、汽车喇叭和警笛等。

5.2. 单语文本到图像生成

单语文本到图像生成侧重于在给定英文文本提示的情况下与 LDM 进行比较。我们将流程分为两个阶段：对齐和微调。GlueNet 有助于对齐新的语言模型，而通过进一步微调图文对，可以更充分地发挥其潜力。本节通过定量和定性评估，验证了基础模型和微调模型的优势。**T5+GlueNet**指的是将 T5 和 GlueNet 堆叠起来作为文本编码器、且 LDM 图像解码器保持不变的 LDM 模型。**T5+微调** 表示在固定 T5 提供的文本条件下，使用去噪扩散概率模型损失对 LDM 图像解码器进行微调。**T5+GlueNet+微调** 是最终版本，包含 GlueNet 以及与图像解码器的联合微调。微调使用 ∼ 100GPU 天，对 Laion-400M 中经 BLIP分数筛选出的 1.16 亿个图文对进行训练 [27]。UNet 微调是可选的，在第5.3 节和第5.4节的其余实验中并非必需。

5.2.1 定量评估

表。1 报告的零样本FID分数 [22] 在 COCO 数据集上 [29] 基于 pyorch-fid [53] 软件包。我们重新实现的 LDM 分数略高于论文中的分数 [46]，这可能是由于实现方式不同。表格显示，T5+GlueNet 相较于 LDM 略有下降，因为我们当前的 GlueNet 模型仍存在一些

[4]https://urbansounddataset.weebly.com/urbansound8k.html

未对齐之处，这些问题可以通过联合微调或收集更多文本语料库来训练 GlueNet 加以解决。使用 ∼100 GPU 天对 LAION 数据集的一个子集（即，1.16 亿图文对）进行微调后，整体性能显著提升，甚至在一定程度上超过了 LDM。

我们还通过用户研究比较了图 6中由 Amazon Turk [5] 进行的文本可控性研究。共收集了∼1K 个来自 Drawbench 的提示 [48] 和 Winoground 的提示 [60]，全面涵盖了各种场景和许多困难案例。

表 1. COCO 上的 FID。* 或 ** 表示我们分别使用 100 或 200 GPU 天实现的结果。ZS 表示零样本。

方法	FID↓	ZS
CogView [13]	27.10	✗
LAFTTE [76]	26.94	✗
GLIDE [37]	12.24	✗
Make-A-Secne [16]	**11.84**	✗
潜在扩散模型 [46]	12.63	✓
LDM*	13.55	✓
T5+微调*	23.30	✓
T5+微调**	12.41	✓
T5+GlueNet	14.32	✓
T5+GlueNet+微调*	12.05	✓

我们在图6中报告了用户对三种方法的投票百分比。在可控性用户研究中，我们的方法优于潜在扩散模型 <0 3%，这是相当显著的提升。此外，进行了三轮用户研究后，计算得出的 p 值为 0.04；根据 <00.05 的标准，这一结果具有统计显著性。

5.2.2 视觉比较

在图8中，我们展示了不同方法生成的一些图像。通过比较潜在扩散模型和 T5+GlueNet，我们可以得出结论：即使替换条件，所提出的 GlueNet 也有助于生成质量相当高的图像。此外，与潜在扩散模型相比，T5+GlueNet+微调在许多情况下表现出更强的可控性，例如“由砖块制成的立方体”；而潜在扩散模型的结果未能呈现“立方体”这一概念。

5.3. 多语言文本到图像生成

我们提出的框架具有很强的通用性，适用于许多语言模型，包括多语言骨干网络。我们选择 SDM 作为基础模型，并采用预训练的 clip-vit-large-patch14[5] 作为文本编码器，再将其升级为多语言编码器，即即。，XLM-Roberta-L。GlueNet 使用平行语料库作为训练数据，帮助对齐这两个骨干网络。因此，旧的 SDM 图像解码器能够成功理解由 GlueNet 和 XLM-Roberta-L 编码的其他语言的文本嵌入表示，且除 GlueNet 外，无需更新任何额外参数。

[5]https://huggingface.co/openai/clip-vit-large-patch14

表 2. 与 AudioClip 模型在声音到图像生成方面的比较。此处报告了 Urban-Sound8K 基准上生成结果的 CLIP↑ 得分。我们评估了三种不同设置，包括 1）单一声音、2）声音混合和 3）声音-文本混合。

	声音	混合声音	声音-文本
AudioCLIP+SDM	16.07	13.25	17.22
GlueNet+SDM	**20.85**	**20.50**	**22.02**

表3. 与翻译模型的比较，即，M2M100- 418M [15], 用于多语言生成。此处报告了多语言 CLIP↑ 在跨模态 [59] 基准上的得分。GlueN- etR 表示第4.3节所述的重新加权目标。

	M2M100 [15]	GlueNet	GlueNetR
中文	**24.50**	22.01	23.17
法语	**25.08**	23.09	23.91
西班牙语	23.83	22.91	**24.18**
日语	23.73	23.99	**24.23**
意大利语	21.08	22.40	**22.88**
平均值	23.64	22.88	**23.67**

5.3.1 比较

图 5展示了使用多种语言输入提示词“印象派画家创作的午后花园油画”生成的示例结果。大多数结果都呈现了“午后花园”的内容。“印象派画家”是一个颇具挑战性的概念，在我们的训练数据中鲜少出现。因此，更多高质量的平行数据将大有裨益。与在多语言图像文本对上微调的模型相比，我们的方案效率高得多，成本也更低。在定量评估中，我们将其与一种热门的多语言翻译模型 M2M100-418M [15] (4.18 亿个参数)进行了比较，该模型使用 75 亿个平行句子训练。相比之下，我们的 GlueNet 使用的训练数据少得多（每对语言 200 万个句子）。结果列于表 3中，我们报告了在从 Crossmodal 数据集中随机选取的 100 个多语言图像-描述对上计算得出的多语言 CLIP分数 [59]。经过重新加权的版本以显著更低的训练成本，取得了与基于翻译的模型相当的结果。

5.3.2 混合多语言生成

混合多语言生成更具挑战性，原始 SDM 和第三方翻译工具箱都无法应对这一问题。我们提出的 GlueNet 在解决这一问题上展现出令人惊讶的能力。根据图 7 中的结果，即使提示采用三种不同的语言且随机混合，与 GlueNet 结合的多语言骨干网络也能为 SDM 解码器提供精确的引导。



图 7. 分辨率为 512 × 512时的混合多语言生成。输入描述中包含三种不同语言的文本，即中文、日语和英文。(a) 的描述为“彩色的、一只由毕加索绘制的猫坐在桌子上吃东西”，(b) 的描述为“一辆白色轿车撞向一栋建筑物”。加入我们预先注入的 GlueNet 后，XLM-Roberta 能够引导 SDM 解码器生成合理的结果，而原始 SDM 则无法正常工作。



引导权重为 7.5、DDIM 步数为 200 时，分辨率 256 × 256 下的单语生成。各方法的详情请参阅第5.1节。

5.4. 声音到图像

除了文本信号之外，所提出的 GlueNet 还通过将 CLIP 文本编码器与 AudioCLIP [20] 音频编码器对齐，实现了声音到图像生成。由于模态差距以及词元长度不匹配（CLIP 为 77，AudioCLIP 为 1），这是一项颇具挑战性的任务。为应对这些挑战，我们采用了重加权目标，更加关注排名靠前且信息丰富的信号，如第 4.3 节所述。GlueNet 的输入是配对数据$< sound, label >$分别由 AudioCLIP音频编码器（ESResNeXt [19]）和 CLIP 文本编码器编码。随后，GlueNet 根据第 4 节所述的目标进行训练，使声音嵌入与文本嵌入对齐。为评估 GlueNet 的有效性，我们在 Urbansound8k 测试集上进行了定量比较和视觉比较。我们的对比基线是部署 AudioCLIP音频和文本模型，在输入音频的情况下检索标签。随后将检索到的标签输入 Stable Diffusion 以生成结果。

表2 中的 CLIP分数有力地证明了我们的 GlueNet 优于基线方法，尤其是在混合声音方面。与单源声音相比，混合声音更具挑战性，有时甚至难以让人分辨。我们还将跨模态信号融合作为本文的另一项贡献。表2 和图10中对声音-文本融合的分析表明，我们提出的 GlueNet 取得了最佳结果。

5.5. 对齐分析

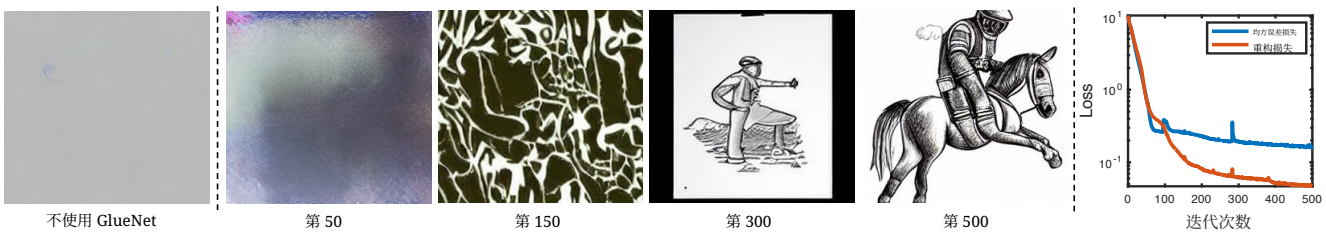
为了理解 GlueNet 如何映射跨模型特征，我们在图中展示了其轨迹以及不同迭代次数下的生成结果。9起初，旧图像解码器完全无法理解新的文本嵌入。随后，经过 300 次训练迭代并且损失显著下降后，“铅笔画”等特定模式开始出现。最终，当 $\mathcal{L}_{mse}$ 和 $\mathcal{L}_{rec}$ 都收敛到较小值时，这种差距得以弥合。

5.6. 与参数高效微调方法的比较

GlueNet 属于适配器系列，但其设计用于离线对齐。ControlNet [75] 或 T2I-adapter[36] 等热门方法需要新的条件图像对来训练适配器模型。

我们将低秩适配与 [25]，一种最具代表性的参数高效微调（PEFT）方法进行比较。我们测试了低秩适配+T5 和 GlueNet+T5 在潜在扩散模型（LDM）上进行图像生成（遵循表格中的设置：1， +微调（表示微调耗时 10 个 GPU 日）。我们训练 Lora+T5，使其与 LDM-bert 对齐，然后使用 LDM-Unet 对 Lora+T5 进行微调。显然，在量化指标上，我们的 GlueNet 优于低秩适配和 Lora+微调。

方法	FID ↓
Lora	16.72
GlueNet	**14.32**
低秩适配+微调	14.80
GlueNet+微调	**13.19**



(a) (b) (c)

图9. 对齐分析。(a)：直接替换新文本编码器的初始结果，即 LDM-Bert → T5-3B。(b)：GlueNet 使用新文本骨干网络在不同迭代次数下的结果。提示词为“一个宇航员骑着马的铅笔画”。(c)：训练过程中$\mathcal{L}_{mse}$ 和 $\mathcal{L}_{rec}$ 的曲线。



(a) SingleSound (b) MixedSounds

图11. 声音到图像生成中，以 (a) 单源声音和 (b) 混合声源为输入时，最佳结果的用户投票百分比。

5.7. 基于Transformer的GlueNet

我们参照表2，在多语言设置下实现了基于 Transformer 的 GlueNet（GlueNet-Trans）。我们的结果表明，GlueNet-MLP 的表现略优于 GlueNet-Trans。我们打算在这一背景下进一步研究自注意力模型。

语言	GNMLP	GNTrans
法语	**23.91**	23.87
西班牙语	24.18	**24.55**
意大利语	**22.88**	21.72
Avg	**23.66**	23.38

5.8. 点云到图像

为拓展本研究的范围，我们纳入了3D到图像生成。我们采用基于点云的预训练模型即，ULIP [69]，作为3D条件编码器，并使用我们提出的GlueNet将其与CLIPText对齐。

因此，通过替换 ULIP-点云编码器，我们实现了基于点云的图像生成。我们还按照以下设置，将其与基于检索的基线方法进行了比较：

方法 CLIP ↑
ULIP 17.29
GlueNet **21.67**

表。2 （其中声音被替换为点云对象），结果表明我们的方法取得了显著提升。

6. 结论

将预训练的条件编码器融入现有的文本到图像生成器，是构建更强大人工智能系统的一条令人振奋的途径。然而，由于两者紧密匹配，当前的编码器难以轻松升级。本文尝试打破图文模型之间这一强约束，以实现灵活的模块化和高效升级。为解决严重的对齐偏差，我们提出了 GlueNet，同时兼顾跨模型对齐和原创性保护这两个目标。实验证明，该方法有利于提升整体性能，并能在有限预算内为 X到图像生成提供多样化功能。我们希望这项工作能为社区设计大规模人工智能系统带来启发。

参考文献

[1] Eirikur Agustsson、Fabian Mentzer、Michael Tschannen、Lukas Cavigelli、Radu Timofte、Luca Benini 和 Luc Van Gool。用于图像和神经网络端到端学习压缩的软到硬向量量化。*arXiv* 预印本 *arXiv:1704.00648*，3，2017年。[2] Namhyuk Ahn、Byungkon Kang 和 Kyung-Ah Sohn。采用级联残差网络实现快速、准确且轻量级的超分辨率。载于欧洲计算机视觉会议，2018年。

[3] Yogesh Balaji、Seungjun Nah、Xun Huang、Arash Vahdat、Jiaming Song、Karsten Kreis、Miika Aittala、Timo Aila、Samuli Laine、Bryan Catanzaro 等。ediffi：采用专家去噪器集成的文本到图像扩散模型。*arXiv* 预印本 *arXiv:2211.01324*，2022年。[4] Tim Brooks、Aleksander Holynski 和 Alexei A Efros。In- structpix2pix：学习遵循图像编辑指令。*arXiv* 预印本 *arXiv:2211.09800*，2022年。[5] Michael Buhrmester、Tracy Kwang 和 Samuel D. Gosling。亚马逊土耳其机器人：一种新的低成本、高质量数据来源？《心理科学展望》，2011年。[6] Ziyi Chang、George A Koulieris 和 Hubert PH Shum。扩散模型的设计基础：综述。*arXiv* 预印本 *arXiv:2306.04542*，2023年。[7] Xinyang Chen、Sinan Wang、Mingsheng Long 和 Jianmin Wang。可迁移性与可判别性：用于对抗域适应的批次谱惩罚。载于国际机器学习会议，2019年。[8] Zhongzhi Chen、Guang Liu、Bo-Wen Zhang、Fulong Ye、Qinghong Yang 和 Ledell Wu。Altclip：改变 CLIP 中的语言编码器以扩展语言能力。*arXiv* 预印本 *arXiv:2211.06679*，2022年。[9]Alexis Conneau、Kartikay Khandelwal、Naman Goyal、Vishrav Chaudhary、Guillaume Wenzek、Francisco Guzm´an、Edouard Grave、Myle Ott、Luke Zettlemoyer 和 Veselin Stoyanov。大规模无监督跨语言表征学习。*arXiv* 预印本 *arXiv:1911.02116*，2019年。[10] Florinel-Alin Croitoru、Vlad Hondru、Radu Tudor Ionescu 和 Mubarak Shah。视觉领域中的扩散模型：综述。*arXiv* 预印本 *arXiv:2209.04747*，2022年。[11] Shuhao Cui、Shuhui Wang、Junbao Zhuo、Liang Li、Qing- ming Huang 和 Qi Tian。迈向可判别性与多样性：标签不足情形下的批次核范数最大化。载于计算机视觉与模式识别会议，2020年。[12] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。Bert：用于语言理解的深度双向变换器预训练。*arXiv* 预印本 *arXiv:1810.04805*，2018年。[13] Ming Ding、Zhuoyi Yang、Wenyi Hong、Wendi Zheng、Chang Zhou、Da Yin、Junyang Lin、Xu Zou、Zhou Shao、Hongxia Yang 等。Cogview：通过变换器掌握文本到图像生成。*NeurIPS*，2021年。[14] Benj Edwards。借助 Stable Diffusion，你或许再也无法相信网上所见，2022年9月。[15] Angela Fan、Shruti Bhosale、Holger Schwenk、Zhiyi Ma、Ahmed El-Kishky、Siddharth Goyal、Mandeep Baines、Onur Celebi、Guillaume Wenzek、Vishrav Chaudhary 等。超越以英语为中心的多语言机器翻译。《机器学习研究期刊》，2021年。[16] Oran Gafni、Adam Polyak、Oron Ashual、Shelly Sheynin、Devi Parikh 和 Yaniv Taigman。Make-A-Scene：采用人类先验的基于场景的文本到图像生成。*arXiv* 预印本 *arXiv:2203.13131*，2022年。[17] Yaroslav Ganin、Evgeniya Ustinova、Hana Ajakan、Pascal Germain、Hugo Larochelle、Franc¸ois Laviolette、Mario Marchand 和 Victor Lempitsky。神经网络的域对抗训练。《机器学习研究期刊》，2016年。

[18] Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。生成对抗网络。*NeurIPS*，2014年。[19] Andrey Guzhov、Federico Raue、J¨orn Hees 和 Andreas Dengel。Esresne (x) t-fbsp：学习稳健的音频时频变换。发表于*IJCNN*，2021年。[20] Andrey Guzhov、Federico Raue、J¨orn Hees 和 Andreas Dengel。AudioCLIP：将 CLIP 扩展至图像、文本和音频。发表于 *ICASSP*，2022年。[21] Dan Hendrycks 和 Kevin Gimpel。高斯误差线性单元（GELU）。*arXiv* 预印本 *arXiv:1606.08415*，2016年。[22]Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bernhard Nessler 和 Sepp Hochreiter。采用双时间尺度更新规则训练的生成对抗网络收敛至局部纳什均衡。*NeurIPS*，2017年。[23] Jonathan Ho、Ajay Jain 和 Pieter Abbeel。去噪扩散概率模型。*NeurIPS*，2020年。[24] Jonathan Ho 和 Tim Salimans。无分类器扩散引导。*arXiv* 预印本 *arXiv:2207.12598*，2022年。[25] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。LoRA：大语言模型的低秩适配。*arXiv* 预印本 *arXiv:2106.09685*，2021年。[26] Seung Hyun Lee、Wonseok Roh、Wonmin Byeon、Sang Ho Yoon、Chanyoung Kim、Jinkyu Kim 和 Sangpil Kim。声音引导的语义图像操作。发表于 计算机视觉与模式识别会议，2022年。[27] Junnan Li、Dongxu Li、Caiming Xiong 和 Steven Hoi。BLIP：通过自举式语言-图像预训练实现统一的视觉-语言理解与生成。*arXiv* 预印本 *arXiv:2201.12086*，2022年。[28] Bee Lim、Sanghyun Son、Heewon Kim、Seungjun Nah 和 Kyoung Mu Lee。用于单幅图像超分辨率的增强型深度残差网络。发表于*CVPRW*，2017年。[29] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Doll´ar 和 C Lawrence Zitnick。Microsoft COCO：上下文中的常见物体。发表于 欧洲计算机视觉会议，2014年。[30] Luping Liu、Yi Ren、Zhijie Lin 和 Zhou Zhao。流形上扩散模型的伪数值方法。*arXiv* 预印本 *arXiv:2202.09778*，2022年。[31] Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettle-moyer 和 Veselin Stoyanov。RoBERTa：一种稳健优化的 Bert 预训练方法。*arXiv* 预印本 *arXiv:1907 11692*，2019。。[32] Mingsheng Long、Zhangjie Cao、Jianmin Wang 和 Michael I Jordan。条件对抗域适配。*NeurIPS*，2018年。[33] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。*arXiv* 预印本 *arXiv:1711.05101*，2017年。[34] Xu Ma、Can Qin、Haoxuan You、Haoxi Ran 和 Yun Fu。重新思考点云中的网络设计与局部几何：一种简洁的残差多层感知机框架。发表于 *ICLR*，2021年。[35] Yi Ma、Doris Tsao 和 Heung-Yeung Shum。论简约性和自洽性原则在智能涌现中的作用。信息技术与电子工程前沿，2022年。

[36] Chong Mou、Xintao Wang、Liangbin Xie、Jian Zhang、Zhonggang Qi、Ying Shan 和 Xiaohu Qie。T2I-adapter：学习适配器以挖掘文本到图像扩散模型更强的可控能力。*arXiv* 预印本 *arXiv:2302.08453*，2023年。[37] Alex Nichol、Prafulla Dhariwal、Aditya Ramesh、Pranav Shyam、Pamela Mishkin、Bob McGrew、Ilya Sutskever 和 Mark Chen。GLIDE：借助文本引导扩散模型实现照片级真实感图像生成与编辑。*arXiv* 预印本 *arXiv:2112.10741*，2021年。[38] Guy Parsons。Dall-e 2 提示词手册。https://dallery.gallery/the-dalle-2-prompt-book/，2022年。[39] Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga、Alban Desmaison、Andreas Kopf、Edward Yang、Zachary DeVito、Martin Rai-son、Alykhan Tejani、Sasank Chilamkurthy、Benoit Steiner、Lu Fang、Junjie Bai 和 Soumith Chintala。PyTorch：一种命令式风格、高性能的深度学习库。发表于 *NeurIPS*。2019年。[40] Can Qin、Haoxuan You、Lichen Wang、C-C Jay Kuo 和 Yun Fu。PointDAN：一种用于点云表示的多尺度 3D 域适配网络。*NeurIPS*，2019年。[41]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。发表于国际机器学习会议，2021年。[42] Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li、Peter J Liu 等。探索通过统一的文本到文本转换器进行迁移学习的极限。《机器学习研究期刊》，2020年。[43] Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。利用 CLIP 潜变量进行分层文本条件图像生成。*arXiv* 预印本 *arXiv:2204 06125*，2022。。[44] Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。零样本文本到图像生成。发表于国际机器学习会议，2021年。[45] Scott Reed、Zeynep Akata、Xinchen Yan、Lajanugen Logeswaran、Bernt Schiele 和 Honglak Lee。生成对抗式文本到图像合成。发表于国际机器学习会议，2016年。[46]Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Ommer。使用潜在扩散模型进行高分辨率图像合成。发表于 计算机视觉与模式识别会议，2022年。[47] Olaf Ronneberger、Philipp Fischer 和 Thomas Brox。U-Net：用于生物医学图像分割的卷积网络。发表于 *MICCAI*，2015年。[48] Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily Denton、Seyed Kamyar Seyed Ghasemipour、Burcu Karagol Ayan、S Sara Mahdavi、Rapha Gontijo Lopes 等。具备深度语言理解能力的照片级真实感文本到图像扩散模型。*arXiv* 预印本 *arXiv:2205.11487*，2022年。[49] Justin Salamon、Christopher Jacoby 和 Juan Pablo Bello。用于城市声音研究的数据集与分类体系。发表于 *ACMMM*，2014年。[50] Christoph Schuhmann、Romain Beaumont、Richard Vencu、Cade Gordon、Ross Wightman、Mehdi Cherti、Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman 等。Laion-5b：用于训练新一代图文模型的开放式大规模数据集。*arXiv* 预印本 *arXiv:2210.08402*，2022年。[51] Christoph Schuhmann、Richard Vencu、Romain Beaumont、Robert Kaczmarczyk、Clayton Mullis、Aarush Katta、Theo Coombes、Jenia Jitsev 和 Aran Komatsuzaki。Laion-400m：由 CLIP 筛选的 4 亿图文对开放数据集。*arXiv* 预印本 *arXiv:2111.02114*，2021年。[52] Holger Schwenk、Vishrav Chaudhary、Shuo Sun、Hongyu Gong 和 Francisco Guzmán。WikiMatrix：从维基百科中挖掘 1620 种语言对的 1.35 亿个平行句。*arXiv* 预印本 *arXiv:1907.05791*，2019年。[53] Maximilian Seitzer。pytorch-fid：适用于 PyTorch 的 FID 分数。https://github.com/mseitzer/pytorch-fid，2020年8月。版本 0.2.1。[54] Abhishek Sinha、Jiaming Song、Chenlin Meng 和 Stefano Ermon。D2C：用于少样本条件生成的扩散解码模型。神经信息处理系统进展，34:12533–12548，2021年。[55] Jascha Sohl-Dickstein、Eric Weiss、Niru Maheswaranathan 和 Surya Ganguli。利用非平衡热力学进行深度无监督学习。发表于国际机器学习会议，2015年。[56] Yang Song 和 Stefano Ermon。通过估计数据分布的梯度进行生成建模。*NeurIPS*，32，2019年。[57] Yang Song、Jascha Sohl-Dickstein、Diederik P Kingma、Abhishek Kumar、Stefano Ermon 和 Ben Poole。通过随机微分方程进行基于分数的生成建模。*arXiv* 预印本 *arXiv:2011.13456*，2020年。[58] Ming Tao、Hao Tang、Fei Wu、Xiao-Yuan Jing、Bing-Kun Bao 和 Changsheng Xu。DF-GAN：一种简单有效的文本到图像合成基线。发表于计算机视觉与模式识别会议，2022年。[59] Ashish V Thapliyal、Jordi Pont-Tuset、Xi Chen 和 Radu Soricut。Crossmodal-3600：一个大规模多语言多模态评估数据集。*arXiv* 预印本 *arXiv:2205 12522*，2022。。[60] Tristan Thrush、Ryan Jiang、Max Bartolo、Amanpreet Singh、Adina Williams、Douwe Kiela 和 Candace Ross。Winoground：探究视觉与语言模型的视觉语言组合能力。发表于计算机视觉与模式识别会议，2022年。[61] Ilya O Tolstikhin、Neil Houlsby、Alexander Kolesnikov、Lucas Beyer、Xiaohua Zhai、Thomas Unterthiner、Jessica Yung、Andreas Steiner、Daniel Keysers、Jakob Uszkoreit 等。MLP-Mixer：一种用于视觉的全 MLP 架构。*NeurIPS*，2021年。[62] Arash Vahdat、Karsten Kreis 和 Jan Kautz。潜在空间中的基于分数的生成建模。*NeurIPS*，34:11287–11302，2021年。[63] Aaron Van Den Oord、Oriol Vinyals 等。神经离散表示学习。*NeurIPS*，2017年。[64] Laurens Van der Maaten 和 Geoffrey Hinton。使用 t-SNE 可视化数据。《机器学习研究期刊》，2008年。[65] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan Gomez 和 Laukasz Kaiser。注意力就是你所需要的一切。*NeurIPS*，2017年。[66] Chenfei Wu、Jian Liang、Xiaowei Hu、Zhe Gan、Jianfeng Wang、Lijuuan Wang、Zicheng Liu、Yuejian Fang 和 Nan

Duan。Nuwa-infinity：通过自回归之上的自回归生成实现无限视觉合成。*arXiv* 预印本 *arXiv:2207.09814*，2022年。[67] Chenfei Wu、Jian Liang、Lei Ji、Fan Yang、Yuejian Fang、Daxin Jiang 和 Nan Duan。NUWA：用于神经视觉世界创建的视觉合成预训练。发表于欧洲计算机视觉会议，2022年。[68] Tao Xu、Pengchuan Zhang、Qiuyuan Huang、Han Zhang、Zhe Gan、Xiaolei Huang 和 Xiaodong He。AttnGAN：利用注意力生成对抗网络进行细粒度文本到图像生成。发表于计算机视觉与模式识别会议，2018年。[69] Le Xue、Mingfei Gao、Chen Xing、Roberto Martín-Martín、Jiajun Wu、Caiming Xiong、Ran Xu、Juan Carlos Niebles 和 Silvio Savarese。ULIP：学习语言、图像和点云的统一表征，以实现三维理解。发表于计算机视觉与模式识别会议，2023年。[70] Yue Yang、Kaipeng Zhang、Yuying Ge、Wenqi Shao、Zeyue Xue、Yu Qiao 和 Ping Luo。Align, Adapt and Inject：声音引导的统一图像生成。*arXiv* 预印本 *arXiv:2306.11504*，2023年。[71] Guojun Yin、Bin Liu、Lu Sheng、Nenghai Yu、Xiaogang Wang 和 Jing Shao。文本到图像生成中的语义解耦。发表于计算机视觉与模式识别会议，2019年。[72] Jiahui Yu、Yuanzhong Xu、Jing Yu Koh、Thang Luong、Gunjan Baid、Zirui Wang、Vijay Vasudevan、Alexander Ku、Yinfei Yang、Burcu Karagol Ayan 等。扩展自回归模型，实现内容丰富的文本到图像生成。*arXiv* 预印本 *arXiv:2206.10789*，2022年。[73]Jiahui Yu、Yuanzhong Xu、Jing Yu Koh、Han Zhang、Ruoming Pang、James Qin、Alexander Ku、Yuanzhong Xu、Baldridge Jason 和 Yonghui Wu。采用改进 VQGAN 的向量量化图像建模。发表于*ICLR*，2022年。[74] Han Zhang、Tao Xu、Hongsheng Li、Shaoting Zhang、Xiaogang Wang、Xiaolei Huang 和 Dimitris Metaxas。StackGAN：利用堆叠式生成对抗网络将文本合成为照片级真实感图像。发表于*ICCV*，2017年。[75] Lvmin Zhang 和 Maneesh Agrawala。为文本到图像扩散模型添加条件控制。*arXiv* 预印本 *arXiv:2302.05543*，2023年。[76] Yufan Zhou、Ruiyi Zhang、Changyou Chen、Chunyuan Li、Chris Tensmeyer、Tong Yu、Jiuxiang Gu、Jinhui Xu 和 Tong Sun。LAFITE：探索文本到图像生成的无语言训练。*arXiv* 预印本 *arXiv:2111.13792*，2021年。[77] Minfeng Zhu、Pingbo Pan、Wei Chen 和 Yi Yang。DM-GAN：用于文本到图像合成的动态记忆生成对抗网络。发表于计算机视觉与模式识别会议，2019年。

附录

A. GlueNet 详情

本文的主要贡献是提出 GlueNet，它同时解决了跨模型对齐和特征保护问题。本节将进一步介绍 GlueNet 的细节，包括其实现与配置、规模分析以及其他消融研究。

A.1. 架构与配置

在主论文的第4.2节中，我们介绍了 GlueNet 的基本架构。GlueNet 有不同的规模可供选择，其配置详见表4。具有三个残差模块的 GlueNet 是较小的变体，其编码器包含 34M 个参数，而 GlueNet-5RMs 包含 51M 个参数。然而，较大的模型往往速度较慢，计算成本也较高。此外，针对不同编码器进行对齐时，架构可能略有差异。为了在微调阶段保持推理速度和效率，GlueNet 作为注入模块时，其规模不应超过图像解码器和文本编码器。根据我们的经验观察，将 RM 数量设为五可获得令人满意的结果。另一方面，GlueNet-3RMs 在表征学习方面表现较弱。除了 MLP-Mixer，也可以使用基于自注意力的模型来实现 GlueNet。未来，我们计划探索更适合 GlueNet 的架构。

表4。我们的 GlueNet 模型配置。我们介绍了 GlueNet-3RMs 和 GlueNet-5RMs 两种配置。 [LN] 表示尾部网络中的层归一化是可选的。对于 Stable Diffusion v1，输出维度和令牌输出分别为 77 和 1024。输入维度和输入令牌数取决于要替换的目标编码器。若输入令牌数等于令牌输出数，则 N 设为 1；若两者不相等，则 N 设为大于 1 的值（2 或 3）。

阶段	维度	模块	GlueNet-3RMs	GlueNet-5RMs
头部网络	输入令牌 → 令牌输出	词元 MLP	[线性层、层归一化 σ 线性层, LN, σ 线性层、层归一化] $\times N$	[线性层，层归一化 σ 线性层, LN, σ 线性层，层归一化] $\times N$
	输入维度 →输出维度	序列 MLP	[线性层，层归一化 σ 线性层，层归一化 σ 线性层，层归一化] $\times N$	[线性层，层归一化 σ 线性层，层归一化 σ 线性层，层归一化] $\times N$
主体网络	令牌输出 → 令牌输出	词元 MLP	[线性层，层归一化 σ 线性层, LN, σ 线性层，层归一化] $\times 3$	[线性层，层归一化 σ 线性层, LN, σ 线性层，层归一化] $\times 5$
	输出维度 →输出维度	序列 MLP	[线性层，层归一化 σ 线性层, LN, σ 线性层，层归一化] $\times 3$	[线性层，层归一化 σ 线性层, LN, σ 线性层，层归一化] $\times 5$
尾部网络	令牌输出 → 令牌输出	词元 MLP	[线性层, [层归一化], σ 线性层, [LN], σ 线性层, [层归一化]	[线性层, [层归一化], σ 线性层, [LN], σ 线性层, [层归一化]
	输出维度 →输出维度	序列 MLP	[线性层, [层归一化], σ 线性层, [LN], σ 线性层, [层归一化]	[线性层, [层归一化], σ 线性层, [LN], σ 线性层, [层归一化]

B. 文本编码器替换分析

B.1. GlueNet 尺寸分析

图 13展示了三种不同方法在示例提示词上的视觉比较。使用检查点 [6] 的原始 LDM 模型称为 LDM Ori。我们的模型 GlueNet-3RMs 和 GlueNet-5RMs，其主体网络分别包含三个或五个残差模块。文本编码器替换为 T5-3B，而图像解码器与 LDM Ori 相同。两个 GlueNet 模型均使用同一个文本语料库进行训练，其中包含 1800 万个英文句子。图中显示，GlueNet-3RMs 难以处理一些复杂提示，例如“布面油画，一只病毒怪物正在弹吉他”和“有一只长着狗头的企鹅站着”，而 GlueNet-5RMs 的表现更好。因此，我们可以得出结论：深层 GlueNet 对实现精准对齐是必要的。GlueNet-5RMs 通常是默认配置。

[6] https://github.com/CompVis/latent-diffusion

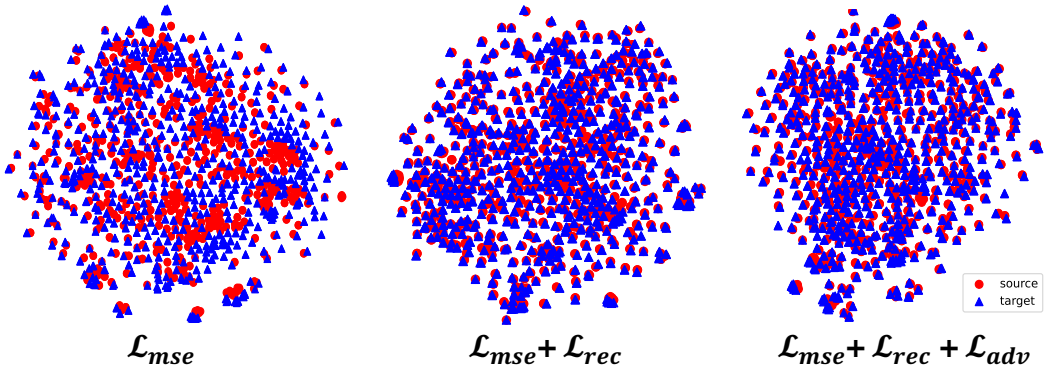


图 12. t-SNE [64] 跨模型特征图可视化。

表 5. GlueNet 消融研究　　ueNet (T5+GlueNet+LDMUnet) 在 5K COCO 子集上的结果。大多数情况下不进行微调　　s.

消融实验				s=1.5		s=5		s=7.5	
均方误差损失	重构损失	对抗损失	微调	CLIP↑	FID↓	CLIP↑	FID↓	CLIP↑	FID↓
✗	✗	✗	✗	10.70	98.17	10.75	141.19	11.00	156.70
✗	✗	✗	✓	18.98	35.76	22.05	49.14	22.49	53.11
✓	✗	✗	✗	20.40	34.41	23.01	48.67	23.40	52.78
✓	✓	✗	✗	20.53	33.14	23.23	45.96	23.57	48.67
✓	✓	✓	✗	20.67	32.80	23.24	45.48	23.74	48.51
✓	✓	✓	✓	**21.14**	**30.93**	**23.88**	**41.92**	**24.17**	**44.58**

B.2. 损失消融

为了分析不同损失的影响，我们在图 12中展示了它们的 t-SNE 映射可视化结果。从图中可以明显看出，仅使用均方误差损失不足以准确对齐两个模型的特征。重建损失对于保持判别能力并避免过拟合至关重要。相比之下，图中对抗损失似乎没有带来显著改善。通过实证研究，我们发现对抗损失带来的收益有限。

我们从 COCO 数据集中随机选取了 5,000 张图像的子集，进行定量消融研究。在对齐方面，我们使用 T5-3B 作为文本编码器，并通过 GlueNet 将其与 LDM UNet 对齐。UNet 的微调记为 FT（这里只需微调 UNet）。我们在下表中报告 FID 和 CLIP 分数，以比较图像质量以及图像与文本的对齐情况。测试数据通过 DDIM 进行推理，步数为 200，图像尺寸为 256 ×256。为进行全面分析，我们根据公式（ s=12）使用了三种无分类器引导强度：0.5、5 和 7.5。表 5 总结了消融研究的结果。

表格第一行展示了 T5 与 LDM 的直接组合，由于二者严重不对齐，结果毫无意义。第二行报告了从头开始训练的 T5+LDM 的结果。T5+LDM 的 UNet 微调耗费了 100 GPU 天，而 GlueNet 训练仅需 5 GPU 天。然而，即使成本高出十倍，其结果仍不如 GlueNet 网络。通过比较第三、第四、第五和第六行，我们可以轻易得出结论：完整版本的模型更具优势。

B.3. 文本编码器分析

我们提出的框架兼容多种文本编码器。在本小节中，我们尝试将几种语言模型用作文本编码器，包括 Bert-base（1.1 亿个参数）、Roberta-Large（3.55 亿个参数）、Clip-text（1.23 亿个参数）、T5-large（7.7 亿个参数）和 T5-3B（28 亿个参数）。我们使用提出的模型，将这些预训练文本编码器与潜在扩散模型（LDM）对齐，而无需对 LDM UNet 进行任何微调。我们在 COCO 的一个子集上评估了这些即插即用模型，该子集包含随机选取的 5,000 个样本。下表报告了 FID 和 CLIP分数，用于比较图像质量和图文对齐效果。GlueNet 网络是在从 Laion-400M 的图像说明中抽取的 1,800 万个句子上训练的。测试数据使用 DDIM 进行推理，步数为 200，图像尺寸设为 256×256。为了进行全面分析，我们针对

表6. 5K COCO 子集上不同文本编码器的分析。

	Bert-B（110M）		Roberta-L（355M）		CLIP-Text（123M）		T5-L（770M）		T5-3B（2.8B）	
	CLIP↑	FID↓	CLIP↑	FID↓	CLIP↑	FID↓	CLIP↑	FID↓	CLIP↑	FID↓
s=1.5	19.42	37.46	20.03	34.06	19.02	38.93	19.78	35.17	**20.67**	**32.80**
s=5	21.97	55.35	22.85	45.88	21.85	47.32	22.76	47.91	**23.24**	**45.48**
s=7.5	22.68	55.21	23.25	48.96	22.38	49.66	23.23	50.02	**23.74**	**48.51**

表 7. 在 5K COCO 子集上，采用不同词元长度（源词元长度 →目标词元长度）进行模型迁移，引导值为 5。Roberta-L [31] 作为新的文本编码器，替换潜在扩散模型文本编码器。

	77→77	128→77	256→77
CLIP↑	22.85	23.19	**23.37**
FID↓	45.88	45.67	**45.53**

表 8. 随着训练数据规模增加，GlueNet（T5-3B → 潜在扩散模型）的训练成本分析。FID 和 CLIP分数是在 COCO 的 5K 子集上计算的。

	数据规模		
	5M	18M	116M
CLIP↑	20.92	23.24	**23.71**
FID↓	48.63	45.48	**43.17**
GPU 天↓	**1.67**	5.89	41.20

三种无分类器引导设置进行了实验，设置值分别为 s=1.5、5 和 7.5。结果汇总于表6。如表所示，随着文本模型规模增大，FID 和 CLIP分数均有所提高，这与 Imagen [48]报告的发现一致。不过，由于与 Bert 等纯文本模型之间存在较大的领域差距，CLIP-Text 在我们的实验中表现不佳。

B.4. 可变长度词元

GlueNet 展现出处理不同长度文本的强大能力。LDM 和 SDM 的词元数量均固定为 77。我们提出的 GlueNet 无需对 Unet 模型进行任何微调，即可处理可变长度的文本编码器。为验证这一点，我们开展了一项实验研究，结果见表 7。在这项实验中，我们使用 Roberta-L 及其分词器对文本进行编码，词元上限分别为 77、128 和 256。引导设置为 5，其他实验设置与表 6 一致。

B.5. 数据规模

条件生成是一项具有挑战性的任务，文本编码器与图像 Unet 之间的对齐仍是一个未解问题。我们通过实证发现，替换文本编码器可以取得与源模型相当的结果，但要超越源模型则需要付出大量努力。瓶颈可能在于 Unet 架构。本节在表 8 中给出了不同训练集规模下的精确计算成本。基准测试采用 COCO-5K 数据集，并使用 T5-3B 和 GlueNet 替换 LDM 文本编码器，与此前的实验保持一致。如表所示，随着训练集规模和计算成本的增加，性能也随之提升。

C. 单语（英文）文本到图像生成

为展示我们提出的框架具有普适性，我们还将 Stable Diffusion（v1- 4）的 CLIP 文本编码器替换为 T5-Large。如图所示，14与标准 Stable Diffusion 模型相比，我们的模型展现出精准的可控性和出色的视觉质量。然而，由于存在细微的不匹配，其表现仍不及原始 Stable Diffusion 模型。



图 13. 使用引导权重 7.5 和 200 个 DDIM 步骤，在 256×256 中对示例提示词进行单语生成（T5 + GlueNet + LDMUnet）。T5 和 LDMUnet 均为预训练模型。我们仅训练不同模型规模的 GlueNet 来实现这一点。



图 14. 单语 生成示例提示词 512×512，引导权重为 7.5，采用 50 个 PLMS [30] 采样步 s.

D. 多语言文本到图像生成

图19、20、21、22 和23 中的补充结果有助于验证我们提出的框架具备多语言文本到图像生成能力。每种语言都对应一个专用的 GlueNet。为了构建自动化流程，模型应能够根据检测到的语言选择合适的 GlueNet。

E. 声音到图像生成

我们在图16 和17中提供了更多视觉结果，比较原始 GlueNet（不含重新加权目标）和自适应GluNet（含重新加权目标），具体如主论文第4.3 节所述。这些图清楚地表明，与原始 GlueNet 相比，自适应GluNet提高了生成图像的稳定性和准确性，从而证明了第4.3节提出的目标重新加权技术是有效的。

F. 多模态到图像生成

根据主论文第4.3 节，可以通过非参数操作融合来自不同编码器的多模态条件特征，例如拼接排名前 K 的信号，并对其余信号（不包括最后 K 个）取平均。图18展示了更多多模态生成结果。在我们的实验中，我们使用文本编码器（CLIPText）提取文本嵌入（输入为“毕加索绘画风格”）。我们还将音频数据输入 AudioCLIP 模型，并通过附加的 GlueNet 将其映射为嵌入表示。随后，我们应用所提出的特征融合算子，融合文本和声音嵌入，使稳定扩散模型能够生成合理的结果。

我们进行了消融研究，以确定使用正文第 4.3 节所述的非参数操作融合多模态条件特征时，键 K 的最优取值。图 15 展示的结果表明，当 $K <= 6$ 时，音频信号主导生成过程。随着 K 增大，文本信号逐渐开始显现，并最终在 $K = 10$ 时主导条件输入。因此，为了在两种跨模态信号之间取得良好平衡，我们建议根据经验观察选择适当的 K。

<声音：狗叫声>  ＋ <文本：“采用文森特·梵高的绘画风格”>



图15. 多模态（声音和文本）特征融合中选出的排名靠前的 K 信号分析。此操作是非参数的（也无需训练），只需拼接排名靠前的 K 词元信号，并对其余信号取平均值。

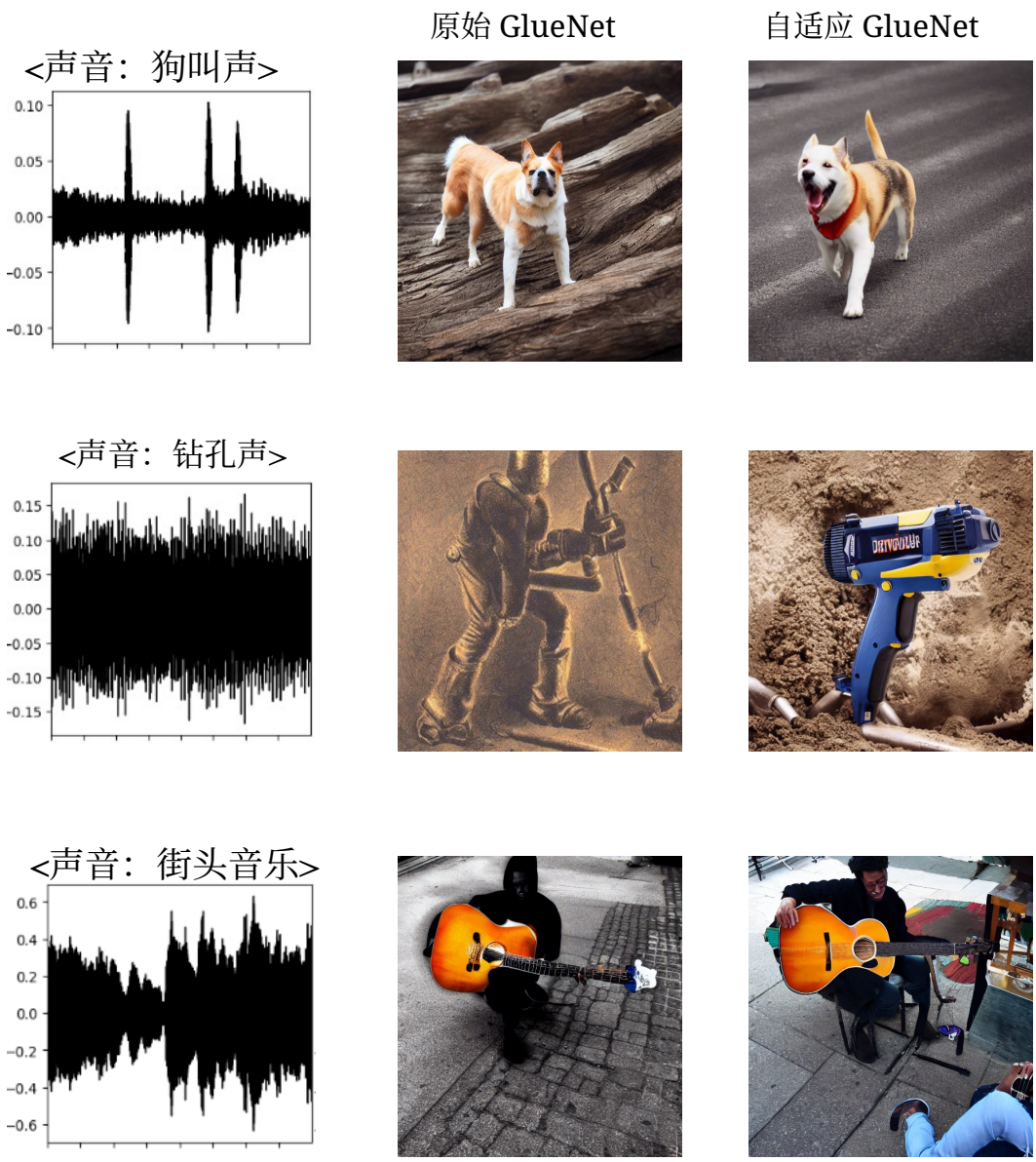


图 16. 使用原始 GlueNet 和自适应 GlueNet 进行声音到图像生成（第 1/2 部分）。

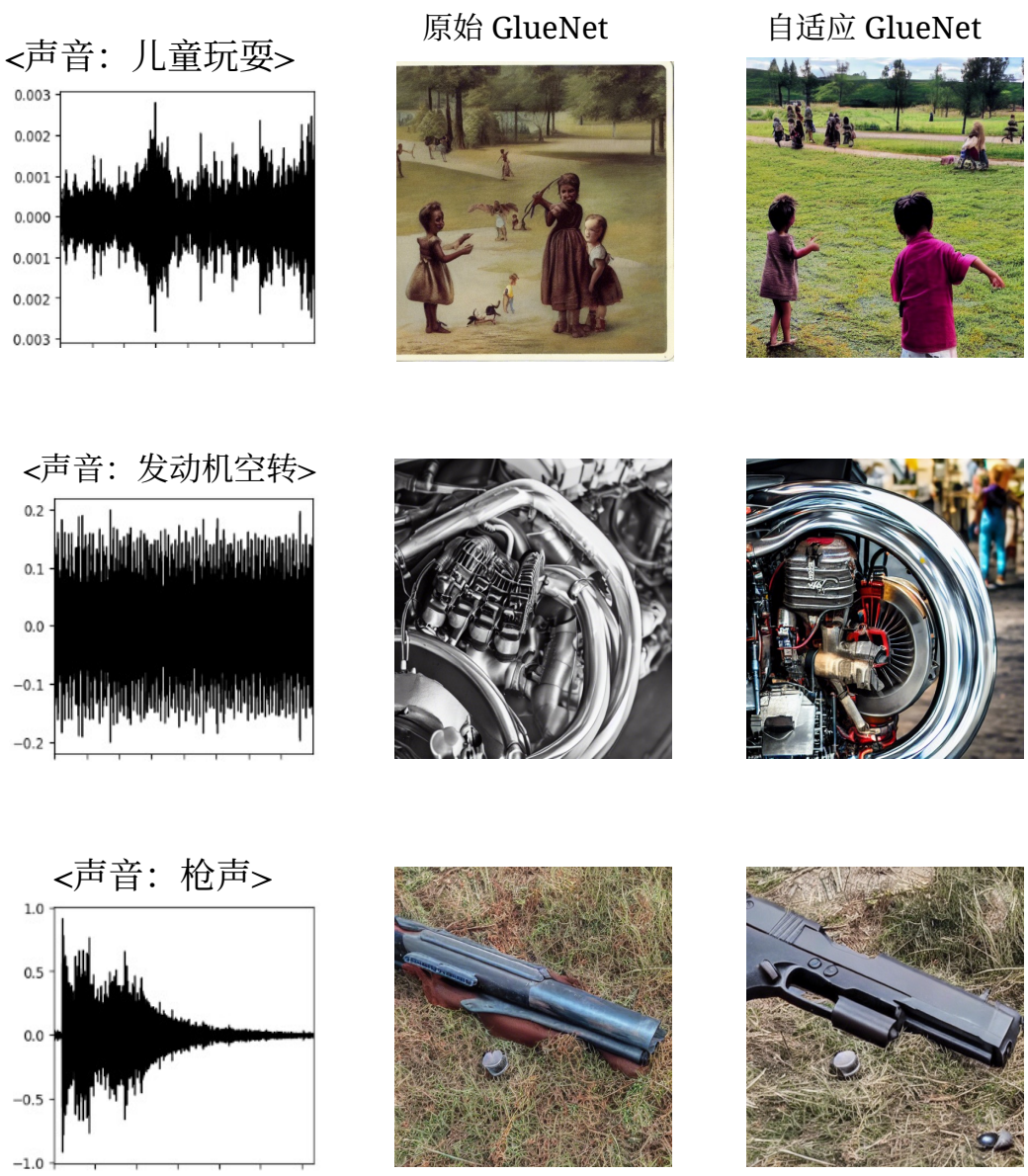


图 17。使用原始和自适应 GlueNet 进行声音到图像生成（第 2/2 部分）。

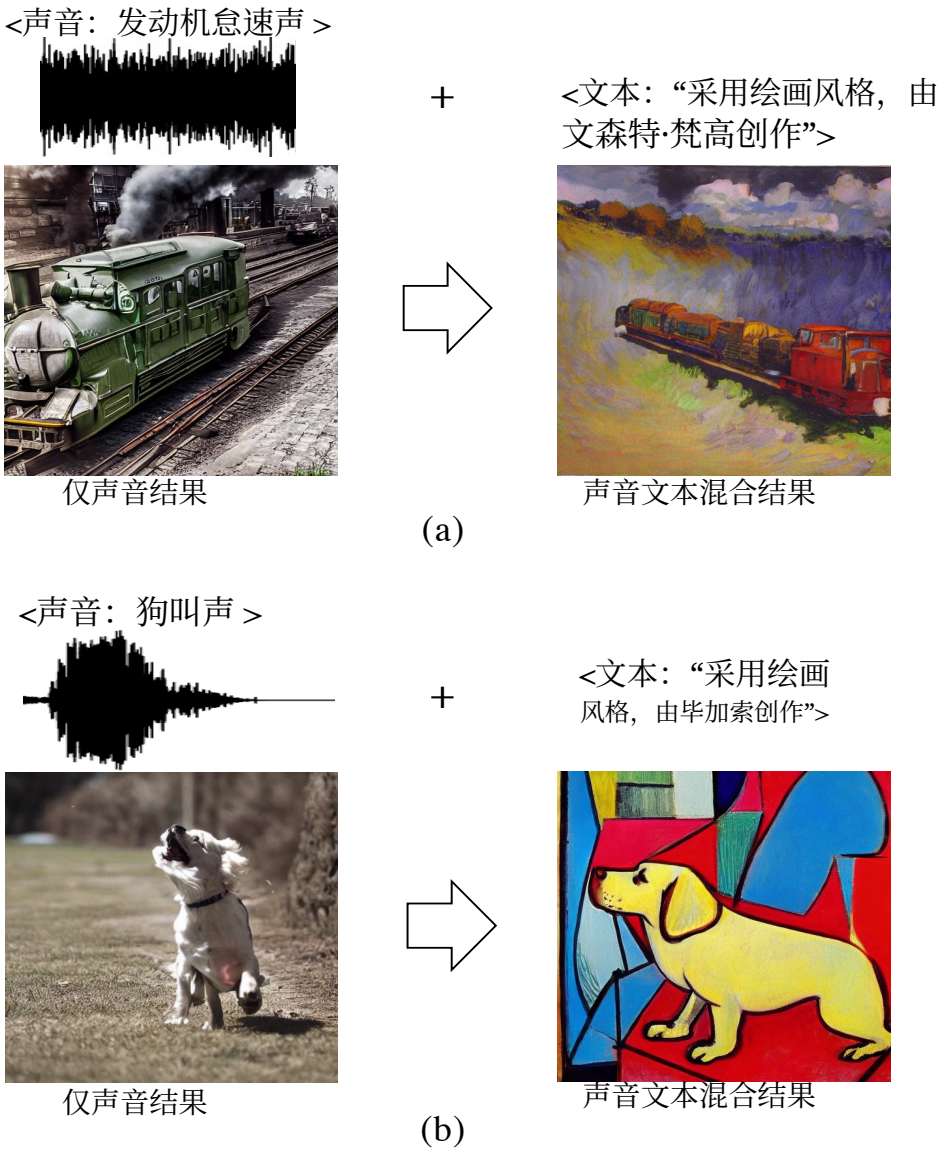


图18. 多模态到图像生成（声音文本混合）结果。条件编码器为 ClipText 和 AudioCLIP+GlueNet。两种模态特征融合后输入 Stable Diffusion Unet，以生成右侧所示结果。

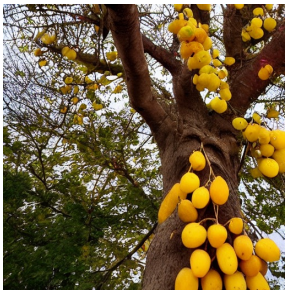
一些花
秋海棠在
一座花园里





一些水果
黄色的挂在
一棵树上





特写
一簇
粉色天竺葵





图 19. 使用法语图像说明的 512 × 512 XLM-Roberta + GlueNet + Stable Diffusion Unet (v1-5) 多语言生成结果。三个结果由不同的随机噪声生成。

黄色柠檬树的
黄色柠檬树
黄色





面条汤的
面条汤
放在桌子上的
木制





架子上摆着
音乐唱片
音乐
由……分隔的
流派分类





图 20. 使用西班牙语图像说明时， 512 × 512 中 XLM-Roberta + GlueNet + Stable Diffusion Unet (v1-5) 的多语言生成结果。这三种结果由不同的随机噪声生成。

一块木板上盛放着刚出炉的圆形披萨，上面撒了一些绿叶





两个古董意大利车





站在山区中两只鸡，一只黄色，另一只黑色黄色，它们俩站着看向同一个个方向





图 21. 使用中文标题的 XLM-Roberta + GlueNet + Stable Diffusion Unet (v1-5) 的多语言生成结果。三张图分别使用不同的随机噪声生成。

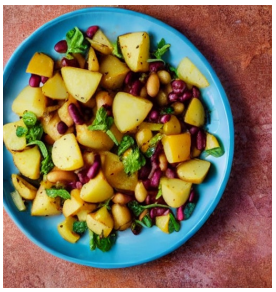
两台经典
胶片相
机的黑白
の写真



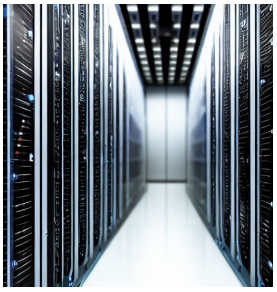


虾和蔬菜沙
拉以及土豆泥
薯条等
附上……的一
皿の料理





服务器路由器
器的服务器
设备、缠绕着
たコード



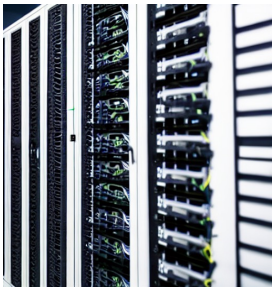
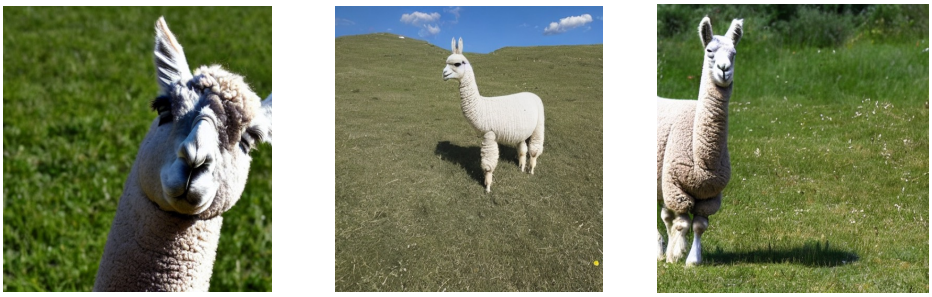


图22。使用日语图注，通过 XLM-Roberta + GlueNet + Stable Diffusion Unet (v1-5) 生成的多语言生成结果。三个结果由不同的随机噪声生成。

跑车
老式的
在展览中
白色墙壁的展厅
白色



一只美洲驼在一张
草地
下面的水桶
一片蓝天



汤配
肉、意大利面
和蔬菜一起上桌
在白盘中



图23。使用意大利语描述，通过 512 × 512 XLM-Roberta + GlueNet + Stable Diffusion Unet (v1-5) 得到的多语言生成结果。三张图分别由不同的随机噪声生成。