arXiv:2305.03610v2 [cs.CV] 2 Feb 2024

数据整理在图像描述中的作用

Wenyan Li[1], **Jonas F.Lotz**[1,2],**Chen Qiu**[3],**Desmond Elliott**[1]1哥本哈根大学计算机科学系2ROCKWOOL基金会研究部3武汉科技大学计算机科学与技术学院
{weli,jonasf.lotz,de}@di.ku.dk, chen@wust.edu.cn

摘要

图像描述模型通常在训练时将所有样本一视同仁，忽略了数据点之间的不匹配或其他难以处理的情况。相比之下，近期研究表明，采用课程学习策略对数据进行调度来训练模型是有效的。本文通过主动筛选数据集中的困难样本，为这一研究方向作出了贡献，而不增加样本总数量。我们探究了在训练过程中使用三种数据筛选方法的效果：完全移除样本、描述替换，或通过文本到图像生成模型进行图像替换。在 Flickr30K 和 COCO 数据集上，使用 BLIP 和 BEiT-3 模型开展的实验表明，这些筛选方法确实能够提升图像描述模型的性能，凸显了它们的有效性。

1 引言

图像描述是一项为视觉数据生成语法正确且准确的描述的任务，这需要理解显著物体的身份及其相互关系（Bernardi 等人，2016；Baltrušaitis 等人，2018）。尽管现有模型在这一问题上取得了显著进展，但仍面临一个固有挑战：如何应对不同图像-描述对所带来的学习难度差异（Sharma 等人，2018；Schuhmann 等人，2021）。

图像描述模型通常采用对整个训练数据集一视同仁的方式进行训练，这忽视了各个数据点在复杂程度上的差异。为解决这一问题，一种尝试是在大规模数据集的预处理舞台应用数据过滤，从预训练过程中移除噪声数据（Li 等人，2022a；Nguyen 等人，2023）。其他一些图像描述技术则采用了课程学习策略（Bengio 等人，2009），按照复杂程度逐步递增的顺序安排训练数据，从而有效地使学习过程适应任务的难度（Liu 等人，2021；Dong 等人，2021；Zhang 等人，2022；Alsharid 等人，2021；Ayyubi 等人，2023）。本文旨在回答一个根本性问题：图像描述模型能否不仅通过识别数据中的变化，还通过主动筛选困难样本来得到改进?

我们提出了三种数据整理方法，旨在改善学习过程，同时保持训练数据集的整体规模。这些方法包括完全移除样本、替换图注，或使用文本生成图像模型替换图像。这些方法针对的是在当前模型参数下，相较于训练数据集中其他样本具有异常高损失的图像-描述训练样本。换言之，我们的方法聚焦于那些难以建模的样本（Bengio 等人，2009；Kumar 等人，2010）。

本文的主要发现如下：

- 动态数据筛选可提升图像描述生成的性能。最佳策略因数据集而异，但可推广至不同的视觉语言模型。[1]

- 筛选程度是一个关键因素，且取决于数据集。我们发现，筛选超过 50% 的数据会降低数据筛选的有效性。

- 基于图像生成的数据筛选结合特定技术具有潜在优势，但其潜在收益受到生成错误的限制。人工研究发现了这些错误，而自动评估指标（例如 CLIPScore（Hessel 等人，2021）无法识别这些错误。

[1]我们已在此发布数据筛选框架的代码：https://github.com/lyan62/data-curation/

2 相关工作

自然语言处理中的数据整理 虽然图像描述领域尚未得到充分探索，Rogers(2021) 强调了数据整理对于深度学习和自然语言处理的重要性。多项研究已将数据整理应用于大语言模型：Chen 等人(2023) 基于大语言模型开发了一种通用文本筛选框架； Kandpal 等人 (2022) 和 Lee 等人 (2022) 探讨了去重对训练的影响；Chang 和 Jia (2023) 表明，仅凭谨慎筛选就能稳定上下文学习。

图像描述与学习策略课程学习（Bengio 等人，2009）和自适应学习（Kumar 等人，2010）是根据学习样本的差异调整学习过程的技术，通过损失值估计模型的能力。对于图像描述，已有多项研究引入了多种学习技术，根据样本难度定制模型训练过程，并兼顾文本和视觉特征（Alsharid 等人，2021；Dong 等人，2021；Zhang 等人，2022）。这些方法通过对数据排序来调整模型训练，而我们的方法提出了一种创新视角：通过筛选损失值异常的样本来调整训练，同时保持数据集的整体规模不变。

文本到图像生成模型 文本到图像生成模型（包括扩散模型（Song 等人，2021；Nichol 和 Dhariwal，2021））迅速流行起来，并展现出强大的能力。尽管近期的大规模潜在扩散模型擅长生成具有艺术感和照片级真实感的高分辨率图像（Rombach 等人，2022；Nichol 等人，2022；Ramesh 等人，2022；Saharia 等人，2022a），但它们在多模态任务中的应用仍未得到探索。与我们同期开展的研究中，Azizi 等人（2023）和 Jain 等人（2023）表明，通过从微调后的生成模型生成的增强图像中学习，可以提升图像分类器的性能；Xiao 等人（2023）和 Caffagni 等人（2023）则使用生成模型扩充了用于训练图像描述模型的数据集。

据我们所知，我们是首批探索动态数据筛选方法如何在不扩大现有数据集的情况下影响下游图像描述，以及如何在这一过程中应用文本到图像生成模型的研究者。

3 图注的数据整理

我们的主要目标是评估在训练期间主动筛选图像-描述对是否能够改进图像描述模型。难样本的产生有多种原因，包括图像与图注不匹配，或图像与图注之间存在不一致（Atliha 和 Šešok，2020），例如图注提及了图像中不可见的实体。为清楚起见，下文设 $\mathcal{D}$ 为一个包含 K 张图像的图像描述训练数据集，设 I_k 为第k张图像。每张图像配有 J 条图注；设 C_k^j 为图像 j的第 k条图注，因此，$(\mathrm{I}_k,\mathrm{C}_k^j)$ 即为一个图像-图注样本。

3.1 识别困难样本

受课程学习中的调度方法启发（Bengio 等人，2009；Kumar 等人，2010），我们假设在整个训练过程中都可以自动识别困难的训练样本。我们提出使用正在训练的图注模型 $\mathcal{M}$ 在数据集 $\mathcal{D}$ 上自动识别此类样本。我们可以随时使用该模型计算 $\mathcal{D}$ 中每个样本的损失，例如在每个训练轮次结束时 t： $\mathcal{L}_{\mathcal{M}}^t(\mathrm{I}_k, \mathrm{C}_k^j)\ \forall j, k$。可以根据样本各自的损失对其进行排名，从而找出可能受益于数据整理的候选样本。特别是，损失最高的样本是我们数据整理方法的目标。我们关注损失距离均值两个标准差的样本，或排名前 $X\%$ 的高损失样本。数据整理会动态更新训练数据集$\mathcal{D} \to \mathcal{D}_1 \to \cdots \to \mathcal{D}_T$。这样，训练数据集会在每个训练轮次结束时，根据模型在时间 t的当前图注能力进行动态更新。我们通过实证观察发现，如果不进行数据整理，高损失样本在五个训练轮次中始终保持高损失。[2]

3.2 筛选方法

我们研究了三种动态处理高损失图像图注对的方法：R移除，R替换C图注，以及 R替换I图像。图1概述了这些方法。

移除 数据整理最简单的方法是移除高损失样本，以防这些样本干扰模型。在移除方法中，

[2]图 5 最左侧的图展示了 Flickr30K 训练样本中损失的经验分布。

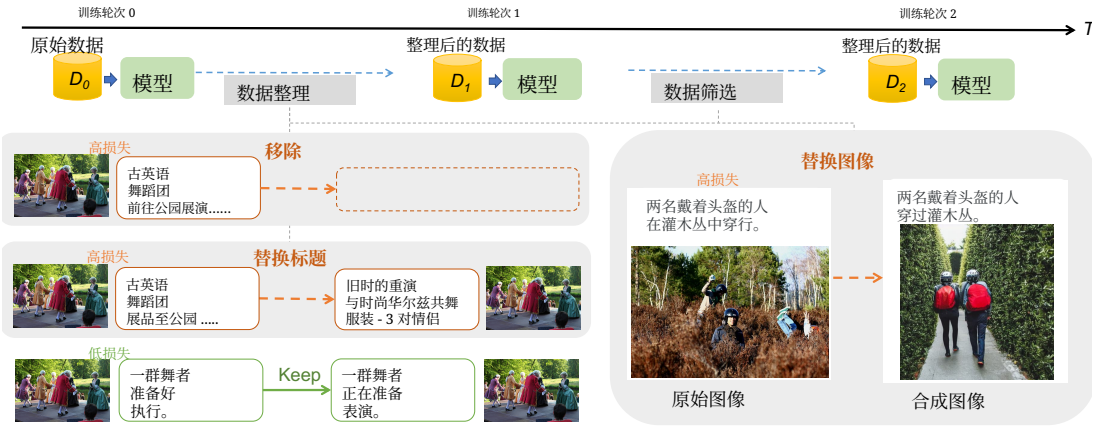


图 1：我们采用的数据整理方法概览。对于 REMOVE，我们会移除高损失图文对；对于 REPLACECAP，会将图像与原始数据集中的另一条图注配对；对于 REPLACEIMG，则将原始图像的图注用作文本到图像生成的提示词，以合成新的图文对。我们尝试了两种方案：仅替换图像，或为合成图像配上另一条相关图注。

高损失样本会从后续训练过程中彻底移除，从而减少图像–图注训练样本的总数量。

REPLACECAP 在 REPLACECAP 中，我们只需将图像–图注样本中的图注替换为原始数据集中另一条描述该图像的图注，从而有效地创建一个副本。采用这种方法后，用于训练模型的样本总数以及不同图像的总数均保持不变。这为我们的实验设置了一个对照条件。作为替代方案，我们还尝试将原始图注替换为语言模型生成的图注，我们在第6。

REPLACEIMG 在 REPLACEIMG 中，我们使用文本生成图像模型进行数据整理。这样做的好处是，在训练模型时使用的样本总数量不变，同时让模型接触到更多不同的图像。在一个快速的模型内循环步骤中，我们使用文本到图像生成模型，根据描述该图像的其他句子合成图像。我们将此方法整合到训练中的方式如下：给定训练数据中的图像 I_k 及其图注 $\{(I_k, C_k^1)$ ……，$(I_k, C_k^j)\}$，我们合成一张新图像 I_k，而不增加原始数据集的样本总数量。具体而言，对于图像 I_k，我们用合成的图像—文本对 $(\hat{} I_k, C_k^j)$ 替换一个原始的高损失样本 (I_k, C_k^j)。

给定一组描述图像的图注，可以通过几种方式提示图像生成模型（见附录中的图11 ）。我们尝试了三种方案：

- 单条图注：每条图注单独用于生成一幅新图像。
- Sentence-BERT 筛选：不同图注描述同一图像的方式多种多样。我们不必使用所有图注，而是可以从中选取一个有代表性的图注。这可以通过 Sentence-BERT（Reimers 和 Gurevych，2019）模型实现，该模型会找出与所有图注的平均嵌入最接近的图注。
- 拼接：所有五条图注拼接在一起，作为生成用的文本提示词。

对于上文提到的三种方法，我们可以在提示词后追加一个字符串，作为风格器，以强制生成的图像采用特定风格（+风格器）。此处使用的风格器是：“《国家地理》、高质量摄影、佳能 EOS R3、Flickr”。[3] 使用此技术生成的图像的一些代表性示例见附录中的图13 。

4 实验设置

4.1 数据与指标

我们在广泛使用的 MS COCO（Lin 等人，2014）和 Flickr30K（Young 等人，2014）数据集上进行微调，以评估我们的数据整理方法。我们使用以下指标报告结果：**BLEU**（Papineni 等人，2002）、**METEOR**（Denkowski 和 Lavie，2014）、**CIDEr**（Vedantam 等人，2015），以及**CLIPScore**（Hessel 等人，2021）。

[3]风格器是通过检查生成的图像来选择的，偏好摄影作品，而不选择素描和电脑艺术等“艺术性”输出。

	方法	比例	BLIP B	BLIP M	BLIP C	BLIP CS	BEiT-3 B	BEiT-3 M	BEiT-3 C	BEiT-3 CS
Flickr30K	基线	-	37.6	27.2	92.8	78.6	28.9	27.2	79.3	**80.4**
	+移除	2个标准差	38.6	**27.4**	**95.8**	**79.2**	31.4	27.1	**83.7**	80.0
	+替换标题	1%	37.9	**27.4**	94.5	78.9	29.6	**27.5**	80.1	80.3
	+替换图像	40%	**39.0**	27.3	95.7	79.1	**32.0**	26.9	82.4	79.1
COCO	基线	-	39.9	30.8	132.0	77.3	39.4	31.1	133.7	77.4
	+移除	1%	40.1	30.9	132.5	77.3	39.3	31.1	133.2	77.3
	+替换标题	1%	**40.2**	30.9	132.7	77.3	39.4	31.0	133.6	76.5
	+替换图像	10%	**40.2**	**31.0**	**133.1**	77.3	**39.6**	**31.1**	**134.4**	**77.5**

表1：在 Flickr30K 和 COCO 数据集上，使用我们的数据整理方法进行微调，与 BLIP 和 BEiT-3 的标准微调结果对比。我们报告**B**LEU、**M**eteor、**C**IDEr 和**C**LIP**S**core。最佳分数以**粗体**显示。

4.2 模型与实现

图像描述模型 我们使用两种最先进的预训练视觉语言模型——BLIP（Li 等人，2022a）和 BEiT-3（Wang 等人，2023）——研究数据整理的有效性。

我们注意到，BLIP 在预训练期间会进行图注生成与筛选（CapFilt）数据增强，其中两个组件都在 COCO 数据集上进行了微调。因此，我们在实验中分别使用预训练检查点 BLIP$_{CapFilt}$（用于 Flickr30K）和 BLIP$_{base}$（用于 COCO），以消除 CapFilt 流程的影响。我们使用总批次大小为 128，在 4×A100 GPU 上对 BLIP 进行 5 个训练轮次的微调。BEiT-3 基础模型则采用默认设置进行微调：总批次大小为 256，在 8×A100 GPU 上训练 10 个轮次。

整理比例我们使用 BLIP 模型，在各数据集的验证集上调整每种方法的数据筛选量。关于数据整理量与模型性能之间权衡的更多讨论，请参见第6 节。

ReplaceImg 文本到图像生成 在 ReplaceImg中进行文本到图像生成时，我们使用开源的稳定扩散模型（Rombach 等人，2022），该模型能够根据文本提示词生成图像。我们使用 MS COCO（Lin 等人，2014）数据集对稳定扩散 v1.5 模型进行微调，提示词由全部 5 条图注拼接而成，训练 15,000 步，学习率恒定为 $1e{-}5$，批量大小为 32。我们尝试了不同版本的已发布稳定扩散模型以及多种用于



图2：COCO 数据集中的定性示例。上方为 BLIP 模型生成的图注，下方为使用我们数据整理方法训练的相同模型。经过筛选后，许多错误（红色标示）都可以避免或修正（蓝色标示）。

生成高质量替换图像。[4]我们发现，使用经过微调的文本到图像模型可以提升图像描述性能。更多分析和消融实验请参见第7 节。

5 结果

数据整理能够提升图注质量 表 1展示了 BLIP 和 BEiT-3 模型在 Flickr30K 和 COCO 数据集上的结果。主要结论是，更好的模型性能

[4]也可以使用基于 API 的模型，但我们选择了稳定扩散，原因如下：(i) 稳定扩散可以利用开源代码直接集成到我们的训练流程中；(ii) 截至 2024 年 2 月 1 日，我们估计使用 DALL·E-2 在 Flickr30K 数据集上运行一次实验的成本为 4,176 美元。

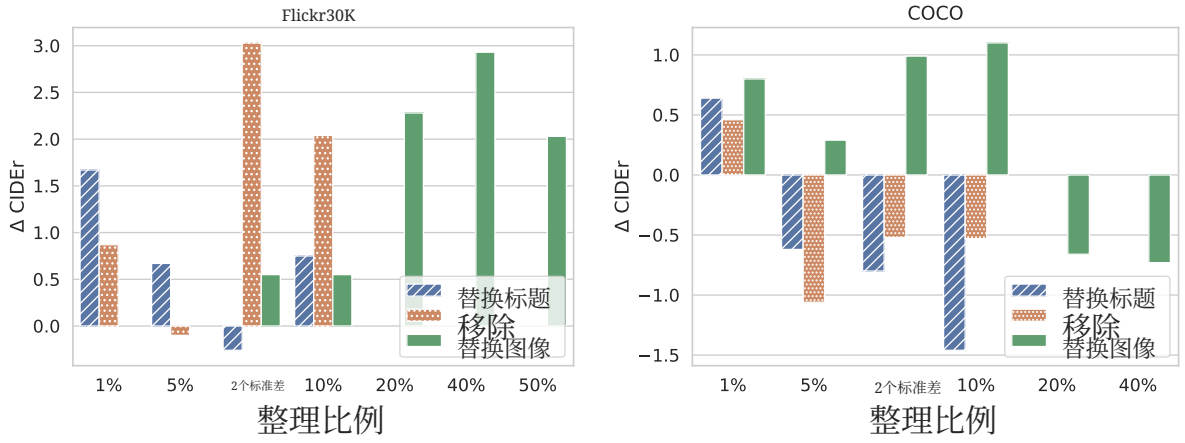


图3：不同数据筛选量的影响。我们观察到，与 COCO（10% 替换图像或 1% 替换标题）相比，Flickr30K 需要更多筛选（40% 替换图像或 2个标准差移除）。移除高损失训练样本对 Flickr30K 的收益更大，这表明原始数据集可能比 MS COCO 含有更多噪声。对于 2个标准差方法，每个训练轮次后筛选的样本数量并不固定，而是在 5% 到 10% 之间变化。

几乎总是可以通过数据整理来实现。

对于 Flickr30K，可以看出，REMOVE（2个标准差）和 REPLACEIMG （40%）的表现相近，CIDEr 分数均提升了 2.9–3 分。REPLACECAP 方法仅在针对高损失样本中排名前 1% 的样本时提升了 1.7 个 CIDEr 分数。对于 COCO，表现最佳的方法是 REPLACEIMG ，其筛选比例为 10%，相较于基线提升了 1.1 个 CIDEr 分数。REPLACECAP 和 REMOVE 在筛选高损失样本中排名前 1% 的样本时效果最佳，但提升幅度较小，仅为 0.5–0.7 个 CIDEr 分数。改进的定性示例见图2。

推广至不同的视觉语言模型 我们还通过在 BEiT-3 模型中实施这些数据整理方法，验证了它们能否推广至其他模型。具体而言，我们使用了与 BLIP 获得性能提升时完全相同的整理比例。如表 1 所示，在 Flickr30K 上，REMOVE 也是提升图注效果最高效的方法；而在 COCO 上，REPLACEIMG 的提升幅度最大。这表明，这些数据整理方法可以直接应用于其他先进的视觉语言模型，且整理比例具有可迁移性。我们注意到，由于 BEiT-3 在预训练中包含 COCO，REMOVE 和 REPLACE-CAP 方法并无益处。

6 讨论

筛选数量很重要筛选的数据量是一个重要的超参数。除

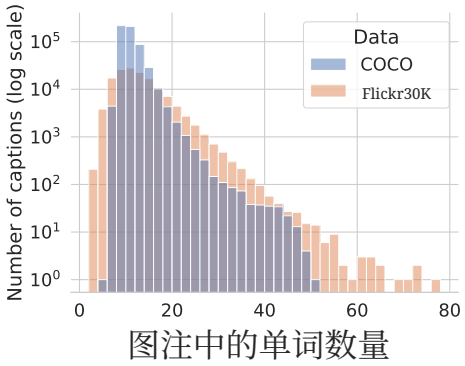


图4：图注长度分布。

了上文报告的最佳结果外，我们还展示了对数据整理量进行更细致调整所得的结果。对于移除和替换标题，我们探索了筛选高损失样本中排名前 1% 、 5% 和 10% 的样本。对于替换图像，我们探索了 10%–80% 的整理比例。除了固定的 X% 比例外，我们还对损失比均值高出两个标准差的样本进行干预。

该分析结果如图3所示。对于 Flickr30K，不同筛选方法的有效整理比例为 1%–50%；而对于 COCO，在损失最高的样本中，移除替换图像 EPLACEIMG 仅需处理不到 10% 的样本即可获益，并且移除EMOVE 和替换EPLACECAP 图注的有效整理比例在 1% 时便达到上限。这表明，Flickr30K 中的噪声样本可能比 MS COCO 数据集更多。与 MS COCO 相比，Flickr30K 包含更多带有较长图注的样本（图4），其中可能包含过于具体的细节，这些细节与其他图注不一致，也较难被模型学习（图12）。通过基于筛选的微调，可以有效识别、

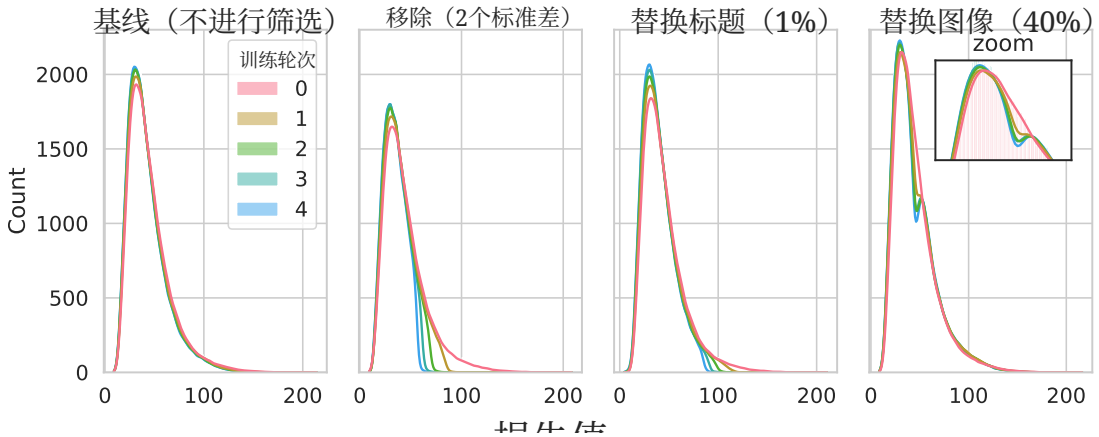


图 5：对于 Flickr30K，不同的筛选方法会改变训练样本在各训练轮次中的损失分布。相比之下，在未进行数据整理的情况下（最左侧的图），高损失样本在整个训练过程中始终保持高损失状态。

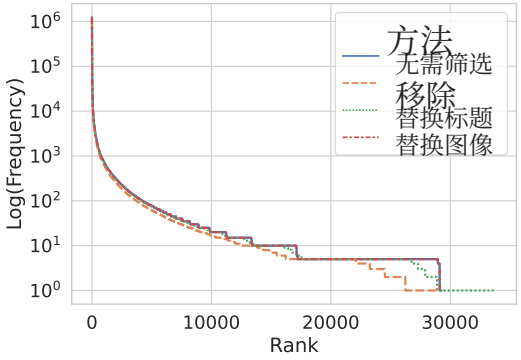


图 6：不同筛选方法下 Flickr30K 训练样本中词语的 Zipf 分布。请注意，REMOVE 对分布尾部所做的明显改变。

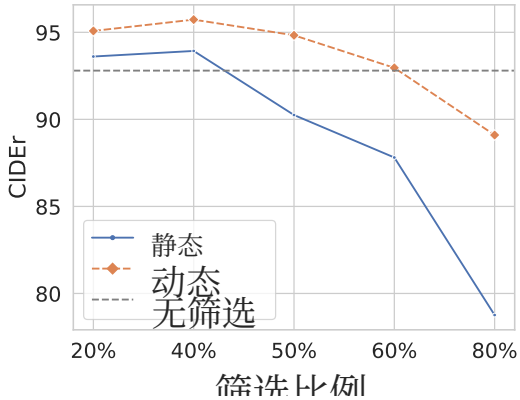


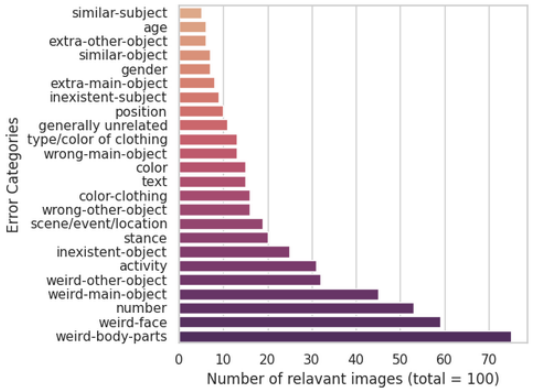
图 7：在 Flickr30K 数据集上使用 BLIP 对 RE-PLACEIMG 进行动态与静态替换的比较，随替换样本数量的变化。

移除或替换这些样本，这表明我们的方法在使用含噪数据集进行训练时十分高效。我们注意到，整理超过 50% 的数据不仅无法促进训练，反而会损害性能。

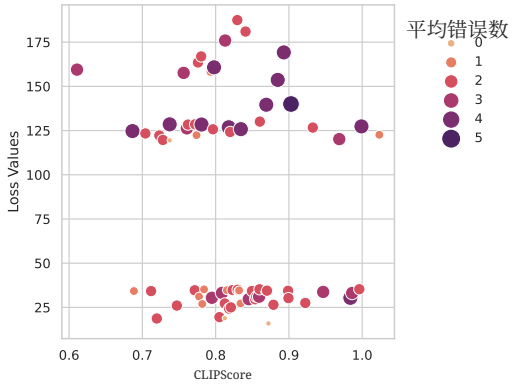
筛选改变训练分布 我们考察了每种筛选方法下训练样本的损失分布随训练轮次的变化，以了解其对训练过程的影响（图5）。这些损失是在每个训练轮次结束后，使用当前模型参数计算得出的；损失较高的样本会在后续筛选步骤中成为处理对象。对于 REMOVE 方法，训练期间会动态移除损失比均值高出两个标准差的训练样本，从而使损失分布的尾部逐渐缩短。REPLACEIMG 会逐渐降低损失，使损失形成由原始图文对和含有合成图像的图文对组成的高斯混合分布。除此之外

我们还考察了各种筛选方法下训练图注中的词语分布，而不只是训练样本的损失。图6展示了这些分布，可以看出，REMOVE 会在训练期间减少低频词和仅出现一次的词，而 REPLACECAP 则会增加一些较低频词的出现次数，同时移除仅出现一次的词。根据定义，REPLACEIMG 只会改变用于训练模型的图像分布，因此不会改变训练数据中的词语分布。

动态替换的有效性 我们将训练损失值作为有效指标，动态筛选被识别为具有挑战性的训练样本。在 REPLACEIMG 中，另一种静态方法是在训练前，将相同的图像，即 I_k中的 $\{(I_k, C_k^1), \ldots\ldots, (I_k, C_k^J)\}$，替换为独特的合成图像，而不是在训练过程中更新训练样本。采用静态图像替换时，对于每条参考图注，



(a) 文本到图像生成错误的分布。



(b) 人工评估与 CLIPScore 的比较。

图8：人工研究中，稳定扩散模型在100张图像中所犯错误的结果。研究中使用的图像选取自模型损失较低或较高的样本。(a) 各类别中标注错误数量的直方图。最常见的标注涉及预期物体或人物的奇怪变形。(b) 人工标注在每张合成图像中识别出的错误平均数量与其相对于原始图注的图注损失之间的关系。损失较高的图像中识别出的错误更多。然而，CLIPScore似乎无法验证合成图像的质量，因为错误较多的样本，其分数范围几乎完全相同。

我们都会将其原始图像替换为生成图像。整理比例为 20%–80% 的静态替换，对应于替换原始五条图注中一至四条所对应的图像。50% 的替换比例模拟公平的抛硬币：对于每个文本 - 图像样本，其图像有 50% 的概率被替换为合成图像。

我们在图 7 中比较了这两种方法的效果。在原始的 1k 验证集上进行评估时，我们发现，对于这两种方法，加入 20% 或 40% 的合成图像均有助于微调，并能获得更高的 CIDEr 分数。然而，动态图像替换始终优于静态方法，这表明聚焦于困难样本是有效的。对于这两种替换方法，当整理比例过高时，性能都会开始下降。这可能表明，当从合成分布中引入过多图像时，训练集与评估集之间的差距会进一步扩大。

用语言模型生成内容替换图注 作为 REPLACECAP 方法的替代方案，我们研究了用语言模型（LM）生成的图注替换原图注的效用。受Ramos 等人 (2023) 的方法启发，我们使用少样本示例提示 XGLM-2.9B 模型（Lin 等人，2022），使其生成新的图注。随后，将 LM 生成的图注与图像配对，作为筛选后的样本。我们在 Flickr30K 上使用两个模型进行评估，并采用相同的整理比例

	BLIP		BEiT3	
方法	B	C	B	C
基线	37.6	92.8	29.8	79.3
+替换标题	**37.9**	**94.5**	29.6	80.1
+替换LMCap	37.5	93.4	**31.2**	**83.2**

表 2：在 Flickr30K 上，将描述替换与 LM 生成同 REPLACECAP 进行比较。两种方法对 BLIP 和 BEiT-3 的表现均优于基线。

为 1%，作为 REPLACECAP。表2 中的结果表明，这种方法可以作为 REPLACECAP的可行替代方案，并且在两个模型上均持续优于基线。更多实现细节请参阅附录A.3。

人工研究：文本生成图像模型所犯的错误为了评估生成图像的质量及其与人类判断的一致性，我们开展了一项人工研究，以评估合成图像中存在的错误。这有助于更好地了解自动评估指标未能反映的 REPLACEIMG 筛选不足之处。

我们首先根据模型损失，对 COCO 验证集中的 1K 张图像生成的合成图像进行排名。然后，我们使用经过微调的图注模型，根据损失选取排名最高和最低的各 50 张图像，抽取其中一部分供人工标注。这些图像被均匀分为 5 组，

每组包含 20 张图像，其中高损失和低损失图像的数量相同。数据由 12 人标注，他们是某大学研究实验室的成员，对文本到图像生成有基本了解，但不了解图像的双峰分布。我们要求标注者根据合成图像以及用于生成图像的参考句子，对合成图像中的错误进行分类。每位参与者标注一组图像。

从 van Mil- tenburg 和 Elliott (2017) 所定义的类别出发，我们定义了 25 种错误类别，包括颜色、数量不匹配，以及与图像中的人物和物体相关的错误。用户界面及更多详细信息请参见附录 A.1。我们分析了至少有三条标注的图像的人类判断结果，共涉及 74 张不同的图像。

如图 8a所示，合成图像最常见的问题是经常生成奇怪的面部或身体部位，使图像显得不够自然或美观。文生图模型在生成正确数量的人物或物体方面也表现不佳。从图 8b中，我们确认了所收集标注的质量：高损失图像通常平均包含更多错误。此外，我们注意到，CLIPScore 对这类错误并不敏感，说明其评估生成图像质量的能力有限。更多示例请见附录中的图 13 。

7 进一步分析

人工研究揭示了文生图模型的失效模式后，我们现在进一步介绍各种已被证明有助于提升图像相关性的技术，以改进图像描述数据集的整理。

往返图像描述评估 以往大多数文本到图像生成研究使用面向图像的指标，如 FID（Heusel 等人，2017）或 CLIPScore（Hessel 等人，2021）。然而，这些指标并不适合我们的目的，因为有研究称它们与感知质量缺乏一致性（Saharia 等人，2022b）。我们还发现，CLIPScore 无法区分图像描述中的低损失样本和高损失样本（图 8）。

或者，类似于 Hong 等人(2018)，我们使用一个固定模型为合成图像生成图注，然后通过一个三步流程将其与原始图注进行比较（图 9）：(1)Gen-

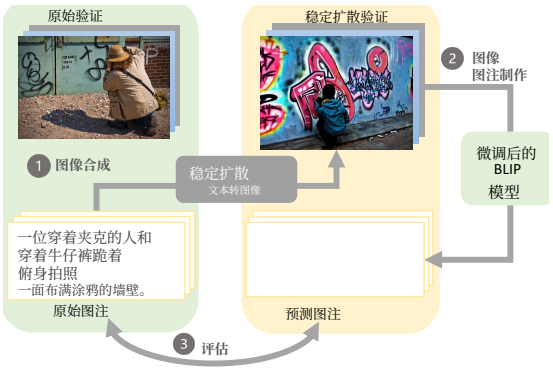


图 9：往返图像描述评估。

模型	FT	提示词	B	C	M
上界			37.6	27.2	57.1
SD 1.5	-	拼接	31.0	24.7	52.5
SD 1.5	-	+ 风格器	30.8	24.2	52.5
SD 1.5	F	+ 风格器	**33.5**	**25.0**	**53.5**
SD 1.5	F	SBERT + 风格器	30.6	24.1	52.0
稳定扩散 2.0	-	拼接 + 风格器	31.2	24.8	52.0

表3：使用不同的稳定扩散模型、提示词和微调设置，在 Flickr30K 上进行往返图注评估。F 表示模型经过微调。我们报告**B**LEU、**C**IDEr、**M**eteor。

从验证集图注生成图像；(2) 使用强大的图像图注生成模型为生成的图像预测图注；此处我们使用在 COCO 数据集上进行微调的 BLIP，但也可以使用其他任何强大的图注生成模型。(3) 将预测的图注与原始图注进行比较。其基本假设是，如果生成的图像与原始图像的质量相近，那么生成的图注也应当相似。

文本到图像生成变体的消融实验我们采用往返式图像描述进行评估，并对文本到图像生成模型的不同变体开展消融研究。表格3 汇总了在 Flickr30K 数据集上的评估结果。具体而言，我们测试了不同版本的稳定扩散模型；采用多种方法向扩散模型提供提示词（第3.2节）；并比较微调后的文生图模型与预训练模型的生成效果。结果表明，在添加风格器并以全部五条图注的拼接作为提示词时，经过 COCO 微调的稳定扩散 v1.5 优于其他变体。有关模型变体的详情，请参阅附录A.2。

8 结论

本文展示了一种简单而有效的数据筛选框架，能够提升图像描述模型的性能。我们研究了三种数据整理方法，这些方法会根据高损失的图像—图注样本动态更新训练数据集。具体方法包括移除样本、替换样本中的图注，或根据现有图注生成新图像。在 Flickr30K 和 MS COCO 数据集上的实验结果表明，这些数据整理方法能够发挥作用，同时不会增加训练数据集的总规模。对文生图模型合成图像的深入分析发现，模型在生成特定数量或颜色的物体时经常出错，也难以准确生成的人体特征。对这些图像中的错误进行人工评估后发现，高损失图像与低损失图像之间存在明显差异。

展望未来，我们预计，更好的文本生成图像模型将使利用合成图像训练图像描述模型取得进一步进展。从附录 A.4中的见解来看，将不同的筛选方法结合起来构建混合模型也大有可为。我们认为，采用多种方法的更复杂学习方案，在筛选数据集时将带来更大的灵活性。我们计划验证这些发现是否也适用于其他图像描述模型。此外，我们也有兴趣将同一框架应用于其他多模态任务，尤其是那些数据集不完整、因众包收集足够数据成本高昂而无法全面覆盖分布空间的任务，例如视觉问答或视觉接地对话。

局限性

正如Nguyen 等人(2023) 使用最先进的 BLIP-2 模型生成更优质的图注，成功提升了预训练数据集的质量，因此我们预计，我们的筛选策略也能够扩展并适配于视觉-语言预训练。然而，受研究资源限制，本文未对此进行探索。目前，我们的数据筛选方法也依赖于用于图像理解和文本到图像生成的最先进预训练模型。

在本研究中，我们探讨了采用不同筛选方法如何影响下游图像描述在不同整理比例下的表现。虽然本文预先设定了实验所用的整理比例，但如果筛选方法能够自动确定整理比例，其应用将会更加便捷。

此外，尽管我们采用在线方式进行数据整理，但当前方法的速度和性能受限于文本到图像生成模型的上限。这可能会成为将该策略应用于更复杂的视觉与语言任务的一大瓶颈。

伦理声明

文本到图像生成在更广泛的人工智能和伦理学界存在争议（Carlini 等人，2023）。例如，它可以根据性别或种族刻板印象生成图像，这可能会对相关群体成员造成伤害（Li 等人，2022b）。尽管在视觉语言领域尚未观察到这一现象，Shumailov 等人 （2023）提供证据表明，使用大型语言模型等生成模型产生的合成数据，可能带来数据质量退化的风险。

本文中，我们利用文本到图像生成来提升图像描述模型的质量，所依据的是一组特定的众包图注。这些图注本身可能包含有害的刻板印象，而这些刻板印象可能会在我们动态更新的训练数据集中变得更加普遍。我们根据损失值动态使用新图像更新模型时，会移除生成图像中的水印，以防止模型泄露信息。合成图像的使用将严格遵循社区准则。

在开发用于图像替换、涉及文本到图像生成的筛选方法时，我们采用了 stable-diffusion v1.5 模型（Rombach 等人，2022），该模型是在 LAION-5B 数据集上训练的。我们注意到，我们此前并未了解到有任何针对该数据集中非法内容的调查（Thiel，2023）。因此，我们强调，我们提出的框架也兼容在更可靠的数据集上训练的任何其他文生图模型。考虑到这一点，我们鼓励研究人员在未来的工作中采用筛选技术时，探索并应用其他替代性的文生图模型。

致谢

我们感谢 Jiaang Li、Lei Li 以及 CoAStal 和 LAMP 团队提供的反馈。Wenyan Li 获得了 Lundbeck基金会（Brain- Drugs资助：R279-2018-1145）的资助，并获得了丹麦创新基金在 AI4Xray 项目框架下的资助。Jonas F. Lotz 获得了theROCKWOOL基金会的资助（资助编号：1242）。本研究获得了 VILLUM基金会提供的一项研究资助（VIL53122）。

参考文献

Mohammad Alsharid、Rasheed El-Bouri、Harshita Sharma、Lior Drukker、Aris T. Papageorghiou 和 J. Alison Noble。2021 年。面向课程的图像描述双重课程。发表于 *2021 IEEE* 第 *18* 届国际生物医学成像研讨会（*ISBI*）。

Viktar Atliha 和 Dmitrij Šešok。2020 年。使用 BERT 进行图像描述的文本增强。应用科学。

Hammad Ayyubi、Rahul Lokesh、Alireza Zareian、Bo Wu 和 Shih-Fu Chang。2023。从儿童身上学习：通过课程学习改进图像-图注预训练。载于 计算语言学协会论文集：*ACL 2023*。

Shekoofeh Azizi、Simon Kornblith、Chitwan Saharia、Mohammad Norouzi 和 David J. Fleet。2023 年。来自扩散模型的合成数据提升了 ImageNet 分类性能。

Tadas Baltrušaitis、Chaitanya Ahuja 和 Louis-Philippe Morency。2018。多模态机器学习：综述与分类体系。*IEEE* 模式分析与机器智能汇刊，41(2)：423–443。

Yoshua Bengio、Jérôme Louradour、Ronan Collobert 和 Jason Weston。2009。课程学习。载于第 *26* 届年度国际机器学习会议论文集，第 41–48 页。

Raffaella Bernardi、Ruket Cakici、Desmond Elliott、Aykut Erdem、Erkut Erdem、Nazli Ikizler-Cinbis、Frank Keller、Adrian Muscat 和 Barbara Plank。2016。图像自动描述生成：模型、数据集和评估指标综述。人工智能研究期刊，55：409–442。

Davide Caffagni、Manuele Barraco、Marcella Cornia、Lorenzo Baraldi 和 Rita Cucchiara。2023 年。Synth-cap：通过合成数据增强 Transformer模型以实现图像描述。载于国际图像分析与处理会议，第 112–123 页。施普林格。

Nicholas Carlini、Jamie Hayes、Milad Nasr、Matthew Jagielski、Vikash Sehwag、Florian Tramèr、Borja Balle、Daphne Ippolito 和 Eric Wallace。2023 年。从扩散模型中提取训练数据。*arXiv* 预印本 *arXiv:2301.13188*。Ting-Yun Chang 和 Robin Jia。2023 年。仅靠数据整理即可稳定上下文学习。载于计算语言学协会第 *61* 届年会论文集（第一卷：长篇论文），第 8123–8144 页。Zui Chen、Lei Cao 和 Sam Madden。2023 年。Lingua manga：一个以大语言模型为中心的通用数据整理系统。*arXiv* 预印本 *arXiv:2306.11702*。

Michael Denkowski 和 Alon Lavie。2014。Meteor universal：适用于任意目标语言的特定语言翻译评估。载于第九届统计机器翻译研讨会论文集，第 376–380 页。

Xinzhi Dong、Chengjiang Long、Wenju Xu 和 Chunxia Xiao。2021。结合 Transformer 和课程学习的图像描述双图卷积网络。载于第*29*届 *ACM*国际多媒体会议论文集。计算机协会。

Jack Hessel、Ari Holtzman、Maxwell Forbes、Ronan Le Bras 和 Yejin Choi。2021。CLIPScore：一种用于图像描述的无参考评估指标。载于*EMNLP*。

Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bernhard Nessler 和 Sepp Hochreiter。2017 年。通过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。载于*NIPS*。

Seunghoon Hong、Dingdong Yang、Jongwook Choi 和 Honglak Lee。2018。为分层文本到图像合成推断语义布局。载于*IEEE*计算机视觉与模式识别会议论文集，第 7986–7994 页。

Saachi Jain、Hannah Lawrence、Ankur Moitra 和 Alek- sander Madry。2023。将模型失效提炼为潜在空间中的方向。载于第十一届国际学习表征会议。

Nikhil Kandpal、Eric Wallace 和 Colin Raffel。2022。去重训练数据可降低语言模型的隐私风险。载于第*39*届国际机器学习会议，《机器学习研究论文集》。PMLR。

M Kumar、Benjamin Packer 和 Daphne Koller。2010。潜变量模型的自步学习。神经信息处理系统进展，23。

Katherine Lee、Daphne Ippolito、Andrew Nystrom、Chiyuan Zhang、Douglas Eck、Chris Callison-Burch 和 Nicholas Carlini。2022。对训练数据进行去重能让语言模型变得更好。载于协会第*60*届年会论文集

计算语言学（第*1*卷：长篇论文），第8424–8445页，爱尔兰都柏林。计算语言学协会。

Junnan Li、Dongxu Li、Caiming Xiong 和 Steven Hoi。2022a。BLIP：为统一的视觉-语言理解与生成引导语言-图像预训练。在国际机器学习大会。

Minghui Li、Yan Wan 和 Jinping Gao。2022b。什么推动了深度合成应用的伦理接受度？模糊集定性比较分析。《人类行为中的计算机》，133:107286。

Tsung-Yi Lin、Michael Maire、Serge J. Belongie、Lubomir D. Bourdev、Ross B. Girshick、James Hays、Pietro Perona、Deva Ramanan、Piotr Doll'a r 和 C. Lawrence Zit nick。2014。Microsoft COCO：上下文中的常见对象。*CoRR*，abs/1405.0312。Xi Victoria Lin、Todor Mihaylov、Mikel Artetxe、Tianlu Wang、Shuohui Chen、Daniel Simig、Myle Ott、Na-man Goyal、Shruti Bhosale、Jingfei Du 等。2022。使用多语言生成式语言模型进行少样本学习。载于*2022* 年自然语言处理实证方法会议论文集。Fenglin Liu、Shen Ge 和 Xian Wu。2021。基于能力的多模态课程学习用于医学报告生成。载于计算语言学协会第 *59* 届年会暨第 *11* 届自然语言处理国际联合会议论文集（第 *1* 卷：长篇论文），第 3001–3012 页，线上发表。计算语言学协会。Thao Nguyen、Samir Yitzhak Gadre、Gabriel Ilharco、Sewoong Oh 和 Ludwig Schmidt。2023。通过图像描述改进多模态数据集。*arXiv* 预印本 *arXiv:2307.10350*。Alex Nichol、Prafulla Dhariwal、Aditya Ramesh、Pranav Shyam、Pamela Mishkin、Bob McGrew、Ilya Sutskever 和 Mark Chen。2022。GLIDE：利用文本引导的扩散模型实现照片级图像生成与编辑。Alexander Quinn Nichol 和 Prafulla Dhariwal。2021。改进的去噪扩散概率模型。载于国际机器学习大会，第 139 卷，第 8162–8171 页。PMLR。Kishore Papineni、Salim Roukos、Todd Ward 和 Wei-Jing Zhu。2002。BLEU：一种机器翻译自动评估方法。载于计算语言学协会第 *40* 届年会论文集，第 311–318 页。Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。2022。利用 CLIP 潜变量进行层次化文本条件图像生成。Rita Ramos、Bruno Martins 和 Desmond Elliott。2023。LMCap：通过检索增强语言模型提示实现少样本多语言图像描述生成。载于计算语言学协会论文集：*ACL 2023*。计算语言学协会。Nils Reimers 和 Iryna Gurevych。2019。Sentence-BERT：使用孪生 BERT 网络的句子嵌入。收录于*2019*年自然语言处理经验方法会议和第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第 3982–3992 页，中国香港。计算语言学协会。Anna Rogers。2021。通过改变数据改变世界。收录于计算语言学协会年会。Robin Rombach、A. Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer。2022。利用潜在扩散模型进行高分辨率图像合成。*2022* 年 *IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*），第 10674–10685 页。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、Tim Salimans、Jonathan Ho、David J Fleet 和 Mohammad Norouzi。2022a。具备深度语言理解能力的逼真文本到图像扩散模型。Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily L. Denton、Seyed Kamyar Seyed Ghasemipour、Burcu Karagol Ayan、Seyedeh Sara Mahdavi、Raphael Gontijo Lopes、Tim Salimans、Jonathan Ho、David J. Fleet 和 Mohammad Norouzi。2022b。具备深度语言理解能力的逼真文本到图像扩散模型。*arXiv* 预印本。Christoph Schuhmann、Richard Vencu、Romain Beaumont、Robert Kaczmarczyk、Clayton Mullis、Aarush Katta、Theo Coombes、Jenia Jitsev 和 Aran Komatsuzaki。2021。LAION-400M：包含 4 亿个经过 CLIP 筛选的图文对的开放数据集。*arXiv* 预印本。Piyush Sharma、Nan Ding、Sebastian Goodman 和 Radu Soric ut。2018。Conceptual Captions：用于自动图像图注生成的清洗后、上位词化的图像替代文本数据集。收录于第 *56* 届计算语言学协会年会论文集（第一卷：长篇论文），第 2556–2565 页，澳大利亚墨尔本。计算语言学协会。Ilia Shumailov、Zakhar Shumaylov、Yiren Zhao、Yarin Gal、Nicolas Papernot 和 Ross Anderson。2023。递归诅咒：使用生成数据进行训练会使模型遗忘。*arXiv* 预印本 *arXiv:2305.17493*。Yang Song、Jascha Sohl-Dickstein、Diederik P. Kingma、Abhishek Kumar、Stefano Ermon 和 Ben Poole。2021。通过随机微分方程实现基于得分的生成建模。收录于 *ICLR*。

David Thiel。2023。调查发现，AI图像生成模型使用儿童虐待内容进行训练。

Emiel van Miltenburg 和 Desmond Elliott。2017。《自动图像描述的改进空间：一项错误分析》。*CoRR*, abs/1704.04198。

Ramakrishna Vedantam、C Lawrence Zitnick 和 Devi Parikh。2015。Cider：基于共识的图像描述评估。载于*IEEE*计算机视觉与模式识别会议论文集，第 4566–4575 页。

Wenhui Wang、Hangbo Bao、Li Dong、Johan Bjorck、Zhiliang Peng、Qiang Liu、Kriti Aggarwal、Owais Khan Mohammed、Saksham Singhal、Subhojit Som 和 Furu Wei。2023。图像作为一种外语：用于视觉和视觉语言任务的 BEiT 预训练。载于*IEEE/CVF*计算机视觉与模式识别会议论文集。

Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、Rémi Louf、Morgan Funtowicz 等。2019。Hugging Face 的 Transformer模型：最先进的自然语言处理。*arXiv*预印本 *arXiv:1910.03771*。

ChangrongXiao、Sean XinXu 和 Kunpeng Zhang。2023。使用扩散模型进行图像描述的多模态数据增强。*arXiv* 预印本 *arXiv:2305.01855*。

Peter Young、Alice Lai、Micah Hodosh 和 Julia Hock- enmaier。2014。从图像描述到视觉指称：用于事件描述语义推理的新相似度指标。*TACL*，2：67–78。

Hongkuan Zhang、Saku Sugawara、Akiko Aizawa、Lei Zhou、Ryohei Sasano 和 Koichi Takeda。2022。基于跨模态相似度的图像描述课程学习。载于*2022* 年自然语言处理实证方法会议论文集。

A 附录

A.1 用于对文本到图像生成错误进行分类的人类研究用户界面

我们的用户界面如图10所示。我们要求标注人员勾选他们在给定合成图像中发现的错误。

错误类别包括：

- **人物**：年龄、性别、服装类型、服装颜色、异常的面部、异常的身体
- **主要物体**：错误、相似、不存在、多余、奇怪
- **其他物体**：错误、相似、不存在、多余、奇怪
- **一般类别**：姿势、活动、位置、数量、指代不一致、场景/事件/地点、文本、颜色，通常无关

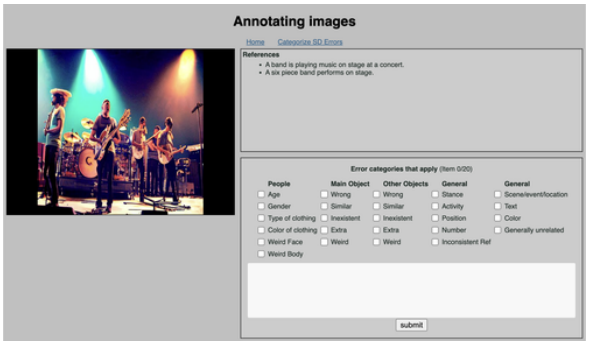


图10：用于对稳定扩散错误进行分类的标注界面。

A.2 文本到图像生成的提示词方法

图11 展示了我们用于提示文本生成图像模型的不同方法。我们通过检查生成的视觉示例，手动设计风格器。

A.3 使用 XGLM 生成替代图像描述

我们遵循（Ramos 等人，2023）中使用的提示词模板来获取语言模型生成的图像描述，即：

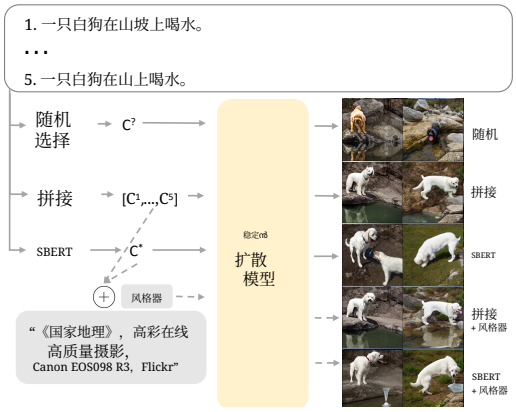


图11：使用文本生成图像进行合成图像生成的不同提示策略及代表性示例。根据我们的往返图像描述评估，使用拼接后的图像描述和风格器进行提示，能为该任务生成最佳图像。

“我是一个智能图像描述机器人。相似图像的图注如下： <图注> 我可以生成一条富有创意的简短图注来描述这张图像： <生成>”。这里我们使用四条真实图注作为 <图注>，并将另一条用于 <生成>，为图像构建三示例提示词。我们使用了 HuggingFace 上提供的“facebook/xglm-2.9B”模型（Wolf 等人，2019）。我们将最大生成长度设为 30 个词元，波束数量设为 5，以避免生成重复词元。

A.4 结合多种筛选方法

为了评估融合多种筛选方法的混合模型的有效性，我们首先以 BEiT-3 为基础，在 Flickr30K 数据集上进行了实验。在组合策略方面，我们选择了该数据集上效果最好的两种方法，即 REMOVE 和 REPLACEIMG。在每个训练轮次结束后，我们对训练样本进行筛选：剔除损失最高的样本中的一半，并替换其余一半样本对应的图像。在此过程中，我们筛选损失超过均值两个标准差的样本。实验取得了 83.8 的 CIDEr 分数和 32.8 的 BLEU4 分数，优于此前在该数据集上采用单一筛选方法所取得的性能。我们认为，通过采用更精细的组合策略，混合筛选方法有望带来更大的收益，这将留待未来工作进一步探索。

A.5 高损失训练样本

在图12中，我们展示了 COCO 数据集经过第一个训练轮次的微调后，高损失训练样本的情况。这些样本是我们数据整理技术的目标。与平均 11 个词的图注长度相比，排名靠前的样本图注都很长，约有 30 个词，这使模型难以学习。在接下来的微调训练轮次中，我们通过以下方式整理这些样本：完全移除文本-图像对（REMOVE）、替换图注（REPLACECAP），或将图像替换为合成的未见图像（REPLACEIMG）。

A.6 合成图像示例

在图13中，我们展示了文生图模型生成的高损失和低损失图像示例，并附上针对这些图像中识别出的错误的人工标注。

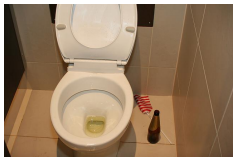
图像	图注	长度	Loss
	一张明显不尊重他人的人乱扔垃圾的照片， 酗酒，还把糟糕的选择留在厕所里不冲掉，并且 最糟糕的是，竟让星条旗碰到了浴室地板	26	213.24
	一张有许多自行车的被雨淋湿的街景照片 骑自行车的人，周围是看起来挤在一起的建筑物 and 战斗 for 空间 可能 look gray and 不起眼，但部分情况下并非如此	33	200.14
	一张展示户外场景的图片，远处有】【。 距离更近、灌木丛比运动场更多，且至少有两个 身着制服的年轻运动员，以及最近处的一名蓝色..... 栅栏和一张长长的长椅，搭配着	33	200.02
	一张户外拍摄的照片，画面中是一栋低矮屋顶建筑的外观 住所，一片种着纤细热带树木的小树林 为超现代风格的蓝白色搭建一个单坡棚屋 摩托车	30	199.90
	一张照片，紫蓝色的天空中有一个看起来像 风筝或松散的滑翔伞飘浮其中，占据了远景的大部分 画面最底部显示着一片草地 银行	33	197.36

图 12：COCO 在第一个训练轮次后的高损失训练样本，按损失从高到低排名。排名靠前的样本图注都很长，约有 30 个词，而数据集的平均值为 11 个词。

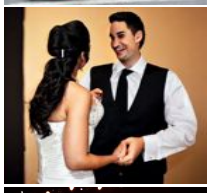
图像	图注	CLIPScore	Loss	分类错误
	一张两位女性的照片，其中一位穿着白色蕾丝连衣裙，带着手提包并穿着打底裤，另一位戴着高高的红帽子，穿着黑色中长连衣裙，身形像。上面穿着塑料连衣裙。	84.1	181.0	服装类型/颜色， 服装颜色， 奇怪的脸
	一位三轮车司机骑着自行车等候。	89.3	169.2	奇怪的主要物体， 奇怪的其他物体， 奇怪的身体部位， 姿势
	一名穿着黑色西装、打着领带并佩戴胸花的男子面带微笑看着一个也回以微笑的女孩，两人都坐在一张半正式活动（例如婚礼）上的餐桌或聚会。	77.6	163.5	服装颜色， 奇怪的身体部位， 错误-主要物体， 场景/事件/地点
	两名男子正在弹奏吉他，其中一名男子正在舞台上对着麦克风唱歌， 聚光灯照在他们身上。	74.7	26.0	奇怪的脸， 奇怪的身体部位， 奇怪的主要物体， 奇怪的其他物体
	昏暗的酒吧式房间里有几个人，包括一个坐在凳子上的女孩。	84.9	26.5	数量， 奇怪的脸， 奇怪的主体对象， 奇怪的身体部位
	许多孩子正在水中玩耍和游泳。 水中。	78.2	26.9	奇怪的脸， 奇怪的身体部位

图13：高损失合成图像（上）和低损失合成图像（下）的示例。人工标注表明，高损失样本中存在一致的错误类型，而 CLIPScore 与人类判断不一致。低损失合成图像在视觉上没有高损失图像那么复杂，但往往仍显得怪异，并且在颜色或物体方面存在错误。