SynC: 用于零样本图像字幕生成的一对多映射的合成图像字幕数据集优化

Si-Woo Kim
Hanyang University
boreng0817@hanyang.ac.kr

MinJu Jeon
Hanyang University
mnju5026@hanyang.ac.kr

Ye-Chan Kim
Hanyang University
dpcksdl78@hanyang.ac.kr

Soeun Lee
AI R & D Division, CJ Group
dlth508@cj.net

Taewhan Kim
Hanyang University
taewhan@hanyang.ac.kr

Dong-Jin Kim*
Hanyang University
djdkim@hanyang.ac.kr

Abstract

零样本图像描述（ZIC）日益利用由文本到图像（T2I）模型生成的合成数据集，以减轻昂贵的手动标注需求。然而，这些T2I模型常常生成与其对应输入描述语义不对齐的图像（例如，缺失对象，错误属性），导致嘈杂的合成图像-描述对，可能阻碍模型训练。现有的数据集剪除技术主要设计用于去除抓取自网络的嘈杂文本数据。然而，这些方法不适合合成数据的独特挑战，那里描述通常构造良好，但图像可能是错误的表示。为了解决这一问题，我们介绍了 *SynC*，这是一个专为ZIC优化合成图像-描述数据集而设计的新框架。不同于传统的过滤或重生成，*SynC* 专注于将描述重新分配给在合成图像集中已经存在的语义最为对齐的图像。我们的方法通过最初为每个描述检索多个相关候选图像，采用多对一映射策略。然后，我们应用一个受循环一致性启发的对齐评分器，选出最佳图像，通过图像到文本检索验证其能否检索回原始描述。广泛的评估表明，*SynC* 在多个ZIC模型和标准基准（MS-COCO，Flickr30k，NoCaps）上一致地显著提高了性能，并在若干场景中达到了最先进的结果。*SynC* 为策划精炼的合成数据以增强ZIC提供了有效的策略。[1]

CCS Concepts

• **Computing methodologies** → **Natural language generation**; *Visual content-based indexing and retrieval*; *Matching*.

Keywords

Zero-shot Image Captioning, Dataset Pruning, Synthetic Dataset

1 介绍

高质量的图像-描述配对数据集 [6] 对于训练高效的图像描述生成模型至关重要。然而，通过人工标注大规模创建这样的数据集是资源密集且繁重的。因此，零样本图像描述生成 (ZIC) 已成为一个重要的研究方向。最初的ZIC方法 [10, 20, 23, 33] 主要利用了像 CLIP [37] 这样的模型的跨模态理解能力进行仅文本训练。最近，强大的文本到图像（T2I）生成模型的出现，这些模型能够合成逼真的图像 [38, 40]，引发了一波新的ZIC方法浪潮 [26, 28, 29]。这些方法利用像 Stable Diffusion (SD) [40] 这样的模型从文本语料库生成合成图像，从而为训练描述生成模型创建图像-描述对，而无需人工标注。

尽管T2I模型已经取得了进展，但生成准确反映复杂输入字幕语义的图像仍然是一个显著挑战。正如 Fig. **??** 左侧部分所示，像SD这样的T2I模型通常会生成与源提示在语义上不对齐的图像，例如省略指定的物体或生成不正确的属性。现有的基于T2I的ZIC方法 [28, 29] 尝试在训练过程中缓解物体生成不良的问题 [28]，或通过利用大型语言模型（LLM）进行总结来优化提示工程 [29]。然而，这些策略并未明确保证或验证最终合成图像与其对应字幕之间的语义对齐。因此，这些方法使用的合成数据集 [28, 29] 不可避免地包含未对齐的图像-字幕对，这可能阻碍模型训练。

这就需要有效的策略来策划合成图像-标题数据集。然而，现有的数据集筛选技术并不适合这个特定的挑战。之前对标题数据集的筛选工作主要集中在不同的目标上，例如在真实世界数据集中消除性别偏见 [14] 或从未配对的来源中策划配对数据 [11, 16, 17]。此外，为剪除噪声的网络抓取视觉语言模型 (VLM) 数据集 [9, 12, 13, 19, 30] 开发的技术也不直接适用。这是因为现有的方法通常解决的是由与真实世界图像配对的低质量替代文本引起的噪声 [32, 42]。相反，合成数据集展示了相反的特征：标题通常格式良好，但由于生成不准确，图像成为噪声的来源。应用这些VLM修剪方法 [9, 12, 19] 甚至可能降低ZIC性能，并带来显著的计算开销，部分原因是一些方法 [12, 19, 30] 依赖于强大的预训练标题模型（例如，BLIP [21]，LLaVA [25]），这会引发公平性问题或利用大型语言模型 (LLMs) [9]，而这些模型无法内在地修复基于图像的对齐问题。

为了应对ZIC合成数据集编排的独特挑战，我们提出了 *SynC*，这是一个用于优化合成图片-标题对的新框架。*SynC* 不再是通过反复生成图片以实现对齐 [41]，而是侧重于重新分配标题给现有的合成图片，以最大化语义对齐。我们的方法涉及两个关键组件：（1）基于文本到图像（T2I）检索机制的一对多映射策略，以标题为查询，每个标题使用T2I检索多个潜在相关的合成图片。（2）多模态对齐评分函数，用于在使用建议的一对多映射策略选择的候选图片中挑选出最相关的合成图片。受到循环一致性学习框架 [22, 51] 的启发，这个评分器通过检查候选图片是否能够通过图像到文本（I2T）检索可靠地检索其对应的标题（或语义相似的标题）来评估对齐。结合基于T2I检索的一对多映射策略和基于I2T检索的多模态对齐评分函数，*SynC* 通过双交叉模态检索为查询标题识别出高度相关的合成图片，其中T2I用于候选，I2T用于评分。*SynC* 通过保留具有高对齐分数的新分配合成图片-标题对来优化噪声合成图像配对数据集，从而进一步提高零样本图像标题生成模型的性能。

尽管概念上简单，*SynC* 在标准图像字幕基准测试中，例如MS-COCO、Flickr30k和NoCaps [1, 6, 47]，在各种ZIC模型 [10, 20, 28, 33] 中表现出一致且显著的性能提升。我们的大量实验验证了 *SynC* 在优化合成图像字幕数据集方面的有效性，尤其是在增强ZIC模型训练方面。此外，*SynC* 提供了在

*Corresponding author.
[1]代码：https://github.com/boreng0817/SynC

零样本图像字幕背景下利用由 T2I 模型生成的大规模合成数据集的实用指导。

总之，我们的贡献如下：

- 我们通过提出一个新颖的细化框架 *SynC* 来解决合成图像在零样本图像描述中的质量和策展挑战，该框架采用一种灵活的、受循环一致性启发的方法，以细化更准确的图像-文本对。
- 我们设计了一对多映射策略和对齐评分函数用于合成图像选择，以评估图像-标题对齐，通过选择语义上对齐的图像和过滤噪声图像，实现准确的图像-标题配对。
- 大量评估表明，*SynC* 在各种零样本图像字幕模型中实现了一致的性能提升，甚至在各种场景中达到了最先进的性能。

2 相关工作

零样本图像描述。最近的零样本图像描述（ZIC）方法越来越多地利用文本到图像（T2I）模型，如 Stable Diffusion [40] 来合成训练数据 [26, 28, 29]，从而避免了对手动标注的需求。例如，ICSD [29] 在生成图像之前使用大型语言模型（LLMs）进行提示总结和选择，创建与标准基准类似的合成数据集 [6, 47]。PCM-Net [28] 也生成合成数据集，但直接使用描述作为提示，并结合训练策略以减轻合成图像固有的质量问题。尽管以前基于合成图像的 ZIC 方法解决了生成图像的质量问题，但它们常常忽略生成图像与其相应描述之间的细微不对齐，尤其是在捕捉细粒度细节时。相反，我们的 *SynC* 结合了灵活的一对多映射策略和循环一致性启发的对齐评分，能够实现更准确的图像-文本配对，显著提高零样本描述的性能。

数据集裁剪。过滤通过网络爬取的数据集对于训练高性能视觉语言模型（VLMs）至关重要。现有的剪枝方法通常修改文本（基于增强的方法）或基于初始相似度分数过滤对（基于度量的方法）；这些方法主要关注于文本模态，因为通过网络爬取的标题和替代文本通常是噪声数据。这种关注已被证明对网络数据有效，并增强了 VLM 性能。然而，合成图像标题数据集具有不同的特征。标题通常格式良好，但挑战在于生成的图像与其配对文本之间可能存在视觉语义不匹配。因此，先前的剪枝方法不太适合于精炼这些合成数据集。此外，传统的剪枝方法通常依赖于严格的一对一映射，丢弃低于阈值的配对。这可能导致潜在有用的标题被删除，仅仅因为它们最初的合成配对是不完美的，即便在数据集中可能存在更好的匹配。为了应对这些限制，我们引入了一种灵活的一对多映射，将格式良好的标题重新分配给最具语义对齐的图像。这使得更可靠的合成图像文本精炼成为可能，并提升了用于零样本标题生成的训练数据质量。

3 方法

在本节中，我们详细介绍了我们提出的方法 *SynC*，该方法专为将嘈杂的合成数据集精炼为良好对齐的图像-字幕对而设计。我们首先描述数据集修剪的基本概念和合成数据集 $\mathcal{D}_{syn}$（Sec. 3.1）的初始生成。然后，我们介绍 *SynC* 的核心组件：一种基于文本到图像检索（Sec. 3.2）的新颖一对多选择策略 $S_{\text{T2I}}(\cdot)$，以及用于策划用于零样本图像字幕的数据集的专用对齐评分函数 $f_{\text{ret}}(\cdot,\cdot)$（Sec. 3.3）。*SynC* 的整体流程如 Fig. 1 所示。

3.1 预备部分

我们首先建立一个通用的数据集修剪框架。我们的目标是将初始图像-字幕数据集 $\mathcal{D} = \{(I_i, C_i)\}_{i=1}^{N}$（包含 N 对图像-字幕）优化成一个更高语义对齐和信息丰富的精选数据集 $\mathcal{D}^\star$。最终的目标是生成一个能够增强零样本图像字幕模型训练效果的数据集。这个修剪过程包括两个关键组件：选择函数 $S(\cdot)$ 和多模态评分函数 $f(\cdot,\cdot)$。

选择函数 $S(\cdot)$ 为给定的查询标题识别一组相关的候选图像。形式上，对于一个查询标题 C，$S(\cdot)$ 从数据集的图像池中输出一个 K 候选图像的子集：$S(C) = \{I_i\}_{i\in\text{cand}}$，其中 $\text{cand} = \{i_1, i_2, ..., i_K\}$ 是所选图像的索引集。这个函数将一个标题映射到一个或多个候选图像。

多模态评分函数 $f(\cdot,\cdot)$ 量化了图像 I 和标题 C 之间的对齐程度。它将图像和标题对作为输入，并产生一个标量相似度分数 $s = f(I, C)$。较高的分数 s 表示图像 I 和标题 C 在不同模态之间有更强的语义对应关系。这个评分机制对于识别和保留良好对齐的图像和标题对至关重要。

合成数据集生成。给定一个未标记的文本语料库 $\mathcal{C} = \{C_i\}_{i=1}^{N}$，包括 N 标题，我们使用文本到图像 (T2I) 生成模型（例如，SD [40]）合成对应的图像集 $\mathcal{I}_{syn} = \{I_i^{syn}\}_{i=1}^{N}$。我们直接使用标题 $C_i \in \mathcal{C}$ 作为输入提示，不作修改。每个合成图像 I_i^{syn} 通过将相应的标题 C_i 输入到 T2I 模型中生成。这个过程生成了一个初步的、可能有噪声的成对合成数据集，记作 $\mathcal{D}_{syn} = \{(I_i^{syn}, C_i)\}_{i=1}^{N}$。

3.2 合成图像选择

初始合成数据集 $\mathcal{D}_{syn}$ 在每个标题 C_i 与其生成的图像 I_i^{syn} 之间建立了一一对应关系。这种合成数据分配可以通过一个简单的选择函数 $S_{\text{One}}(\cdot)$ 建模，其中一个标题 C_i 只选择其直接生成的图像：

$$S_{\text{One}}(C_i) = \{I_i^{syn}\}. \tag{1}$$

这一一对应的方法在现有的数据集筛选方法 [9, 13, 19, 30] 和基于 T2I 的零样本图像描述模型 [26, 28, 29] 中很常见。

然而，T2I 模型可能会生成与输入标题不相关或无法捕捉完整语义的合成图像（Fig. ??）。此外，简单的标题可能主要描述主导概念，同时省略较细的细节 [29]。相反，人类标注的数据集如 COCO [6] 和 Flickr30k [47] 通常会为每张图像提供多个参考标题。

受到人类标注一对多特性的启发，且旨在缓解不完美 T2I 生成带来的问题，我们提出了一种新颖的一对多选择策略用于合成图像标题配对数据集 $S_{\text{T2I}}(\cdot)$。我们的方法不局限于一开始生成的 I_i^{syn}，而是允许每个标题 C_i 从预生成的图像池 $\mathcal{I}_{syn}$ 中进行选择。至关重要的是，虽然 $S_{\text{One}}(\cdot)$ 放弃了不匹配的对，但 $S_{\text{T2I}}(\cdot)$ 旨在主动为标题 C_i 从现有池中寻找一个更好匹配的合成图像，从而有可能改进那些因为初始 T2I 生成不佳而被丢弃的有价值标题。

具体来说，对于每个标题 $C_i \in \mathcal{C}$，我们在整个生成的合成图像集 $\mathcal{I}_{syn}$ 上执行文本到图像（T2I）检索。我们使用一个预训练视觉-语言模型（例如，CLIP [37]，SigLIP [43, 50]）的图像编码器 $\mathcal{E}_I(\cdot)$ 和文本编码器 $\mathcal{E}_T(\cdot)$，以在预生成的图像池中找到与查询标题 C_i 最相关的 K 合成图像。这个选择过程不需要额外的图像生成；它只是从现有的图像池 $\mathcal{I}_{syn}$ 中检索。我们定义了这个一对多选择函数，

$$S_{\text{T2I}}(C_i) = \{I_r^{syn}\}_{r\in R_i}, \tag{2}$$

$$\text{where } R_i = \underset{j\in[1,N]}{\text{Top-K}}\{\mathcal{E}_I(I_j^{syn}), \langle \mathcal{E}_T(C_i)\rangle\}. \tag{3}$$

在这里，$\langle\cdot,\cdot\rangle$ 表示余弦相似度 $\cos(\cdot,\cdot)$，$R_i = [r_1, ..., r_K]$ 是与标题 C_i 的前 K 个检索图像对应的索引集。因此，对于每个标

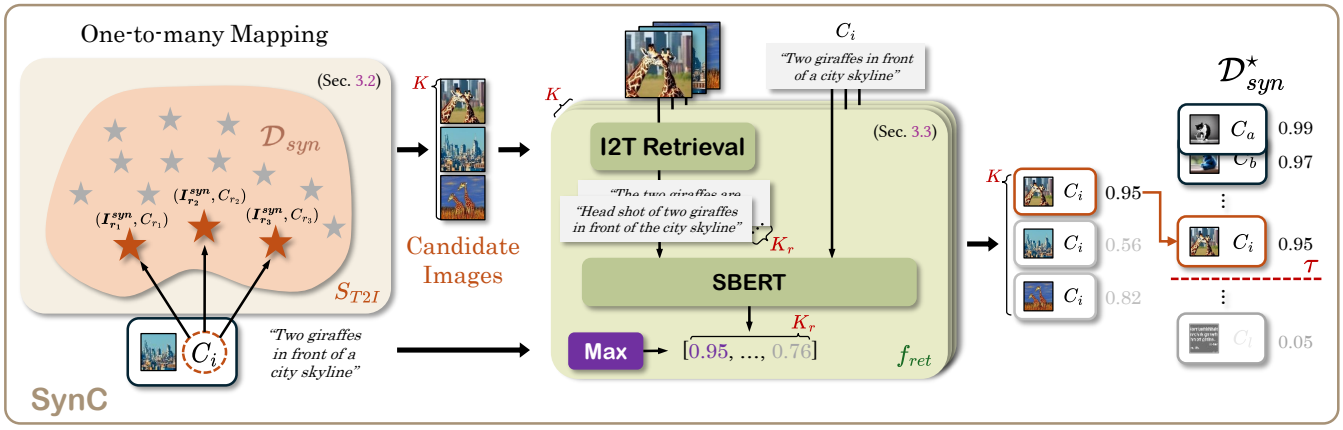

Figure 1: *SynC* 的整体流程将合成图像字幕对数据集 $\mathcal{D}_{syn}$ 处理为精炼的 $\mathcal{D}^{\star}_{syn}$。通过选择函数 $S_{\text{T2I}}(\cdot)$，*SynC* 执行文本到图像的检索，为每个字幕 $C_i \in \mathcal{C}$ 找到与查询字幕相关的合成图像。有了 K 候选图像，*SynC* 计算 C_i 和候选图像之间的对齐分数，挑选出最佳对与相似度分数。如果来自 C_i 的生成图像与输入提示不匹配，*SynC* 会重新分配一个合适的合成图像。通过遍历文本语料库 $\mathcal{C}$ 中的字幕 C_i，我们得到一组由合成图像、字幕和对齐分数组成的三元组。保留上层 τ 部分，*SynC* 最终输出精炼的数据集 $\mathcal{D}^{\star}_{syn}$。

题 C_i，我们考虑 K 候选图像 $\{I^{syn}_r\}_{r\in R_i}$。这形成了一组候选对 $\{(I^{syn}_r, C_i)\}_{r\in R_i}$，我们将使用评分函数选择其中对齐最好的一对。我们提出的方法 *SynC* 采用 $S(\cdot) := S_{\text{T2I}}(\cdot)$ 作为其选择函数。

3.3 对齐评分函数

为图像-字幕对齐进行评分的标准方法是 CLIPScore [13]。使用 CLIP 的图像编码器 $\mathcal{E}^{CLIP}_I(\cdot)$ 和文本编码器 $\mathcal{E}^{CLIP}_T(\cdot)$，CLIPScore 函数 $f_{\text{CLIP}}(\cdot,\cdot)$ 计算图像 I 和字幕 C 的嵌入之间的余弦相似度：

$$f_{\text{CLIP}}(I, C) = w \cdot \langle \mathcal{E}^{CLIP}_I(I), \mathcal{E}^{CLIP}_T(C) \rangle, \tag{4}$$

其中 w 是缩放因子，为简单起见，我们使用 $w = 1.0$。

我们的初步实验表明，使用 $f_{\text{CLIP}}(\cdot,\cdot)$ 进行过滤，通过去除一些不匹配的对，能在零样本图像描述模型上带来细微的提升。值得注意的是，CLIP 通常优先考虑整体的对齐，可能忽略细粒度的细节，或者难以进行精确的组合理解 [30, 31, 35, 49]。

为了克服这些局限性，特别是在改进嘈杂的合成数据集的情况下，我们引入了一种新的基于检索的多模态评分函数，$f_{\text{ret}}(\cdot,\cdot)$。在选择功能中的 T2I 检索之后，$f_{\text{ret}}(\cdot,\cdot)$ 使用候选合成图像进行图像到文本（I2T）的检索，然后在单模态文本空间内将检索到的标题与原始目标标题进行比较。其基本原理是，原始标题 C_i 通常比可能存在缺陷的合成图像 I^{syn} 更可靠地表示预期的语义。因此，我们并不直接在多模态空间中比较图像和文本的嵌入（如在 CLIPScore 中），而是利用检索方向的一致性。受循环一致性概念的启发，我们的方法结合了用于选择的 T2I 检索（在 Eq. (3) 中）和用于评分的 I2T 检索。

具体来说，对于候选对 (I^{syn}, C)，我们首先使用合成图像 I^{syn} 通过 I2T 检索从原始语料库 $\mathcal{C}$ 中检索出最语义相似的字幕，使用相同的 VLM 编码器 $\mathcal{E}_I$ 和 $\mathcal{E}_T$。然后我们测量查询字幕 C 与这些图像检索字幕之间的语义相似性。对于这种仅文本的比较，我们采用专用的单模态文本编码器 Sentence Transformer（SBERT）[39]，记为 $\mathcal{E}_S(\cdot)$，以其在捕捉句子级语义的有效性而著称。

我们定义了基于检索的评分函数 $f_{\text{ret}}(\cdot,\cdot)$ 如下：

$$f_{\text{ret}}(I^{syn}, C) = \max_{r\in\hat{R}(I^{syn})} \langle \mathcal{E}_S(C_r), \mathcal{E}_S(C) \rangle, \tag{5}$$

$$\text{where } \hat{R}(I^{syn}) = \underset{j\in[1,N]}{\text{Top-K}_r} \{\langle \mathcal{E}_I(I^{syn}), \mathcal{E}_T(C_j) \rangle\}. \tag{6}$$

这里，$\hat{R}(I^{syn})$ 表示从语料库 $\mathcal{C}$ 中基于它们与 VLM 嵌入空间中的合成图像 I^{syn} 的相似性检索到的前 K_r 个标题集（使用编码器 $\mathcal{E}_I$ 和 $\mathcal{E}_T$ 与 Eq. (4)）。最终得分是目标标题 C 与图像相关标题之间最高的 SBERT 相似度。我们提出的方法 *SynC* 使用 $f(\cdot,\cdot) := f_{\text{ret}}(\cdot,\cdot)$ 作为其评分函数。

通过使用合成图像选择策略 $S(\cdot)$ 和评分函数 $f(\cdot,\cdot)$，我们迭代计算 C_i 和 $I^{syn} \in S(C_i)$ 之间的对齐分数。相似度 s_i 是最高的对齐分数，然后我们将最相关的合成图像 $I^{syn}_{i^*}$ 匹配到 C_i，并形成一个精炼的合成图像字幕对 $(I^{syn}_{i^*}, C_i)$。

SynC 接收一个输入 $\mathcal{D}_{syn}$，对 $\mathcal{D}_{syn}$ 中的每对进行该程序的迭代，输出 $\mathcal{D}'_{syn} = \{(I^{syn}_{i^*}, C_i, s_i)\}^{N}_{i=1}$，这是一组包括精炼的合成图像、标题和对齐分数 s_i 的三元组。由于一些精炼的对可能因为句子的复杂性或错误书写的标题（例如，语法错误、打字错误）而未对齐，*SynC* 过滤掉 $\mathcal{D}'_{syn}$ 的下半部分。在按对齐分数降序排序 $\mathcal{D}'_{syn}$ 之后，*SynC* 保留上部的一部分，$\tau \in [0, 1]$。我们得到一个精炼的合成图像标题对数据集 $\mathcal{D}^{\star}_{syn} = \{(I^{syn}_{i^*}, C_i)\}^{\lfloor N\cdot\tau \rceil}_{i=1}$ 作为最终输出。我们执行一个 $\lfloor\cdot\rceil$ 楼层操作用于整数转换 $N \cdot \tau$。整个 *SynC* 的过程可以写为

$$SynC(\mathcal{D}_{syn}, \tau) = \mathcal{D}^{\star}_{syn}. \tag{7}$$

。

4 实验

4.1 实现细节和设置

我们采用了 PCM-Net [28] 作为基准零样本图像描述模型。虽然存在几个相关的方法 [26, 28, 29]，但 PCM-Net 是唯一公开可用的实现，能够实现可再现的比较。我们通过将其应用于公开源码的 PCM-Net [28] 零样本描述基准所用的数据集，来评估我们的合成数据优化方法 *SynC*。对于 MS-COCO [6] 和

Table 1: MS-COCO 和 Flickr30k 中的领域内描述整体结果 [6, 47] 。我们测量了最常用的描述指标：BLEU@4 (B@4)、METEOR (M)、ROUGE (R)、CIDEr (C) 和 SPICE (S) [2, 3, 24, 36, 44] 。使用经过 *SynC* 优化的合成数据集训练的 *Baseline* 表现出相对于 *Baseline* 的显著性能提升，达到了当前最先进的性能。最佳得分在 bold 。† 表示在训练期间使用真实图像。

Method	COCO					Flickr30k				
	B@4	M	R	C	S	B@4	M	R	C	S
ViT-B/32										
DeCap [23] ICLR'23	24.7	25.0	–	91.2	18.7	21.2	21.8	–	56.7	15.2
ViECap [10] ICCV'23	27.2	24.8	–	92.9	18.2	21.4	20.1	–	47.9	13.6
ICSD [29] AAAI'24	29.9	25.4	–	96.6	–	25.2	20.6	–	54.3	–
SynTIC [26] AAAI'24	29.9	25.8	53.2	101.1	19.3	22.3	22.4	47.3	56.6	16.6
IFCap [20] EMNLP'24	30.8	**26.7**	–	108.0	20.3	23.5	23.0	–	64.4	17.0
Baseline [28] ECCV'24	31.5	25.9	53.9	103.8	19.7	26.9	23.0	50.1	61.3	16.8
+*SynC*	**33.6**	**26.7**	**55.3**	**112.0**	**20.5**	**28.2**	**23.4**	**50.5**	**65.8**	**17.1**
ViT-L/14										
CgT-GAN † [48] ACMMM'23	30.3	26.9	54.5	108.1	20.5	24.1	22.6	48.2	64.9	16.1
Baseline [28] ECCV'24	33.6	26.9	55.4	113.6	20.8	28.5	24.3	51.4	69.5	18.2
+*SynC*	**35.2**	**27.8**	**56.5**	**119.8**	**21.9**	**29.6**	**24.8**	**52.5**	**75.6**	**18.5**

Table 2: 跨领域描述生成的结果。*SynC* 优于 *Baseline* ，在大多数指标中达到最新的性能水平。

Method	COCO ⟹ Flickr30k					Flickr30k ⟹ COCO				
	B@4	M	R	C	S	B@4	M	R	C	S
ViT-B/32										
DeCap [23]	16.3	17.9	–	35.7	11.1	12.1	18.0	–	44.4	10.9
ViECap [10]	17.4	18.0	–	38.4	11.2	12.6	19.3	–	54.2	12.5
SynTIC [26]	17.9	18.6	42.7	38.4	11.9	14.6	19.4	40.9	47.0	11.9
IFCap [20]	17.8	19.4	–	47.5	12.7	14.7	**20.4**	–	**60.7**	**13.6**
Baseline [28]	20.8	19.2	45.2	45.5	12.9	17.1	19.6	43.2	54.9	12.8
+ *SynC*	**22.4**	**20.0**	**45.9**	**49.5**	**13.4**	**18.4**	20.0	**44.1**	58.4	13.4
ViT-L/14										
CgT-GAN † [48]	17.3	19.6	43.9	47.5	12.9	15.2	19.4	40.9	58.7	13.4
Baseline [28]	23.9	21.2	47.8	55.9	14.2	17.9	20.3	44.0	61.3	13.5
+ *SynC*	**25.2**	**22.1**	**49.6**	**60.3**	**15.8**	**19.9**	**21.3**	**45.6**	**66.8**	**14.9**

Flickr30k [47] 的评估，我们优化了 PCM-Net 的公共合成数据集 $\mathcal{D}_{\text{SynthImgCap}}$ [28] 。

对于在 CC3M [5] 和 SS1M [11] 上的实验，我们利用来自 huggingface [46] 的稳定扩散 v1.4 [2] 来生成带有 CC3M 和 SS1M 来源说明的合成图像。我们将图像分辨率设置为 512×512，使用 DPMSolver [27] 进行 20 步采样。我们的方法应用了 $S(\cdot) := S_{\text{T2I}}(\cdot)$ 选择函数，使用 SigLIP2 ViT-B/16@256 [43] 从 $K = 15$ 候选图像中进行基于 Fig. 4 的检索。评分函数 $f(\cdot,\cdot) := f_{\text{ret}}(\cdot,\cdot)$ 结合了 SigLIP2 ViT-B/16@256 和轻量级蒸馏句子变换器 all-MiniLM-L6v2 [45] 。我们选择保留比例 $\tau = 0.9$ ，如 Fig. 3 和基于 Fig. 4 的 $K_r = 2$ 。

评估使用 MS-COCO 和 Flickr30k (Karpathy split [15]) 进行领域内任务，并使用 NoCaps [1] 验证集进行领域外泛化。我们报告标准指标：BLEU@4 [36] 、METEOR [3] 、ROUGE-L [24] 、CIDEr [44] 和 SPICE [2] 。

[2] https://huggingface.co/CompVis/stable-diffusion-v1-4

Table 3: 在跨域设置下 NoCaps [1] 验证集上的结果。*SynC* 显示了在域外场景中的性能提升。

Method	NoCaps							
	In		Near		Out		Entire	
	C	S	C	S	C	S	C	S
ViECap [10]	61.1	10.4	64.3	9.9	65.0	8.6	66.2	9.5
SYN-ViECap [28]	61.7	**10.5**	68.5	10.3	71.4	**9.4**	70.5	10.0
+*SynC*	**65.7**	10.4	**71.2**	**10.5**	**71.6**	**9.4**	**72.7**	**10.1**

Table 4: 使用从 CC3M [5] 和 SS1M [11] 的文本源分别生成的 20 万对合成图像字幕对进行跨领域图像字幕生成的结果。在每种设置中，*SynC* 都表现出一致的性能提升。

Dataset	Dataset ⟹ COCO				
	B@4	M	R	C	S
Baseline ViT-L/14 [28]					
$\mathcal{D}_{\text{CC3M}}$	9.6	14.0	34.7	41.3	10.1
+*SynC*	**11.0**	**15.3**	**36.9**	**47.7**	**11.3**
$\mathcal{D}_{\text{SS1M}}$	10.7	13.4	29.9	47.5	**11.3**
+*SynC*	**11.3**	**14.2**	**30.9**	**49.1**	**11.3**

4.2 主要结果

我们通过比较基线 PCM-Net 模型 [28] 在其原始合成数据集 $\mathcal{D}_{\text{SynthImgCap}}$ 上进行训练时的性能与使用我们的方法改进后的数据集 $\mathcal{D}^{*}_{\text{SynthImgCap}}$ 上的性能来评估 *SynC* 的有效性。我们将基线记为 *Baseline* ，而将使用我们改进的数据集进行训练的基线记为 +*SynC* 。

与当前最先进方法的比较。我们将 +*SynC* 与当代最先进的零样本图像描述模型进行比较。这些方法包括仅在文本数据上训练的模型，如 DeCap [23] 、ViECap [10] 和 IFCap [20] ，以及利用合成图像的方法，如原始 PCM-Net [28] 、SynTIC [26] 、ICSD [29] 和 CgT-GAN [48] 。遵循常规做法，我们根据主干视觉变压器分组结果：CLIP [37] ViT-B/32 和 ViT-L/14。

Figure 2: 关于 $S_{\text{One}}(\cdot)$ 和我们 $S_{\text{T2I}}(\cdot)$ 之间比较的定性结果。$S_{\text{T2I}}(\cdot)$ 可以为给定的查询标题搜索合适的合成图像。结合 $f_{\text{ret}}(\cdot,\cdot)$，*SynC* 可以优化未对齐的对，而 $S_{\text{One}}(\cdot)$ 则丢弃该对。

Table 5: 将 *SynC* 与网络数据剪枝方法进行比较。基于扩充的剪枝方法会降低 *Baseline* 的性能。基于度量的调优方法略微提升了 *Baseline*，而加入我们提出的 $S_{\text{T2I}}(\cdot)$ 显示了额外的性能提升。

Method	COCO				
	B@4	M	R	C	S
Baseline [28]	31.5	25.9	53.9	103.8	19.7
+ *SynC*	**33.6**	**26.7**	**55.3**	**112.0**	**20.5**
Augment-based pruning					
VeCLIP [19] ECCV'24	27.0	25.9	52.5	93.9	19.8
LaCLIP [9] NeurIPS'23	23.7	22.6	49.6	80.5	16.3
Recaptioning [12] ArXiv'24	27.8	25.8	52.8	95.1	19.2
Metric-based pruning					
Sieve [30] CVPR'24	32.4	25.8	54.1	106.3	**19.9**
Sieve + S_{T2I}	**32.9**	**26.1**	**54.6**	**108.1**	**19.9**
CLIPScore [13] EMNLP'21	31.6	25.8	53.7	104.4	19.5
CLIPScore + S_{T2I}	**32.3**	**26.0**	**54.3**	**106.5**	**19.8**

域内字幕生成。结果展示在 Table 1 中。应用我们的 *SynC* 在所有评估指标上，对于两个数据集和两种骨干网络，相较于 *Baseline* 都有显著的改善。特别是在 MS-COCO 上，+*SynC* 的 CIDEr 分数相较于 *Baseline* 提高了 +8.2 (ViT-B/32) 和 +6.2 (ViT-L/14)。在 Flickr30k 上同样观察到显著的 CIDEr 提升，分别为 +4.5 (ViT-B/32) 和 +6.2 (ViT-L/14)。这些一致的提升验证了通过 *SynC* 精炼合成的图像字幕对，提高了零样本字幕生成模型的性能。

跨域和域外泛化。我们进一步在跨域和域外环境中评估 *SynC*。正如 Table 2 所示，+*SynC* 在跨数据集的传输中在所有指标上始终优于 *Baseline*。对于 COCO 到 → Flickr30k（包括 ViT-B/32 和 ViT-L/14 骨干网络）以及 Flickr30k 到 → COCO（ViT-L/14），它达到了最先进的结果。对于 Flickr30k 到 → COCO (ViT-B/32)，它也显著超过了基线，并在与先前技术相比时展现出强大的竞争力。

Table 6: 网页数据集修剪方法和 *SynC* 的计算成本测量。成本以 RTX A6000 GPU 小时数表示。Δ 表示 MS-COCO 测试集上的 CIDEr 评分增益，相较于 *Baseline*。

Method	Preprocess	$\mathcal{E}(\cdot)$	S	f	Total	Δ
VeCLIP [19]	46.5h	–	–	–	46.5h	-9.9
LaCLIP [9]	28h	–	–	–	28h	-23.3
Recaptioning [12]	36h	–	–	–	36h	-8.7
Sieve [30]	36h	0.7h	–	0h	36.7h	+2.5
CLIPScore [13]	–	0.9h	–	0h	0.9h	+0.6
SynC	–	1.5h	0.4h	0.4h	2.3h	+8.2

在 NoCaps 验证集（Table 3）上，我们使用经过 $\mathcal{D}_{\text{SynthImgCap}}$ 训练的 SYN-ViECap [10]。SYN-ViECap 在 *SynC* 精炼的数据集上训练后，展示出增强的领域外泛化能力，与 [28] 中报告的基准表现相比，在所有分割中的 CIDEr 和大多数分割中的 SPICE 都有所提高。这些结果证实了使用 *SynC* 精炼合成数据的广泛适用性和泛化优势。

利用网络数据进行跨领域描述生成。我们进一步验证了我们提出的 *SynC* 从人工注释的描述中获得的好处。由于 COCO 和 Flickr30k 的每张图像包含多个描述，注释风格可能会影响 *SynC* 的性能提升。为了模拟稀疏环境，我们使用 CC3M 和 SS1M [5, 11] 的部分训练描述来创建合成图像描述数据集 $\mathcal{D}_{\text{CC3M}}$ 和 $\mathcal{D}_{\text{SS1M}}$。我们从每个数据集中随机抽取 200,000 个描述，然后进行如 Sec. 3.1 中所示的图像生成。我们分别使用 $\mathcal{D}_{\text{CC3M}}$ 和 $\mathcal{D}_{\text{SS1M}}$ 与 ViT-L/14 一起训练 *Baseline*，然后在 COCO 测试集上测试，在网络爬取数据集来源的文本与合成图像描述配对数据集上的 *SynC* 结果如 Table 4 所示。*SynC* 在大多数指标上成功提升了 *Baseline* 的性能。这验证了 *SynC* 在稀疏的描述环境中也能得到改进。

4.3 与其他剪枝方法的比较

与网络数据过滤方法的比较。我们将 *SynC* 与应用于合成数据集 $\mathcal{D}_{\text{SynthImgCap}}$ 的标准网络数据过滤方法进行比较。这些方法

Table 7: 在不同的零样本图像描述模型上，*SynC* 表现出一致的性能提升。

Method	COCO				
	B@4	M	R	C	S
CapDec [33]	26.4	25.1	51.8	91.8	–
$\mathcal{D}_{\text{SynthImgCap}}$	25.4	24.7	50.4	89.3	18.6
+*SynC*	**27.7**	**25.7**	**52.1**	**96.4**	**19.3**
ViECap [10]	27.2	**24.8**	–	92.9	18.2
$\mathcal{D}_{\text{SynthImgCap}}$	26.3	24.1	49.9	91.4	17.9
+*SynC*	**27.7**	**24.8**	**51.6**	**96.4**	**18.5**
IFCap [20]	30.8	**26.7**	–	108.0	**20.3**
$\mathcal{D}_{\text{SynthImgCap}}$	29.9	25.7	53.0	106.3	19.6
+ *SynC*	**31.1**	26.4	**53.6**	**108.6**	**20.3**

包括基于增强的方法（VeCLIP [19]，LaCLIP [9] 和 Recaptioning [12]）和基于指标的方法（Sieve [30] 和 CLIPScore [13]）。如 Table 5 中所示，当应用于我们的基线模型时：1）基于增强的方法，旨在处理文本噪声，其表现下降可能是因为合成数据中的主要噪声来源是视觉而非文本。2）基于指标的方法（Sieve, CLIPScore）相比基线仅提供了微小的改进，显著不如我们的 +*SynC* 方法有效，突显了其严格一对一对过滤的局限性。

然而，我们的核心一对多选择策略（$S_{\text{T2I}}(\cdot)$，在 Table 5 中）补充了现有的基于指标的筛选器。在用 Sieve 或 CLIPScore 进行初步筛选后应用 $S_{\text{T2I}}(\cdot)$，在大多数指标上均带来了进一步的一致性提升（例如，CIDEr 分别提升了 +1.8 和 +2.1）。这显示了我们选择策略的附加价值和兼容性。

计算成本。我们以 RTX A6000 GPU 小时数来衡量计算成本。我们的分析考虑了剪枝方法中可能涉及的四个阶段：1）预处理（例如，大型语言模型推理和图像描述生成），2）使用 VLM 编码器（例如，CLIP [37] 和 SigLIP [43, 50]）或单模态文本编码器（例如，句子变换器 [45]）生成嵌入 $\mathcal{E}(\cdot)$，3）选择策略 $S(\cdot)$，以及 4）使用多模态评分函数 $f(\cdot,\cdot)$ 计算对齐分数。

对于 *SynC*，需要两个嵌入步骤：生成完整合成数据集的 SigLIP2 嵌入（$\sim$ 1.5 小时）和文本语料库的 Sentence Transformer 嵌入（$\sim$ 2 分钟）。*SynC* 涉及每个 $S_{\text{T2I}}(\cdot)$ 和 $f_{\text{ret}}(\cdot,\cdot)$ 的两个检索阶段（T2I, I2T），每个阶段大约需要 0.4 小时。使用 $f_{\text{ret}}(\cdot,\cdot)$ 计算最终对齐分数的时间不到 2 分钟。

总共使用 *SynC* 对所有的 $\mathcal{D}_{\text{SynthImgCap}}$ 进行精炼，需要约 2.3 小时的 RTX A6000 GPU 时间，这包括了 COCO 的 542,401 张图像-说明对。这比 CLIPScore 多花费 1.4 小时，但性能提升了 7.6 个 CIDEr 分数。其他网络数据剪枝方法涉及像 Vicuna、LLaVA 或 BLIP [7, 21, 25] 这样的大模型；预处理需要更多的 GPU 时间。

4.4 *SynC* 应用

我们还将 *SynC* 应用于其他零样本图像描述模型，以验证其泛化能力。我们修改了仅用于文本训练的图像描述模型 CapDec、ViECap 和 IFCap [10, 20, 33]，以便与合成的图像描述对进行训练。对于 CapDec 和 ViECap，我们将 CLIP 文本编码器修改为 CLIP 图像编码器，并消除了噪声注入过程。对于 IFCap，我们也执行与 CapDec 和 ViECap 相同的方法，将 IFCap 的类似图像检索过程替换为标准的图像到文本检索。

Table 8: 选择函数 $S(\cdot)$ 和多模态评分函数 $f(\cdot,\cdot)$ 的消融研究。

S	f	COCO				
		B@4	M	R	C	S
Baseline [28]		31.5	25.9	53.9	103.8	19.7
S_{One}	f_{SigLIP2}	31.7	25.6	53.6	104.0	19.4
	f_{ret}	31.7	25.7	53.8	105.2	19.6
S_{T2T}	f_{SigLIP2}	32.3	26.0	54.2	106.8	20.0
	f_{ret}	32.8	26.2	54.6	108.4	20.3
S_{T2I}	f_{SigLIP2}	32.9	26.2	54.7	108.2	20.3
	f_{ret}	**33.6**	**26.7**	**55.3**	**112.0**	**20.5**
S_{I2T}	f_{SigLIP2}	31.6	25.6	53.9	104.0	19.4
	f_{ret}	31.5	25.7	53.9	104.8	19.7
S_{I2I}	f_{SigLIP2}	32.0	25.8	54.1	104.4	19.6
	f_{ret}	31.8	25.7	53.9	105.2	19.6

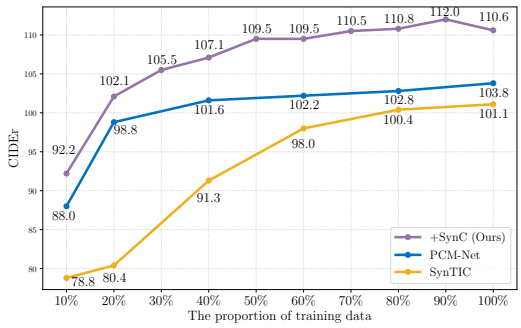

Figure 3: 在 MS-COCO 上关于保留比例 τ 的消融研究。

其他零样本图像描述模型中的*SynC*。我们通过比较在 $\mathcal{D}_{\text{SynthImgCap}}$ 上训练并用 *SynC* 精炼的修改后 CapDec、ViECap 和 IFCap [10, 20, 33] 的性能，展示了 *SynC* 的有效性。我们在 Table 7 中报告了 MS-COCO 领域内的性能。与之前的实验一样，*SynC* 提升了各种零样本图像描述模型，这验证了 *SynC* 的普遍适用性。

4.5 消融研究

选择函数。我们将主要的一对多字幕到图像策略（$S_{\text{T2I}}(\cdot)$，详见 Sec. 3.2）与几个替代方案进行比较：基准的一对一映射 $S_{\text{One}}(\cdot)$，以及其他一对多变体，包括文本到文本 S_{T2T}，检索相似的字幕，然后选择它们对应的图像，S_{I2T} 和 S_{I2I}。在 Table 8 中呈现的结果表明，以字幕为主要查询的策略（$S_{\text{T2I}}(\cdot)$ 和 S_{T2T}）显著优于以图像为查询的策略。这支持了我们的假设，即专注于高质量字幕作为查询并可能丢弃对齐不良的合成图像，更有助于优化图像-字幕数据集。

多模态评分函数。为了单独衡量所提出的 $f_{\text{ret}}(\cdot,\cdot)$ 的效果，我们将其与 $f_{\text{SigLIP2}}(\cdot,\cdot)$ 进行比较，后者是一个修改过的 CLIPScore [37]，其中的 CLIP 编码器被 SigLIP2 ViT-B/16@256 [43] 替代（以便与我们方法中使用的编码器相匹配，实现公平比较）。这种比较是在不同的选择策略下进行的。如 Table 8 中所报告，$f_{\text{ret}}(\cdot,\cdot)$ 在大多数指标上表现优异，尤其是在与有效的标题-查询选择策略 $S_{\text{T2I}}(\cdot)$ 和 $S_{\text{T2T}}(\cdot)$ 配对时表现更好。即便在选择方法不太理想的情况下，$f_{\text{ret}}(\cdot,\cdot)$ 也表现得与 $f_{\text{SigLIP2}}(\cdot,\cdot)$ 相当甚至稍优。基于这些结果，我们确认 $S_{\text{T2I}}(\cdot)$ 选择策略与 $f_{\text{ret}}(\cdot,\cdot)$ 评分函数的组合是 *SynC* 的最有效配置。

Table 9: 关于 VLMs 编码器 $\mathcal{E}_I$ 和 $\mathcal{E}_T$ 的消融研究。

VLMs Encoder Method	Variant	COCO B@4	M	R	C	S
Baseline [28]		31.5	25.9	53.9	103.8	19.7
CLIP [37]	ViT-B/32	31.9	25.7	54.2	105.9	19.6
	ViT-L/14	32.3	25.9	54.4	106.7	19.5
	RN50x4	32.0	25.8	54.2	106.8	19.6
SigLIP [50]	ViT-B/16	33.3	26.6	55.1	109.9	20.6
	ViT-L/16@384	33.3	26.5	55.1	110.0	20.3
SigLIP2 [43]	ViT-B/16@256	33.6	26.7	55.3	**112.0**	20.5
	ViT-L/16@512	**34.0**	**26.8**	**55.5**	**112.0**	**20.8**

Table 10: 关于文本编码器 $\mathcal{E}_S$ 用于 $f_{\text{ret}}(\cdot,\cdot)$ 的消融研究。

$\mathcal{E}_S$ for f_{ret}	COCO B@4	M	R	C	S
Baseline [28]	31.5	25.9	53.9	103.8	19.7
Sentence Transformer [45]	**33.6**	**26.7**	**55.3**	**112.0**	**20.5**
SigLIP2 ViT-B/16@256 [43]	33.1	26.4	54.9	108.6	20.3

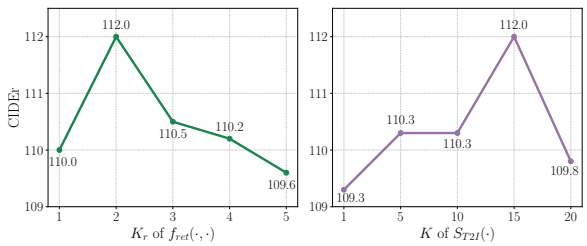

Figure 4: 关于超参数 K 和 K_r 的消融研究。

跨模态操作的 VLM 编码器选择。在 Table 9 中，我们评估了用于跨模态检索的 VLM 编码器的影响。我们比较了 3 个具有不同变体的 VLM。虽然所有测试的编码器在性能上均优于 *Baseline*，但 SigLIP2 模型效果最佳。在比较 SigLIP2 的变体时，ViT-L/16 相比 ViT-B/16 仅提供了微小的改善，但为嵌入带来了大约 5 倍的计算成本的显著增加。考虑到这一权衡，我们选择 SigLIP2 ViT-B/16@256 作为 *SynC* 的 VLM 编码器，因为它表现出色且效率更高。

的文本编码器 $f_{\text{ret}}(\cdot,\cdot)$。我们特别探讨了在 Table 10 内选择生成这些嵌入的文本编码器 $\mathcal{E}_T(\cdot)$。我们比较了使用我们选择的 VLM（SigLIP2 ViT-B/16）[43] 中的文本编码器与一款单模态文本编码器，即句子转换器 [45]。与之前研究的发现 [30] 一致，建议在语义相似任务中使用单模态编码器具有优势，在 $f_{\text{ret}}(\cdot,\cdot)$ 中应用句子转换器在所有指标中均取得了更高的得分。超参数，τ，K 和 K_r。最后，我们检查 *SynC* 对关键超参数的敏感性。对于保持比例 τ，我们通过在 COCO 测试集中增加每 10 % 从 10 % 到 100 % 的训练数据来测试。结果如 Fig. 3 所示，表明仅使用 30 % 的训练数据集即在 *Baseline* 的性能上占优。最佳性能出现在 $\tau = 0.9$。我们评估了 K 和 K_r 的各种设置，结果如 Fig. 4，表明 *SynC* 在不同设置中表现稳健。配置 $K_r = 2$ 和 $K = 15$ 显示了整体最佳性能。

这项工作解决了 T2I 生成的合成数据集中语义错位的关键挑战，这限制了零样本文本生成（ZIC）的性能。我们引入了 *SynC*，一个新颖的框架，采用一对多映射策略 $S_{\text{T2I}}(\cdot)$ 和多模态评分函数 $f_{\text{ret}}(\cdot,\cdot)$ 来优化嘈杂的合成图像-文本配对数据集。我们的大量实验验证了 *SynC*，展示了在各种 ZIC 模型和基准测试中显著且一致的性能提升，包括达到最新的结果。*SynC* 提供了一种实用的方法来提高合成数据在 ZIC 中的质量和实用性。其原理也有可能在更广泛的应用中实现，并且未来的工作将研究将 *SynC* 扩展到其他视觉与语言的下游任务（例如，分割 [18]，视觉问答 [8]）以及采用先进的图像生成模型 [4, 34]。

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8948–8957.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer, 382–398.

[3] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.

[4] SeungJu Cha, Kwanyoung Lee, Ye-Chan Kim, Hyunwoo Oh, and Dong-Jin Kim. 2025. VerbDiff: Text-Only Diffusion Models with Enhanced Interaction Awareness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 8041–8050.

[5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3558–3568.

[6] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

[7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/

[8] Jae Won Cho, Dong-Jin Kim, Hyeonggon Ryu, and In So Kweon. 2023. Generative bias for robust visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11681–11690.

[9] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems* 36 (2023), 35544–35575.

[10] Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. 2023. Transferable decoding with visual entities for zero-shot image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3136–3146.

[11] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4125–4134.

[12] Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. 2024. SynthCLIP: Are We Ready for a Fully Synthetic CLIP Training? *arXiv preprint arXiv:2402.01832* (2024).

[13] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 7514–7528.

[14] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2023. Model-agnostic gender debiased image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15191–15200.

[15] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[16] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. 2019. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. *arXiv preprint arXiv:1909.02201* (2019).

[17] Dong-Jin Kim, Tae-Hyun Oh, Jinsoo Choi, and In So Kweon. 2024. Semi-supervised image captioning by adversarially propagating labeled data. *IEEE Access* 12 (2024), 93580–93592.

[18] Ye-Chan Kim, SeungJu Cha, Si-Woo Kim, Taewhan Kim, and Dong-Jin Kim. 2025. SIDA: Synthetic Image Driven Zero-shot Domain Adaptation. In *Proceedings of the 33rd ACM International Conference on Multimedia*.

[19] Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, et al. 2024. Veclip: Improving clip training via visual-enriched captions. In *European Conference on Computer Vision*. Springer, 111–127.

[20] Soeun Lee, Si-Woo Kim, Taewhan Kim, and Dong-Jin Kim. 2024. IFCap: Image-like Retrieval and Frequency-based Entity Filtering for Zero-shot Captioning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 20715–20727. https://aclanthology.org/2024.emnlp-main.1153

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[22] Tianhong Li, Sangnie Bhardwaj, Yonglong Tian, Han Zhang, Jarred Barber, Dina Katabi, Guillaume Lajoie, Huiwen Chang, and Dilip Krishnan. 2023. Leveraging unpaired data for vision-language generative models via cycle consistency. *arXiv preprint arXiv:2310.03734* (2023).

[23] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. 2023. Decap: Decoding clip latents for zero-shot captioning via text-only training. *arXiv preprint arXiv:2303.03032* (2023).

[24] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[25] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26296–26306.

[26] Zhiyue Liu, Jinyuan Liu, and Fanrong Ma. 2024. Improving Cross-Modal Alignment with Synthetic Pairs for Text-Only Image Captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 3864–3872.

[27] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems* 35 (2022), 5775–5787.

[28] Jianjie Luo, Jingwen Chen, Yehao Li, Yingwei Pan, Jianlin Feng, Hongyang Chao, and Ting Yao. 2024. Unleashing Text-to-Image Diffusion Prior for Zero-Shot Image Captioning. In *European Conference on Computer Vision (ECCV)*.

[29] Feipeng Ma, Yizhou Zhou, Fengyun Rao, Yueyi Zhang, and Xiaoyan Sun. 2024. Image captioning with multi-context synthetic data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4089–4097.

[30] Anas Mahmoud, Mostafa Elhoushi, Amro Abbas, Yu Yang, Newsha Ardalani, Hugh Leather, and Ari S Morcos. 2024. Sieve: Multimodal dataset pruning using image captioning models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22423–22432.

[31] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. 2023. Verbs in Action: Improving Verb Understanding in Video-Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 15579–15591.

[32] Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. 2023. Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems* 36 (2023), 22047–22069.

[33] David Nukrai, Ron Mokady, and Amir Globerson. 2022. Text-Only Training for Image Captioning using Noise-Injected CLIP. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 4055–4063.

[34] Hyunwoo Oh, SeungJu Cha, Kwanyoung Lee, Si-Woo Kim, and Dong-Jin Kim. 2025. CatchPhrase: EXPrompt-Guided Encoder Adaptation for Audio-to-Image Generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*.

[35] Youngtaek Oh, Jae Won Cho, Dong-Jin Kim, In So Kweon, and Junmo Kim. 2024. Preserving Multi-Modal Capabilities of Pre-trained VLMs for Improving Vision-Linguistic Compositionality. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 19060–19076.

[36] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[38] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*. Pmlr, 8821–8831.

[39] N Reimers. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084* (2019).

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[41] Mert Bülent Sarıyıldız, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. 2023. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8011–8021.

[42] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2556–2565.

[43] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786* (2025).

[44] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.

[45] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems* 33 (2020), 5776–5788.

[46] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 38–45.

[47] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.

[48] Jiarui Yu, Haoran Li, Yanbin Hao, Bin Zhu, Tong Xu, and Xiangnan He. 2023. CgT-GAN: clip-guided text GAN for image captioning. In *Proceedings of the 31st ACM International Conference on Multimedia*. 2252–2263.

[49] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why Vision-Language Models behave like Bags-of-Words, and what to do about it?. In *International Conference on Learning Representations*. https://openreview.net/forum?id=KRLUvxh8uaX

[50] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11975–11986.

[51] Rui Zhao, Weijia Mao, and Mike Zheng Shou. 2025. DoraCycle: Domain-Oriented Adaptation of Unified Generative Model in Multimodal Cycles. *arXiv preprint arXiv:2503.03651* (2025).