PolyGen：通过多生成器集成实现全合成视觉-语言训练

Leonardo Brusini[1], CristianSbrolli[*1]、Eugenio Lomurno[1], Toshihiko Yamasaki[2],和 Matteo Matteucci[1]

[1]米兰理工大学，*AIRLab*，意大利[2]东京大学，计算机视觉与媒体实验室，日本

摘要

合成数据为视觉-语言预训练提供了一种可扩展的解决方案，但当前最先进的方法通常依赖于扩展单一生成式骨干网络，这会带来特定于生成器的频谱偏差，并限制特征多样性。在本文中，我们提出***PolyGen***，一种通过优先考虑流形覆盖和组合严谨性，而非单纯扩大数据集尺寸，来重新定义合成数据构建的框架。*PolyGen*采用多体方法，在架构各异的生成器交集上进行训练，从而有效地消除模型特定伪影的影响。此外，我们提出一种程序化困难负样本课程，以强化对细粒度句法的理解。通过将相同的数据预算从独特图像描述结构性地重新分配到多源变体，*PolyGen*获得了更稳健的特征空间，在汇总多任务基准上的表现超过领先的单源基线（*SynthCLIP*） *+19.0%*，并在*SugarCrepe++* 组合性基准上超过*+9.1%*。这些结果表明，与单纯增加单源样本数量相比，结构多样性是一种数据效率更高的缩放定律。

关键词： 视觉-语言模型 · 合成数据 · 对比学习 · 生成式人工智能 · 难负样本挖掘 ·集成学习

arXiv:2602.01370v1 [cs.CV] 1 Feb 2026

1 引言

人工智能领域的一项重大突破，是视觉-语言模型（VLMs）的出现。这类模型能够在共享嵌入空间中对齐视觉表征与文本表征。CLIP[27] 和 ALIGN [13]等模型解锁了卓越的零样本能力，有效地摆脱了视觉识别对固定标签集的依赖。然而，这些模型的效能建立在一个日益岌岌可危的基础之上：由数十亿图文对构成的海量网络规模数据集。

对网络抓取数据的依赖带来了两项生存性挑战。首先，近期研究记录了一些令人担忧的趋势，表明高质量人工生成数据的供应可能会成为未来扩展的瓶颈 [35]。其次，“真实世界”数据集本质上噪声较多，充斥着错位、社会偏见以及隐私或版权问题，需要耗费大量资源进行整理。一个颇具吸引力的替代方案是转向完全合成数据 [10, 20, 21, 33]。鉴于现代文本到图像（T2I）扩散模型的保真度，理论上可以按需生成规模可无限扩展、安全且可控的数据集 [8]。

尽管前景可期，持续存在的“合成差距”仍未消除：仅使用合成数据训练的模型，表现始终逊于其真实数据对应模型，在泛化任务上尤其如此。Fan 等人近期的研究 [8] 指出了一个根本原因：单生成器训练固有地存在多样性与可识别性 之间的权衡。此外，实证证据表明——即便是使用相同数据训练的同构模型集成，也能胜过单模型方法，这揭示了单个生成器受限于流形空间中的局部最优解 [17, 19]。

这种权衡的表现因架构而异。较早的扩散模型（例如 Stable Diffusion 1.5）能产生较高的语义方差，但会受到感知伪影和频域特征的影响；而较新的模型（例如 SDXL-Turbo）则能实现更高的照片级真实感，代价是输出多样性降低。若仅使用单一生成器生成的图像进行训练，视觉语言模型不可避免地会依赖这些低层次、特定于生成器的线索，例如纹理模式或谱指纹 [3], ，而非开放世界泛化所必需的抽象语义特征。

为应对这些挑战，我们推出了**PolyGen**（Polylithic生成框架），这是一个用于在完全由合成数据构成的数据集上训练视觉-语言模型的端到端流程。我们的方法以生成器不变性假设为基础：通过使用从多个架构各异的生成器中采样的图像进行训练，我们迫使模型学习不受生成器特定伪影影响的表征，从而有效消除虚假相关性，并将语义信号作为学习目标单独提取出来。

我们调整了方法，使其也能生成多个困难负样本，从而解决对比学习的另一项局限：标准训练目标往往会退化为“词袋”匹配，模型通过简单的共现统计来识别对象，而非进行细粒度的组合式理解 [41]。

最后，PolyGen 将生成器不变性

*通讯作者：cristian.sbrolli@polimi.it

假设通过三种互补机制付诸实践：

1. **多正样本集成：** 我们通过使用异构池（SD 1.5、SD 2、SDXL-Turbo、SANA），为每个图像描述生成 n^+ 种多样化的视觉呈现，从而打破单一生成器的权衡。通过将这些架构各异的输出视为正样本对，并采用多正样本目标，我们迫使模型学习对特定频谱伪影不变、但对共享语义内容敏感的表征。

2. **程序化困难负样本：** 我们利用大语言模型为基础描述构造结构化且可控的语义扰动（反事实样本）。使用这些困难负样本进行训练，迫使模型摒弃“词袋”捷径，转而进行细粒度的组合式理解，从而区分不同的语义与组合绑定关系。

3. **课程调度器：** 我们在训练过程中逐步提高困难负样本的比例，使模型能够先建立粗粒度的语义结构，再完善判别边界。

实证表明，PolyGen 证明了平衡合成数据的多样性与质量至关重要。通过设计训练分布，而非仅仅对其进行扩展，我们在 Delta 多任务学习方面较强大的单源基线实现了 **+19.0%** 的提升，缩小了与真实数据性能之间的差距，并为可持续的大规模视觉-语言训练提供了一条可行路径。

2 相关工作

混合合成策略。 一类主流研究通过利用真实世界字幕生成合成图像，缓解数据稀缺问题。Sta- bleRep [32] 证明，将由同一图注生成的多张合成图像视为正样本对的多正样本对比学习，其性能可以优于使用真实数据训练的模型。LaCLIP [7] 和 VeCLIP [16]则利用大语言模型（LLM）改写图像描述，在保持原始图文基础对齐的同时，提升对齐度和多样性。然而，这些“混合”方法仍受制于网络抓取数据的限制，继承了其中的语言/视觉偏差、隐私风险和有限规模。

全合成流程。 近期研究尝试通过同时生成图像描述和图像[33]来实现闭环。Fan 等人 [8] 提出了合成数据的缩放定律，表明合成预训练可与真实世界基线相媲美。一项里程碑式研究 SynthCLIP [10], 引入了端到端流程，使用大语言模型生成的图像描述和 Stable Diffusion 图像，从头开始训练 CLIP 模型，完全不使用人工数据。近期研究进一步表明，合成数据在图像分类任务中能够达到与真实数据集相当的效果，同时提供隐私保障 [28], ，但这一结果仍局限于单任务监督学习。尽管这些方法具有开创性，但它们存在架构同质性问题：通常依赖单一生成式模型，且往往依赖过时的架构（例如 Stable Diffusion v1.5），导致下游模型对生成器特定伪影过拟合 [3]。我们认为，生成器多样性的缺乏是视觉语言任务中“合成差距”持续存在的主要原因之一，因为模型学会利用低级纹理模式，而非抽象语义特征。

难负样本挖掘。 对比式视觉-语言模型的一个众所周知的局限，是它们往往依赖词汇共现模式，而非真正的组合推理 [41]。ARO 和 SugarCrepe [12] 等基准通过要求模型完成属性绑定、空间推理和物体关系理解等任务，系统地评估这一失效模式；结果表明，即使是最先进的模型，也存在显著的性能差距。为增强判别能力，难负样本挖掘已从重要性加权 [26, 29], 发展到通过基于规则的替换 [41]和大语言模型驱动改写 [25]生成困难负样本。然而，现有方法要么依赖外部解析器，要么对非结构化文本进行不受约束的大语言模型生成，因而存在语义漂移的风险。尽管已有研究提出内在数据挖掘 [36], ，但其计算成本仍然高昂，且受静态数据分布限制。我们提出一种构造式替代方案：先合成具有结构依据的基础描述，再通过编程方式引导大语言模型扰动特定语义轴，从而确保生成更安全、保真度更高的反事实样本。

3 方法

我们提出 **PolyGen** （Polylithic GenerationFramework），这是一个端到端流程，用于使用完全合成的数据训练视觉-语言模型。该框架分为两个阶段（图 1）：*Polylithic*合成将结构化图像描述生成与多生成器图像渲染相结合，随后进行基于课程学习的对比训练。下面将详细介绍各个组成部分。

3.1 结构化描述对

*Polylithic*合成 的第一个组成部分（阶段 1，图 1）会生成经过语义控制的标题对。以往的合成数据生成工作通常依赖“创意型”提示词，在缺乏语义控制的情况下生成描述性图像描述。我们则采用构造式方法，通过受控语义修改生成训练对（T^+，T^-）。

属性条件描述生成。 我们使用 Mistral-V0.2-7B [14] 生成基础描述 T^+，并以控制元组 ($c \in \mathcal{C}$，$a \in \mathcal{A}$) 为条件进行生成。这里，c 是基础概念，即从 MetaCLIP概念库 [39], 中采样得到的主要主体；a则是生成困难负样本所沿用的语义修改轴 。 $\mathcal{A}$包含视觉属性（照明、材质、透视、颜色、背景、位置、风格，在附录 E 中进行了消融），以及用于“主体更换”负样本的显式概念轴。当扩展生成导致多次使用相同概念时，额外的属性指示也有助于提高多样性。（完整的生成提示词见附录 A）。

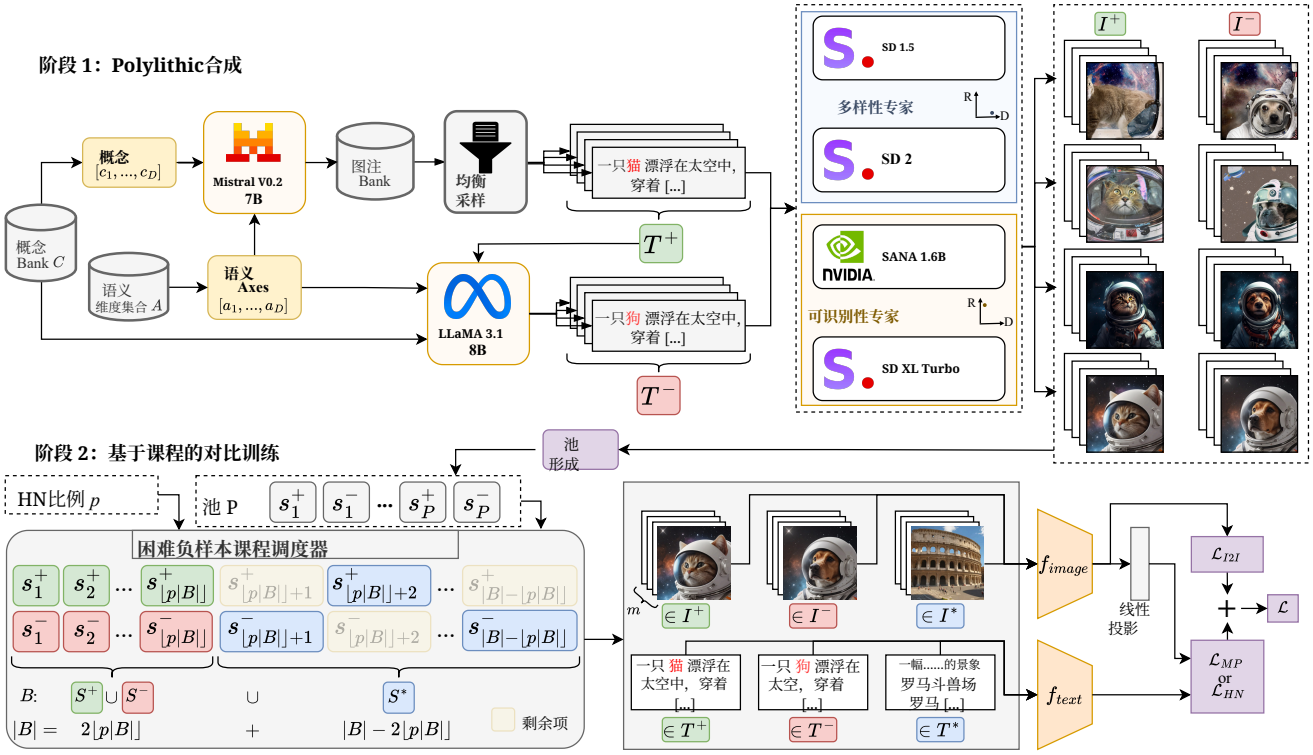


图 1：PolyGen 流程概览。阶段 *1*（*Polylithic*合成）：MetaCLIP 库中的概念和语义轴输入双大语言模型系统。Mistral V0.2 根据概念-属性对生成基础描述 T^+，而 Llama 3.1 通过修改特定语义维度生成困难负样本 T^-。图像描述对由四种扩散模型生成：多样性专家（SD 1.5、SD 2）尽管保真度较低，仍能确保广泛的流形覆盖；可识别性专家（SANA、SDXL-Turbo）虽变化有限，却能提供较高的提示词遵循度。阶段 *2*（基于课程的训练）：调度器在训练过程中将困难负样本的比例 p 从 0 线性提高至 0.5。未使用的样本会放入剩余队列，以维持数据效率。批次结合多正样本（I^+，T^+）和困难负样本（I^-，T^-），以计算 $\mathcal{L}_{I2I}$ 和 $\mathcal{L}_{HN}$。

通过 MetaCLIP 的均衡采样算法（附录 H）对更大的图像描述池进行下采样后，即可获得 T^+。

困难负样本生成。得益于结构化的基础描述，我们可以利用大语言模型生成可控的负样本。我们使用 Llama-3.1-8B [5]（在第 4.5 节中进行消融）来生成反事实困难负样本 T^-。模型接收 $t \in T^+$、相应的语义轴 a 和概念 c，并遵循严格指令，仅修改指定的语义轴，同时保留句法结构。如图 3 所示，该策略确保了全局语义连贯性。与僵化的程序化替换不同，后者将词语孤立地修改，可能导致逻辑荒谬；大语言模型会动态调整上下文依赖属性，使其与新概念相符。

3.2 多生成器合成集成

第一阶段的第二个组成部分是通过多个扩散模型渲染每组图注对。合成训练的一项核心挑战是确定最优生成器$\mathcal{G}$。依赖单一模型可能会导致模型学到有偏的数据流形表征：每个生成器只能捕捉视觉语义谱的一部分，并引入架构特定频率模式和压缩伪影，使下游模型可能将其当作虚假的特征利用，而非学习稳健的语义内容。

指标驱动的选择。为了系统地评估候选生成器，我们建立了一个定量框架，用于衡量开放词汇对比设置中合成数据集的两种互补属性。Fan 等人 [8] 为监督分类数据集（具有固定标签集）定义了可识别性 和多样性，我们则针对 CLIP（自监督）范式对这些指标进行调整：

- 可识别性（保真度）。在没有类别标签的情况下，我们衡量生成图像x 与其条件标题 t之间的对齐。我们将可识别性定义为数据集 $\mathcal{D}$上的期望 CLIPScore [11]：

$$R(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(x,t)\in\mathcal{D}} \text{CLIPScore}(f_{img}(x), f_{txt}(t)) \quad (1)$$

高可识别性表明，生成器能够可靠地遵循提示词的语义约束。

- 多样性：由于视觉空间的高维性，以及何为有意义变化所具有的语义相对性，衡量开放世界数据中的多样性具有挑战性。我们提出一种能够捕捉概念内方差的基于聚类的方法。我们基于图注嵌入将数据集划分为 $K = 1000$ 个语义聚类，然后将多样性定义为每个聚类 k 内图像嵌入的平均标准差：

$$D(\mathcal{D}) = \frac{1}{K} \sum_{k=1}^{K} \sqrt{\frac{1}{|\mathcal{D}_k|} \sum_{x\in\mathcal{D}_k} ||f_{img}(x) - \mu_k||^2} \quad (2)$$

其中，μ_k 是簇 k 中图像嵌入的质心。

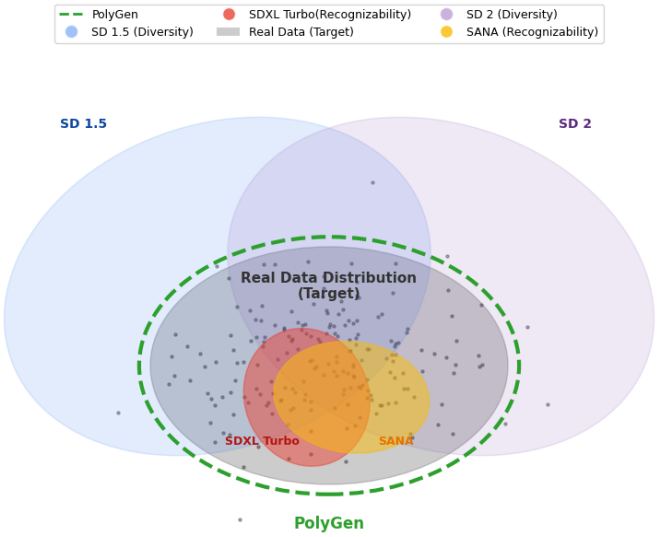


图 2：生成器不变性假设的概念示意图。单个生成器（彩色圆圈）充当具有特定偏差的“专家”：较旧的模型（如 SD 1.5（蓝色））多样性高，但保真度低；而现代蒸馏模型（如 SDXLT（红色））保真度高，却遭受模式崩溃。PolyGen（绿色虚线）在这些流形的并集上进行训练，通过消除生成器特定伪影的影响，近似覆盖真实数据分布（灰色）。

该指标专门衡量生成器为同一语义概念或概念组生成视觉上各不相同的诠释的能力。

生成器不变性策略。在生成器不变性假设（图 2）的指导下，我们构建了一个集成 $\mathcal{G}$，以最大限度地覆盖视觉-语义谱系。我们选择了四种架构各异的模型：

- 可识别性专家：SDXL-Turbo [31] 和 SANA- 1.6B [38], 因提示词遵循度高且具有照片级真实感而入选，尽管其输出多样性较低。
- 多样性专家：选择 Stable Diffusion v1.5 和 v2 [30], , 以确保较高的方差和广泛的流形覆盖，尽管保真度较低。

对于每一对图像描述（t^+，t^-），我们使用全部四个模型生成对应的图像。这迫使下游模型学习不受生成器特有频谱特征影响的表征，例如 SD1.5 中的压缩伪影或 SDXL-Turbo 中的低频偏差（附录 F）；因为集成中唯一一致的信号就是语义概念本身。

3.3 基于课程的对比训练

阶段 2 利用结构化数据集，通过专门的训练流程发挥图像描述到图像的一对多映射关系（每条图像描述最多对应 $n^+ = 4$ 张正样本图像）。

多正样本目标。为了利用针对每条图注 t_i 生成的 n^+ 张图像集成，我们采用多正样本目标 [32]。我们不使用独热目标，而是使用软目标，将概率质量均匀分配 $(1/n^+)$ 至来自同一提示词的所有图像。该目标 $\mathcal{L}_{MP}$ 使文本嵌入与生成器集成所形成的多样化视觉流形的语义质心对齐。除了文本到图像的对齐，我们还评估了显式图像到图像正则化的影响。我们加入对比项 $\mathcal{L}_{I2I}$，以惩罚模型学习生成器特有的个性化特征。对于任意图像锚点 i，我们定义一个真实分布 p_i，该分布在同一图注生成的其他图像上均匀分布（$1/(n^+ - 1)$）。给定预测的类别分布：

$$q_{i,j} = \frac{e^{(s(i,j)/\tau)}}{\sum_{k \in I \setminus \{i\}} e^{s(i,k)/\tau)}}, \tag{3}$$

损失按交叉熵计算：

$$\mathcal{L}_{I2I} = -\frac{1}{N} \sum_{i \in I} \left(\sum_{j \in I \setminus \{i\}} p_{i,j} \log q_{i,j} \right). \tag{4}$$

这一项充当结构正则化项，迫使视觉编码器舍弃频谱伪影或风格偏差等“伪相关”特征，并使表征趋近于提示词所定义的不变语义核心。

困难负样本目标。标准 InfoNCE 会将困难负样本 T^- 视作批次分母中的其他任意样本，从而抑制它们。为实现细粒度判别，我们采用 TripletCLIP 损失 [25]。

我们定义图像到文本的困难负样本 NCE 损失如下：

$$\mathcal{L}_{NCE}^{I^+ \to T^+; T^-} = -\frac{1}{N} \sum_{k=1}^{N} \log \frac{e^{s(i_k, t_i)/\tau}}{\sum_{t \in T^+ \cup T^-} e^{s(i_k, t)/\tau}}, \tag{5}$$

文本到图像的情况类似。NegCLIP 损失为：

$$\mathcal{L}_{NegCLIP}^{I^+, T^+, T^-} = \mathcal{L}_{NCE}^{I^+ \to T^+; T^-} + \mathcal{L}_{NCE}^{T^+ \to I^+}. \tag{6}$$

最后，TripletCLIP 总损失定义如下：

$$\mathcal{L}_{TripletCLIP} = \mathcal{L}_{NegCLIP}^{I^+, T^+, T^-} + \mathcal{L}_{NegCLIP}^{I^-, T^-, T^+}. \tag{7}$$

这种公式迫使模型明确区分基础概念与其语义反事实。

困难负样本课程调度器。朴素地注入困难负样本会导致训练初期不稳定，因为细粒度验证任务会干扰粗粒度语义对齐，妨碍形成稳定的概念簇。我们引入一种课程调度器，在训练轮次中线性提高批内困难负样本比例 p，从 $0 \to 0.5$ 开始。为确保整个轮次中充分利用数据，我们实现了一个 剩余队列：未与相应困难负样本配对的样本会进入队列，并在后续步数中组成批次。该调度器与逐渐衰减的学习率结合使用时，可让模型先学习更稳健的知识基线，再在训练后期利用困难负样本学习细粒度判别。由于批次组成会发生变化，我们调整 $\mathcal{L}_{HN}$ 以平衡各子集的贡献：

$$\mathcal{L}_{HN} = \frac{1}{|B|} (|T| \mathcal{L}_{NegCLIP}^{I,T,T^-} + |T^-| \mathcal{L}_{NegCLIP}^{I^-,T^-,T}), \tag{8}$$

以及 $T = T^* \cup T^+$、$I = I^* \cup I^+$ 和 $|B| = |T| + |T^-|$。



图 3：多生成器集成中的困难负样本生成。通过四种扩散模型呈现的两个受控语义扰动示例。上： 概念级修改（“奥利奥” → “巧克力曲奇”， $a =$ 概念）。下： 视角变换（“鸟瞰视角” → “人行道视角”， $a =$透视）。多样性专家（SD 1.5、SD 2）能产生较高的概念内方差，但照片级真实感较低；可识别性专家（SANA、SDXL-Turbo）则能生成照片级真实感的输出，但风格变化较少。架构异质性迫使模型学习生成器不变语义表示，因为各列之间唯一一致的信号是基础概念修改。

4 实验与结果

4.1 实验设置

训练配置。为确保与先前最先进的 SynthCLIP 进行严格比较， [10], 我们在从 CC3M 数据集中采样的 50 万条图注构成的受控子集上进行主要评估。我们从头开始训练 ViT-B/16 CLIP 架构，共训练 40 个 epoch。我们使用 AdamW 优化器，学习率为 5×10^{-4}，并采用包含 1 个预热 epoch 的余弦衰减调度。对于 PolyGen，我们使用 1536 的批量尺寸。为公平处理多正样本输入，我们固定训练图像数量，并相应调整图注数量：这对应于 $500k/n^+$ 条图注（在 n^+ 个正样本的情况下）。困难负样本比例 p 遵循我们的课程学习调度器，在整个训练过程中从 0 线性增加至 0.5。

下游任务 我们采用 SynthCLIP 的评估设置，在纯视觉任务和视觉语言任务中评估模型。纯视觉任务用于评估图像编码器提取特征的质量（即线性探测、少样本分类）；视觉语言任务则包括零样本检索和分类。我们还沿用 SynthCLIP 的做法，将结果汇总为一个综合 Δ_{MTL} 指标 [34] （附录 I），该指标综合了视觉任务的 Top-1 准确率、零样本分类的 Top-5 准确率以及检索的 Recall@5，用以表示相对于基线模型的性能下降或提升；各任务的完整结果见附录 C。

数据集。 这些模型在一系列全面的下游基准测试上进行评估，以考察其纯视觉和视觉-语言能力：在线性探测和少样本评估中，我们使用：CIFAR10 和 CIFAR100[15], FGVC Aircraft [22], DTD（纹理） [2], Oxford Flowers [23],Oxford-IIIT Pets [24], SUN397 [37], Caltech-101[9], 食物-101 [1]。我们还使用 ImageNet [4]进行零样本分类（准确率），使用 Flickr8k 和 Flickr30k [40], 以及 MS COCO [18] 进行文本到图像和图像到文本检索。最后，在困难负样本任务中，我们使用 SugarCrepe [12] 和 SugarCrepe++ [6]。本节按任务报告结果，即各数据集的平均值。各数据集的结果见附录 C。

4.2 可识别性与多样性的影响

图 4 展示了训练数据的分布属性（可识别性与多样性）和下游性能（以颜色表示）之间的关系。为单独考察流形覆盖的影响，我们在固定的 50 万组图文对上训练一系列 CLIP 模型（$n^+ = 1$）。对于集成实验（$m > 1$），每条图注对应的图像均从所选生成器池中均匀采样。这确保任何性能提升都归因于来源间的多样性，而非训练样本数量的增加。图中呈现出生成行为的明显两极分化。较早的架构，特别是 Stable Diffusion (SD) v1.5，处于多样性较高、可识别性较低的区域。相比之下，SDXL-Turbo 和 SANA-1.6B 等现代蒸馏模型的可识别性显著更高，但概念内多样性也因此明显降低。这一现象

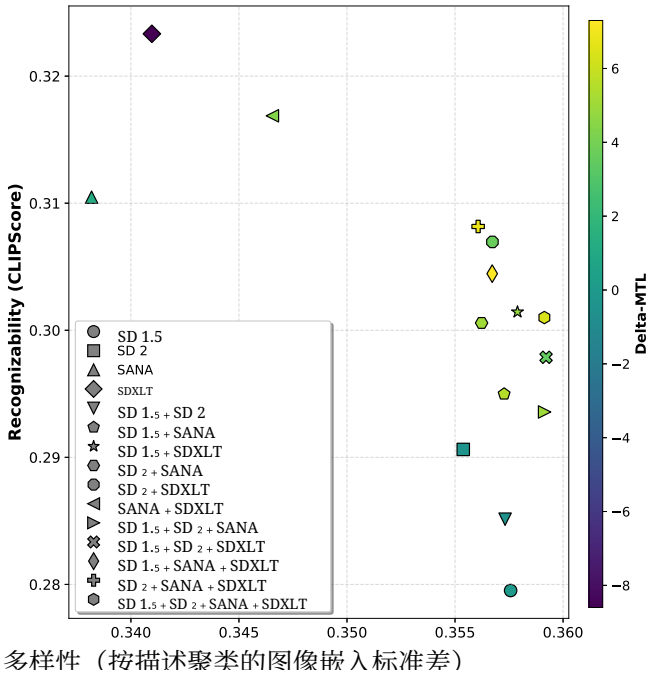


图 4： 可识别性与多样性的关系，颜色表示下游性能。尽管单个模型需要在多样性和可识别性之间取舍，PolyGen 集成却突破了这一帕累托前沿，在两个维度上均实现最大化，从而获得更优的 Δ_{MTL}。

n^+	m	$\mathcal{L}_{I2I}$	*L.P.*	*Few-shot*	*Img Ret.*	*Text Ret.*	*0-shot*	Δ_{MTL} 对比 SynthCLIP	Δ_{MTL} 对比 相同的 n^+, $m=1$
1	1	✗	49.1	61.6	11.6	15.9	8.15	-	-
1	2	✗	47.5	61.1	12.1	17.5	8.53	+3.1%	+3.1%
1	3	✗	48.5	61.3	13.3	17.9	8.56	+6.1%	+6.1%
1	4	✗	48.9	61.5	12.8	17.6	8.92	+6.2%	+6.2%
2	1	✗	49.5	62.4	13.8	18.7	9.41	+11.0%	-
2	2	✗	47.9	61.1	14.2	20.2	9.54	+12.7%	+1.3%
3	1	✗	49.3	62.7	14.7	21.1	9.66	+16.1%	-
3	3	✗	48.2	61.1	16.2	**22.4**	9.38	+18.7%	+1.7%
4	1	✗	**49.8**	**62.8**	14.8	19.4	9.56	+14.1%	-
4	4	✗	48.9	62.0	16.1	21.4	9.86	**+19.0%**	+3.8%
2	1	✓	43.4	57.3	13.3	17.6	10.14	+6.3%	-
2	2	✓	41.5	55.7	14.4	19.8	11.19	+12.3%	+4.8%
3	1	✓	44.4	57.3	14.2	18.9	10.88	+11.8%	-
3	3	✓	41.3	55.2	**16.5**	21.7	10.98	+17.4%	+4.1%
4	1	✓	45.2	57.7	15.0	19.9	11.24	+15.6%	-
4	4	✓	42.9	56.4	15.7	20.6	**11.34**	+16.6%	+0.4 %

表 1： 异质性（m）与密度（n^+）的平均结果。联合扩展 m 和 n^+ 可使 Δ_{MTL}最大化。值得注意的是，强不变性（$m=4$, $\mathcal{L}_{I2I}$）提升了抽象推理能力（零样本），但代价是基于纹理的分类性能下降（线性探测：线性探测）。

体现了高保真度生成器中的“模式崩溃”：模型收敛到一组有限的高概率视觉解读，以确保审美吸引力。我们的方法有效地将训练分布推向右上象限，从而提升了下游任务的性能。这一结果通过实证验证了我们的生成器不变性假设：尽管单个生成器受其特定频谱偏差的限制，集成却能消除这些偏差的影响。通过迫使模型学习在不相交流形间保持一致的特征，我们实现了稳健的泛化（Δ_{MTL}），这是任何一种生成器单独使用都无法达到的。

4.3 异质性-密度权衡

我们分别考察两个核心设计变量的影响：集成规模（m）和正样本密度（n^+）。表 1 展示了结果，依次从 $m=1$（SD 1.5）开始，$m=2$ 包括 SDXLT，$m=3$ 再加入 SANA，并以 $m=4$ 表示完整的 PolyGen 集成。

异质性增益 ($m\uparrow$)。将集成尺寸从单一来源 ($m=1$) 扩大到完整集成 ($m=4$)，在所有配置中都能普遍提升 Δ_{MTL}。关键在于，这一增益由视觉-语言任务驱动。例如，在 $n^+=1$ 模块中，零样本准确率从 8.15 ($m=1$) 上升至 8.92 ($m=4$)。这证实了生成器不变性假设：让模型看到同一概念由不同引擎生成的结果，会迫使模型将语义与生成器特定伪影区分开来，从而形成更稳健的开放世界表征。

密度增益 ($n^+\uparrow$)。 多样性提升泛化能力，而密度则增强鲁棒性。增加正样本对的数量（n^+）可在两个维度（纯视觉和视觉-语言）上提升性能。通过聚类单个图注周围的多张图像，模型能够学习到稳定的“质心”表征。两种因素结合后会产生乘法效应：完整的 PolyGen 配置（$n^+=4$, $m=4$）达到 Δ_{MTL} 的峰值 +19.0%，显著优于仅依赖密度的方法（$n^+=4$, $m=1$, +14.1%）。

不变性作为正则化项。 分析显式不变性机制的影响时，我们发现了一个值得关注的规律：图像到图像损失（$\mathcal{L}_{I\to I}$）和集成规模（m）。虽然增大 m 或添加 $\mathcal{L}_{I\to I}$可使零样本准确率最大化（后者达到表中的最高值），但这会导致线性探测性能下降。例如，单一来源模型的探测得分往往略高于其集成模型（$m=4$）。我们认为这是由该框架的“语义纯粹主义者”特性所致。简单的线性分类器往往会利用低级纹理捷径或生成器特定的频谱伪影来区分类别。

通过强制模型在多样流形之间保持不变性（借助多正样本），或显式惩罚特征差异（借助 $\mathcal{L}_{I\to I}$），PolyGen 去除了这些非语义特征。虽然这会消除线性探针所利用的纹理线索，却促使模型学习能够更好迁移到开放世界任务的稳健抽象表示。最终，即使存在这种权衡，完整的 PolyGen 配置仍能带来最高的 Δ_{MTL} 提升，提供了一种帕累托最优解：优先实现语义对齐——这是视觉-语言预训练的首要目标，同时保持具有竞争力的视觉表征。

4.4 困难负样本与课程学习

表2对训练动态进行了细致剖析，尤其是 $m=2, n^+=2$ 模块中的动态。

被动与延迟学习的局限。我们首先

n^+	m	*HN samples*	$\mathcal{L}_{HN}$	*scheduler*	*L.P.*	*Few-shot*	*Img Ret.*	*Text Ret.*	*0-shot*	Δ_{MTL}	SugarCrepe	SugarCrepe++	Δ^{HN}_{MTL}
1	1	✗	✗	✗	**49.1**	**61.6**	11.6	15.9	8.15	-	55.4	36.7	-
2	2	✓	✗	✗	41.5	55.7	13.2	17.9	9.95	+4.7%	56.9	40.0	+5.9%
2	2	✓	✗	✓	40.8	55.8	13.6	18.8	10.23	+7.0%	55.0	38.8	+2.6%
2	2	✓	✓	✗	42.3	56.8	14.0	19.0	10.27	+9.0%	55.8	39.2	+3.7%
2	2	✓	✓	✓	41.4	55.8	14.6	**20.5**	10.8	+12.5%	56.4	40.5	+6.1%
3	3	✓	✓	✓	42.7	56.5	**15.7**	20.4	**11.1**	**+15.5%**	56.8	40.7	+6.8%
4	4	✓	✓	✓	42.7	56.5	**15.7**	20.4	**11.1**	**+15.5%**	56.8	40.7	+6.8%

表2： 使用困难负例的模型在各任务上的平均结果。从第二行到最后一行，模型使用 $\mathcal{L}_{I2I}$。启用时， $\mathcal{L}_{HN}$ 会替换 $\mathcal{L}_{MP}$。

通过采用“朴素”的困难负样本建立基线表现，即仅将反事实图像描述加入批次，而不设置针对性目标。这种方法带来了出色的组合表现（SugarCrepe++ 40.0%），因为模型在整个训练过程中都接触到了细粒度变化。然而，这种早期的复杂性会干扰整体语义对齐，因此，与上一节中更简洁的训练相比，标准任务上的提升较为有限（Δ_{MTL} +4.7%）。仅依靠课程调度器来解决这一问题，会引入相反的问题。通过延后引入困难负样本，我们保护了训练早期阶段，使标准任务的表现得以提升（Δ_{MTL} +7.0%）。然而，这种“延后接触”是以牺牲组合推理能力为代价的：模型接触困难反事实样本的频率不足，无法学会属性绑定的细微语法。

困难负样本损失的影响。单独使用$\mathcal{L}_{HN}$ 的效果已得到验证，其 Δ_{MTL}高于采用 $\mathcal{L}_{MP}$的相同配置。我们认为，这归因于TripletCLIP 损失的一项关键变化；$\mathcal{L}_{HN}$ 基于该损失以及标准 CLIP。

由于生成图像往往无法正确呈现图像描述中的细微差异（附录D.3），TripletCLIP 会从损失中移除 $T^+ \to I^-$ 和 $T^- \to I^+$ 比较。这样，模型就能继续利用困难负样本（$I^+ \to T^-$ 和 $I^- \to T^+$），同时不受困难负图像所提供的噪声更大的信号误导。CLIP 损失与 TripletCLIP 损失之间的关系在附录B中有清晰展示。另一方面，移除损失中的这些组成部分会导致组合基准测试上的性能提升未达最优。

优化协同效应。 从经验结果来看，完整的 PolyGen 配置解决了稳定性-可塑性困境。通过结合调度器和三元组损失，我们实现了 Δ_{MTL} 的 +12.5%，超过了仅使用调度器和仅使用损失的配置。这表明存在一种纠正机制：调度器可保护早期特征形成过程免受噪声梯度影响，而三元组损失则通过在引入困难负样本后施加更严格的间隔，弥补曝光时间缩短带来的影响。这种恢复使我们能够在不牺牲整体性能的情况下，达到朴素方法较高的组合得分。

扩展与帕累托前沿。 与我们的主要结果一致，扩大集成规模（m）并提高正样本密度（n^+），可使一般开放世界任务的表现持续提升，其中 $m = 4$ 配置达到Δ_{MTL} 峰值。然而，对于组合推理，我们观察到其峰值出现在 $m = 3$，表现优于 $m = 4$。我们将此归因于方差饱和点：虽然最大化多样性（$m = 4$）有助于扩大语义覆盖，但也会为细粒度句法任务引入过多视觉噪声，使模型更难在剧烈的风格变化中分辨细微的属性变化。最终， $m = 4$ 配置是帕累托最优的选择：它在最大限度提升一般迁移能力的同时，显著改善了组合推理，在语义广度与句法深度之间实现了最稳健的平衡。

模型	**准确率**	**平均F1**	**平均精确率**	**平均召回率**
Mistral V0.2	61%	60%	76%	61%
Mistral V0.3	62%	62%	77%	62%
Qwen 2.5	67%	67%	78%	67%
Falcon 7B	63%	65%	79%	63%
Llama 3.1	**74%**	**75%**	**80%**	**74%**

表 3：困难负样本的LLM选择。LLM负样本生成评估的语义轴汇总结果。

4.5 消融实验：语义合成的LLM选择

我们的流程采用了两种不同的大语言模型。对于基础图注，我们使用 Mistral-v0.2-7B，以便与 SynthCLIP 基线保持一致。相比之下，困难负样本合成更为关键，因为它要求模型能够高度忠实地遵循指令，在不改变整体场景结构的情况下修改特定语义属性。为此，我们评估了一组参数量为 7B/8B 的大语言模型，考察其生成正确负样本的能力：给定由模型针对一对（c， a）生成的正负图注对，我们验证大语言模型集成能否从该图注对中还原正确的 a。如表 3 所示，Llama-3.1-8B 显著优于其他模型，其生成的负样本分类准确率达到 74%。

5 讨论与局限性

多样性与可识别性的协同作用。 我们的结果凸显了当前生成式骨干网络用于合成训练视觉语言管线时所面临的一项根本性权衡：高多样性模型会引入噪声，削弱学习信号；而高保真模型则会导致对狭窄视觉模式的过拟合。PolyGen 通过整合这些互补的流形解决了这一问题。高保真生成器提供了强有力的“语义锚点”，确保文本概念得到正确实例化；而高多样性生成器则扩展了该概念的视觉支持集，防止模型拘泥于特定模式。这种协同作用迫使模型忽略每种架构独有的光谱指纹，并最终聚焦于唯一共享的信号：不变语义内容。

为什么选择合成训练？转向合成预训练，源于对精细筛选而非单纯收集的需求。与网页抓取不同，合成流程能够实现精准控制，从而生成隐私保护、偏差感知且可多模态扩展的数据集，

但合成数据效率较低，阻碍了其应用。PolyGen 致力于推动这一领域向前发展：造成这种“低效”的一个重要因素是单生成器同质性。我们证明，流形多样性能够显著提升单个样本的效用，并由此确立：合成预训练不仅关乎规模，还可以通过结构化的集成采样来优化效率。

局限性与未来工作。 我们的方法涉及特定的设计权衡。首先，尽管不一定会带来更高的计算成本，但维护异构集成相比单模型基线会引入更高的流水线复杂度。其次，为遵循固定预算，我们的多正样本策略优先考虑语义稳健性，而非独特的概念覆盖；在更大规模的设置下，这种平衡需要重新校准。最后，尽管当前的概念库和语义轴确保了广泛覆盖，但它仍然是对真实图像世界视觉复杂性的简化。未来的工作将沿三个维度扩展这一策略：（1）增大集成规模，从而同时增加 m 和 n^+，并采用在可识别性和多样性属性方面表现各异的架构；（2）扩展数据预算，以评估更大规模设置下的广度-深度帕累托前沿；（3）扩展程序化困难负样本，以覆盖多概念关系、计数和因果关系等复杂轴，最终获得一种能够充分支撑开放世界推理的合成课程。

6 影响声明

本文旨在推动深度学习领域的发展，具体而言，通过合成数据实现更安全、可控的视觉-语言预训练。我们采用此方法的一个主要动机，是减少对大规模网络爬取的依赖，因为这类做法本身就有可能收集到私人、未经同意或有害的内容。

为减轻有害偏见的传播，我们的流程采用了经过安全对齐和数据过滤的标准生成式骨干网络（例如 Llama 3、Stable Diffusion、SANA）。此外，我们的文本生成以 MetaCLIP概念库为基础；该库是经过整理的元数据集，旨在提供比原始网络文本更均衡、更干净的语义分布。我们承认，合成生成器仍可能反映其底层训练数据中的偏见，但我们的框架提供了一种通过程序显式过滤这些偏见的途径。鉴于本研究性质可控且方法严谨（以有限的数据预算开展概念验证），我们预计不会立即造成负面的社会影响。

致谢

本文得到 FAIR（未来人工智能研究）项目的支持。该项目由 NextGenerationEU 计划资助，属于 PNRR-PE-AI 方案（M4C2，投资项目 1.3，人工智能方向）。我们还感谢 CINECA 提供计算资源，以支持本研究的实验和训练工作。

参考文献

[1]Lukas Bossard、Matthieu Guillaumin 和 Luc Van Gool。Food-101：利用随机森林挖掘判别性组件。发表于欧洲计算机视觉会议，2014年。

[2] M. Cimpoi、S. Maji、I. Kokkinos、S. Mohamed、和 A. Vedaldi。描述野外纹理。发表于*IEEE* 计算机视觉与模式识别会议论文集（*CVPR*），2014年。

[3] Riccardo Corvi、Davide Cozzolino、Giada Zingarini、Giovanni Poggi、Koki Nagano 和 Luisa Verdoliva。关于检测由扩散模型生成的合成图像。发表于*ICASSP 2023-2023 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第1–5页。IEEE，2023年。

[4] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li和Li Fei-Fei。ImageNet：一个大规模层次化图像数据库。载于*2009 IEEE*计算机视觉与模式识别会议，第248–255页。IEEE，2009。

[5] Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Letman、Akhil Mathur、Alan Schelten、Amy Yang、Angela Fan等。Llama 3模型群。*CoRR*，2024。

[6] Sri Harsha Dumpala、Aman Jaiswal、Chandramouli Shama Sastry、Evangelos Milios、Sageev Oore和Hassan Sajjad。SugarCrepe++数据集：视觉-语言模型对语义改动和词汇改动的敏感性。神经信息处理系统进展，37：17972–18018，2024。

[7] Lijie Fan、Dilip Krishnan、Phillip Isola、Dina Katabi 和 Yonglong Tian。通过语言改写改进 CLIP训练。神经信息处理系统进展，36:35544–35575，2023。

[8] Lijie Fan、Kaifeng Chen、Dilip Krishnan、Dina Katabi、Phillip Isola 和 Yonglong Tian。合成图像用于模型训练的缩放定律……暂且如此。在*IEEE/CVF*计算机视觉与模式识别会议论文集中，第7382–7392页，2024。

[9] Li Fei-Fei、Robert Fergus 和 Pietro Perona。物体类别的一次性学习。*IEEE*模式分析与机器智能汇刊，28(4)：594–611，2006。

[10] Hasan Abed Al Kader Hammoud、Hani Itani、Fabio Pizzati、Philip Torr、Adel Bibi 和 Bernard Ghanem。SynthCLIP：我们准备好进行完全合成的 CLIP训练了吗？*arXiv*预印本 *arXiv:2402.01832*，2024。

[11] Jack Hessel、Ari Holtzman、Maxwell Forbes、Ronan Le Bras 和 Yejin Choi。CLIPScore：一种无需参考的

图像描述评估指标。载于*2021*年自然语言处理实证方法会议论文集，第 7514–7528 页，2021 年。

[12] Cheng-Yu Hsieh、Jieyu Zhang、Zixian Ma、Aniruddha Kembhavi 和 Ranjay Krishna。《Sugarcrepe：修正视觉-语言组合性方面可操纵的基准》。神经信息处理系统进展，36：31096–31116，2023。

[13] Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc Le、Yun-Hsuan Sung、Zhen Li 和 Tom Duerig。通过噪声文本监督扩展视觉与视觉-语言表征学习。载于国际机器学习会议，第 4904–4916 页。机器学习研究论文集，2021。

[14]Albert Q. Jiang、Alexandre Sablayrolles、Arthur Mensch、Chris Bamford、Devendra Singh Chap-lot、Diego de las Casas、Florian Bressand、Gianna Lengyel、GuillaumeLample、Lucile Saulnier、LélioRe-nard Lavaud、Marie-Anne Lachaux、Pierre Stock、Teven Le Scao、Thibaut Lavril、Thomas Wang、Timo-thée Lacroix 和 William El Sayed。Mistral 7b，2023。网址 https://arxiv.org/abs/2310.06825。

[15] Alex Krizhevsky、Geoffrey Hinton 等人。从微小图像中学习多层特征。（2009），2009。

[16] Zhengfeng Lai、Haotian Zhang、Bowen Zhang、Wen- tao Wu、Haoping Bai、Aleksei Timofeev、Xianzhi Du、Zhe Gan、Jiulong Shan、Chen-Nee Chuah 等人。Ve- clip：通过视觉增强型说明文字改进 CLIP 训练。载于欧洲计算机视觉会议，第 111–127 页。Springer，2024。

[17] Andrea Lampis、Eugenio Lomurno、Matteo Matteucci 和 Via Ponzio。弥合差距：通过后处理技术提升合成数据的效用。2023。

[18] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C Lawrence Zitnick。微软COCO：上下文中的常见物体。载于欧洲计算机视觉会议，第 740–755 页。Springer，2014。

[19] Eugenio Lomurno 和 Matteo Matteucci。联邦知识回收：隐私保护型合成数据共享。《模式识别快报》，191：124–130，2025。

[20] Eugenio Lomurno 和 Matteo Matteucci。合成图像学习：在保持性能的同时防止成员推理攻击。《模式识别快报》，190：52–58，2025。

[21] Eugenio Lomurno、Matteo D'Oria 和 Matteo Matteucci。用于下游分类任务的 Stable Diffusion 数据集生成。*arXiv* 预印本 *arXiv:2405.02698*，2024。

[22] Subhransu Maji、Esa Rahtu、Juho Kannala、Matthew Blaschko 和 Andrea Vedaldi。飞机细粒度视觉分类。*arXiv* 预印本 *arXiv:1306.5151*，2013。[23] Maria-Elena Nilsback 和 Andrew Zisserman。对大量类别的花卉进行自动分类。载于印度计算机视觉、图形学与图像处理会议，2008年12月。[24] Omkar M. Parkhi、Andrea Vedaldi、Andrew Zisserman 和 C. V. Jawahar。猫和狗。载于*IEEE* 计算机视觉与模式识别会议，2012年。

[25] Maitreya Patel、Naga Sai Abhiram Kusumba、Sheng Cheng、Changhoon Kim、TejasGokhale、Chitta Baral 等。TripletCLIP：通过合成视觉-语言负样本提升 CLIP 的组合推理能力。神经信息处理系统进展，37：32731–32760，2024。

[26] Filip Radenovic、Abhimanyu Dubey、Abhishek Kadian、Todor Mihaylov、Simon Vandenhende、Yash Patel、Yi Wen、Vignesh Ramanathan 和 Dhruv Mahajan。视觉-语言预训练中的过滤、蒸馏与困难负样本。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第 6967–6977 页，2023。

[27] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。《从自然语言监督中学习可迁移视觉模型》。载于国际机器学习会议，第 8748–8763 页。机器学习研究论文集，2021。

[28] Nicolò Francesco Resmini、Eugenio Lomurno、Cristian Sbrolli 和 Matteo Matteucci。《你的图像生成器就是你的新私有数据集》。《style id='2'>《图像与视觉计算》，第 105727 页，2025。

[29] Joshua Robinson、Ching-Yao Chuang、Suvrit Sra 和 Stefanie Jegelka。《使用困难负样本进行对比学习》。《style id='2'>arXiv 预印本 arXiv:2010.04592，2020。

[30]Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer。《使用潜在扩散模型进行高分辨率图像合成》。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 10684–10695 页，2022。

[31] Axel Sauer、Dominik Lorenz、Andreas Blattmann 和 RobinRombach。《对抗性扩散蒸馏》。载于欧洲计算机视觉会议，第 87–103 页。Springer，2024。

[32] Yonglong Tian、Lijie Fan、Phillip Isola、Huiwen Chang 和 Dilip Krishnan。StableRep：来自文生图模型的合成图像能够训练出强大的视觉表征学习器。神经信息处理系统进展，36：48382–48402，2023。

[33] Yonglong Tian、Lijie Fan、Kaifeng Chen、Dina Katabi、Dilip Krishnan 和 Phillip Isola。从模型中学习视觉，与从数据中学习视觉的效果不相上下。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 15887–15898 页，2024。

[34] Simon Vandenhende、Stamatios Georgoulis、Wouter Van Gansbeke、Marc Proesmans、Dengxin Dai 和 Luc Van Gool。密集预测任务的多任务学习：一项综述。*IEEE*模式分析与机器智能汇刊，44(7):3614–3633，2021。

[35]Pablo Villalobos、Jaime Sevilla、Lennart Heim、Tamay Besiroglu、Marius Hobbhahn 和 Anson Ho。我们会用光数据吗？对机器学习中数据集扩展上限的分析。*arXiv* 预印本 *arXiv:2211.04325*，1:1，2022。

[36] Haonan Wang、Minbin Huang、Runhui Huang、Lanqing Hong、Hang Xu、Tianyang Hu、Xiaodan Liang、Zhenguo Li、Hong Cheng 和 Kenji Kawaguchi。从数据中榨取更多价值：硬配对细化提升视觉语言模型，且无需额外数据。载于*2025*年美洲语言学学会计算语言学分会会议论文集：人类语言技术（第*1*卷：长篇论文），第7854–7873页，2025年。

[37] Jianxiong Xiao、James Hays、Krista A Ehinger、Aude Oliva 和 Antonio Torralba。SUN 数据库：从修道院到动物园的大规模场景识别。载于*2010* 年 *IEEE* 计算机学会计算机视觉与模式识别会议，第 3485–3492 页。IEEE，2010 年。

[38] Enze Xie、Junsong Chen、Junyu Chen、Han Cai、Haotian Tang、Yujun Lin、Zhekai Zhang、Muyang Li、Ligeng Zhu、Yao Lu 和 Song Han。SANA：采用线性扩散变换器的高效高分辨率图像合成，2024 年。网址https://arxiv.org/abs/2410.10629。

[39] Hu Xu、Saining Xie、Xiaoqing Ellen Tan、Po-Yao Huang、Russell Howes、Vasu Sharma、Shang-Wen Li、Gargi Ghosh、Luke Zettlemoyer 和 Christoph Feichtenhofer。揭开 CLIP 数据的奥秘。*arXiv* 预印本 *arXiv:2309.16671*，2023 年。

[40] Peter Young、Alice Lai、Micah Hodosh 和 Julia Hockenmaier。从图像描述到视觉指称：用于事件描述语义推断的新相似度指标。计算语言学协会学报，2：67–78，2014 年。

[41]Mert Yuksekgonul、Federico Bianchi、Pratyusha Kalluri、Dan Jurafsky 和 James Zou。《视觉语言模型何时以及为何表现得像词袋，以及该如何应对？》*arXiv* 预印本 *arXiv:2210.01936*，2022。

你的任务是撰写一段图像描述，其中包含一个概念，并对该概念周围的场景进行视觉描述。该概念必须明确出现在图像描述中。你的概念是 {c}。在图像描述中加入 {attribute}。只输出一个语法正确且不超过 20 个词的图像描述。不要输出任何注释、词数、事实等内容。只输出一个句子。

属性中的["一个背景"、"一种颜色"、"一个与照明相关的术语"、"一种材质"、"一种位置关系"、"透视或相机角度"、"一种艺术风格"]

(a) Mistral 用于生成基线描述集的提示词。仅当添加属性指示时才会包含红色文本。

根据以下图像描述，仅通过改变[概念/ {attribute1}]。要改变的概念是 {c}。
图注：{caption}
尽可能少地修改图注，使其与原图注相似，同时仍然能与新的[概念/ {attribute2}]。此外，不要改变图注的主要概念为 {c}。 不要输出任何注释、字数统计，事实等内容。只输出新图注。

属性1 的 ["背景"、"颜色"、"照明（含可能相反的含义）"、"材质"、"位置关系"、"透视"、"艺术风格"]属性2 的 ["背景"、"颜色"、"照明"、"材质"、"位置关系"、"透视"、"风格"]

(b) Llama 用于生成困难负例描述集的提示词。若生成原始图注时未使用任何属性，则使用 c 作为要改变的语义轴。

图 5：零样本程序化条件控制。(a) 基线提示通过强制使用结构化概念-属性元组（c，a）来确保语义正交性。(b) 困难负样本提示施加严格的反事实约束，以生成有效的 T^- 样本，同时避免句法漂移。

A 图注生成提示

图 5 展示了我们用于生成图像描述的零样本提示。基线图像描述集由图 5a 所示的提示生成，该提示要求模型根据起始概念 $c \in C$ 生成图像描述，并可选择要求其从预定义的语义修改轴列表 $\mathcal{A}$ 中选取一个属性 a 加以包含。随后，使用图 5b 中的提示生成困难负样本标题，该提示指示大语言模型仅沿所选语义轴修改基础描述，同时保持句子结构的其余部分不变。

B CLIP与TripletCLIP损失之间的关系

本节将形式化推导标准 CLIP 目标与我们在困难负样本挖掘策略中使用的 TripletCLIP 损失之间的联系。我们明确说明，Triplet 公式如何通过过滤掉含噪的文本到图像负样本来优化对比学习空间，同时在信息丰富的图像到文本对上保持严格间隔。

给定 $T^+ = \{t_1^+, t_2^+, \ldots, t_N^+\}$，从文本编码器中提取的基础文本特征集合：

$$t_k^+ = f_{txt}(\text{caption}_k), \quad \forall k \in [1, N], \tag{9}$$

以及 $T^- = \{t_1^-, t_2^-, \ldots, t_N^-\}$，使用大语言模型生成的困难负面描述，从 f_{txt} 中提取的困难负样本文本特征集合：

$$\text{caption}_k^- = \text{LLM}(\text{caption}_k, \text{hard negative prompt}). \tag{10}$$

类似地，设 $I^+ = \{i_1^+, i_2^+, \ldots, i_N^+\}$ 为从图像编码器中提取的基础图像特征集合：

$$i_k^+ = f_{img}(\text{image}_k), \quad \forall k \in [1, N], \tag{11}$$

以及 $I^- = \{i_1^-, i_2^-, \ldots, i_N^-\}$，使用扩散生成图像，从 f_{img} 中提取的困难负样本图像特征集合：

$$\text{image}_k^- = \text{T2I}(\text{caption}_k^-). \tag{12}$$

我们将 $\tilde{T} = T^+ \cup T^-$ 和 $\tilde{I} = I^+ \cup I^-$ 分别定义为所有文本特征和视觉特征的拼接批次，其尺寸为 $2N$。在该拼接批次上，标准 CLIP 损失定义为：

$$\mathcal{L}_{CLIP}(\tilde{I}, \tilde{T}) = \frac{1}{2}(\mathcal{L}_{NCE}^{\tilde{I}\to\tilde{T}} + \mathcal{L}_{NCE}^{\tilde{T}\to\tilde{I}}). \tag{13}$$

我们可以将对图像到文本项 $\mathcal{L}_{NCE}^{I\to T}$ 的求和拆分为对 $2N$ 正锚点 (I^+) 和负锚点 (I^-) 的求和：

$$\mathcal{L}_{NCE}^{\tilde{I}\to\tilde{T}} = -\frac{1}{2N}\sum_{k=1}^{2N}\log\frac{\exp(s(\tilde{i}_k, \tilde{t}_k)/\tau)}{\sum_{j=1}^{2N}\exp(s(\tilde{i}_k, \tilde{t}_j)/\tau)} \tag{14}$$

$$= -\frac{1}{2N}\left(\sum_{k=1}^{N}\log\frac{\exp(s(i_k^+, t_k^+)/\tau)}{\sum_{t\in\tilde{T}}\exp(s(i_k^+, t)/\tau)} + \sum_{k=1}^{N}\log\frac{\exp(s(i_k^-, t_k^-)/\tau)}{\sum_{t\in\tilde{T}}\exp(s(i_k^-, t)/\tau)}\right) \tag{15}$$

$$= \frac{1}{2}\left(\mathcal{L}^{I^+\to T^+;}_{NCE} \quad {}^{T^-} + \mathcal{L}^{I^-\to T^-;}_{NCE} \quad {}^{T^+}\right). \tag{16}$$

设置					SugarCrepe								SugarCrepe++ (ITT)					
					替换			Swap		Add			替换			Swap		
n^+	m	HN	$\mathcal{L}_{HN}$	S	Att	Obj	Rel	Att	Obj	Att	Obj	**Avg**	Att	Obj	Rel	Att	Obj	**Avg**
1	1	✗	✗	✗	60.0	64.2	53.2	51.2	48.2	59.0	52.0	55.4	42.1	46.7	33.2	31.5	29.8	36.7
2	2	✓	✗	✗	64.6	60.6	53.7	52.9	**57.1**	58.7	50.9	56.9	49.0	47.6	35.9	33.0	34.3	40.0
2	2	✓	✗	✓	64.9	61.7	**54.6**	49.1	46.9	56.5	51.5	55.0	48.9	49.2	36.7	31.2	28.2	38.8
2	2	✓	✓	✗	63.1	61.4	51.4	52.6	49.8	61.3	50.8	55.8	49.8	47.3	33.9	32.1	32.7	39.2
2	2	✓	✓	✓	64.6	61.4	52.4	49.0	53.5	**61.9**	51.8	56.4	**50.6**	47.6	35.5	31.1	**37.6**	40.5
3	3	✓	✓	✓	**65.0**	**64.9**	53.6	**54.7**	**57.1**	58.2	**52.9**	**58.1**	50.3	**52.9**	36.0	**36.0**	32.7	**41.6**
4	4	✓	✓	✓	63.8	64.2	52.4	52.1	51.0	**61.9**	51.9	56.8	**50.6**	51.7	**37.0**	33.3	31.0	40.7

表4：细粒度句法绑定评估。SugarCrepe基准上的结果表明，最大化集成异质性（m）对于消解复杂的组合歧义至关重要，并可防止模型依赖词汇捷径。

		HN组合			图像检索				文本检索				零样本
n^+	m	Smpl损失		调度	MS Coco	F8K	F30K	**Avg**	MS Coco	F8K	F30K	**Avg**	ImageNet
1	1	✗	✗	✗	6.4	15.0	13.3	11.6	8.7	21.5	17.5	15.9	8.15
2	2	✓	✗	✗	6.7	16.7	16.1	13.2	8.9	23.3	21.5	17.9	9.95
2	2	✓	✗	✓	7.5	16.4	16.9	13.6	10.4	23.0	22.9	18.8	10.23
2	2	✓	✓	✗	7.8	16.9	17.4	14.0	9.8	22.8	24.4	19.0	10.27
2	2	✓	✓	✓	7.7	18.3	17.8	14.6	10.5	**26.2**	**24.9**	**20.5**	10.80
3	3	✓	✓	✓	8.7	**19.5**	18.8	**15.7**	10.1	22.9	23.6	18.9	10.72
4	4	✓	✓	✓	**8.8**	18.9	**19.4**	**15.7**	**11.5**	24.9	24.7	20.4	**11.01**

表 5：困难负样本策略分析。课程调度器与三元组损失的结合优化了稳定性与可塑性之间的权衡，在缓解训练早期梯度噪声的同时，为检索任务设置了更严格的间隔。

同样，对于文本到图像方向 $\mathcal{L}_{NCE}^{T\to I}$，我们可以推导出：

$$\mathcal{L}_{NCE}^{\tilde{T}\to\tilde{I}} = \frac{1}{2}\left(\mathcal{L}_{NCE}^{T^+\to I^+;I^-} + \mathcal{L}_{NCE}^{T^-\to I^-;I^+}\right). \tag{17}$$

将这些项结合起来，总 CLIP 损失可表示为：

$$\mathcal{L}_{CLIP}(\tilde{I},\tilde{T}) = \frac{1}{4}\left(\mathcal{L}_{NCE}^{I^+\to T^+;T^-} + \mathcal{L}_{NCE}^{I^-\to T^-;T^+} + \mathcal{L}_{NCE}^{T^+\to I^+;I^-} + \mathcal{L}_{NCE}^{T^-\to I^-;I^+}\right). \tag{18}$$

然而，TripletCLIP 损失去除了涉及困难负样本的噪声文本到图像比较（即 $T^+ \to I^-$和 $T^- \to I^+$）。我们可以通过增加一项代价 $C \geq 0$来表示标准 CLIP 项与“干净”的三元组项之间的关系：

$$\mathcal{L}_{NCE}^{T^+\to I^+;I^-} = \mathcal{L}_{NCE}^{T^+\to I^+} + C_{T^+\to I^-} \tag{19}$$

$$\mathcal{L}_{NCE}^{T^-\to I^-;I^+} = \mathcal{L}_{NCE}^{T^-\to I^-} + C_{T^-\to I^+} \tag{20}$$

其中 C 表示 分母中额外负样本所对应的对数和指数项。因此，我们得出结论 标准 CLIP 损失的下界为 TripletCLIP 损失：

$$\mathcal{L}_{CLIP}(\tilde{I},\tilde{T}) = \frac{1}{4}\left(\mathcal{L}_{TripletCLIP}(\tilde{I},\tilde{T}) + C_{T^+\to I^-} + C_{T^-\to I^+}\right). \tag{21}$$

C PolyGen 完整结果

此处我们详细列出了所有下游任务的性能。表4展示了 SugarCrepe 和 SugarCrepe++上的组合推理得分。表 5分析了困难负样本策略对跨模态检索的影响。最后，表 6 和表 7 分别报告了纯视觉迁移（线性探测、少样本）以及零样本分类/检索在各数据集上的完整指标，而表 8 则展示了视觉任务上的消融研究。

方法	n^+	m	$\mathcal{L}_{I2I}$	视觉数据集									
				CIFAR10	CIFAR100	Aircraft	DTD	Flowers	Pets	SUN397	Caltech	Food	**Avg**
线性探测													
	1	1	✗	72.5	50.8	24.7	**47.2**	69.7	38.9	39.6	51.9	46.5	49.1
	1	2	✗	72.9	51.0	20.4	44.1	67.6	36.7	38.0	51.7	45.4	47.5
	1	3	✗	73.0	51.6	23.5	44.3	70.9	37.4	38.1	51.9	45.8	48.5
	1	4	✗	73.1	51.4	23.8	44.7	71.1	36.8	39.1	53.5	46.8	48.9
	2	1	✗	72.2	50.3	23.4	47.1	**71.6**	39.2	40.3	54.5	47.4	49.5
	2	2	✗	73.0	49.0	22.2	45.9	68.3	36.2	39.4	51.0	46.2	47.9
	3	1	✗	72.8	51.3	24.6	46.9	70.6	**39.8**	**41.1**	50.9	46.1	49.3
	3	3	✗	73.6	51.6	22.8	45.2	68.5	35.4	39.0	51.1	46.8	48.2
	4	1	✗	72.8	51.3	**25.3**	46.9	70.2	38.9	40.9	**54.9**	47.3	**49.8**
	4	4	✗	**73.9**	**51.8**	23.9	45.1	69.0	37.7	40.1	50.5	**47.6**	48.9
	2	1	✓	70.5	47.2	20.7	41.5	57.7	31.9	34.4	42.8	43.6	43.4
	2	2	✓	69.8	47.6	19.4	37.3	54.6	31.0	32.6	40.2	40.7	41.5
	3	1	✓	71.7	49.2	20.9	43.5	58.4	34.9	36.0	42.8	42.8	44.4
	3	3	✓	70.5	48.3	19.5	37.6	52.8	31.5	31.8	39.1	40.5	41.3
	4	1	✓	72.4	50.7	21.4	44.5	57.8	34.9	36.2	43.7	45.0	45.2
	4	4	✓	71.6	49.7	19.9	41.1	56.0	31.5	33.2	39.8	43.0	42.9
少样本（*5*样本，*5*类）													
	1	1	✗	46.3	57.8	**39.8**	63.5	85.6	48.5	84.2	75.4	53.4	61.6
	1	2	✗	47.8	59.5	39.1	60.9	85.1	47.0	83.8	75.3	51.2	61.1
	1	3	✗	48.2	59.2	39.1	61.7	85.8	47.3	83.8	75.4	51.5	61.3
	1	4	✗	47.7	**59.7**	39.3	61.7	85.9	47.5	84.3	75.4	52.4	61.5
	2	1	✗	46.2	58.1	38.8	65.1	**87.0**	48.9	86.9	77.4	53.4	62.4
	2	2	✗	48.0	59.3	37.7	61.4	84.9	47.1	85.1	74.6	52.1	61.1
	3	1	✗	47.0	58.5	38.9	65.4	86.0	**49.2**	87.4	78.1	52.1	62.7
	3	3	✗	47.9	59.1	37.4	62.5	84.1	46.3	85.7	76.0	51.3	61.1
	4	1	✗	47.0	58.3	38.6	**65.7**	86.0	48.7	**88.0**	**78.7**	**54.5**	**62.8**
	4	4	✗	**49.5**	**59.7**	38.1	62.9	84.3	47.7	86.7	77.0	52.2	62.0
	2	1	✓	43.9	53.9	36.3	57.7	78.3	42.7	84.5	71.6	47.1	57.3
	2	2	✓	43.0	54.0	35.5	54.8	75.9	40.6	82.8	69.6	45.1	55.7
	3	1	✓	42.9	53.0	36.0	57.8	78.1	42.7	85.3	72.1	48.2	57.3
	3	3	✓	43.1	52.6	34.8	54.2	76.0	39.1	82.4	69.5	45.1	55.2
	4	1	✓	43.2	54.2	34.5	58.6	78.3	42.5	86.0	73.2	48.8	57.7
	4	4	✓	44.8	53.9	34.8	55.3	77.5	40.3	84.3	70.0	46.8	56.4

表 6：纯视觉表征质量。 线性探测的增益与正样本密度（n^+）高度相关，这表明，与引入困难负样本判别器相比，稳定视觉簇质心更有利于实现线性可分性。

D 生成模型选择

D.1 用于困难负样本图像描述属性分类的大语言模型选择

生成困难负样本图注的任务要求严格遵循指令。为了选出最适合该任务的大语言模型，我们在困难负样本属性分类集成任务上评估了一组 M 规模相近的大语言模型：Mistral-V0.2-7B、Mistral-V0.3-7B、Llama-3.1-8B、Falcon-3-7B 和 Qwen-2.5-7B。给定大语言模型 $m \in M$、概念 c、基础描述 c_1以及由 m 通过修改属性 $a \in A$生成的困难负面描述 c_2，我们使用图 6 中定义的提示词，要求每个大语言模型判断图像 A 的哪个属性被用于生成 c_2。最终分类结果通过对各模型的预测结果进行多数投票得出。对于每个 $m \in M$，我们生成 $4.8k$ 对基础描述和困难负样本描述。具体而言，我们针对一组有限的维度生成 960 条基础描述$A = [$背景、颜色、位置、尺寸、概念$]$。在这 960 条描述中，有 160 条用于生成未改变的图注，要求模型仅对原始图注进行改写（图 6）。这被用作对照标签，以评估模型能否理解有意义的信息是否发生了变化。其余 800 条描述则用于生成困难负样本。表 9 显示，Llama 在此任务中的表现远超其他大语言模型，生成的图注能够清晰且严格地修改指定维度。

D.2 基线描述的大语言模型选择

选择 Mistral-V0.2-7B 生成基础描述集合有两个原因。首先，为了更公平地与 SynthCLIP 流程进行比较——该流程使用相同的模型生成其描述集合——从而将重点转向提示策略和我们的具体贡献。其次，我们对大语言模型进行消融实验，分别使用 Mistral-V0.2-7B 和 Llama-3.1-8B 生成 $500k$图像-描述对数据集，并使用 SDXLT 生成相应的图像。

			图像检索				文本检索				零样本
n^+	m	$\mathcal{L}_{I2I}$	MS Coco	Flkr 8K	Flkr 30K	**Avg**	MS Coco	Flkr 8K	Flkr 30K	**Avg**	ImageNet
1	1	✗	6.4	15.0	13.3	11.6	8.7	21.5	17.5	15.9	8.15
1	2	✗	7.1	15.6	13.8	12.1	9.8	22.1	20.6	17.5	8.53
1	3	✗	7.4	17.1	15.3	13.3	9.4	24.1	20.2	17.9	8.56
1	4	✗	7.2	17.0	14.3	12.8	9.4	23.3	20.2	17.6	8.92
2	1	✗	7.5	17.2	16.8	13.8	9.5	23.8	22.9	18.7	9.41
2	2	✗	7.6	17.7	17.2	14.2	10.4	25.8	24.3	20.2	9.54
3	1	✗	8.9	17.6	17.5	14.7	**12.1**	26.6	24.7	21.1	9.66
3	3	✗	8.9	**20.6**	19.1	16.2	**12.1**	**29.7**	25.4	**22.4**	9.38
4	1	✗	8.1	17.9	18.4	14.8	10.5	23.0	24.7	19.4	9.56
4	4	✗	**9.0**	19.9	19.4	16.1	11.4	27.1	25.8	21.4	9.86
2	1	✓	7.3	16.2	16.4	13.3	9.1	21.6	22.2	17.6	10.14
2	2	✓	8.2	17.4	17.7	14.4	11.1	24.5	23.7	19.8	11.19
3	1	✓	7.9	17.5	17.3	14.2	10.0	22.0	24.7	18.9	10.88
3	3	✓	**9.0**	20.2	**20.3**	**16.5**	11.5	26.4	**27.0**	21.7	10.98
4	1	✓	8.0	18.6	18.3	15.0	10.4	24.4	24.8	19.9	11.24
4	4	✓	8.8	18.9	19.4	15.7	11.1	24.5	26.2	20.6	**11.34**

表 7：零样本和检索性能。视觉-语言泛化与集成规模（m）和不变性学习（$\mathcal{L}_{I2I}$）相关，这证实了边缘化生成器特定伪影能够提升与开放世界语义流形的对齐。

		HN 组件			视觉数据集									
n^+	m	Smpl	Loss	调度	C10	C100	Air	DTD	Flwr	Pets	SUN	Cal	Food	**Avg**
线性探测														
1	1	✗	✗	✗	**72.5**	**50.8**	**24.7**	**47.2**	**69.7**	**38.9**	**39.6**	**51.9**	**46.5**	**49.1**
2	2	✓	✗	✗	70.4	46.9	19.8	38.4	53.4	30.6	31.9	41.6	40.4	41.5
2	2	✓	✗	✓	69.6	45.7	19.2	37.7	54.4	29.8	31.9	39.4	39.3	40.8
2	2	✓	✓	✗	71.0	48.9	18.9	38.7	55.8	31.4	33.2	42.1	41.2	42.3
2	2	✓	✓	✓	70.2	47.4	20.0	38.0	54.6	29.0	32.3	41.1	39.8	41.4
3	3	✓	✓	✓	71.1	48.6	20.0	34.2	53.4	29.0	31.7	38.1	41.6	40.8
4	4	✓	✓	✓	71.8	49.9	19.5	39.5	55.8	31.8	33.4	39.8	43.2	42.7
少样本														
1	1	✗	✗	✗	**46.3**	**57.8**	**39.8**	**63.5**	**85.6**	**48.5**	**84.2**	**75.4**	**53.4**	**61.6**
2	2	✓	✗	✗	42.4	53.9	34.8	54.6	76.8	41.5	82.7	70.0	44.9	55.7
2	2	✓	✗	✓	43.2	54.5	34.1	54.3	76.7	41.6	83.0	69.7	45.1	55.8
2	2	✓	✓	✗	43.4	55.3	34.9	55.5	78.0	42.5	84.0	70.7	46.4	56.8
2	2	✓	✓	✓	42.2	54.0	34.8	54.5	77.0	41.2	83.1	69.7	45.5	55.8
3	3	✓	✓	✓	44.3	54.5	34.5	53.9	76.4	39.0	82.3	69.3	45.3	55.5
4	4	✓	✓	✓	44.8	55.2	35.8	54.7	77.8	40.1	83.9	70.0	46.4	56.5

表 8：困难负样本集成的消融实验。结果表明，课程调度器是防止标准视觉任务中出现特征退化的关键组件，它能在早期对齐阶段保护表征免受噪声负梯度的影响。

我们发现，使用 Mistral 生成的图像描述在 ImageNet 零样本准确率上略有优势（top-1 为 1.8%，top-5 为 5.97%），优于 Llama（top-1 为 1.66%，top-5 为 5.61%），这进一步验证了我们的选择。

D.3 扩散模型选择

图像质量指标。 PolyGen 所用图像生成器（SD 1.5、SD 2、SANA、SDXLT）的选择，是通过对更大模型池进行比较后得出的；该模型池还包括 Stable Diffusion v2.1（SD 2.1）、Stable Diffusion 3.5 Large Turbo（SD 3.5 LT）和 FLUX.1 Schnell（FLUX）。我们从 MS COCO 2017 验证集中随机抽取 $5k$ 个样本各自的一条图像描述，并用每个模型生成相应图像。我们利用这些图像集，根据 FID、CLIPScore 和 TIFA 分数对模型进行比较。尽管 FID 最初被提出用于衡量在狭窄领域上训练的 GAN 模型所生成图像的保真度，但这一指标与多样性相关，因为它衡量的是一组图像特征相对于相应真实数据的分布偏移。在许多采用 CFG 的扩散模型中，较低的 CFG 尺度通常与较低（更好）的 FID 相关，但也与较低（更差）的输入图像描述对齐度相关；后者由 CLIPScore 衡量，因此也意味着较低的感知质量。基于这些原因，“多样性专家”模型与较低的 FID 相关，而“可识别性专家”则与较高的 CLIPScore 和 TIFA 分数相关。图 8 展示了在不同分辨率和去噪步数下进行这些评估的结果。虽然我们在选择最终集成时考虑了计算约束，但也可以考虑将 FLUX 和 SD 3.5 用作“可识别性专家”，将 SD 2.1 用作“多样性专家”。

通过重新排列所提供图注中的词语来改写图注，同时保持含义不变。
图注：{caption}
不要输出任何注释、词数、事实等内容。
只输出新图注。

将两张图像中主要的差异归类为以下类别之一：

概念：场景的主体不同。
背景：场景的背景不同。
位置：一个或多个元素的位置关系不同。
颜色：一个或多个元素的颜色不同。
尺寸：一个或多个元素的尺寸不同。
未改变：两张图像展示的是同一个场景。
[其他标签]

必须且只能选择一个标签。你不能避免从这些选项中选择一个，
即使差异似乎模棱两可。

输出标签，除此之外不要输出任何内容。不要包含解释、注释、推理等。

将两条图像描述之间的语义差异分类为以下类别之一
类别：

概念：场景的概念有所不同。
背景：场景的背景有所不同。
位置：描述的位置关系有所不同。
颜色：描述的颜色有所不同。
尺寸：尺寸形容词有所不同。
未改变：两条图像描述描述的是同一个场景。
[其他标签]

必须恰好选择一个标签。您无法避免从以下
选项中选择一个，即使差异看起来模棱两可。

第一条图像描述的概念是 {c}。
第一条图像描述：{base_caption}
第二条图注：{hn_caption}

只输出标签，不要包含解释、注释、推理等内容。

图 6：合成数据的验证协议。 用于定量评估指令遵循程度并排除生成流水线中幻觉的零样本分类提示词。

模型	准确率	平均F1	平均精确率	平均召回率
Mistral V0.2	61%	60%	76%	61%
Mistral V0.3	62%	62%	77%	62%
Qwen 2.5	67%	67%	78%	67%
Falcon 7B	63%	65%	79%	63%
Llama 3.1	**74%**	**75%**	**80%**	**74%**

表 9：大语言模型指令遵循基准测试。Llama 3.1 在遵循语义约束方面表现更出色，因此在反事实图像描述生成过程中，逻辑矛盾（假阴性）的风险更低。

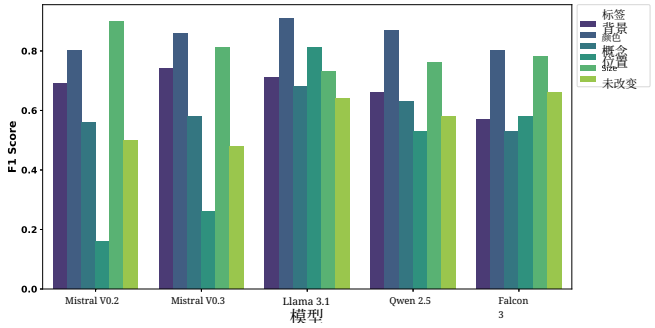


图 7：各语义轴的 F1 分数（大语言模型）。不同修改类型下的分类性能。有些维度（例如纹理）对大语言模型来说天生更难修改，同时又不改变全局上下文。

语义轴分类。 我们还测试了扩散模型能否正确、清晰地报告困难负面描述中包含的变化。为此，我们采用与大语言模型实验类似的方式，对语义轴进行分类。这次，我们使用 Qwen 2.5 VL 作为分类器，并将基础描述生成的图像、困难负面描述生成的图像，以及图 6 所示的图像分类提示作为输入。这次，我们总共使用了 2.4k 对图像；对于用于测试大语言模型的相同 5 个维度，每个维度有 480 对，其中 80 对用于未改变类别。表 10 显示，对于图像-文本描述对齐度和感知质量更高的模型，生成有用的困难负图像这一任务更容易完成。总体而言，由于图像提供的信号噪声更大，这些分数低于针对大语言模型进行的类似实验；这也印证了在使用困难负样本的对比模型中采用 TripletCLIP 损失所带来的性能提升。

E 语义轴选择

根据语义轴分类的结果（分别用于评估大语言模型和文生图模型），我们注意到，模型生成有效困难负样本的能力会因所修改的具体维度而显著不同（图 7、9）。随后，我们从更广泛的维度集合中进行第二轮分类，以确定最终的维度集合$A = [$背景、颜色、照明、材质、透视、位置、尺寸、风格、纹理、概念$]$, 这些分类既基于 Llama 生成的图像描述对，也基于 SDXLT 生成的图像对。图 10 展示了这些实验的结果。我们可以看到，当要改变的属性是纹理时，Llama 难以生成困难负样本。同时，SDXLT 生成的图像难以呈现图像描述之间的尺寸差异。因此，这两个属性从最终的语义轴集合中移除，该集合见图 5。

频域中的文生图模型比较

为了进一步了解生成图像的特征，并尝试理解为什么不同的扩散模型会带来不同的训练性能，我们在频域中分析这些图像。该分析将每个模型生成的图像的频谱特征与真实图像的基线数据集进行比较。为此，我们对每张标准化图像应用二维快速傅里叶变换（FFT），将其从空间域转换到频域。根据每张图像的二维幅度谱，我们提取两组关键指标：

模型	Res.	步数	FID ↓	CLIPScore ↑	TIFA ↑
Real	512^2	-	-	30.319	88.8
SD 1.5	512^2	50	23.014	29.962	78.7
SDXLT	512^2	2	32.593	**31.748**	87.9
SD 2	512^2	20	53.091	27.310	69.0
		50	55.348	27.244	69.0
	768^2	20	23.613	30.065	80.6
		50	**21.095**	30.162	80.4
SD 2.1	512^2	20	61.443	27.035	68.5
		50	60.213	27.128	68.8
	768^2	20	23.047	29.991	81.3
		50	23.481	30.068	82.5
SD 3.5 LT	512^2	4	35.448	31.586	88.1
		8	36.008	31.109	87.7
	1024^2	4	36.732	31.484	89.4
		8	36.337	31.358	89.1
SANA	512^2	20	27.316	31.163	88.1
	1024^2	20	27.573	31.171	88.0
FLUX	512^2	4	31.583	31.634	90.5
	1024^2	4	28.581	31.403	**90.9**

图 8：生成流形的定量分析。模型分为多样性专家（高方差）和可识别性专家（提示词遵循度高），以构建涵盖视觉语义帕累托前沿的异构集成。

模型	Res.	步数	准确率	平均F1	平均精确率	平均召回率
SD 1.5	512^2	50	25%	23%	28%	25%
SDXLT	512^2	2	37%	33%	40%	37%
SD 3.5 LT	512^2	4	34%	30%	36%	36%
		8	32%	33%	32%	32%
	1024^2	4	35%	31%	35%	35%
		8	35%	30%	36%	35%
FLUX	512^2	4	36%	33%	42%	36%
	1024^2	4	**38%**	**35%**	**43%**	**39%**

表 10：语义扰动的视觉定位。高保真度架构（可识别性专家）展现出显著更强的属性绑定能力，这对于正确呈现困难负样本课程所要求的细粒度差异至关重要。

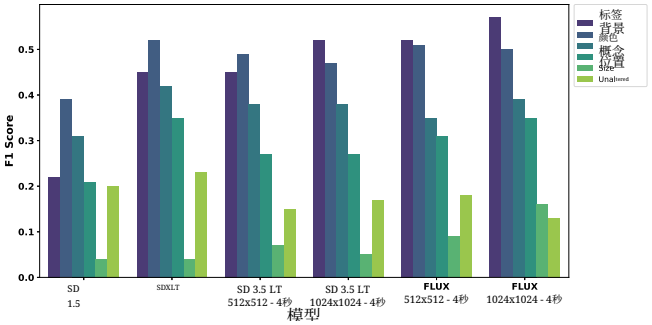


图 9：各语义轴的评估指标（文生图）。不同修改类型下的视觉定位性能。与较早的架构相比，SDXL-Turbo 和 FLUX 在呈现空间变化和对象级变化方面表现出更强的能力。

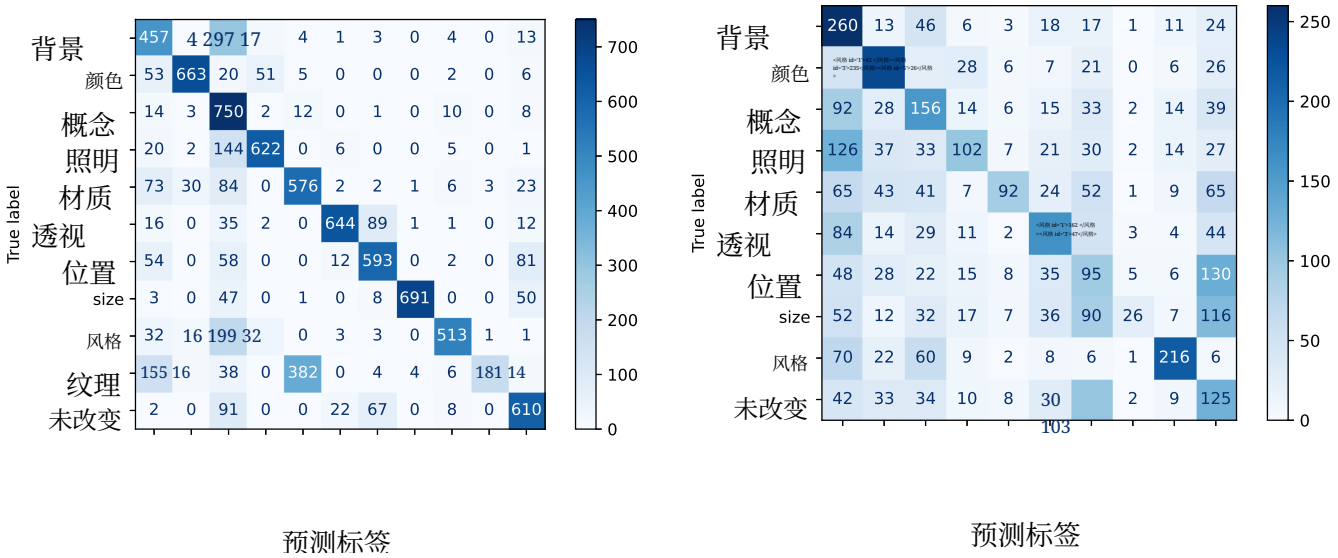
- **径向频率谱：** 通过径向平均函数将二维频谱缩减为一维向量。该向量表示幅度随空间频率（即距中心的半径）变化的平均值。对集合中的所有 $5k$ 图像重复此过程，并计算该集合的平均径向频谱。

- **频率能量：**我们对每张图像分别量化其能量分布，并计算高频分量的总能量 $[0.5, 1]$（归一化单位）。

图 11 展示了这项分析的结果，涵盖与 MS COCO 真实图像进行比较的 4 种扩散模型。显然，扩散过程的作用类似于低通滤波器，会衰减高频分量。不过，不同模型的频率响应显然各不相同，其中 SDXLT 的高频分量衰减最为陡峭。这或许可以归因于各模型截然不同的训练流程和目标（例如，SDXLT 通过蒸馏实现了强压缩，以支持低步数推理）；我们还观察到，这可能与多样性和分布偏移相关，而这两者通过 FID 分数进行衡量。唯一的例外是 SANA：它的衰减较为平缓，但记录到的多样性最低。这项分析表明，生成图像的频率特征可能是此前观察到的可识别性与多样性的权衡，以及由此导致的使用此类数据训练的模型下游性能表现的影响因素之一。不过，要充分理解这些关系，仍需进一步研究。

G 困难负样本感知损失函数

图12展示了针对一组能感知困难负样本的损失函数进行的消融实验，对比了 $\mathcal{L}_{CLIP}$在有无$\mathcal{L}_{margin}$、 $\mathcal{L}_{HN-NCE}$和 $\mathcal{L}_{TripletCLIP}$时的表现。从这些结果来看，TripletCLIP 似乎能够克服困难负图像带来的过度嘈杂信号，同时仍能利用困难负样本，并保留图像到文本方向。相比之下，间隔损失似乎也有助于提升性能，尝试在锚点、其困难负样本以及其他“更容易”、相似度较低的负样本之间建立相对层级。不过，当锚点与困难负样本之间存在一对多关系时，这种损失才能得到更有效的利用，而这并非我们当前数据集设计的情况。HN-NCE 则表现出截然不同的行为：其重要性加权策略似乎在训练的极早期达到峰值，随后迅速退化，这可能是因为在这些设置下加权方案的惩罚过重。



(a) Llama 结果。纹理属性已从最终语义轴集合中移除。**(b)** SDXLT 结果。尺寸已从最终语义轴集合中移除。

图 10：用于语义轴验证的混淆矩阵。 该分析推动了对修改空间 $\mathcal{A}$ 的裁剪，消除了混淆率较高的轴（例如“纹理”和“尺寸”），以确保训练信号在结构上保持一致。

H 均衡采样

均衡采样是 Xu等人提出的一种采样算法。 [39] 这是一种数据整理策略，旨在解决原始网络爬取数据集中存在的“富者愈富”问题：少数常见概念出现数百万次，而稀有概念则代表不足，导致分布高度偏斜，限制模型的泛化能力。给定阈值计数 t后，该算法会为C 中的每个概念分配一个采样概率，等于 $t/count(c)$，其中 $count(c)$ 等于 c 在描述集合中出现的次数（若 $count(c) > t$，否则概率为 1 。每条图像描述都根据其中各个概念的概率进行采样。我们生成一个包含 $100M$ 条图像描述的初始描述集合，随后对图像描述去重，并以$t = 30$运行该采样算法，从而得到包含 $13.1M$ 条图像描述的最终描述集合。最后，我们对描述集合进行均匀下采样，得到目标尺寸为 T^+ 的最终描述集合。图13展示了该算法如何对高频概念进行大幅下采样。

I Delta 多任务学习指标

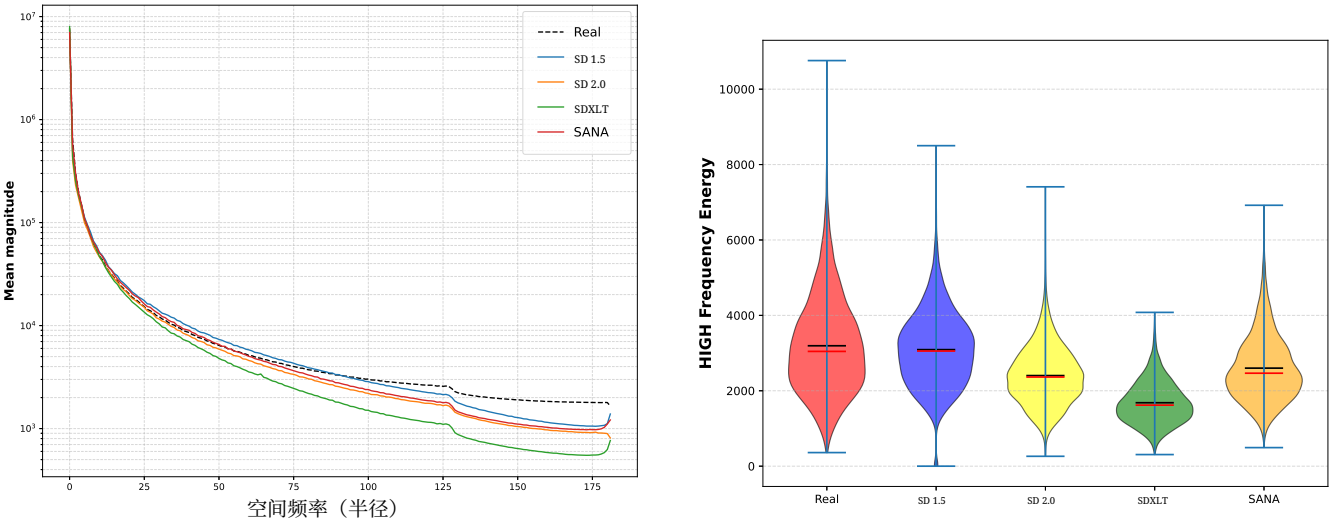
为了总结能够执行多个任务的模型的整体性能，并汇总不可直接比较的指标，Vandenhende等人 [34] 定义了 Delta 多任务学习（Δ_{MTL}）指标。该指标计算模型 m 相对于基线模型 b 在一组任务 T上的平均相对性能提升或下降：

$$\Delta_{MTL} = 100 \times \frac{1}{T} \sum_{i=1}^{T} (-1)^{l_i} \frac{M_{m,i} - M_{b,i}}{M_{b,i}}, \tag{22}$$

其中， l_i 表示对于指标 i，数值越低意味着性能越好时取 1，否则取 0。

J 定性示例

我们展示了定性示例，以直观评估困难负样本生成流程的有效性，以及多生成器集成的影响。图14展示了成功生成的示例，其中不同架构均准确体现了相应的语义修改。相较之下，图15展示了失败案例：特定生成器难以处理细粒度属性绑定，或生成含义模糊的可视化结果，这凸显了采用集成方法来降低个别模型错误影响的必要性。



(a) 所采用模型的径向轮廓。 **(b)** 最高 50% 频率的频率能量分布。

图 11：谱密度估计。 分析揭示了明显的生成器特定的频率偏差（例如，SDXL 的高频滚降）。PolyGen 在这些频谱的并集上进行训练，以防止模型过拟合到低层合成伪影。

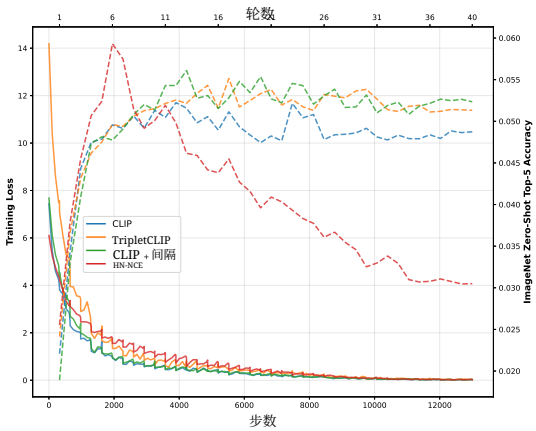


图 12：损失函数动态。 验证曲线表明，TripletCLIP 通过显式忽略噪声 $T \to I$ 比较而优于标准目标函数，从而有效过滤掉由文本到图像错位引入的假阴性。

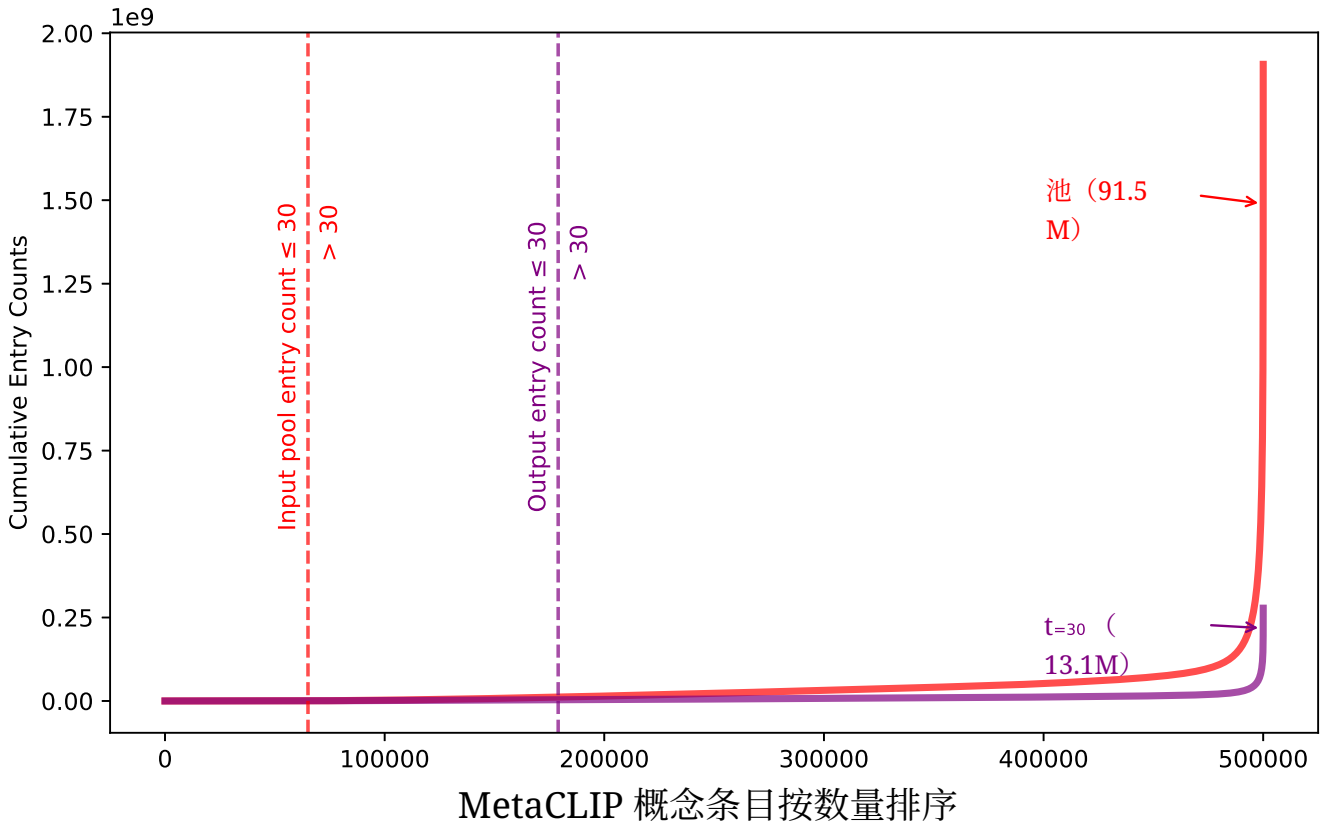


图13：通过均衡采样进行分布校正。该算法将网络数据中典型的幂律分布拉平，提高尾部概念的有效采样率，从而确保语义覆盖均匀。



图14：生成器不变性的可视化。该集成在互不相交的视觉流形上呈现相同的语义变化，促使编码器与共享的概念修改保持一致，而不是依赖特定生成器的风格线索。



图15：失效模式与集成韧性。单生成器中的属性泄漏或视觉歧义。多生成器方法通过对集成中的学习信号取平均，缓解了这些局部错误。