北京大学
PEKING UNIVERSITY

# DOES CLIP'S GENERALIZATION PERFORMANCE MAINLY STEM FROM HIGH TRAIN-TEST SIMILARITY?

**Prasanna Mayilvahanan**[1,2,3*] **Thaddäus Wiedemer**[1,2,3*] **Evgenia Rusak**[1,2,3]

**Matthias Bethge**[1,2] **Wieland Brendel**[2,3,4]

[1]University of Tübingen [2]Tübingen AI Center
[3]Max-Planck-Institute for Intelligent Systems, Tübingen [4]ELLIS Institute Tübingen

ICLR 2024

# IN SEARCH OF FORGOTTEN DOMAIN GENERALIZATION

**Prasanna Mayilvahanan**[1,2,3*] **Roland S. Zimmermann**[1,2,3*] **Thaddäus Wiedemer**[1,2,3]

**Evgenia Rusak**[1,2,3] **Attila Juhos**[1,2,3] **Matthias Bethge**[1,2] **Wieland Brendel**[2,3,4]

ICLR 2025

彭天天

2025/09/20

# 01 CLIP的泛化能力是否来源于训练-测试集的相似性?

模型：CLIP ViT-B/32
训练集：LAION-400M、LAION-200M、ImageNet-Train
ID测试集：ImageNet-Val、ImageNet-V2
OOD测试集：ImageNet-Sketch、ImageNet-R、ObjectNet

内容、风格都相似---------->风格相似

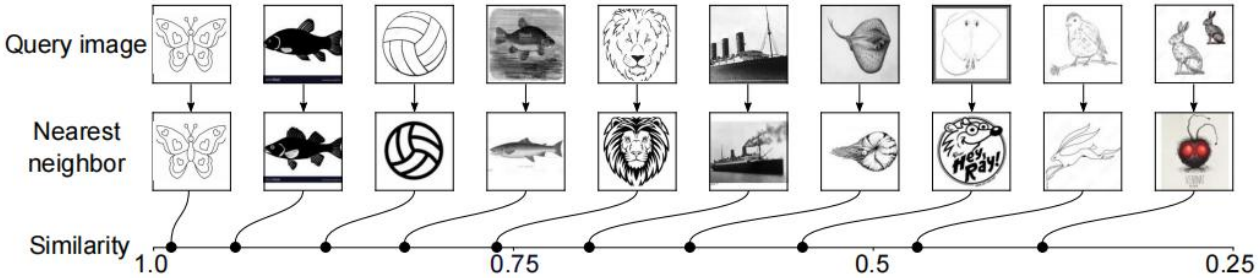
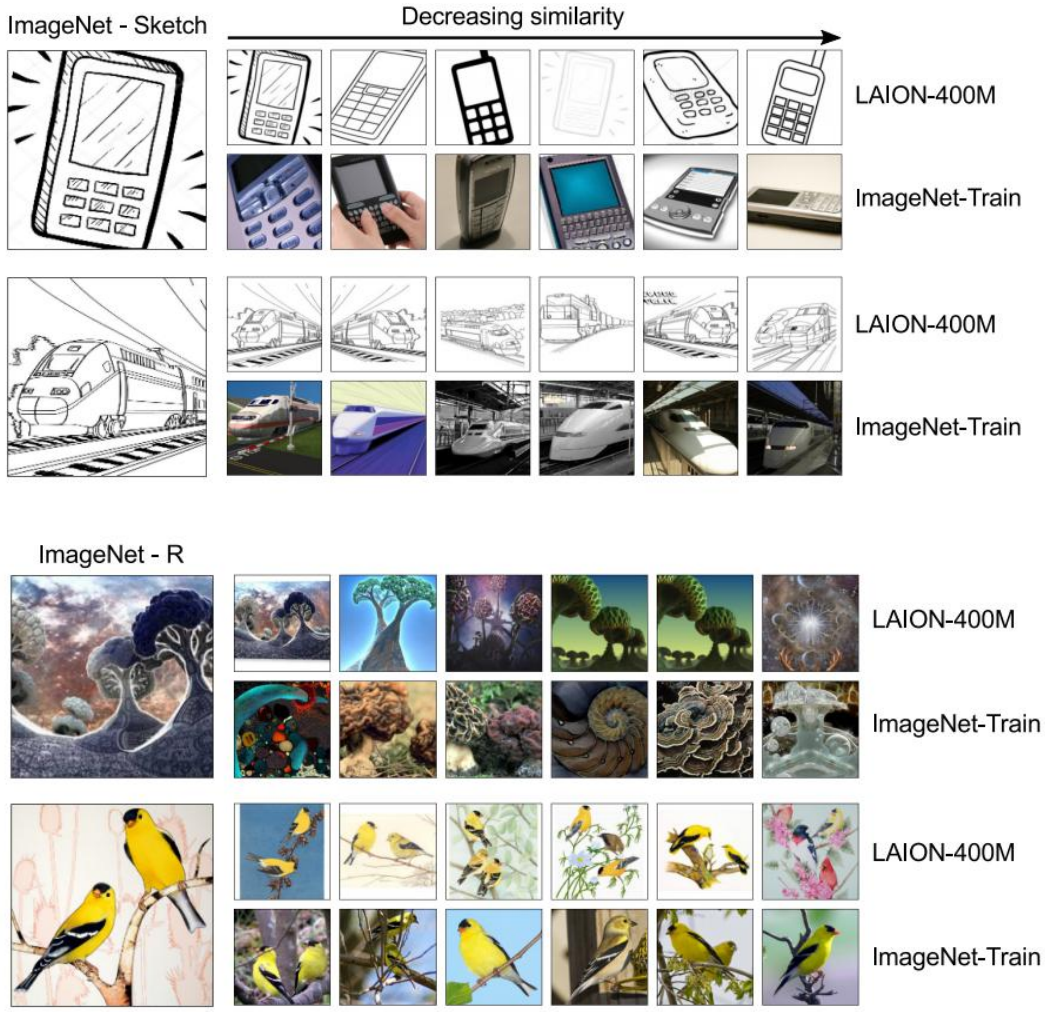



Figure 2: **Relation between *perceptual similarity* and visual closeness of nearest neighbors**. Query images are sampled from ImageNet-Sketch (top row) and are connected to their nearest neighbor in LAION-400M (bottom row). As in Fig. 1, perceptual similarity is simply the cosine similarity measured in CLIP ViT-B/16+'s image embedding space.

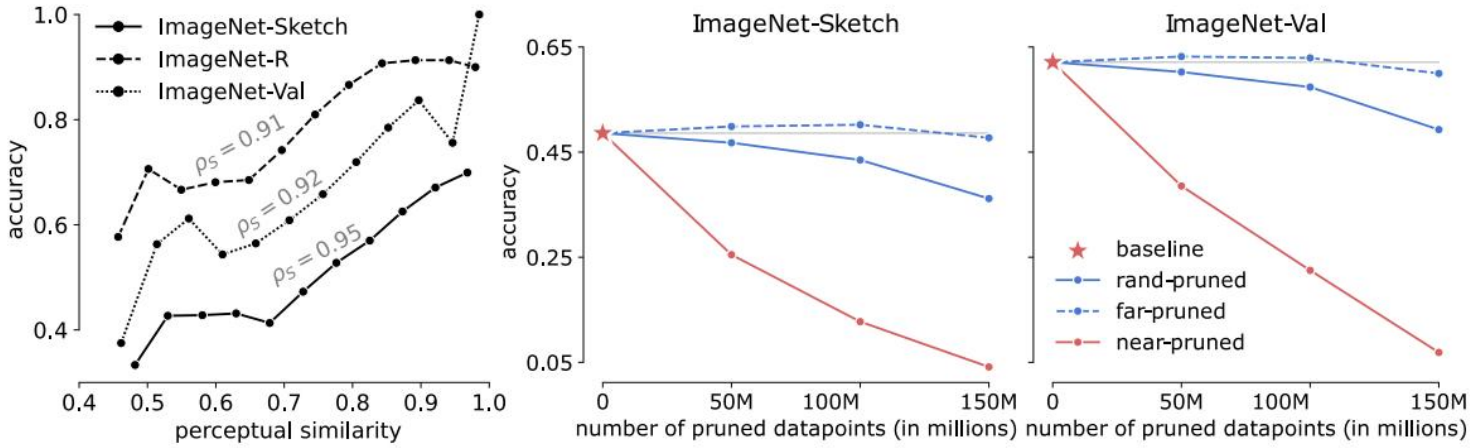
用图像在CLIP嵌入空间的余弦相似度来衡量图片之间的相似度



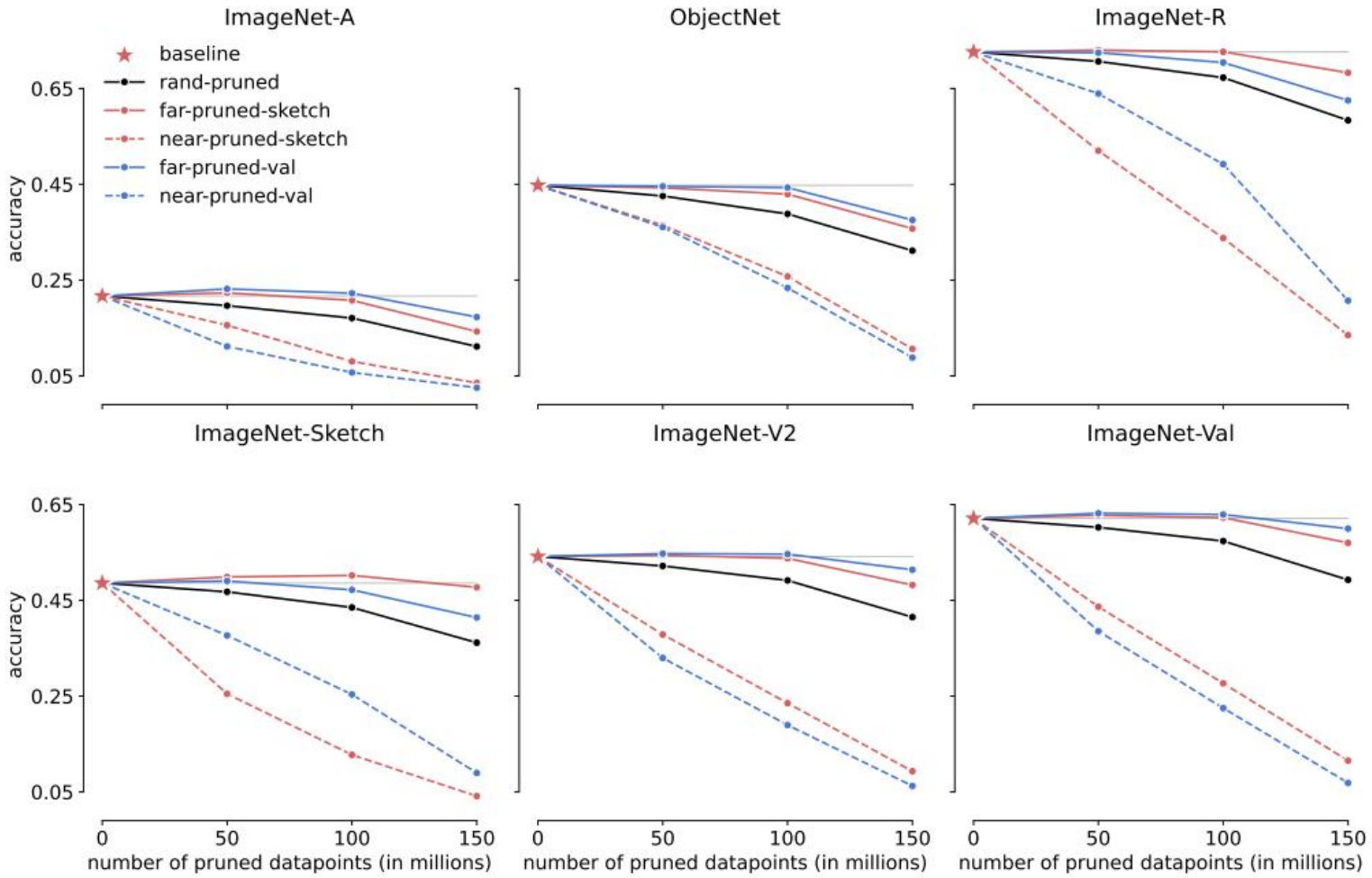
# 01 CLIP的泛化能力是否来源于训练-测试集的相似性?

左：将测试集按与训练集中图像相似度进行划分，发现OOD准确率与相似度高度相关
中：从训练集LAION-200M中逐步移除与ImageNet-Sketch相似的图像
右：从训练集LAION-200M中逐步移除与ImageNet-Val相似的图像

结论：CLIP的分类性能与其训练集和测试集的相似度直接相关



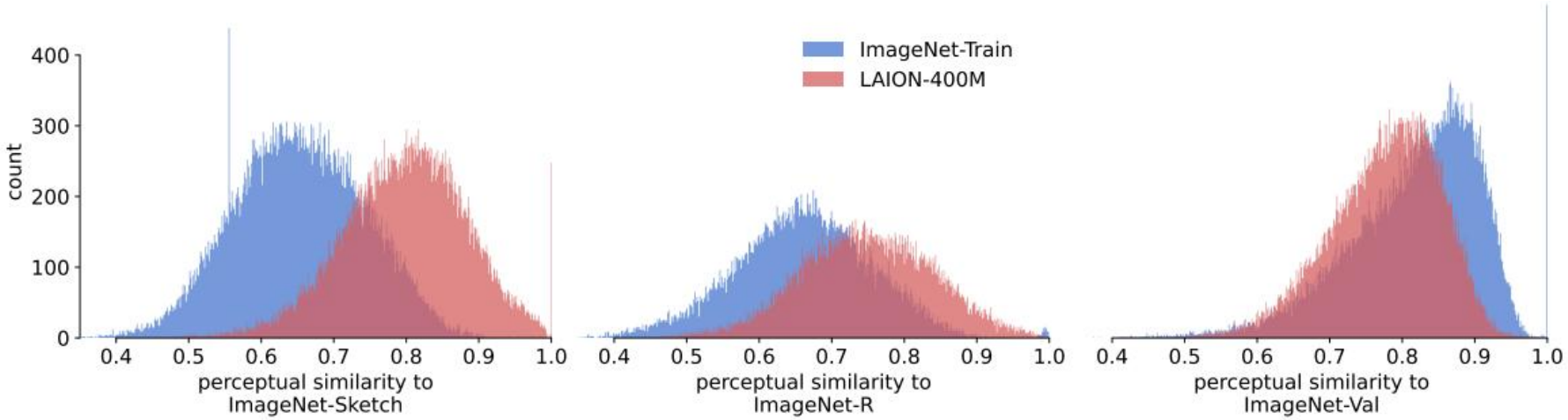
# 01 CLIP的泛化能力是否来源于训练-测试集的相似性?



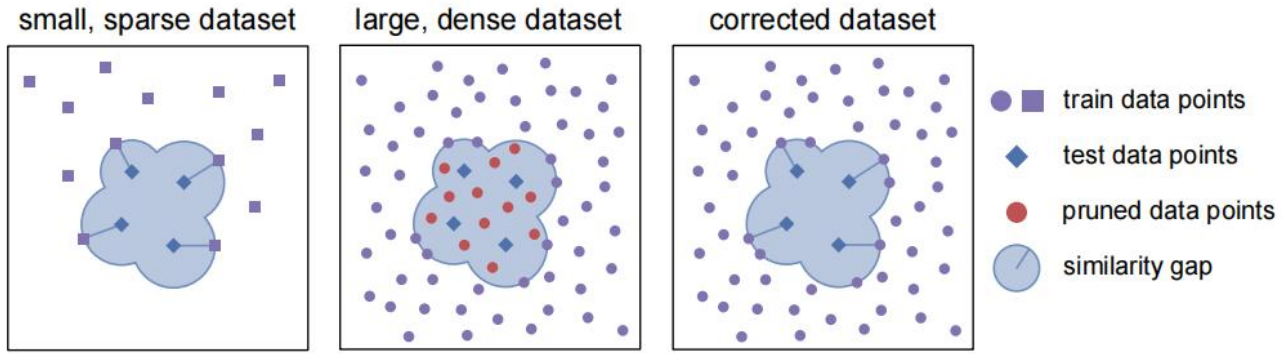

（near-pruned这种剪枝方式相当于删除了训练集中某一特定domain，导致训练集的多样性降低，CLIP的泛化能力也遭受降低）

# CLIP的泛化能力是否来源于训练-测试集的相似性?

CLIP的训练集LAION与传统在ImagNet上训练的模型相比，训练集分布存在明显差异——LAION中包含更多与OOD测试集相似的图像




训练集相似度对齐：移除LAION中与OOD图像相似度高于ImageNet与OOD图像的部分，用对齐后的LAION训练CLIP（只从embedding相似度上判断并移除，合理.....吗？）



# CLIP的泛化能力是否来源于训练-测试集的相似性?

即使是在相似度对齐后的LAION上训练的CLIP模型，OOD表现仍然很好——只有些许降低

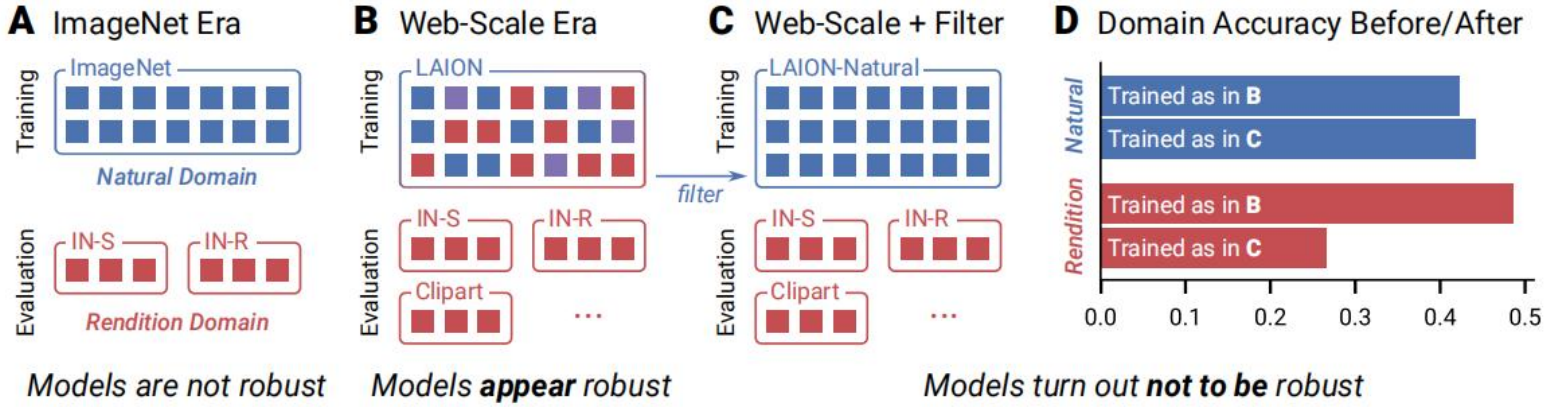
结论：CLIP的泛化能力不全是因为训练集-测试集的相似，而是因为它确实在大量、多样的数据上学到了泛化特征

| | | | Top-1 Accuracy | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | **Dataset** | **Size** | **Val** | **Sketch** | **A** | **R** | **V2** | **ObjectNet** |
| ViT-B/32 | OpenAI | 400 000 000 | 63.38 | 42.32 | 31.44 | 69.24 | 55.96 | 44.14 |
| ViT-B/32 | L-400M | 413 000 000 | 62.94 | 49.39 | 21.64 | 73.48 | 55.14 | 43.94 |
| ViT-B/32 | L-200M | 199 824 274 | 62.12 | 48.61 | 21.68 | 72.63 | 54.16 | 44.80 |
| ViT-B/32 | ├ val-pruned | *−377 340* | 62.12 | 48.38 | 21.45 | 72.2 | 54.76 | 42.79 |
| ViT-B/32 | ├ sketch-pruned | *−8 342 783* | 61.55 | 43.22 | 22.28 | 69.6 | 53.53 | 42.77 |
| ViT-B/32 | ├ a-pruned | *−138 852* | 62.49 | 48.49 | 21.63 | 72.15 | 54.38 | 43.25 |
| ViT-B/32 | ├ r-pruned | *−5 735 749* | 61.73 | 45.66 | 21.67 | 68.28 | 54.1 | 42.90 |
| ViT-B/32 | ├ v2-pruned | *−274 325* | 62.48 | 48.62 | 22.13 | 72.3 | 53.83 | 43.38 |
| ViT-B/32 | ├ objectnet-pruned | *−266 025* | 62.30 | 49.03 | 22.64 | 72.90 | 54.21 | 42.80 |
| ViT-B/32 | └ combined-pruned | *−12 352 759* | 61.5 | 41.97 | 21.72 | 67.25 | 53.65 | 42.23 |
| ResNet-101 | ImageNet-1k | 1 200 000 | 77.21 | 27.58 | 4.47 | 39.81 | 65.56 | 36.63 |

# 域污染多大程度上影响CLIP的领域泛化能力？

将数据划分为Natural和Rendition，LAION-200M作为训练集，由大部分的Natural图像和小部分Rendition图像组成，而OOD测试集（ImageNet-R、ImageNet-Sketch、ImageNet-A等）绝大部分为Rendition


单纯在LAION-Natural上训练的CLIP（C），与在未过滤的LAION上训练的CLIP（B），在Rendition域上准确率明显下降：



先训练了一个分类器，划分数据集中的Natural图像和Rendtion图像：

Table 2: **Domain composition of training sets.** We apply our *natural* and *rendition* domain classifiers with their strict thresholds at 98 % validation-precision to get a lower bound of samples from each domain and with their default thresholds to obtain a more balanced estimate. ImageNet-Train has a much smaller fraction of ***rendition*** samples than LAION-200M. We also note that 'combined-pruned', the training set from Mayilvahanan et al. (2023) that corrected for test set contamination, still contains a large fraction of renditions.

| | | Classifier Precision | | | | | |
|---|---|---|---|---|---|---|---|
| **Dataset** | **# Samples** | *Natural* | *Rendition* | *Natural* | *Ambiguous* | *Rendition* | |
| LAION-200M | 199 663 250 | 0.79 | 0.77 | 60.74 % | 25.41 % | 13.86 % | ⟹ 默认阈值 |
| | | 0.98 | 0.98 | 28.40 % | 63.70 % | 7.90 % | ⟹ 严格阈值 |
| ImageNet-Train | 1 281 167 | 0.79 | 0.77 | 89.20 % | 9.62 % | 1.18 % | |
| | | 0.98 | 0.98 | 36.00 % | 63.60 % | 0.40 % | |
| combined-pruned | 187 471 515 | 0.79 | 0.77 | 62.98 % | 25.18 % | 11.83 % | |
| | | 0.98 | 0.98 | 29.58 % | 64.02 % | 6.40 % | |

LAION-200M中至少包含7.9%的Rendition域图像（严格阈值下），combined-pruned在经过与ImageNet-Train的相似度对齐后，仍然存在相当部分的Rendtion域图像；此外，即使是OOD测试集上，也存在领域混杂现象（5%来自对立域）

——先前基于这些数据集训练或评估模型时得到的领域泛化能力很可能被高估！

# 域污染多大程度上影响CLIP的领域泛化能力？

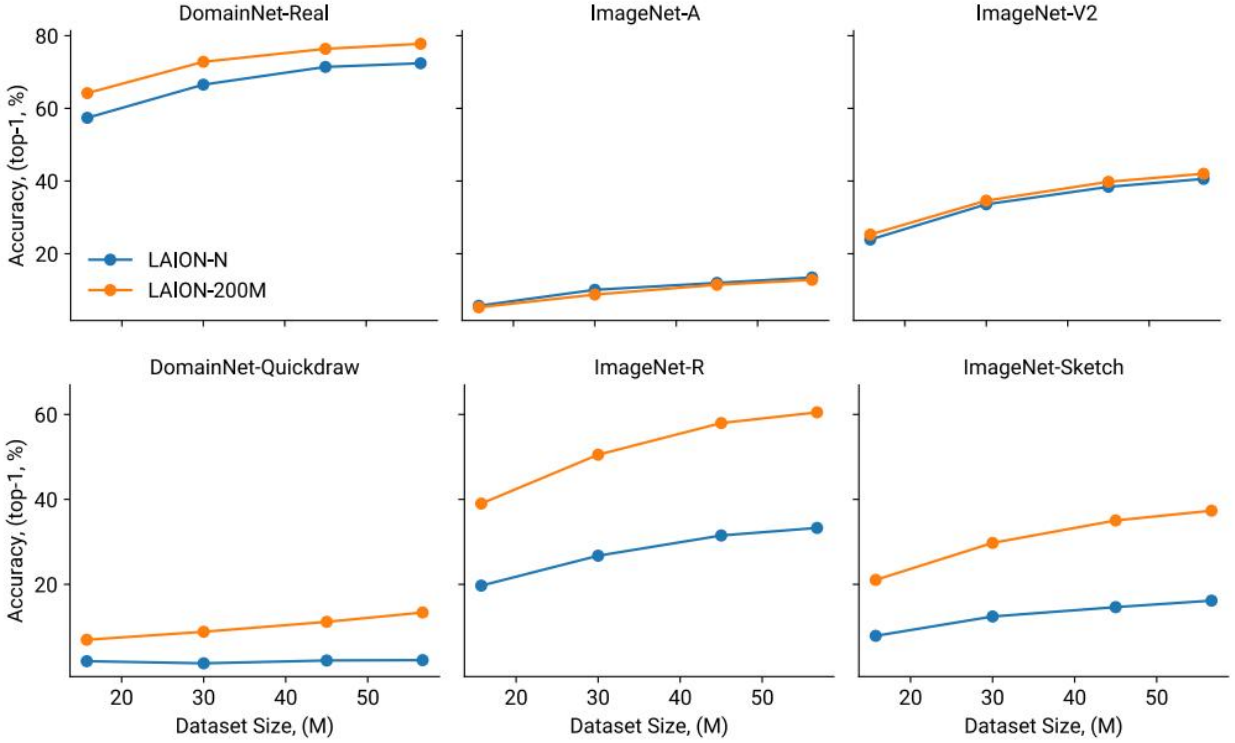
用严格阈值，从LAION-200M中提取单领域纯净子集LAION-Natural和LAION-Rendition：


Figure 3: **Random samples from LAION-Natural and LAION-Rendition.**

在LAION-Natural和LAION-200M下训练CLIP模型，性能对比：

standard natural testsets：相差无几

standard rendition testsets：LAION-N更弱

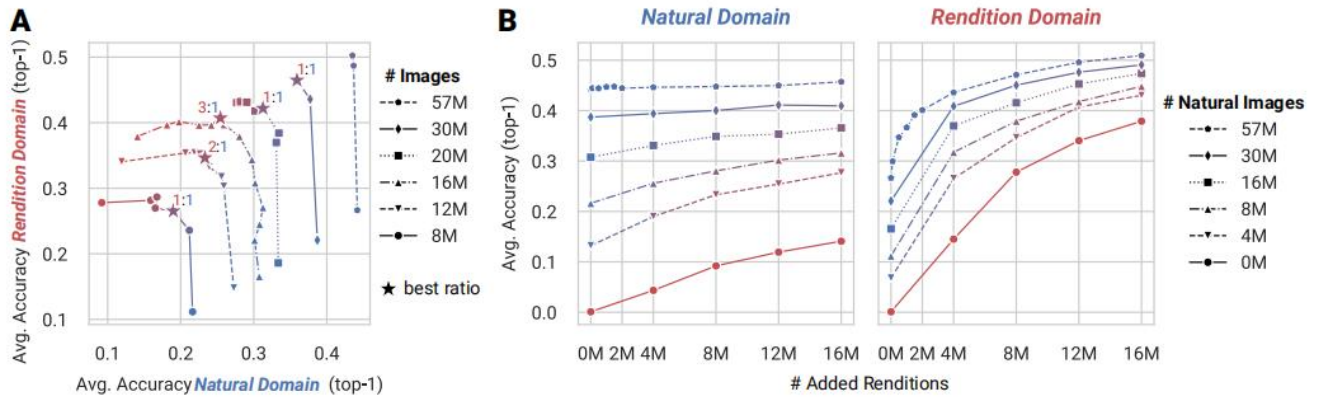


# 域污染多大程度上影响CLIP的领域泛化能力？

第一行在LAION-Natural上训练的模型，在clean rendition testsets上测试准确率比standard rendtion testsets要低，第四行同。

| Dataset | Standard Datasets top-1 Acc. | | Clean Datasets top-1 Acc. | |
|---|---|---|---|---|
| | *Natural* | *Rendition* | *Natural* | *Rendition* |
| LAION-Natural | 36.88 % | 21.98 % | 39.72 % | 17.81 % |
| LAION-Mix-13M | 37.28 % | 40.48 % | 38.97 % | 40.78 % |
| LAION-Mix-16M | 36.92 % | 41.46 % | 38.58 % | 42.07 % |
| LAION-Rand-57M | 37.62 % | 40.66 % | 36.99 % | 39.58 % |

A：控制训练集总图像数量，rendition域数据：natural域数据在1:1至3:1之间取得最佳
B：控制Natural图像数量，像其中逐步添加rendition图像，发现仅需少量rendition图像即可获得较大提升，且所需rendition图像数与初始Natural图像数相关。




结论：CLIP确实具有一定的领域泛化能力，但在之前的研究中被高估了

北京大学
PEKING UNIVERSITY

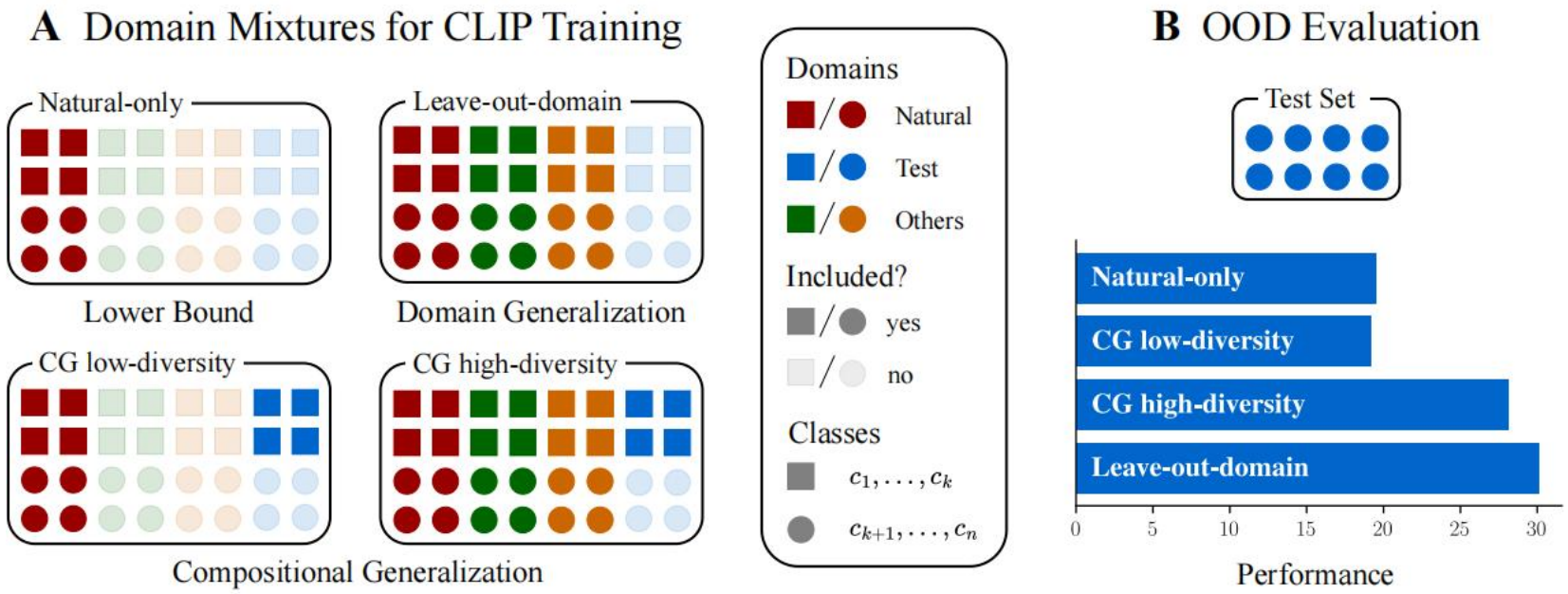
# When and How Does CLIP Enable Domain and Compositional Generalization?

Elias Kempf [*1] Simon Schrodi [*1] Max Argus [1] Thomas Brox [1]

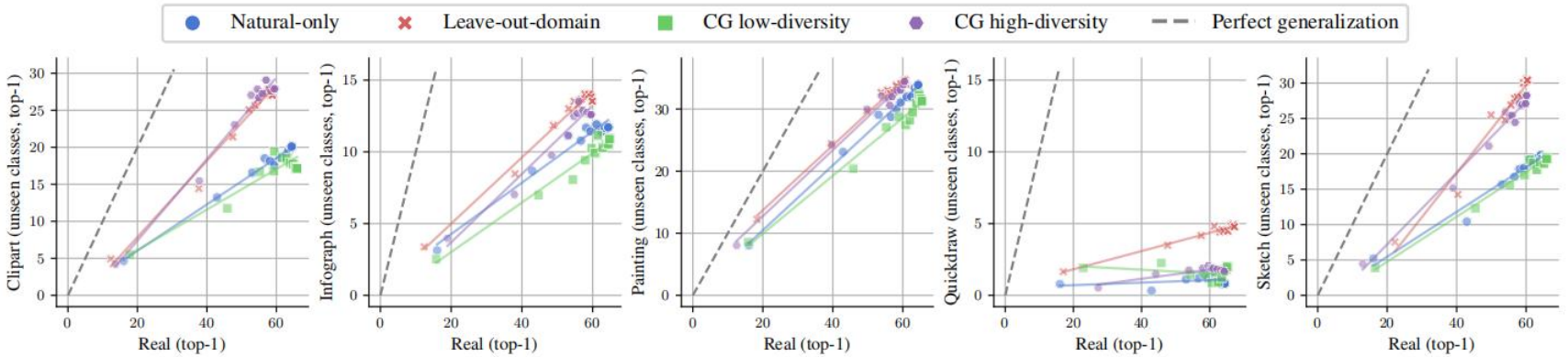
ICML 2025

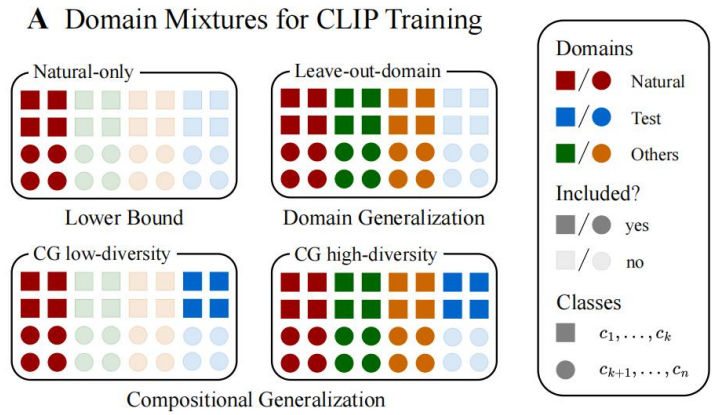
2025/09/20

Domain Generalization vs Compositional Generalization




模型：CLIP Resnet-50、ViT-S-32，Swin-T
基础数据集：ImageNet-Captions、CC3M、CC12M、DomainNet-real
领域数据集：Clipart、Infograph、painting、Quickdraw、Sketch（均来自DomainNet）

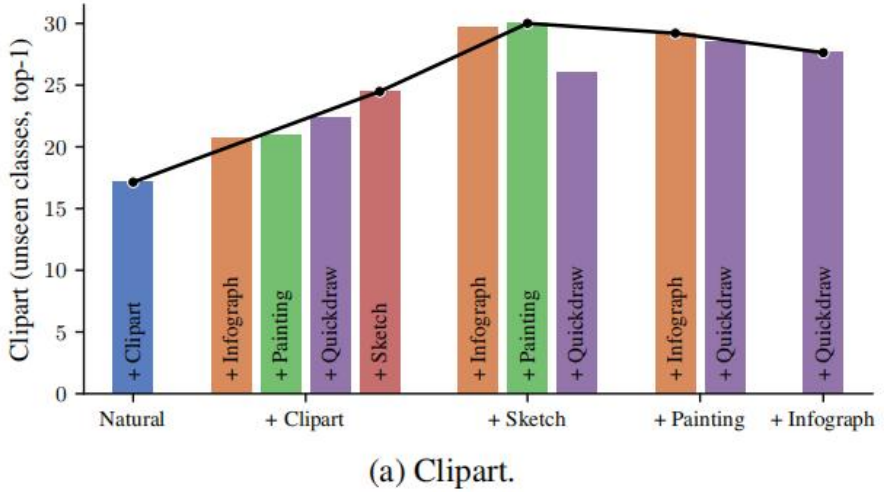



与低领域多样性场景（Natural-only、CG low-diversity）相比，
高领域多样性场景下的领域泛化和组合泛化都更强

每个domain对泛化能力的贡献相同吗?
向CG low-diversity中逐步添加领域多样性直至CG high-diversity，会发现不同的领域有不同的贡献

值得关注的是，Quickdraw和Sketch虽然在第二步提升最大，但当添加Sketch后，再添加Quickdraw，提升十分微小——可能原因二者存在视觉相似性



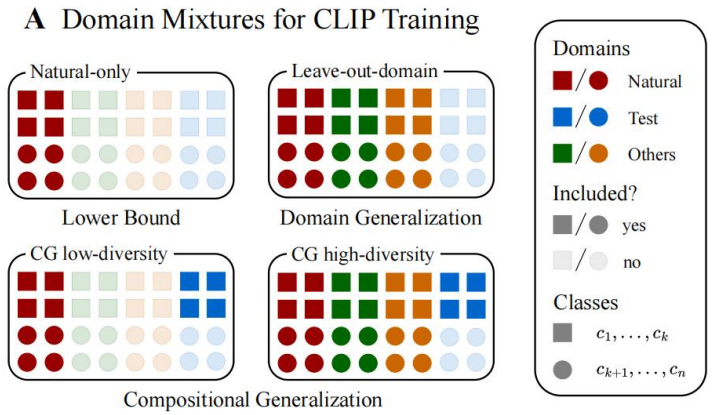

(a) Clipart.

猜想：CLIP学到了一种“捷径”，比如认为所有Sketch都属于某些已经见过的类

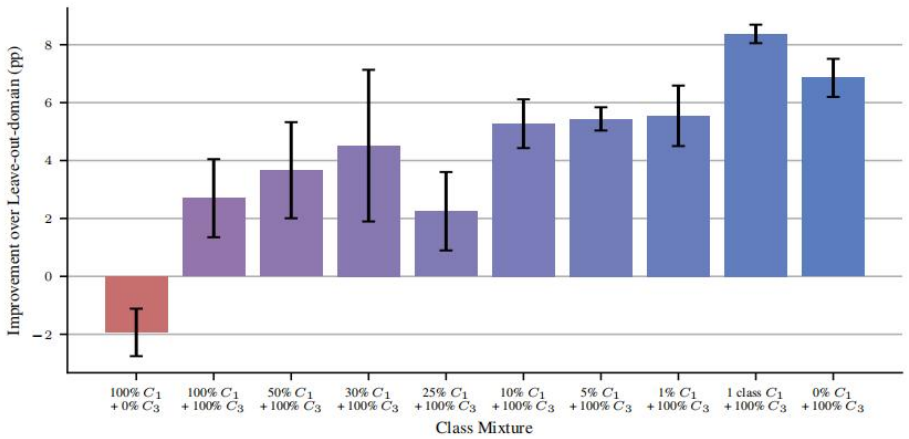
实验：找一个新数据集ImageNet-Sketch，移除其中所有与先前数据集重叠的类别，作为右图中蓝色正方形

结果：类不重叠确实提升了性能

| Training data setup (Figure 1) | Sketch |
|---|---|
| Natural-only | 19.5 |
| Leave-out-domain | 30.1 |
| CG low-diversity | 19.2 |
| w/ sketches of non-queried classes only | 27.1 (+7.9) |
| CG high-diversity | 28.1 |
| w/ sketches of non-queried classes only | 36.9 (+8.8) |




右图研究了类重叠程度对于组合泛化性能的影响——总体而言，重叠程度越低，表现越好



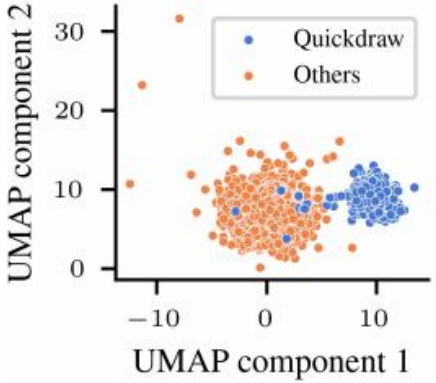

(b) Effect of different mixtures of the test domain's classes $C_1 \cup C_3$ seen during training.

在已见类上表现正常——说明能正确区分quickdraw不同的类别
在未见类上糟糕——无泛化能力

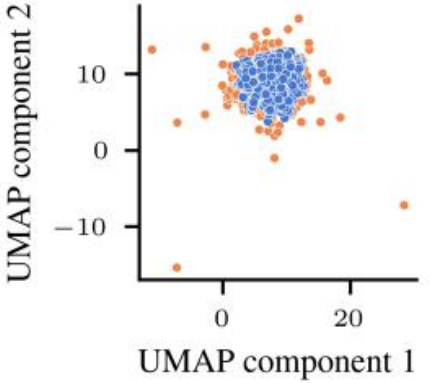
猜想1：captions存在bias，且对齐类别过于困难？导致模型专注于图像风格的统一（domain-specific），而不是图像本身类别的对齐（domain-invariant）

猜想2：虽然视觉embedding呈现对齐，但并不意味着产生嵌入的计算过程是对齐的——insufficient sharing of intermediate representations and circuits——CLIP可能为qiuckdraw学习了一个独立的路径，缺乏与其它领域的共享

| Captions | | Classes | |
|---|---|---|---|
| domain-specific | domain-invariant | seen | unseen |
| 50% | 50% | 50.7 | 1.7 |
| 0% | 100% | 46.2 | 0.7 |




(a) Domain-invariant and specific captions.




(b) Only domain-invariant captions.

提出machanistci similarity，度量底层回路的相似程度

实验证实Quickdraw域的中间特征相似性与共享神经元数量远低于其它域——机制相似性低

因此，组合泛化需要中间特征与底层回路的充分共享