

# 基于语言增强的图像新类别发现

**团队名字：**交个朋友

**比赛成绩：**初赛第二名，决赛第六名

初赛A榜评审

| 排名 | 团队        | 分数                | 提交次数 | 最佳成绩提交时间         | 最后提交时间           |
|----|-----------|-------------------|------|------------------|------------------|
| 1  | 交个朋友      | 75.1106302991214  | 17   | 2023/10/06 21:55 | 2023/10/06 21:55 |
| 2  | NJUST-KMG | 75.07193875558148 | 14   | 2023/10/06 15:28 | 2023/10/06 15:28 |
| 3  | 美奥大同      | 73.5403733316318  | 5    | 2023/10/05 21:48 | 2023/10/06 09:50 |

初赛B榜评审

| 排名 | 团队         | 分数                | 提交次数 | 最佳成绩提交时间         | 最后提交时间           |
|----|------------|-------------------|------|------------------|------------------|
| 1  | NJUST-KMG  | 75.34775675484535 | 2    | 2023/10/09 11:43 | 2023/10/09 11:43 |
| 2  | 交个朋友       | 74.81654053253102 | 2    | 2023/10/09 11:13 | 2023/10/09 11:13 |
| 3  | SparkSquad | 74.1465780546989  | 2    | 2023/10/09 11:55 | 2023/10/09 11:55 |

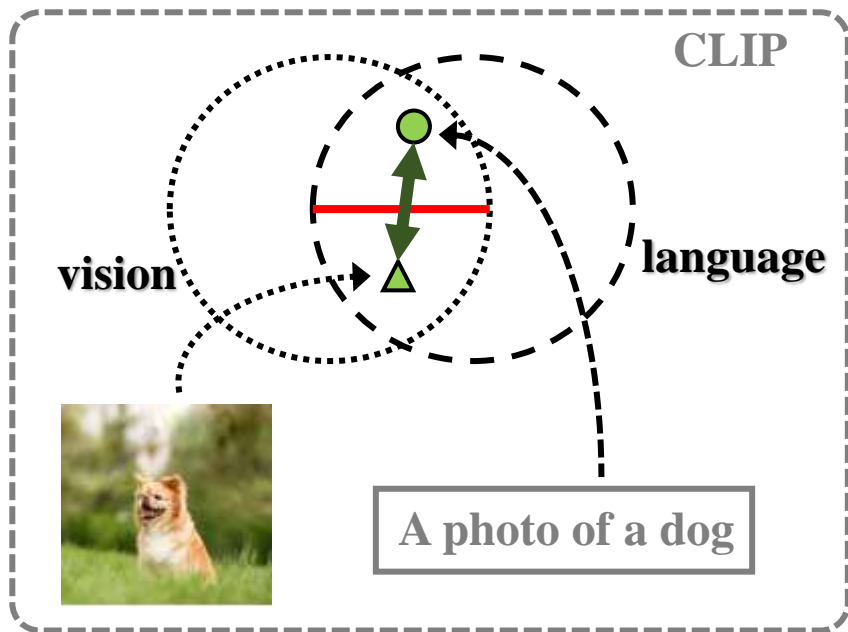
决赛A榜

| 排名 | 团队        | 分数                | 提交次数 | 最佳成绩提交时间         | 最后提交时间           |
|----|-----------|-------------------|------|------------------|------------------|
| 1  | 交个朋友      | 81.80186607246702 | 5    | 2023/11/18 11:58 | 2023/11/18 11:58 |
| 2  | NJUST-KMG | 81.73739904097388 | 5    | 2023/11/17 18:23 | 2023/11/17 18:23 |
| 3  | CB        | 81.0754647232672  | 7    | 2023/11/15 19:28 | 2023/11/15 19:28 |

|       | Score | Rank |
|-------|-------|------|
| 初赛-A榜 | 75.11 | 1    |
| 初赛-B榜 | 74.81 | 2    |
| 决赛-A榜 | 81.80 | 1    |
| 决赛-B榜 | 82.71 | 5    |

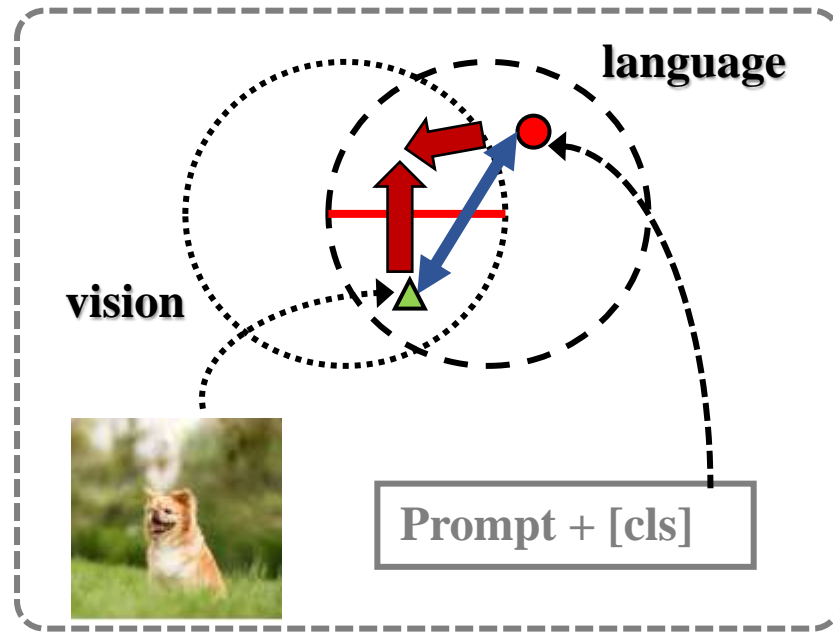
决赛B榜

| 排名 | 团队         | 分数                | 提交次数 | 最佳成绩提交时间         | 最后提交时间           |
|----|------------|-------------------|------|------------------|------------------|
| 1  | CB         | 84.38531131772201 | 2    | 2023/11/20 11:58 | 2023/11/20 11:58 |
| 2  | NJUST-KMG  | 84.0276988540994  | 2    | 2023/11/20 11:57 | 2023/11/20 11:57 |
| 3  | SparkSquad | 82.98573348196445 | 2    | 2023/11/20 11:41 | 2023/11/20 11:41 |
| 4  | MOCT       | 82.7255180067433  | 2    | 2023/11/20 11:45 | 2023/11/20 11:45 |
| 5  | 交个朋友       | 82.7006883400388  | 2    | 2023/11/20 11:45 | 2023/11/20 11:45 |



## 科学问题

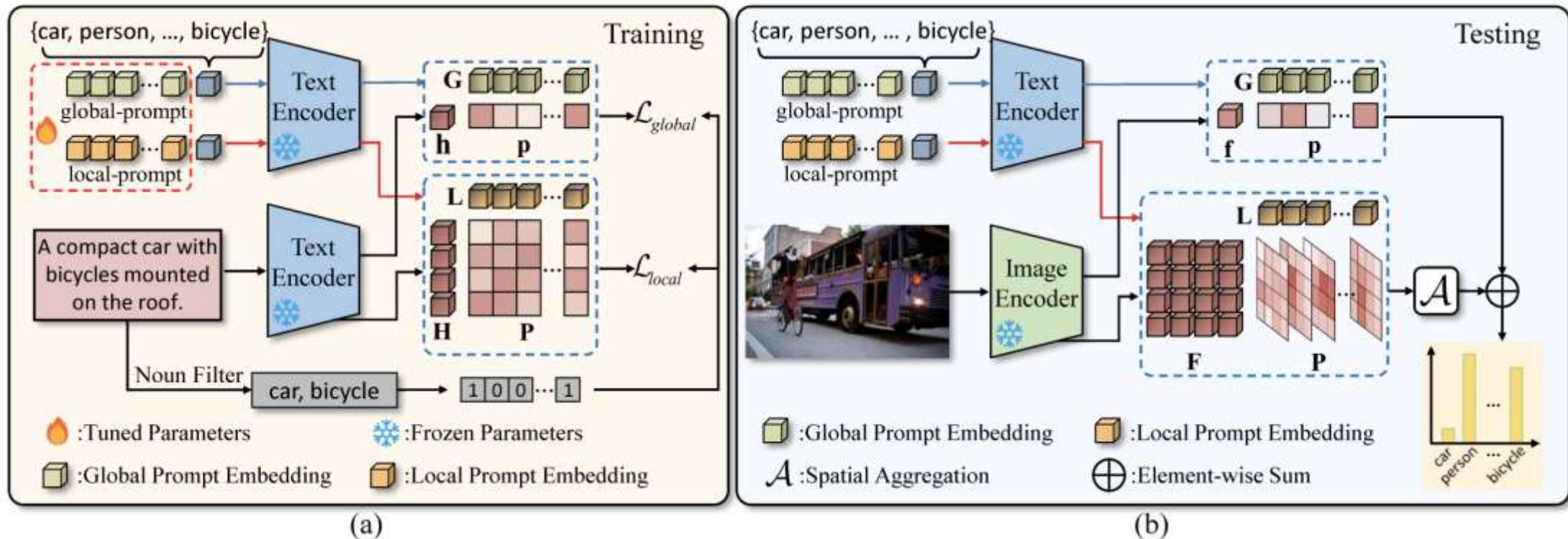
文本-图像之间存在模态差异



## 解决思路

训练：增加对文本的训练难度

推理：图像特征向文本模态靠拢



Text-as-Image (TaI) prompting 框架

该工作对我们解决方案启发：

1、学习文本到图像的可迁移参数

2、针对粗、细粒度特征的特点设计相适应的模型结构

- Prompt设计

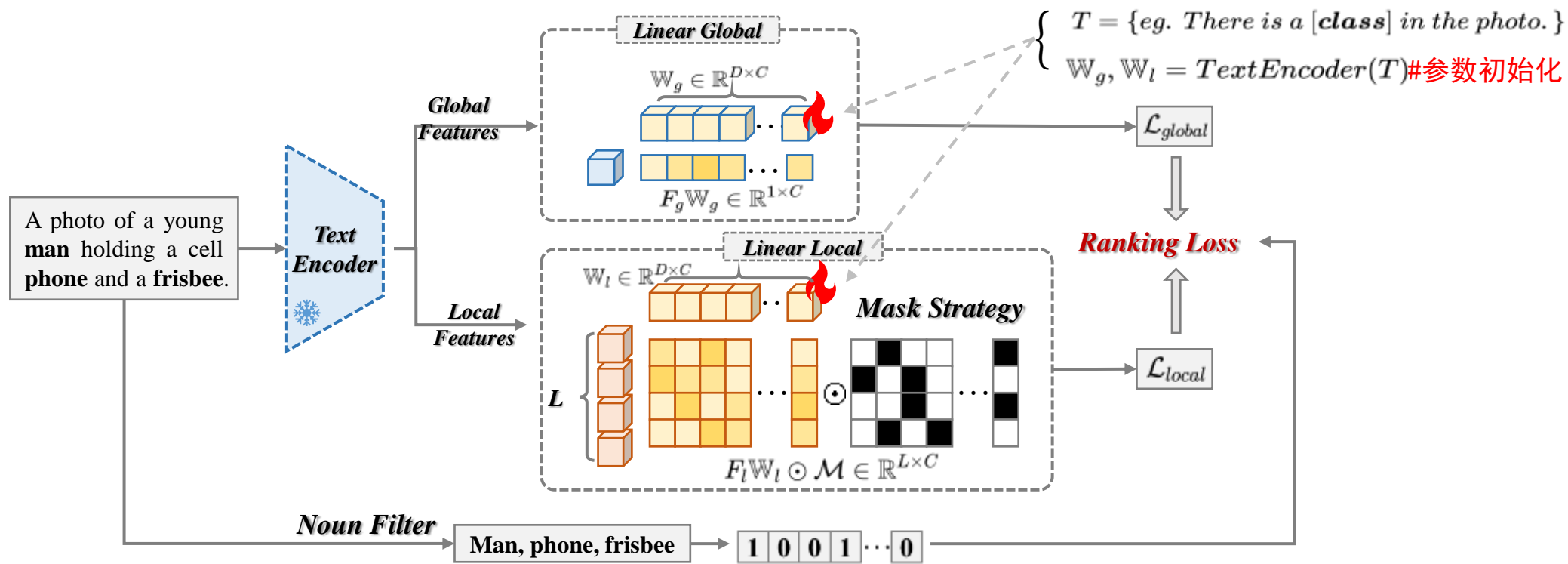
“Please start with ‘A photo of’ to generate three image captions, which should contain the following objects: [class name].

The scene of the caption should be as rich as possible, and the caption should not exceed fifty words.”

e.g., [class name] = ‘truck’ or ‘truck and person’

- 后处理

- “A photo of ” 句式提取
- 对同义词表进行关键词匹配，给文本数据打标签
- 获取了3w条的文本数据



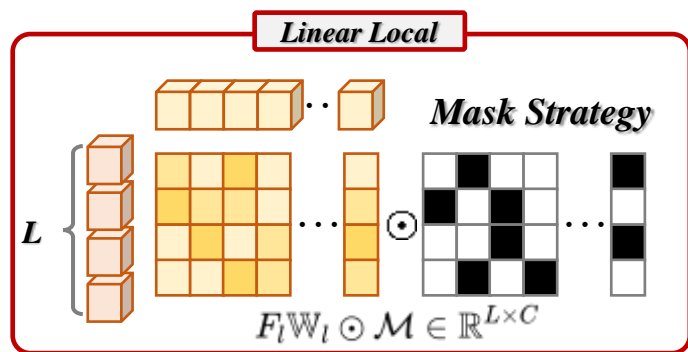
□ **主要思路**: 学习更高质量的细粒度特征表示

□ **模型创新**: 针对细粒度特征提出的分类器模型，新的结构包括:

(1) 带温度系数的注意力机制模块; (2) 特征的随机掩码模块

[1] Zhang, Yuhui et al. Diagnosing and Rectifying Vision Models using Language. ICLR'23

[2] Dunlap, Lisa et al. Using Language to Extend to Unseen Domains. ICLR'23



$Attention_t(Q, K, V) = \text{Softmax}(QK^T/t)V$  #带温度系数的注意力机制

$M \sim \text{Bernoulli}(p)$  #掩码矩阵采样

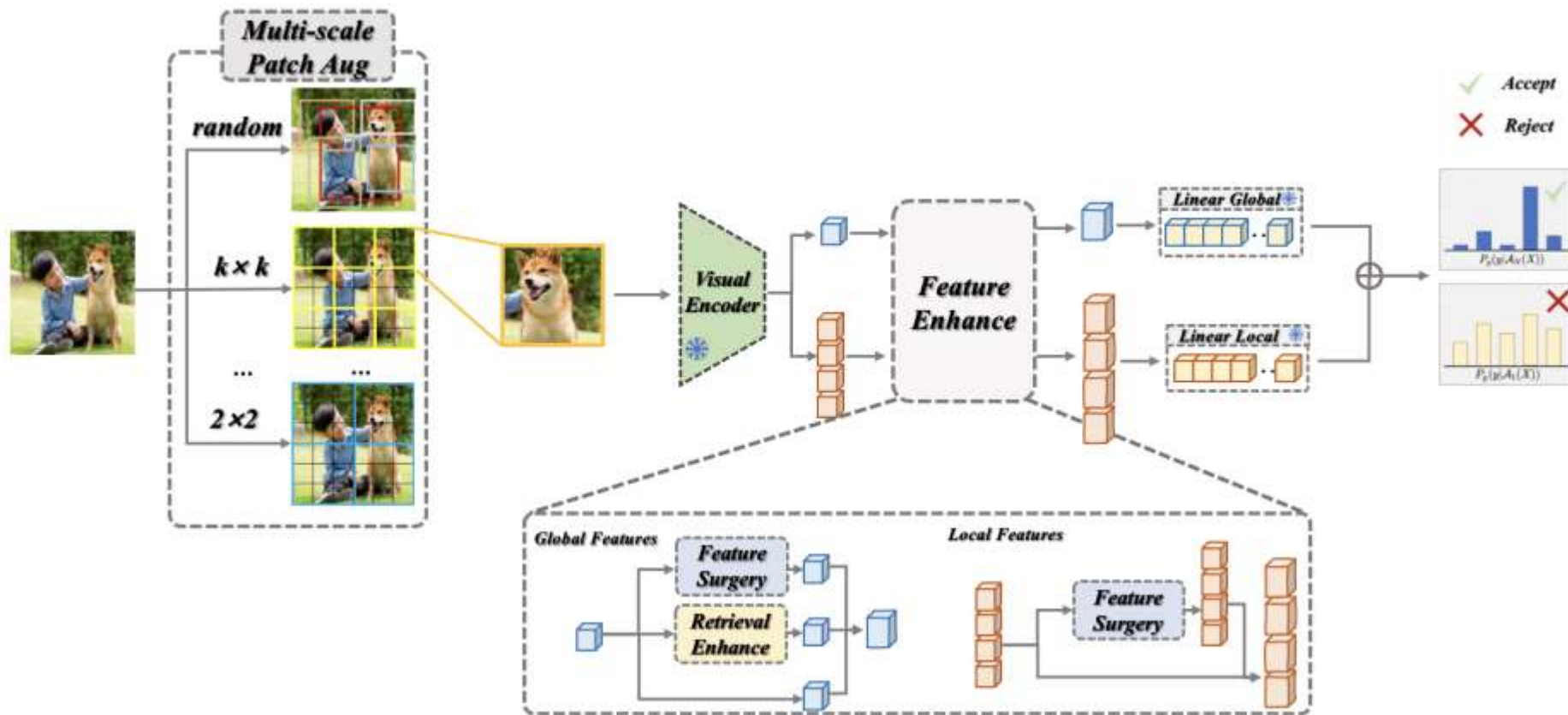
$\hat{W}_l = Attention_t(W_l, F_l \odot M, F_l \odot M)$

$Logits_l = \text{Cosine}(\hat{W}_l, W_l)$

### □ 模型优势:

- 注意力机制: 筛选出包含类判别信息的细粒度特征
- Mask策略: 防止对文本特征过拟合





➤ 主要思路

编码更细粒度的视觉特征

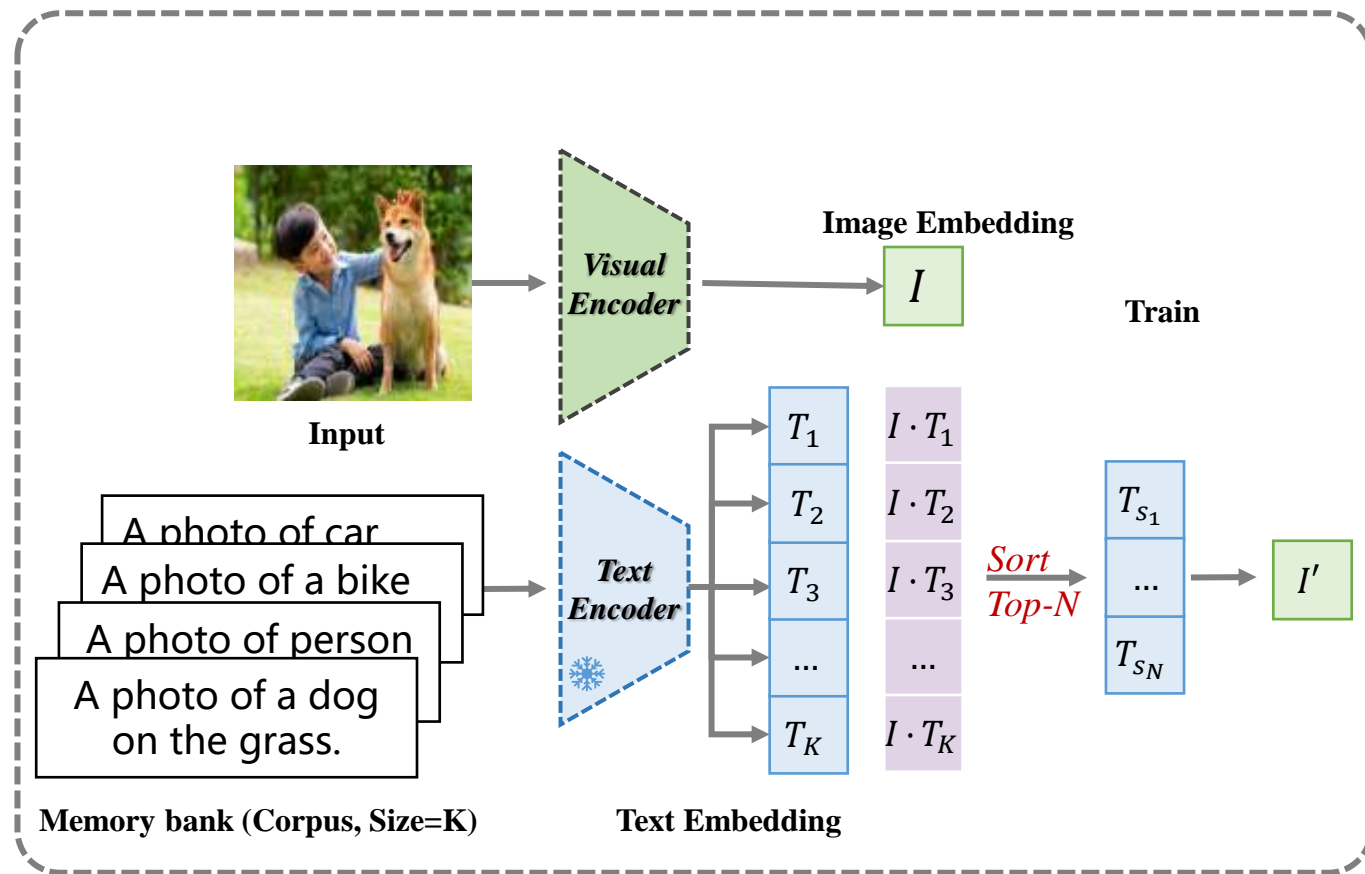
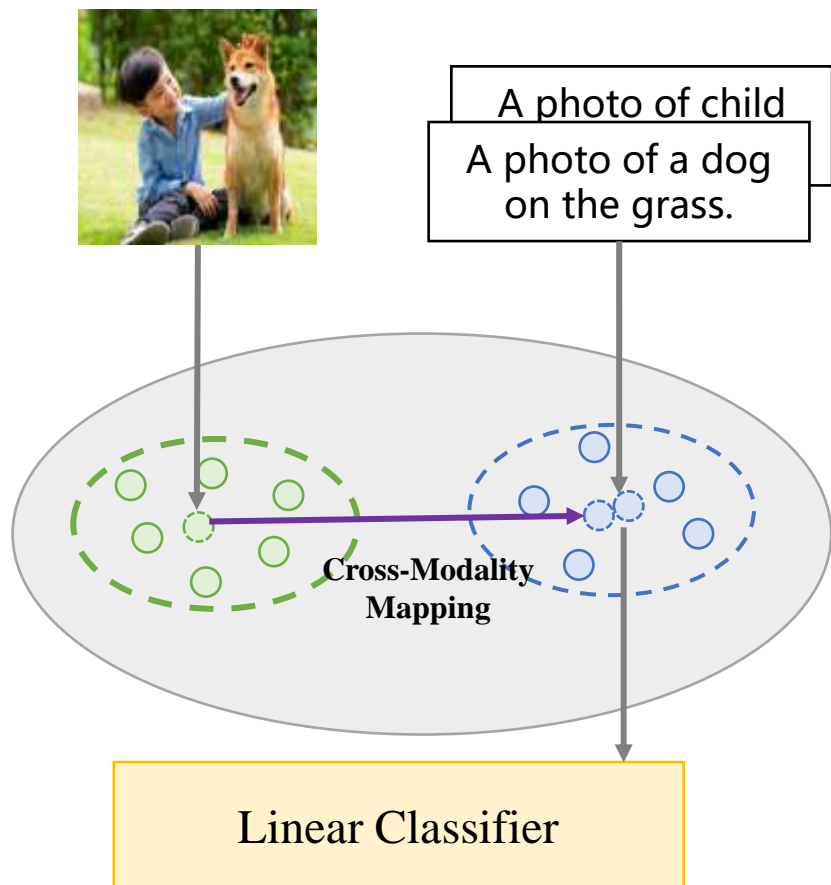
拉近图像特征和文本空间的距离

➤ 模型创新

多尺度的图像Patch

文本检索增强



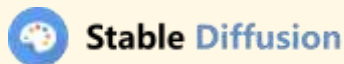


思路：使用文本代替图像进行分类→减小模态GAP

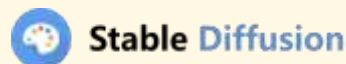
具体实现：N近邻文本检索→向量聚合



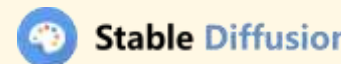
A photo of a group of friends enjoying a picnic lunch, with a delicious sandwich being the highlight of the meal.



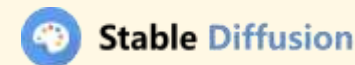
A photo of a group of people, including a person sitting at a keyboard and several others standing around, enjoying a meal at a restaurant.



A photo of a cozy living room, with a comfortable sofa and a set of large windows that face out into the countryside.



A photo of a person holding a cell phone in their hand, looking off into the distance as they walk through a busy city street, with a street performer in the background.



- 实验：在COCO2014的val数据上的多标签分类结果
- 比较的基准方法：TaI<sup>[1]</sup>
- 消融实验设置：训练阶段的改进评估（Linear），推理阶段的改进评估（TTA）

|                          | mAP   | $\Delta$      |
|--------------------------|-------|---------------|
| TaI [1]                  | 65.18 | -             |
| Linear                   | 69.05 | <b>+3.87</b>  |
| Linear <sub>TTA</sub>    | 74.01 | <b>+8.83</b>  |
| zero-shot*               | 68.08 | <b>+2.90</b>  |
| zero-shot <sub>TTA</sub> | 73.80 | <b>+5.72</b>  |
| Merge                    | 75.95 | <b>+10.77</b> |

实验结果

|                 | Time(H) | Memory(G) |
|-----------------|---------|-----------|
| Text generation | 24      | 30        |
| Training        | 0.083   | 8         |
| Testing         | 1.25    | 12        |

算法效率分析

[1] Guo Z, Dong B, Ji Z, Bai J, Guo Y, Zuo W. Texts as images in prompt tuning for multi-label image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 (pp. 2808-2817).

## □ 模型创新

### ➤ 训练阶段：

带温度系数的注意力机制、特征掩码模块

### ➤ 推理阶段：

多尺度patch的数据增强、文本检索增强

## □ 可提升方向

- 本方案只采用了简单的训练语料，可以向大语言模型的查询更丰富的文本知识。

**感谢各位评委老师的批评指正**