基于语言增强的图像新类别发现

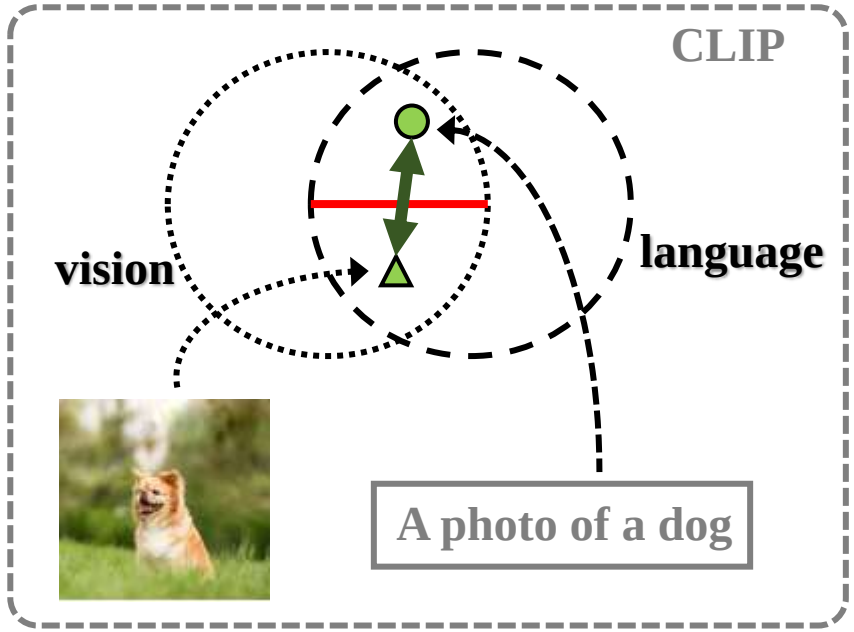
团队名字：交个朋友

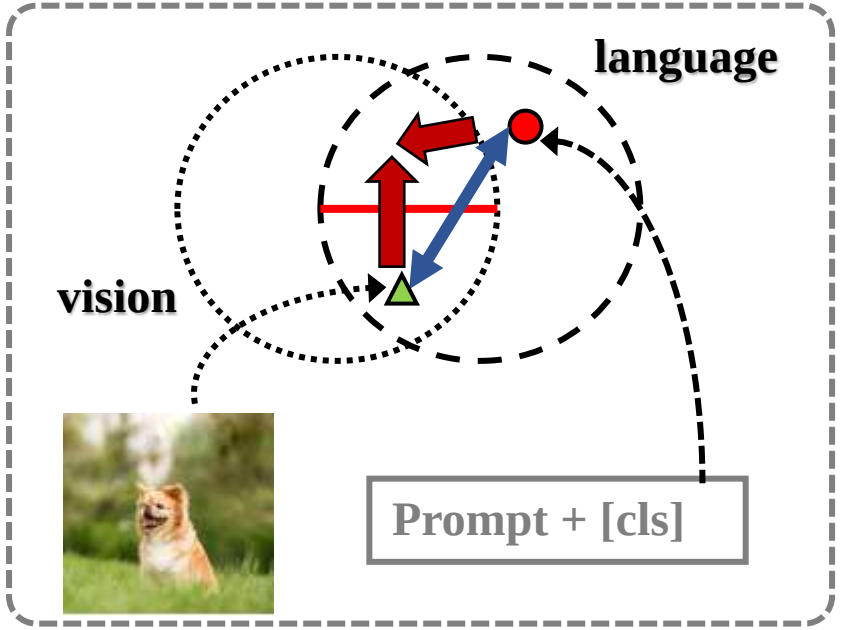
比赛成绩：初赛第二名，决赛第六名

团队成绩

	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

1、赛题挑战&解决思路





科学问题

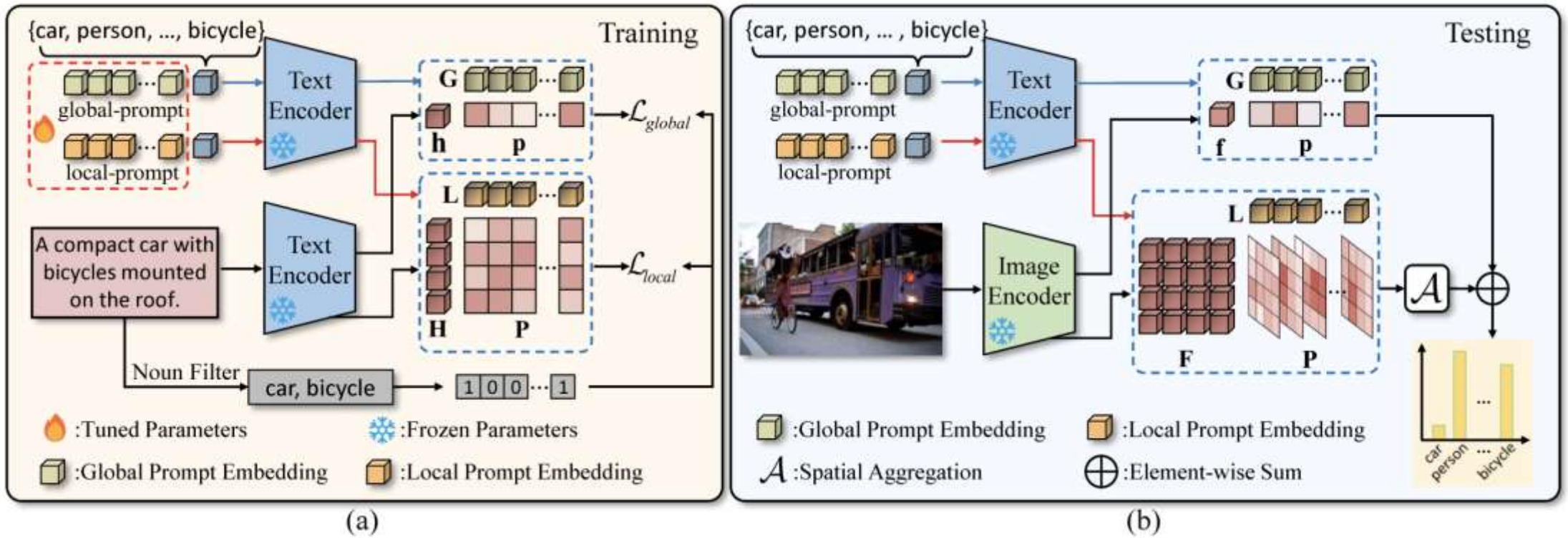
文本-图像之间存在模态差异

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

2、相关工作



Text-as-Image (Tal) 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

[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).

3.1 文本训练数据生成

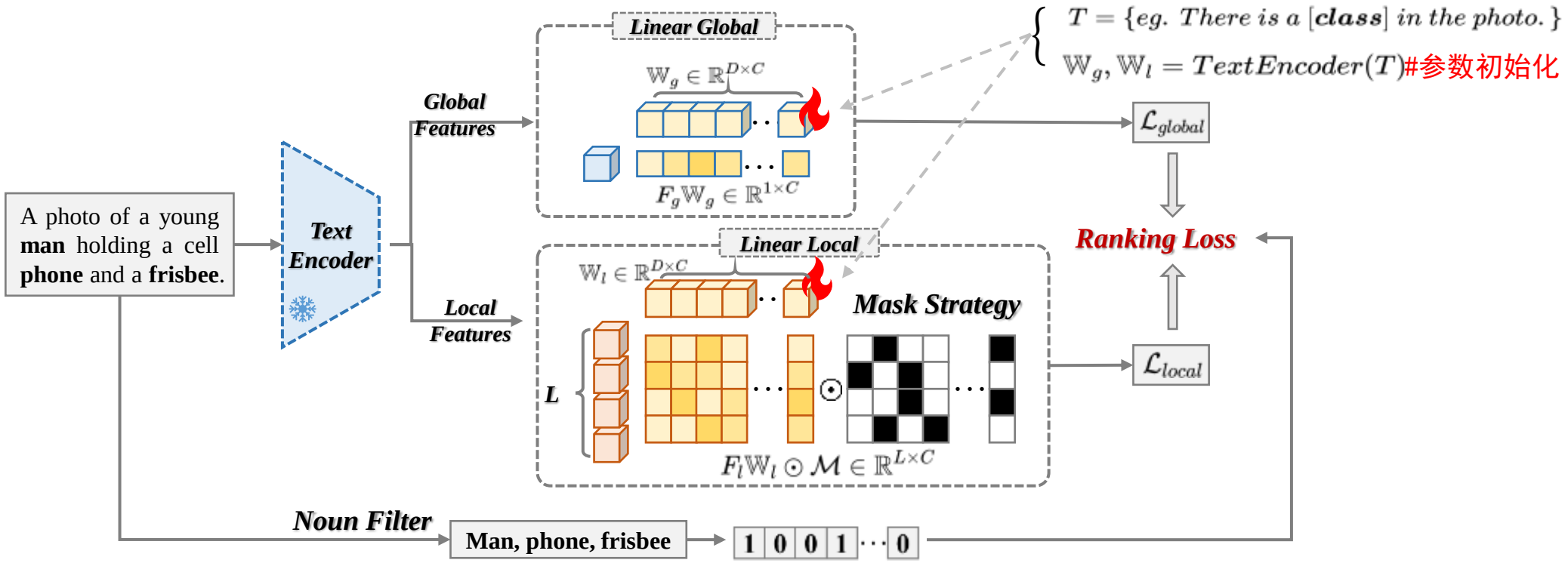
- Prompt设计

  “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

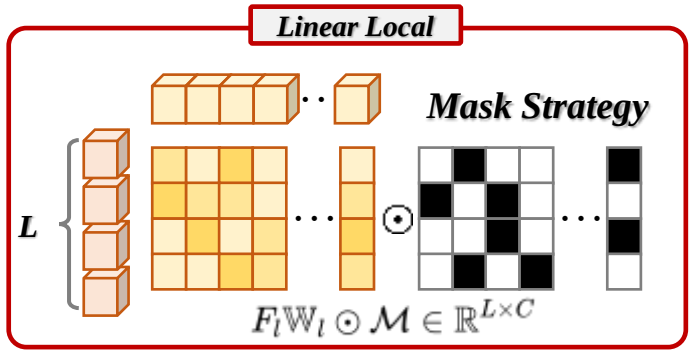
3.2 训练框架



- **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



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

$$M \sim Bernoulli(p)$$ #掩码矩阵采样

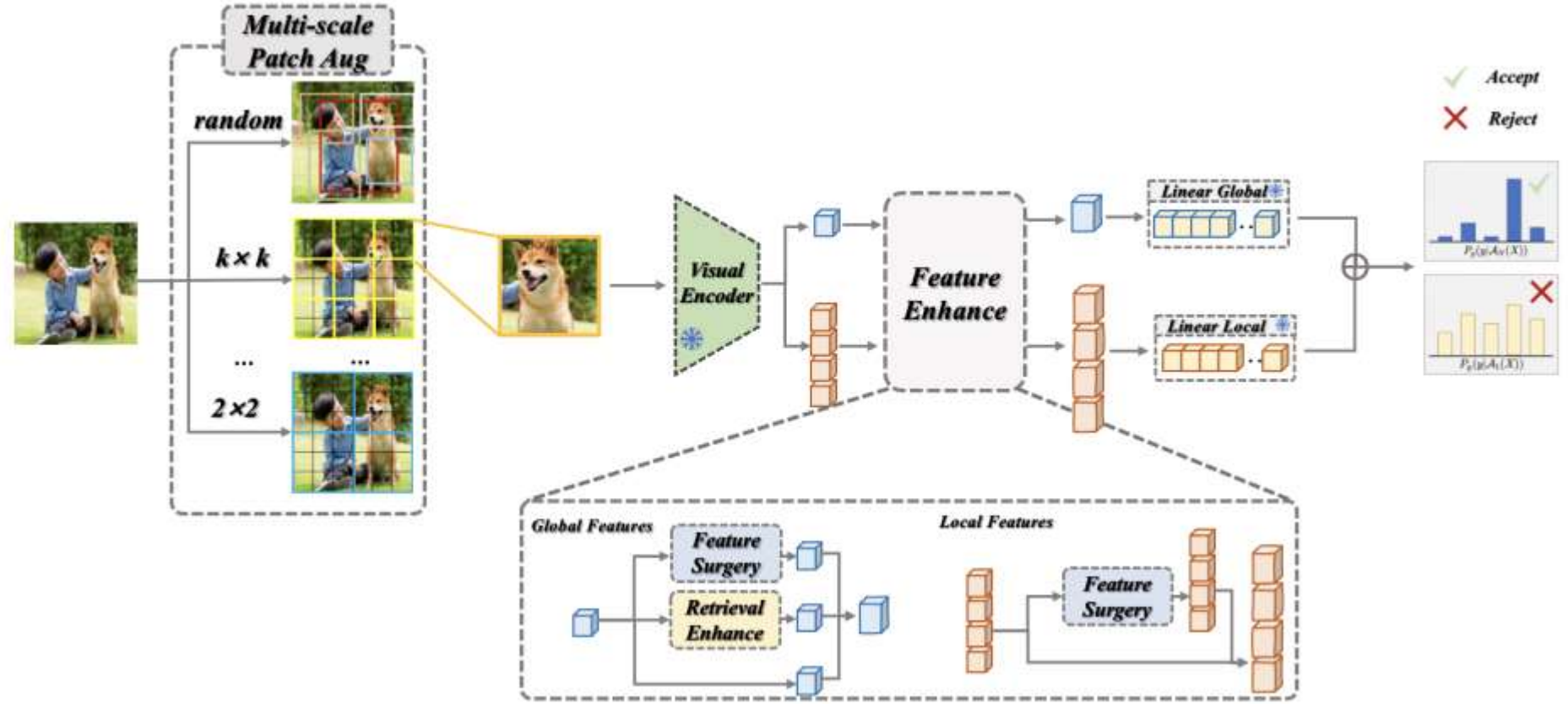
$$\hat{\mathbb{W}}_l = Attention_t(\mathbb{W}_l, F_l \odot M, F_l \odot M)$$

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

□ 模型优势：

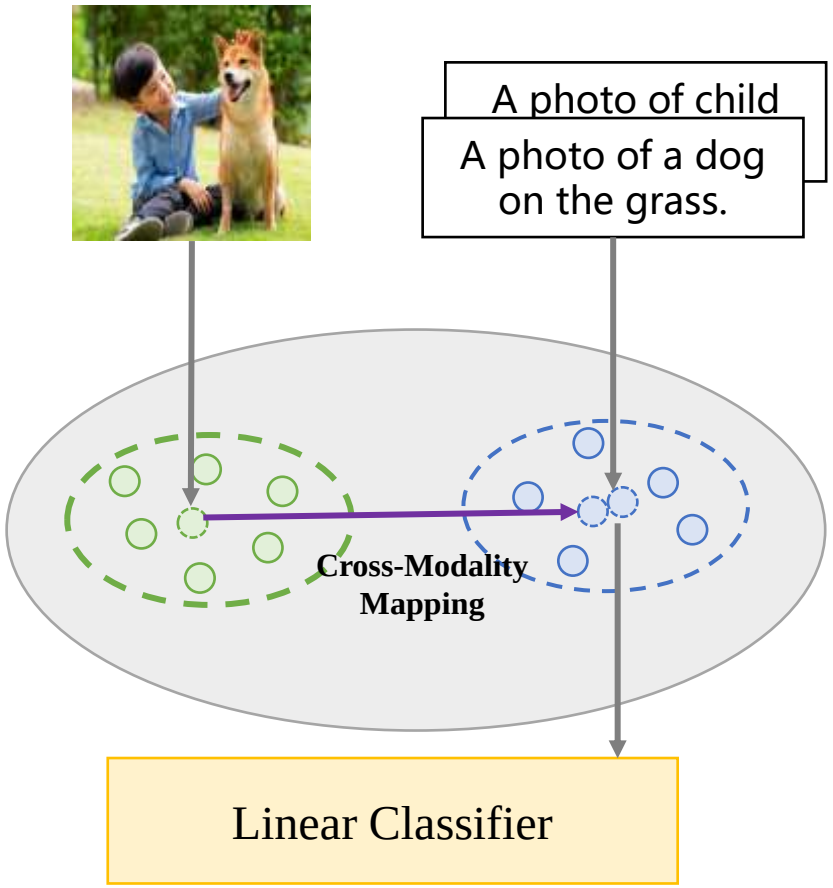
- **注意力机制：** 筛选出包含类判别信息的细粒度特征
- **Mask策略：** 防止对文本特征的过拟合

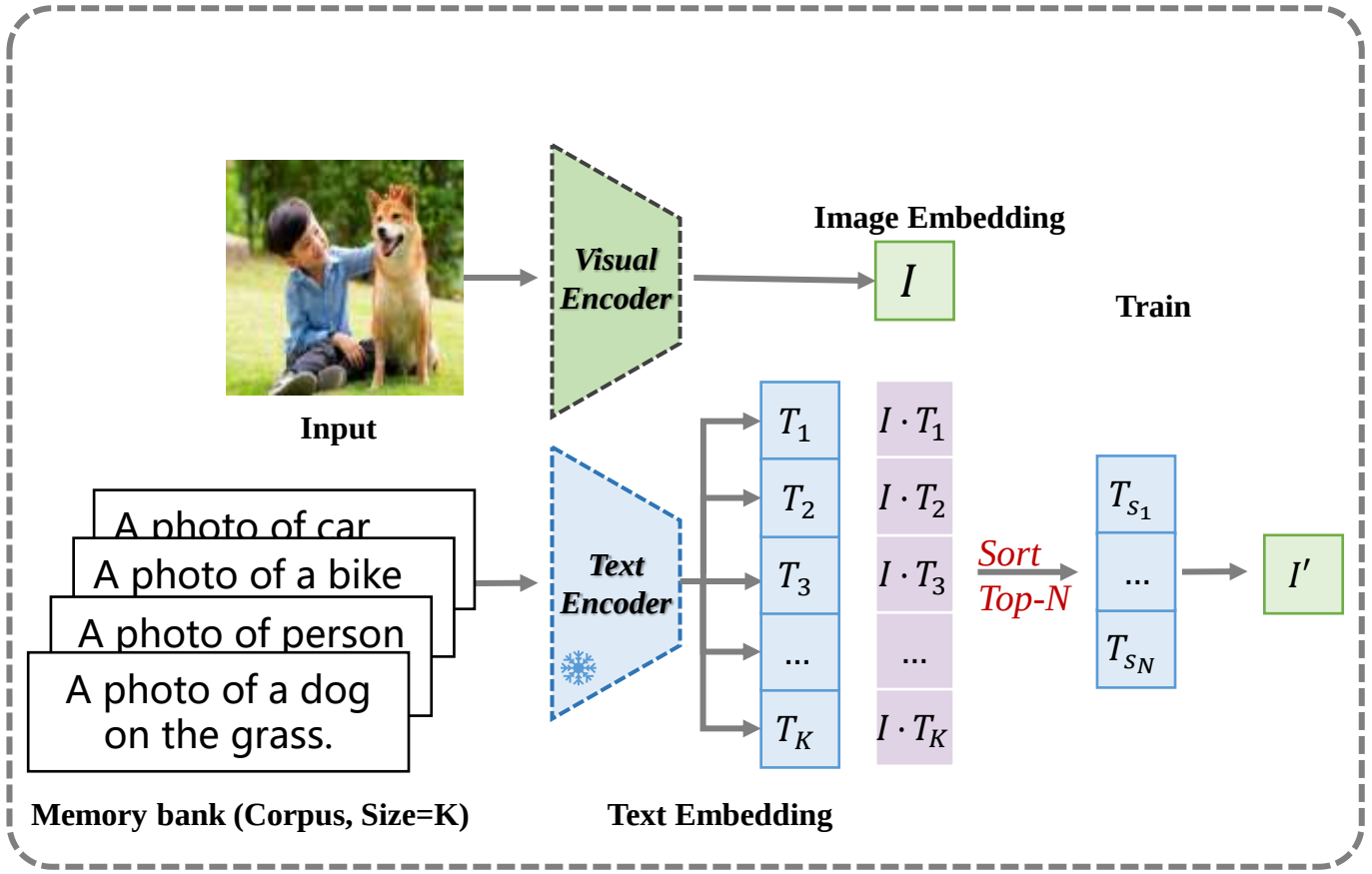
推理框架



- **主要思路**
  - 编码更细粒度的视觉特征
    - 多尺度的图像Patch
  - 拉近图像特征和文本空间的距离
    - 文本检索增强
- **模型创新**

推理框架：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

文本数据可视化

ChatGLM

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

Stable Diffusion

ChatGLM

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.

Stable Diffusion

ChatGLM

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

Stable Diffusion

ChatGLM

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.

Stable Diffusion

实验结果

- 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[1]
- 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	Δ
TaI [1]	65.18	-
Linear	69.05	**+3.87**
$\text{Linear}_{\text{TTA}}$	74.01	**+8.83**
zero-shot*	68.08	**+2.90**
$\text{zero-shot}_{\text{TTA}}$	73.80	**+5.72**
Merge	75.95	**+10.77**

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).

总结

口 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

口 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