Language-driven Cross-modal Classifier for Zero-shot Multi-label Image Recognition

Yicheng Liu [1] Jie Wen [*1] Chengliang Liu [1] Xiaozhao Fang [*2] Zuoying Li [3] Yong Xu [1] Zheng Zhang [*1]

[1] Harbin Institute of Technology, Shenzhen
[2] Guangdong University of Technology
[3] Minjiang University
[*] Corresponding author

Background

CLIP has show impressive zero-shot transfer capabilities.

Transfer the capability of CLIP to multi-label recognition (MLR) faces two challenges:

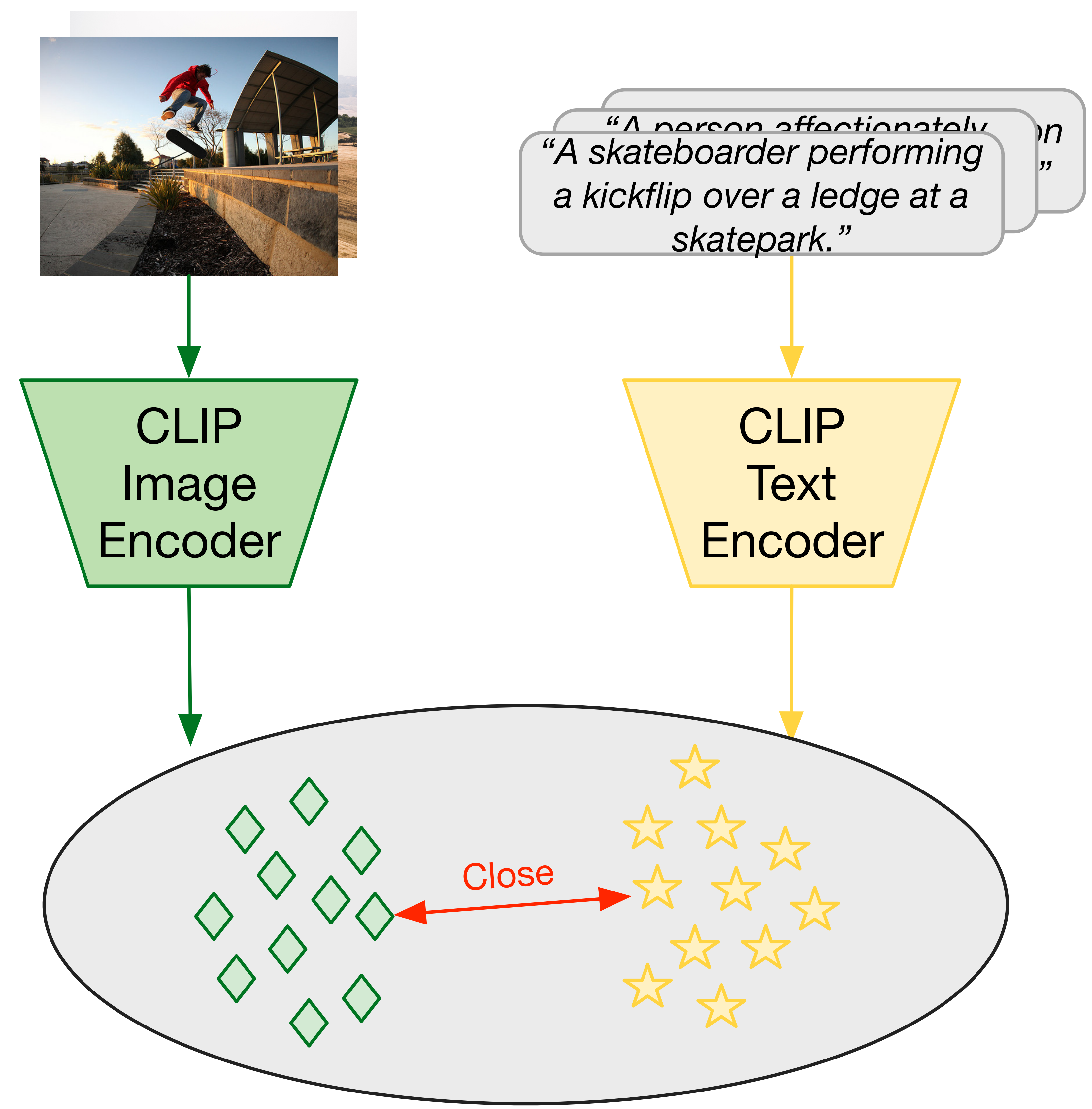
- Collecting sufficient multi-label annotated image data in real-world application is challenging and not scalable.

- CLIP only focuses on matching each image with a single label during its training, hence it is not suitable to handle the multi-label cases.

Motivation

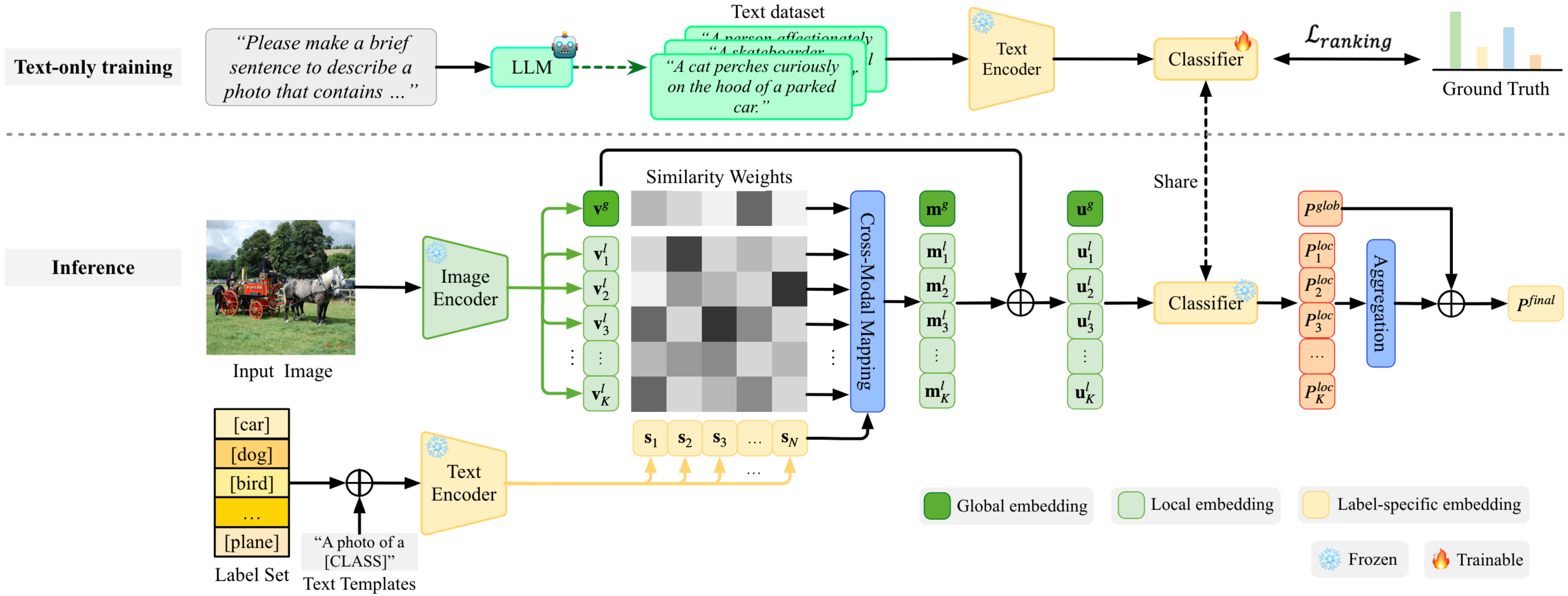
- Pretrained vision-language model learned an shared multi-modal embedding space via contrastive learning.

- Language data is much easier to collect. Large Language Model (LLM) can generate a large scale multi-label language dataset automatically.



Multi-modal shared embedding space

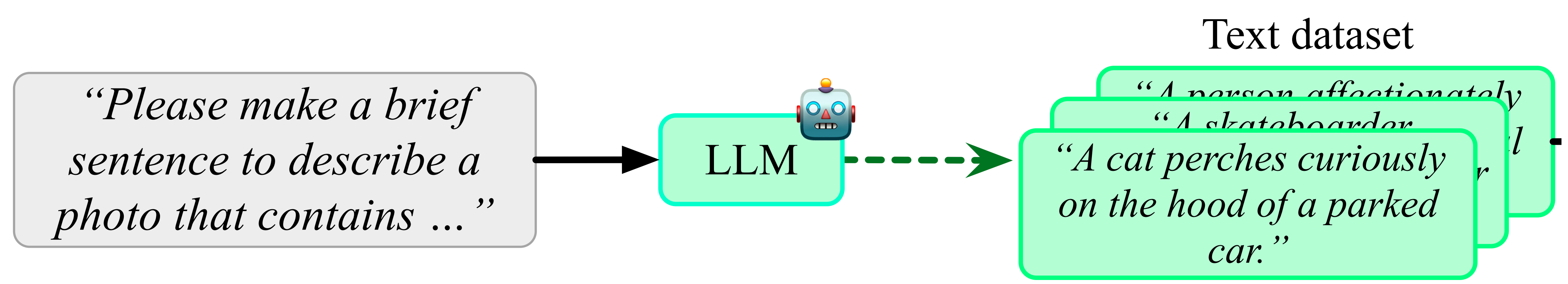
Method Overview



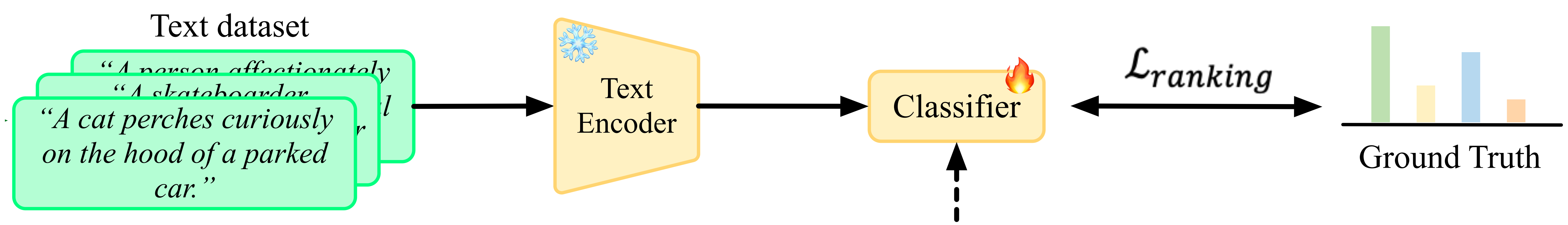
Method

ICML
International Conference
On Machine Learning

- Text data generation



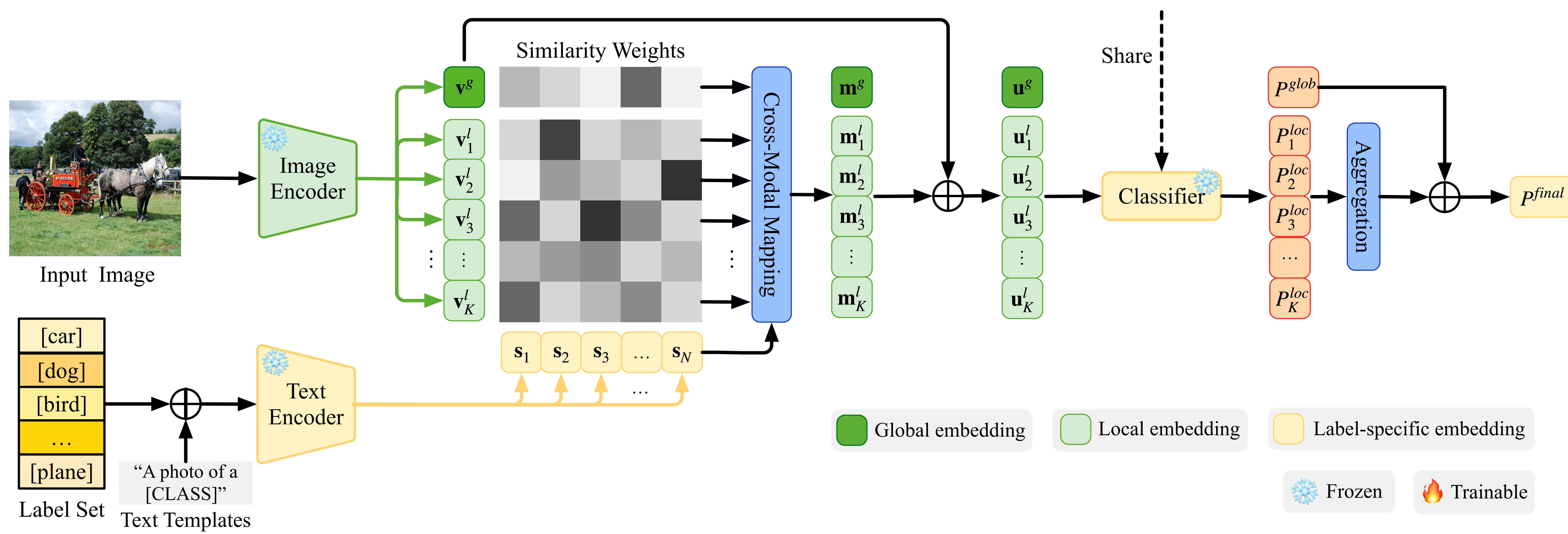
- Text only training



$$\mathcal{L}_{ranking} = \sum_{j \in \{c^+\}} \sum_{k \in \{c^-\}} \max(0, m - (p_{i,j} - p_{i,k}))$$

Method

- Inference stage
  - ➢ Cross-modal mapping
  - ➢ Fine-grained image embeddings



Experiments

- Our method shows good results on both zero-shot and few-shot MLR tasks.

Table 1. Comparison with zero-shot learning methods without image training on MS-COCO, VOC2007, and NUS-WIDE. The evaluation is based on mAP (%).

Method	MS-COCO	VOC2007	NUS-WIDE
Zero-shot CLIP	47.3	76.2	36.4
CLIP-DPT	49.7	77.3	37.4
TaI-DPT	65.1	88.3	46.5
CoMC	**68.7**	**89.4**	**48.2**

Table 2. Comparison with related multi-label zero-shot learning methods with image training on the NUS-WIDE dataset. We report the results in terms of mAP, as well as precision (**P**), recall (**R**), and **F1** score at $K \in \{3, 5\}$.

Method	Top-3 P	Top-3 R	Top-3 F1	Top-5 P	Top-5 R	Top-5 F1	mAP
CONSE (Norouzi et al., 2013)	17.5	28.0	21.6	13.9	37.0	20.2	9.4
LabelEM (Akata et al., 2015)	15.6	25.0	19.2	13.4	35.7	19.5	7.1
Fast0Tag (Zhang et al., 2016)	22.6	36.2	27.8	18.2	48.4	26.4	15.1
One Attention per Label (Kim et al., 2018)	20.9	33.5	25.8	16.2	43.2	23.6	10.4
LESA (M=10) (Huynh & Elhamifar, 2020)	25.7	41.1	31.6	19.7	52.5	28.7	19.4
BiAM (Narayan et al., 2021)	-	-	33.1	-	-	30.7	26.3
SDL (M=7) (Ben-Cohen et al., 2021)	24.2	41.3	30.5	18.8	53.4	27.8	25.9
MKT (He et al., 2023)	27.7	44.3	34.1	21.4	57.0	31.1	37.6
DualCoOp (Sun et al., 2022)	**37.3**	46.2	**41.3**	**28.7**	59.3	**38.7**	43.6
CoMC	33.5	**53.5**	41.2	24.8	**66.1**	36.1	**48.2**

Table 3. Comparison with multi-label few-shot methods on VOC2007 and MS-COCO. The evaluation is based on mAP (%) for 0-shot, 1-shot, 2-shot, 4-shot, 8-shot, and 16-shot with treating all classes as novel classes.

Method	VOC2007 0-shot	1-shot	2-shot	4-shot	8-shot	16-shot	MS-COCO 0-shot	1-shot	2-shot	4-shot	8-shot	16-shot
CoOp	-	79.3	83.2	83.8	84.5	85.7	-	52.6	57.3	58.1	59.2	59.8
CoOp-DPT	-	83.2	88.1	88.2	90.0	90.1	-	65.8	66.2	67.6	68.1	68.9
CoMC	**89.4**	**89.7**	**90.1**	**90.6**	**91.4**	**92.1**	**68.7**	**68.9**	**69.3**	**70.4**	**70.9**	**71.4**

Thank you!

The full paper can be found here.
Code is available at https://github.com/yic20/CoMC