FLAME: Frozen Large Language Models Enable Data-Efficient Language-Image Pre-training

Anjia Cao[1], Xing Wei[1], Zhiheng Ma[2,3,4*]

[1]School of Software Engineering, Xi'an Jiaotong University

[2]Shenzhen University of Advanced Technology

[3]Guangdong Provincial Key Laboratory of Computility Microelectronic

[4]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

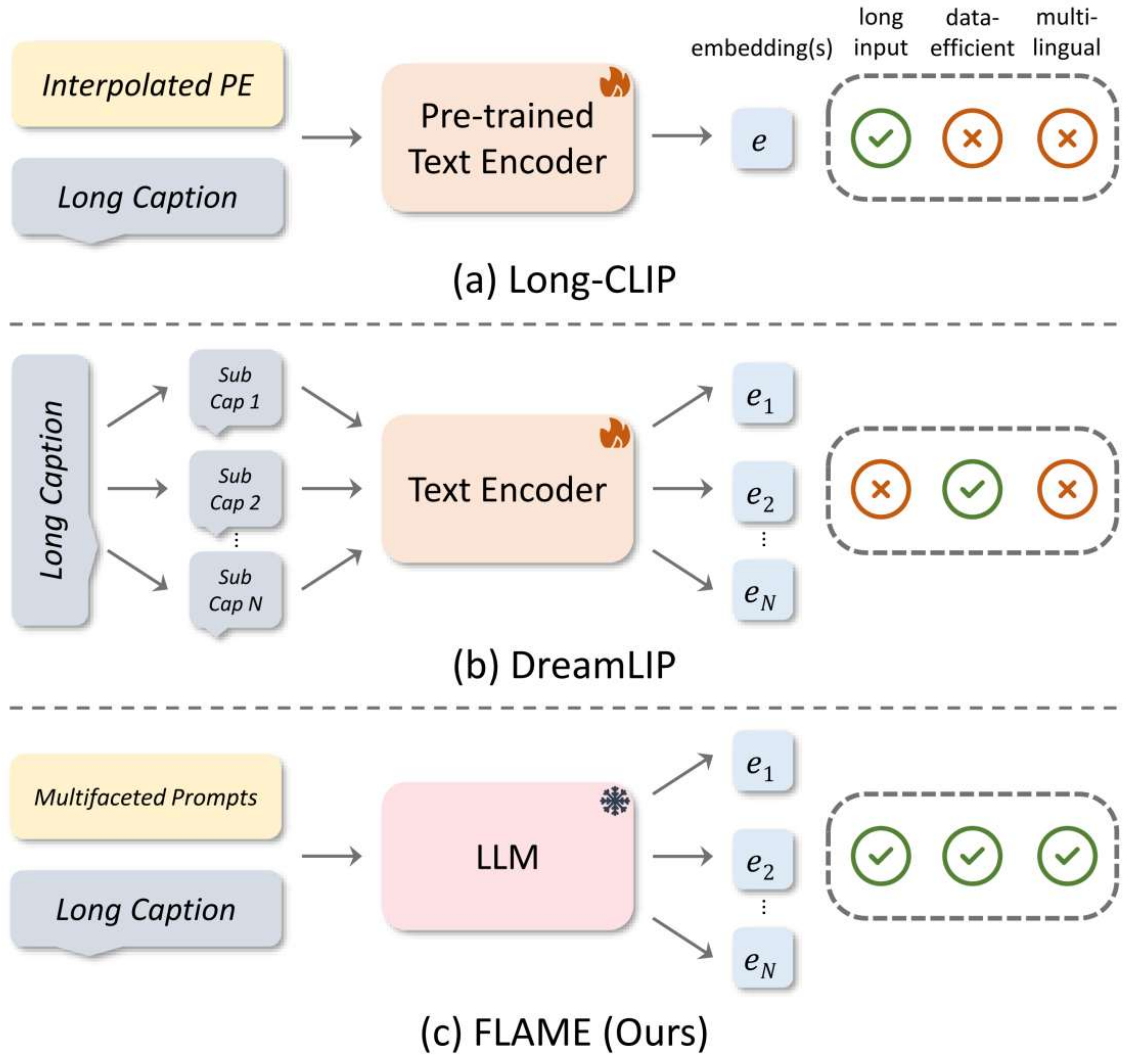
The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2025

Motivation

Challenges in Language-Image Pre-training

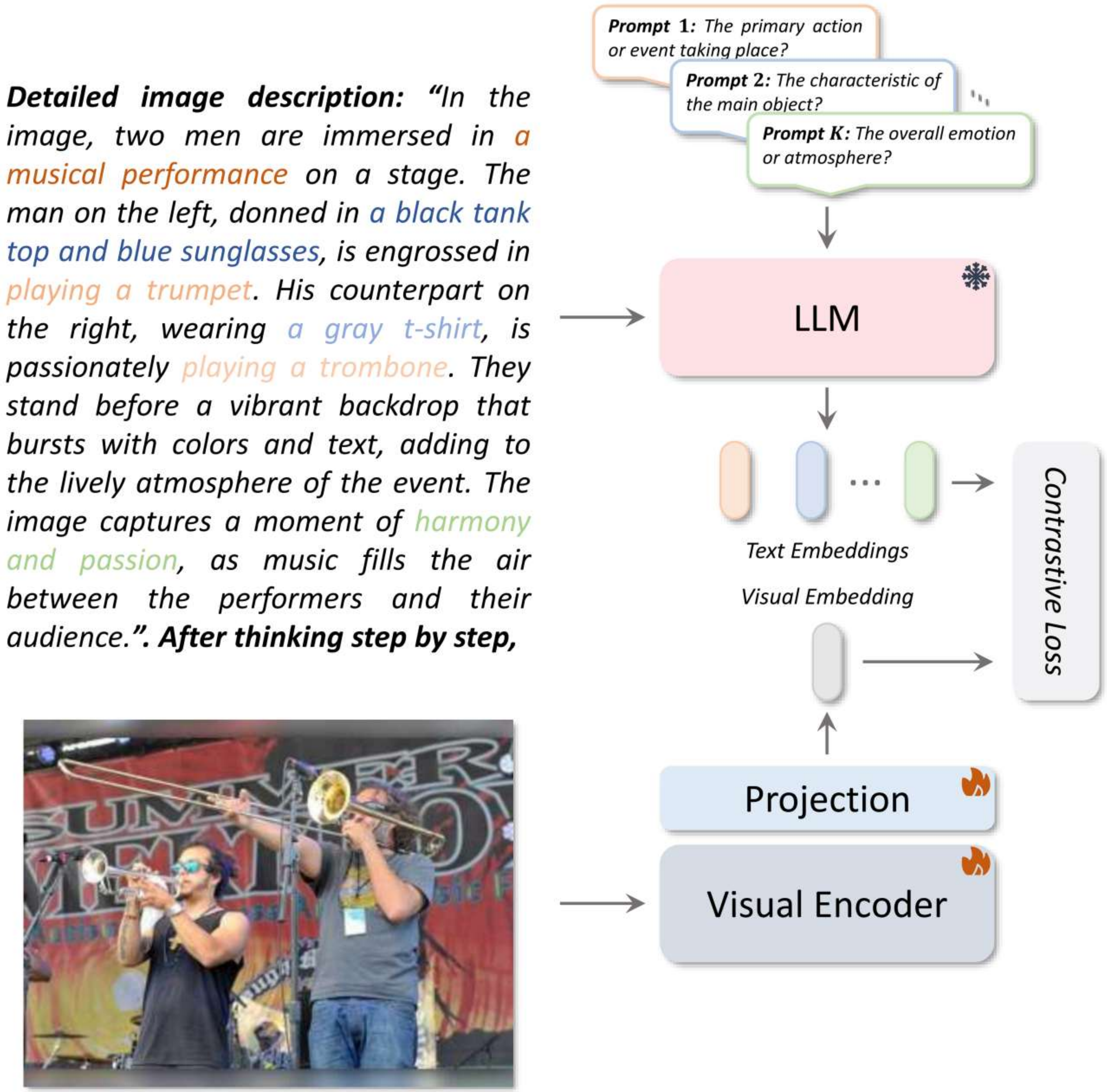
- **Data Scarcity:** High-quality image-text pairs (e.g., long text, multi-lingual) are rare.
- **Constrained Text Input Length:** Standard CLIP-style text encoders choke on long texts (>77 tokens).
- **Limitations in Prior Works:** Approaches such as long text decomposition and positional encoding interpolation fail to resolve the core limitation imposed by the text encoder's capacity bottleneck.
- **Conventional Wisdom:** Frozen text encoders lead to suboptimal performance.

Leveraging Frozen LLMs as Text Encoders



FLAME: Frozen Large Language Models Enable Data-Efficient Language-Image Pre-training

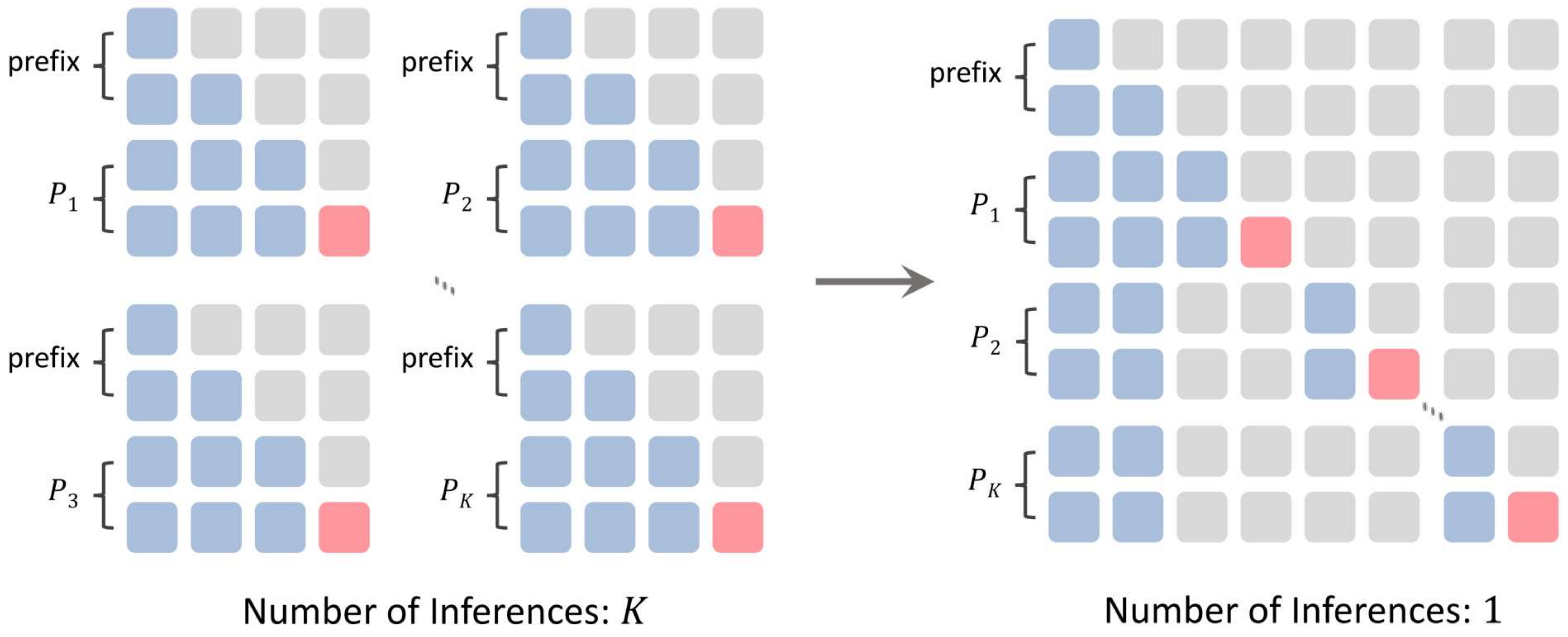
Multifaceted Prompt Distillation



- **Objective:** To capture diverse semantic representations from long captions, aligning better with the multifaceted nature of images.
- **Mechanism:** Employing a set of structured prompts targeting different semantic levels (e.g., entity, interaction, scene) to guide the frozen LLM in extracting distinct features.

FLAME: Frozen Large Language Models Enable Data-Efficient Language-Image Pre-training

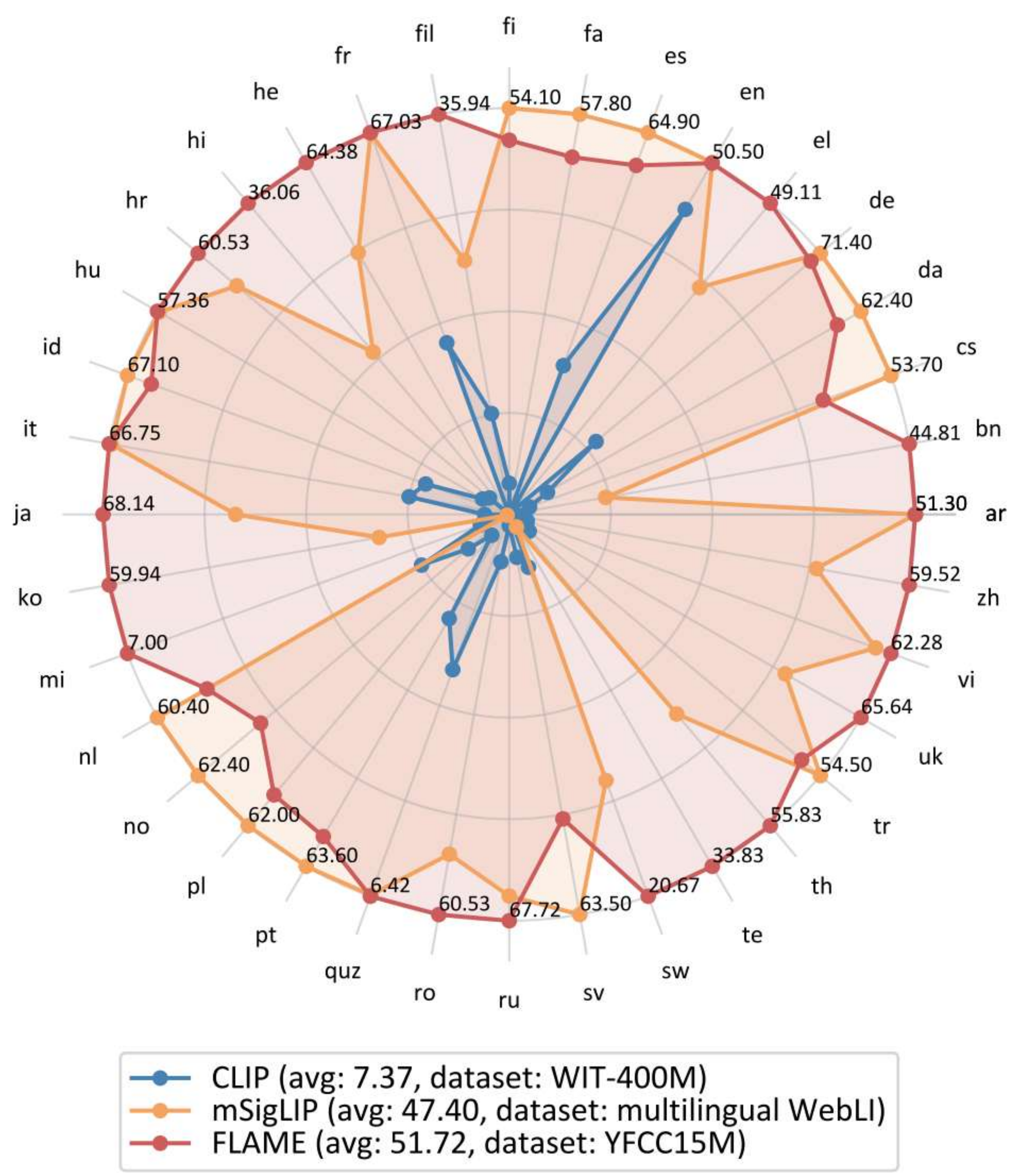
Facet-Decoupled Attention



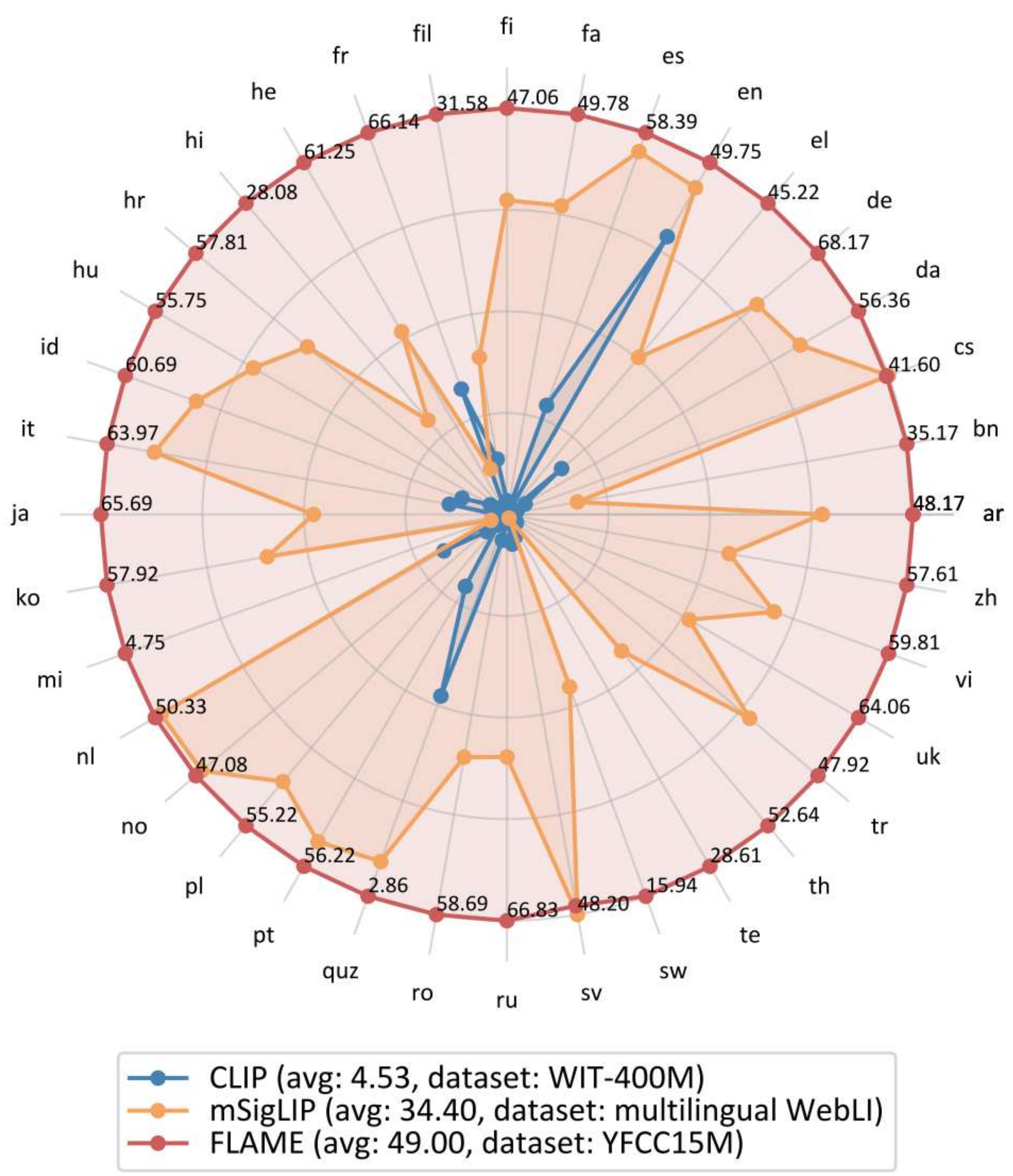
- **Practical Efficiency:** A specialized attention mask allowing parallel computation of all prompt embeddings within a single forward pass, utilizing a shared prefix structure. The frozen nature of the LLM enables to pre-compute the text embeddings offline.

Image-Text Retrieval

Multilingual Retrieval



(a) Text Retrieval



(b) Image Retrieval

Long-Context Retrieval

Method	Dataset	S4V-val I2T	S4V-val T2I	Urban-1k I2T	Urban-1k T2I	DCI I2T	DCI T2I	DOCCI-test I2T	DOCCI-test T2I
CLIP [41]	CC3M	21.4	20.2	10.7	9.6	8.3	7.5	8.6	7.0
Long-CLIP [62]	CC3M+S4V	51.3	46.1	15.5	18.5	13.8	14.2	14.7	12.6
FLAME	CC3M	**85.6**	**80.0**	**65.3**	**66.6**	**50.8**	**49.3**	**54.5**	**51.9**
CLIP [41]	YFCC15M	57.1	45.9	30.4	23.6	22.0	19.1	26.5	23.3
Long-CLIP [62]	YFCC15M+S4V	77.2	77.6	40.5	46.1	29.4	29.7	35.0	33.5
FLAME	YFCC15M	**94.1**	**93.2**	**84.0**	**87.9**	**66.1**	**68.1**	**75.8**	**76.2**
CLIP [41]	WIT-400M	78.2	79.6	67.5	53.3	45.4	43.0	60.7	57.0

Short-Context Retrieval

Dataset	Method	Text Retrieval MSCOCO R@1	R@5	R@10	Text Retrieval Flickr30k R@1	R@5	R@10	Image Retrieval MSCOCO R@1	R@5	R@10	Image Retrieval Flickr30k R@1	R@5	R@10
CC3M	CLIP [41]	8.7	23.9	33.7	7.1	19.7	28.6	17.4	37.9	50.1	13.9	30.8	40.5
	MLLM-A [36]	35.9	62.4	73.9	63.5	86.6	91.7	26.5	51.1	62.7	49.3	74.8	83.1
	DreamLIP [63]	<u>39.9</u>	<u>67.2</u>	<u>78.1</u>	<u>66.8</u>	**89.6**	**94.4**	**29.8**	**55.2**	**66.3**	<u>50.7</u>	<u>76.7</u>	<u>83.6</u>
	FLAME	**43.3**	**69.1**	**78.9**	**67.3**	<u>87.6</u>	<u>93.1</u>	<u>28.6</u>	<u>54.5</u>	<u>65.7</u>	**53.6**	**79.9**	**87.1**
YFCC15M	CLIP [41]	30.7	56.2	67.4	54.9	80.0	88.4	19.1	40.9	52.5	37.2	64.3	74.3
	SoftCLIP [16]	30.9	56.2	68.3	56.2	82.1	88.6	19.2	41.2	52.6	37.2	64.3	74.5
	DreamLIP [63]	<u>55.8</u>	<u>80.7</u>	<u>88.7</u>	<u>84.9</u>	**97.3**	**99.1**	<u>42.3</u>	<u>68.9</u>	<u>78.0</u>	<u>65.3</u>	<u>86.7</u>	<u>91.8</u>
	FLAME	**60.5**	**82.9**	**89.3**	**86.4**	**97.3**	<u>98.6</u>	**43.9**	**70.4**	**79.7**	**73.3**	**91.7**	**95.5**
WIT-400M	CLIP [41]	52.4	76.7	84.7	81.9	96.2	98.8	33.1	58.4	69.0	62.1	85.6	91.8

FLAME: Frozen Large Language Models Enable Data-Efficient Language-Image Pre-training

More Results

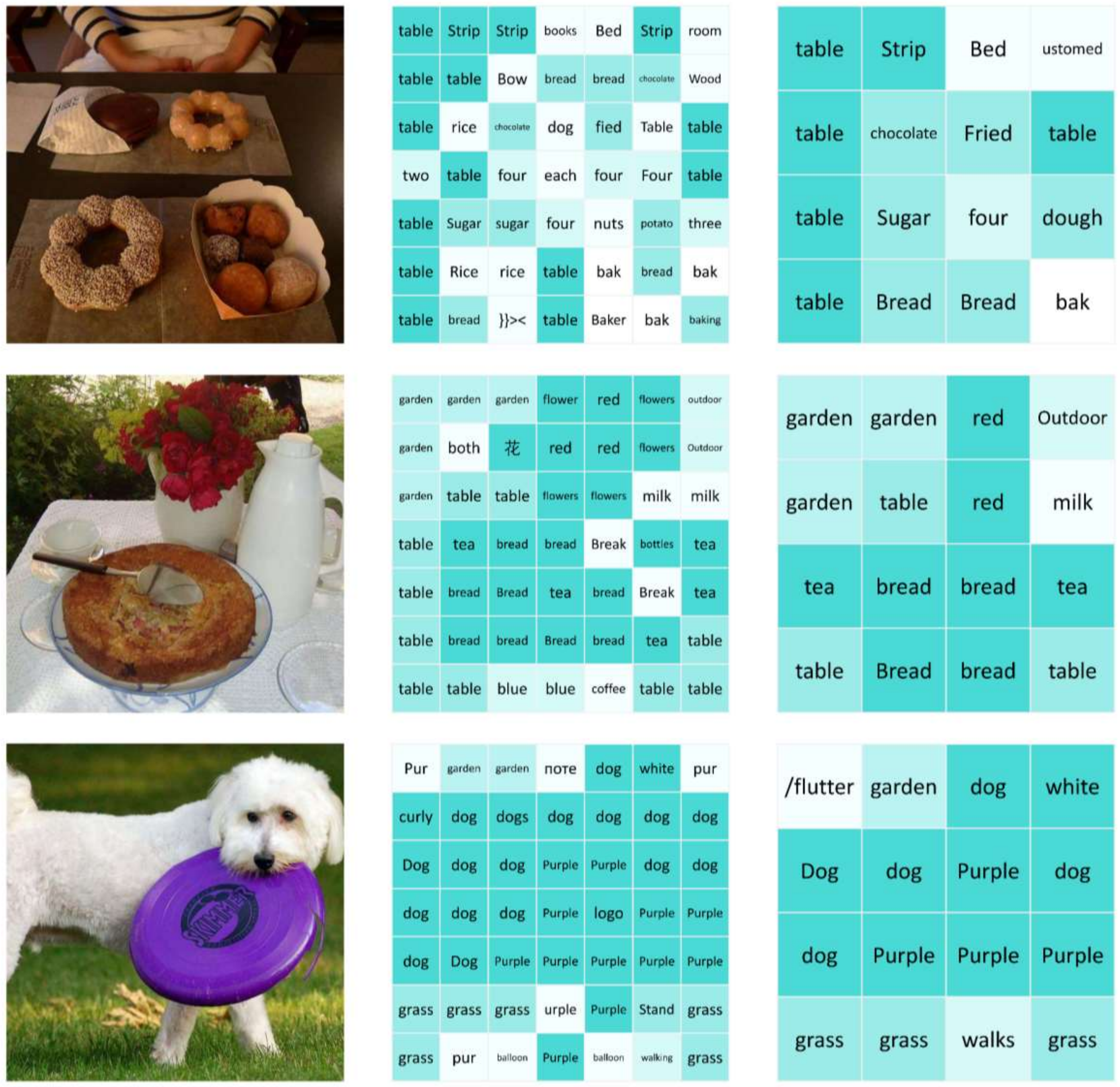
Zero-Shot Classification

Dataset	Method	Food-101	CIFAR-10	CIFAR-100	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	Average	ImageNet
CC3M	CLIP [41]	10.6	53.9	20.4	31.2	1.2	1.1	10.4	11.7	43.2	12.9	19.7	16.0
	LaCLIP [14]	14.2	57.1	27.5	35.1	1.6	1.6	16.6	15.6	52.7	14.7	23.7	21.5
	MLLM-A [36]	18.7	58.4	32.4	43.8	3.9	1.5	20.2	32.1	63.5	17.5	29.2	25.0
	DreamLIP [63]	19.4	**74.3**	**44.2**	45.9	2.8	1.0	17.0	27.1	63.1	14.7	31.0	31.1
	FLAME	**32.1**	73.6	42.0	**56.6**	**6.7**	**6.9**	**43.8**	**41.2**	**74.1**	**26.3**	**40.3**	**36.0**
YFCC15M	CLIP [41]	35.0	67.1	34.8	42.0	5.1	6.3	13.9	20.4	54.5	44.3	32.3	34.1
	DreamLIP [63]	44.2	**89.0**	**62.0**	57.1	9.2	6.4	30.5	32.6	**79.8**	40.2	45.1	48.2
	FLAME	**61.8**	86.1	56.7	**66.8**	**10.7**	**10.3**	**54.9**	**40.7**	78.9	**51.7**	**51.9**	**51.5**

Linear-Probe Classification

Dataset	Method	Food-101	CIFAR-10	CIFAR-100	Cars	Aircraft	DTD	Caltech-101	Average
CC3M	CLIP [41]	62.6	86.8	68.1	32.8	40.9	63.4	82.0	62.4
	LaCLIP [14]	63.8	87.7	69.5	32.4	**42.7**	64.0	83.3	63.3
	MLLM-A [36]	64.0	87.7	68.5	34.5	32.1	60.4	85.5	61.8
	DreamLIP [63]	71.2	**92.2**	**74.0**	31.5	26.7	**70.4**	88.5	64.9
	FLAME	**72.0**	90.3	72.9	**45.2**	38.6	69.7	**89.5**	**68.3**
YFCC15M	CLIP [41]	77.2	88.5	66.4	29.0	25.5	65.2	82.4	62.0
	HiCLIP [17]	81.0	89.1	70.4	36.4	32.3	68.7	86.4	66.3
	DreamLIP [63]	83.6	**96.5**	**82.3**	41.8	34.6	74.3	91.2	72.0
	FLAME	**85.9**	95.0	81.0	**54.3**	**39.3**	**76.8**	**92.5**	**75.0**

Semantic Interpretability



Input Image avg pool=2x2 avg pool=4x4, padding = 1

FLAME: Frozen Large Language Models Enable Data-Efficient Language-Image Pre-training

Key Contributions

- We **challenge the conventional wisdom** about frozen text encoders and demonstrate that frozen LLMs can effectively enhance language-image pre-training.
- We introduce a novel framework that leverages frozen LLMs for data-efficient language-image pretraining via **multifaceted prompt distillation**.
- We propose a **facet-decoupled attention** mechanism, complemented by an offline embedding strategy, to enhance computational efficiency.
- Extensive experiments validate that FLAME significantly outperforms existing methods in data-scarce scenarios, excelling in **long-context understanding** and **multilingual** tasks.

Thank you!

Project page:

https://github.com/MIV-XJTU/FLAME

For any further questions, please contact us:

caoanjia7@stu.xjtu.edu.cn weixing@mail.xjtu.edu.cn
mazhiheng@suat-sz.edu.cn