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

Supplementary Material

This document is structured as follows:

- In Section A, we provide details of the hyper-parameters.
- In Section B, we present the full list of multifaceted prompts used in our experiments.
- In Section C, we conduct additional experiments, including: 1) multilingual image-to-text retrieval, 2) vision-language compositionality, 3) few-shot image comprehension, 4) ablations on individual semantic levels, and 5) preliminary exploration of scalability.
- In Section D, we provide visualizations of text-to-image retrieval.

A. Hyper-Parameters

Table A1 and A2 provide the pre-training hyperparameters used for CC3M and YFCC15M, respectively. These hyper-parameters are set similarly to those used in previous methods [1, 12] for fair comparisons.

Config	Value
Batch size	4,096
Optimizer	AdamW [4]
Learning rate	5e-4
Weight decay	0.5
Adam β_1, β_2	0.9, 0.98
Adam ϵ	1e-8
Total epochs	32
Warm up iterations	2,000
Learning rate schedule	cosine decay

Table A1. **Hyper-parameters for CC3M.**

Config	Value
Batch size	8,192
Optimizer	AdamW [4]
Learning rate	5e-4
Weight decay	0.2
Adam β_1, β_2	0.9, 0.98
Adam ϵ	1e-8
Total epochs	32
Warm up iterations	2,000
Learning rate schedule	cosine decay

Table A2. **Hyper-parameters for YFCC15M.**

B. Full List of Multifaceted Prompts

We provide all the designed multifaceted prompts here. The distinct part in each prompt is marked. We also annotate the **default** prompts for long and short text input.

Entity Level:

- Detailed image description: “y_i”. After thinking step by step, the category of the main object in this image means in just one word:” **(default for long input)**
- Detailed image description: “y_i”. After thinking step by step, the prominent characteristic or pattern of the main object in this image means in just one word:” **(default for long input)**
- Detailed image description: “y_i”. After thinking step by step, the category of the minor object in this image means in just one word:” **(default for long input)**
- Detailed image description: “y_i”. After thinking step by step, the prominent characteristic or pattern of the minor object in this image means in just one word:” **(default for long input)**

Interaction Level:

- Detailed image description: “y_i”. After thinking step by step, the primary action or event taking place in this image means in just one word:” **(default for long input)**
- Detailed image description: “y_i”. After thinking step by step, the positioning layout or spatial relationship in this image means in just one word:”

Scene Level:

- Detailed image description: “y_i”. After thinking step by step, this image description means in just one word:” **(default for long and short inputs)**
- Detailed image description: “y_i”. After thinking step by step, the overall atmosphere or emotion of this image means in just one word:” **(default for long input)**
- Detailed image description: “y_i”. After thinking step by step, the dominant color or color combination of this image means in just one word:”

C. Additional Experiments

C.1. Multilingual Image-to-Text Retrieval

We provide image-to-text retrieval results on Crossmodal-3600 [6] in Figure C1.

C.2. Vision-Language Compositionality

To evaluate the vision-language compositional understanding (i.e. whether the model understands the fine-grained atomic

Method	Dataset	Winoground			SugarCrepe-Add		SugarCrepe-Replace			SugarCrepe-Swap		
		Image	Text	Both	Attribute	Object	Attribute	Object	Relation	Attribute	Object	Avg.
CLIP [5]	WIT-400M	10.8	25.0	7.5	66.8	78.5	81.1	93.4	66.9	64.6	60.0	55.5
DreamLIP [12]	YFCC15M	14.7	26.2	9.7	78.3	80.3	81.3	91.0	72.9	**77.5**	**66.9**	59.9
FLAME	YFCC15M	**18.3**	**34.5**	**13.2**	**82.4**	**87.8**	**85.8**	**94.1**	**79.9**	67.6	66.1	**63.0**

Table C3. **Vision-language compositionality.** FLAME demonstrates better fine-grained scene understanding capability.

Method	N-Way K-shot	2-1	2-3	2-1	2-1	Avg.	5-1	5-3	5-1	5-1	Avg.
	# Repetitions	0	0	1	3		0	0	1	3	
Frozen [8]	-	33.7	66.0	63.0	65.0	56.9	14.5	34.7	33.8	33.3	29.1
LQAE [3]	GPT-3.5	35.2	68.2	68.5	68.7	60.2	15.7	35.9	31.9	36.4	30.0
V2L-Tokenizer [13]	Llama-2-7B	76.3	91.2	84.0	84.4	84.0	44.8	**91.8**	**73.9**	**82.2**	**73.2**
SPAE [11]	PaLM-2-340B	**84.8**	**92.5**	84.8	85.2	**86.8**	**65.1**	73.7	66.4	67.0	68.1
FLAME	Mistral-7B	83.3	91.7	**85.7**	**86.3**	**86.8**	55.7	82.1	65.7	70.1	68.4

Table C4. **Few-shot image comprehension on MiniImageNet.** FLAME's performance is either comparable to or surpasses that of SPAE.

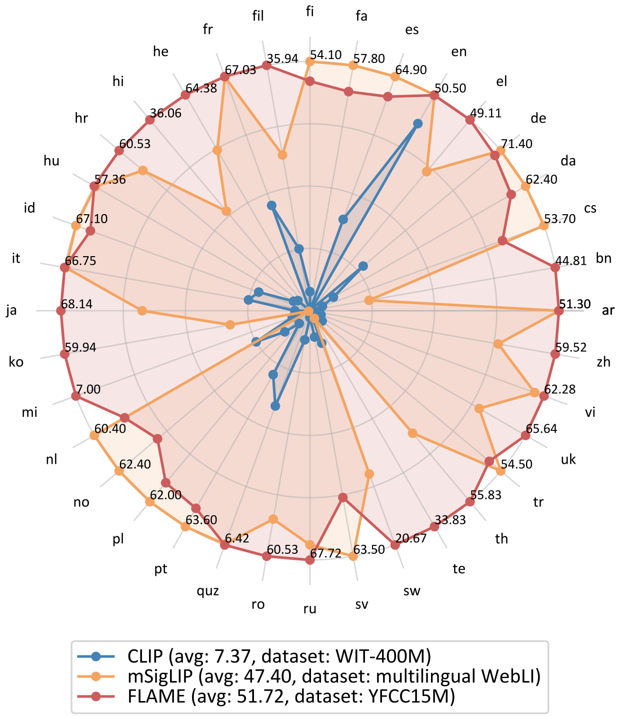


Figure C1. **Multilingual zero-shot image-to-text retrieval recall@1 results on Crossmodal-3600.**

concepts that compose the scene), we conduct experiments on the Winoground [7] and SugarCrepe [2] benchmarks, with results presented in Table C3. As shown, FLAME trained on YFCC15M significantly outperforms WIT-400M-trained CLIP [5] across all tasks, especially in relation understanding. Furthermore, when compared to the recent work [12] that utilizes multiple short captions, FLAME retains an advantage on 8 out of 10 tasks, achieving an average improvement of 3.1%. These results reveal that our multifaceted training promotes fine-grained semantic learning.

C.3. Few-Shot Image Comprehension

Following SPAE [11] and V2L-Tokenizer [13], we conduct few-shot image comprehension experiments on both 2-way and 5-way MiniImageNet benchmarks. Our evaluation methodology involves converting image patches into words and then using the original LLM for in-context reasoning, without additional fine-tuning. Specifically, given a few-shot example image set $\{x_i\}_{i=1}^S$, we perform the vocabulary mapping process to obtain $\{C_i\}_{i=1}^S$, as mentioned in Section **??**. By pairing each C_i with its corresponding text answer A_i, we construct the in-context example $E = \{\langle C_i, A_i \rangle\}_{i=1}^S$. We input this example and the new context $\tilde{C}$ to the LLM, which then outputs the answer $\tilde{A}$ for verification of correctness. This reasoning process can be formulated as $\tilde{A} = \text{LLM}(E, \tilde{C})$. As shown in Table C4, FLAME achieves average accuracies of 86.8% and 68.4% on 2-way and 5-way scenarios, respectively. These results are comparable to or surpass SPAE [11]. FLAME achieves these results using a smaller 7B backbone than SPAE's 340B model.

C.4. Ablations on Individual Semantic Levels

To further clarify the reasonability of our semantic decomposition, we conduct ablations by orderly increasing individual semantic levels. The results are presented in Table C7. Evidently, a systematic increase in semantic levels consistently yields positive outcomes. This also indicates that our rational prompt design is generalizable and does not lead to detrimental caption-prompt misalignment.

Level	Long Avg.	
	I2T	T2I
+Scene	50.5	47.4
+Interaction	52.9	50.3
+Entity	64.1	62.0

Table C7. **Semantic decomposition.**

Backbone	Method	Food-101	CIFAR-10	CIFAR-100	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	Average	ImageNet
ViT-B/16	CLIP [5]	35.0	67.1	34.8	42.0	5.1	6.3	13.9	20.4	54.5	44.3	32.3	34.1
	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**
ViT-L/14	CLIP [5]	42.2	66.8	33.1	45.3	2.7	2.3	19.7	26.4	65.1	53.1	35.7	36.3
	FLAME	**68.3**	**87.5**	**58.4**	**68.3**	**14.6**	**11.3**	**56.3**	**43.8**	**80.3**	**56.3**	**54.5**	**54.8**

Table C5. **Zero-shot classification on YFCC15M with ViT-L/14.** FLAME consistently demonstrates substantial advantages over CLIP.

Backbone	Method	Food-101	CIFAR-10	CIFAR-100	Cars	Aircraft	DTD	Caltech-101	Average
ViT-B/16	CLIP [5]	77.2	88.5	66.4	29.0	25.5	65.2	82.4	62.0
	FLAME	**85.9**	**95.0**	**81.0**	**54.3**	**39.3**	**76.8**	**92.5**	**75.0**
ViT-L/14	CLIP [5]	74.4	88.9	69.7	27.3	26.1	63.7	86.4	62.3
	FLAME	**88.0**	**95.9**	**81.7**	**64.3**	**46.3**	**78.4**	**93.8**	**78.3**

Table C6. **Linear-probe classification on YFCC15M with ViT-L/14.** FLAME benefits from the increased scale of the visual backbone.

C.5. Preliminary Exploration of Scalability

To explore the scalability of FLAME, we increase the size of the visual encoder from the default ViT-B/16 to ViT-L/14 and train it on YFCC15M.

Table C5 presents the zero-shot classification results, where FLAME continues to exhibit a substantial advantage over CLIP when using ViT-L/14, as evidenced by an improvement in ImageNet top-1 accuracy of 18.5% and an increase in downstream average accuracy of 18.8%. We also perform linear-probe classification experiments, with results shown in Table C6. In multilingual scenarios, the use of this larger backbone also proves beneficial, yielding a 2.8% improvement in average accuracy on the multilingual ImageNet1k classification benchmark. Detailed results and comparisons are provided in Table C8.

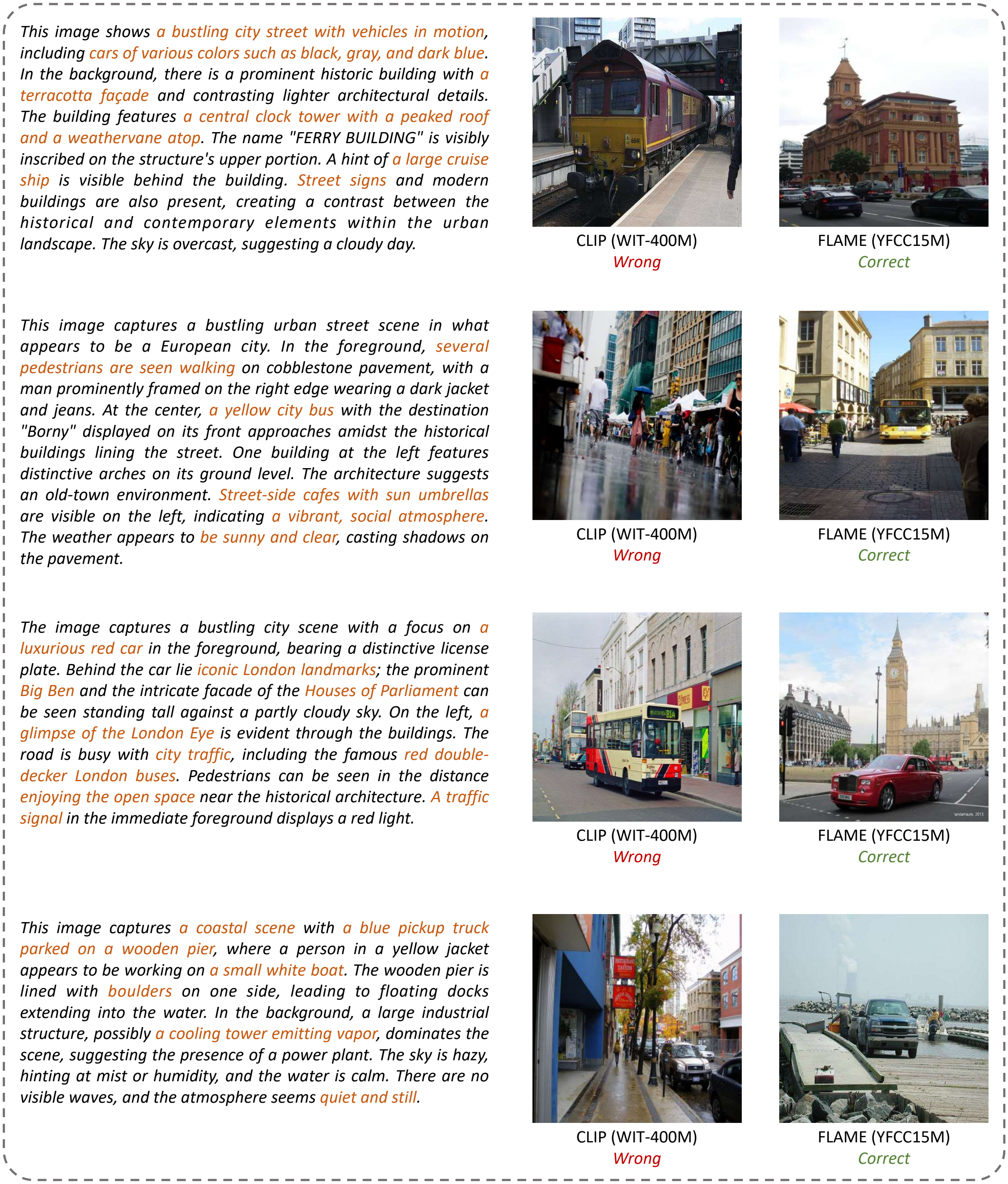
Method	Training	Evaluation					
		ar	en	it	jp	zh	avg
CLIP [5]	en	0.6	**75.5**	27.6	4.5	1.9	22.0
CN-CLIP [10]	en+zh	0.1	32.5	8.0	16.3	**53.4**	22.1
NLLB-CLIP [9]	en+multi	25.8	36.7	27.4	23.9	24.3	27.6
FLAME	en	**34.9**	54.8	**44.9**	**39.6**	42.6	**43.4**

Table C8. **Multilingual ImageNet1k classification with the "large" model variant.** CLIP is trained on WIT-400M. CN-CLIP is pre-trained on WIT-400M and fine-tuned on 200M Chinese data. NLLB-CLIP is pre-trained on LAION-2B and fine-tuned on LAION-COCO-NLLB with 200 languages. FLAME is trained on YFCC15M.

These notable performance enhancements illustrate the scalability of FLAME, paving the way for future large-scale language-image pre-training.

D. Visualizations

Figure D2 presents some visualizations of text-to-image retrieval on Urban-1k, comparing the results of FLAME with those of CLIP.



This image shows a bustling city street with vehicles in motion, including cars of various colors such as black, gray, and dark blue. In the background, there is a prominent historic building with a terracotta façade and contrasting lighter architectural details. The building features a central clock tower with a peaked roof and a weathervane atop. The name "FERRY BUILDING" is visibly inscribed on the structure's upper portion. A hint of a large cruise ship is visible behind the building. Street signs and modern buildings are also present, creating a contrast between the historical and contemporary elements within the urban landscape. The sky is overcast, suggesting a cloudy day.

This image captures a bustling urban street scene in what appears to be a European city. In the foreground, several pedestrians are seen walking on cobblestone pavement, with a man prominently framed on the right edge wearing a dark jacket and jeans. At the center, a yellow city bus with the destination "Borny" displayed on its front approaches amidst the historical buildings lining the street. One building at the left features distinctive arches on its ground level. The architecture suggests an old-town environment. Street-side cafes with sun umbrellas are visible on the left, indicating a vibrant, social atmosphere. The weather appears to be sunny and clear, casting shadows on the pavement.

The image captures a bustling city scene with a focus on a luxurious red car in the foreground, bearing a distinctive license plate. Behind the car lie iconic London landmarks; the prominent Big Ben and the intricate facade of the Houses of Parliament can be seen standing tall against a partly cloudy sky. On the left, a glimpse of the London Eye is evident through the buildings. The road is busy with city traffic, including the famous red double-decker London buses. Pedestrians can be seen in the distance enjoying the open space near the historical architecture. A traffic signal in the immediate foreground displays a red light.

This image captures a coastal scene with a blue pickup truck parked on a wooden pier, where a person in a yellow jacket appears to be working on a small white boat. The wooden pier is lined with boulders on one side, leading to floating docks extending into the water. In the background, a large industrial structure, possibly a cooling tower emitting vapor, dominates the scene, suggesting the presence of a power plant. The sky is hazy, hinting at mist or humidity, and the water is calm. There are no visible waves, and the atmosphere seems quiet and still.

Figure D2. **Visualizations of text-to-image retrieval on Urban-1k.**

References

[1] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. In *NeurIPS*, 2023. 1

[2] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *NeurIPS*, 2023. 2

[3] Hao Liu, Wilson Yan, and Pieter Abbeel. Language quantized autoencoders: Towards unsupervised text-image alignment. In *NeurIPS*, 2023. 2

[4] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3

[6] Ashish V Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *EMNLP*, 2022. 1

[7] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, 2022. 2

[8] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *NeurIPS*, 2021. 2

[9] Alexander Visheratin. Nllb-clip–train performant multilingual image retrieval model on a budget. *arXiv preprint arXiv:2309.01859*, 2023. 3

[10] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*, 2022. 3

[11] Lijun Yu, Yong Cheng, Zhiruo Wang, Vivek Kumar, Wolfgang Macherey, Yanping Huang, David Ross, Irfan Essa, Yonatan Bisk, Ming-Hsuan Yang, et al. Spae: Semantic pyramid autoencoder for multimodal generation with frozen llms. In *NeurIPS*, 2023. 2

[12] Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. In *ECCV*, 2024. 1, 2

[13] Lei Zhu, Fangyun Wei, and Yanye Lu. Beyond text: Frozen large language models in visual signal comprehension. In *CVPR*, 2024. 2