ePALM: Efficient Perceptual Augmentation of Language Models

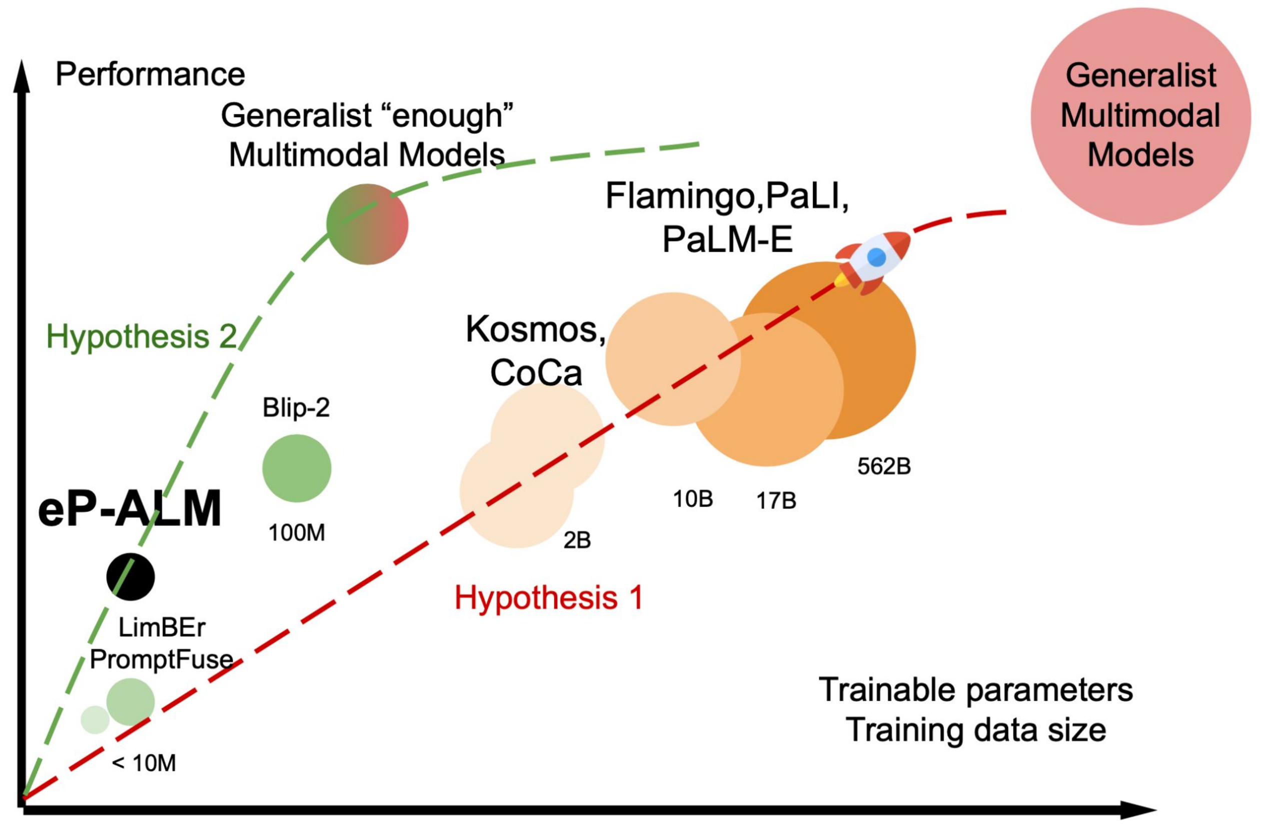
Mustafa Shukor, Corentin Dancette, Matthieu Cord

Sorbonne University, ISIR, Paris, France

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

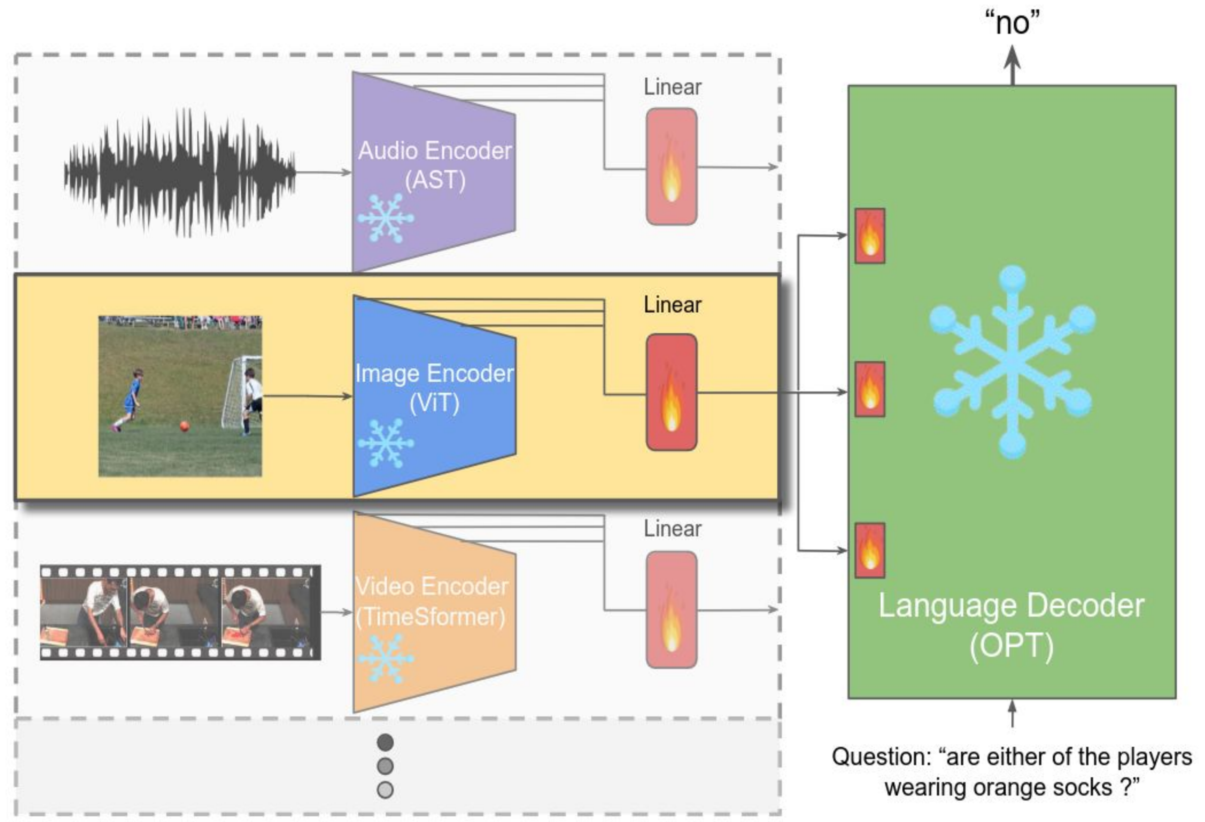
How to build generalist multimodal models?

- **Hypothesis 1**: scaling parameters, data and compute
- **Hypothesis 2:** efficient adaptation of large unimodal-pretrained models



Summary of the work: efficient adaptation (linear projection) of frozen, pretrained, unimodal models (OPT and ViT) to solve multimodal tasks (VQA, Captioning) across image, **video** and **audio** modalities

- Trainable parameters **< 0.06%**
- No multimodal pretraining

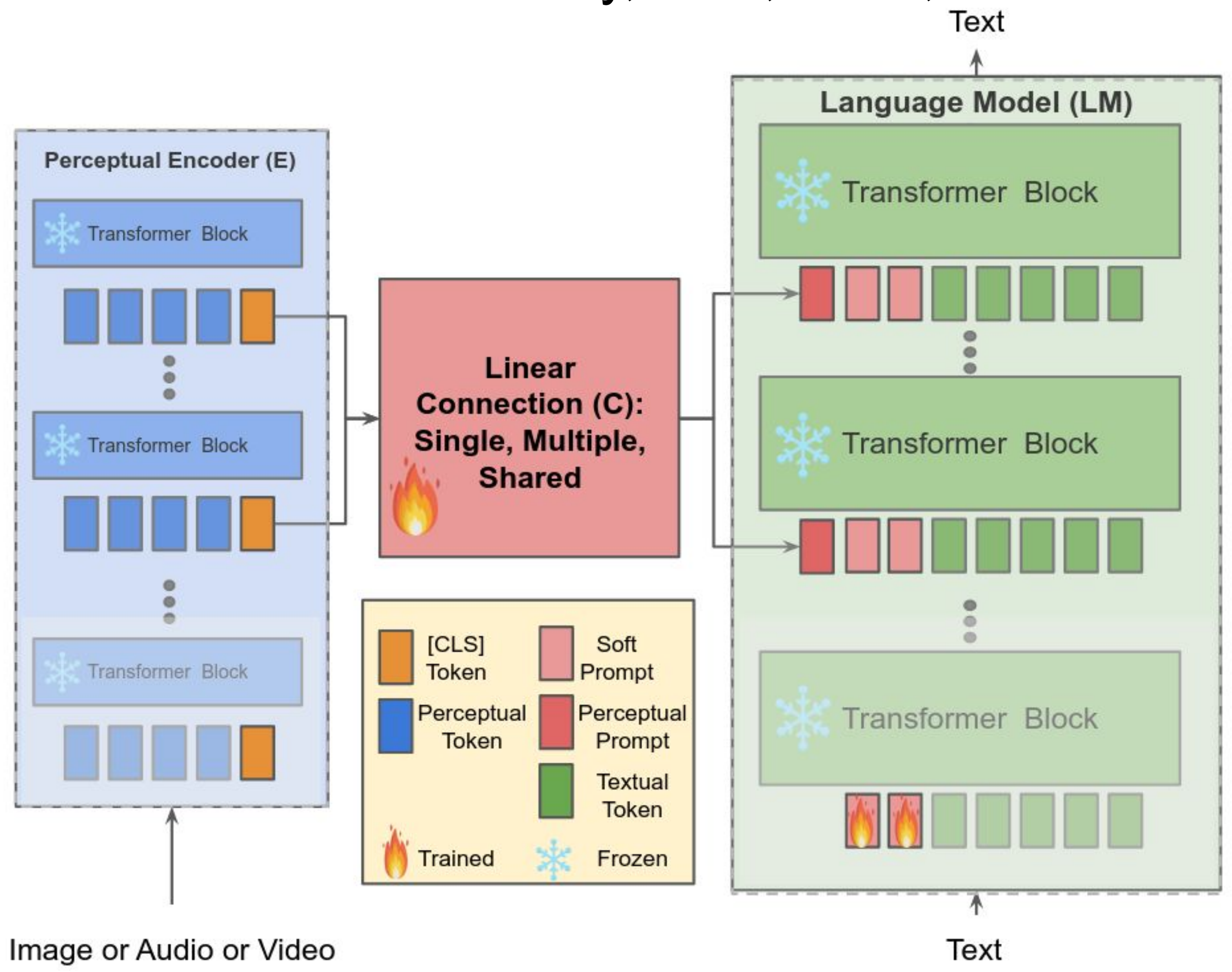


Direct Finetuning (ePALM)

- Efficient to train
- Generally better performance
- Easy to adapt to new tasks/datasets
- Efficient to adapt to new LLMs
- Task-specific finetuning

Pretrain-Zeroshot (e.g. LimBEr, Flamingo)

- Costly pretraining
- Limited performance, saturation with FS ICL
- Finetuning is needed for "new" datasets/tasks
- Pretraining is needed for a new LLM
- One training for many tasks



Proposed approach:

- **Model**:
  - *Language Model*: OPT (2.7B)
  - *Unimodal Encoders*: ViT-Base (ImageNet), TimeSformer-B (kinetics), AST-B (audioset)
  - *Adaptation parameters*:
    - ***Cross-Modal Connection***: linear projection of the encoders' last layers visual/audio **[CLS]** tokens injected in the OPT's last layers
    - ***Soft Prompt***: 10 learnable tokens prepended to the text input
- **Data**: target dataset (e.g. COCO, VQAv2, AudioCaps, MSR-VTT)
- **Training**: training only adaptation parameters for few epochs

Image-Text tasks:

Method	VQA v2		GQA		COCO	
	Val	Test	Val	Test	B@4	CIDEr
PromptFuse [57]	34.1	–	–	–	–	–
B_{LimBEr}	34.1	33.5	30.81	29.4	–	–
$B_{PromptFuse}$	40.4	39.5	33.74	31.51	15.05	48.26
B_{MAGMA}	32.2	31.8	30.98	28.93	–	–
ePALM$_{pt}$	48.8	47.8	43.8	40.3	27.52	91.92
ePALM	**50.7**	**50.2**	**45.0**	**40.4**	**29.47**	**97.22**
ePALM$_{pt}$-L	54.90	54.90	47.19	43.0	33.35	113.0

Main results

Method	Train. data % (# of shots)	VQA v2
PromptFuse* [56]	0.12% (512)	29.40
ePALM	0.12% (512)	35.54
ePALM	1% (4.4K)	42.28
B_{LimBEr}	1% (4.4K)	28.9
$B_{PromptFuse}$	1% (4.4K)	31.9
B_{MAGMA}	1% (4.4K)	34.5

Data Efficiency

Q: what color is the bear?
A: black and blue (blue)

- Consistently better than other baselines that prepend visual tokens to the input layer and use adapters or prompt tuning
- More data-efficient

Video-Text tasks:

- Better zero-shot generalzation on VideoQA (and Image VQA)

Method	Training data	Train. Param. (%)	OE Gen	MSRVTT-QA	MSVD-QA
JustAsk [95]	ActivityNet-QA	89.6%	✗	2.7	-
JustAsk [95]	HowToVQA69M	89.6%	✗	2.9	7.5
LAVENDER [53]	WebVid2.5M+CC3M	100%	✗	4.5	11.6
MERLOT Reserve [100]	YT-Temporal-1B	100%	✗	5.8	-
FrozenBiLM † [96]	400M-CLIP + VQA v2	2.9%	✗	6.9	12.6
Flamingo 3B [2]	M3W+ALIGN+VTP	40%	✓	11.0	**27.5**
ePALM	VQA v2	0.9%	✓	13.17	24.82
ePALM †	VQA v2	0.9%	✓	**14.54**	27.09

Image/Video/Audio-text tasks:

- Comparison with SoTA that trained with large number of parameters and most often with large-scale pretraining

Dataset (Metric)	SoTA (ZS)	ePALM (FT)	SoTA (FT)
AudioCaps (CIDEr)	–	<u>63.6</u>	**66.7** (Liu et al. [59])
MSRVTT-QA (Acc)	17.4 (Flamingo80B [2])	<u>36.7</u>	**44.1** (OmniVL [88])
MSR-VTT (CIDEr)	–	<u>50.7</u>	**60** (MV-GPT [73])
COCO (CIDEr)	84.3 (Flamingo80B [2])	<u>107.0</u>	**145.3** (OFA [89])
VQAv2 (Acc)	<u>56.3</u> (Flamingo80B [2])	53.3	**84.3** (PaLI [14])
GQA (Acc)	29.3 (FewVLM [43])	<u>42.7</u>	**60.8** (VL-T5 [17])

Code

https://github.com/mshukor/ePALM