

# Frozen Transformers in Language Models Are Effective Visual Encoder Layers

ICLR 2024 Spotlight

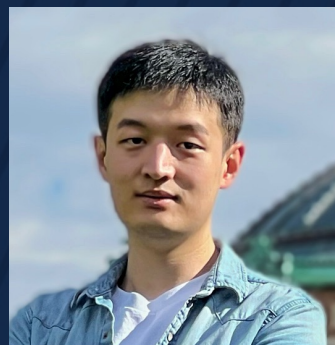
Ziqi Pang



Ziyang Xie\*



Yunze Man\*



Yu-Xiong Wang



UNIVERSITY OF  
**ILLINOIS**  
URBANA-CHAMPAIGN



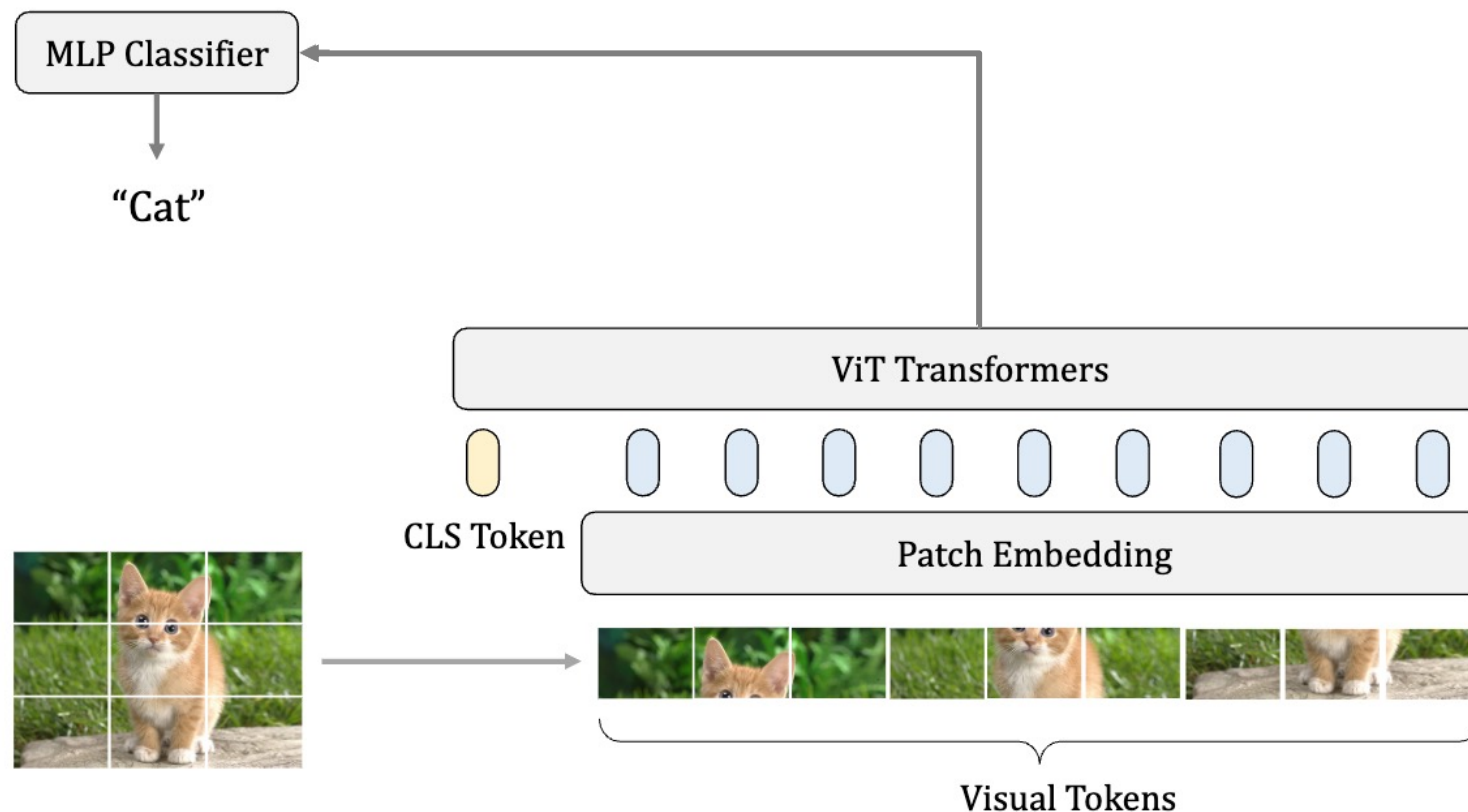
# Our Key Observation

- LLMs, though **trained solely with language**, are surprisingly strong **encoders for purely visual tasks** even in the **absence of language**



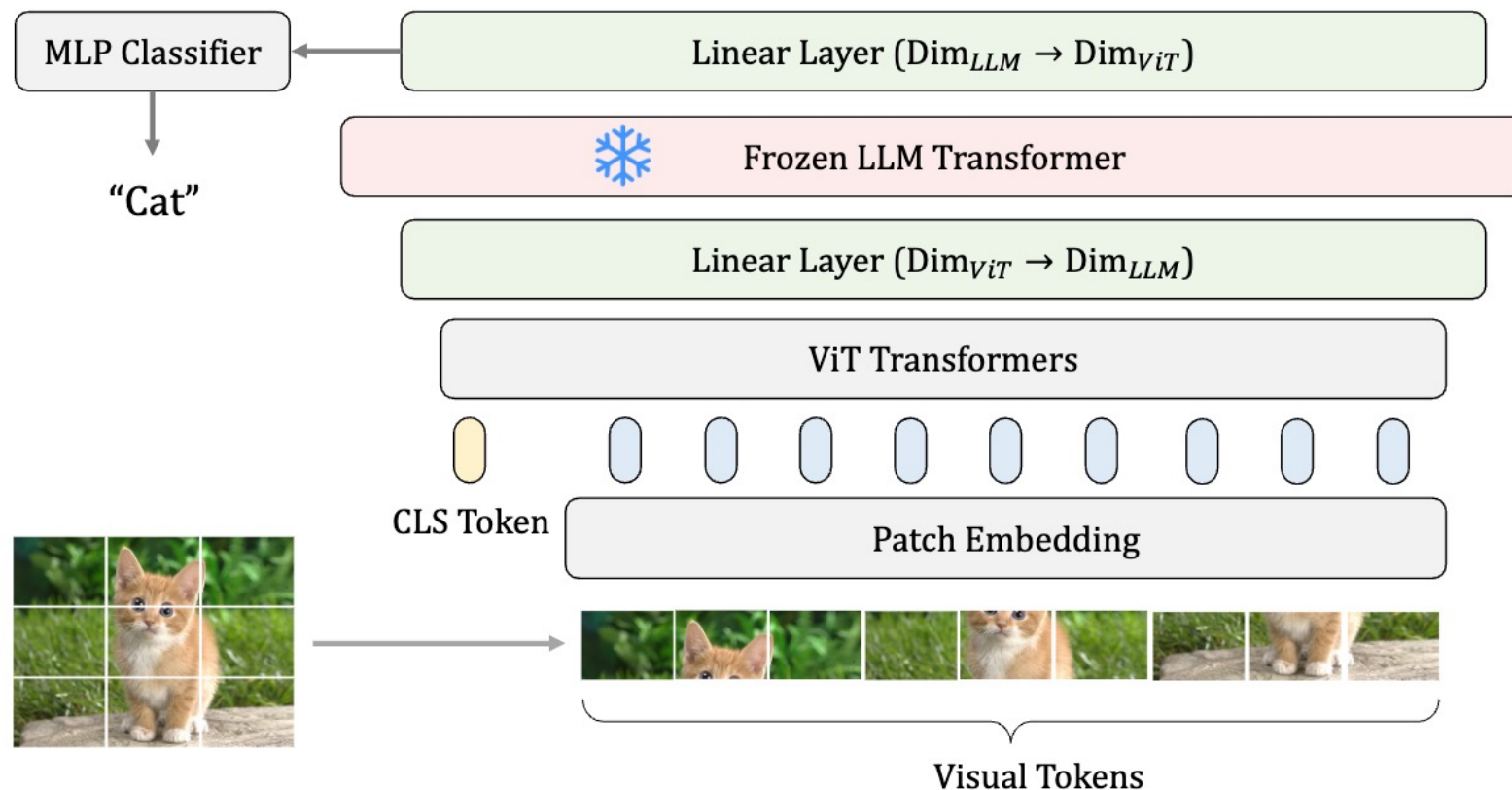
# Our Simple Framework

- LLMs, though **trained solely with language**, are surprisingly strong **encoders for purely visual tasks** even in the **absence of language**
- This is achieved via a simple framework  
→ Appending a **frozen LLM transformer** after the visual encoder



# Our Simple Framework

- LLMs, though **trained solely with language**, are surprisingly strong **encoders for purely visual tasks** even in the **absence of language**
- This is achieved via a simple framework  
→ Appending a **frozen LLM transformer** after the **visual encoder**



# Applicability to Diverse Tasks

- By using the **last layer from LLaMA**, we **significantly improve** various tasks

ImageNet  
Image Classification  
(**2D**)

Model	Acc
ViT-S	80.1%
+ LLaMA	<b>80.7%</b>

SSv2  
Action Recognition  
(**Video**)

Model	Acc
ViT-B	64.9%
+ LLaMA	<b>66.0%</b>

VQAv2,  
2D VQA  
(**2D Vision Language**)

Model	Score
METER	0.696
+ LLaMA	<b>0.702</b>

ScanObjectNN,  
Point Cloud Classification  
(**3D**)

Model	Acc
PointBERT	83.1%
+ LLaMA	<b>83.8%</b>

Argoverse  
Motion Forecasting  
(**Non-semantic**)

Model	MR ↓
VectorNet	13.2%
+ LLaMA	<b>12.7%</b>

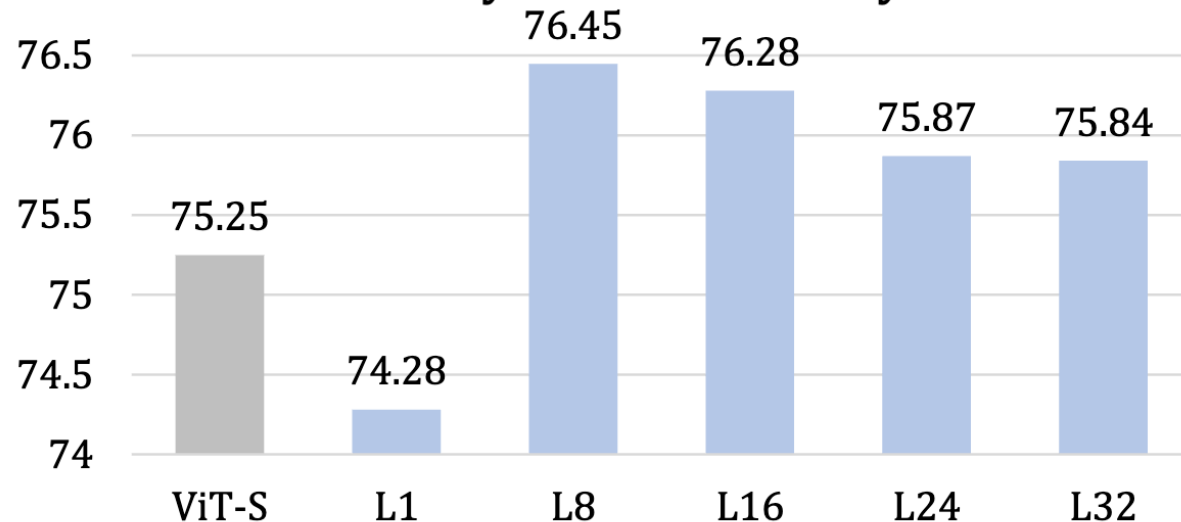
SQA3D,  
3D Grounding  
(**3D Vision Language**)

Model	EM
SQA3D	47.2%
+ LLaMA	<b>48.1%</b>

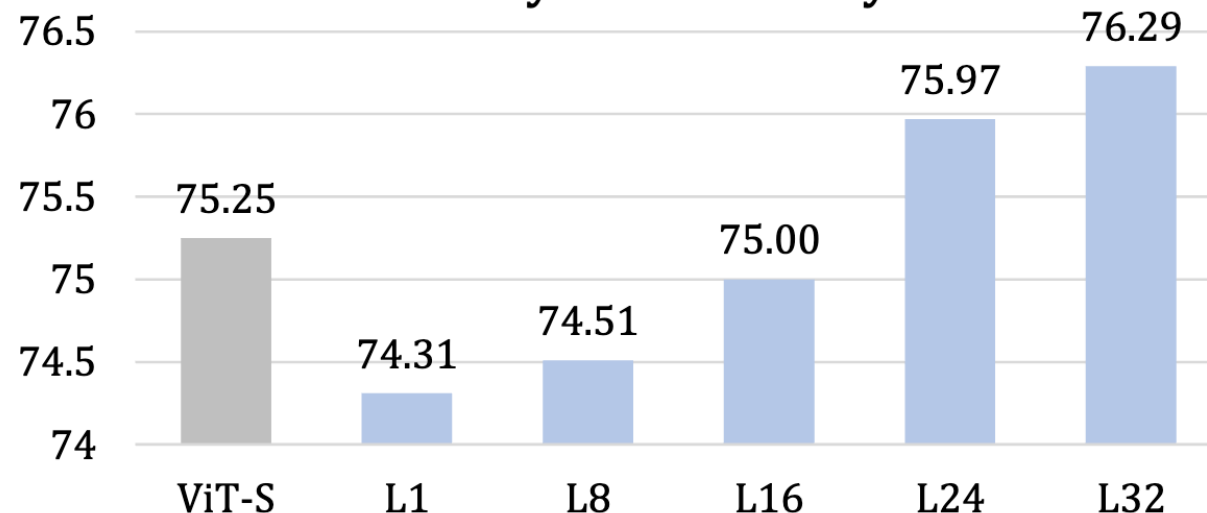
# Ablation and Analysis

- The improvement does not solely come from capacity, but **the LLM transformer**
- The improvement generalizes to **diverse choices of LLM transformers**

Accuracy with LLaMA Layers



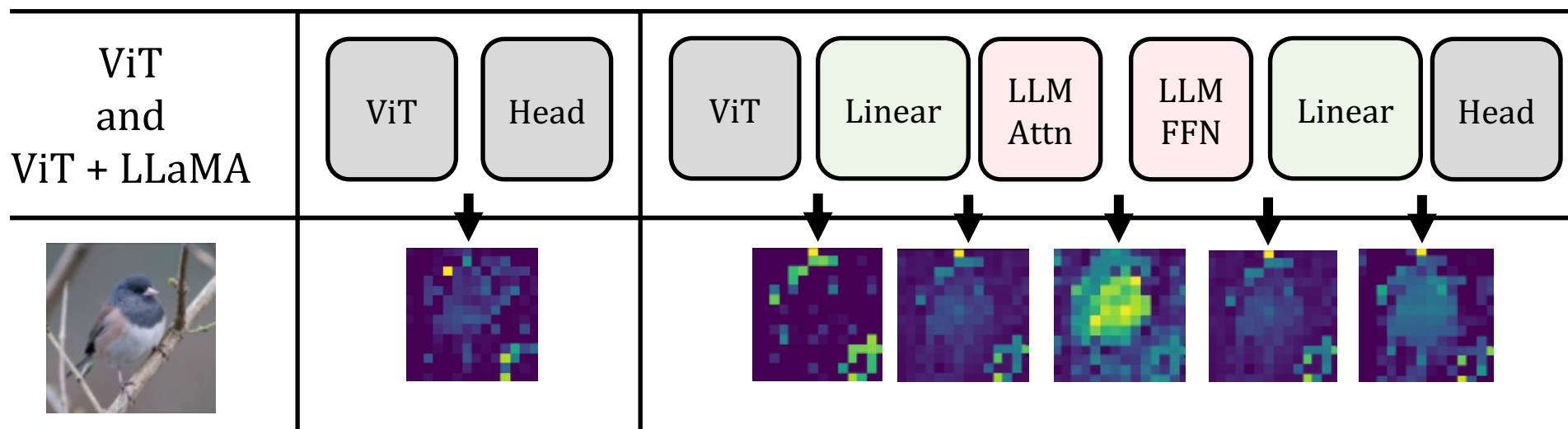
Accuracy with OPT Layers





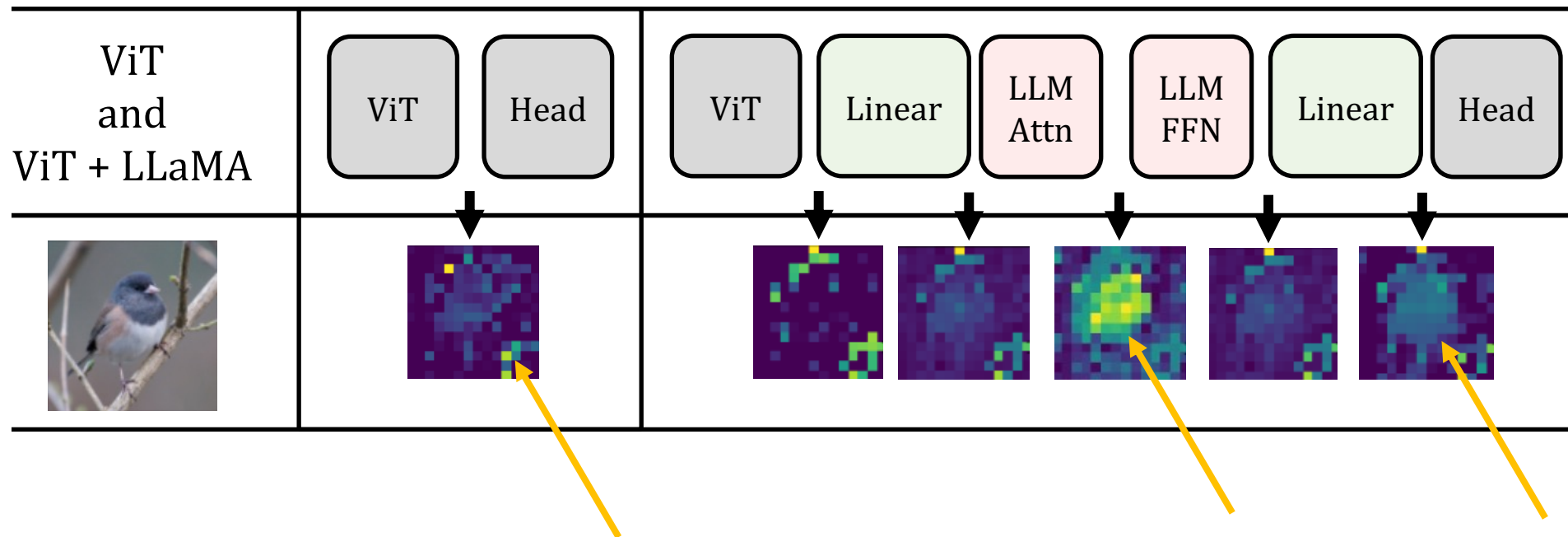
# Explanation: Information Filtering Hypothesis

- Our method benefits from frozen LLM transformers  
*distinguishing and amplifying informative tokens*
- Emergent ability to **concentrate on foregrounds**. Uncommon for supervised ViTs



# Explanation: Information Filtering Hypothesis


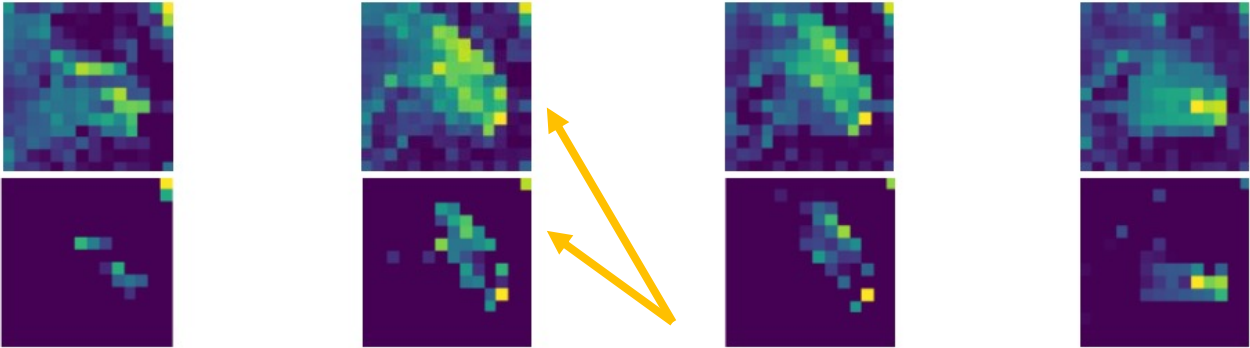
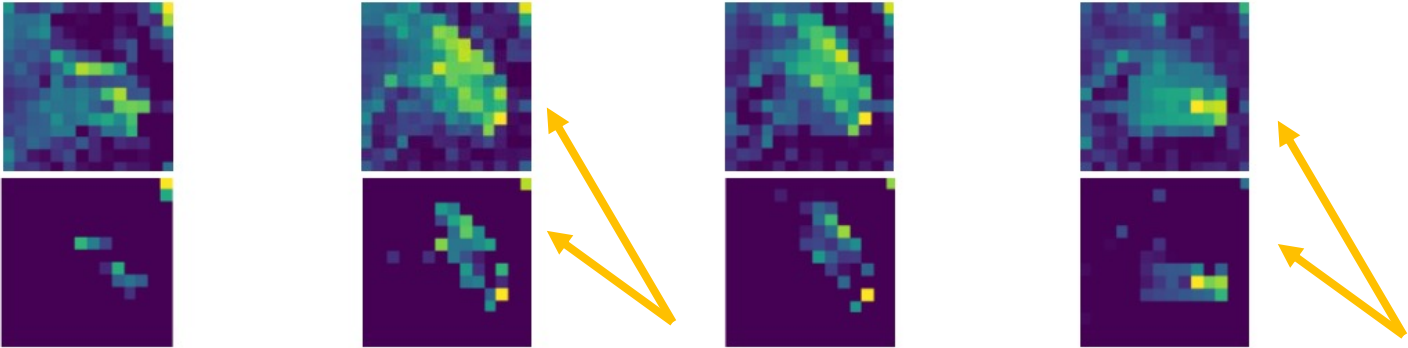
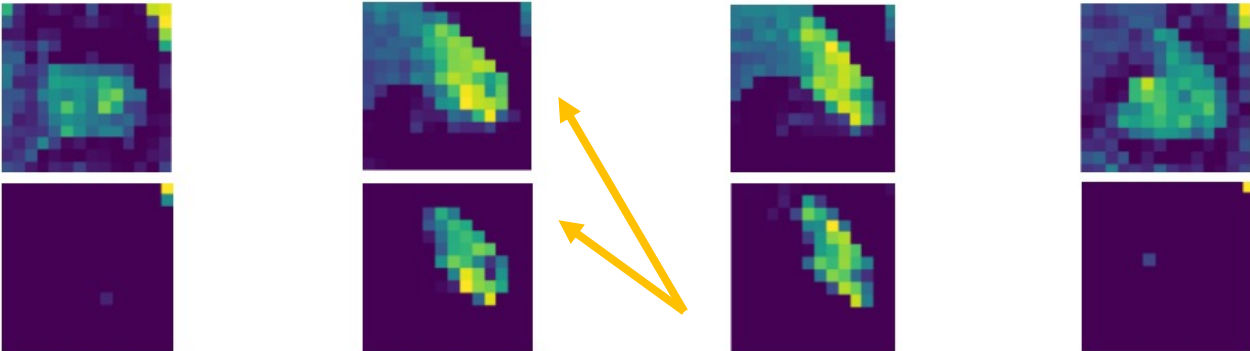
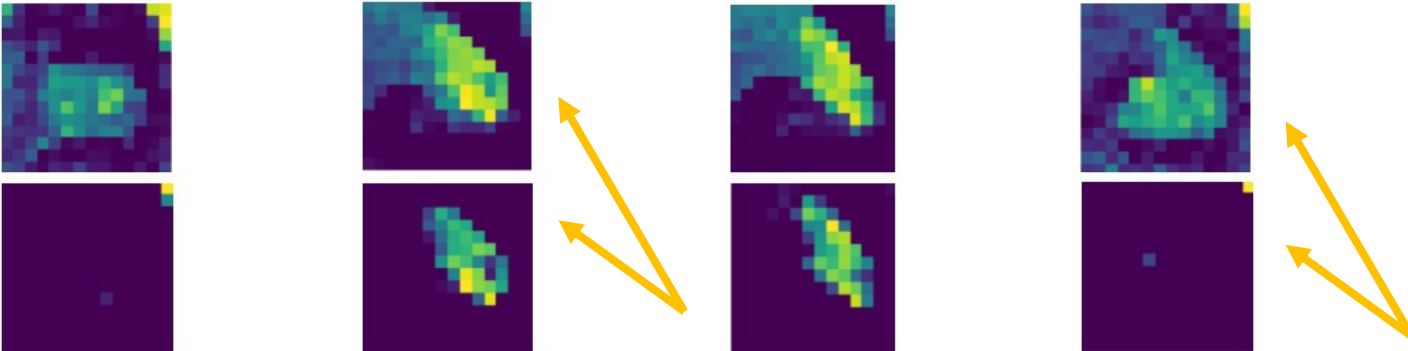
- Our method benefits from frozen LLM transformers  
*distinguishing and amplifying informative tokens*
- Emergent ability to **concentrate on foregrounds**. Uncommon for supervised ViTs





# Our Hypothesis Applies to Various Tasks

- Action recognition: selection of key frames

Model	Threshold								
ViT-S	Low								
	High								
ViT-S + LLaMA	Low								
	High								



# Takeaways

- We discover that using a frozen transformer block from pre-trained LLMs as a visual encoder layer facilitates a diverse range of tasks
- We demonstrate that the benefits of frozen LLM transformers generalize to various LLMs and transformer blocks
- We propose the information filtering hypothesis: the incorporated LLM blocks distinguish the informative tokens and amplify their effect

Project Website  
&  
Thank You!

