UiT The Arctic University of Norway

NORA summer school on multi-modal learning

Multi-modal Fusion

Rwiddhi Chakraborty

UiT Machine Learning Group and Visual Intelligence

Schedule Today

- 10 - 11: Welcome, practical information, motivation
- 11 - 12: The fundamentals of Multi-modal learning
- 12 – 13: Lunch
- 13 – 14: Multi-modal Classification
- 14 – 15: Multi-modal Fusion
- 15 – 17: Practical exercise (Supervised Multi-modal)

In this talk

Why fusion?

Fusion Techniques

Interesting Issues

In this talk

Why fusion?

Fusion Techniques

Interesting Issues

Motivation

Why fusion?

Fusion is the essence of Multi-modality

Motivation

Why fusion?

Fusion is the essence of Multi-modality

Fusion captures *information redundancy*

Motivation

Why fusion?

Fusion is the essence of Multi-modality

Fusion captures *information redundancy*

Fusion captures *semantic overlap*

Motivation

Why fusion?

Fei Zhao, Chengcui Zhang, and Baocheng Geng. 2024. Deep Multi-modal Data Fusion. ACM Comput. Surv. 56, 9, Article 216 (September 2024), 36 pages. https://doi.org/10.1145/3649447

Motivation

Why fusion?

Because humans practice fusion all the time

Fei Zhao, Chengcui Zhang, and Baocheng Geng. 2024. Deep Multi-modal Data Fusion. ACM Comput. Surv. 56, 9, Article 216 (September 2024), 36 pages. https://doi.org/10.1145/3649447

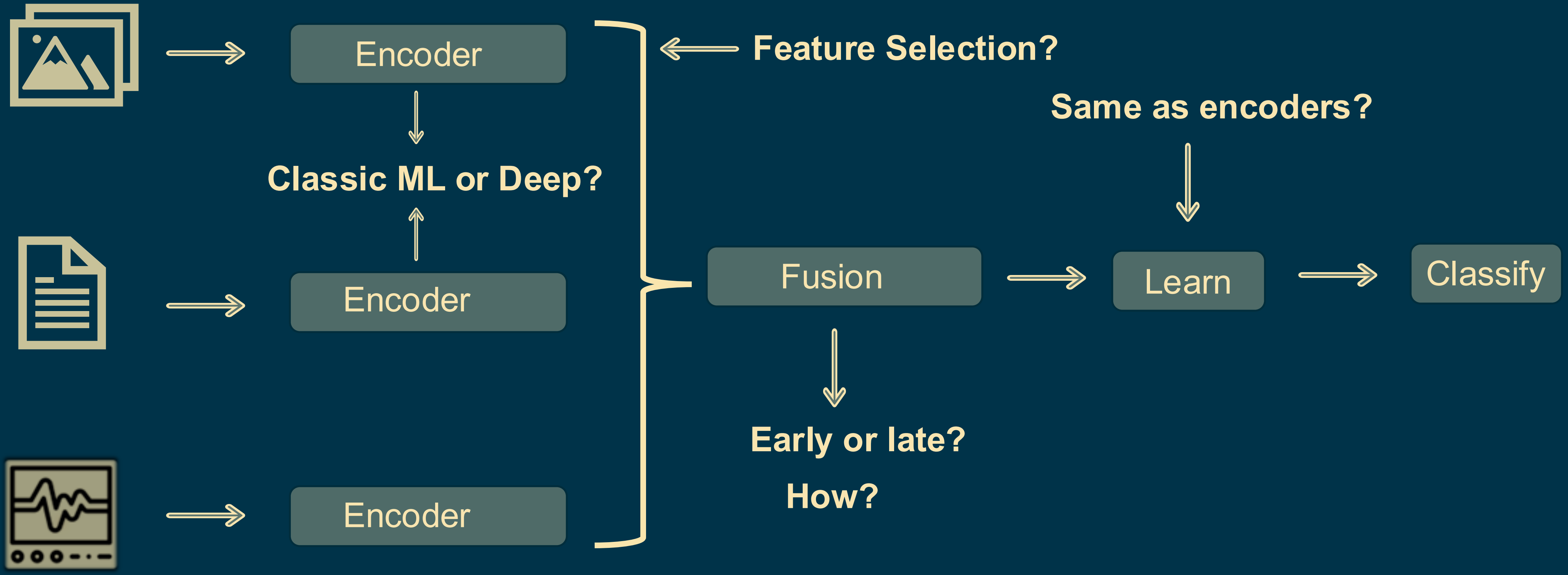
In this talk

Why fusion?

Fusion Techniques

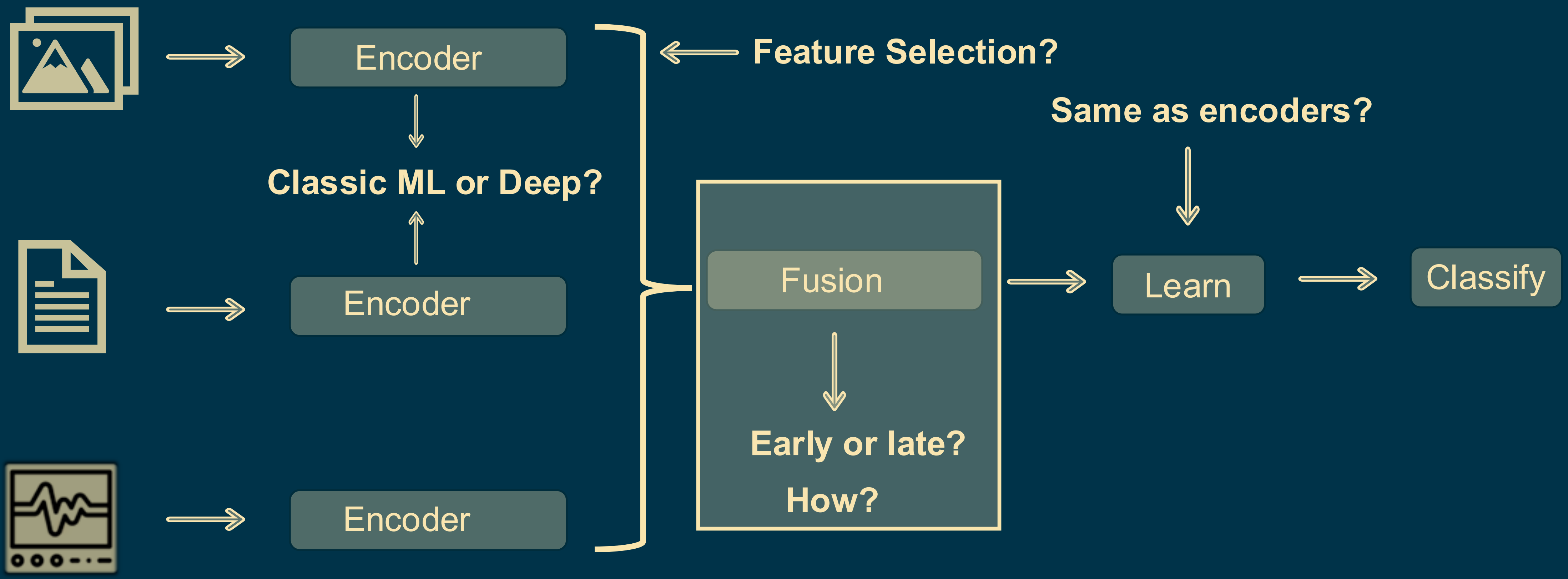
Interesting Issues

Recall



Encoder
Classic ML or Deep?
Encoder
Encoder
Feature Selection?
Same as encoders?
Fusion
Learn
Classify
Early or late?
How?

Recall



Encoder

Encoder

Encoder

Classic ML or Deep?

Feature Selection?

Fusion

Early or late?

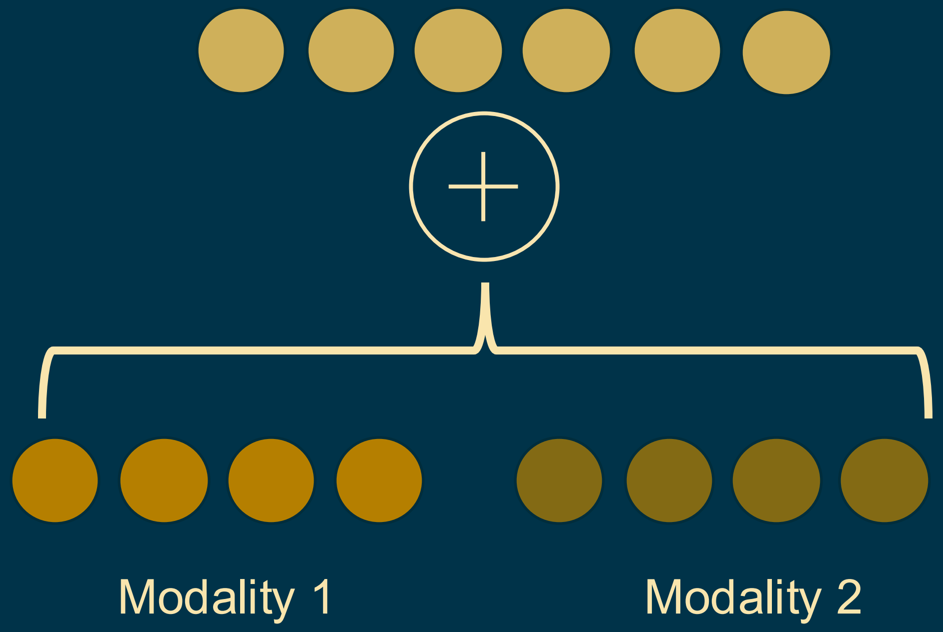
How?

Same as encoders?

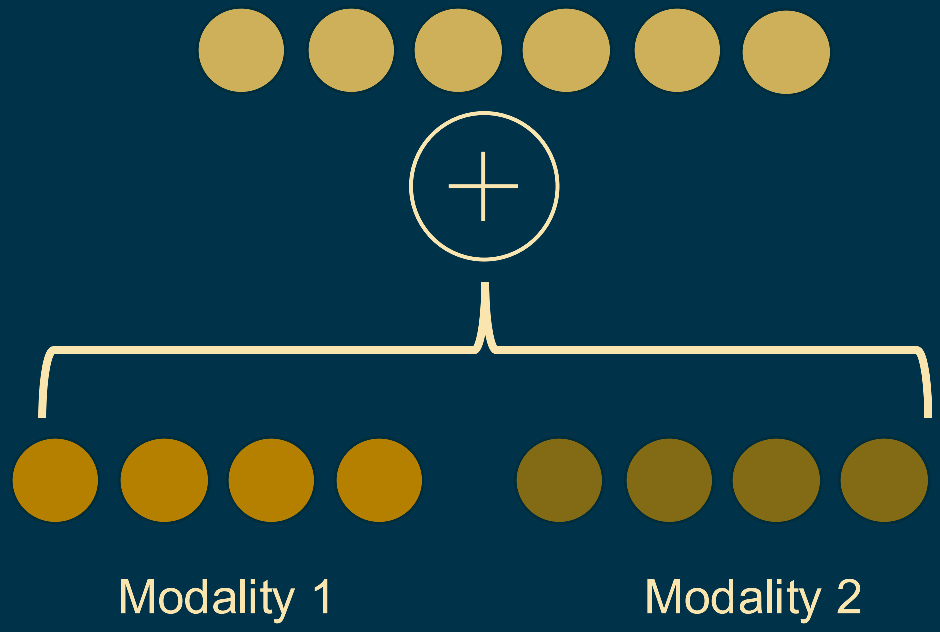
Learn

Classify

Initial attempts

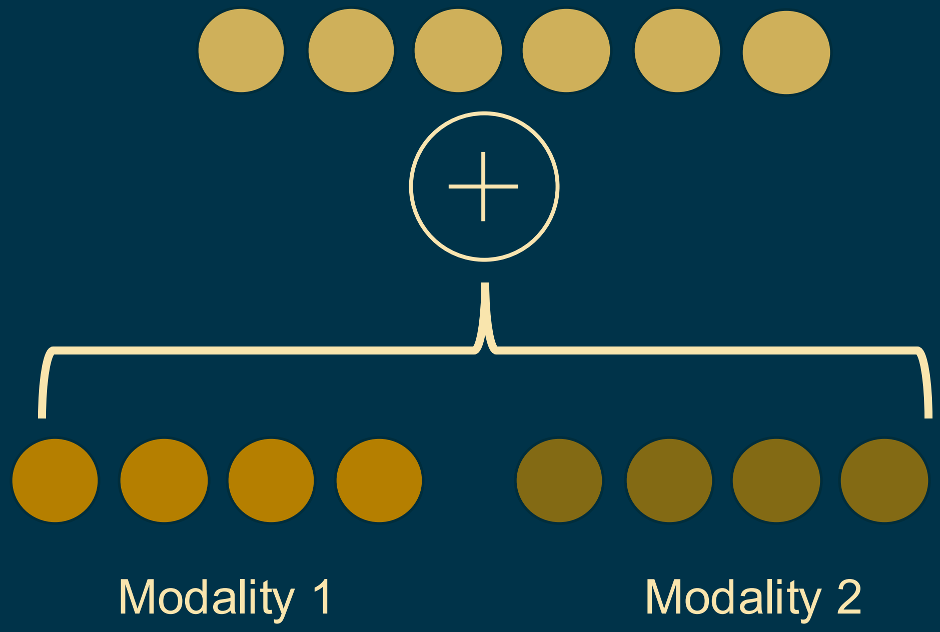


Initial attempts

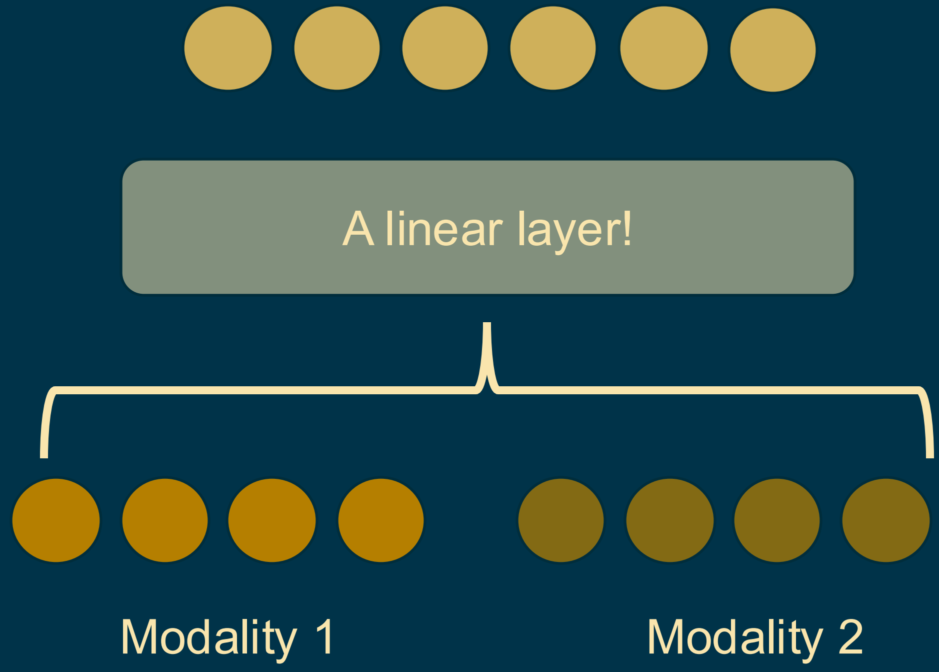


Or?

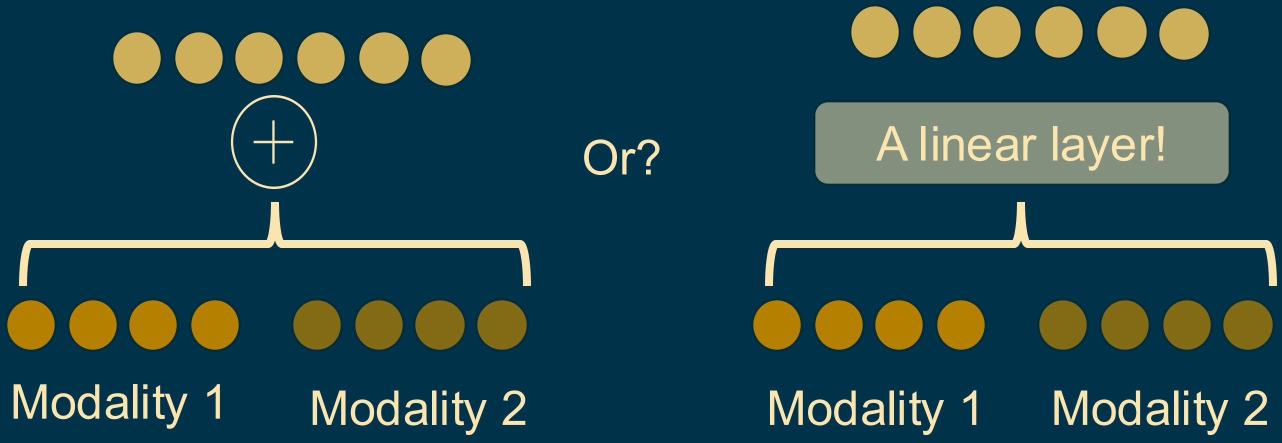
Initial attempts



Or?

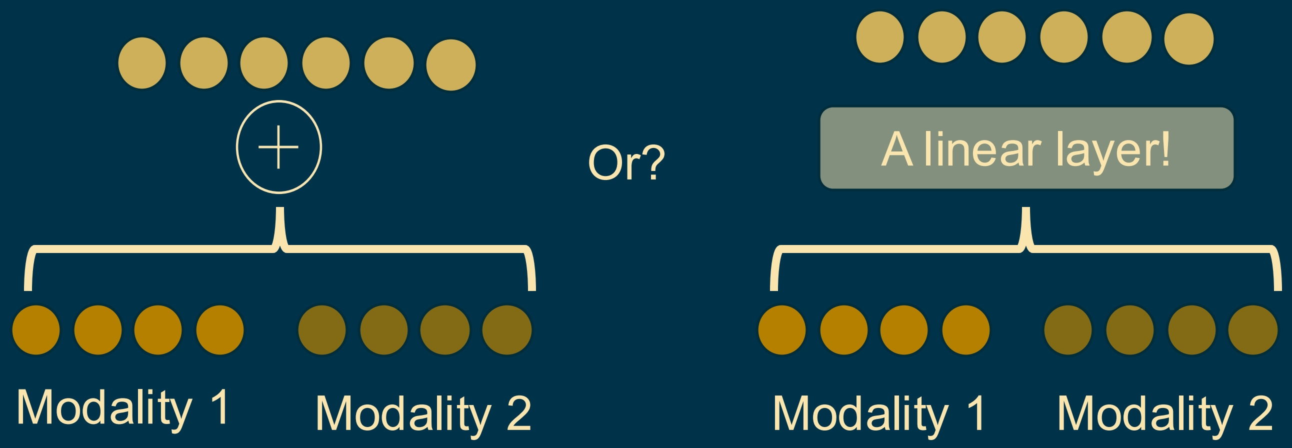


Initial attempts



Features

Initial attempts



Or?

A linear layer!

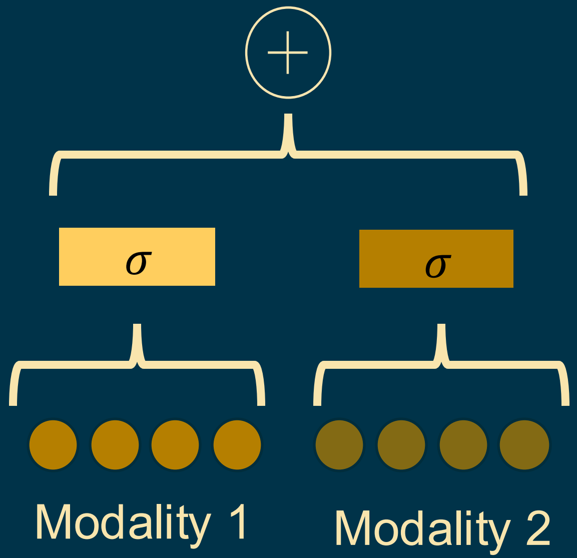
Modality 1

Modality 2

Modality 1

Modality 2

Features



σ

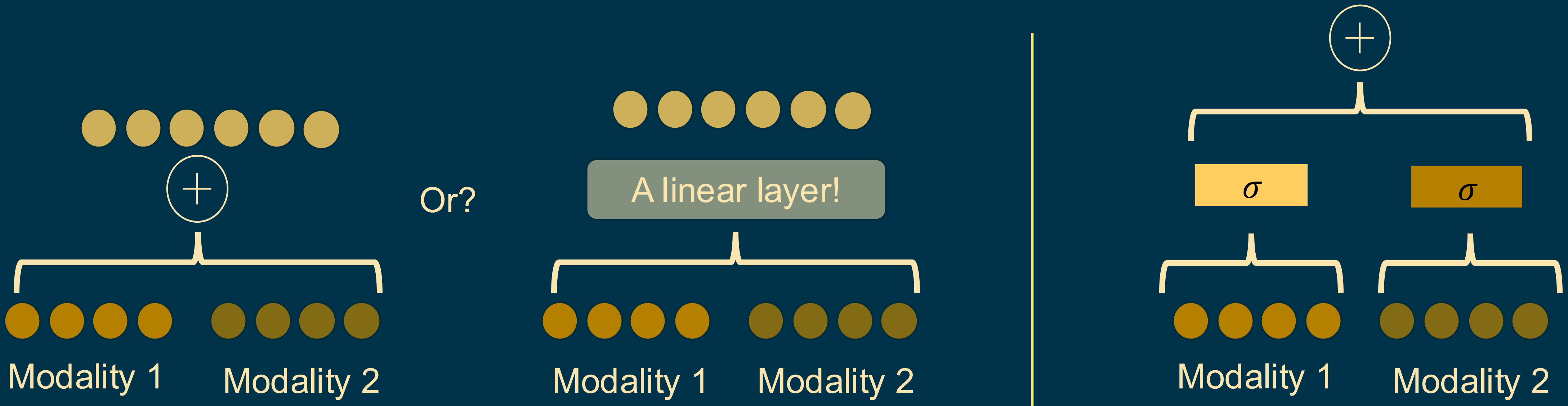
σ

Modality 1

Modality 2

Scores

Initial attempts



Or?

A linear layer!

σ

σ

Modality 1

Modality 2

Modality 1

Modality 2

Modality 1

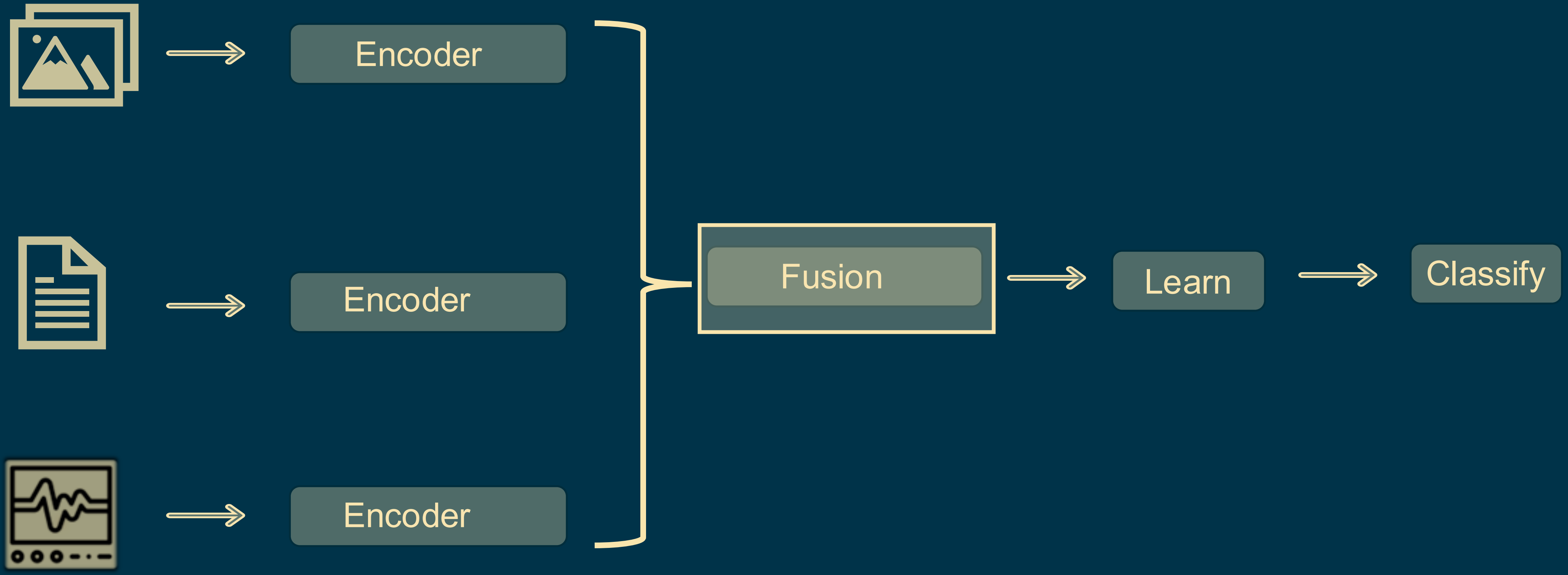
Modality 2

Features

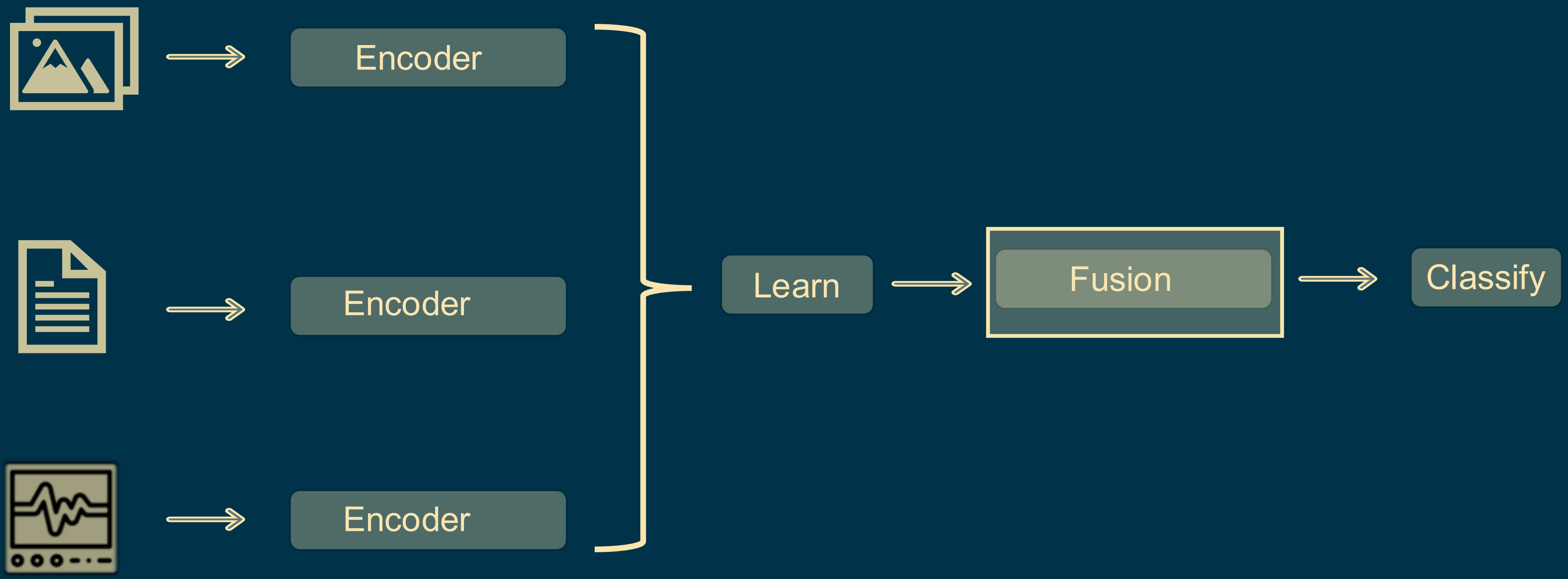
Scores

And when do we do all this?

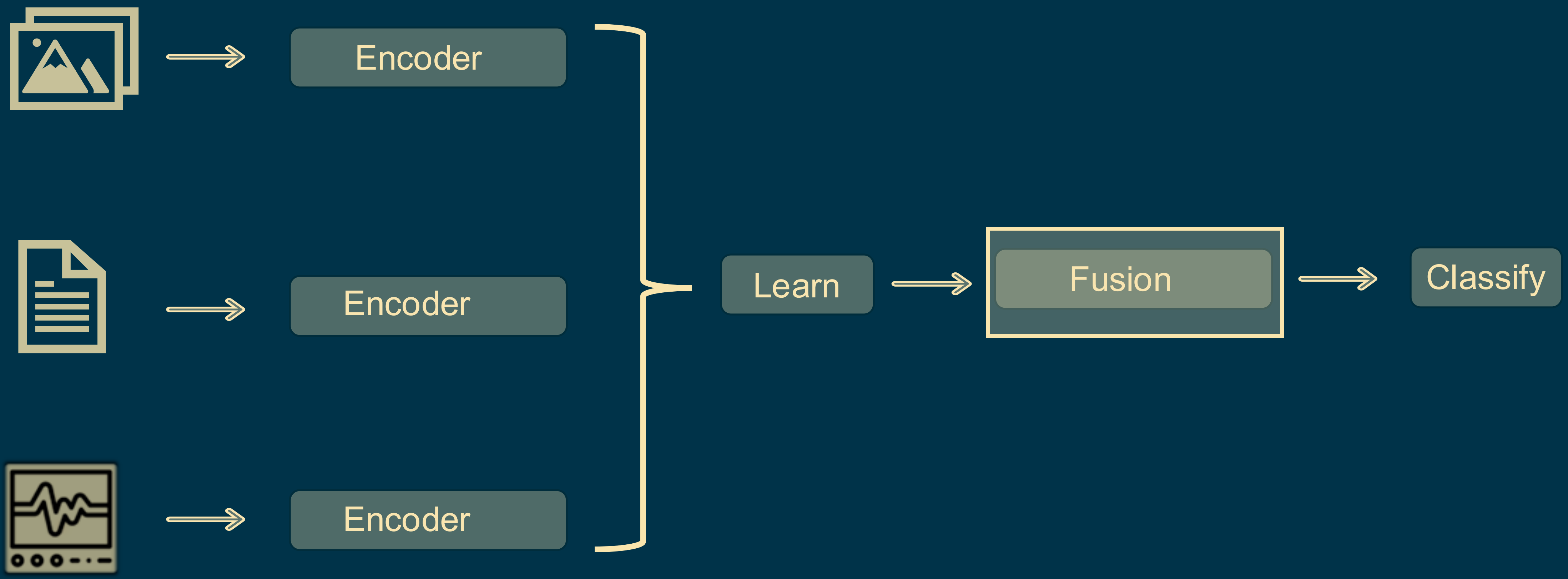
Flexible timing



Flexible timing



Flexible timing



But wait

Is there something deeper going on?

Recall: CLIP is all you need?

Not exactly

CLIP is simply a method of training similar representations

There are other quite a few popular alternatives

You could even pair of a powerful image-only encoder with a large language model!

Recall: CLIP is all you need?

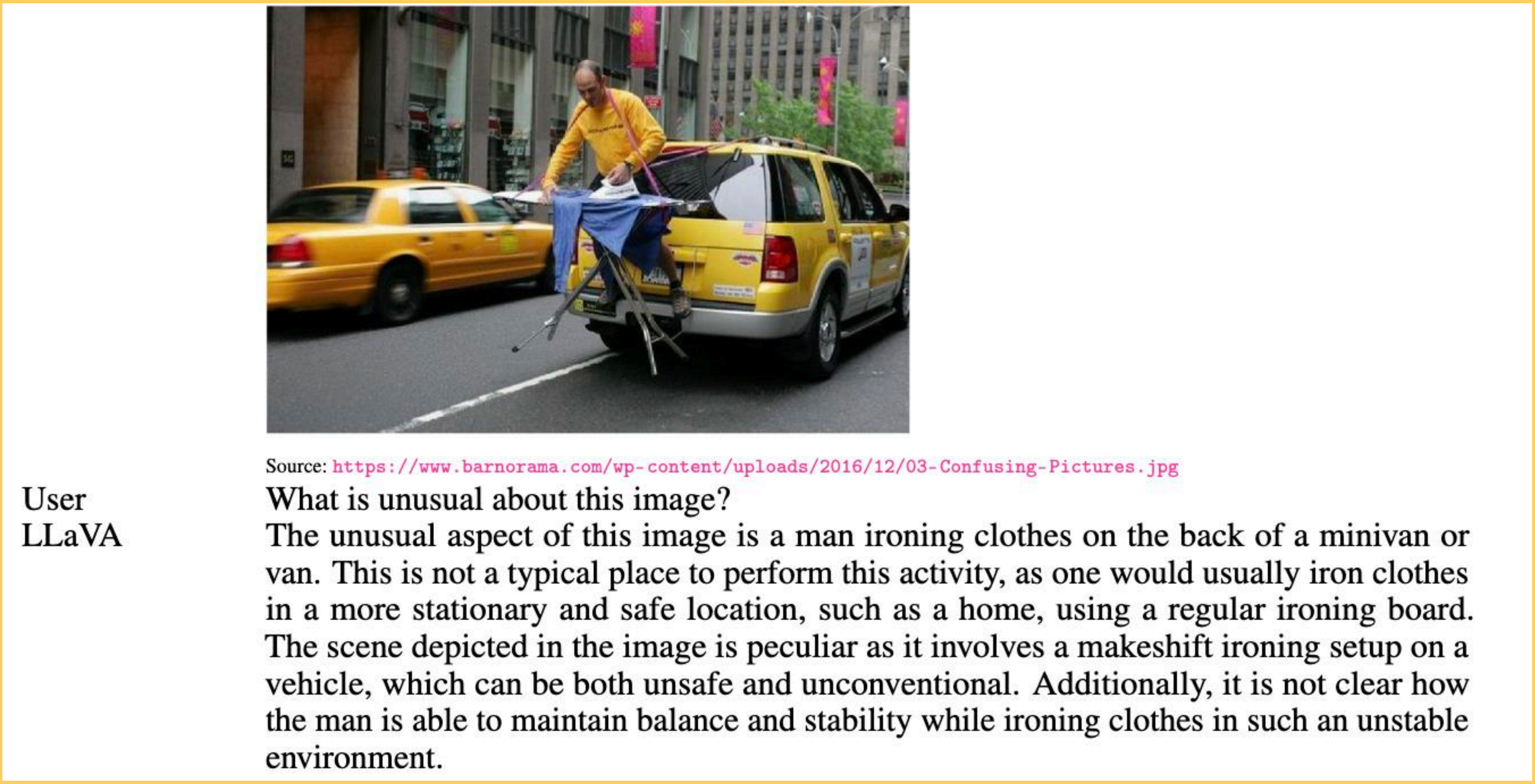
Not exactly

CLIP is simply a method of training similar representations

There are other quite a few popular alternatives

You could even pair of a powerful image-only encoder with a large language model!

Multi-modal Large Language Models

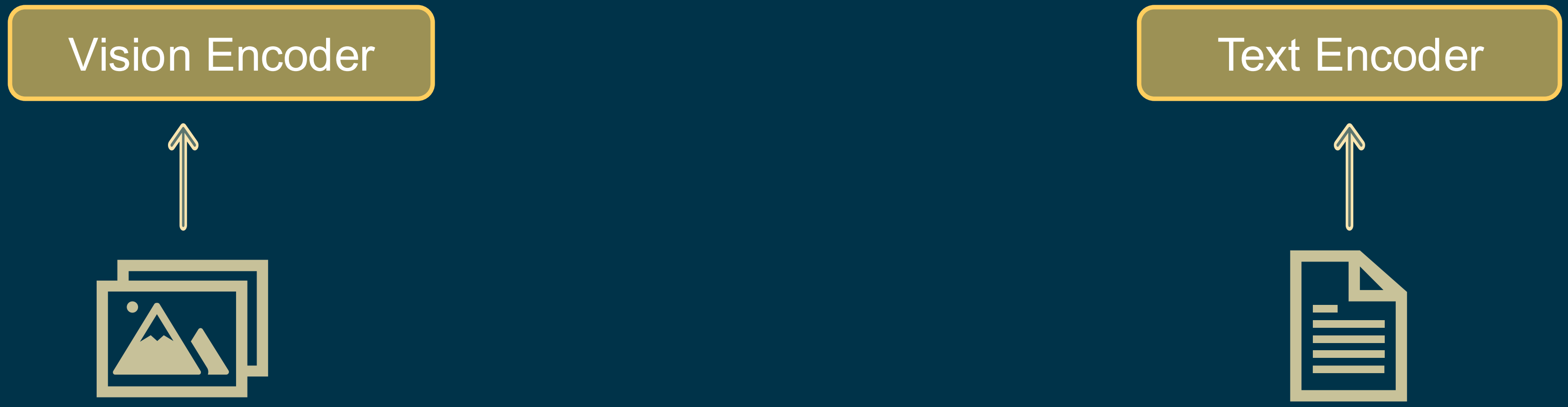


Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

User	What is unusual about this image?
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

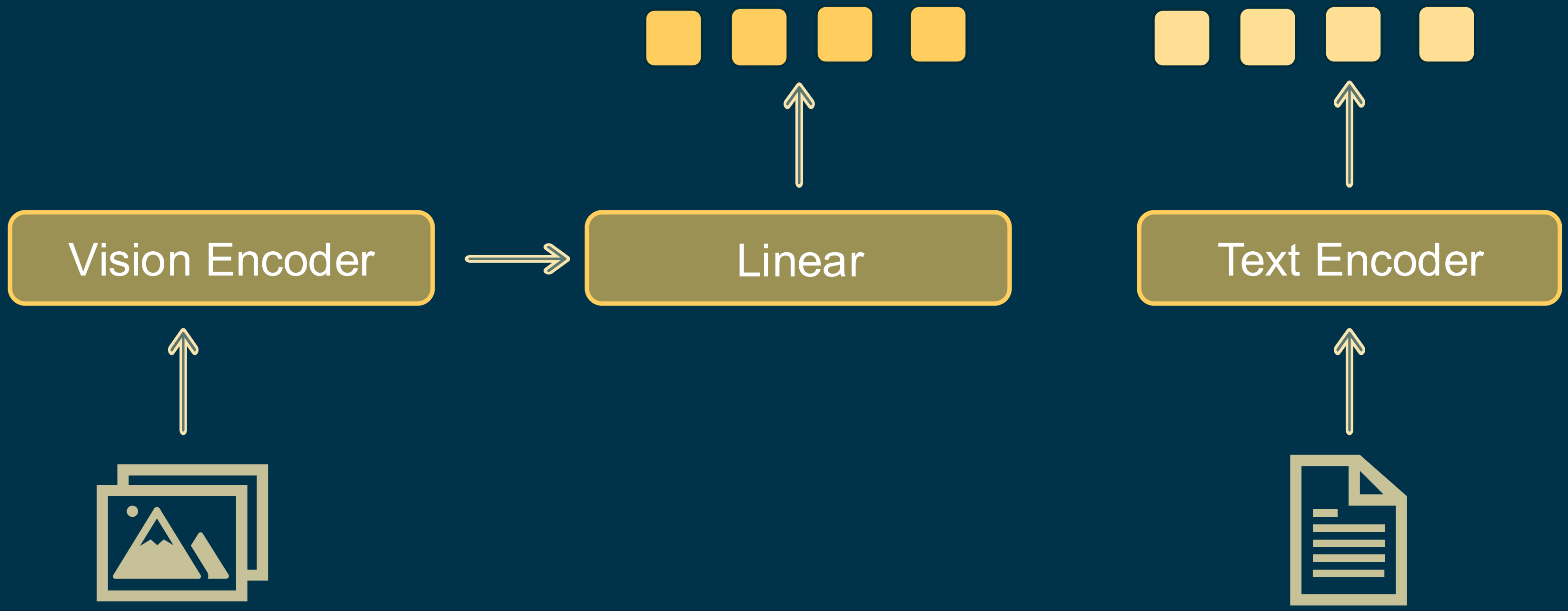
Liu, Haotian, et al. "Visual instruction tuning." Advances in neural information processing systems 36 (2023): 34892-34916.

Multi-modal Large Language Models



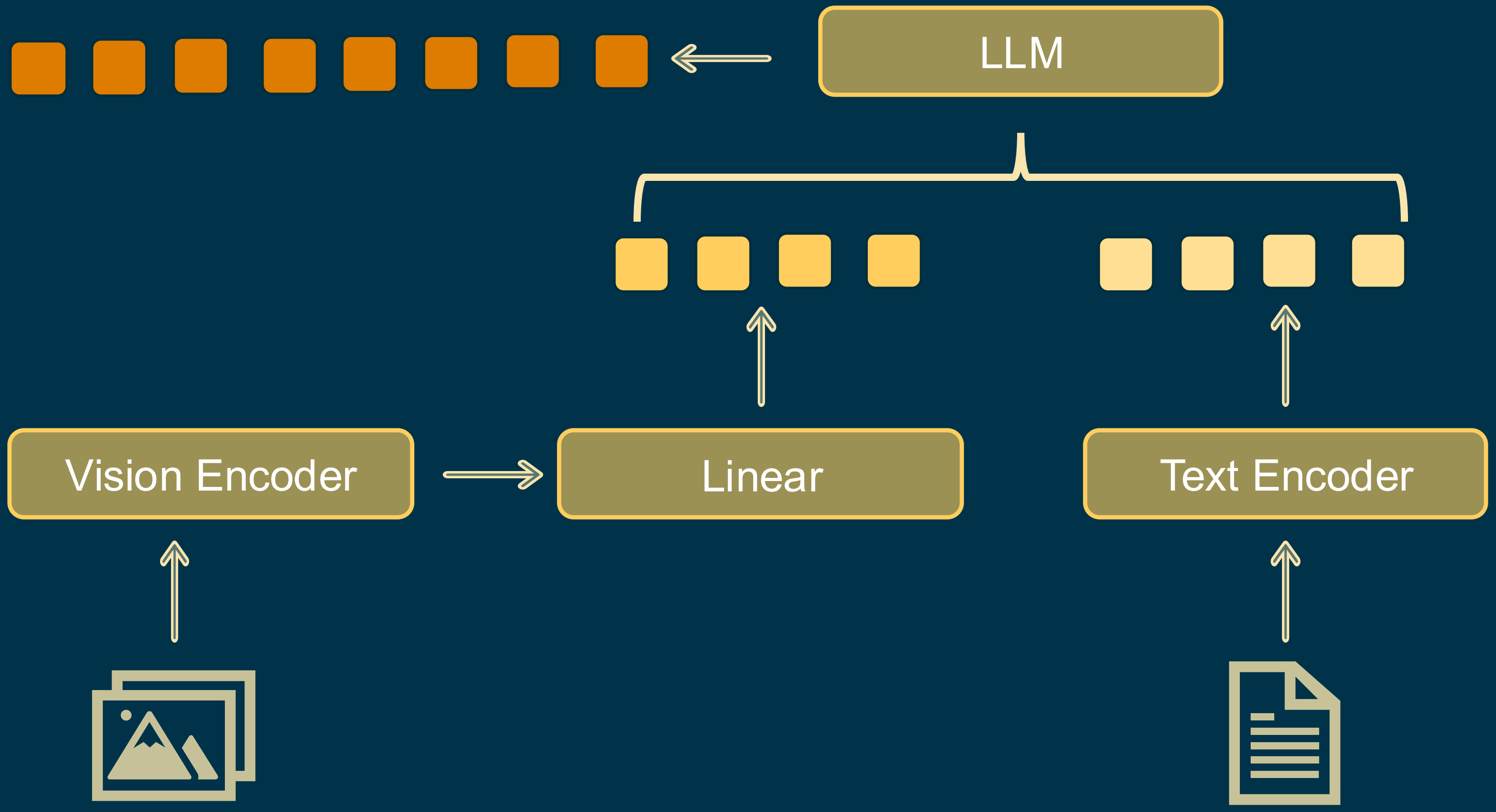
Liu, Haotian, et al. "Visual instruction tuning." Advances in neural information processing systems 36 (2023): 34892-34916.

Multi-modal Large Language Models



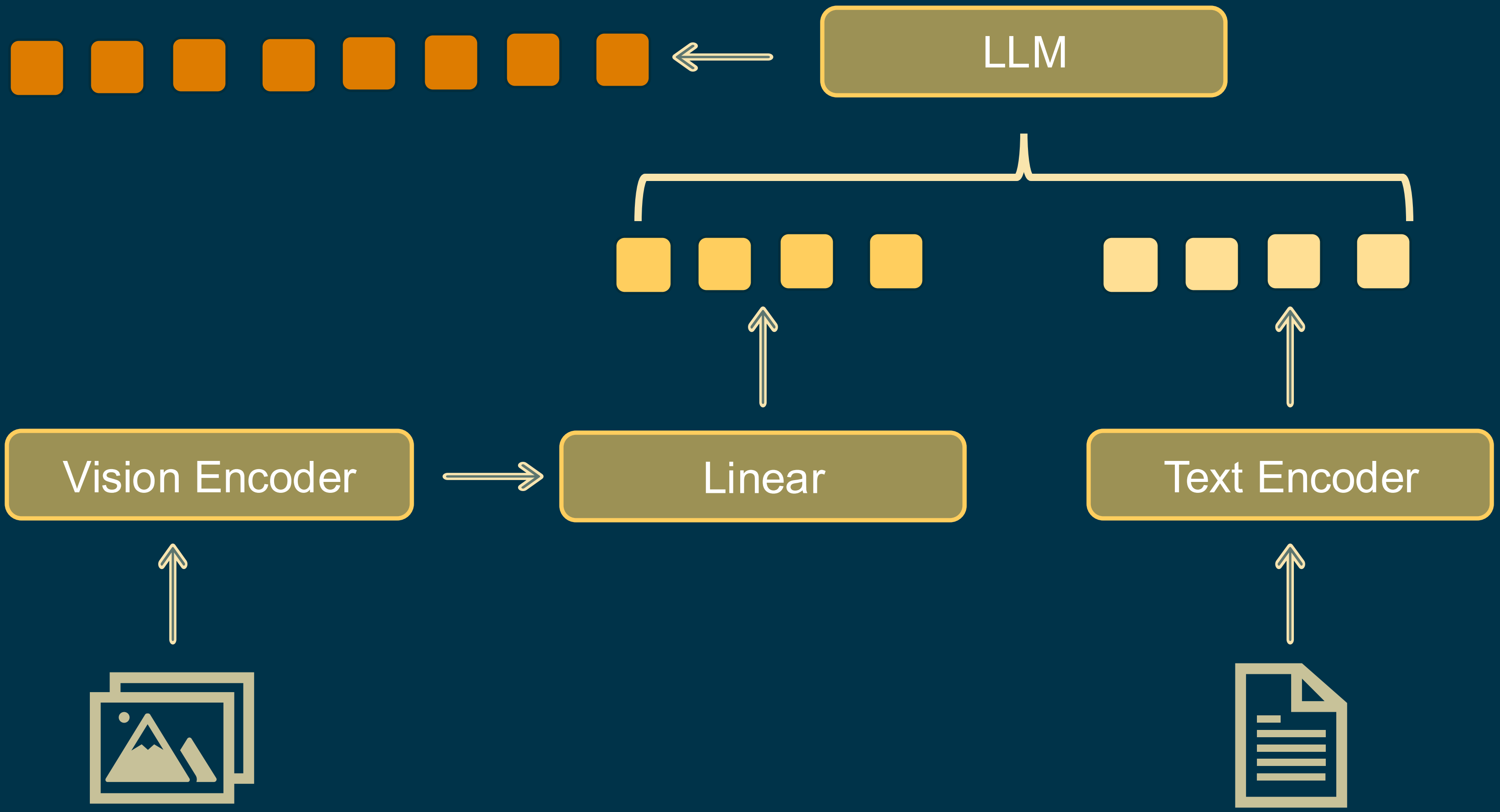
Liu, Haotian, et al. "Visual instruction tuning." Advances in neural information processing systems 36 (2023): 34892-34916.

Multi-modal Large Language Models



Liu, Haotian, et al. "Visual instruction tuning." Advances in neural information processing systems 36 (2023): 34892-34916.

Multi-modal Large Language Models



Astonishingly simple and powerful

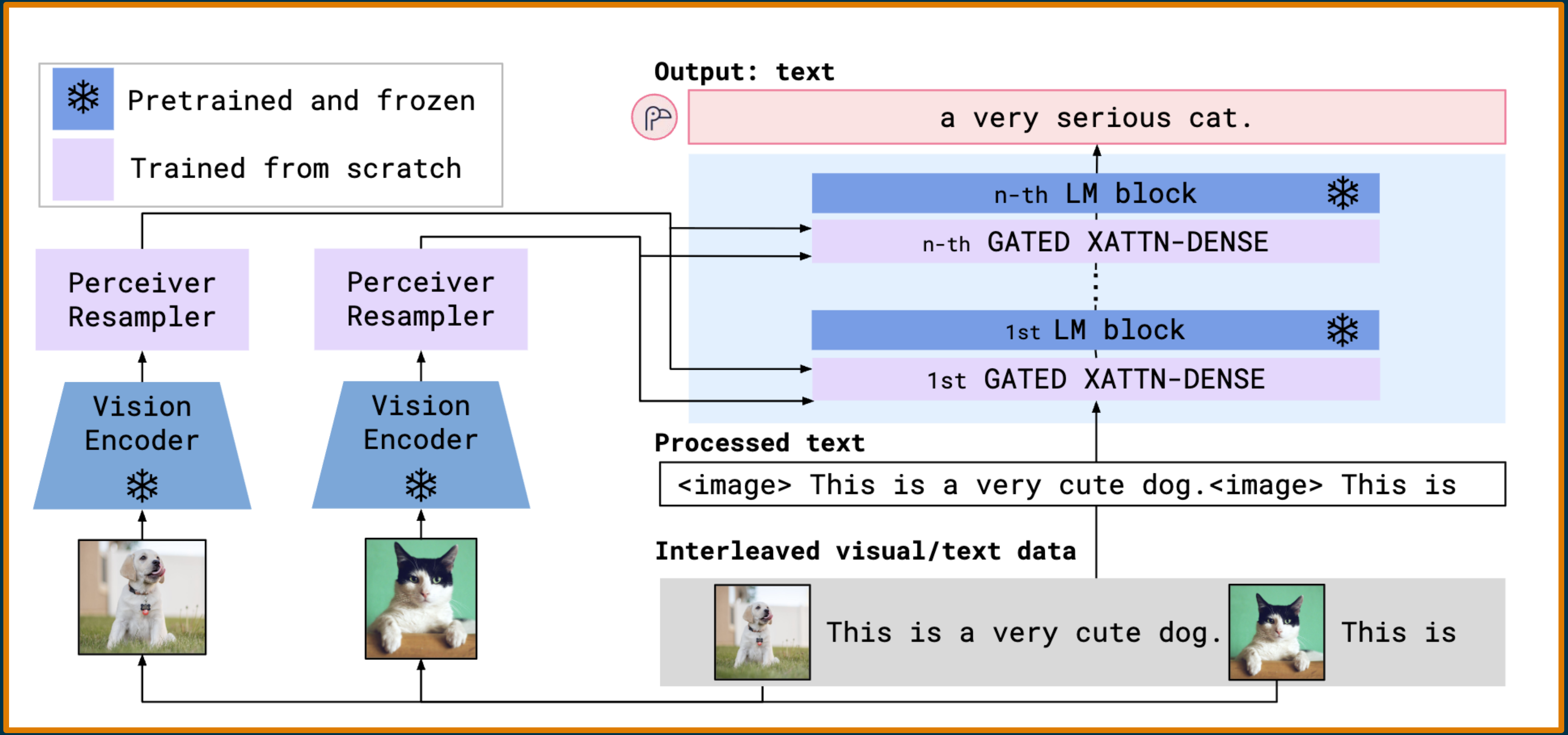
Fusion is simply a linear layer!

Liu, Haotian, et al. "Visual instruction tuning." Advances in neural information processing systems 36 (2023): 34892-34916.

Multi-modal Large Language Models

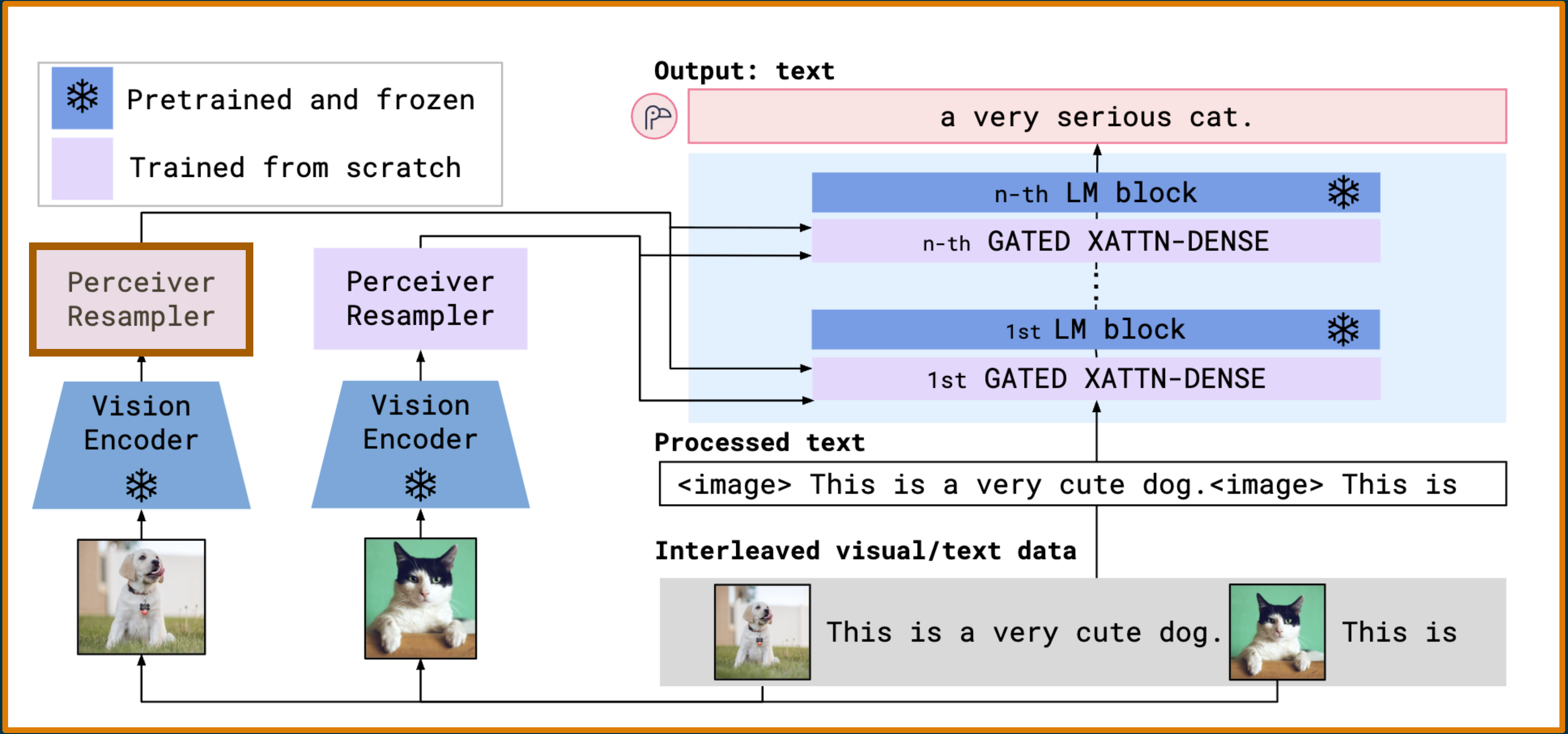
Other ways to fuse?

Multi-modal Large Language Models



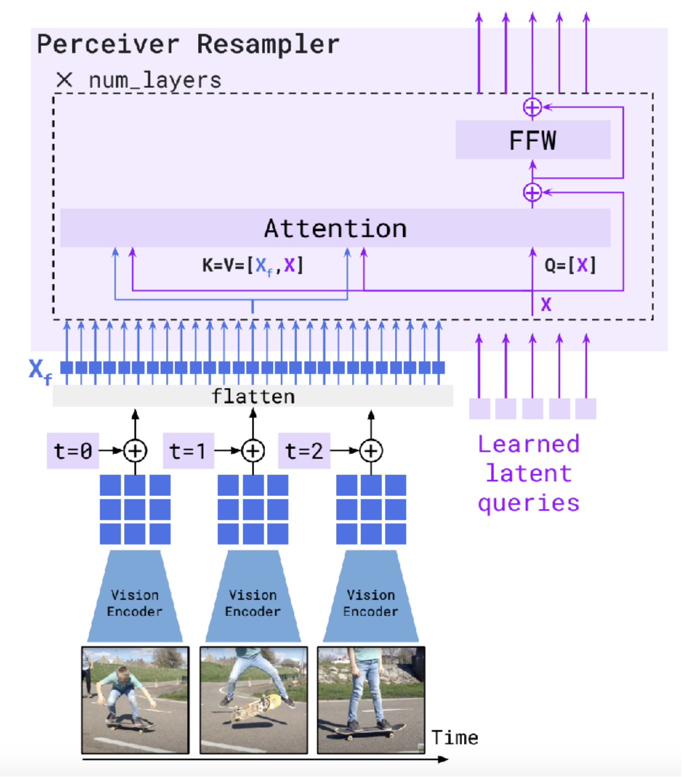
Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems 35 (2022): 23716-23736.

Multi-modal Large Language Models



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems 35 (2022): 23716-23736.

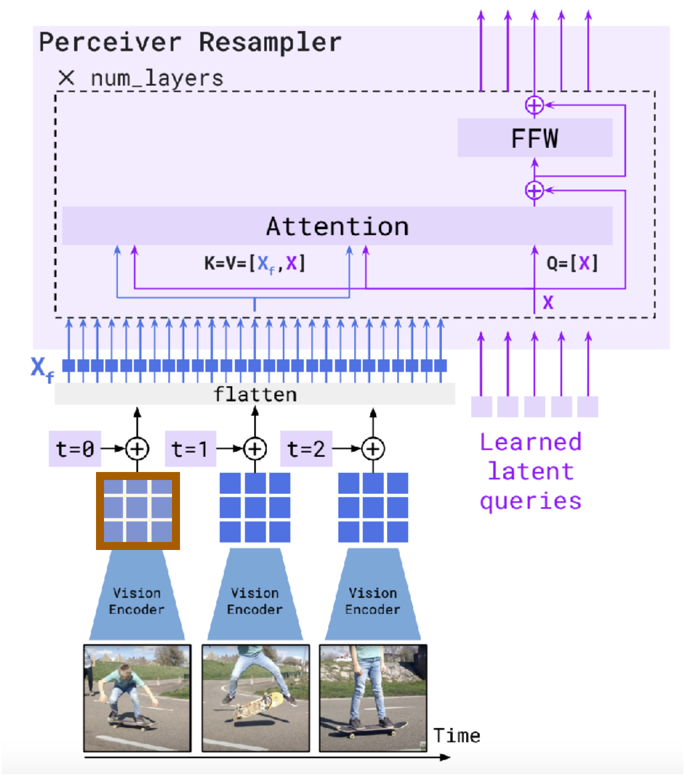
Multi-modal Large Language Models



The Perceiver Resampler

Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems 35 (2022): 23716-23736.

Multi-modal Large Language Models

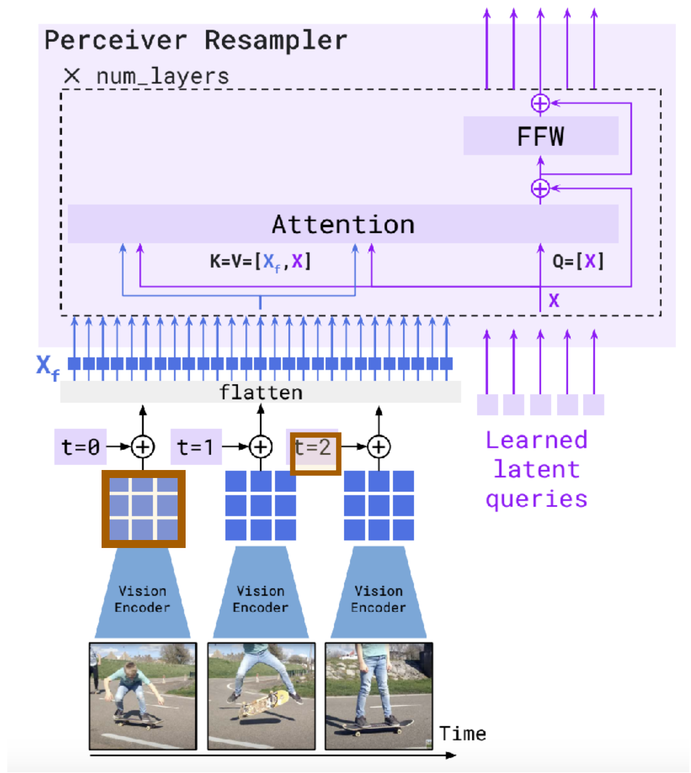


The Perceiver Resampler

Input: Spatio-temporal features

Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems 35 (2022): 23716-23736.

Multi-modal Large Language Models



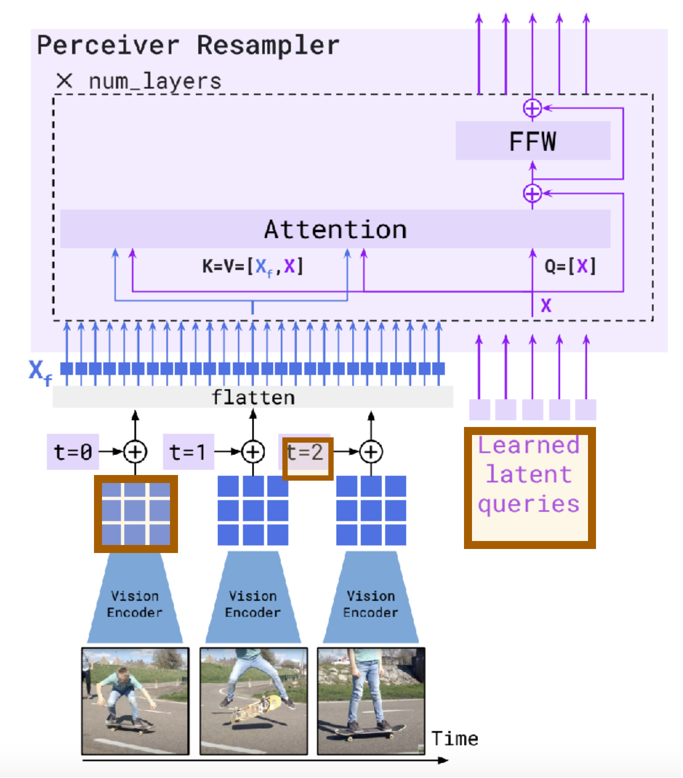
The Perceiver Resampler

Input: Spatio-temporal features

Input: Temporal position embeddings

Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems 35 (2022): 23716-23736.

Multi-modal Large Language Models



The Perceiver Resampler

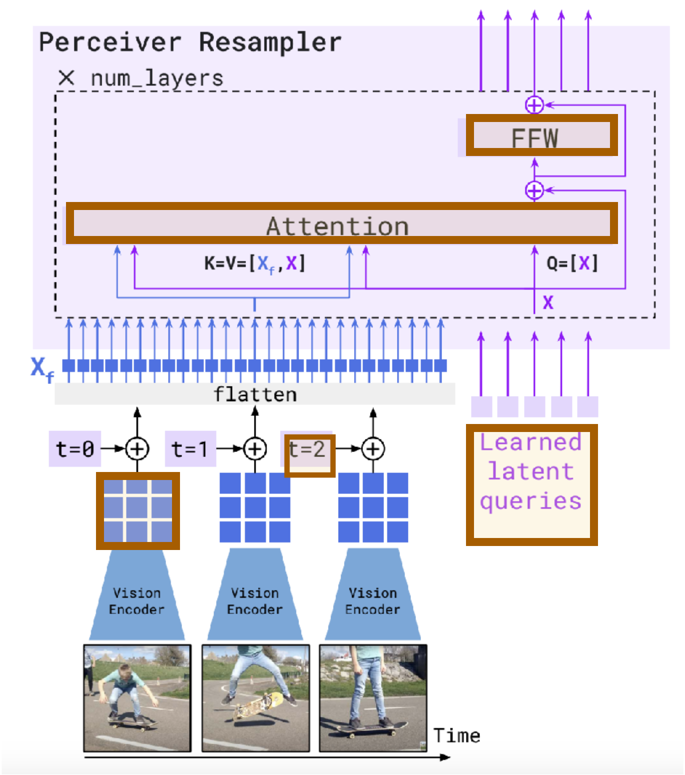
Input: Spatio-temporal features

Input: Temporal position embeddings

Input: R learnable queries

Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems 35 (2022): 23716-23736.

Multi-modal Large Language Models



The Perceiver Resampler

Input: Spatio-temporal features

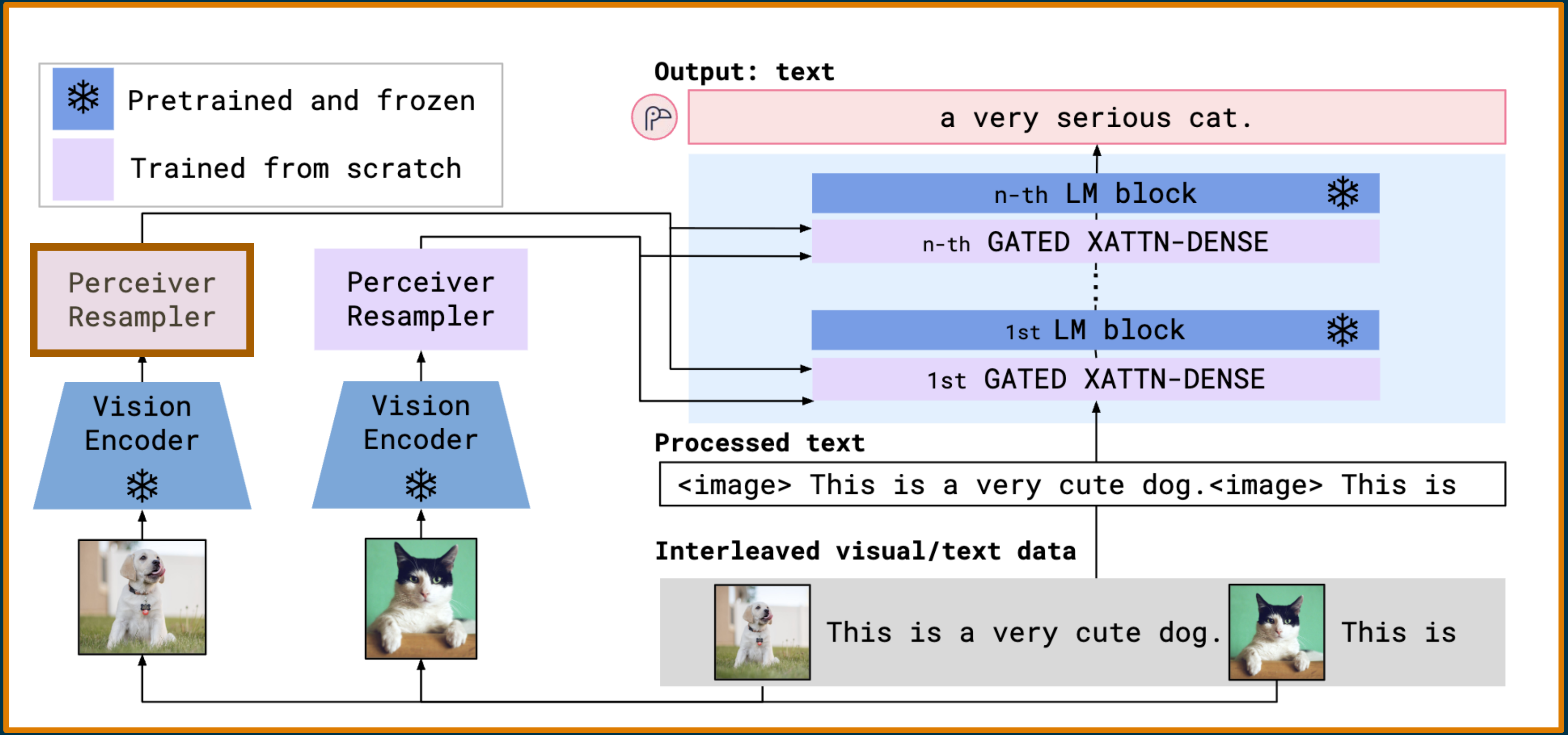
Input: Temporal position embeddings

Input: R learnable queries

Output: Restricted set of visual tokens

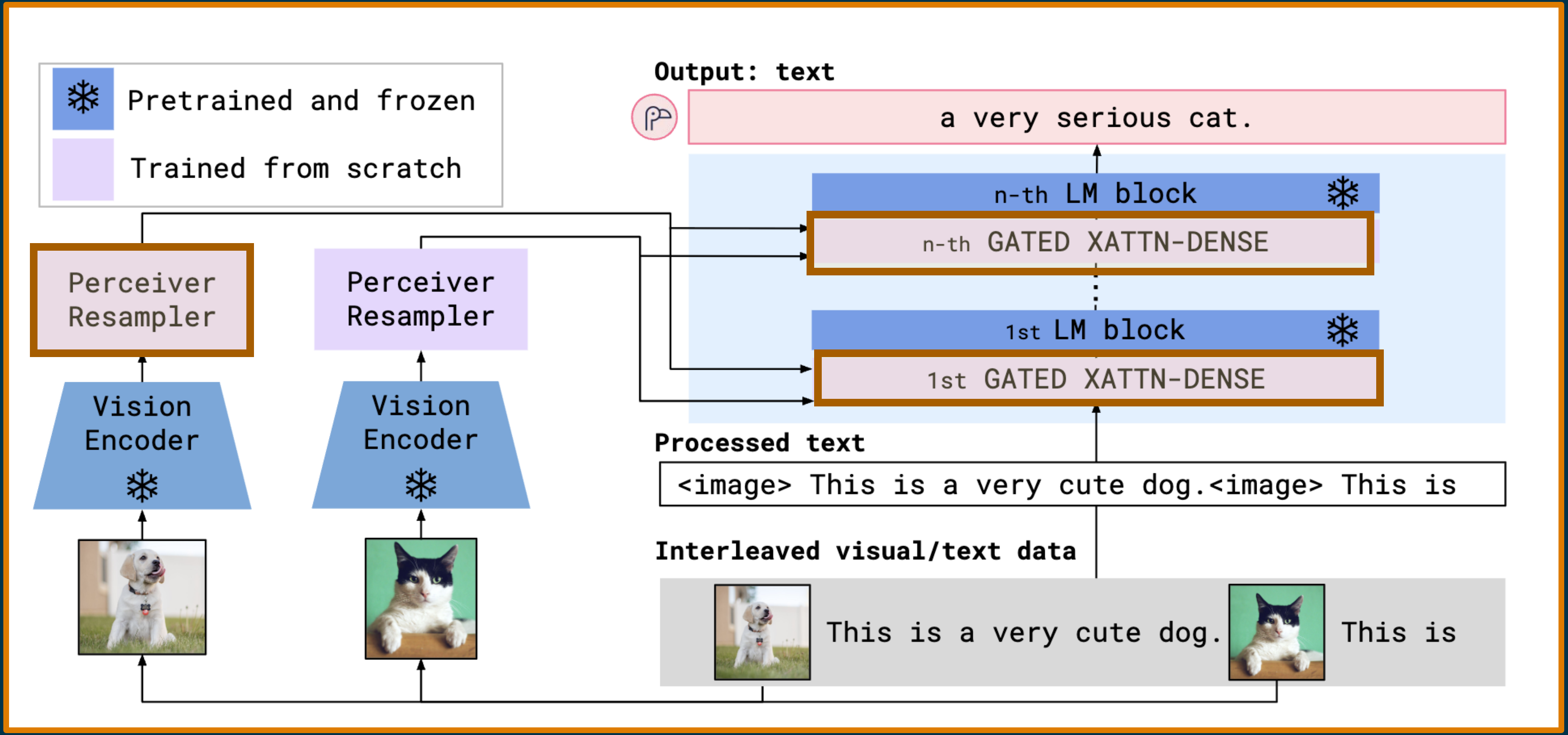
Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems 35 (2022): 23716-23736.

Multi-modal Large Language Models



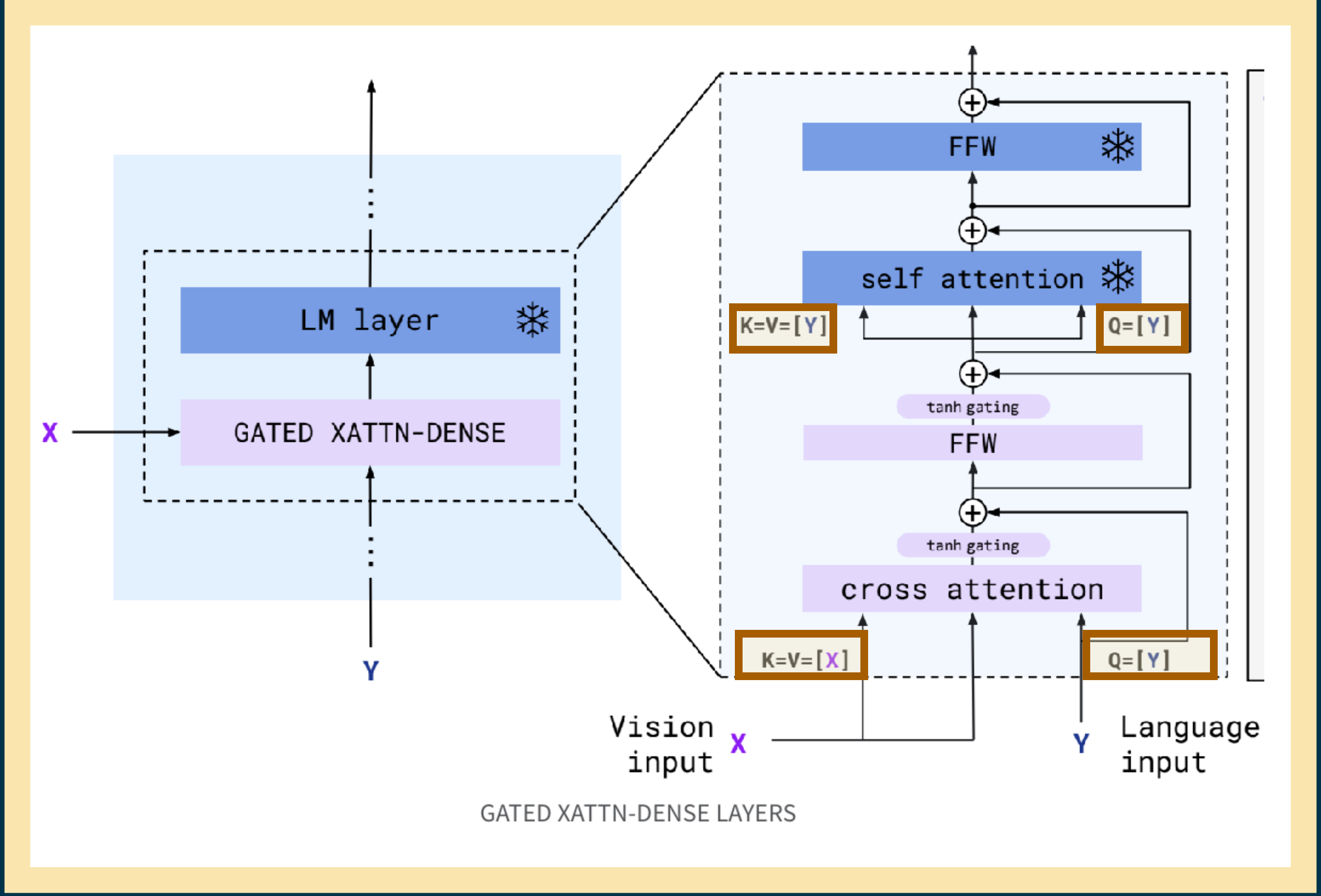
Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems 35 (2022): 23716-23736.

Multi-modal Large Language Models



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems 35 (2022): 23716-23736.

Multi-modal Large Language Models



Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems 35 (2022): 23716-23736.

In this talk

Why fusion?

Fusion Techniques

Interesting Issues

In this talk

Why fusion?

Fusion Techniques

Interesting Issues

Multi-modal Fusion: Interesting Issues

Fusion is a way to “connect” modalities

Multi-modal Fusion: Interesting Issues

Fusion is a way to “connect” modalities

Are some connections better than others?

Multi-modal Fusion: Interesting Issues

Fusion is a way to “connect” modalities

Are some connections better than others?

What problems should we be aware of?

Multi-modal Fusion: Interesting Issues

Fusion is a way to “connect” modalities

Are some connections better than others?

What problems should we be aware of?

Multi-modal Fusion: Grounding

The *Symbol Grounding* Problem

“How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols?”

Harnad, Stevan. "The symbol grounding problem." *Physica D: Nonlinear Phenomena* 42.1-3 (1990): 335-346.

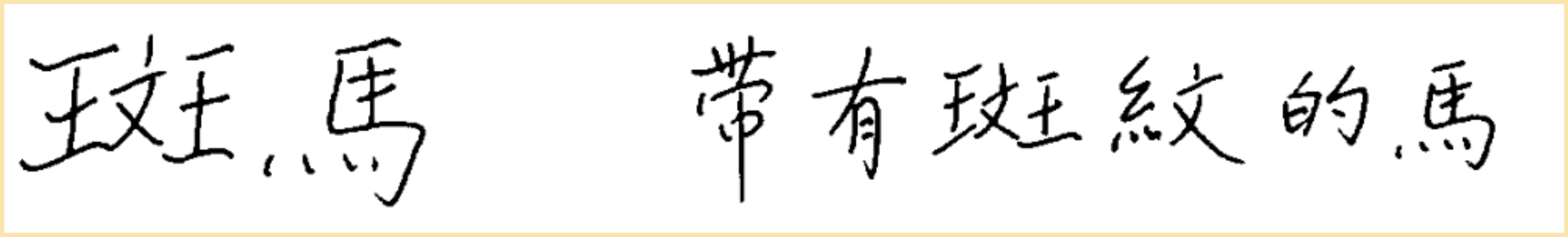
Multi-modal Fusion: Grounding

How do we connect abstract symbols to concrete artefacts in our experience?

Harnad, Stevan. "The symbol grounding problem." *Physica D: Nonlinear Phenomena* 42.1-3 (1990): 335-346.

Multi-modal Fusion: Grounding

Example

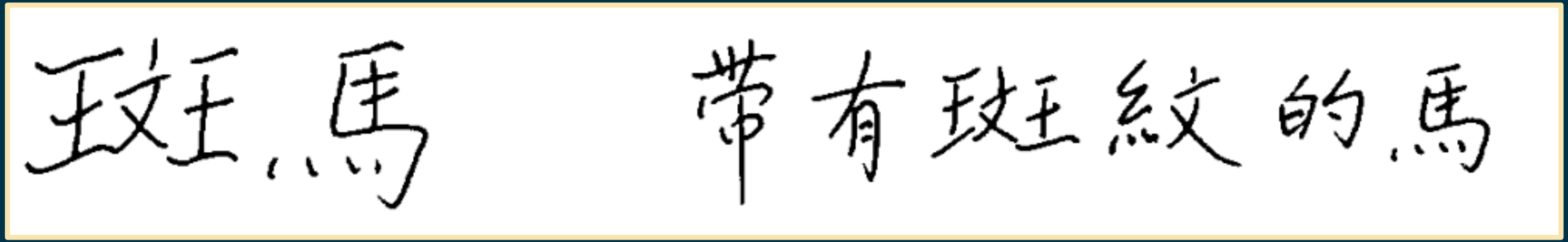


斑馬　带有斑紋的馬

Harnad, Stevan. "The symbol grounding problem." *Physica D: Nonlinear Phenomena* 42.1-3 (1990): 335-346.

Multi-modal Fusion: Grounding

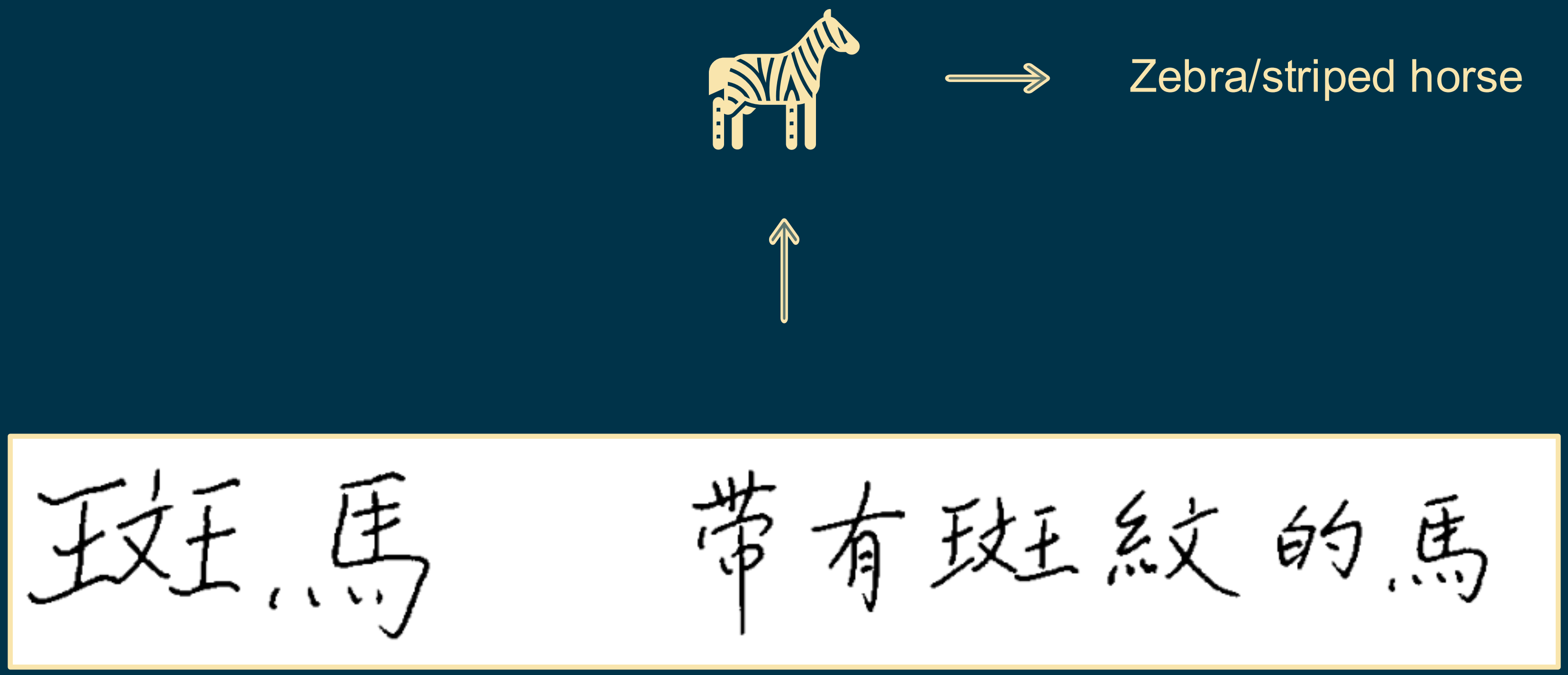
How would a non-native speaker translate this?



斑馬　　带有斑紋的馬

Harnad, Stevan. "The symbol grounding problem." *Physica D: Nonlinear Phenomena* 42.1-3 (1990): 335-346.

Multi-modal Fusion: Grounding



Zebra/striped horse

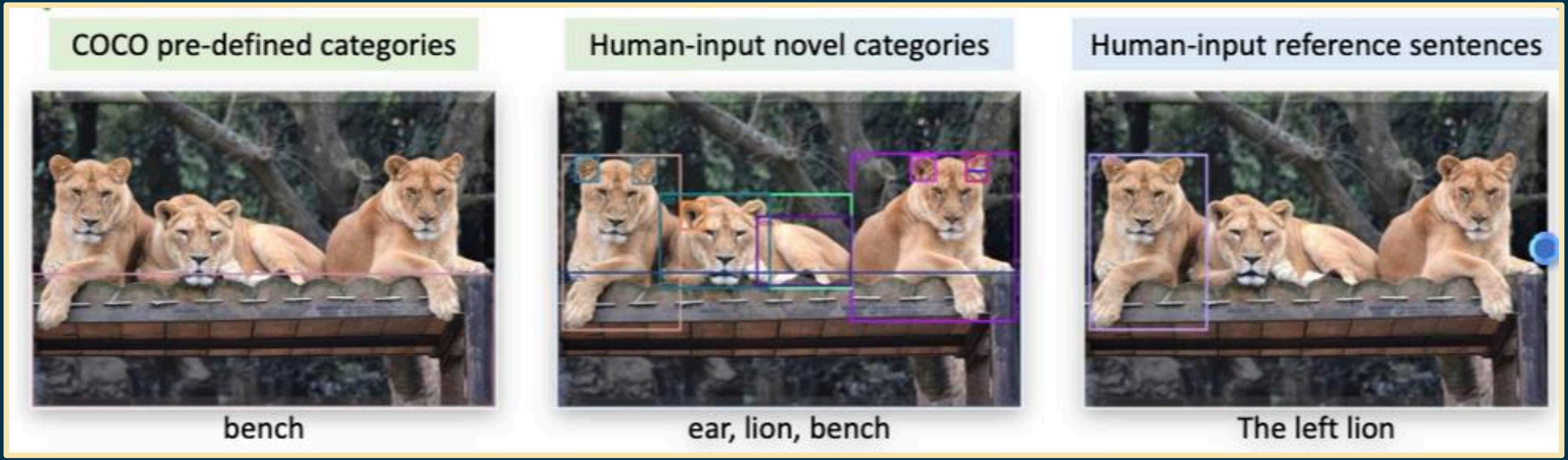
斑馬 带有斑紋的馬

Harnad, Stevan. "The symbol grounding problem." *Physica D: Nonlinear Phenomena* 42.1-3 (1990): 335-346.

Multi-modal Fusion: Grounding

Similarly, to design intelligent systems, one must have experiential grounding baked in

Multi-modal Fusion: Grounding


COCO pre-defined categories
Human-input novel categories
Human-input reference sentences
bench
ear, lion, bench
The left lion

Open-Vocabulary detection with grounded-DINO

Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.

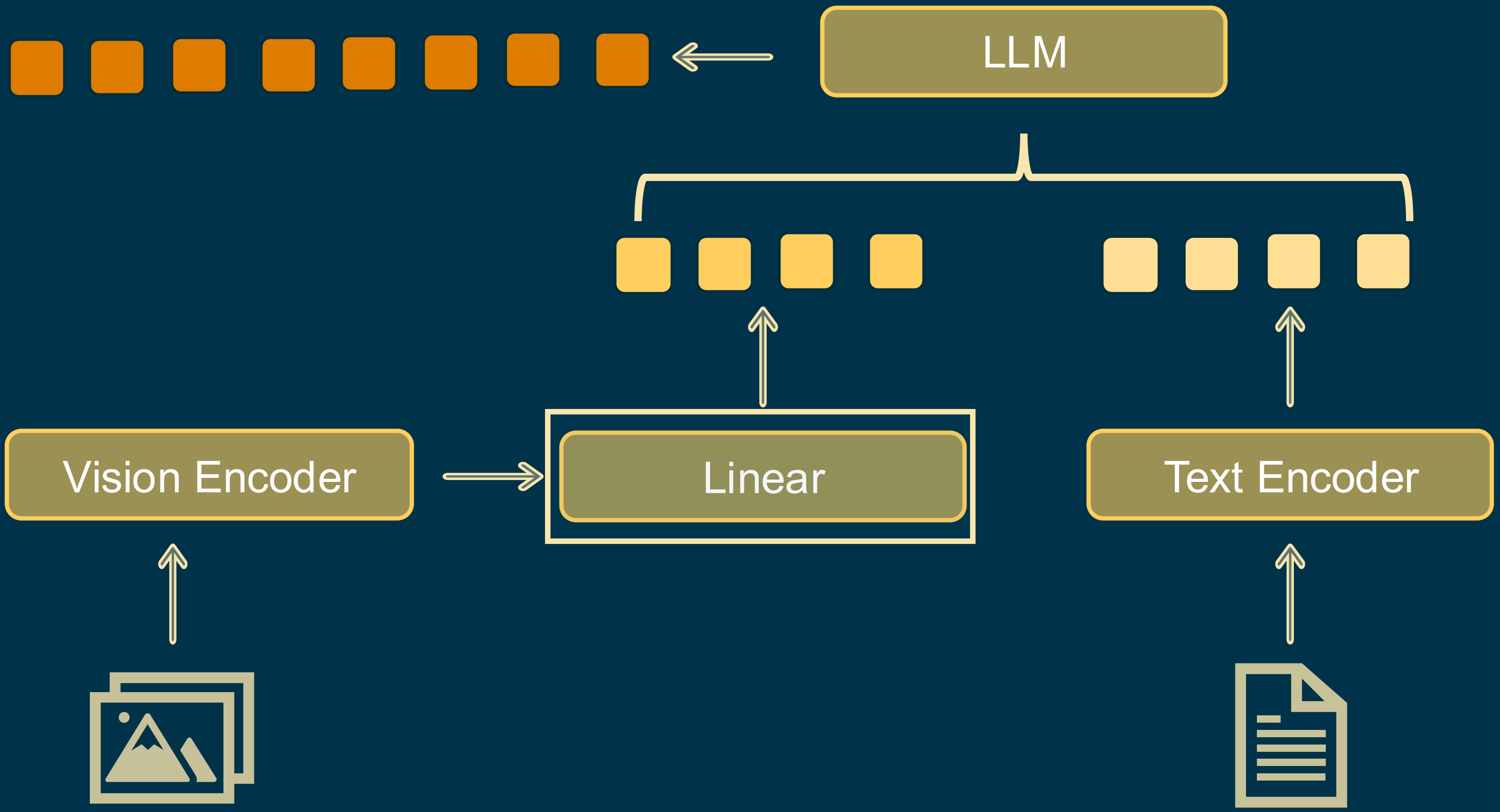
Multi-modal Fusion: Interesting Issues

Fusion is a way to “connect” modalities

Are some connections better than others?

What problems should we be aware of?

Recall the linear layer

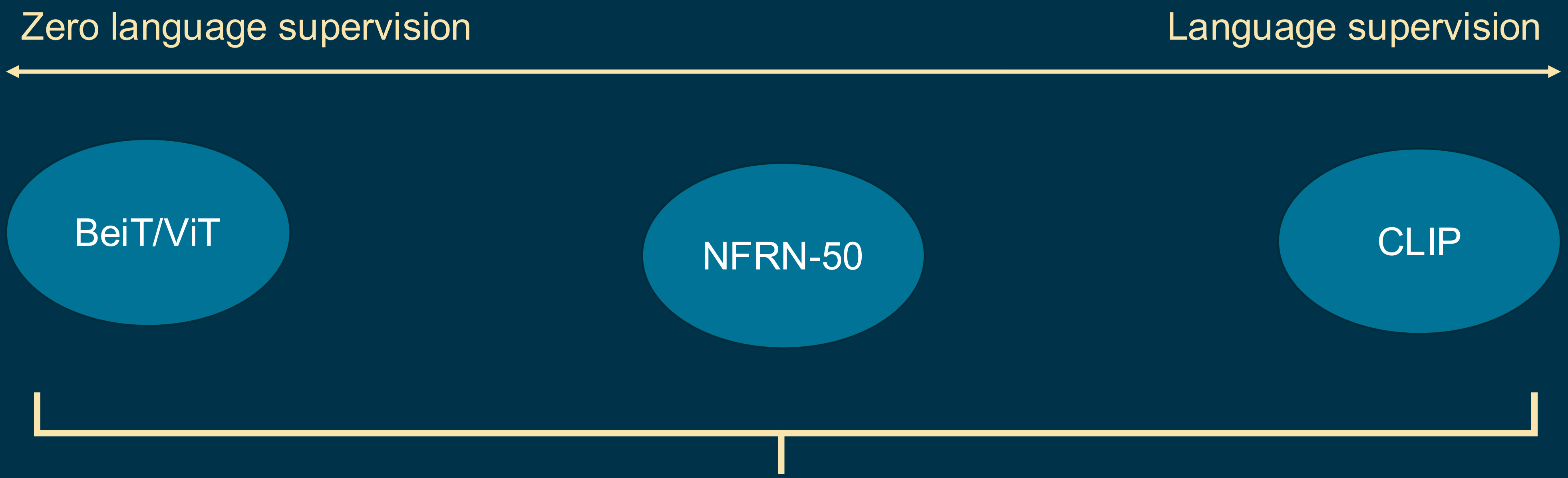


Astonishingly simple and powerful

Fusion is simply a linear layer!

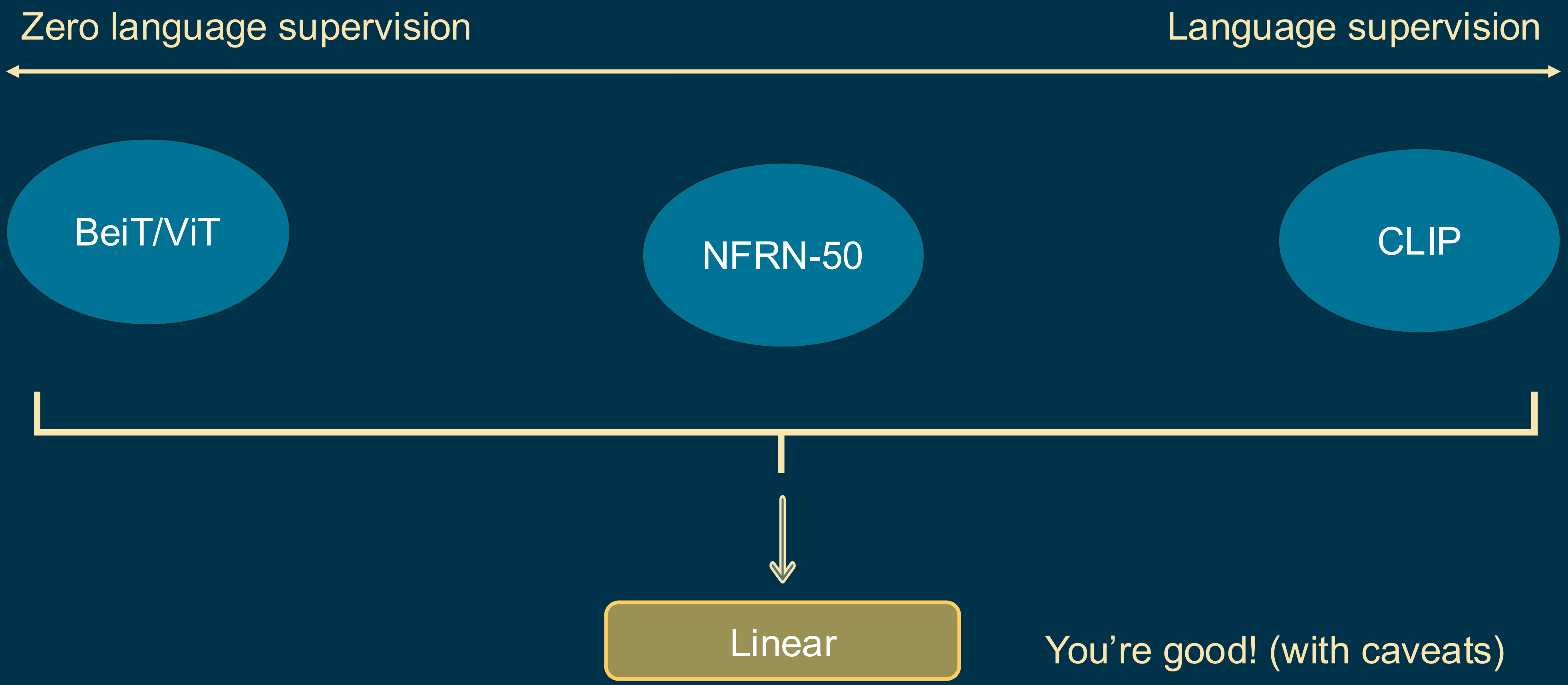
Liu, Haotian, et al. "Visual instruction tuning." Advances in neural information processing systems 36 (2023): 34892-34916.

In fact



Merullo, Jack, et al. "Linearly mapping from image to text space." arXiv preprint arXiv:2209.15162 (2022).

In fact



Zero language supervision

Language supervision

BeiT/ViT

NFRN-50

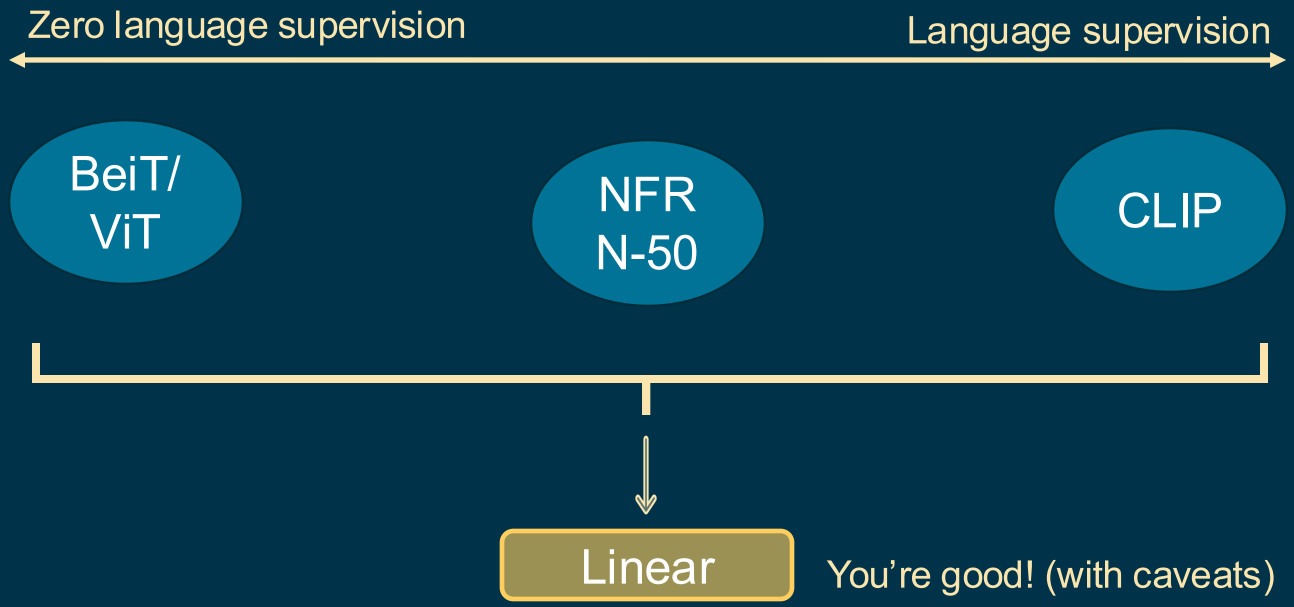
CLIP

Linear

You’re good! (with caveats)

Merullo, Jack, et al. "Linearly mapping from image to text space." arXiv preprint arXiv:2209.15162 (2022).

In fact



You’re good! (with caveats)

Why this is the case is unclear

Active area of research!

Merullo, Jack, et al. "Linearly mapping from image to text space." arXiv preprint arXiv:2209.15162 (2022).

Multi-modal Fusion: Interesting Issues

Fusion is a way to “connect” modalities

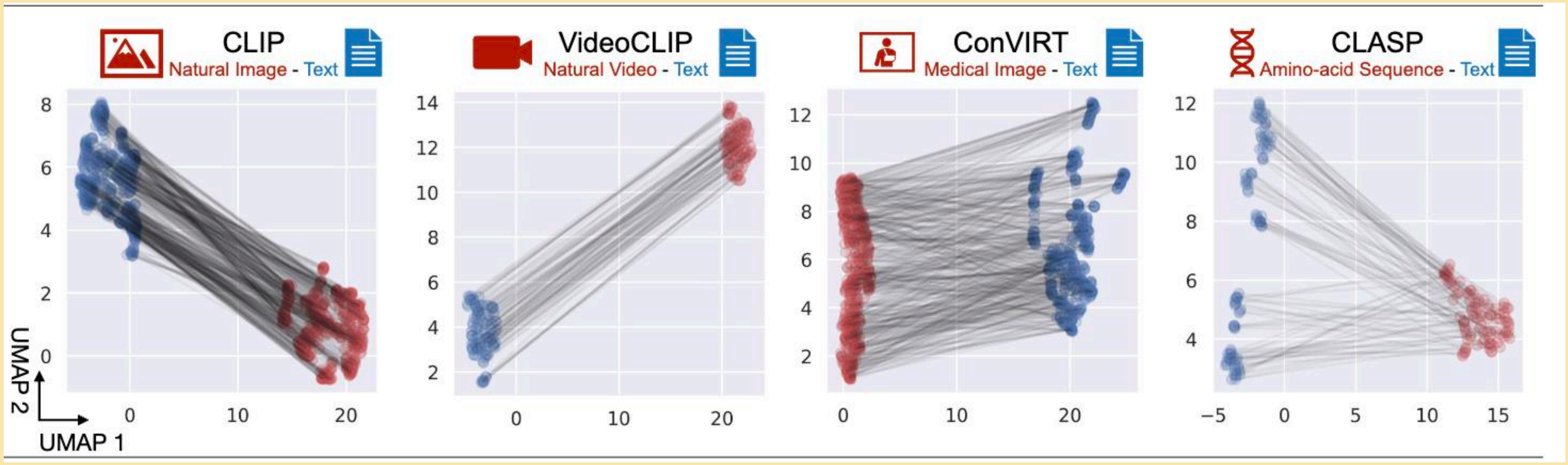
Are some connections better than others?

What problems should we be aware of?

Multi-modal Fusion: The Modality Gap

What happens if we visualize CLIP embeddings?

Multi-modal Fusion: The Modality Gap



Liang, Victor Weixin, et al. "Mind the gap: Understanding the modality gap in Multi-modal contrastive representation learning." Advances in Neural Information Processing Systems 35 (2022): 17612-17625.

Multi-modal Fusion: The Modality Gap

It is currently unclear what to do with this gap

Liang, Victor Weixin, et al. "Mind the gap: Understanding the modality gap in Multi-modal contrastive representation learning." Advances in Neural Information Processing Systems 35 (2022): 17612-17625.

Multi-modal Fusion: The Modality Gap

It is currently unclear what to do with this gap

Modifying the gap affects performance, but unclear how!

Liang, Victor Weixin, et al. "Mind the gap: Understanding the modality gap in Multi-modal contrastive representation learning." Advances in Neural Information Processing Systems 35 (2022): 17612-17625.

Multi-modal Fusion: The Modality Gap

It is currently unclear what to do with this gap

Modifying the gap affects performance, but unclear how!

Tight connections to training dynamics and gradient flow

Liang, Victor Weixin, et al. "Mind the gap: Understanding the modality gap in Multi-modal contrastive representation learning." Advances in Neural Information Processing Systems 35 (2022): 17612-17625.
Yaras, Can, et al. "Explaining and Mitigating the Modality Gap in Contrastive Multi-modal Learning." arXiv preprint arXiv:2412.07909 (2024).

Summary

What is fusion?

Summary

What is *not* fusion?