



Language
Technologies
Institute

Carnegie
Mellon
University

Multimodal Machine Learning

Lecture 6.2: Many Modalities

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis.
Major revision co-developed with Paul Liang. Co-lecturers
included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Administrative Stuff

Upcoming Course Assignments

Second project assignment (due Sunday 10/4 at 8pm ET)

- Look at your data, before modeling it! 😊

Week 7 Reading assignment (due Friday 10/9 at 8pm ET)

- Two papers (with a focus on visual-language and omni models)

Fall break (Oct 12-16): No classes

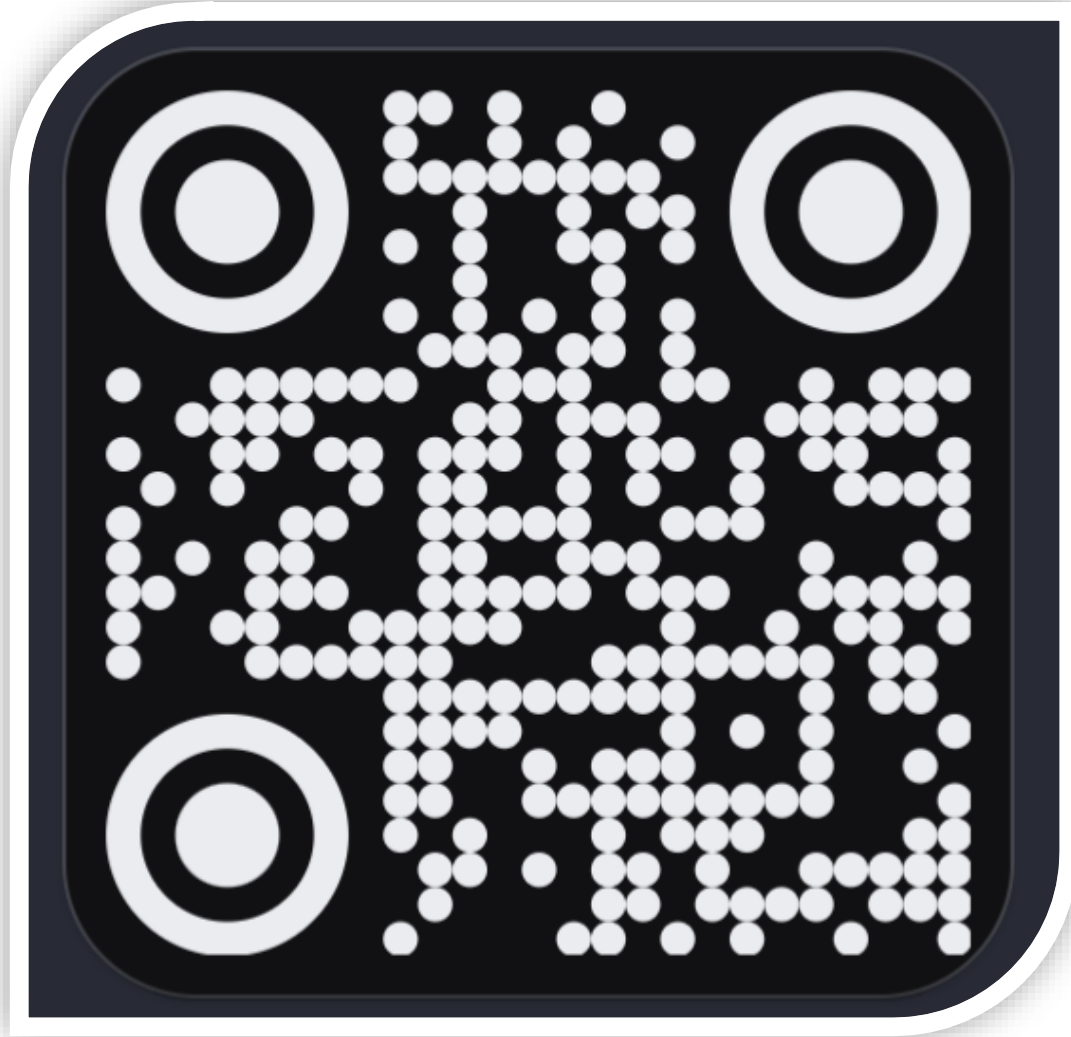
Midterm project assignment (due Sunday 10/25 at 8pm ET)

- For each N-1 research idea, implement a first (baseline) model
- In-class presentations: Tuesday 10/27 and Thursday 10/29

Available Computing Resources

- AWS \$150 credit
 - Submitted IDs to AWS, waiting for response
 - WARNING: Don't go over the limit!
- LiteLLM \$100 credit
 - CMU branded it as AI Gateway
 - Instructions shared via Piazza
- Pittsburgh Super Computers (PSC)
 - Application through NSF's ACCESS program
 - Instructions sent via Piazza. You are the PI of this computing grant.
- Google Colab, AWS SageMaker Studio Lab, ...

Attendance using Poll Everywhere



- 1 Scan QR Code
or
<https://pe.app/mmm1>
- 2 Enter your AndrewID
(as your “name tag”)
- 3 Click “Share location”

✗ *Do not try to sign-in*



Language
Technologies
Institute

Carnegie
Mellon
University

Multimodal Machine Learning

Lecture 6.2: Many Modalities

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis.
Major revision co-developed with Paul Liang. Co-lecturers
included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Objectives of today's class

- 4th VLM Generation: Towards Native VLMs
 - Native-resolution visual encoding (e.g., Qwen2.5-VL)
 - Rotary Position Embedding (RoPE) and Multidimensional RoPE (M-RoPE)
 - Connecting visual tokens at multiple LLM layers (e.g., DeepStack, Qwen3-VL)
 - Natively-pretrained visual-language models (e.g., Qwen 3.5)
- Many Modalities — three challenges
 - Modality alignment: binding through a hub (e.g., ImageBind, LanguageBind)
 - Quantifying heterogeneity: heterogeneity-aware fusion (e.g., HighMMT)
 - Many modalities and many tasks: transfer across partially observable modalities
- Modalities Beyond Sequences
 - Graph neural networks

4th VLM Generation : Towards Native VLMs

Native-Resolution Visual-Language Model (e.g., Qwen2.5-VL)

Three main parts of the model architecture are still the same:

- **Vision encoder** (ViT): Trained to handle images and videos natively
- **Connector**: 2-layer MLP
- **Language model**: From Qwen 2.5 LLM, with continual pretraining

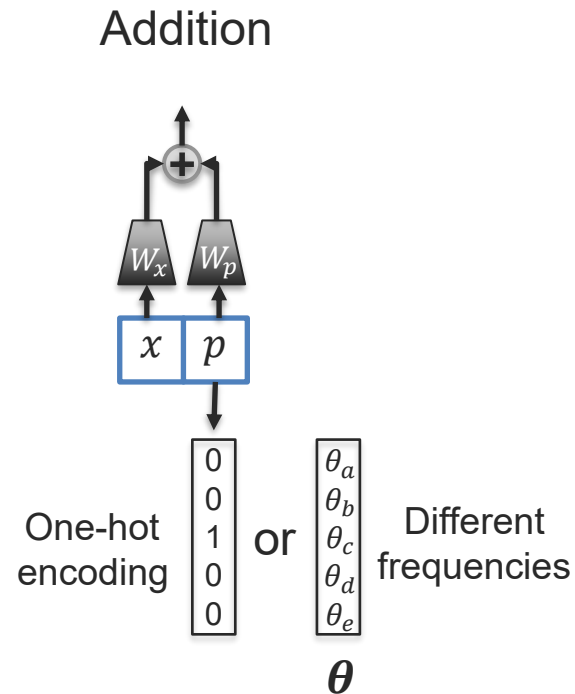
With expanded training procedure:

- Image captions + grounding coordinates, documents/HTML, OCR, videos, GUI logs
 - 1.5T-2.1T tokens, interleaved
- + a post-training phase, for better human-AI alignment (*more later*)

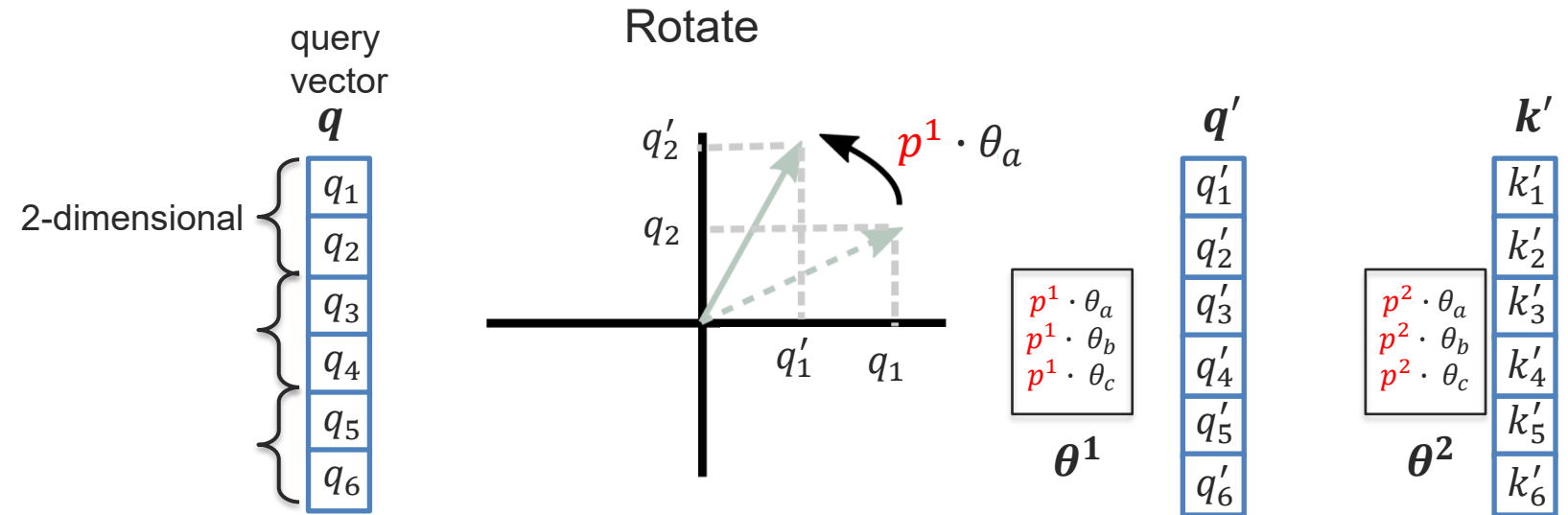
And improved visual encoding...

Preliminary Idea: Rotary Position Embedding (RoPE)

Previous approach:



New approach:



$$\text{For 2D rotations: } R(\theta^1)^T R(\theta^2) = R(\theta^2 - \theta^1)$$

So $q' \cdot k'$ only keeps the relative positions $(p^2 - p^1)$

and each token can be encoded with an **absolute** position index p

Multidimensional-RoPE (M-RoPE)

Three dimensions (aka indices): Temporal (t), Height (h), Width (w)

Token type	t	h	w
Text	sequence index	same	same
Image	constant (one frame)	patch row	patch column
Video	frame index	patch row	patch column

When all three components are equal, M-RoPE collapses exactly to standard 1D RoPE

Qwen2.5-VL t t t t t t h h h h h h w w w w w w

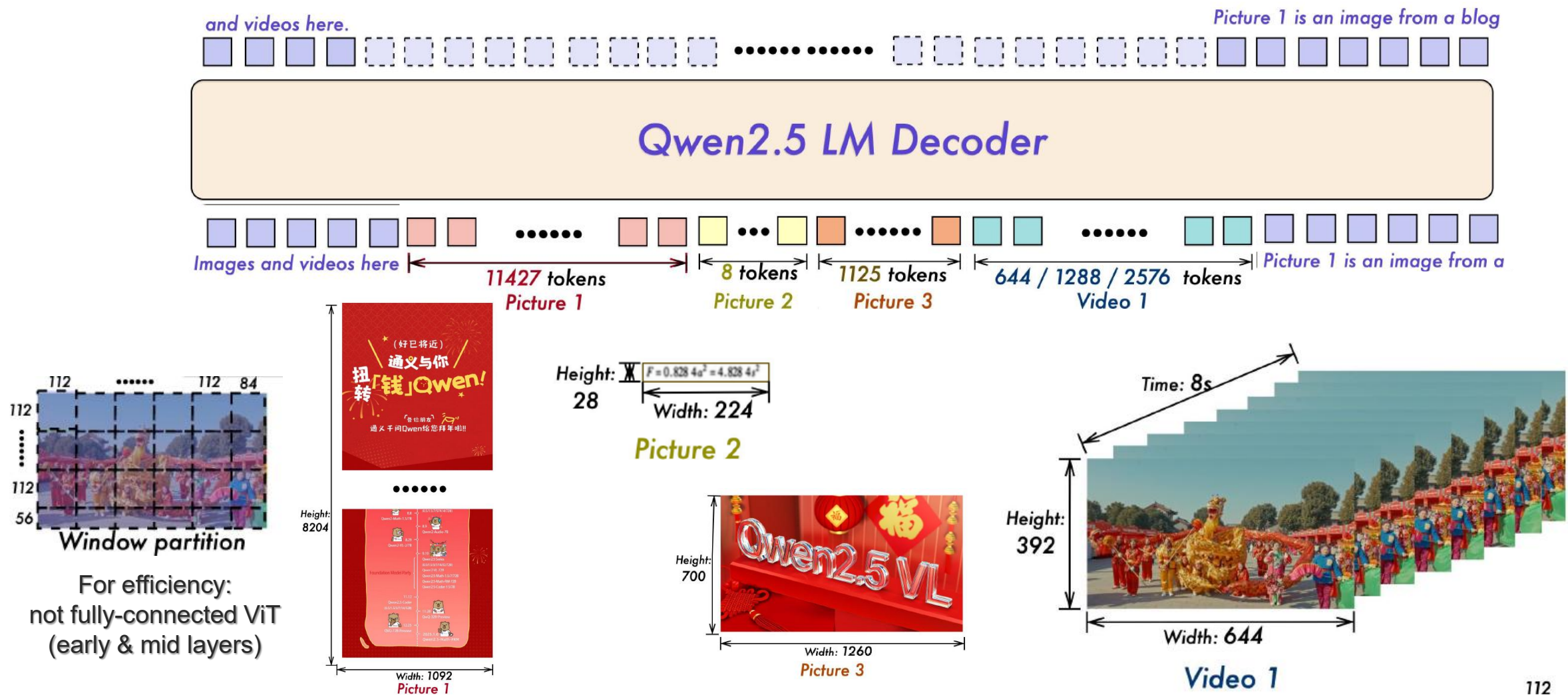
Problem: time gets only the fastest clocks. It became an issue with long videos (loses order info)

Qwen3-VL t h w t h w t h w t h w t h w t h w

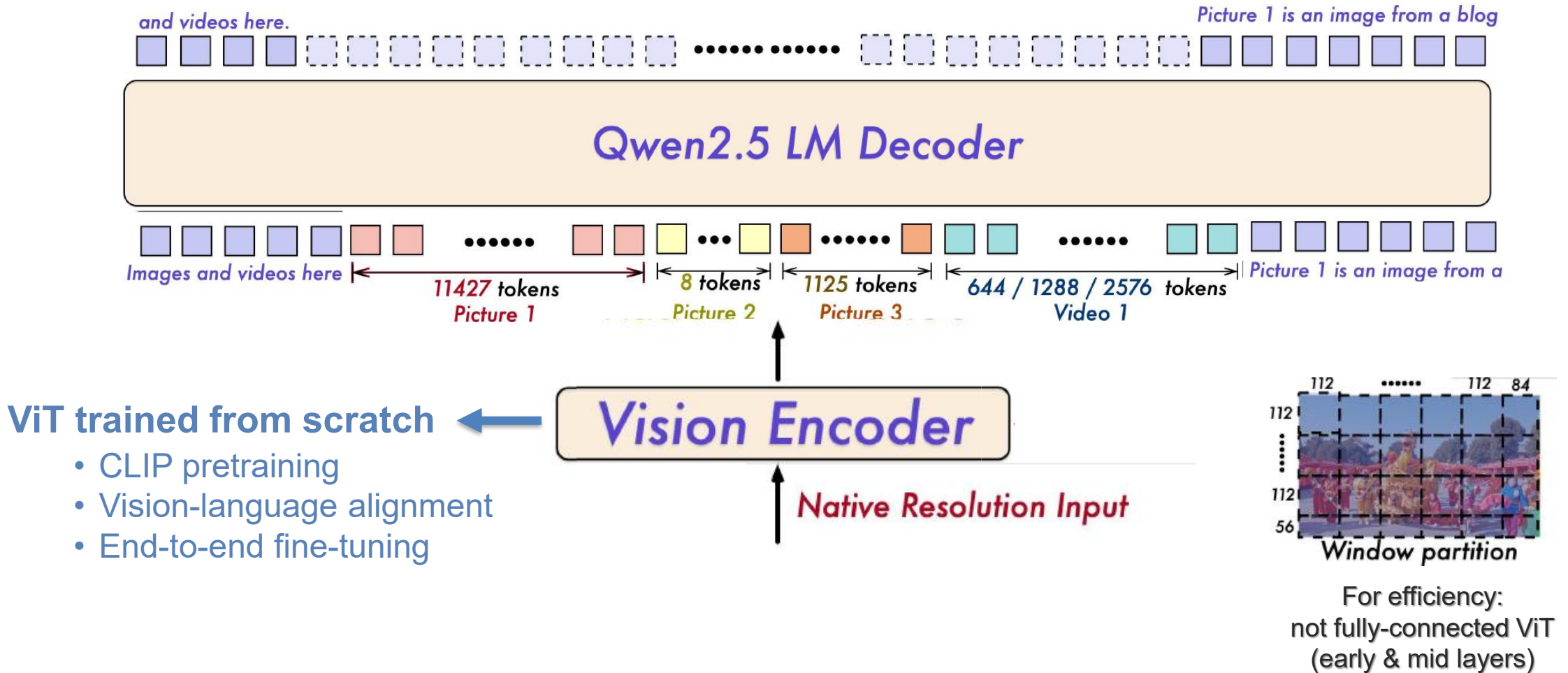
High frequencies Low frequencies

[Qwen2.5-VL Technical Report 2025]

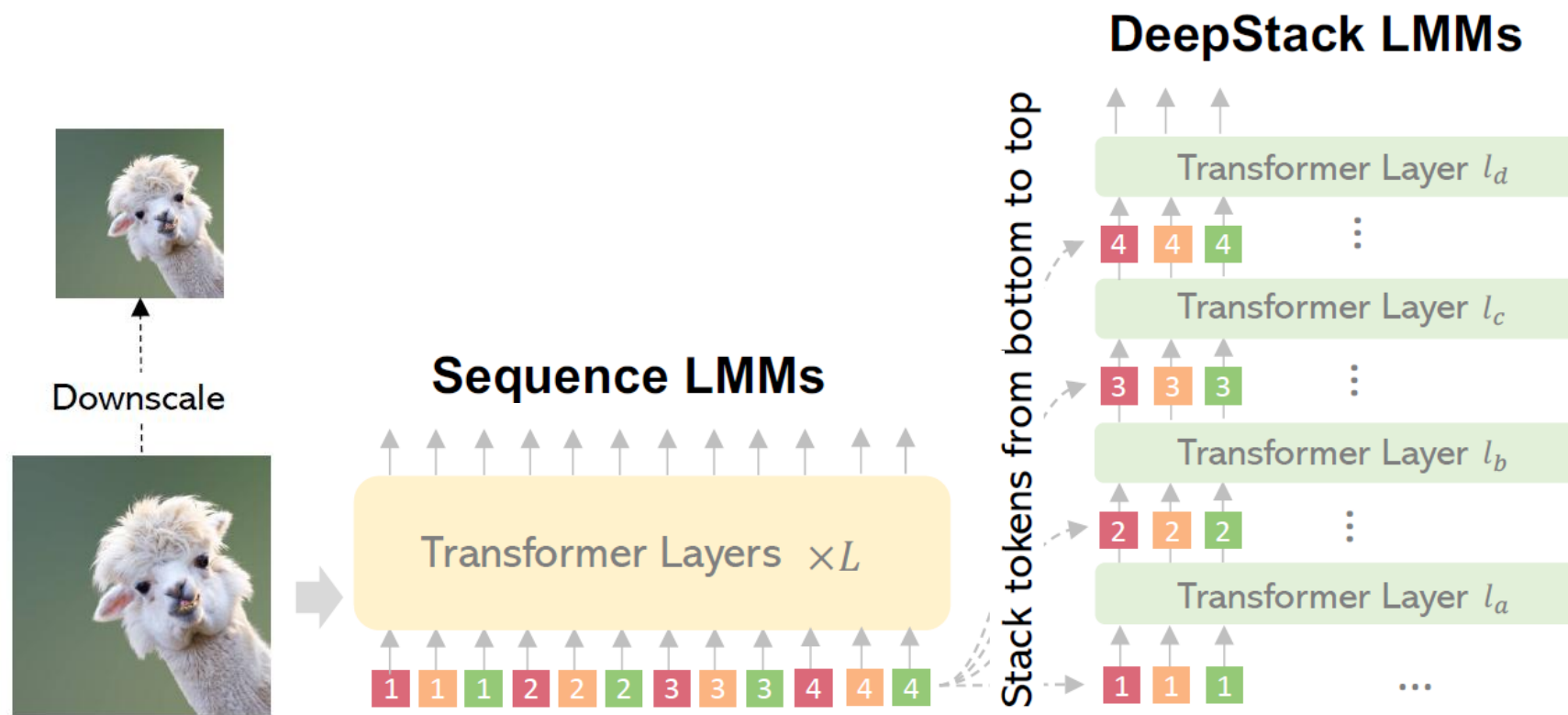
Native-Resolution Visual-Language Model (e.g., Qwen2.5-VL)



Native-Resolution Visual-Language Model (e.g., Qwen2.5-VL)

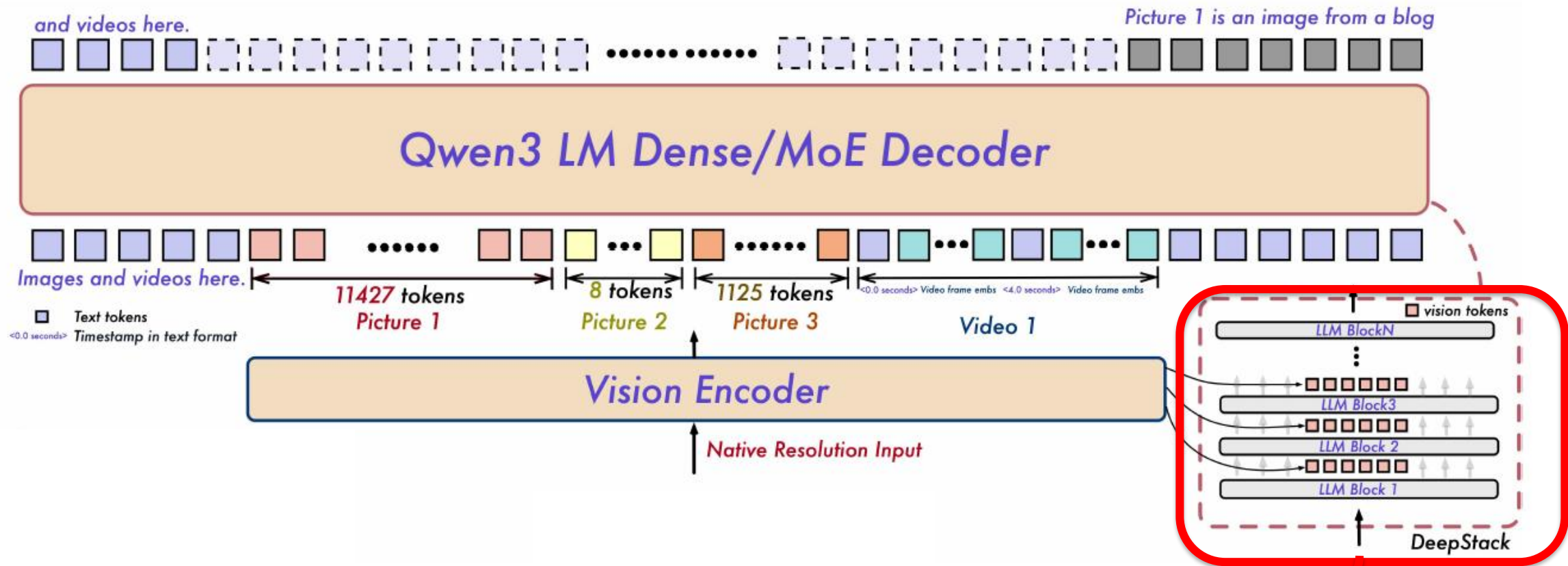


Connecting Visual Tokens at Different LLM Layers (e.g., DeepStack)



Idea: Visual tokens from different resolutions are added at different LLM layears

Connecting ViT Intermediate Outputs to LLM (e.g., Qwen3-VL)



Intermediate ViT outputs are inserted at intermediate LLM layers

- 3 intermediate layers (for both ViT and LLM): 8, 16 and 24

Can We Pretrain Natively with Language and Vision?

What could be a main challenge with jointly training them?

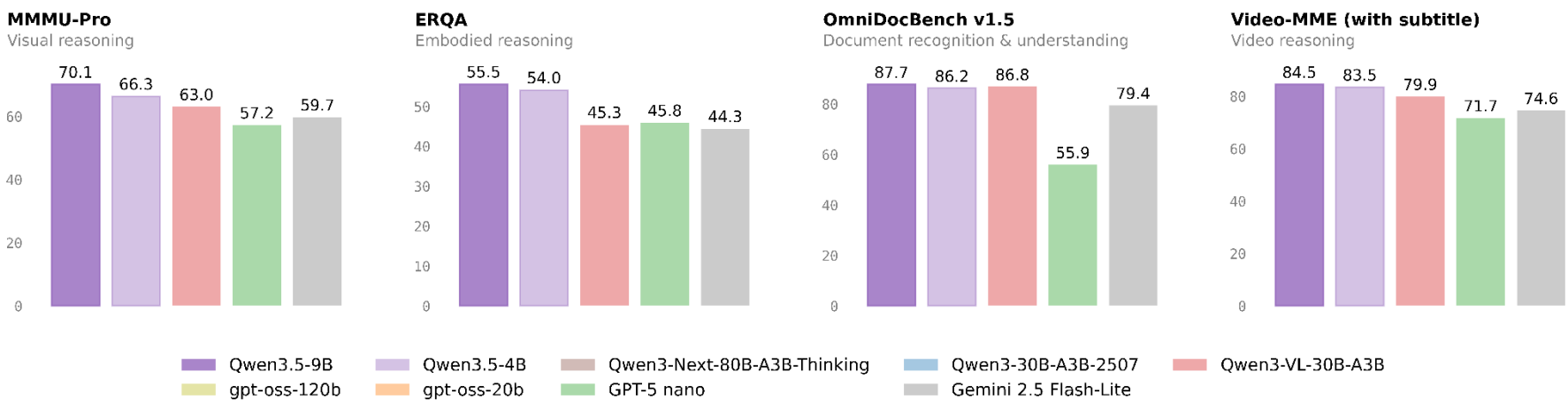
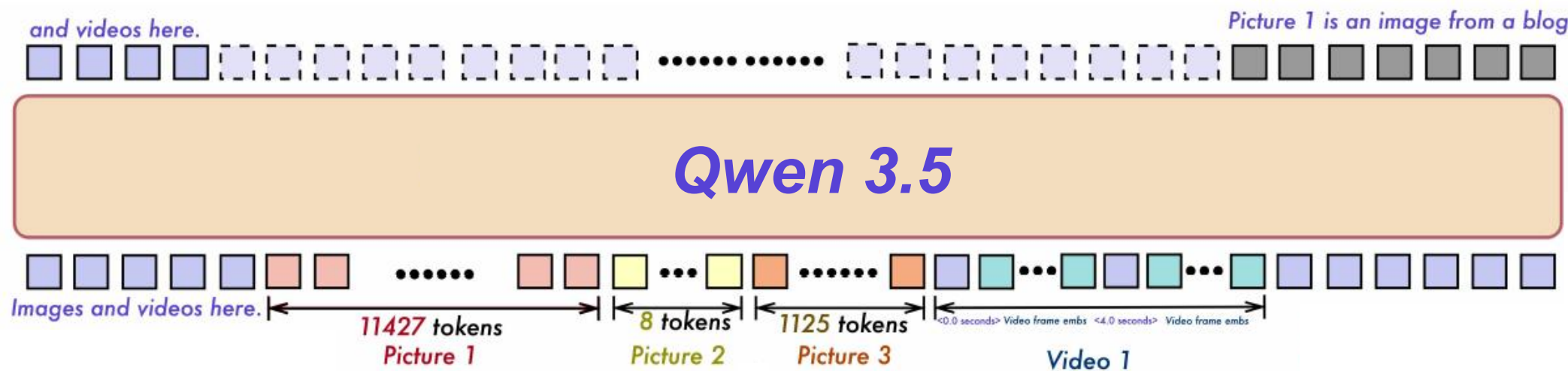
- Text tokens: low entropy, predictable distributions
- Image tokens: high entropy, far more varied

Sharing weights, each modality competes by growing its activation norms

Solution: QK-Norm

- Normalizes queries and keys at the layer level, before the attention dot product. This reduces training failures.

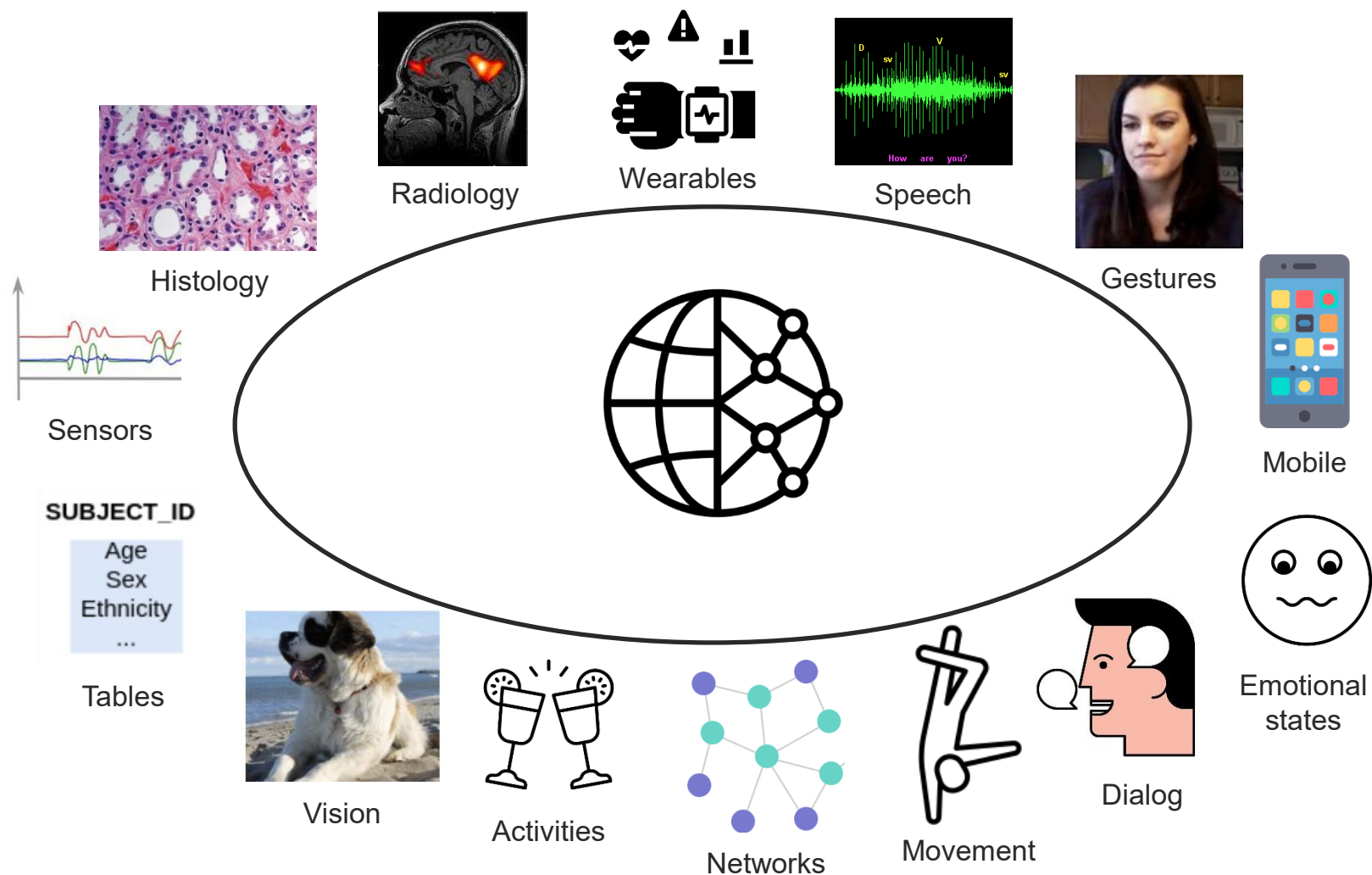
Natively-Pretrained Visual-Language Models (e.g., Qwen 3.5)



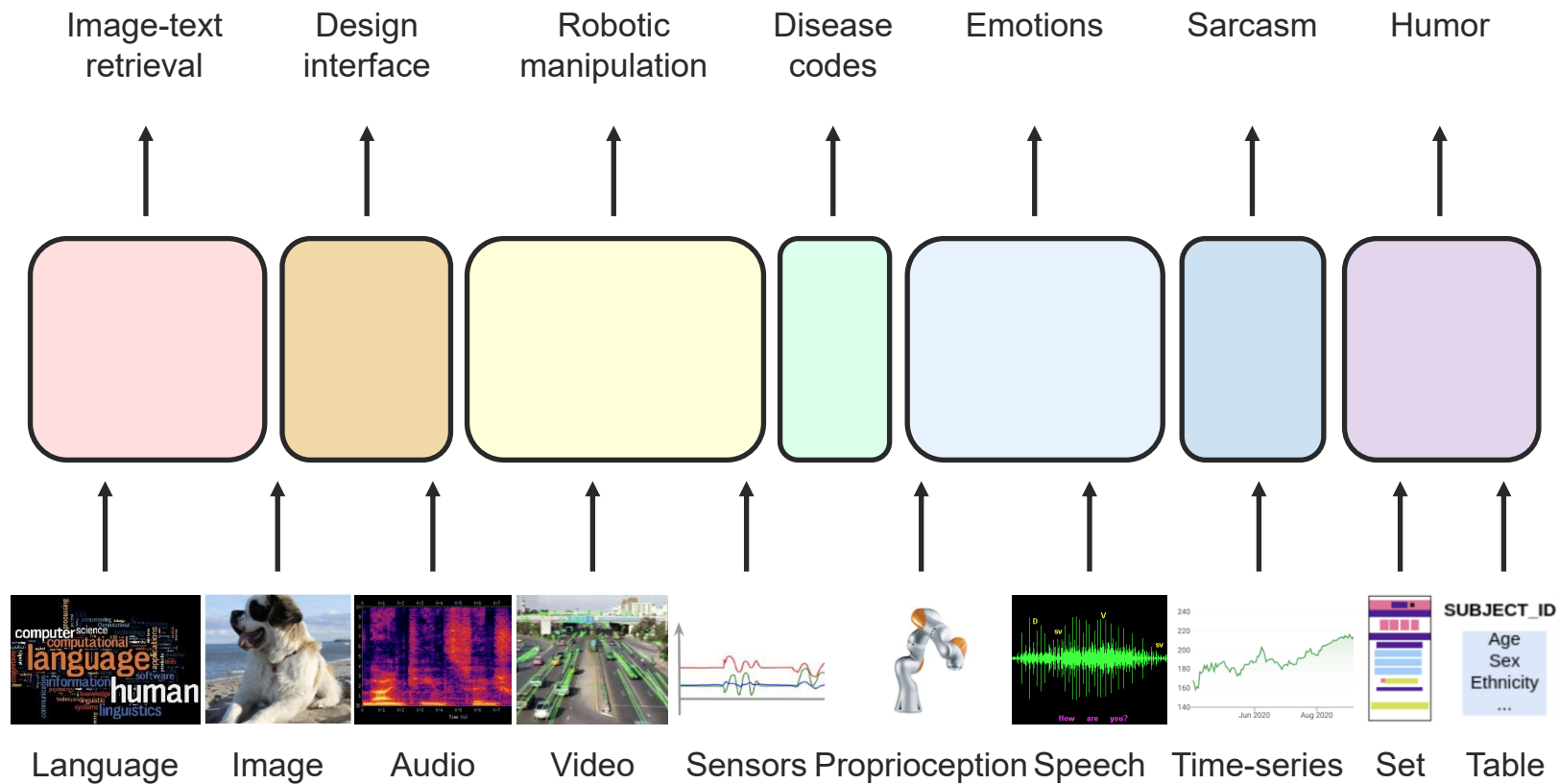
Qwen3-VL is still competitive!

Many Modalities

Multiple Modalities...

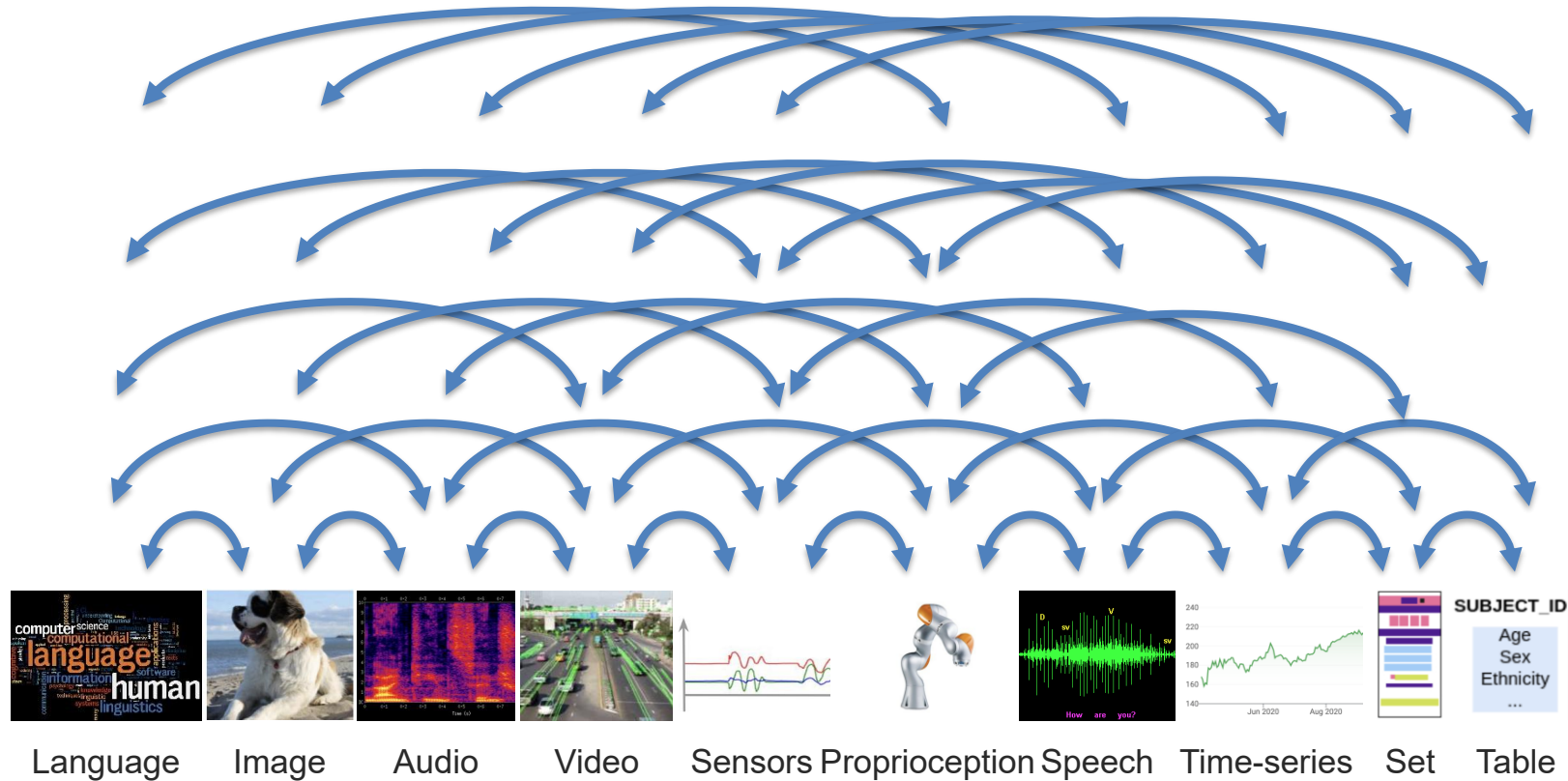


... and Multiple Tasks!

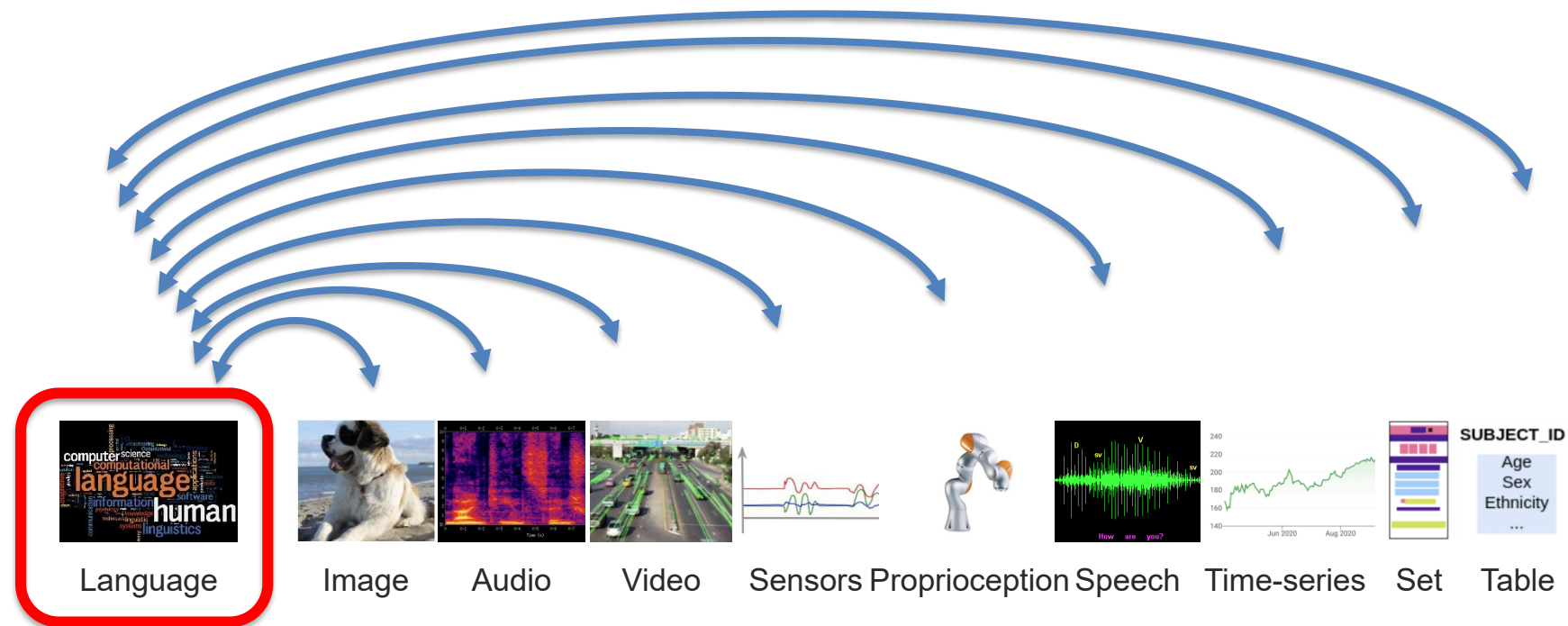


A First Challenge: Modality Alignment

⋮ N^2 paired alignments!!!

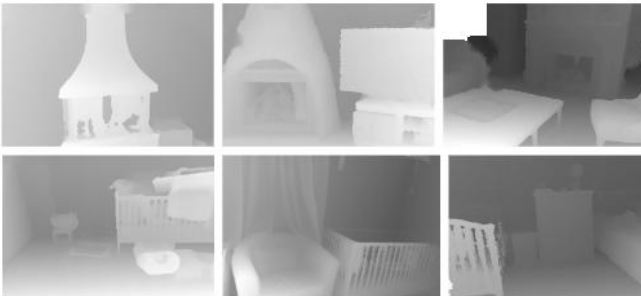


Solution: Pick a Modality at the Main Hub



What's the intuition for ImageBind?

1) Cross-Modal Retrieval

Audio	Images & Videos	Depth	Text
 Crackle of a Fire			"A fire crackles while a pan of food is frying on the fire." "Fire is crackling then wind starts blowing." "Firewood crackles then music..."
 Baby Cooing			"A baby is crying while a toddler is laughing." "A baby is laughing while an adult is laughing." "A baby laughs and something..."

2) Embedding-Space Arithmetic



3) Audio to Image Generation

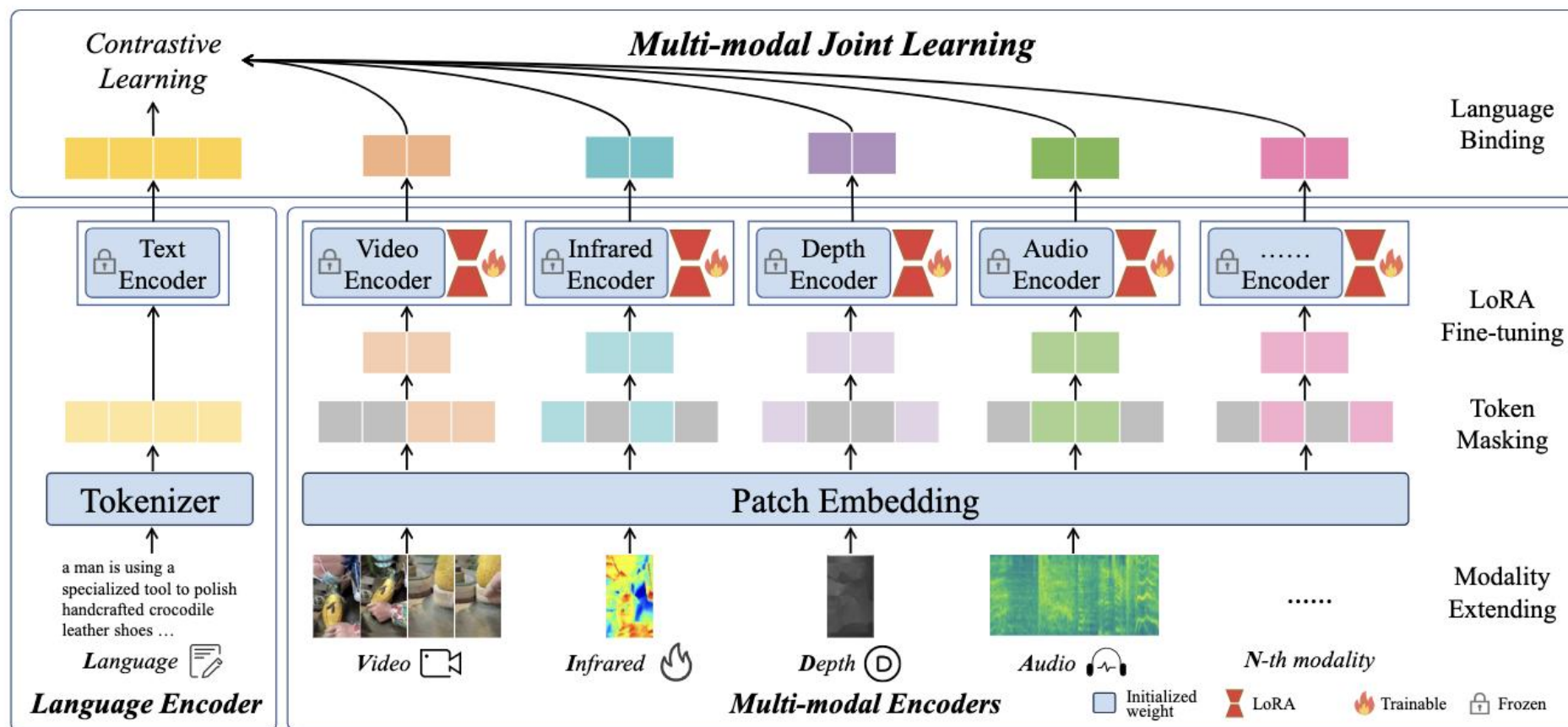


- Freeze the image encoder, and train each other modality's encoder to align to it contrastively.
- Audio and text are aligned with each other despite never seeing a single (audio, text) pair.

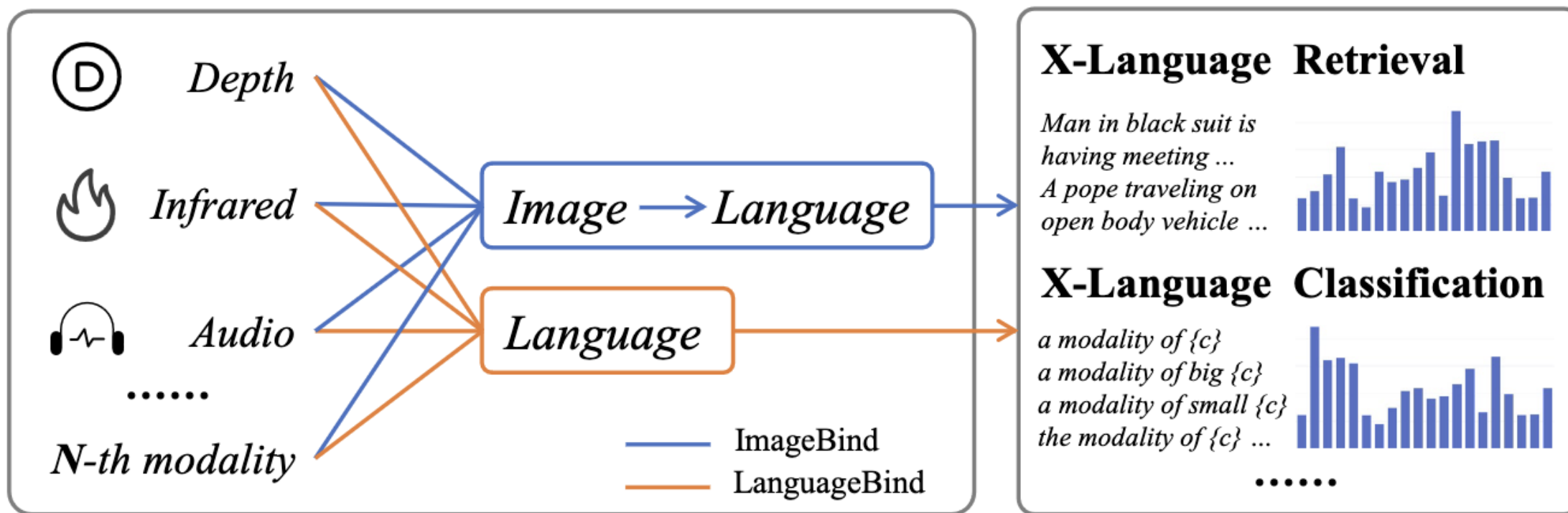
<https://imagebind.metademolab.com/demo>

<https://facebookresearch.github.io/ImageBind>

LanguageBind



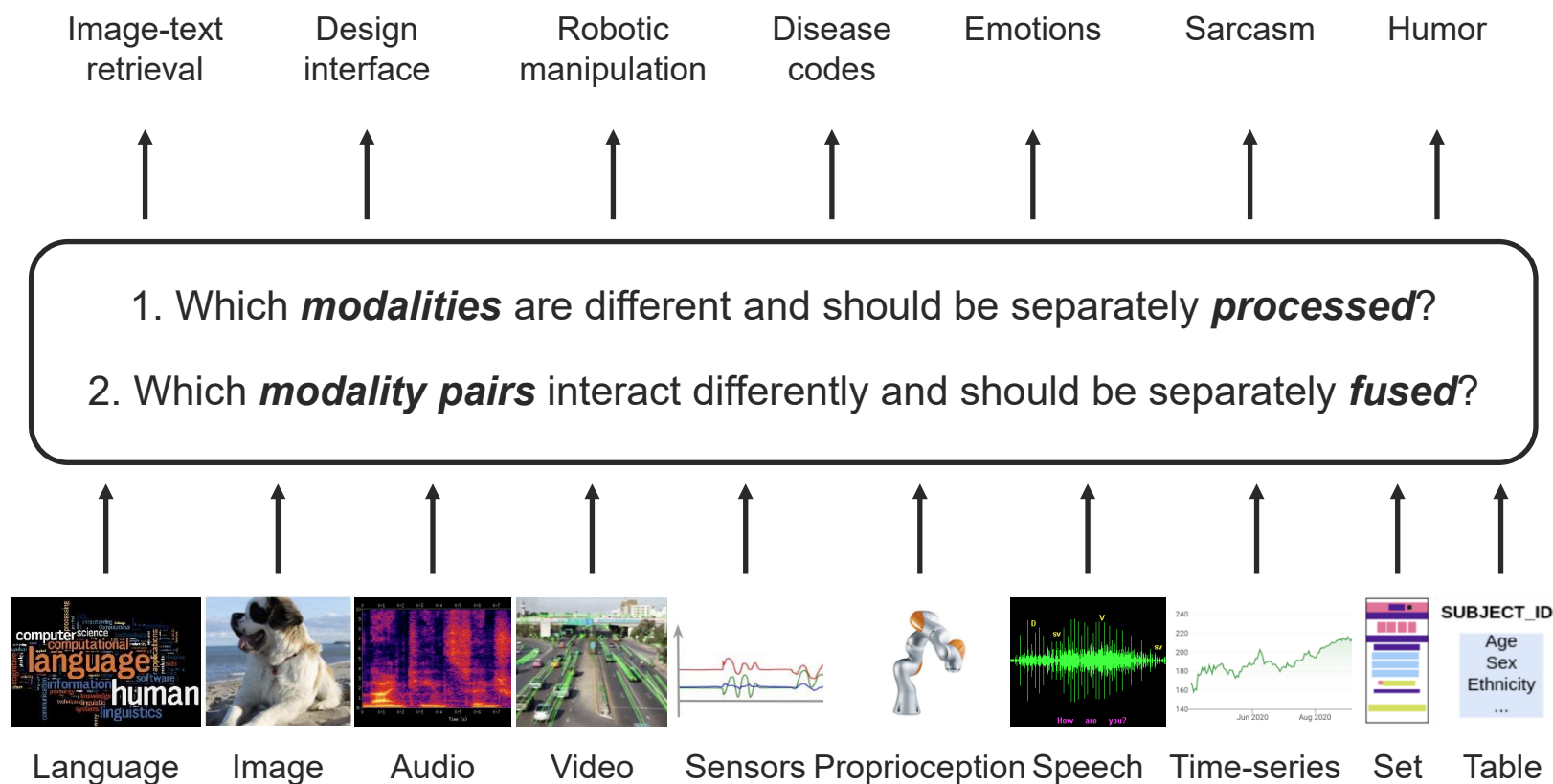
ImageBind vs LanguageBind



“taking the language as the bind across different modalities because the language modality is well-explored and contains rich semantics”

<https://arxiv.org/abs/2310.01852>

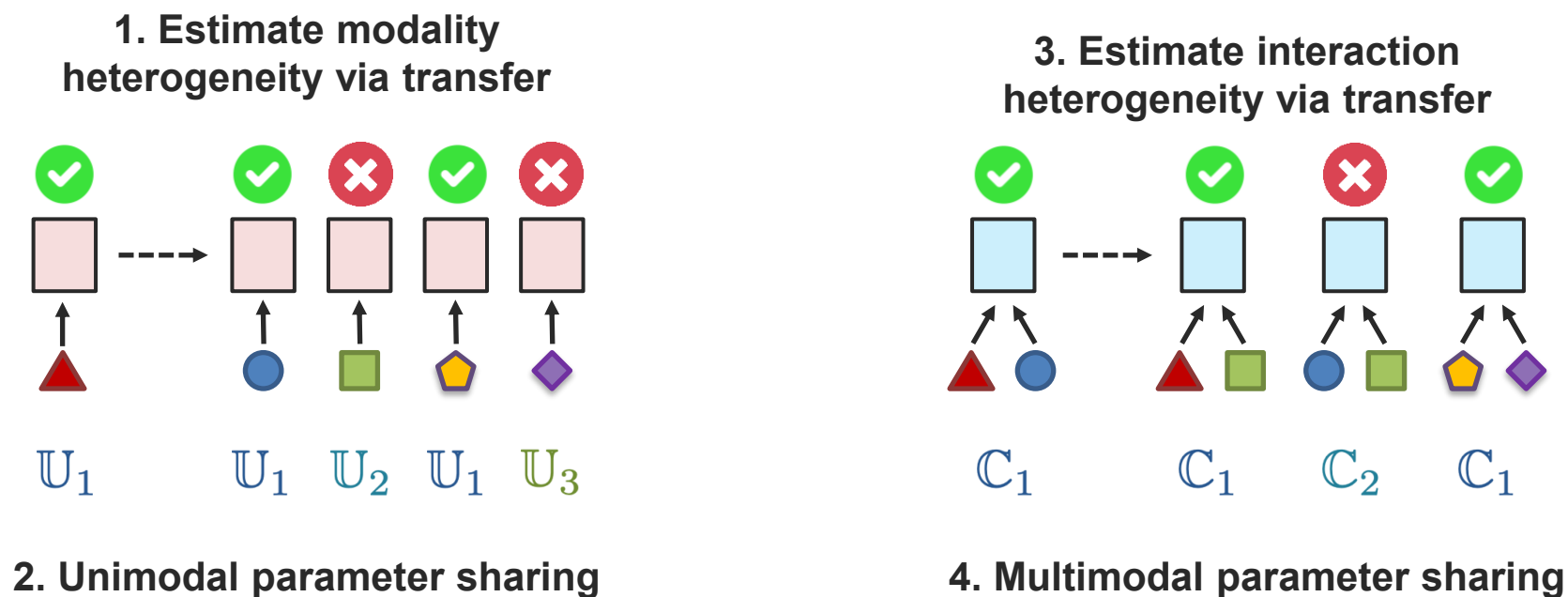
Second Challenge: Quantifying Heterogeneity



[Liang et al., HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Learning. TMLR 2022]

Quantifying Heterogeneity

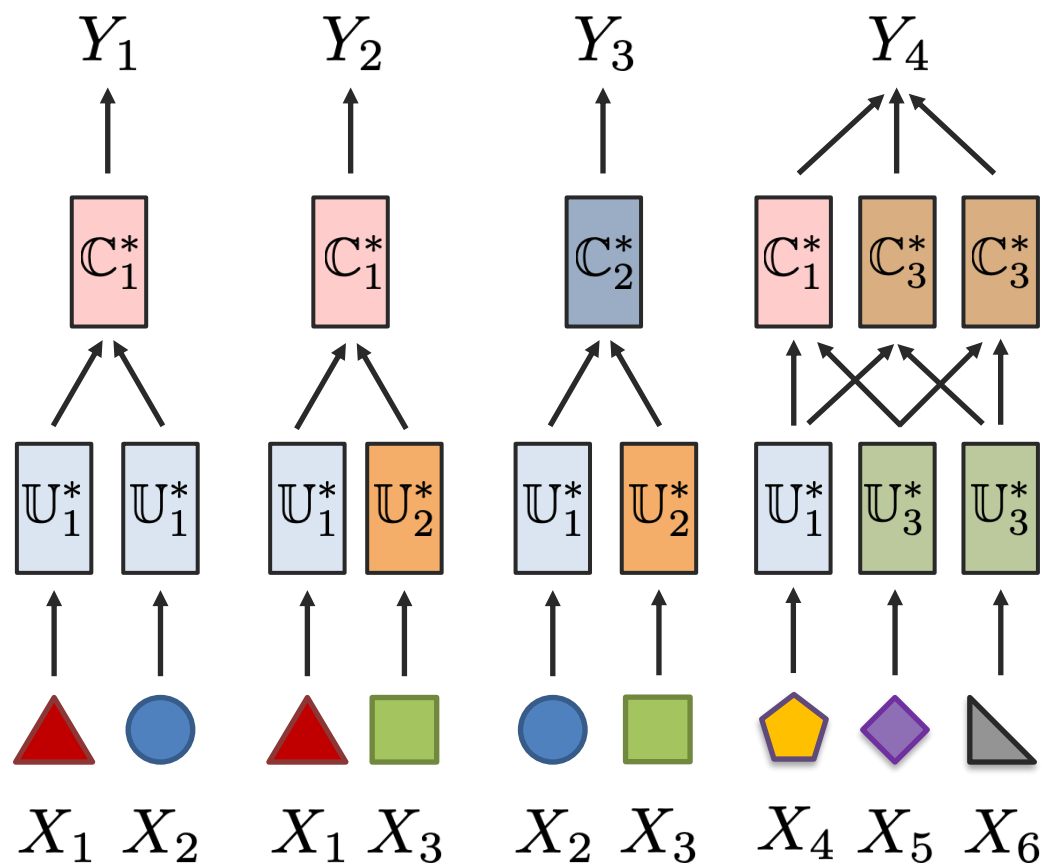
Model transfer perspective



In practice, efficient by pre-trained models and few-shot transfer

Heterogeneity-aware Fusion

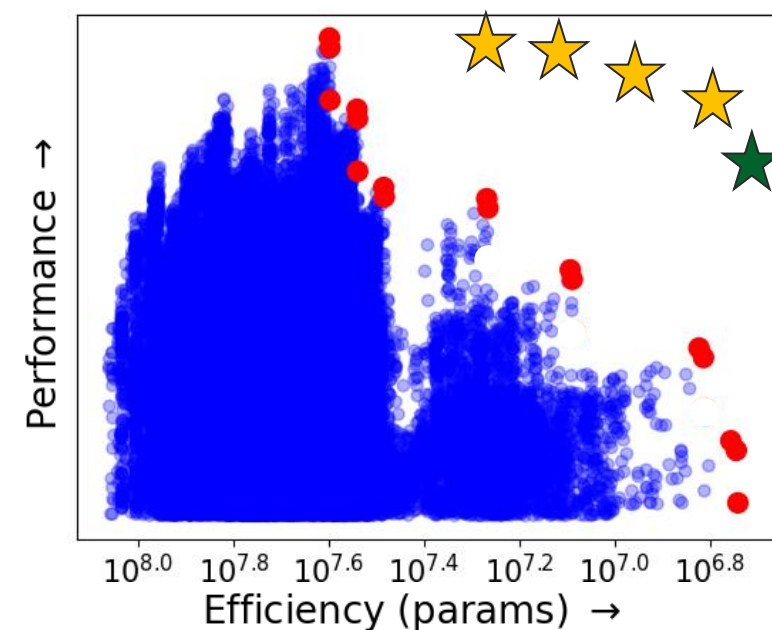
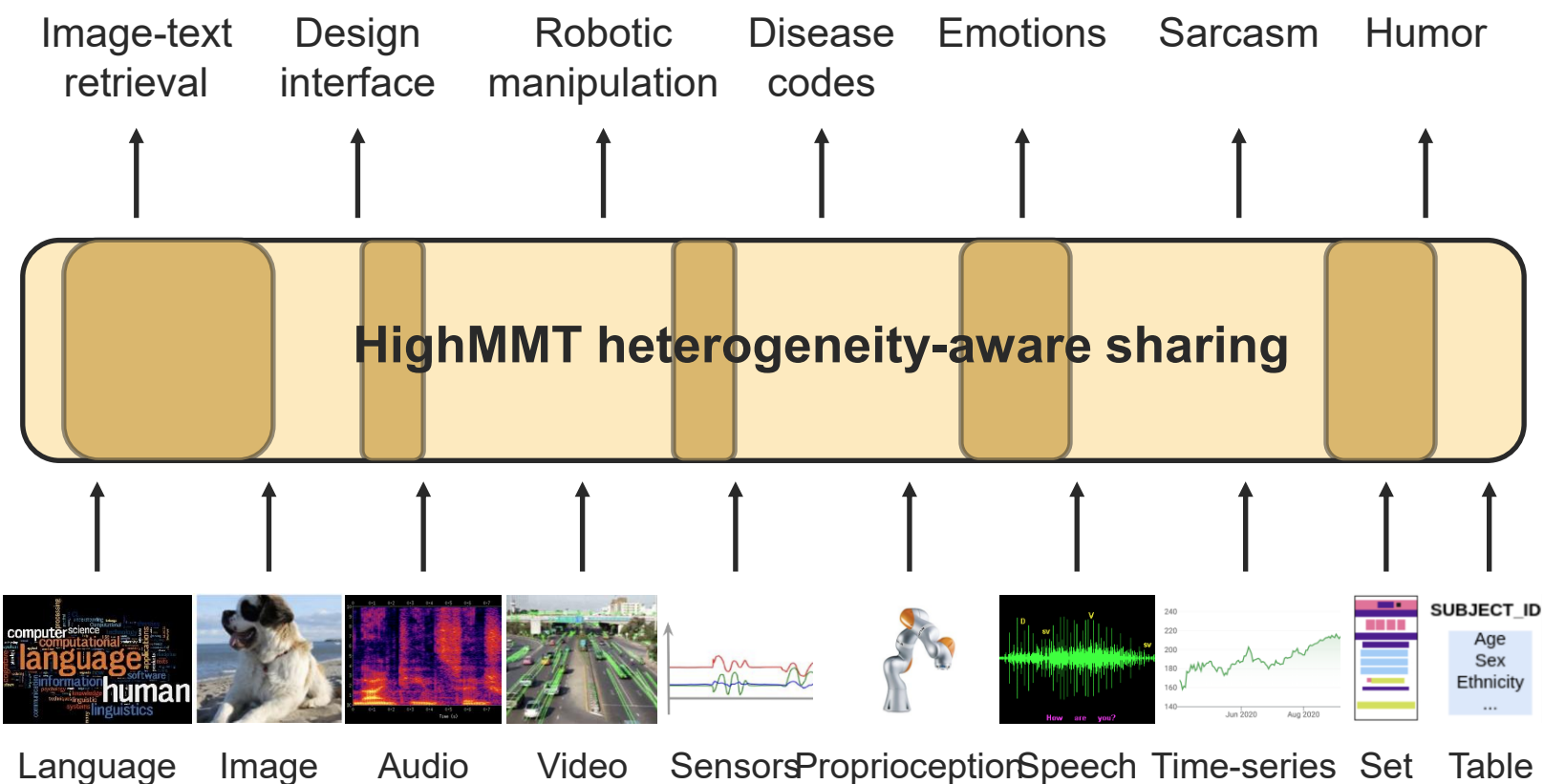
Fusion of many modalities based on how similar or different they are



[Liang et al., HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Learning. TMLR 2022]

High-Modality Multimodal Transformer

HighMMT heterogeneity-aware: estimate heterogeneity to determine parameter sharing

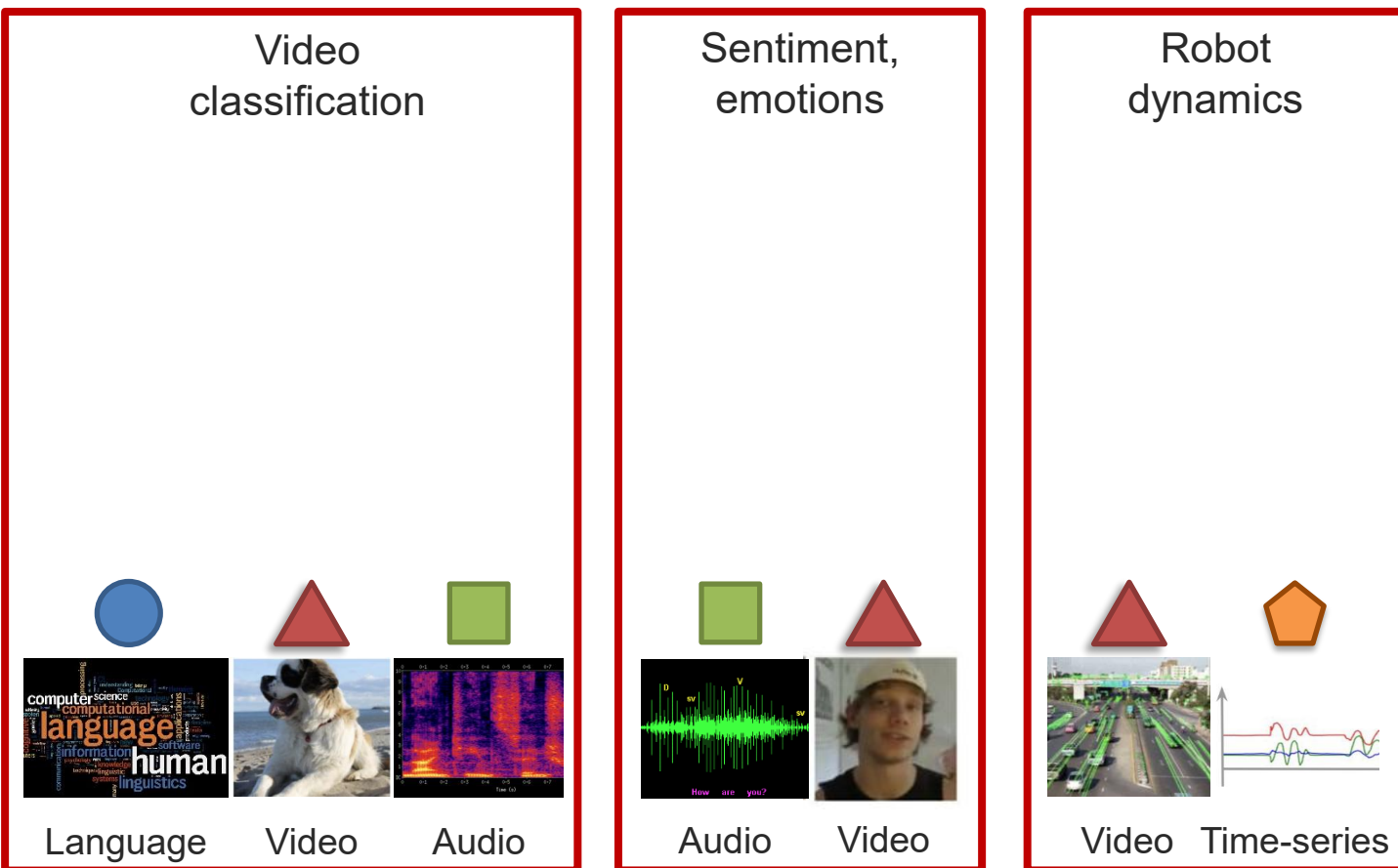


- All model combinations (>10,000)
- Pareto front
- HighMMT multitask
- HighMMT heterogeneity-aware

[Liang et al., HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Learning. TMLR 2022]

Third Challenge: Many Modalities and Many Tasks

How can we transfer knowledge across multiple tasks, each over a different subset of modalities?



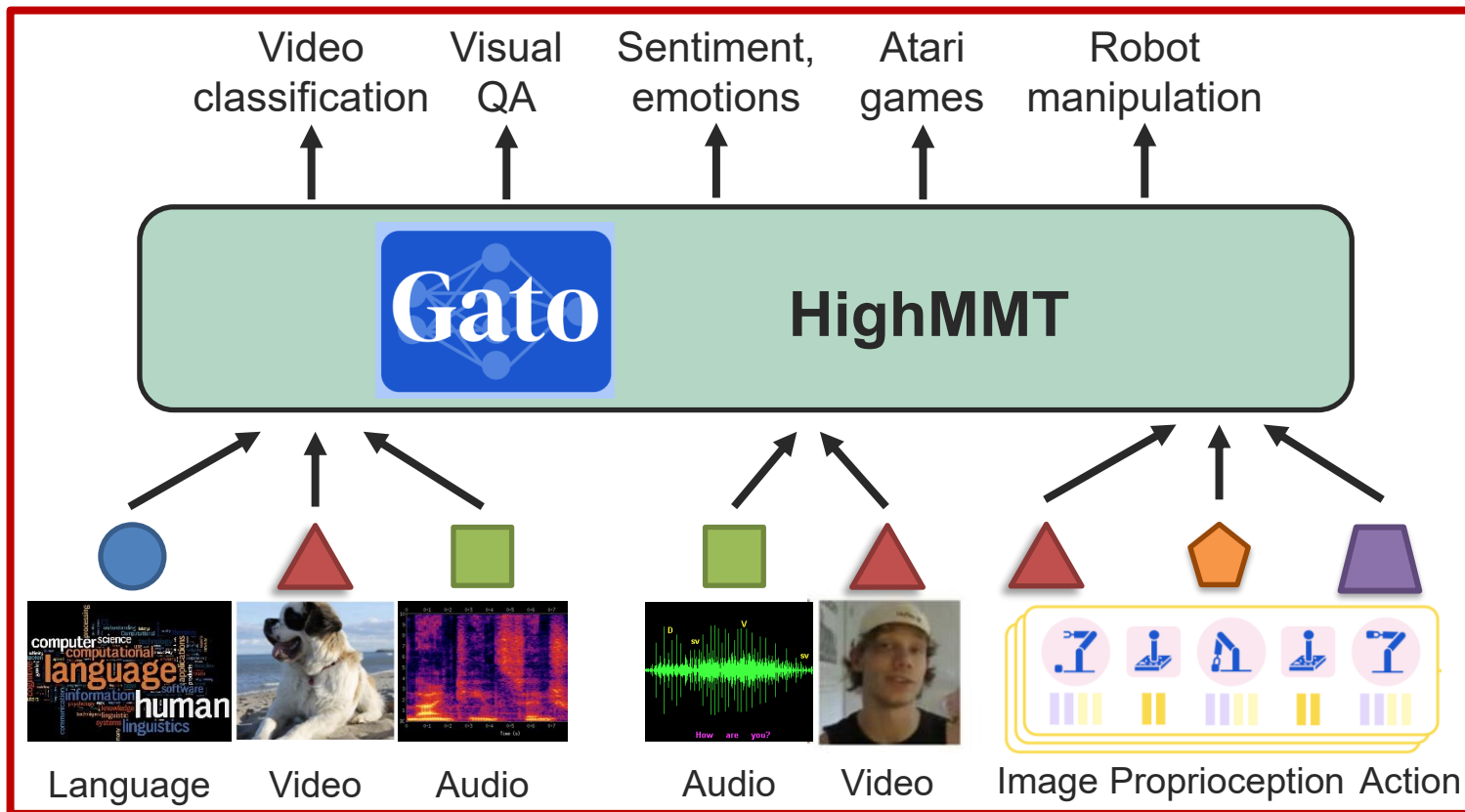
Generalization across modalities and tasks
Important if some tasks are low-resource

[Liang et al., HighMMT: Towards Modality and Task Generalization for High-Modality Representation Learning. TMLR 2022]

High-Modality Multimodal Transformers

Transfer across partially observable modalities

Unified model + parameter sharing + multitask and transfer learning



Non-parallel multitask learning

Task-specific classifiers

Same model architecture!

Shared multimodal model

Same parameters!

Modality-specific embeddings

Standardized input sequence

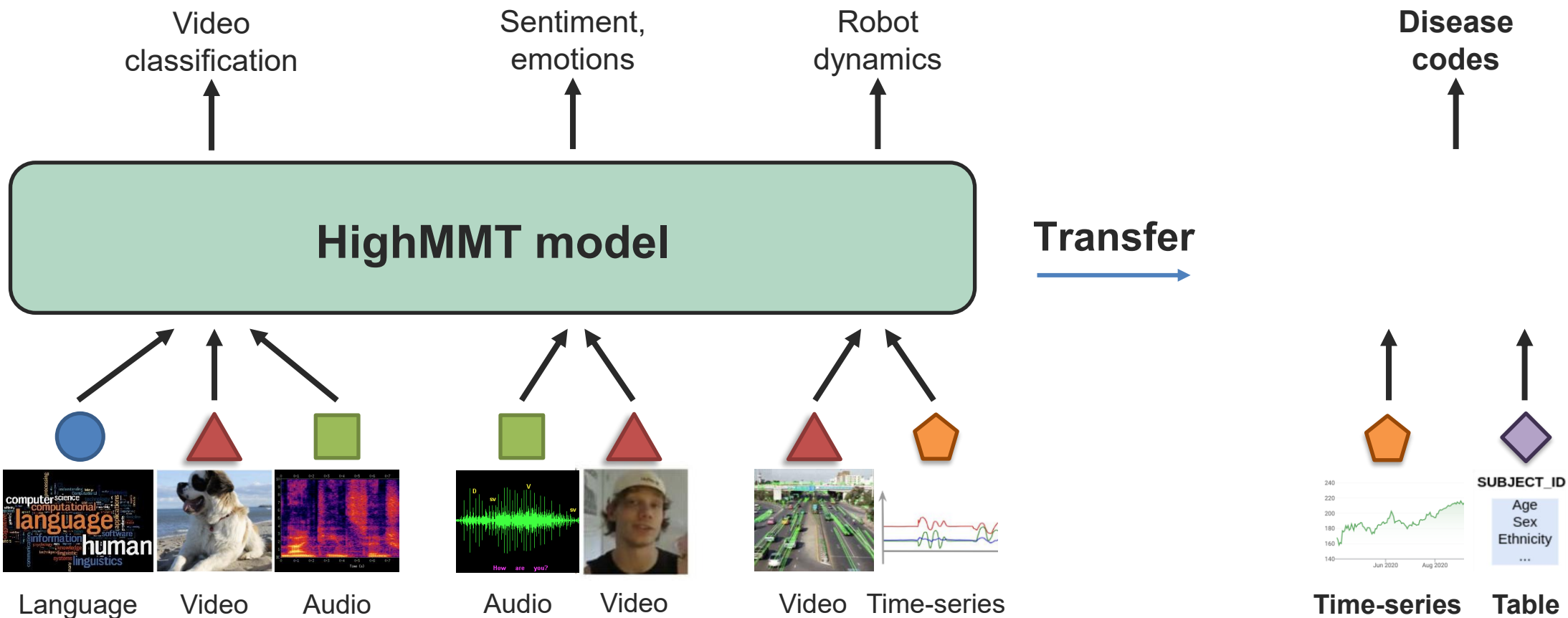
[Reed et al., A Generalist Agent. TMLR 2022]

[Liang et al., HighMMT: Quantifying Modality and Task Heterogeneity for High-Modality Representation Learning. TMLR 2022]

Multitask and Transfer Learning

Transfer across partially observable modalities

HighMMT: unified model + parameter sharing + multitask and transfer learning



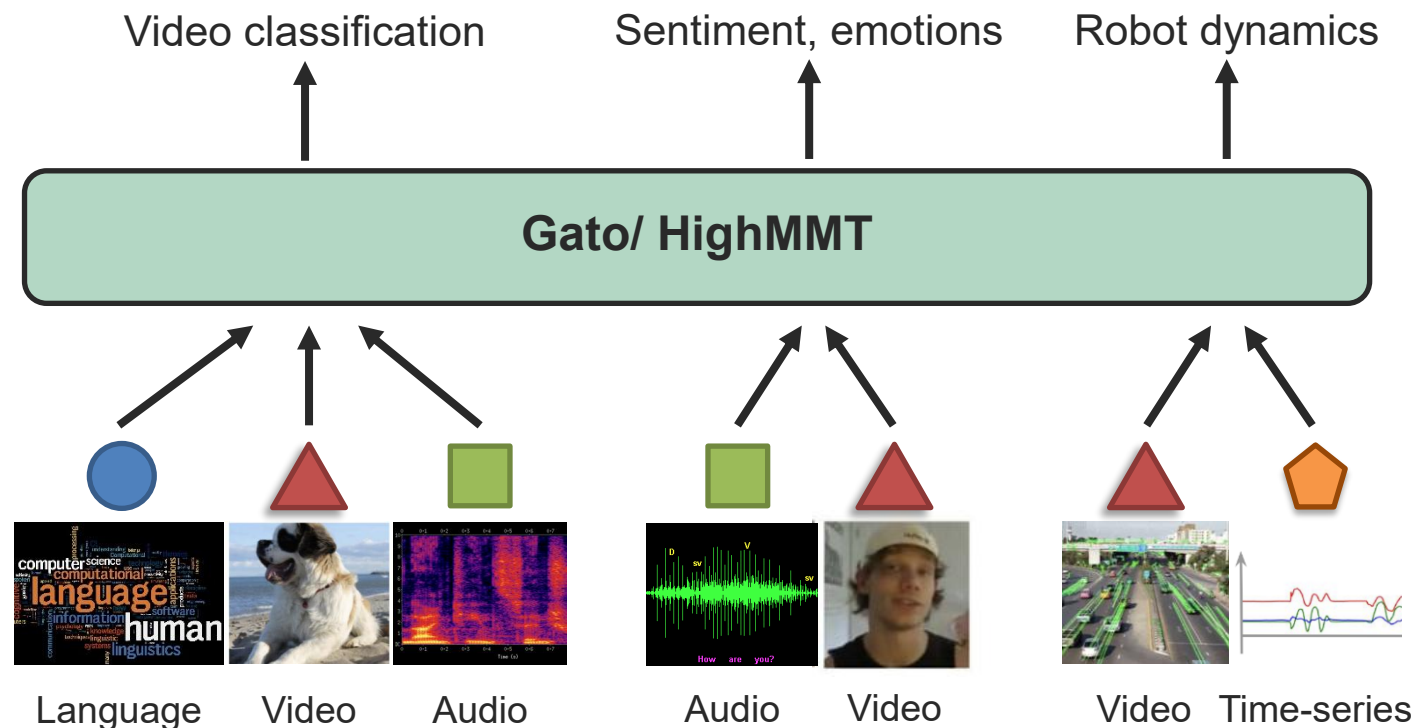
[Liang et al., HighMMT: Towards Modality and Task Generalization for High-Modality Representation Learning. TMLR 2022]

High-Modality Models

Open
challenges

Some implicit assumptions:

- All modalities can be represented as sequences without losing information.
- Dimensions of heterogeneity can be perfectly captured by modality-specific embeddings.
- Cross-modal connections & interactions are shared across modalities and tasks.



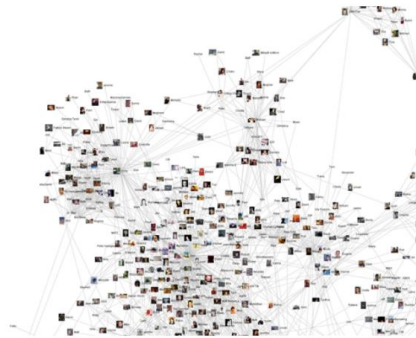
Shared multimodal model?

Modality-specific embeddings?

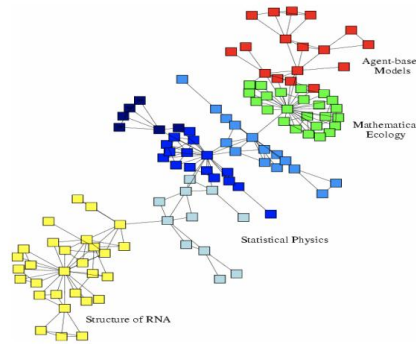
Standardized input sequence?

Modalities Beyond Sequences

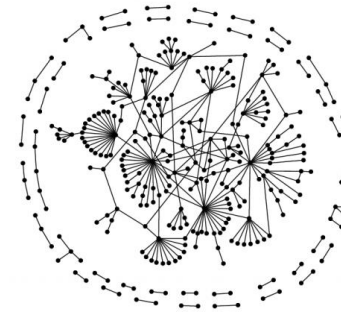
Modalities Naturally Represented as Trees and Graphs



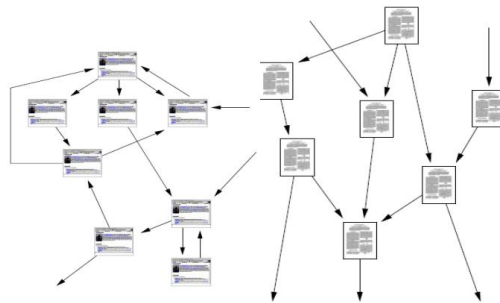
Social networks



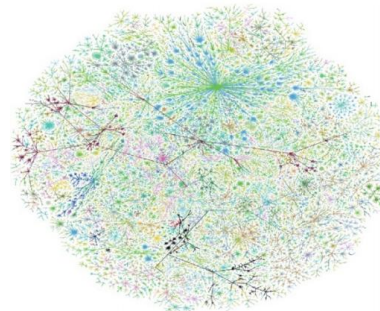
Economic networks



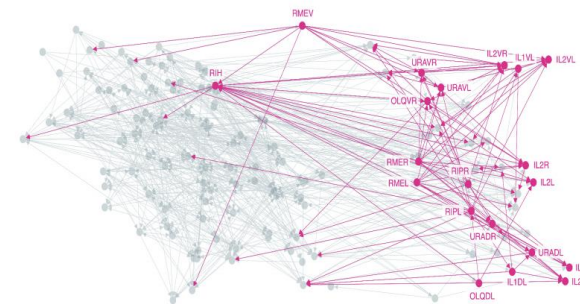
Biomedical networks



Information networks:
Web & citations



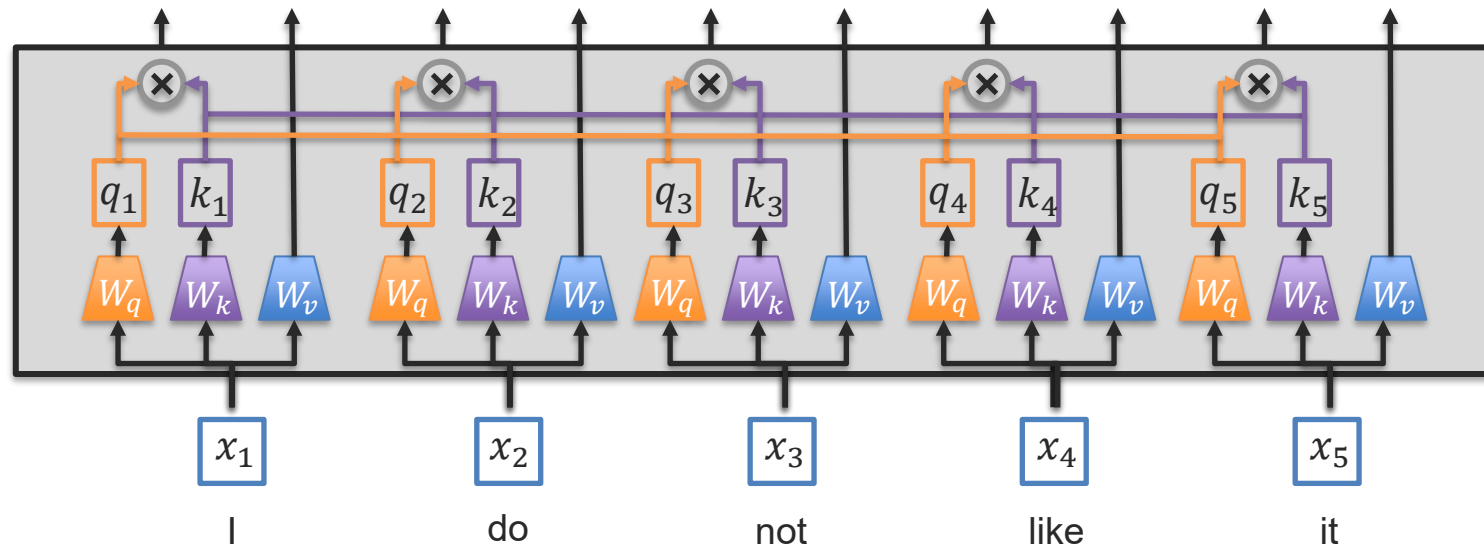
Internet



Networks of neurons

[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

One Option: Fully-Connected Graph - Transformers

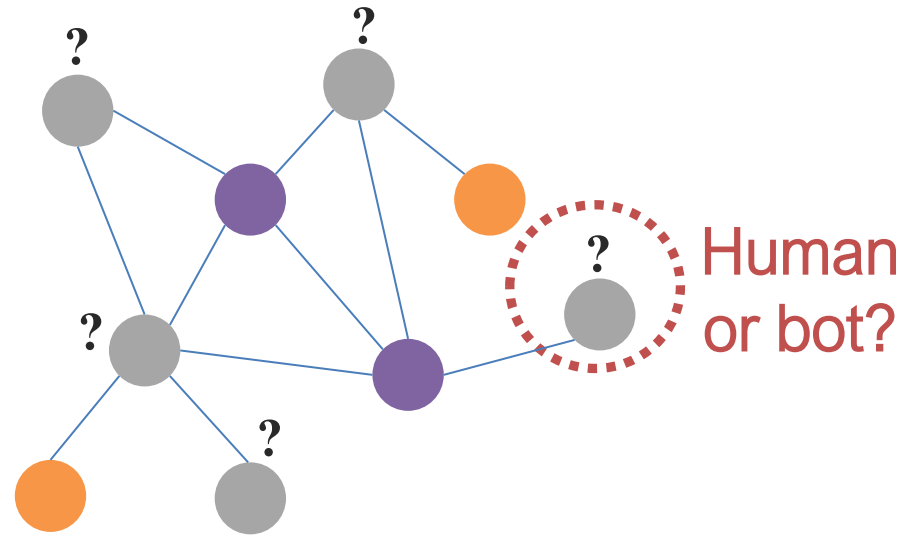


Should everything be connected to everything?

What if we have domain knowledge about connections?

Graphs as a Supervised Task

Goal: Learn from labels associated with a subset of nodes (or with all nodes)

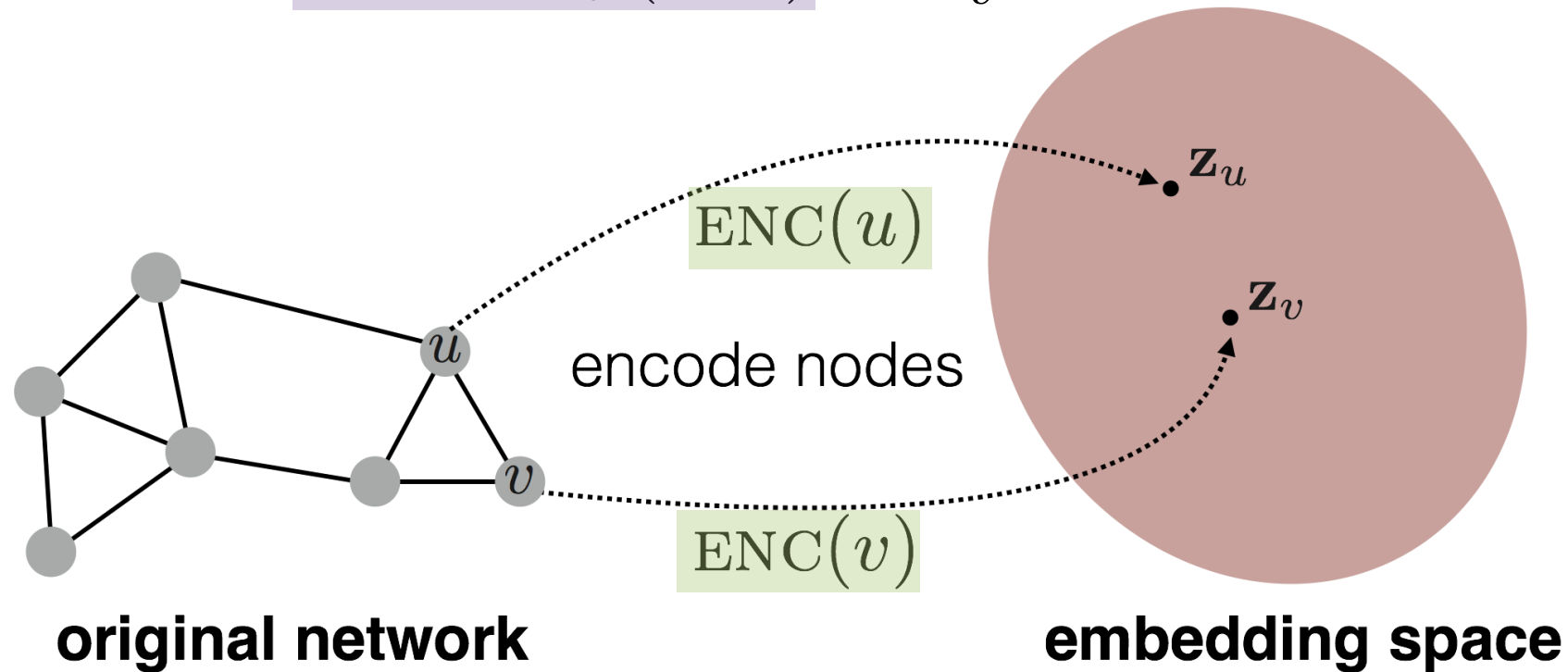


e.g., an online social network

Graphs as a Unsupervised Task

Goal: Learn an embedding space where

$$\text{similarity}(u, v) \approx \mathbf{z}_v^\top \mathbf{z}_u$$



[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Graph Neural Nets

Assume we have a graph $\mathbf{G}$:

$\mathbf{V}$ is the set of vertices

$\mathbf{A}$ is the binary adjacency matrix

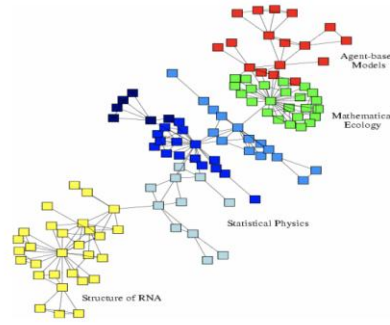
$\mathbf{X}$ is a matrix of node features:

- Categorical attributes, text, image data
e.g. profile information in a social network
- ...

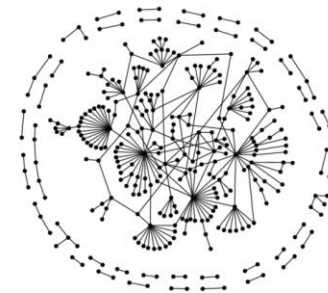
$\mathbf{Y}$ is a vector of node labels (optional)



Social networks



Economic networks

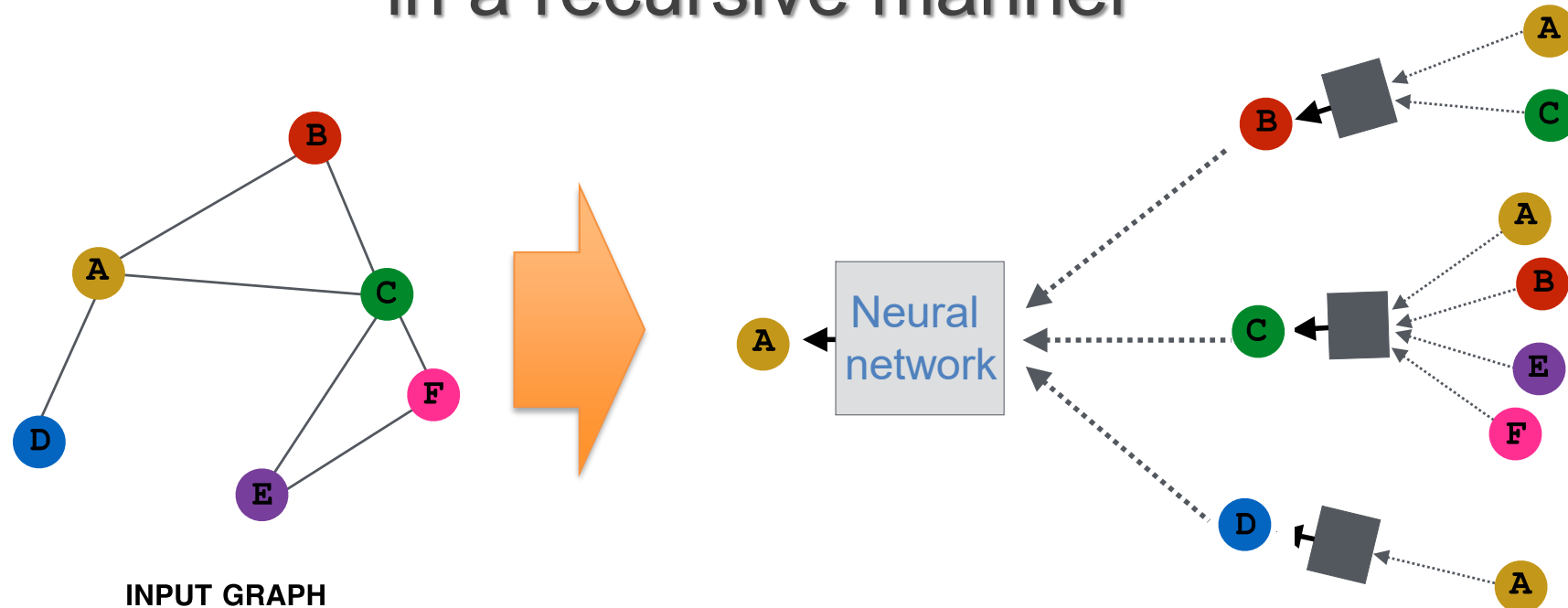


Biomedical networks

[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Graph Neural Nets

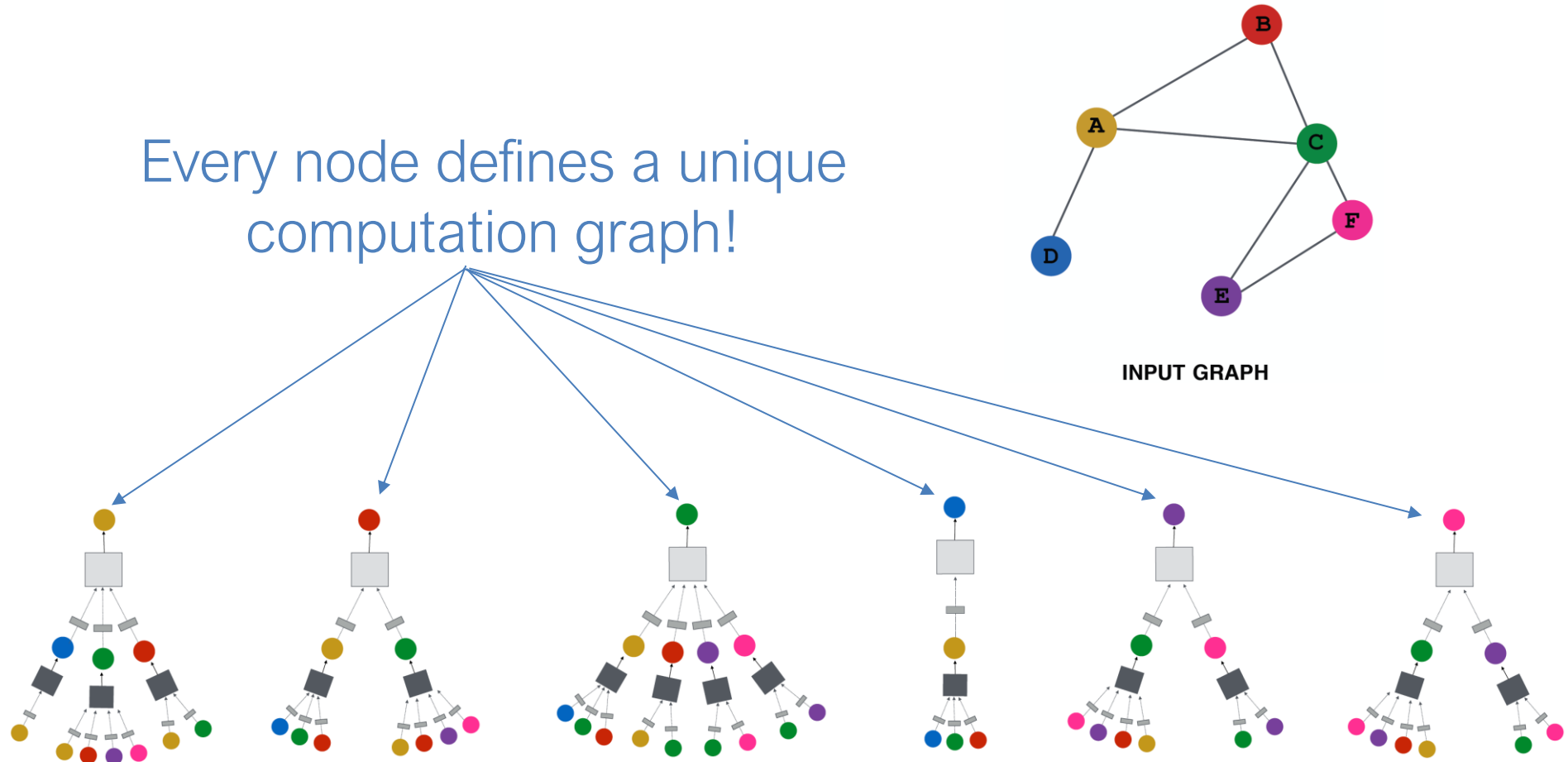
Key idea: Generate node embeddings based on local neighborhoods in a recursive manner



[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Graph Neural Nets

Every node defines a unique computation graph!

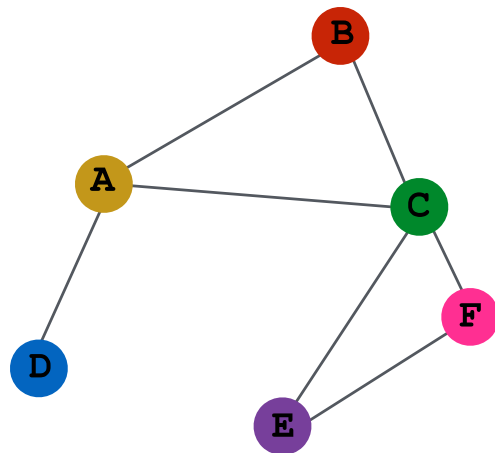


[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

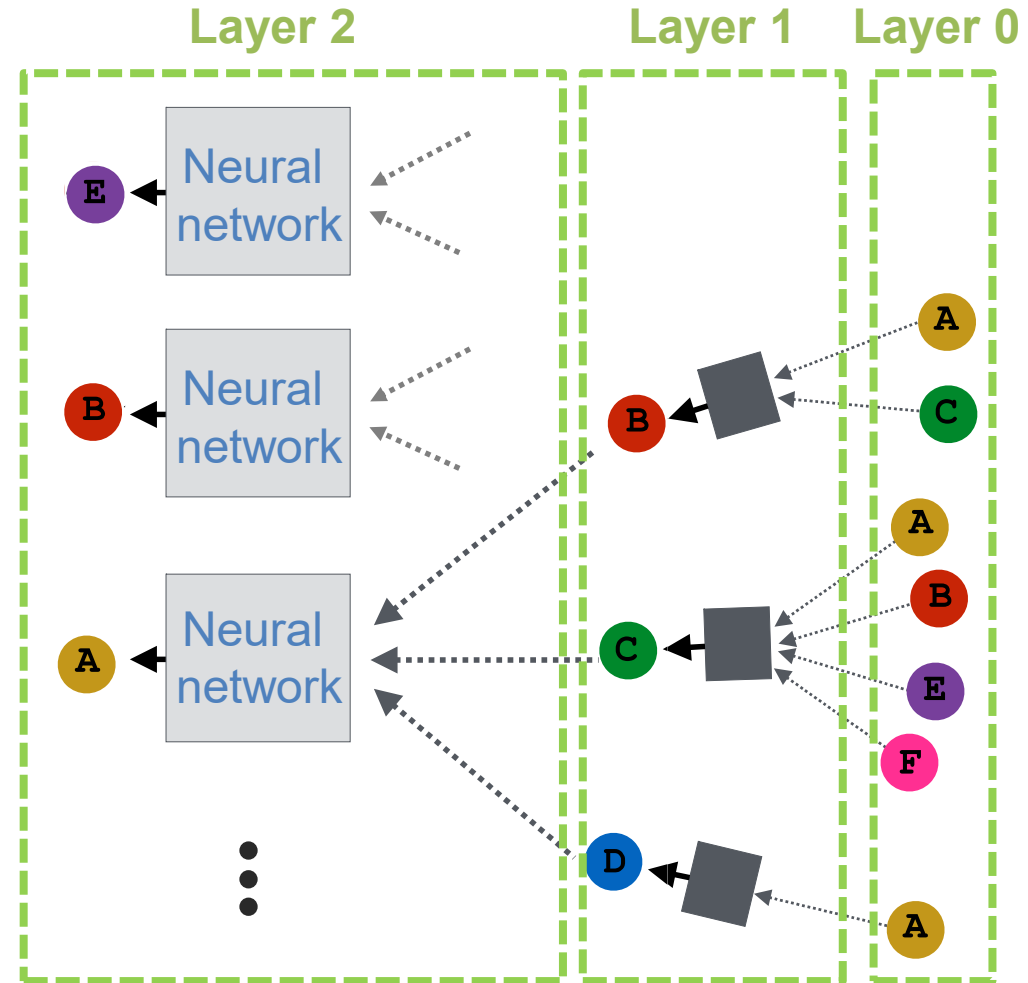
Graph Neural Nets

And multiple layers!

- ➡ Shared parameters within a specific layer
- ➡ “layer-0” is the input feature x_u

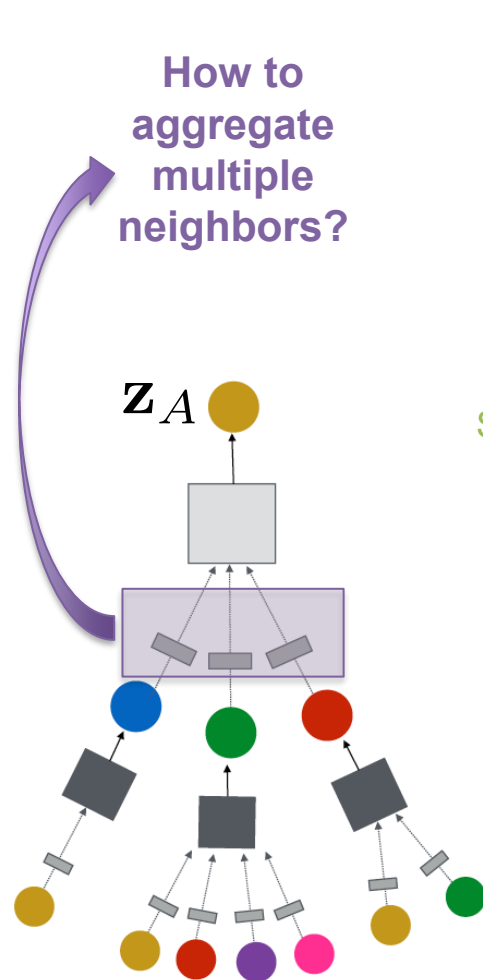


INPUT GRAPH



[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Graph Neural Nets – Neighborhood Aggregation



Average pooling (Scarselli et al., 2005)

$$\mathbf{h}_v^k = \sigma \left(\mathbf{W}_k \sum_{u \in N(v)} \frac{\mathbf{h}_u^{k-1}}{|N(v)|} + \mathbf{B}_k \mathbf{h}_v^{k-1} \right)$$

K is num layers

Different weights for neighbors and self

Graph Convolution Network (Kipf et al., 2017)

$$\mathbf{h}_v^k = \sigma \left(\mathbf{W}_k \sum_{u \in N(v) \cup v} \frac{\mathbf{h}_u^{k-1}}{\sqrt{|N(u)| |N(v)|}} \right)$$

Same weights

Different normalization

It can be efficiently implemented

Graph Attention Network (Velickovic et al., 2018)

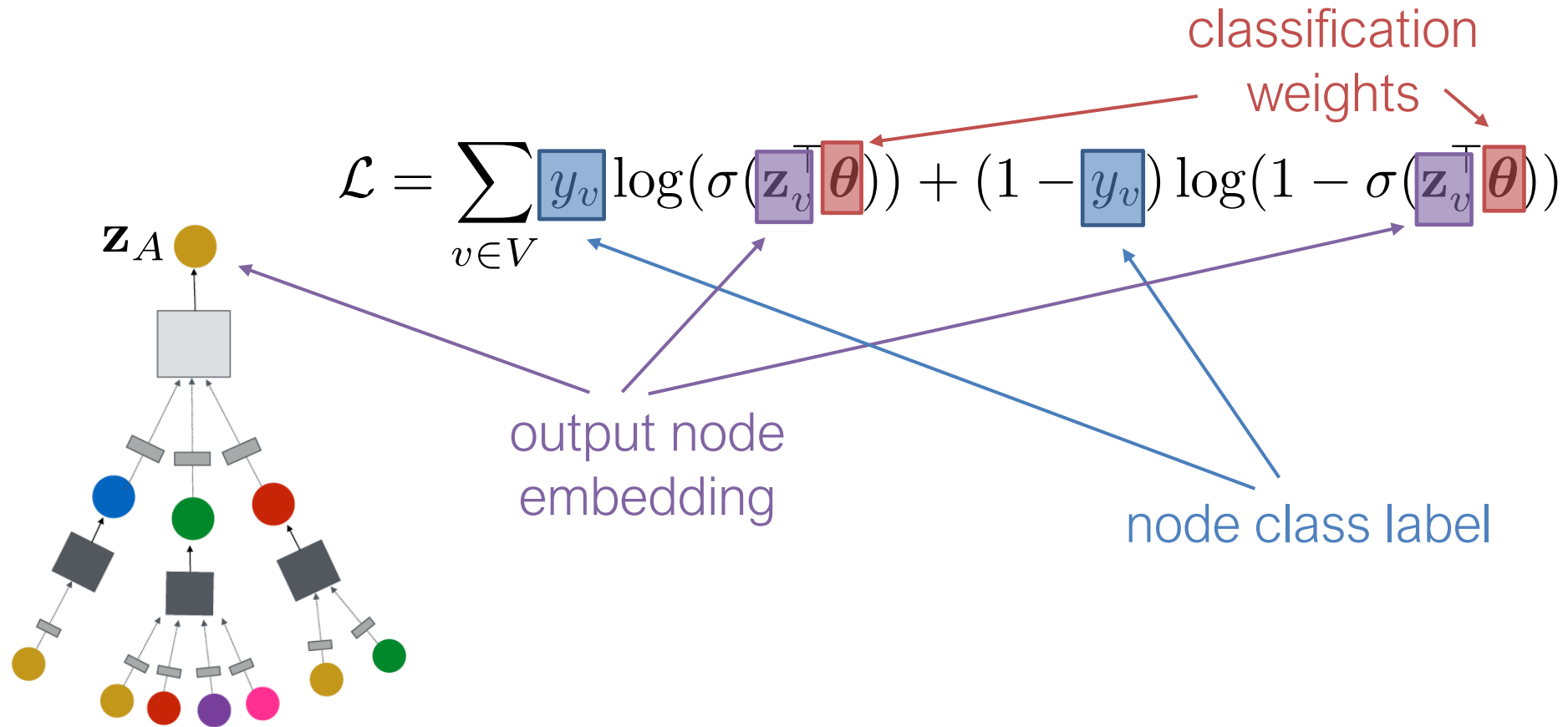
$$\mathbf{h}_v^k = \sigma \left(\mathbf{W}_k \sum_{u \in N(v) \cup v} \frac{\alpha_{uv} \mathbf{h}_u^{k-1}}{\sqrt{|N(u)| |N(v)|}} \right)$$

Attention weights

Very similar to a self-attention transformer

[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Graph Neural Nets – Supervised Training



[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]