Multimodal Machine Learning for Healthcare

Marija Stanojevic, PhD
University of Toronto, 28th July, 2023

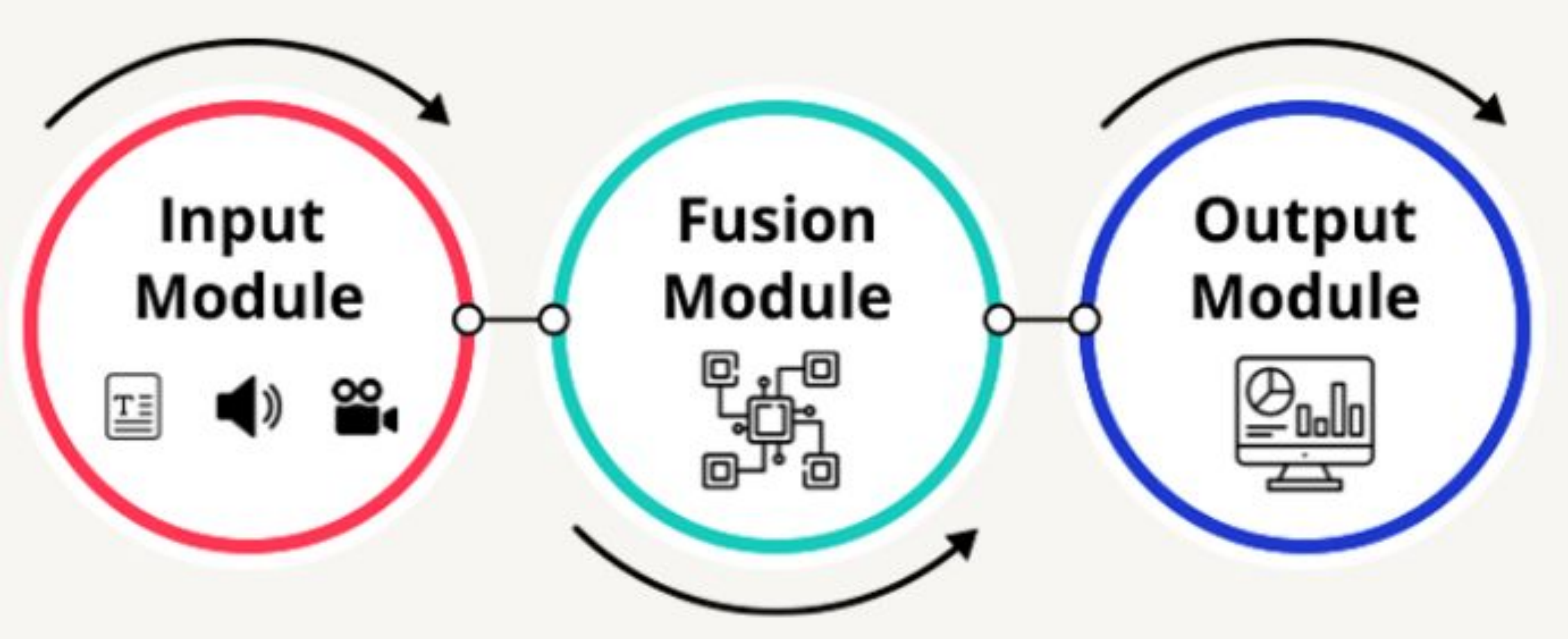
Why multimodal (MML) learning?

- Humans are MML learners
    - vision, smell, touch, taste, sound, text, move,...
- Is it possible to teach machine everything we know with just one modality?
- Data is originally multimodal
    - research articles, youtube, fintech, news, company data, medical data, fitness data, ...

Why are we doing anything else, then?

- Integrating (aligning) modalities has is hard
- Additional modalities add more noise, more parameters, and require more knowledge and data
- Explainability, Generalization and Scalability is harder
- Why should you be interested:
    - pretrained DL models allow for faster MML
    - many open challenges
    - many applications - it would change the world

What is multimodal (MML) learning?



https://insights.daffodilsw.com/blog/multimodal-ai

Fusion module

Similarity

- Inner product: $\mathbf{u}\mathbf{v}$

Linear / sum

- Concat: $W[\mathbf{u},\mathbf{v}]$
- Sum: $W\mathbf{u}+V\mathbf{v}$
- Max: $\max(W\mathbf{u}, V\mathbf{v})$

Multiplicative

- Multiplicative: $W\mathbf{u}\odot V\mathbf{v}$
- Gating: $\sigma(W\mathbf{u})\odot V\mathbf{v}$
- LSTM-style: $\tanh(W\mathbf{u})\odot V\mathbf{v}$

Attention

- Attention: $\alpha W\mathbf{u}+\beta V\mathbf{v}$
- Modulation: $[\alpha\mathbf{u},(1-\alpha)\mathbf{v}]$

Bilinear

- Bilinear: $\mathbf{u}W\mathbf{v}$
- Bilinear gated: $\mathbf{u}W\boldsymbol{\sigma}(\mathbf{v})$
- Low-rank bilinear: $\mathbf{u}U^{T}V\mathbf{v}=P(U\mathbf{u}\odot V\mathbf{v})$
- Compact bilinear: $FFT^{-1}(FFT(\Psi(\mathbf{x},\mathbf{h}_1,\mathbf{s}_1))$ $\odot FFT(\Psi(\mathbf{x},\mathbf{h}_2,\mathbf{s}_2)))$

https://web.stanford.edu/class/cs224n/slides/Multimodal-Deep-Learning-CS224n-Kiela.pdf

Where to place fusion module?

- Early - Inputs fusion:

$$\sigma(W_2\sigma(W_1[\mathbf{u},\mathbf{v}]+b_1)+b_2)$$

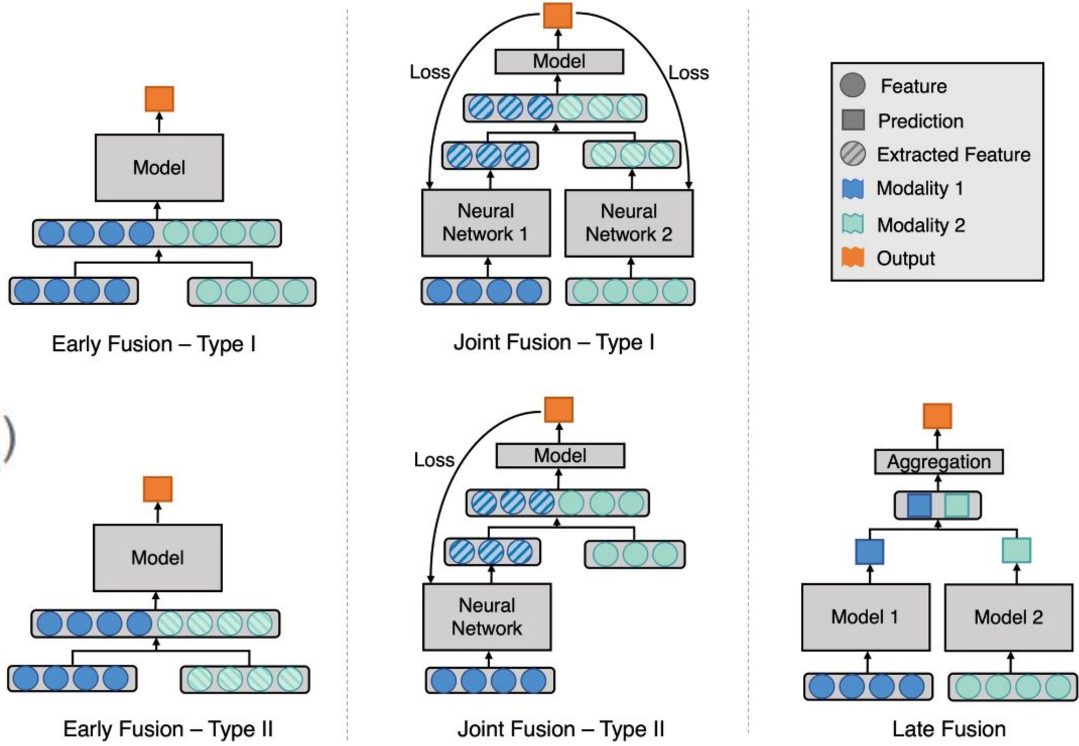
- Joint - Features fusion:

$$\sigma(W_2[\sigma(W_1[\mathbf{v}]+b_1), \sigma(W'_1[\mathbf{v}]+b'_1)] +b_2)$$

- Late - Outputs fusion:

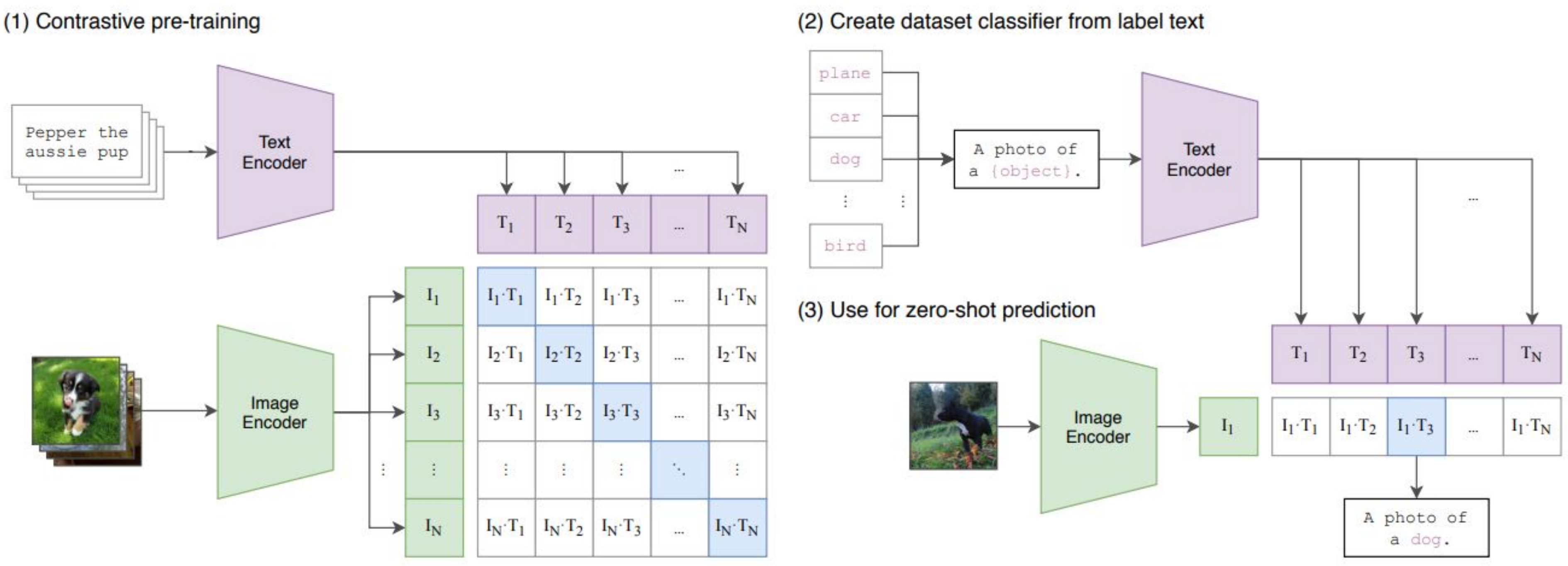
$$1/2\ (\sigma(W_2\sigma(W_1[\mathbf{u}]+b_1)+b_2) + \sigma(V_2\sigma(V_1[\mathbf{u}]+b'_1)+b'_2))$$

Formulas use concatenation for fusion, but modality fusion is flexible.



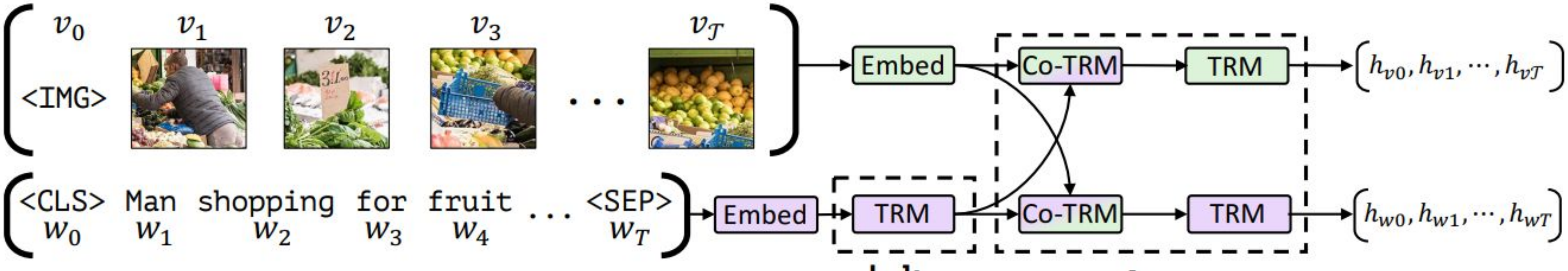
Fusion strategies using deep learning, by Huang et al., 2020

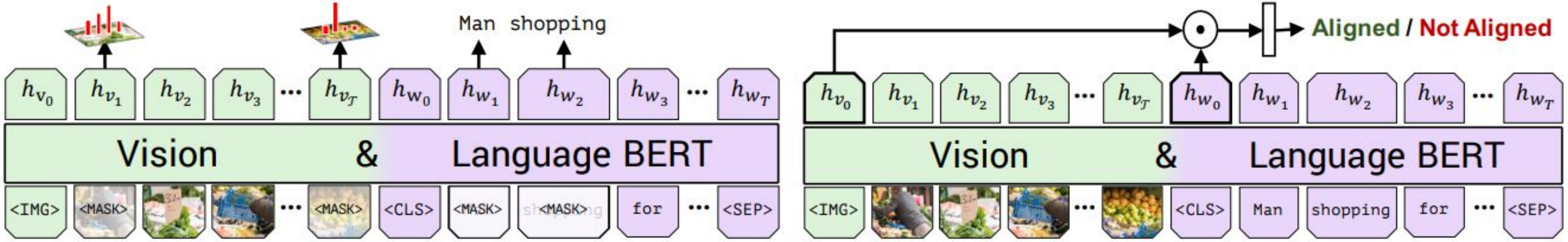
CLIP (Radford et al., 2021) - contrastive



DALL-E, DALL-E 2, Stable Diffusion, Midjourney, LAION (+5B open dataset)

ViLBERT (Lu et al., 2019) - discriminative





(a) Masked multi-modal learning

(b) Multi-modal alignment prediction

VL-PTM	Text encoder	Vision encoder	Fusion scheme	Pre-training tasks	Multimodal datasets for pre-training
Fusion Encoder					
VisualBERT [2019]	BERT	Faster R-CNN	Single stream	MLM+ITM	COCO
Uniter [2020]	BERT	Faster R-CNN	Single stream	MLM+ITM+WRA+MRFR+MRC	CC+COCO+VG+SBU
OSCAR [2020c]	BERT	Faster R-CNN	Single stream	MLM+ITM	CC+COCO+SBU+Flickr30k+VQA
InterBert [2020]	BERT	Faster R-CNN	Single stream	MLM+MRC+ITM	CC+COCO+SBU
ViLBERT [2019]	BERT	Faster R-CNN	Dual stream	MLM+MRC+ITM	CC
LXMERT [2019]	BERT	Faster R-CNN	Dual stream	MLM+ITM+MRC+MRFR+VQA	COCO+VG+VQA
VL-BERT [2019]	BERT	Faster R-CNN+ ResNet	Single stream	MLM+MRC	CC
Pixel-BERT [2020]	BERT	ResNet	Single stream	MLM+ITM	COCO+VG
Unified VLP [2020]	UniLM	Faster R-CNN	Single stream	MLM+seq2seq LM	CC
UNIMO [2020b]	BERT, RoBERTa	Faster R-CNN	Single stream	MLM+seq2seq LM+MRC+MRFR+CMCL	COCO+CC+VG+SBU
SOHO [2021]	BERT	ResNet + Visual Dictionary	Single stream	MLM+MVM+ITM	COCO+VG
VL-T5 [2021]	T5, BART	Faster R-CNN	Single stream	MLM+VQA+ITM+VG+GC	COCO+VG
XGPT [2021]	transformer	Faster R-CNN	Single stream	IC+MLM+DAE+MRFR	CC
Visual Parsing [2021]	BERT	Faster R-CNN + Swin transformer	Dual stream	MLM+ITM+MFR	COCO+VG
ALBEF [2021a]	BERT	ViT	Dual stream	MLM+ITM+CMCL	CC+COCO+VG+SBU
SimVLM [2021b]	ViT	ViT	Single stream	PrefixLM	C4+ALIGN
WenLan [2021]	RoBERTa	Faster R-CNN + EffcientNet	Dual stream	CMCL	RUC-CAS-WenLan
ViLT [2021]	ViT	Linear Projection	Single stream	MLM+ITM	CC+COCO+VG+SBU
Dual Encoder					
CLIP [2021]	GPT2	ViT, ResNet		CMCL	self-collected
ALIGN [2021]	BERT	EffcientNet		CMCL	self-collected
DeCLIP [2021b]	GPT2, BERT	ViT, ResNet, RegNetY-64GF		CMCL+MLM+CL	CC+self-collected
Fusion Encoder+ Dual Encoder					
VLMo [2021a]	BERT	ViT	Single stream	MLM+ITM+CMCL	CC+COCO+VG+SBU
FLAVA [2021]	ViT	ViT	Single stream	MMM+ITM+CMCL	CC+COCO+VG+SBU+RedCaps

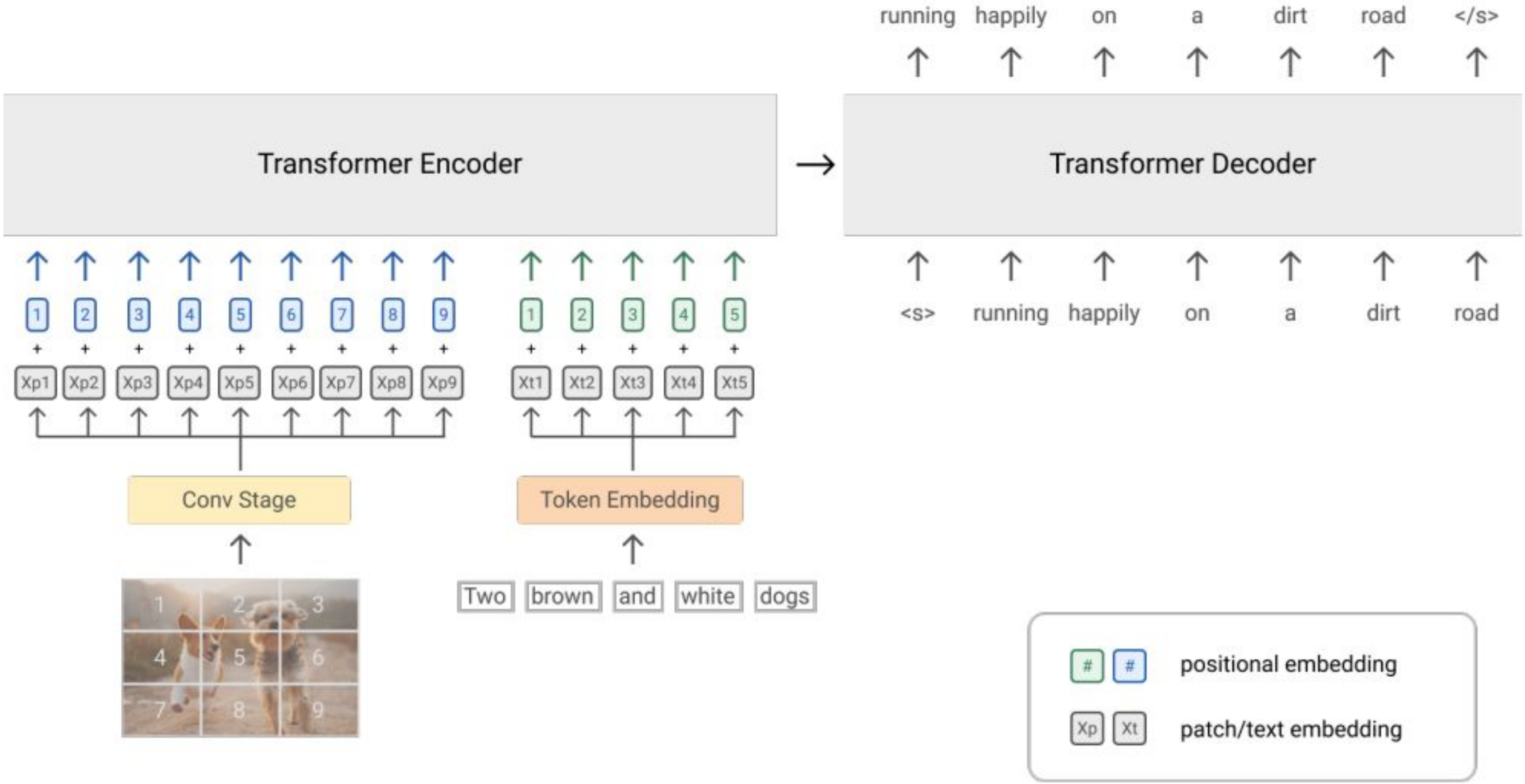
PMD - 70 M dataset

COCO	Visual Genome	SBU captions	Localized narratives	WIT	RedCaps	CC12M	YFCC filtered
A close up view of a pizza sitting on a table with a soda in the back.	a lenovo laptop rebooting	Front view of basket 13, from the sidewalk in front of the basket.	The woman is touching a utensil in front of her on the grill stand.	Typocerus balteatus, Subfamily: Flower Longhorns	Deigdoh falls in india	Jumping girl in a green summer dress stock illustration	In the kitchen at the Muse Nissim de Camondo

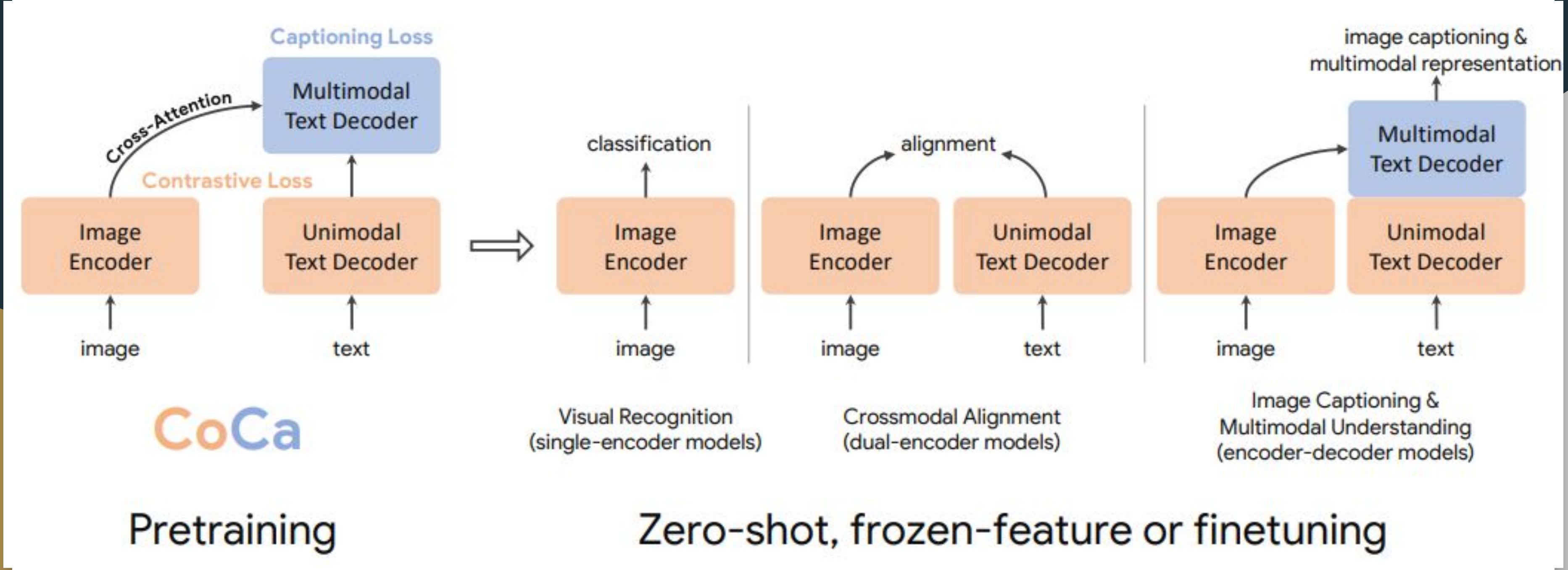
Dataset	Size	Reference
COCO	328,124	[Lin *et al.*, 2014]
VG	108,077	[Krishna *et al.*, 2017]
CC	3.1M	[Sharma *et al.*, 2018]
SBU	1M	[Ordonez *et al.*, 2011]
LAION	400M	https://laion.ai/laion-400-open-dataset/
RedCaps	12M	[Desai *et al.*, 2021]

Table 2: Widely Used Pre-training Datasets

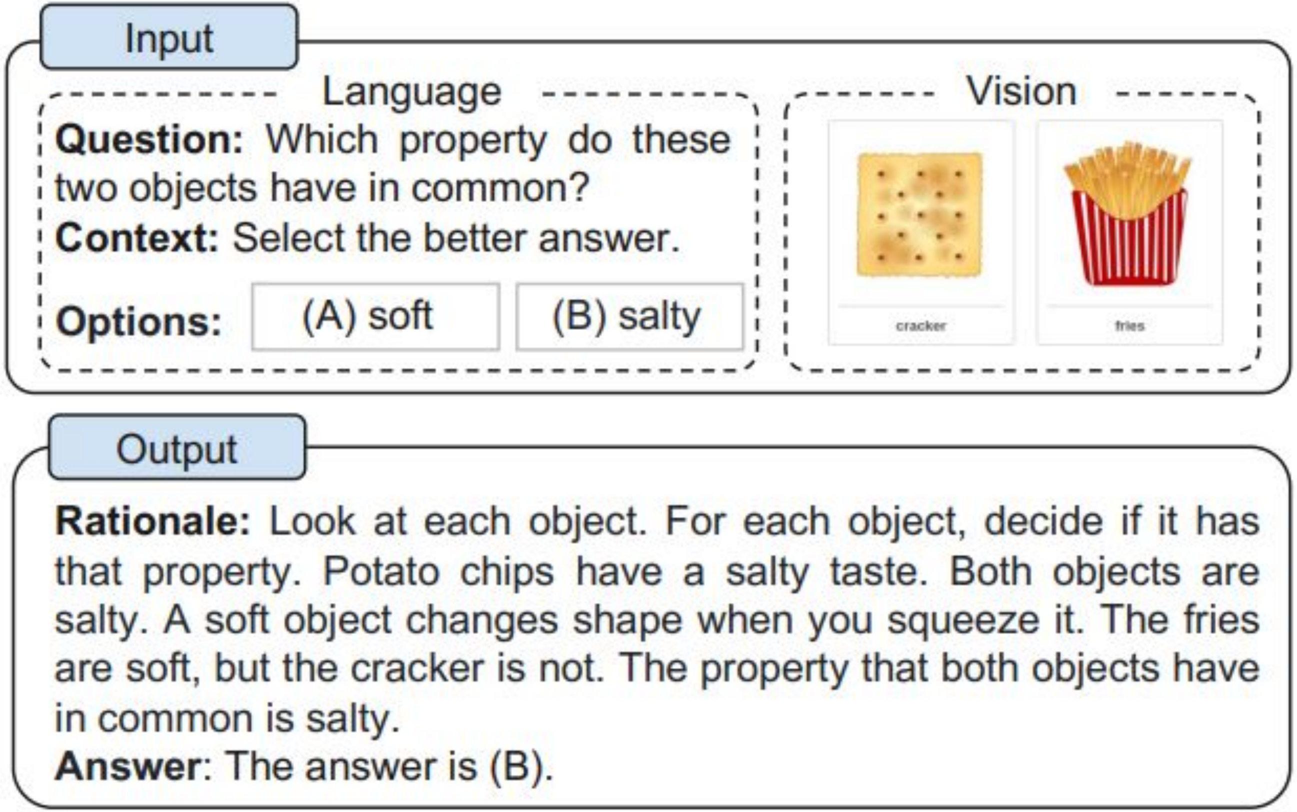
SimVLM (Wang et al., 2022) - generative



CoCa (Yu et al., 2022) - contrastive & generative

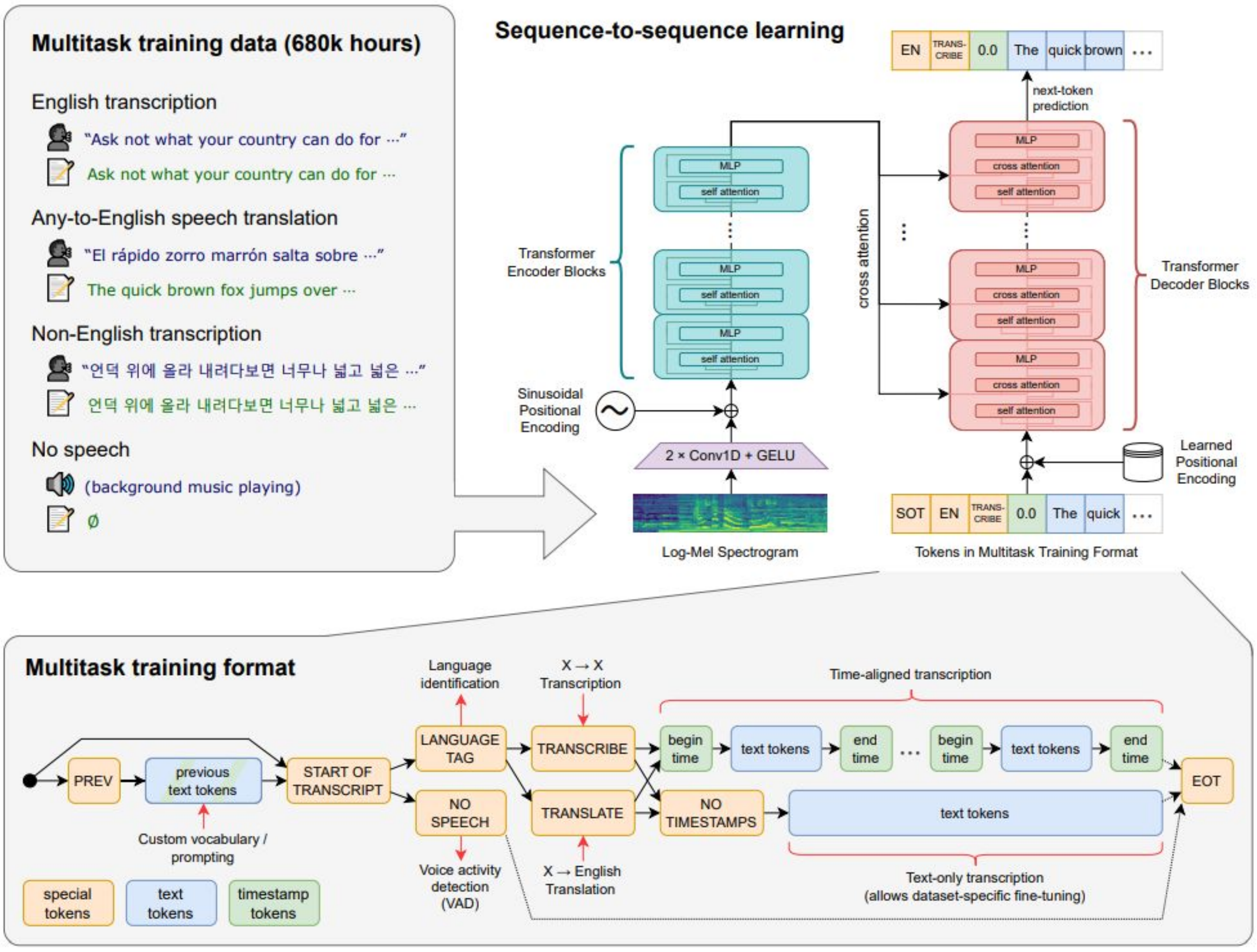


Multimodal Chain-of-Thought (Zhang et al, 2023)

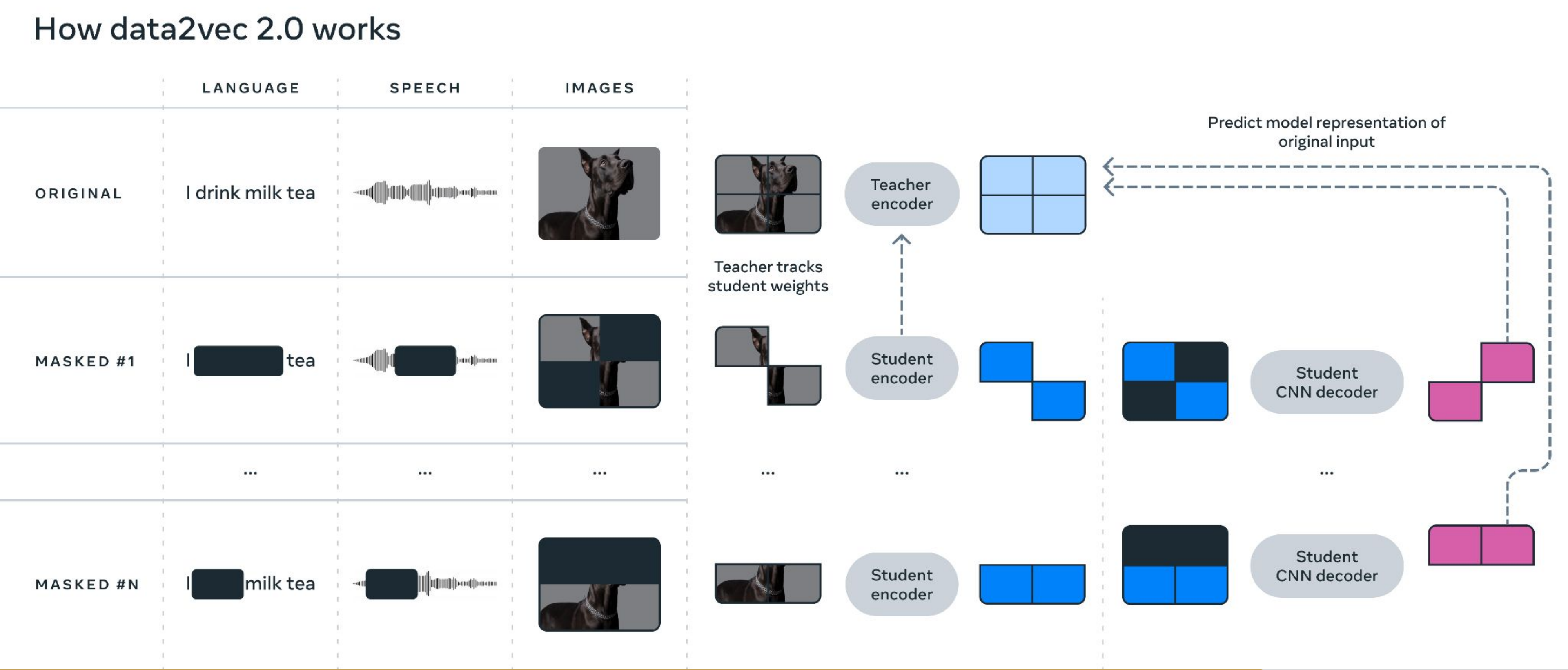


Whisper (Radford et al, 2022) - generative - mixing audio & text - ASR & AST

AudioPalm (Rubenstein et al., 2023)



data2vec & data2vec2 (Baevski et al, 2022; 2023)



MERIOT RESERVE (Zellers et al, 2022)

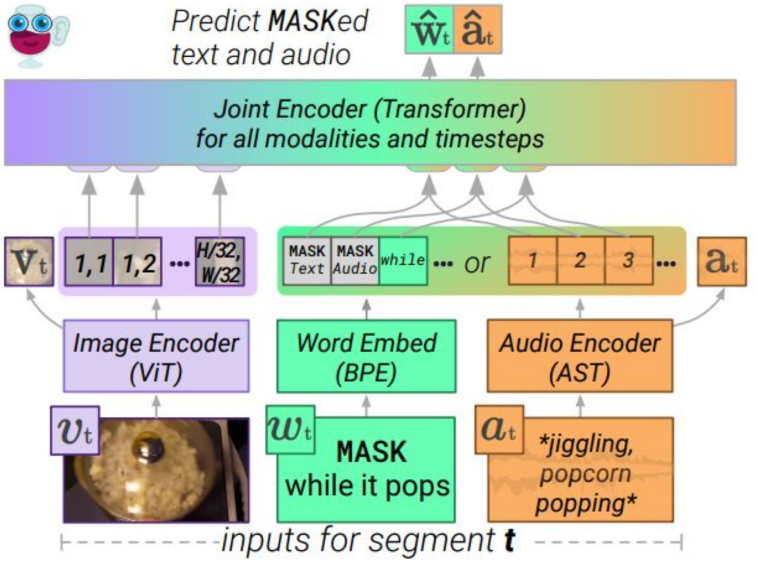


Figure 2: RESERVE architecture. We provide sequence-level representations of video frames, and *either* words or audio, to a joint encoder. The joint encoder contextualizes over modalities and segments, to predict what is behind MASK for audio $\widehat{\mathbf{a}}_t$ and text $\widehat{\mathbf{w}}_t$. We supervise these predictions with independently encoded targets: $\mathbf{a}_t$ from the audio encoder, and $\mathbf{w}_t$ from a separate text encoder (not shown).

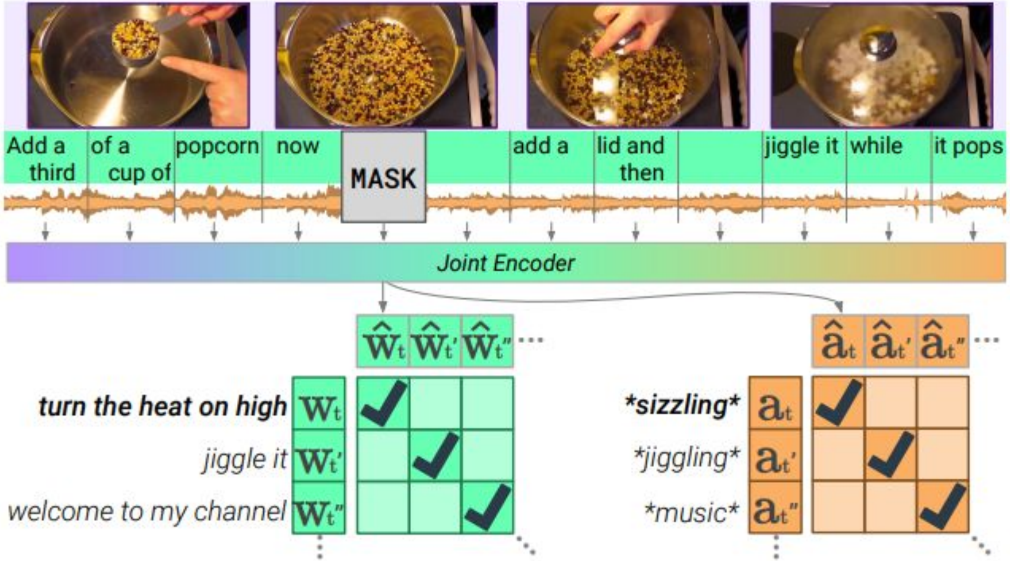
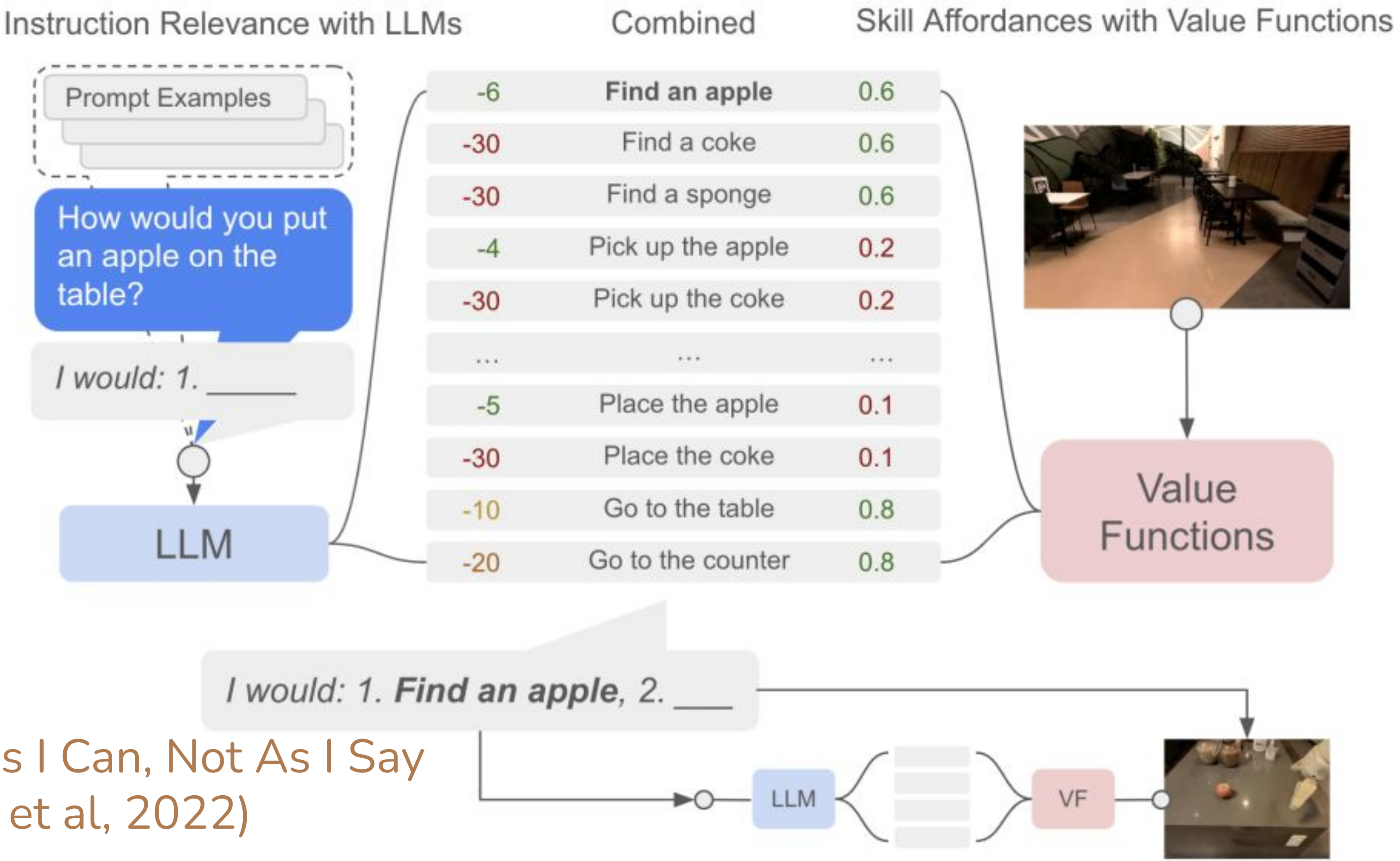
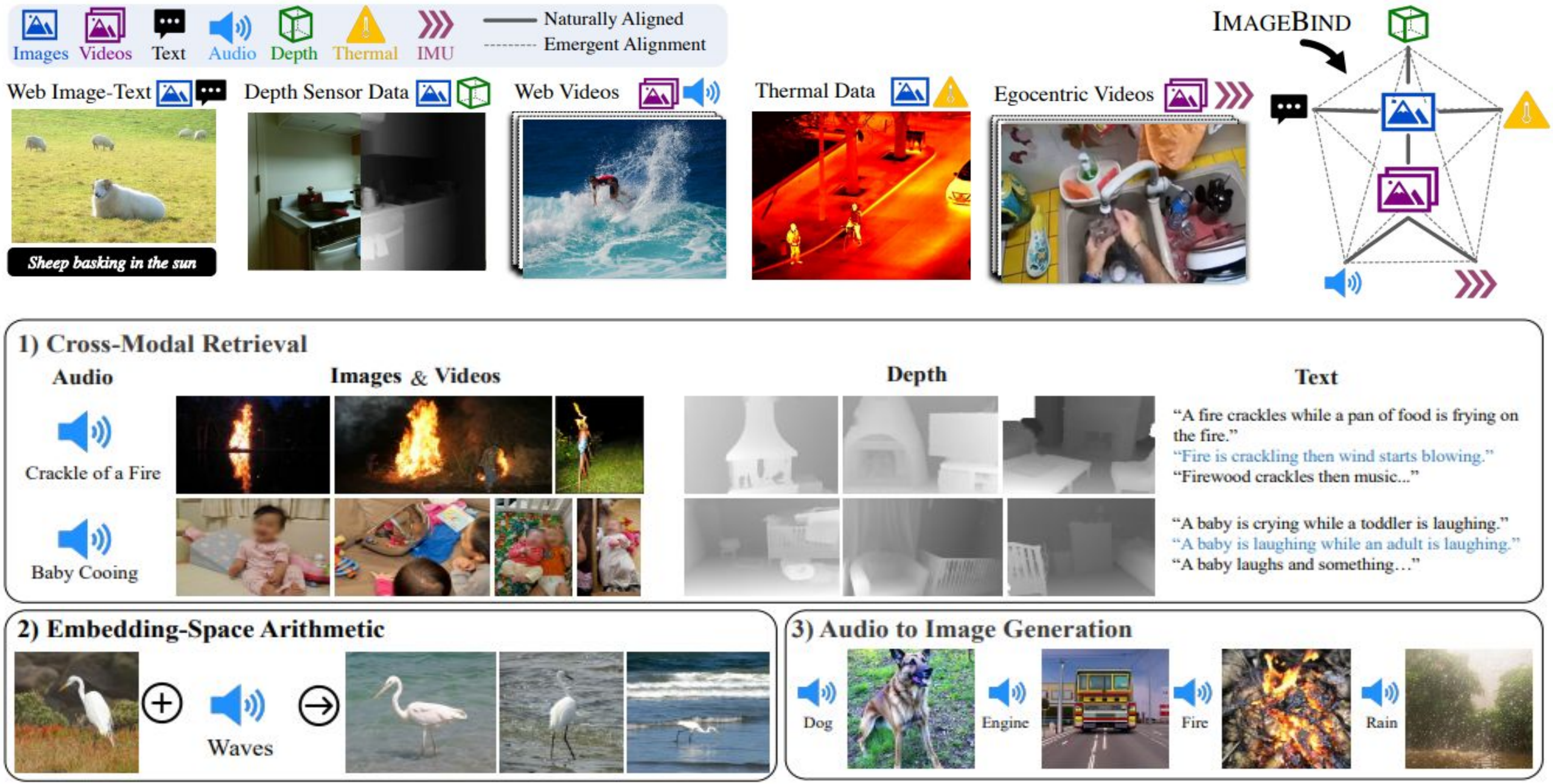


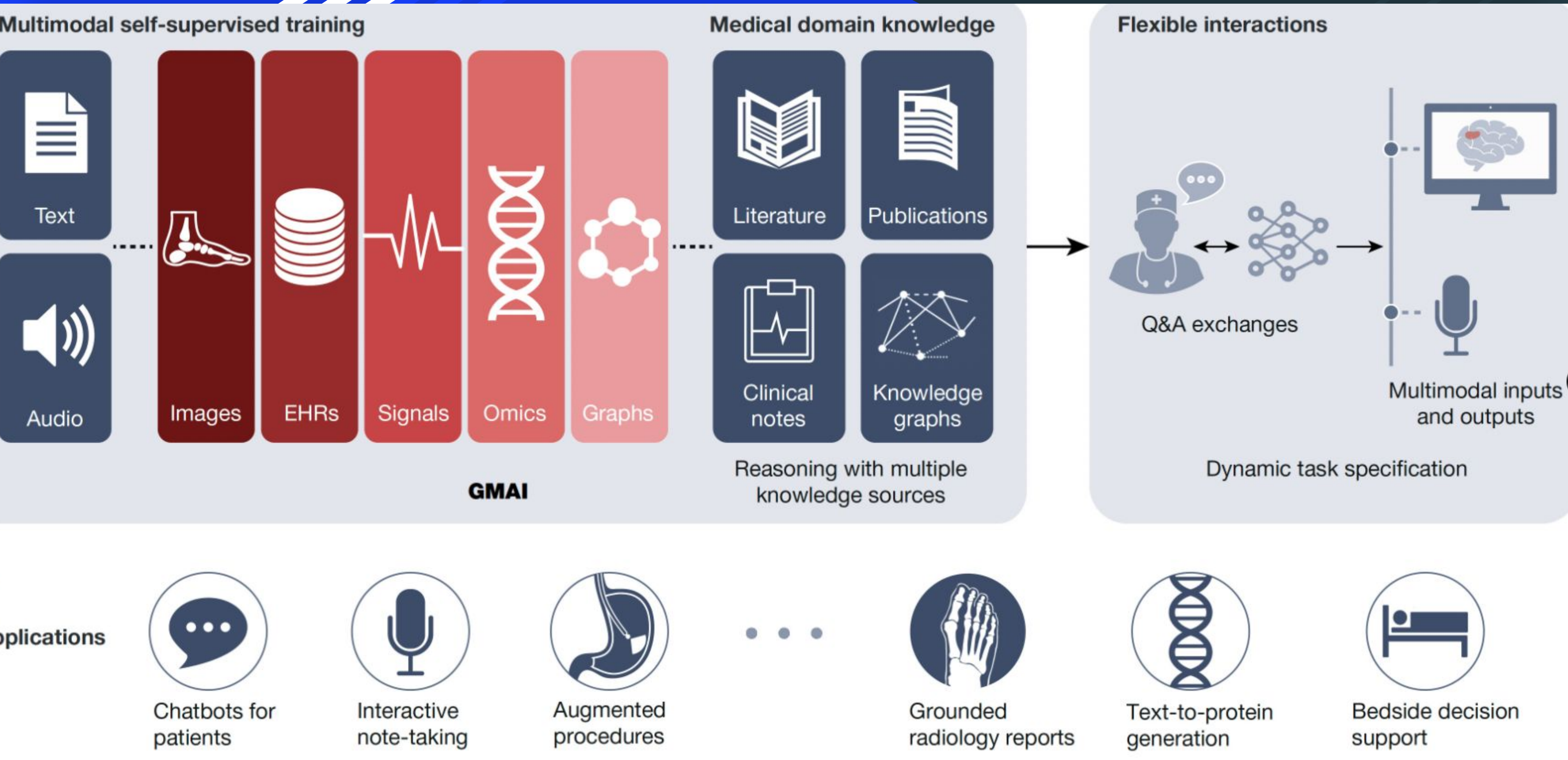
Figure 3: Contrastive span training. Given a video with all modalities temporally aligned, we MASK out a region of text and audio. The model must maximize its similarity *only to* an independent encoding of the text $\mathbf{w}_t$ and audio $\mathbf{a}_t$.



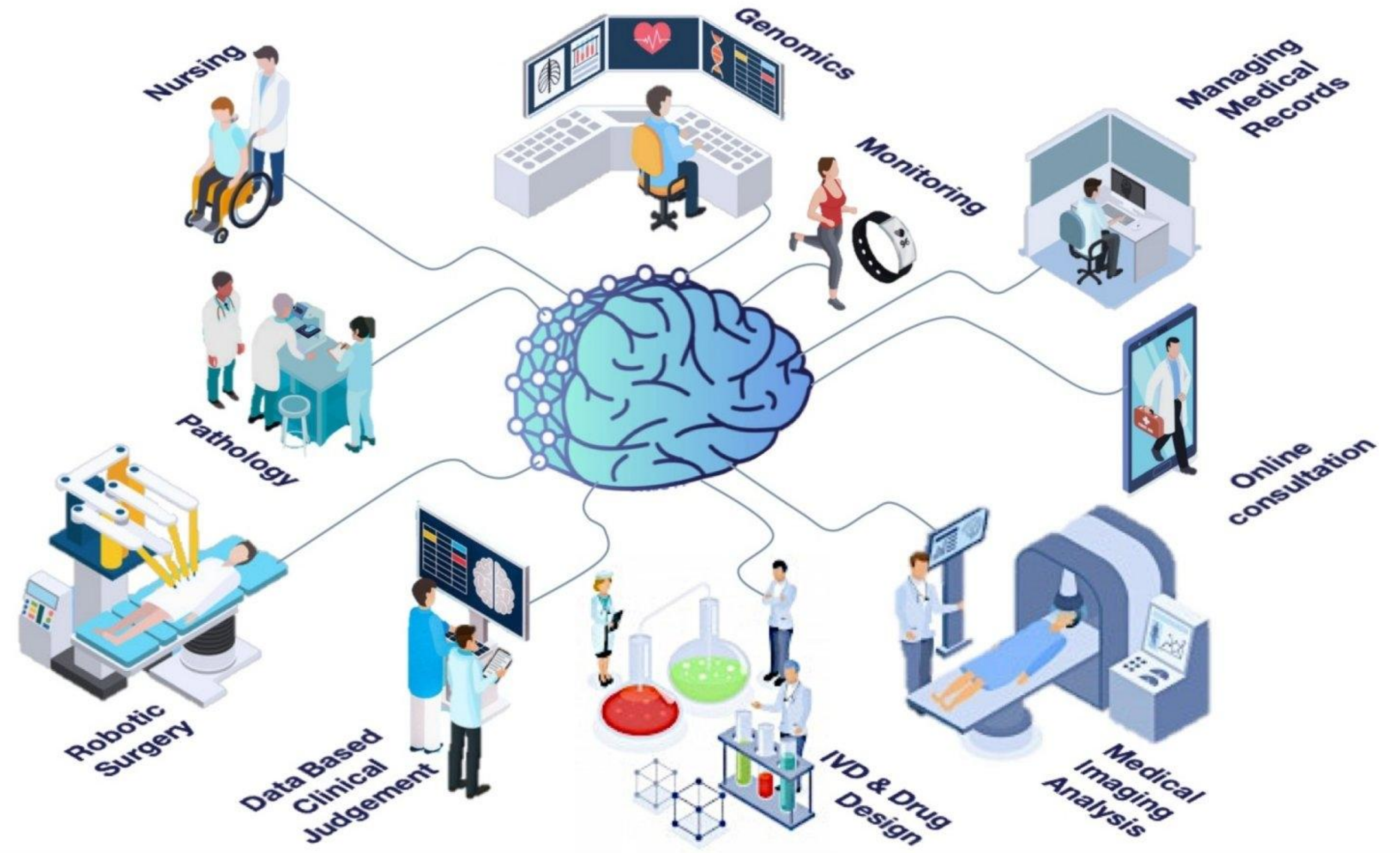
Do As I Can, Not As I Say
(Ahn et al, 2022)

ImageBind (Girdhar et al, 2023)





https://erictopol.substack.com/p/generative-ai-and-the-new-medical



https://mkai.org/artificial-intelligence-framework-reveals-nuance-in-performance-of-multimodal-ai-for-health-care/

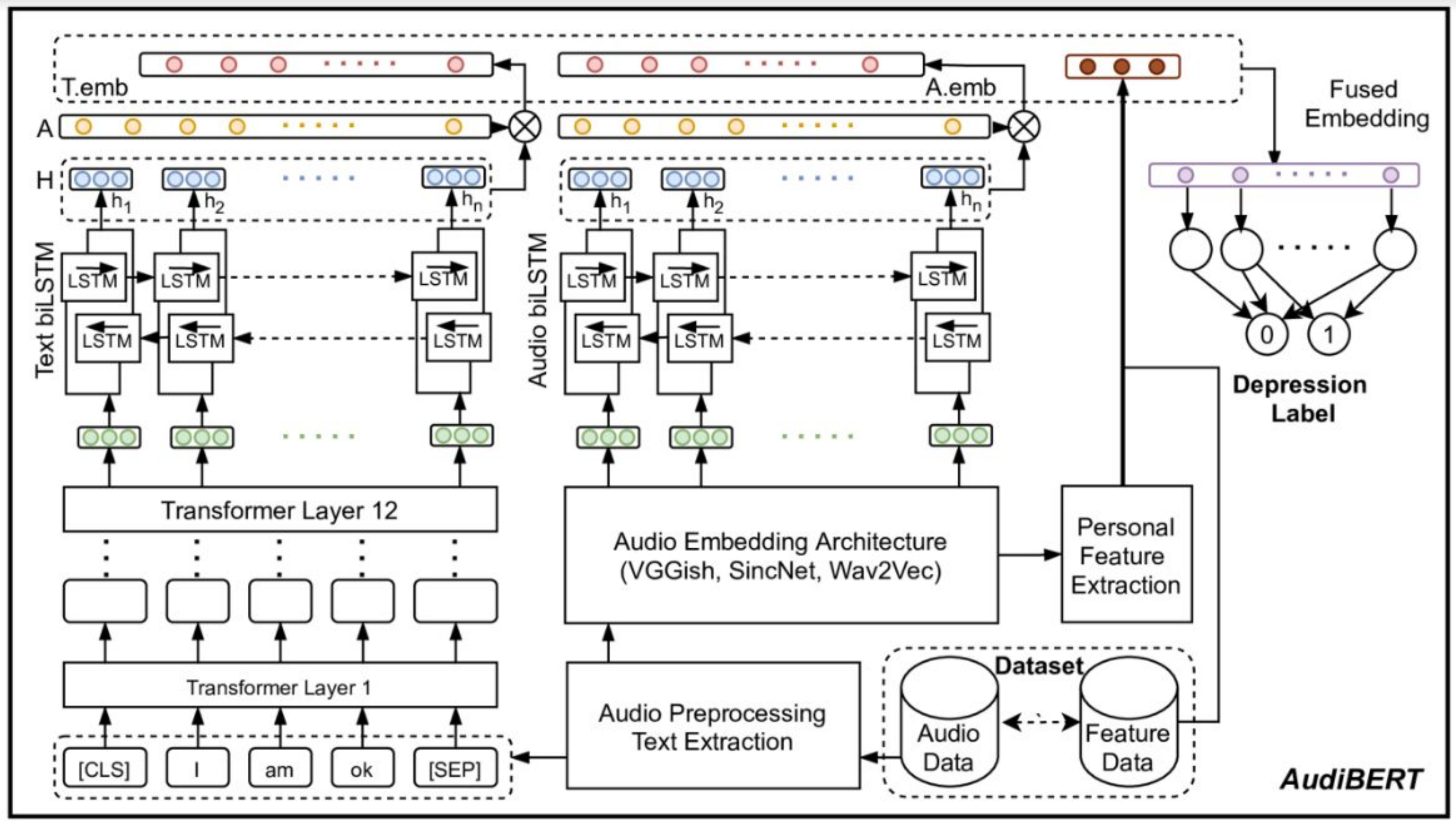
Challenges of MML for Healthcare

- Data Sources - private, hard to share - federated learning
- Temporality - signifies a moment in time of a patient
- Missing values
- Skewness
- Interpretability, fairness, explainability
- Tracking various modalities to model complex human body

Resources

- Soenksen et al, 2022; Acosta, et al, 2022

AudiBERT (Toto et al, 2021)



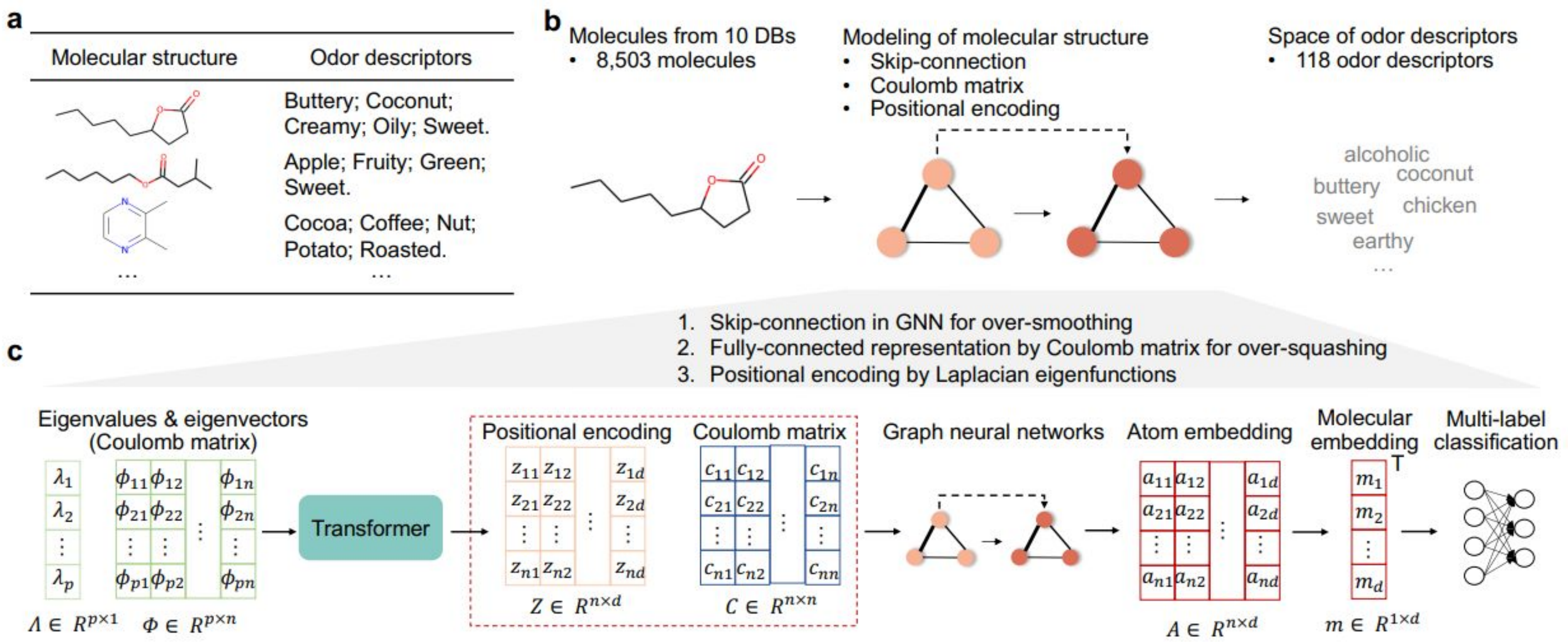
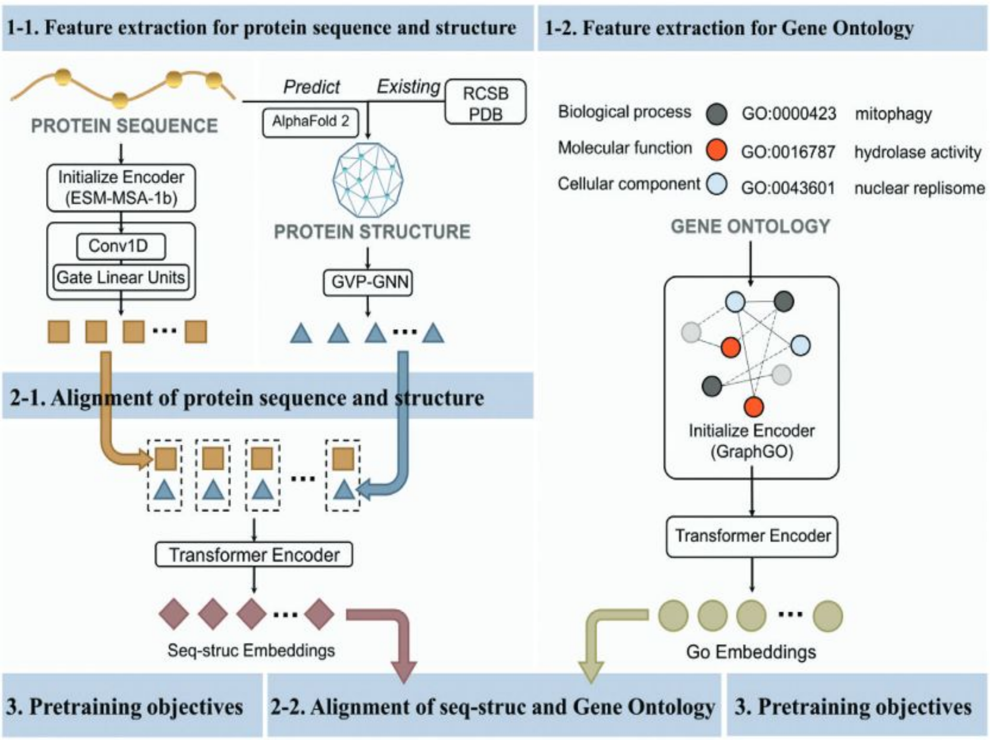


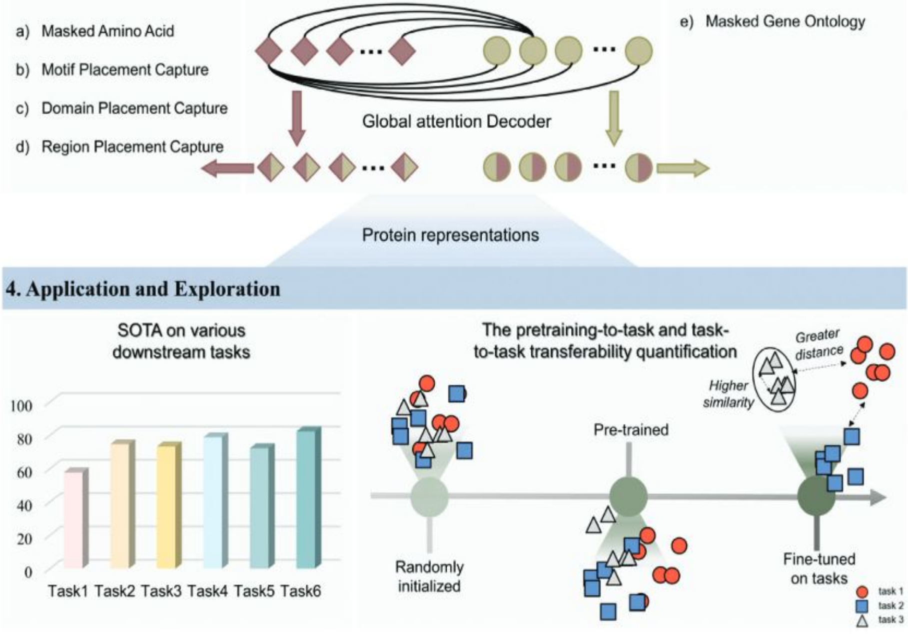
Figure 1: Overview of Mol-PECO. (a) Typical molecular structures and the corresponding odor descriptors are shown as examples. (b) The main workflow of modeling quantitative structure-odor relationship (QSOR). (c) The detailed model architecture of Mol-PECO and its three features: 1) skip-connection in graph neural networks to alleviate over-smoothing, 2) fully-connected molecular representation by Coulomb matrix to suppress over-squashing, and 3) positional encoding by Laplacian eigenfunctions.

Mol-PECO (Zhang et al. 2023) - olfactory

MASSA (Hu et al, 2023)

- proteins, genes, and ontologies





Future of MML

- Modality-agnostic “foundation models” with versatile discriminative and generative capabilities
- Parameters sharing will be solved in better ways
- End-to-end automatic alignment from unpaired unimodal data
- Enhanced scaling (Aghajanyan et al, 2023)
- Retrieval capabilities
- More modalities (structured data, sensors & financial signals, graphs) & more applications
- Better evaluation and benchmarking (Barbarosa-Silva et al. 2022)

Resources

- CMU Multimodal ML course
- CVPR tutorial on Multimodal ML
- UBC Multimodal Learning with Vision, Language and Sound
- Awesome MML
- MML session by Microsoft Applied Scientist - Sep 21st
- Virginia Tech Multimodal Vision
- Stanford Multimodal Deep Learning lecture
- Vision-Language Models