Advances in Multimodal Models: Continual Learning and Trustworthiness

Dianzhi Yu[1], Xinni Zhang[1], Guanzhong Chen[2], Zhuo Wang[3], Yukun Zhang[2], Zenglin Xu[3,4] and Irwin King[1]

[1] The Chinese University of Hong Kong, Hong Kong, China

[2] Harbin Institute of Technology Shenzhen, China

[3] Fudan University, China

[4] Shanghai Academy of AI for Science, China

June 30, 2025

Biography

Prof. Irwin King

- **Ph.D.**, Computer Science, June 1993. **University of Southern California (USC)**, Los Angeles, CA. USA
- **M.Sc.**, Computer Science, **University of Southern California (USC)**, Los Angeles, CA. USA
- **B.Sc.**, Engineering & Applied Science (Computer Science), **California Institute of Technology (Caltech)**, Pasadena, CA. USA

1. **Associate Dean (2012-18), Chair (2020-23), Director of ELITE (2017)**
2. **ACM Fellow (2024); IEEE Fellow (2019)**; **INNS Fellow**; AAIA Fellow; HKIE Fellow; Kavli Fellow; Lee Woo Sing College Fellow; Apple Distinguished Educator
3. Global AI 2000 List since 2000, Top 2% of World Scientists since 2020
4. **ACM WSDM 2022 Test of Time Award**
5. **ACM SIGIR 2020 Test of Time Award**
6. **ACM CIKM 2019 Test of Time Award**
7. **2021 INNS Dennis Gabor Award for work in Neural Engineering**
8. **2020 Asia Pacific Neural Network Society (APNNS) Outstanding Achievement Award**
9. ICONIP 2020 Best Paper Award
10. JCDL 2012 Vannevar Bush Best Paper Award
11. The 4th Beijing-Hong Kong International Doctoral Forum 2009 Best Paper Award
12. CVPR 2019 Best Paper Finalist; CIKM 2016 Best Paper Award Runner-up; ICONIP 2017 Best Paper Award Runner-up
13. Seven patents world-wide in information technology-related areas
14. Over HKD $80 M in competitive and non-competitive research grants, project contracts, education, etc. Grants: Research (HKD $10.78 M), Education (HKD $64.02 M), Industry (HKD $5.85 M)

Machine Intelligence and Social Computing (MISC) Lab

Trustworthy AI

Privacy, Security, Robustness, Fairness, Explainability, Interpretability, Watermarking, Accountability, Policy, etc.

Social Computing

- Big data
- Data mining
- Social recommender systems
- Social media analysis
- Social network analysis
- Graph algorithms
- Community search and detection

Natural Language Processing (NLP)

- Large language model (LLM)
- Sentiment analysis
- Summarization
- Translation
- Language models
- Speech Language Models
- Multilingual modeling
- Fact-checking
- Watermarking in LLMs

Machine Learning

- Foundation models
- Semi-supervised learning
- Online learning
- Self-supervised learning
- Multimodal learning
- Continual learning
- Contrastive learning
- Federated learning
- Hyperbolic embedding

Graph Neural Networks (GNN) & AI for Science

- Heterogeneous GNN
- Graph algorithms
- Knowledge graph
- Bioinformatics

- AI for science
- Gastric cancer diagnosis and prediction
- Bioinformatics

Recent Works

Machine Learning

1. Pretrain Model for Crystal Property Prediction, **AAAI 2024**
2. Hyperbolic Efficient Transformer, **KDD 2024**
3. Geometric View of Soft Decorrelation in Self-Supervised Learning, **KDD 2024**
4. Hyperbolic Temporal Network Embedding Learning, **TKDE 2023**
5. Meta-Learning with Motif-based Task Augmentation for Few-Shot Molecular Property Prediction, **IJCAI 2023**

Social Computing

1. Hierarchical Hyperbolic Product Quantization, **AAAI 2024**
2. Deep Structural Knowledge Exploration, **AAAI 2024**
3. Influential Exemplar Replay for Incremental Learning in Recommender Systems, **AAAI 2024**
4. Shopping Trajectory Representation Learning, **KDD 2024**
5. Mitigating the Popularity Bias of Graph Collaborative Filtering, **NeurIPS 2023**

NLP

1. Entropy-based Text Watermarking Detection, **ACL 2024**
2. Unforgeable Publicly Verifiable Watermark for Large Language Models, **ICLR 2024**
3. Improving Open Relation Extraction With Search Documents, **TKDE 2024**
4. Knowledge Graph Entity Typing, **NAACL 2024**
5. Continuous Rationale Extraction for Relation Extraction, **SIGIR 2023**
6. Multimodal Relation Extraction, **ACL 2023**

Graphs

1. Long-Tail Distribution Issues in GNN, **TKDE 2024**
2. Empowering Graph Neural Networks with Expected Model Change Maximization, **NeurIPS 2023**
3. Optimal Block-wise Asymmetric Graph Construction for Graph-based Semi-supervised Learning, **NeurIPS 2023**
4. Bipartite Graph Convolutional Hashing, **WebConf 2023**
5. Contrastive Cross-scale Graph Knowledge Synergy, **KDD 2023**
6. Doubly Stochastic Graph-based Non-autoregressive Reaction Prediction, **IJCAI 2023**

Our Team

Team Members 2024-2025

Prof. Irwin King

Muzhi Li
KG, NLP

Yueen Ma
Robotics, VLA

Zhihang Hu
Bio, ML

MInda Hu
NLP, KG

Conghao Xiong
CV, MIA

Yixuan Wang
Bio, Graph

Jiahong Liu
HR, Graph

Zhihan Guo
FL, NLP

Zexuan Qiu
NLP, IR

Wenqian Cui
NLP, Speech

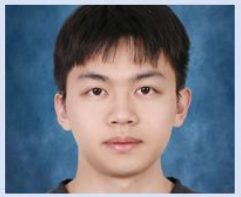
Dianzhi Yu
NLP, CL

Xinni Zhang
RecSys, Graph

Terrence Ng
FL, HCI

Yifan Li
NLP. KG

Ziqian Lin
Bio, ML



Wenaho Yu
Graph, AI4Sci

FL: Federated Learning, HR: Hyperbolic Representation, MIA: Medical Image Analysis, VLA: Vision Language Action, CL: Continual Learning

Biography

Prof. Zenglin Xu

- **Professor, Fudan University**
- **AI Scientist, Shanghai Academy of AI for Science**

Education Background

- **Ph.D., Computer Science and Engineering, 2009. The Chinese of University of Hong Kong, HongKong, China**
- **M.Sc., Computer Software and Theory, Xi'an Jiaotong University, Xi'an, China**
- **B.Sc., Computer Science and Technology, Xi'an Polytechnic University, Xi'an CHina**

Social Services

- **Vice President for Education of INNS**
- **Senior Action Editor of Neural Networks**
- **ICONIP 2023 Best Paper Finalist**
- **ACML 2016 Runner-up for Best Paper Award**
- **AAAI 2015 Outstanding Student Paper Honorable Mention**
- **.APNNS 2016 Young Investigator Award**
- **Consistently listed among the world's top 2% scientists by Stanford University**

Statistical Machine Intelligence & Learning (SMILE Lab)

- Goal: Trustworthy and Autonomous AI

- Research Interests
  - Trustworthy AI: Federated Learning, Secure Multi-party Computation
  - Large Language Models
  - Multi-agent Systems and Reinforcement Learning
  - AI for Health Informatics
  - AI for Finance
  - AI for Education

Recent Works

Federated Learning

1. StoCFL: A stochastically clustered federated learning framework for Non-IID data with dynamic client participation, **NN 2025**
2. Meta-Learning via PAC-Bayesian with Data-Dependent Prior: Generalization Bounds from Local Entropy, **IJCAI 2024**
3. On the Necessity of Collaboration for Online Model Selection with Decentralized Data, **NeurIPS 2024**
4. Topology Learning for Heterogeneous Decentralized Federated Learning Overleaf Unreliable D2D Networks, **IEEE TVT 2024**
5. SecFormer: Fast and Accurate Privacy-Preserving for Transformer vis SMPC, **ACL 2024**
6. Unveiling the Vulnerability of Private Fine-Tuning in Split-Based Frameworks for Large Language Models: A Bidirectionally Enhanced Attack, **CCS 2024**
7. Information-Theoretic Generalization Analysis for Topology-Aware Heterogeneous Federated Edge Learning Over Noisy Channels, **SPL 2024**
8. FEDLEGAL: The First Real-World Federated Learning Benchmark for Legal NLP, **ACL 2023**

Spatial and Temporal Modeling

1. GeoPro-Net: Learning Interpretable Spatiotemporal Prediction Models Through Statistically-Guided Geo-Prototyping , **AAAI 2025**
2. SMARTformer: Semi-Autoregressive Transformer with Efficient Integrated Window Attention for Long Time Series Forecasting, **IJCAL 2023**

Large Language Models

1. Preference-Strength-Aware Self-Improving Alignment with Generative Preference Models, **SIGIR 2025**
2. Preparing Lessons for Progressive Training on Language Models, **AAAI 2024**
3. XMoE: Sparse Models with Fine-grained and Adaptive Expert Selection, **ACL 2024**
4. Preparing lessons for progressive training on language models, **AAAI 2024**
5. APrompt: Attention Prompt Tuning for Efficient Adaptation of Pre-trained Language Models, **EMNLP 2023**

Graph Neural Networks

1. Mitigating Over-Squashing in Graph Neural Networks by Spectrum-Preserving Sparsification, **ICML 2025**
2. Sign is Not a Remedy: Multiset-to-Multiset Message Passing for Learning on Heterophilic Graphs, **ICML 2024**
3. Tackling long-tailed distribution issue in graph neural networks via normalization, **TKDE 2023**
4. Predicting Global Label Relationship Matrix for Graph Neural Networks under Heterophily, **NeurIPS 2023**
5. Self-supervised graph attention networks for deep weighted multi-view clustering, **AAAI 2023**

Our Team

Team Members

Prof. Zenglin Xu

Junfan Li
FL, OL

Jinglong Luo
FL, SMC

Fangfei Lin
NLP, Clustering

Zhuo Zhang
FL, LLMR

Dun Zeng
FL

Jiaxiang Chen
LLMR

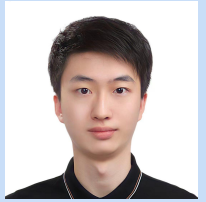
Zheshun Wu
FL

Zhuo Wang
LLMR

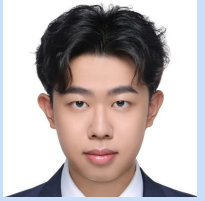
Yingqi Hu
FL

Mengna Hu
TSF

Guanzhong Chen
FL

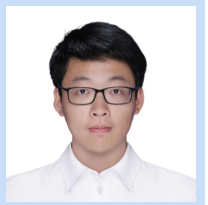


Yunkun Zhang
FL, MLLM



Mingxi Zou
LLMR

Haozhe Shan
RAG



Aotian Luo
MMR

Cunjie He
TSF, NLP

FL: Federate Learning, OL: Online Learning, SMC: Secure Multi-party Computation, LLMR: LLMs Reasonning, TSF: Time Series Forecasting, MLLM: Multi-modal LLMs

Tutorial Speakers

Irwin King
The Chinese University of Hong Kong, Hong Kong, China

Zenglin Xu
Fudan University, Shanghai, China

Dianzhi Yu
The Chinese University of Hong Kong, Hong Kong, China

Guanzhong Chen
Harbin Institute of Technology, Shenzhen

Yukun Zhang
Harbin Institute of Technology, Shenzhen

Zhuo Wang
Fudan University, Shanghai, China

Xinni Zhang
The Chinese University of Hong Kong, Hong Kong, China

News

- Autonomous AI driving is the future, with **multimodal models and data,** such as vision and sensor data
- **A failed example**: Tesla on autopilot crashes into overturned truck
- We need **further improvement on multimodal models**, for auto driving and other applications

Tesla on autopilot crashes into overturned truck

Images Credit: https://www.reddit.com/r/SelfDrivingCars/comments/guoexg/tesla_on_autopilot_crashes_into_overturned_truck/

Issues

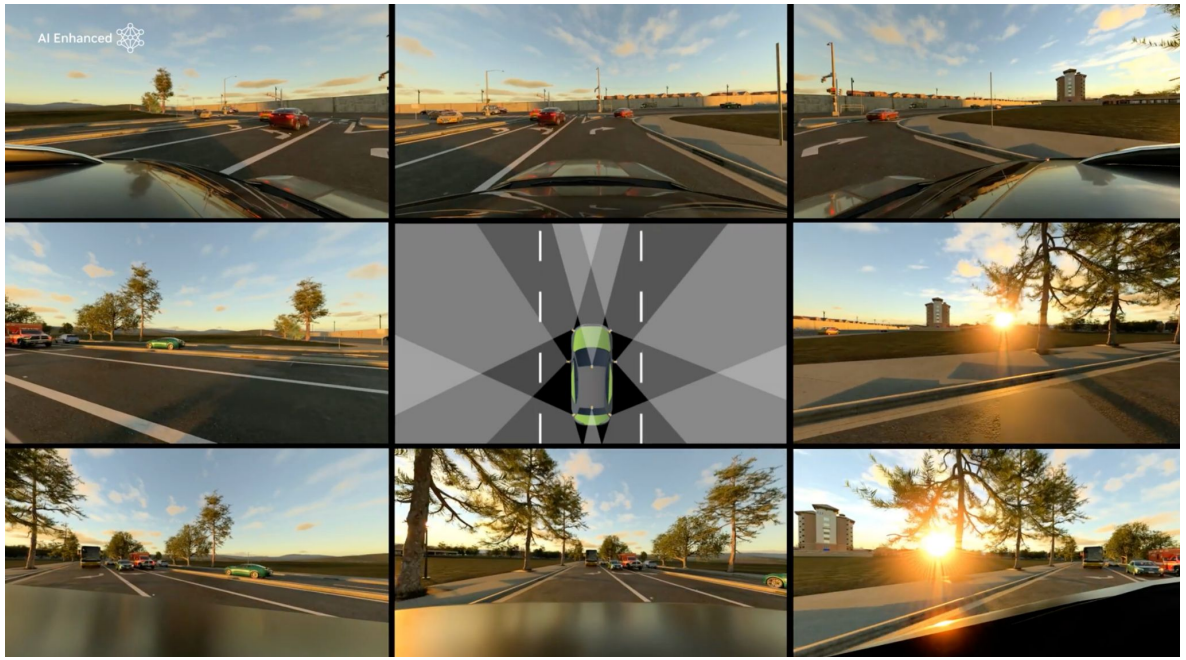
- **Insufficient adaptation to the environment changes on the road**
- **Lack of robustness to deal with various situations**

Tesla on autopilot crashes into overturned truck

Images Credit: https://www.reddit.com/r/SelfDrivingCars/comments/guoexg/tesla_on_autopilot_crashes_into_overturned_truck/

Techniques

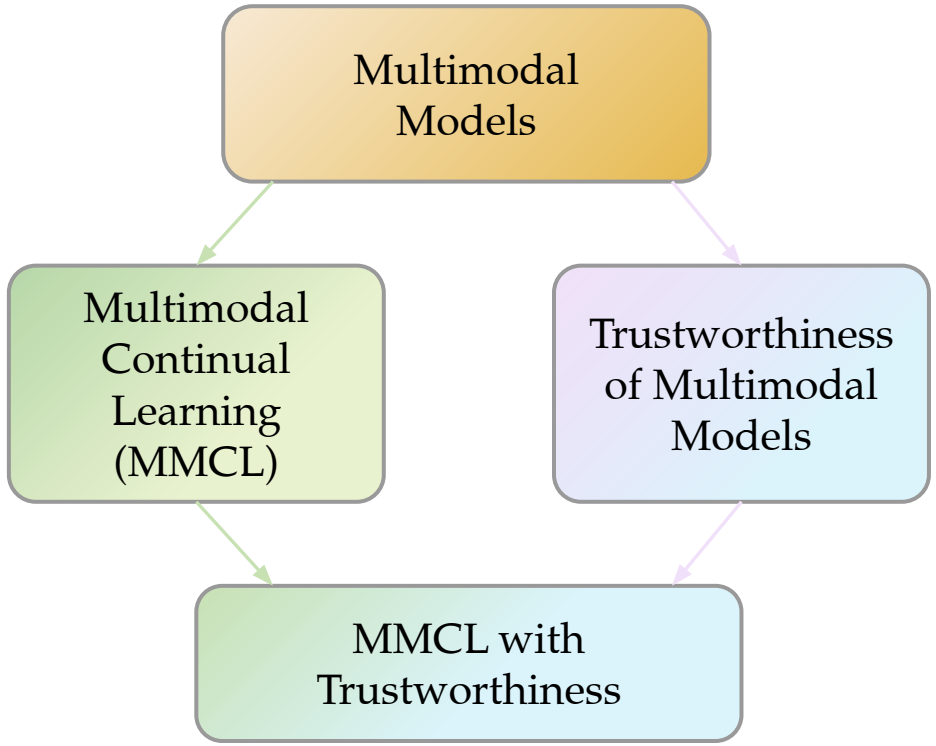
- **Continual learning**
  - Continual model adaptation for autonomous driving models to adapt to evolving and new environments
- **Trustworthiness**
  - Ensure privacy protection and robustness for safe consumer use of autonomous driving models



Images Credit: https://blogs.nvidia.com/blog/auto-research-cvpr-2024/

Contents

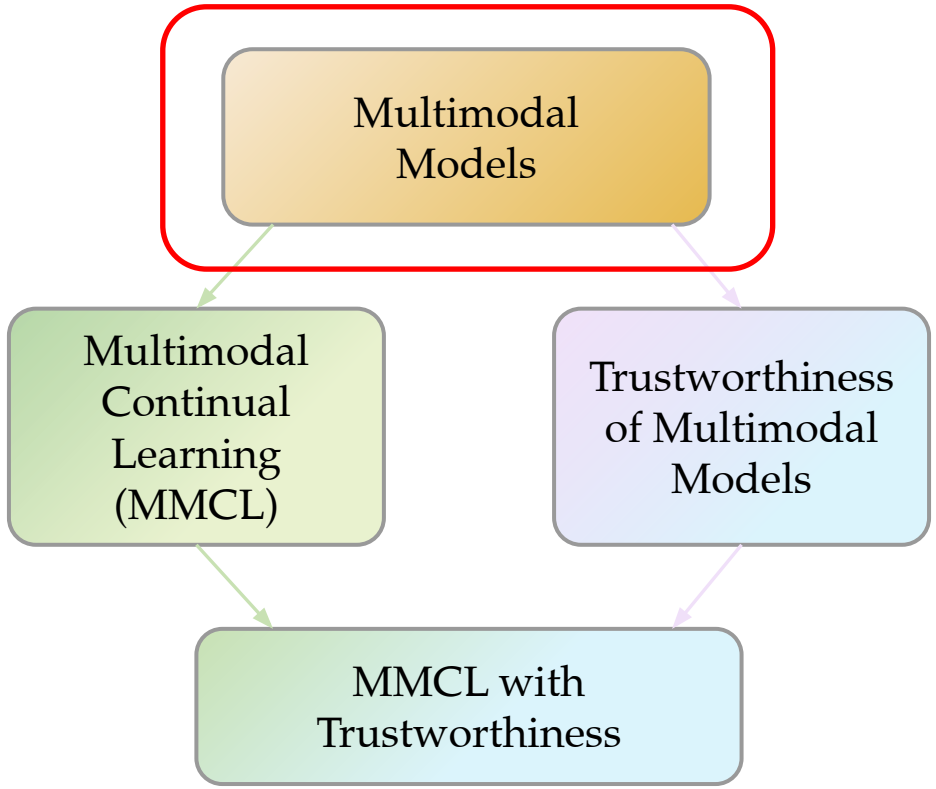
- Introduction to Multimodal Models
- Multimodal Continual Learning (MMCL)
- Trustworthiness of Multimodal Models
- MMCL with Trustworthiness
- Conclusion



Introduction to Multimodal Models

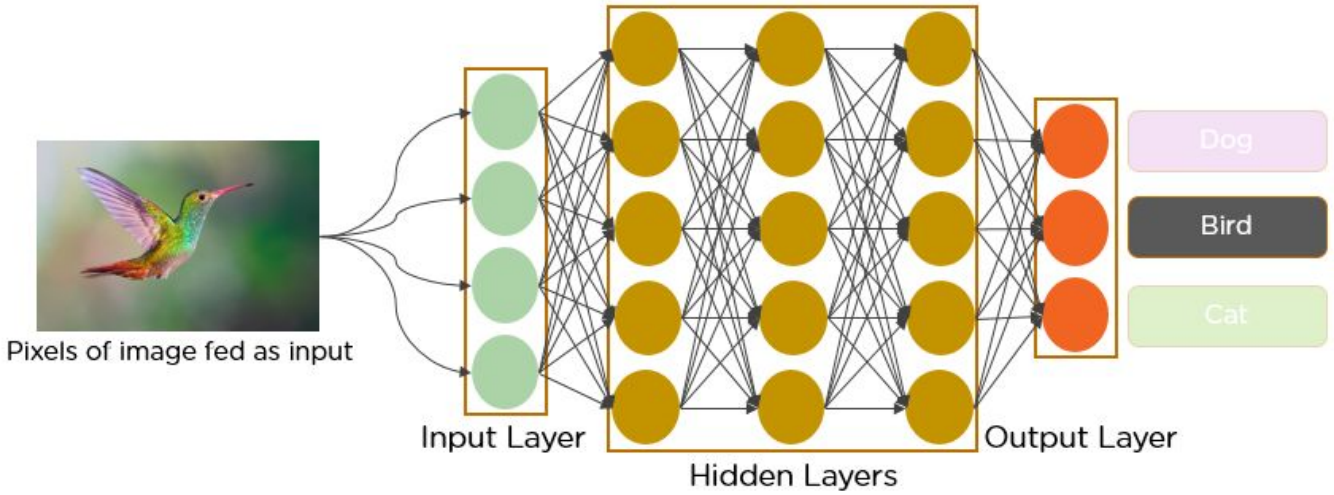
Contents

- Introduction to Multimodal Models
- Definitions of modality and multimodal models
- Overview of multimodal modal architectures
- Challenges of multimodal modals

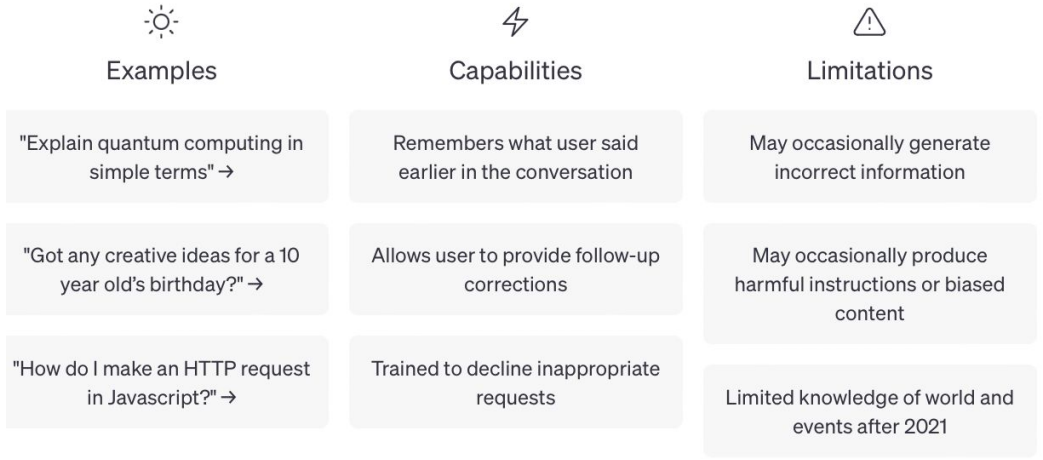


AI: the 5th Industrial Revolution

- Image



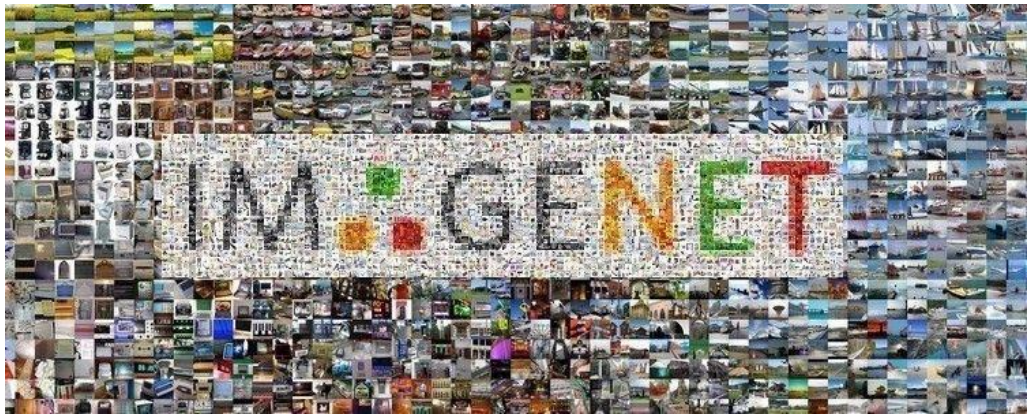
- Language

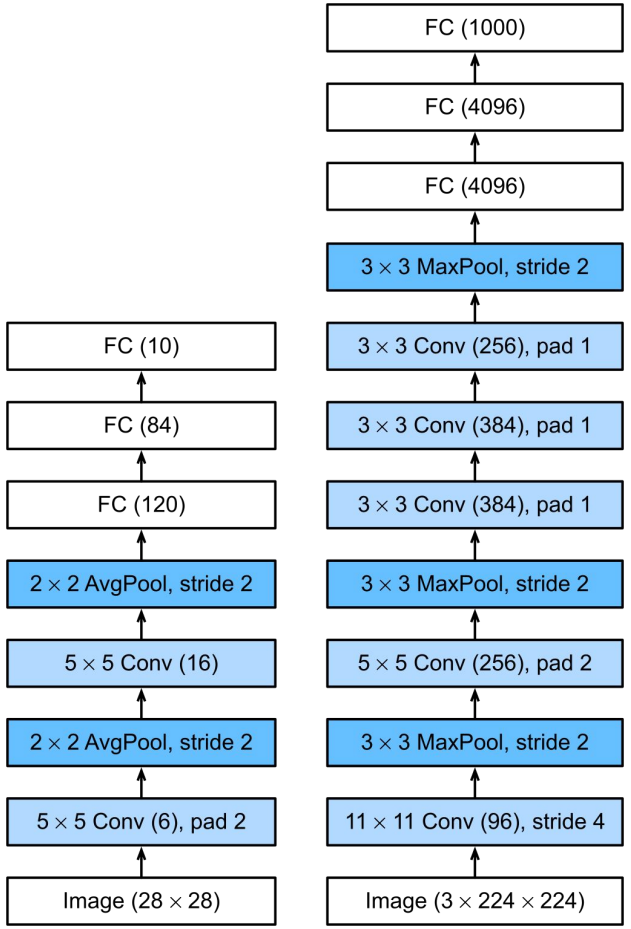


Images Credit: https://www.analyticsvidhya.com/, https://openai.com

Computer Vision

- 2012: AlexNet made a pivotal improvement for computer vision (CV)



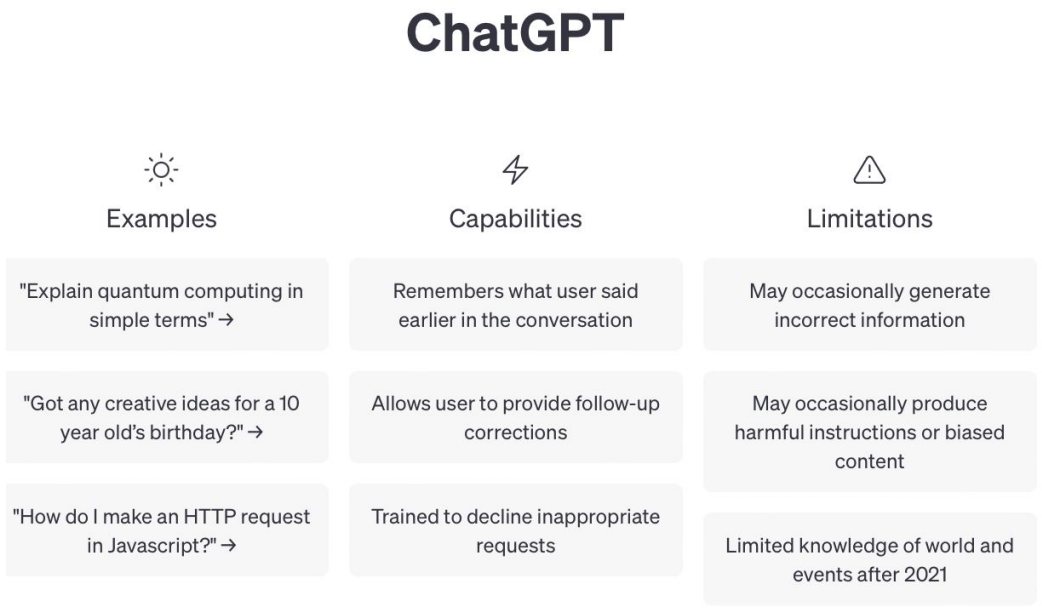


AlexNet block diagram

Image Credit: https://en.wikipedia.org/wiki/AlexNet, https://radekosmulski.com/how-to-train-and-validate-on-imagenet/

Natural Language Processing

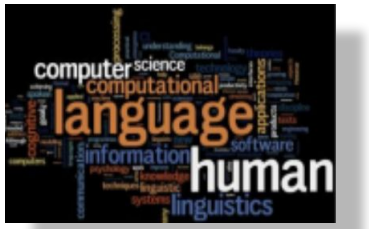
- 2022: ChatGPT was released as a groundbreaking model in natural language processing (NLP)
- Various large language models (LLM) continue to emerge



Images Credit: https://openai.com

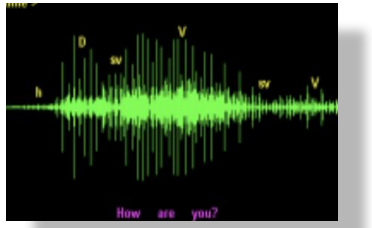
Multimodal Models

- How about a model which can take both text and images as inputs?
- Bridging text, image, audio, and beyond



Language

Vision



Speech

Robotics

Rapid growth of multimodal data

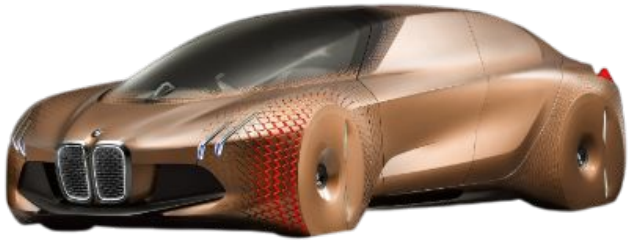
Multimodal Models

Multimodal Models

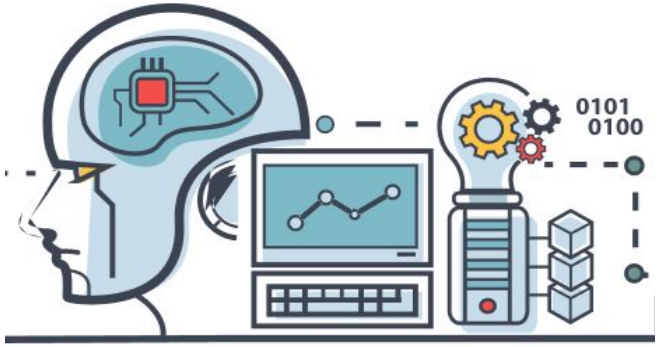
- Multimodal AI Technologies

Robots

Personal Vehicles



Healthcare



Mobile

Wearable

Online

Slide credit: Multimodal Machine Learning, Tutorial @ ICML 2023, https://riseapps.co/machine-learning-in-healthcare/

Multimodal Models

- **Ultimate goals**
- Personalized omnipotent robot
  - Omnipotent capabilities
    - Able to assist with a wide range of tasks—learning, working, entertainment, and daily life
  - A lifelong companion
    - Supports and adapts to a user's changing needs throughout every stage of life
  - Privacy & trust by design
    - Strongly committed to protecting personal data and respecting user privacy at all times

Boston Dynamics

Image credit: https://johnkoetsier.com/boston-dynamics-the-golden-age-of-robotics/

Ultimate Goals of Multimodal Models

- **Ultimate goals**
- Perfect AI doctor – Dynamic Patient Modeling
- Creative AI – Evolving Content Generation
- Smart Cities – Adaptive Sensor Networks
- **However, existing multimodal models face challenges to achieve these ultimate goals**

Image credit: https://www.hudsonregionalhospital.com/services/robotic-surgery/da-vinci-xi-robotic-surgical-system/

Goal of this tutorial

- Introduce multimodal models
- Present challenges in multimodal models
- Provide a detailed overview of current techniques to address these challenges
- Discuss directions for further research
- **By persistently advancing research along this path, we move closer to achieving the ultimate goals!**

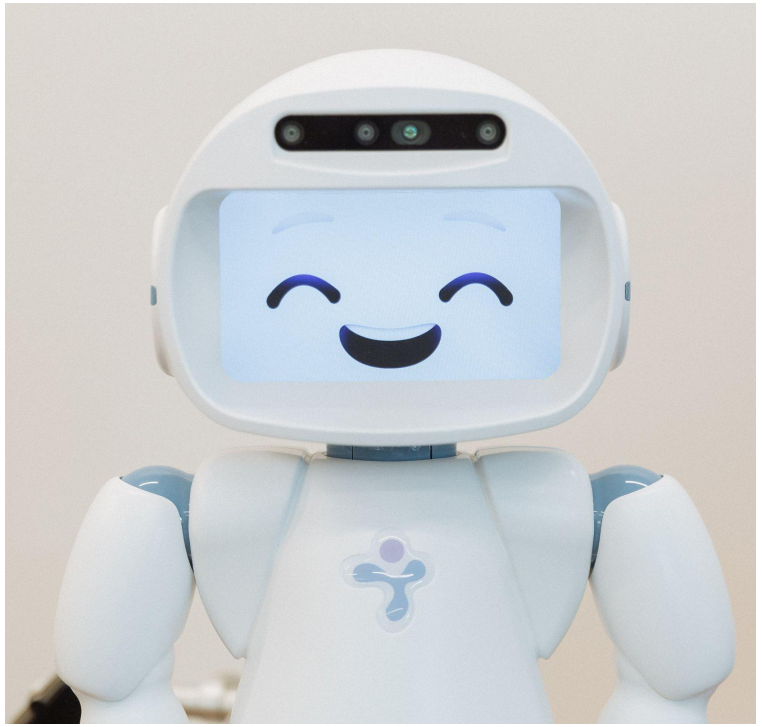


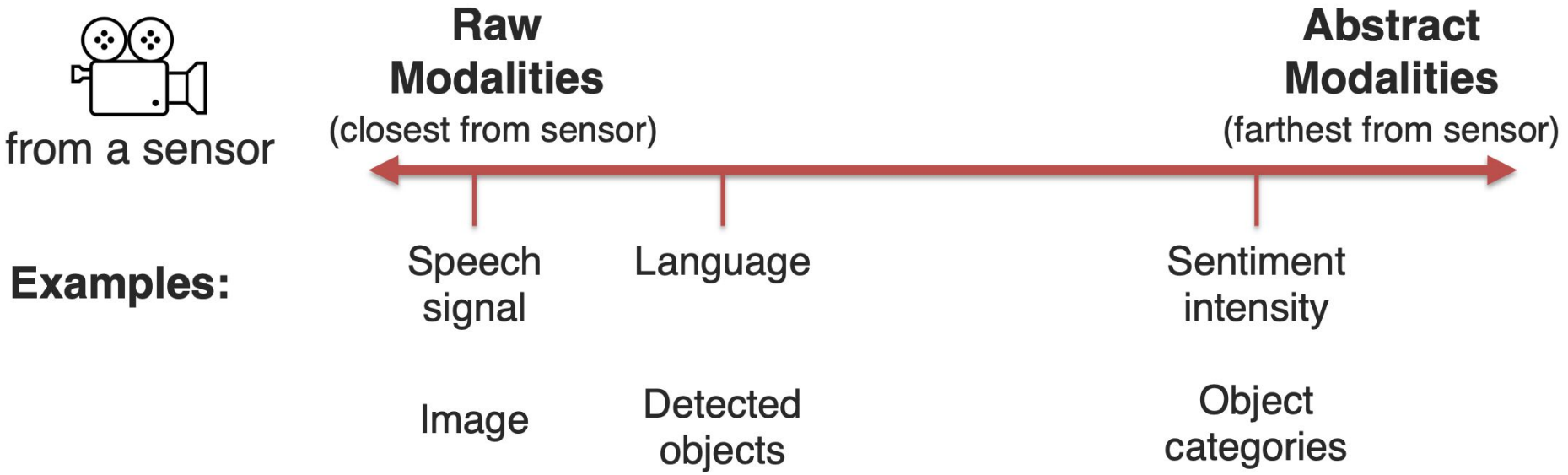
Image credit: https://www.wired.com/story/parents-dementia-robots-warm-technology/

Multimodal Models - Definition

- What is a Modality?

Modality

Modality refers to the way in which something expressed or perceived.



Slide credit: Multimodal Machine Learning, Tutorial @ ICML 2023

Multimodal Models - Definition

Multimodality

With multiple (more than one) modalities.

Multimodal Model

An AI model which is trained on multimodal data.

Multimodal **is the scientific study of heterogeneous and interconnected data**

Connected + Interacting

Overview of Multimodal Modal Architectures

- LLM-based multimodal – Multimodal LLM
- LLMs extended to process multimodal data (text, images, audio, video) simultaneously
- Examples: GPT-4V, Gemini, LLaVA

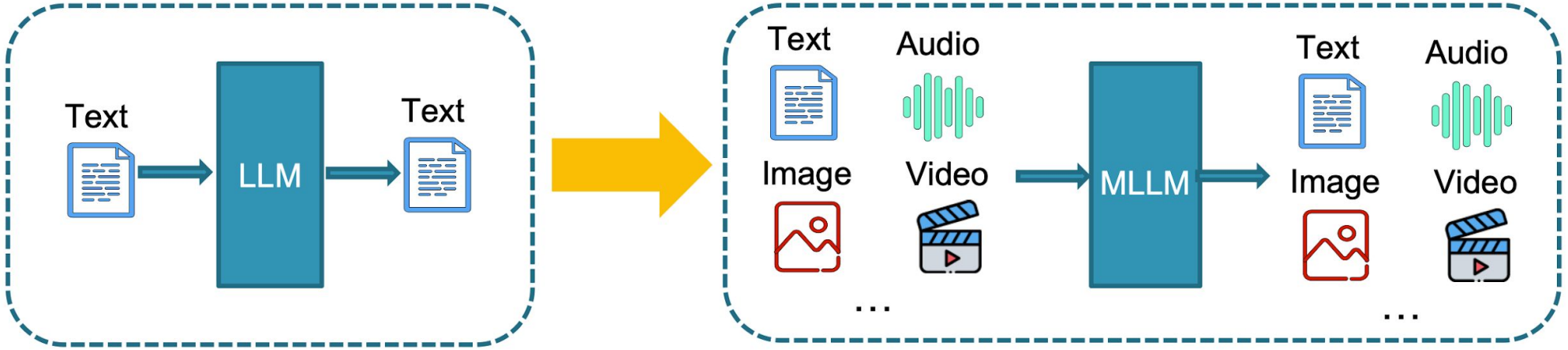


Image Credit: From Multimodal LLM to Human-level AI, Tutorial @ CVPR 2024

Overview of Multimodal Modal Architectures

- Alignment-based multimodal – CLIP

- Align text and images pairs in a shared embedding space
  - Pulling matched pairs closer
  - Pushing mismatched pairs apart

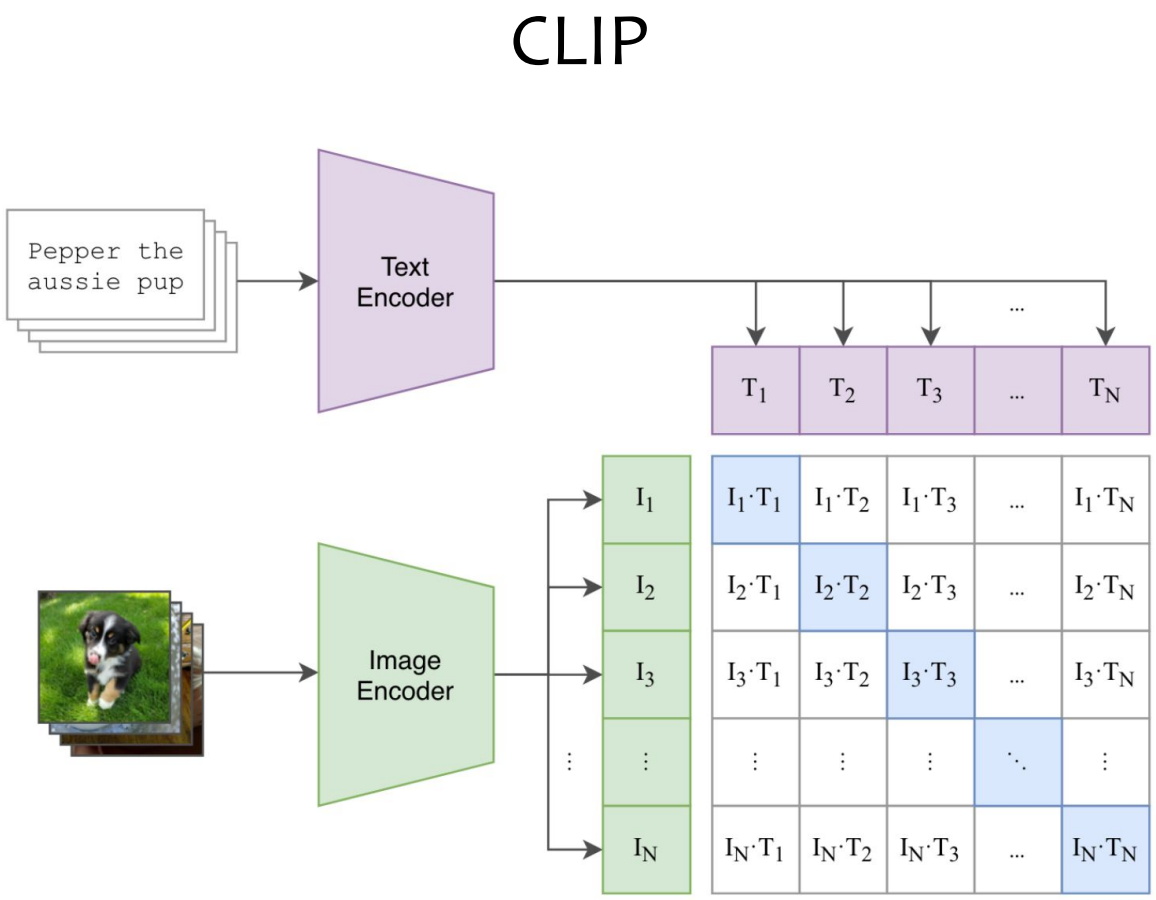


Image credit: Radford et al. Learning transferable visual models from natural language supervision. In ICML, 2021.

Challenges of Multimodal Models

- **Issue 1**
  - The world is **dynamically evolving**

- We want AI to
  - adapt to **new data**
  - perform **new tasks**

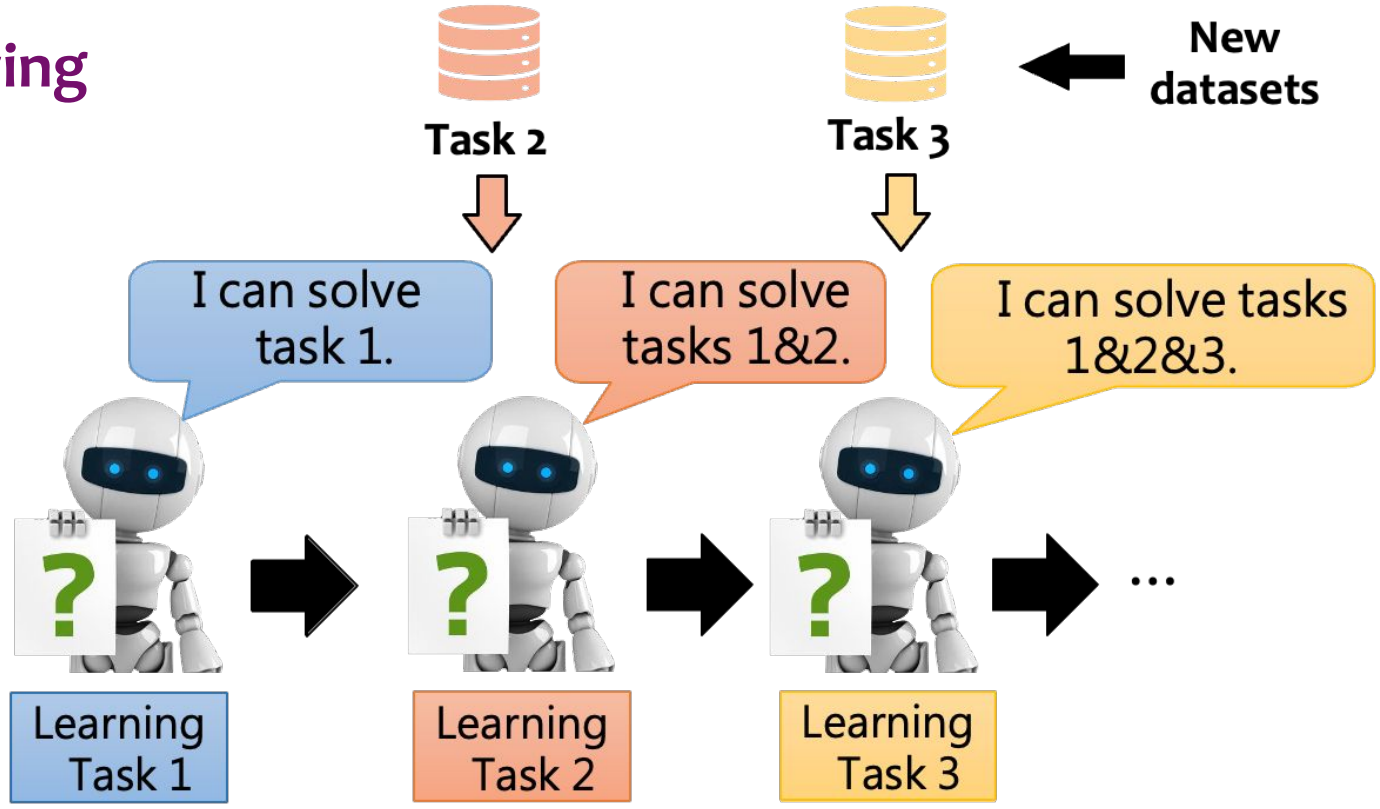


Image credit: Hung-yi Lee, https://www.youtube.com/watch?v=rWF9sg5w6Zk&ab_channel=Hung-yiLee.

Challenges of Multimodal Models

- Example: We want to use CLIP to perform image classification on new image datasets from different domains
- A naive approach: directly finetune the model on new datasets

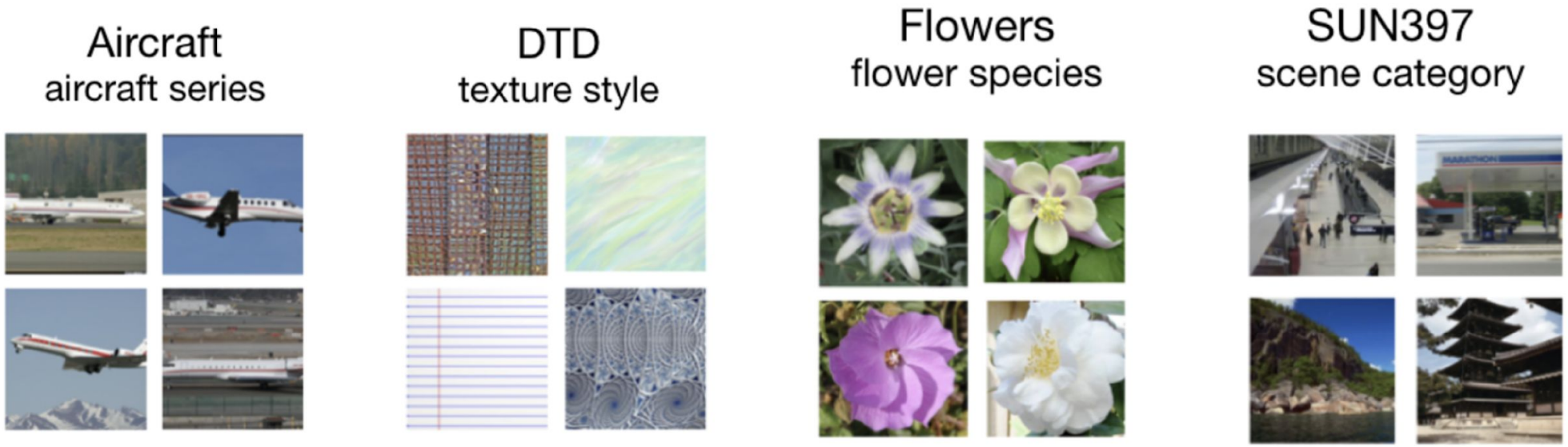
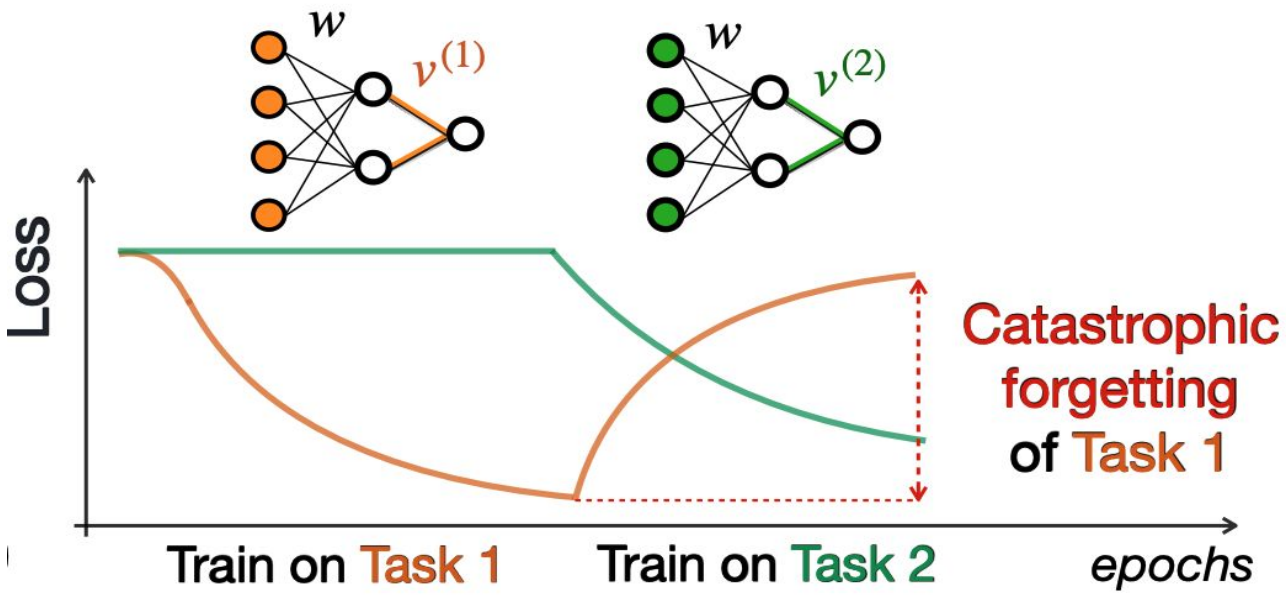
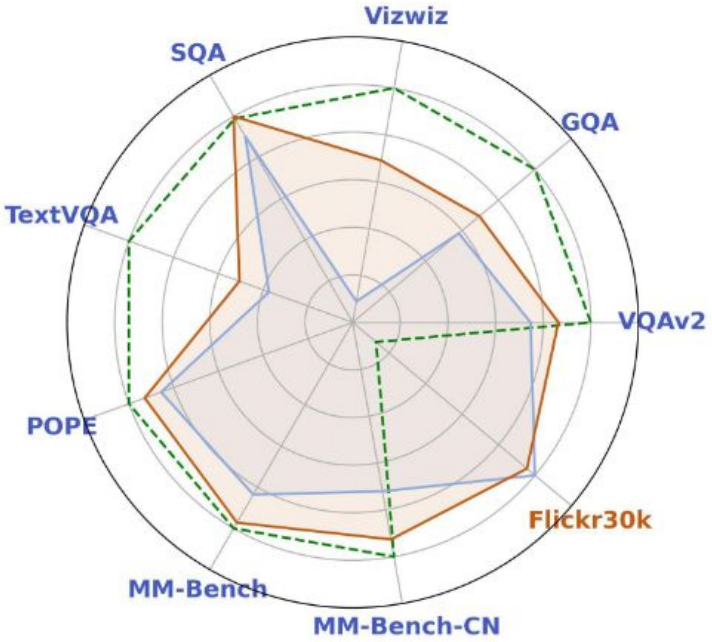


Image credit: Zheng et al., Preventing Zero-Shot Transfer Degradation in Continual Learning of Vision-Language Models, ICCV 2024.

Challenges of Multimodal Models

- **Challenge 1: Catastrophic forgetting** in multimodal models using direct fine-tuning in both new models and pretrained models
- When tasks are trained sequentially, training on the new task **greatly disrupts performance** on previously learned tasks



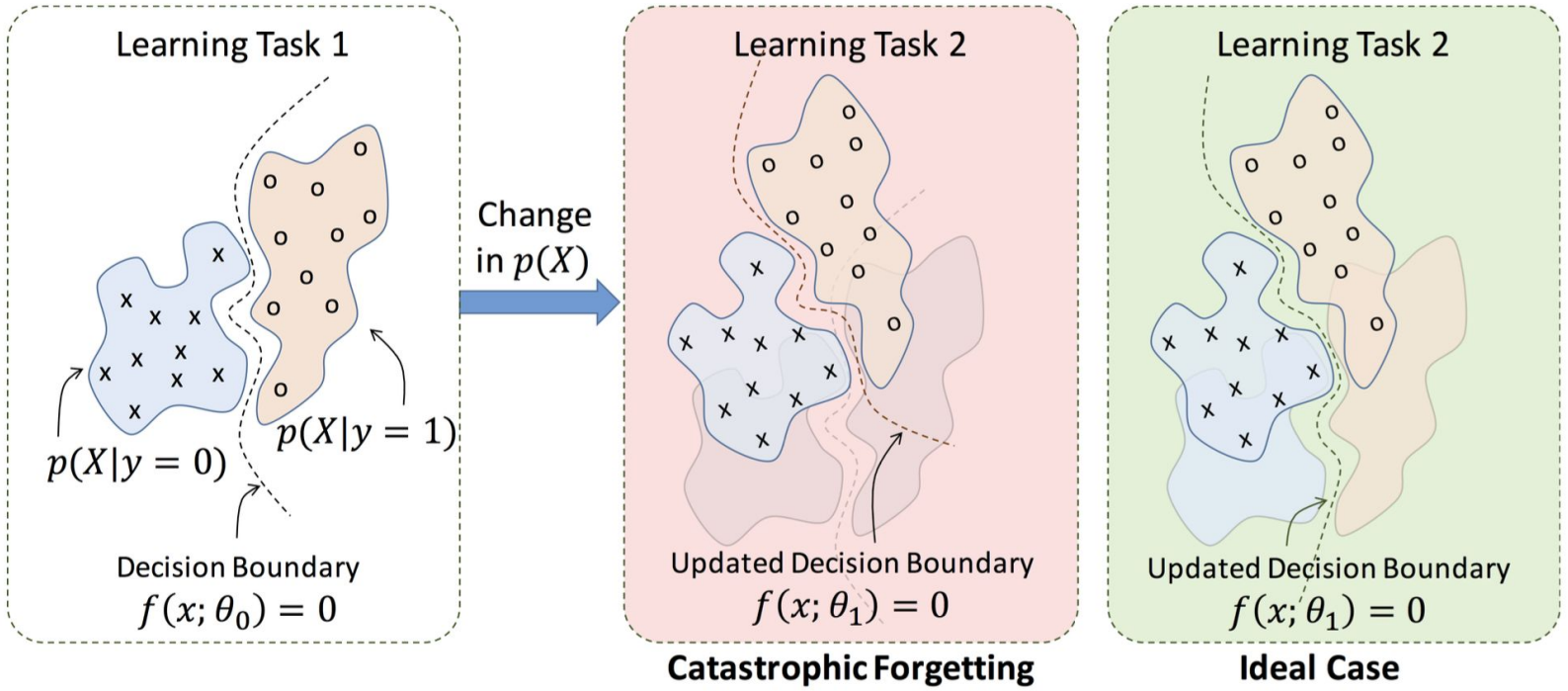


McCloskey and Cohen, "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem", 1989, In Psychology of learning and motivation (Vol. 24, pp. 109-165). Academic Press.
Image credit: Mori et al. "Optimal protocols for continual learning via statistical physics and control theory." arXiv preprint arXiv:2409.18061 (2024).
Zhu et al., Model tailor: Mitigating catastrophic forgetting in multi-modal large language models, ICML 2024

Challenges of Multimodal Models

- **Catastrophic forgetting**
- Reason: Unconstrained fine-tuning drives the uniformly plastic parametric model to adapt fully to the new distribution

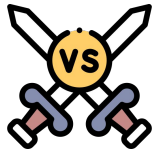
Depiction of catastrophic forgetting in binary classification tasks when there is a distribution shift from an initial task to a secondary task.



Hassabis et al.. Neuroscience-Inspired Artificial Intelligence. In: Neuron 95.2 (July 2017)
Image credit: Kolouri, Soheil, et al. "Attention-based selective plasticity." arXiv preprint arXiv:1903.06070 (2019).

Challenges of Multimodal Models

- **Issue 2 – Need of trustworthiness**
  - Public concern for model safety and privacy increases
  - Governments impose more related regulations (e.g., GDPR,CCPA)



- We want AI to ensure
  - **Explainability**
  - **Privacy & Security**
  - **Robustness**
  - **Fairness**

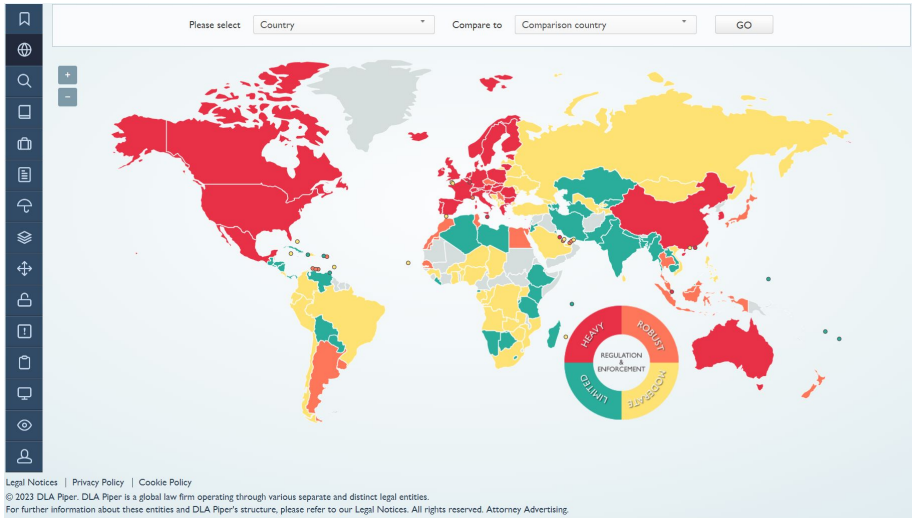


Image credit: https://www.interviewbit.com/blog/big-data-projects/, https://www.dlapiperdataprotection.com/

Challenges of Multimodal Models

- **Challenge 2: Untrustworthiness** of existing multimodal models

- Explainability
  - Make the model white box
  - Force the model to self-explain
  - Model-specific vs -agnostic explanation

- Privacy & Security
  - Attacks
  - Data leakage
  - Model inversion

- Robustness
  - Train-Test modality mismatch
  - Adversarial attack
  - Jailbreaking

- Fairness
  - Stereotypes & gender Bias
  - Regional / cultural Bias
  - Linguistic inequality
  - Racial bias

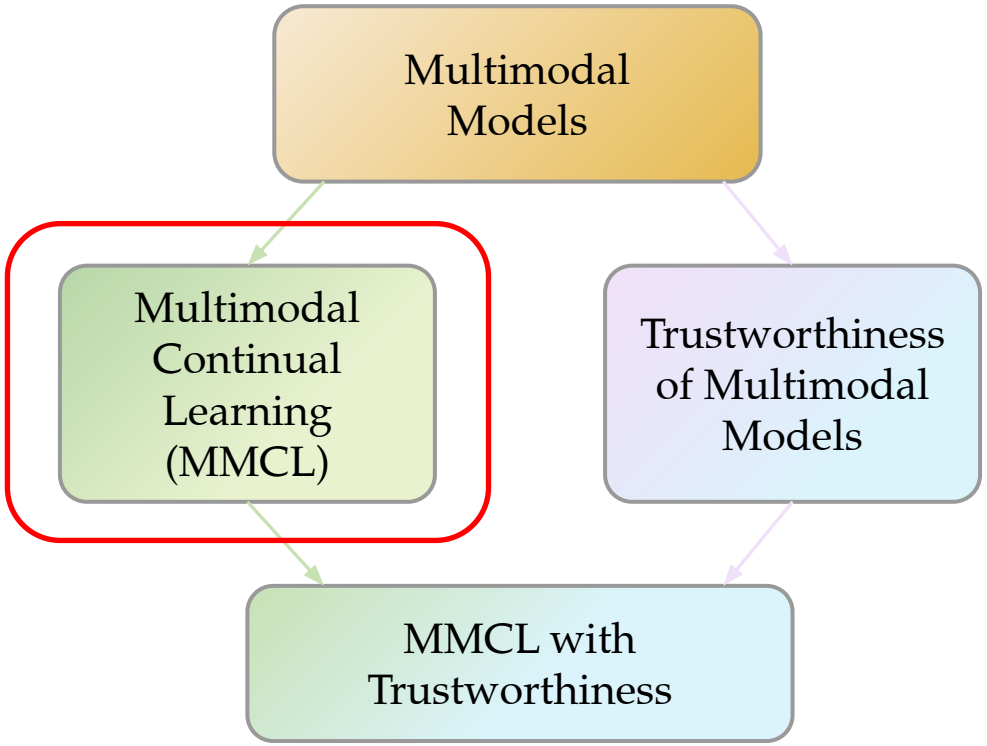
Summary of Introduction

- Definitions of modality and multimodal models
- Overview of multimodal modal architectures
  - LLM-based
  - Alignment-based
- Challenges of multimodal modals
  - Challenge 1: Catastrophic forgetting using direct finetuning
  - Challenge 2: Untrustworthiness of existing multimodal models

Multimodal Continual Learning (MMCL)

Contents

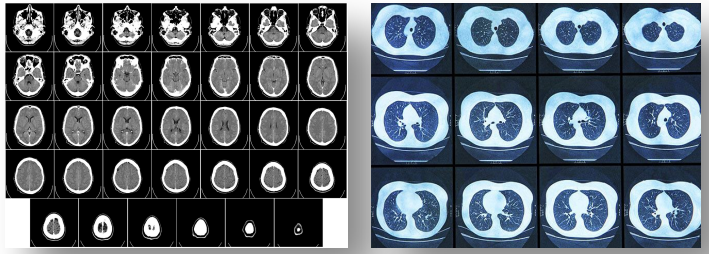
- Multimodal Continual Learning
- Introduction of MMCL
- Challenges of MMCL
- Taxonomy of MMCL
- Open Source Toolkit for MMCL
- Future directions



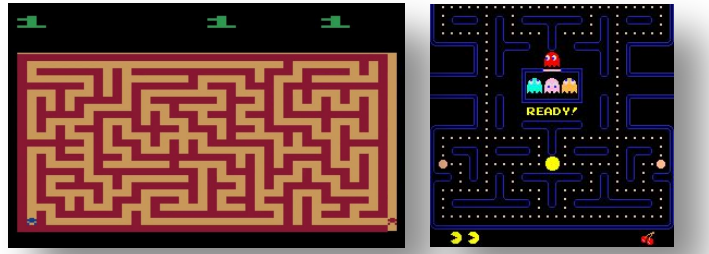
Introduction to Continual Learning

- **Most AI models**: Focus on one task

One kind of chess

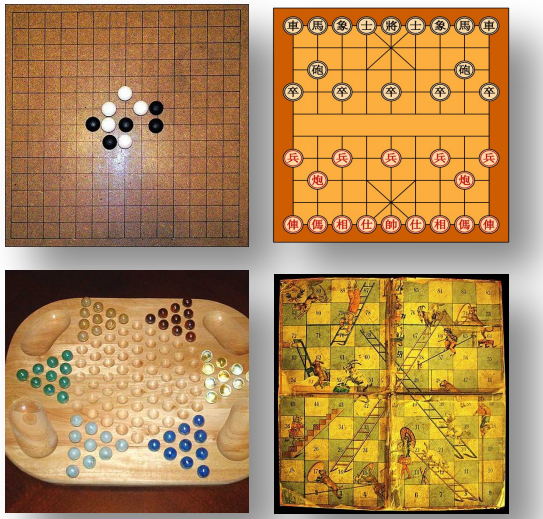


One domain of image



One kind of game

- Real-world scenarios require models to handle multiple tasks or domains



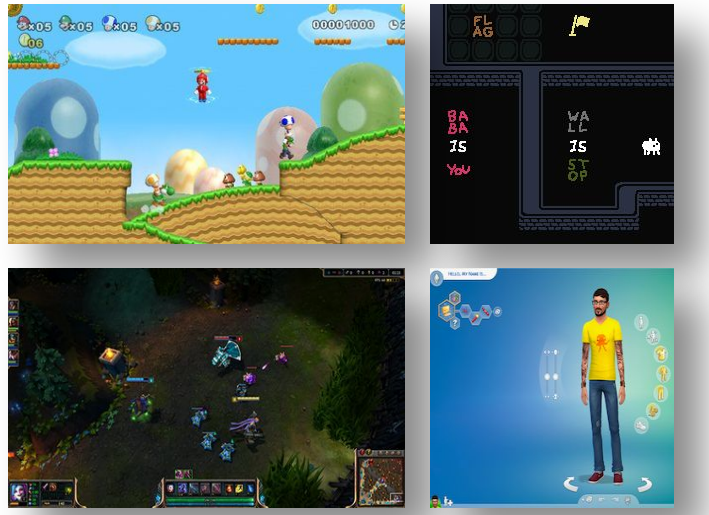


Image credit: wikipedia, gamefaqs, bowtie

Introduction to Continual Learning

- Conventional settings
- "**Single-episode**" paradigm
- Trained on *static* and *single* datasets

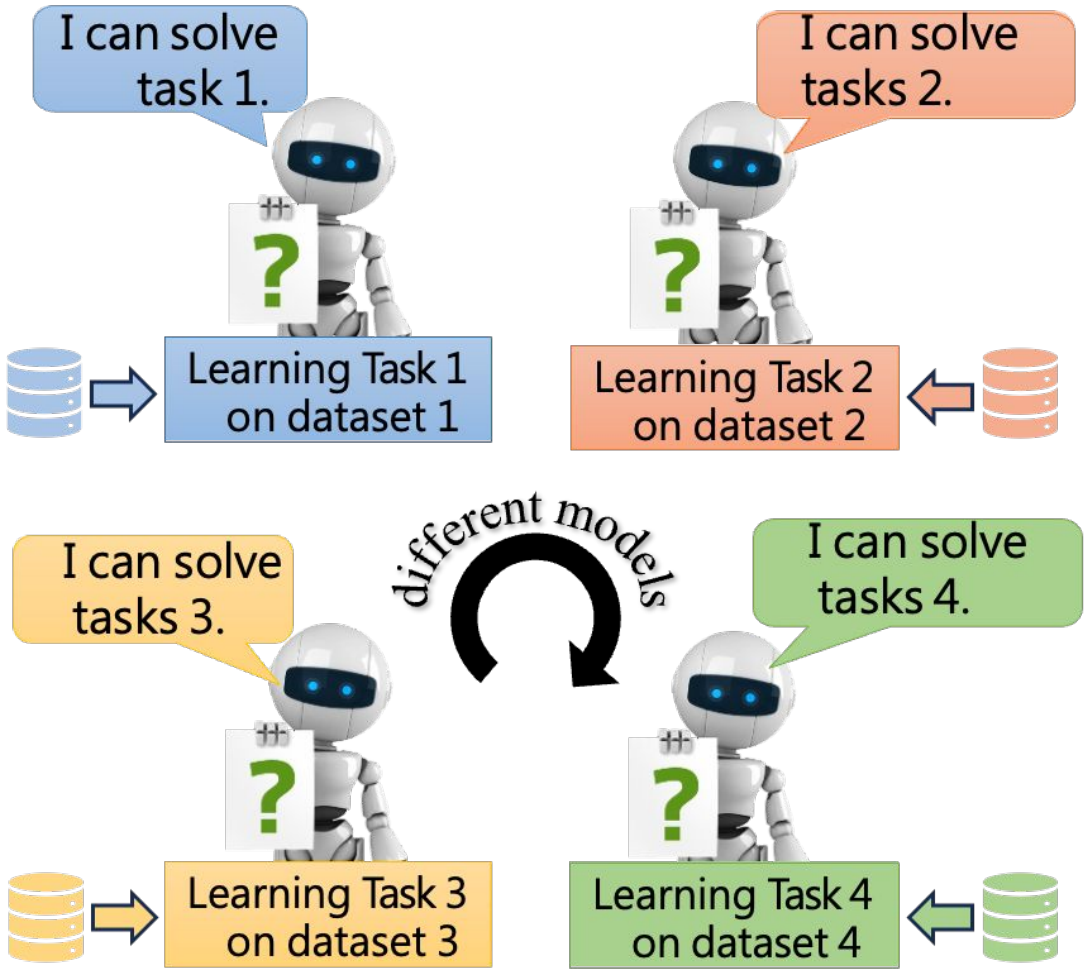
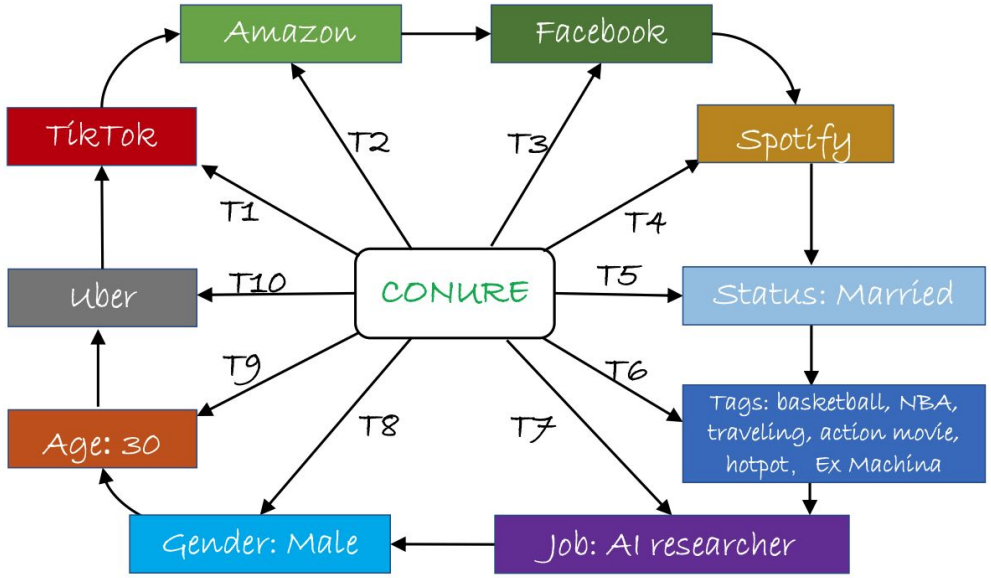


Image credit: Hung-yi Lee, https://www.youtube.com/watch?v=rWF9sg5w6Zk&ab_channel=Hung-yiLee.

Introduction to Continual Learning

- Train once
- Model performance degradation
  - on new data with different distributions
  - out-of-distribution data



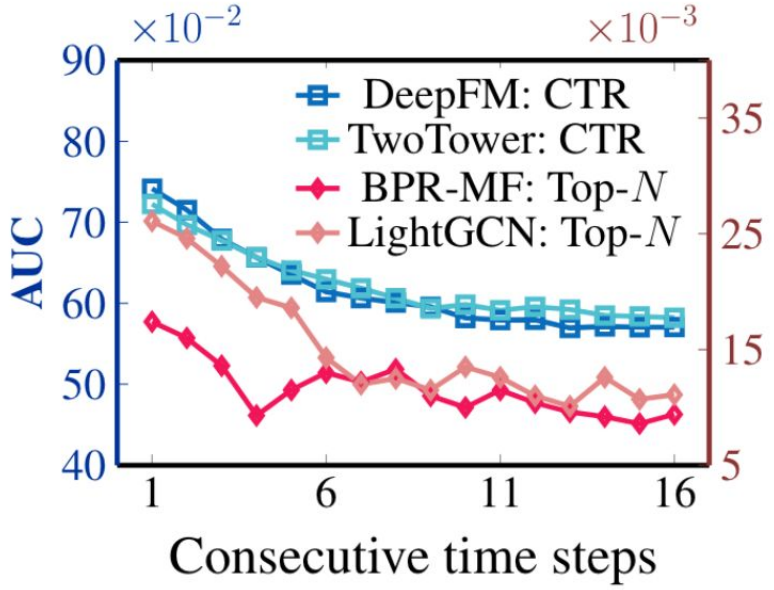
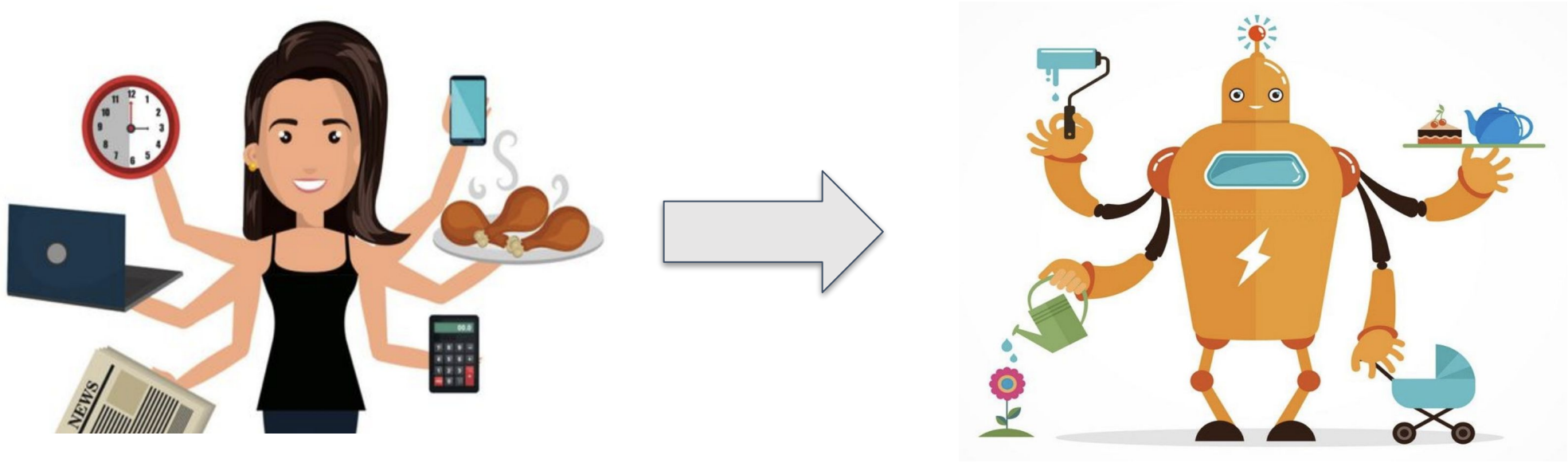


Image credit: Yuan, Fajie, et al., "One person, one model, one world: Learning continual user representation without forgetting", SIGIR. 2021
Zhang et al., “Influential Exemplar Replay for Incremental Learning in Recommender Systems”, AAAI 2024

Introduction to Continual Learning

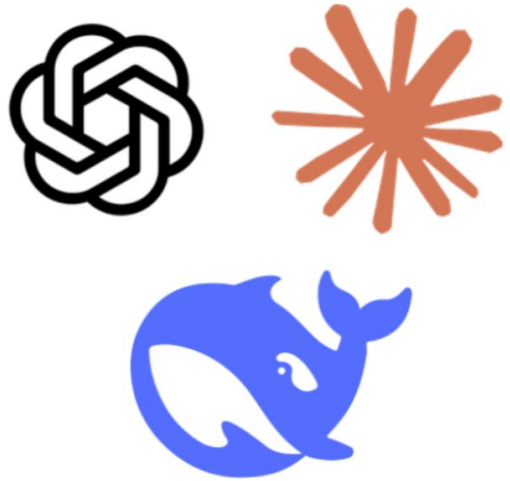
- **Human-Level AI**: Versatile, performs multiple tasks



Slide credit: Continual Learning with Deep Architectures, Tutorial @ ICML 2021

Introduction to Continual Learning

- Are foundation models enough?
  - Not quite
  - A fixed model cannot cope with this dynamically evolving world
  - Example: ChatGPT's knowledge cutoff date is September 2021



Slide credit: Continual Learning with Deep Architectures, Tutorial @ ICML 2021

Introduction to Continual Learning

- Solution: **Continual Learning**

Image credit: https://world.edu/lifelong-learning-part-time-undergraduate-provision-crisis/

Introduction to CL - Biological Mechanisms

- Biological mechanisms that support Continual Learning

- **Neurogenesis**
  - Generation of new neurons in brains throughout life
  - Increased neurogenesis is associated with better adaptation to new environments

Continual Learning with Deep Architectures, Tutorial @ ICML 2021
Image credit: Kudithipudi, Dhireesha, et al. "Biological underpinnings for lifelong learning machines." Nature Machine Intelligence 4.3 (2022): 196-210.

Introduction to CL - Biological Mechanisms

- Biological mechanisms that support Continual Learning

- **Episodic replay**
  - Hippocampus: fast learning of episodic information
  - Knowledge consolidated to the neocortex via "replay" of neural activity episode patterns

Continual Learning with Deep Architectures, Tutorial @ ICML 2021
Image credit: Kudithipudi, Dhireesha, et al. "Biological underpinnings for lifelong learning machines." Nature Machine Intelligence 4.3 (2022): 196-210.

Introduction to CL - Biological Mechanisms

- Biological mechanisms that support Continual Learning

- **Plasticity**
  - Learning
  - -> enlargement of dendritic spines (neural synapses components)
  - -> decreased plasticity of the synapses

Continual Learning with Deep Architectures, Tutorial @ ICML 2021
Image credit: Kudithipudi, Dhireesha, et al. "Biological underpinnings for lifelong learning machines." Nature Machine Intelligence 4.3 (2022): 196-210.

Introduction to CL - Definition

Continual Learning

Continual learning (CL) is the setting to train models incrementally on new tasks/datasets and maintain early knowledge without requiring full-data retraining, **overcoming catastrophic forgetting**.

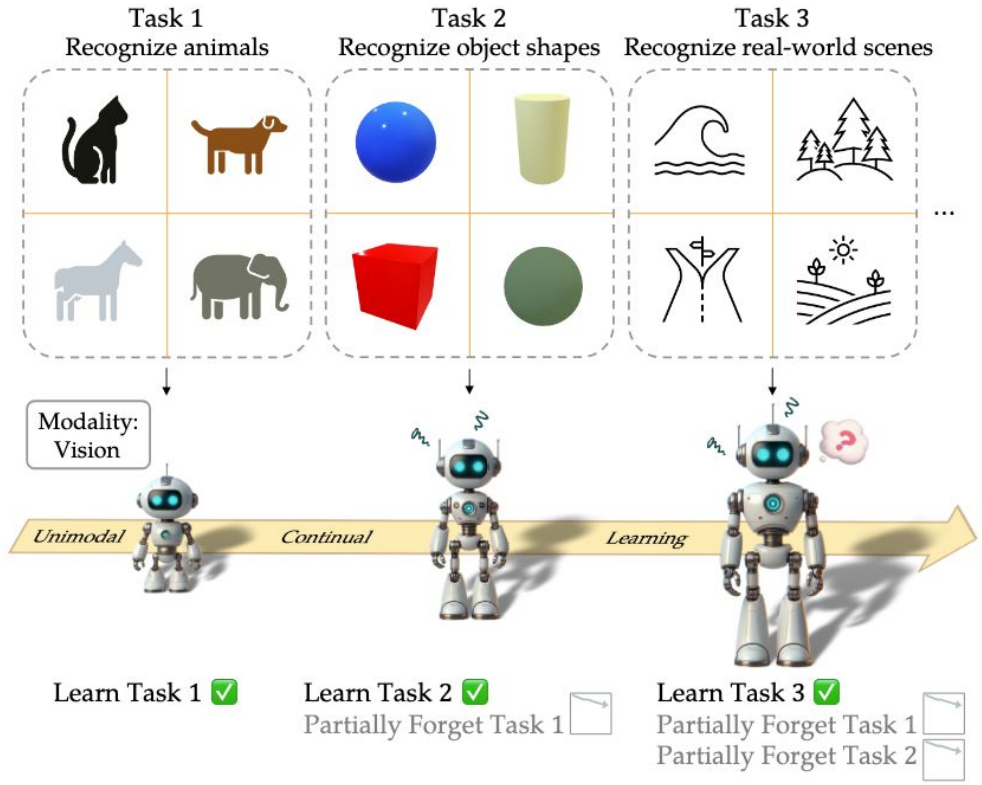


Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Introduction to CL - Definition

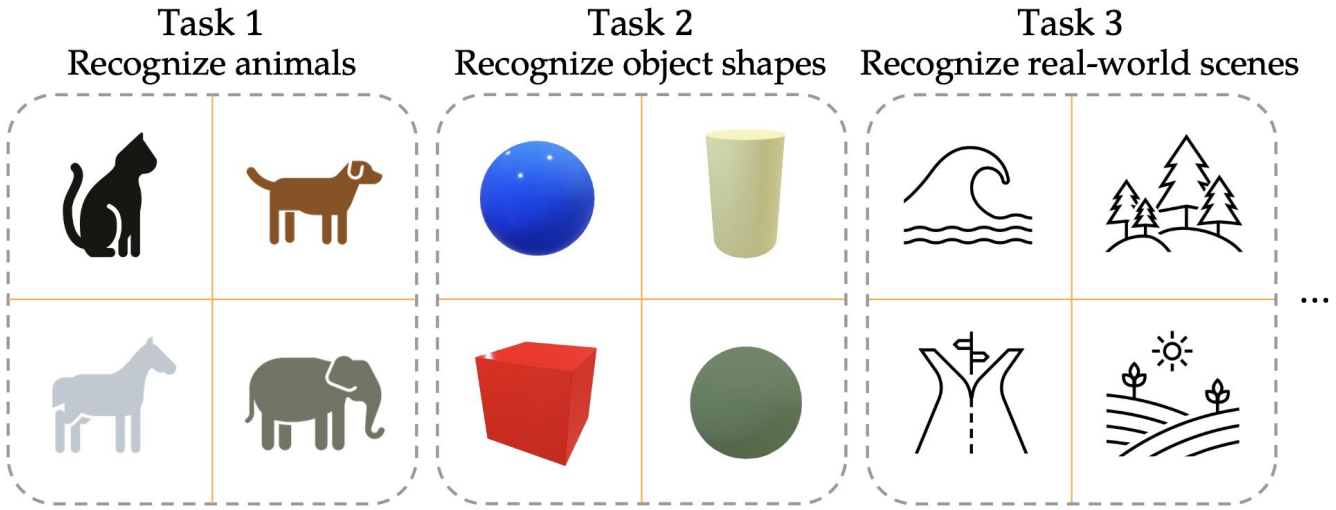
Definition 1 (Task Sequence)

- The **dataset** of the t-th task, denoted as $\mathcal{D}_t$, is defined as:

$$\mathcal{D}_t = \{(\mathbf{x}_{t,i}, y_{t,i}) : i \in \mathbb{N}, 1 \leq i \leq N_t\},$$

- A **task sequence** $\mathcal{TS}$ of size T (where $T > 1$ is required) is a sequence of tasks with their datasets in a certain order, defined as:

$$\mathcal{TS} = [\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_T].$$



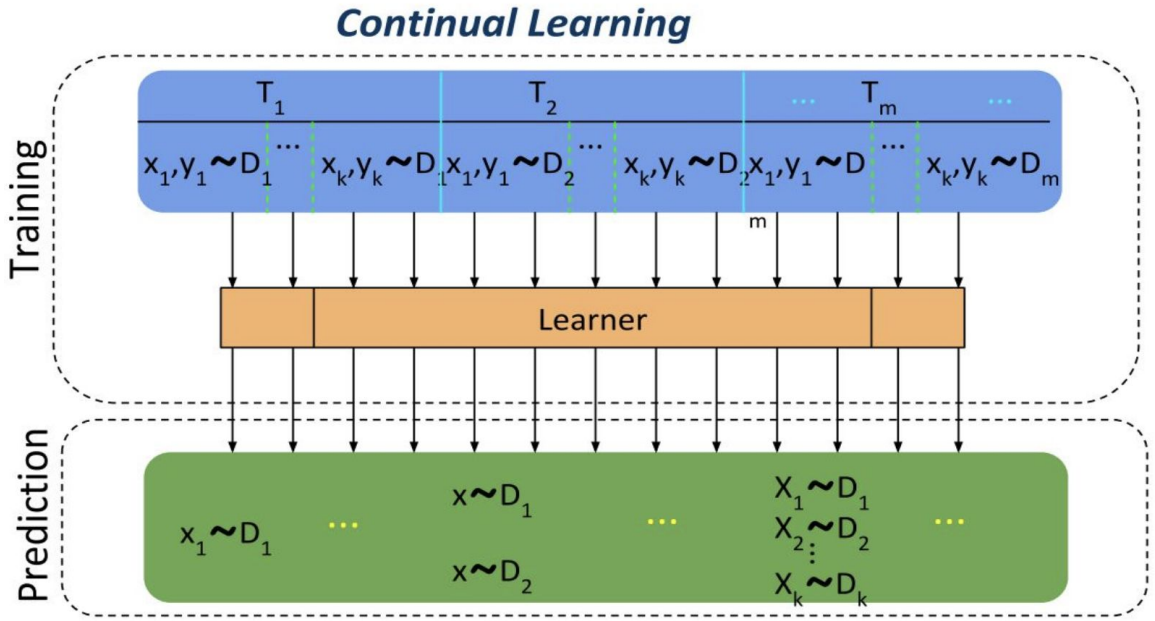
Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Introduction to CL - Definition

Definition 2 (Continual Learning (CL))

- **Continual learning**: for each task t, the model is trained only on data $\mathcal{D}_t$ (or with very limited access to previous datasets in a more relaxed setting).
- **Objective**: learn the new task while maintaining performance on old tasks.

$$\boldsymbol{\theta}^* = argmax_{\boldsymbol{\theta}} \sum_{i=1}^{T} P(\boldsymbol{\theta}, \mathcal{D}_i)$$



Wang et al., Learning to Prompt for Continual Learning, CVPR 2022.
Image credit: de Lange et al. Continual learning: A comparative study on how to defy forgetting in classification tasks, TPAMI 2019.

Continual Learning vs Other Learning Paradigms

CL vs Standard Supervised Learning

- **Static Dataset**: Training data is available all at once, and the model learns from a fixed dataset

- **Fixed Task**: Models are trained for a specific task and do not adapt to new tasks over time

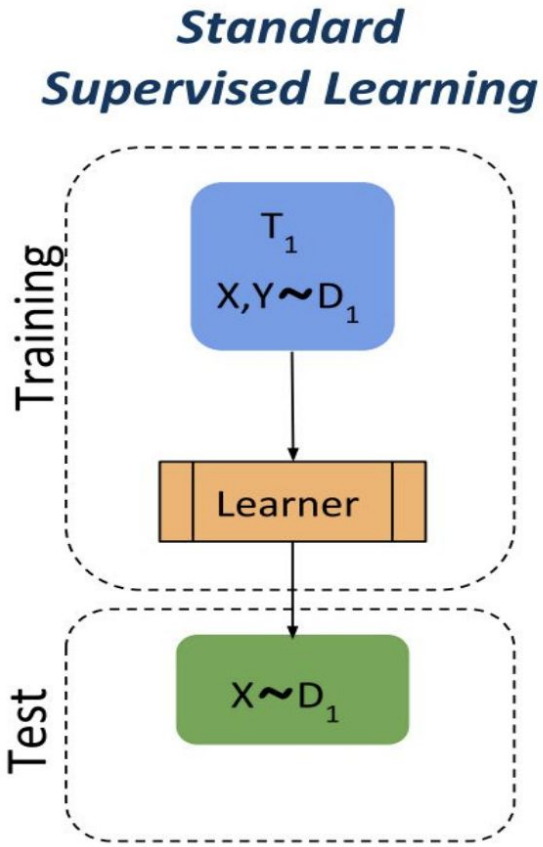
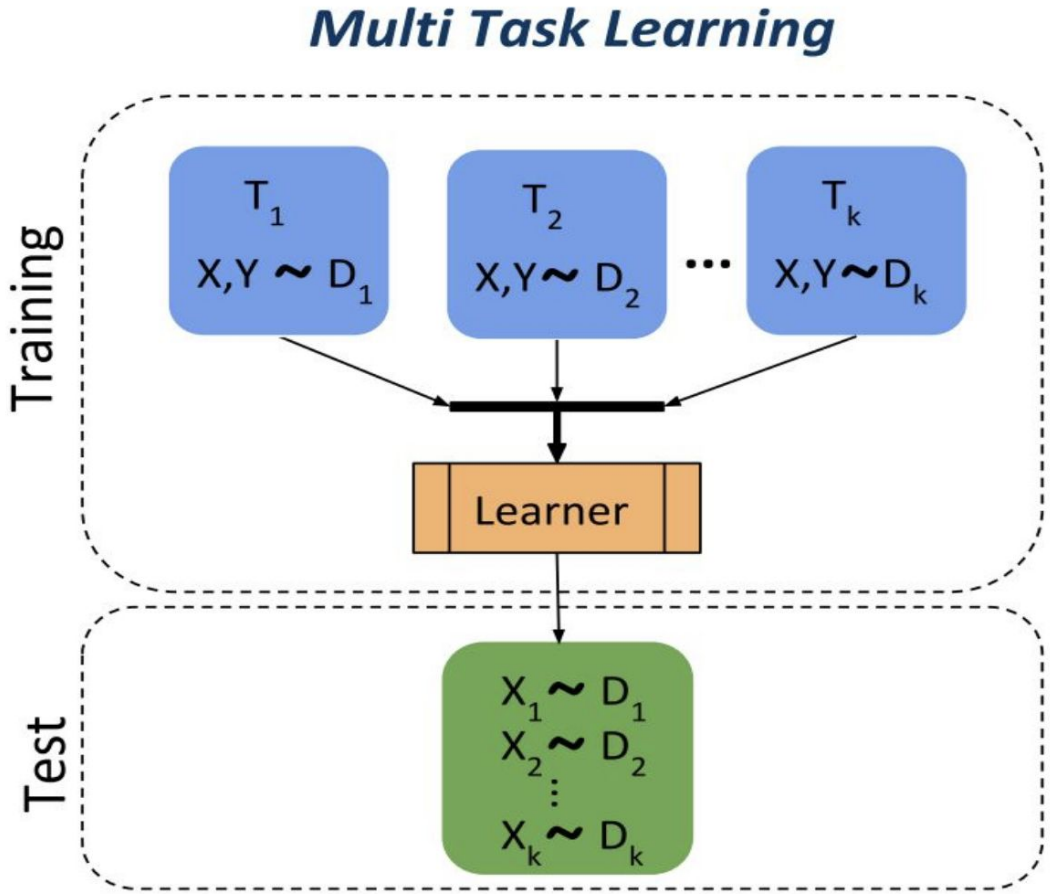


Image credit: de Lange et al. Continual learning: A comparative study on how to defy forgetting in classification tasks, TPAMI 2019.

Continual Learning vs Other Learning Paradigms

CL vs Multi-Task Learning

- Learning of multiple related tasks **offline**, simultaneously
- Using a set or subset of shared parameters
- **No continual model adaptation**

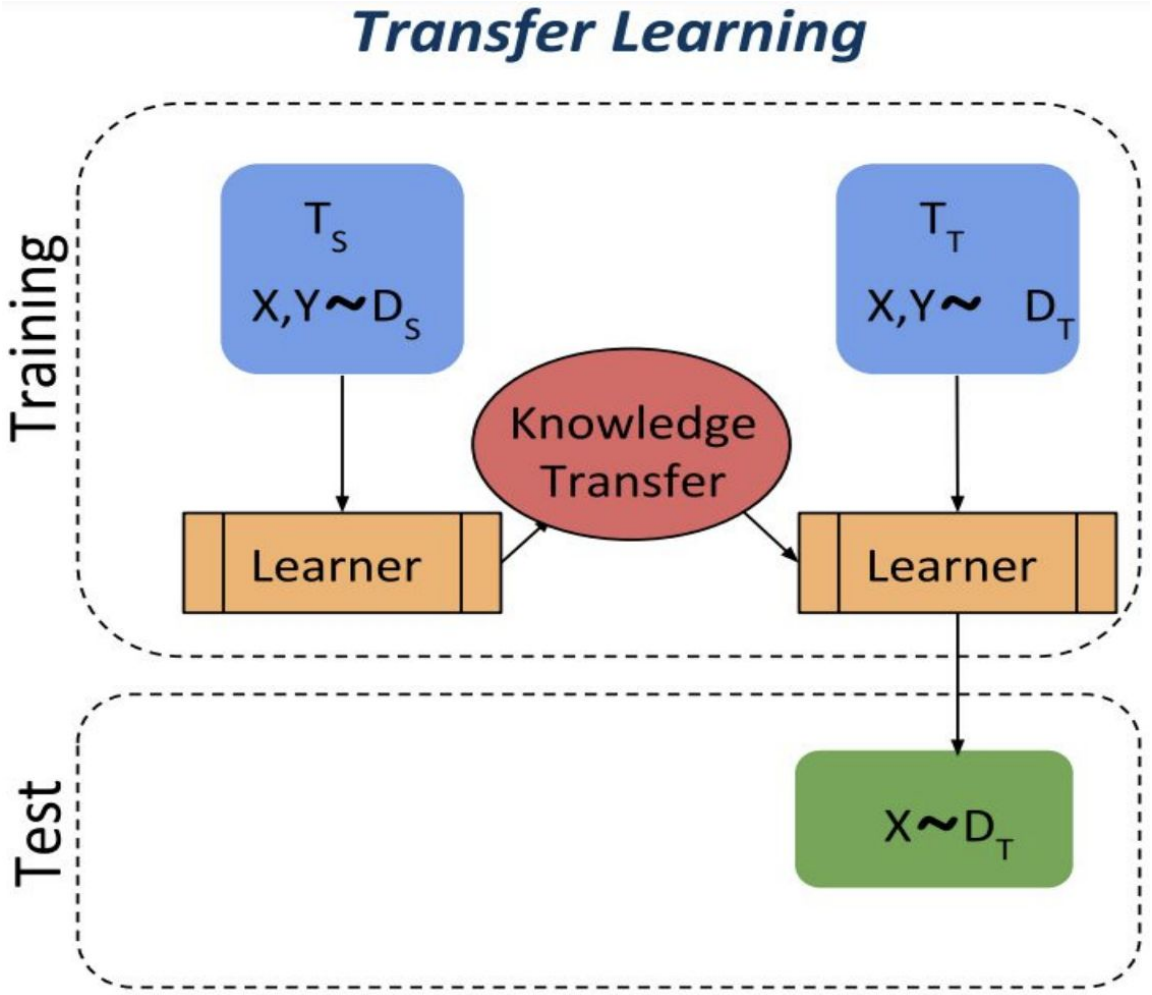


Slide credit: Continual Learning with Deep Architectures, Tutorial @ ICML 2021
Image credit: de Lange et al. Continual learning: A comparative study on how to defy forgetting in classification tasks, TPAMI 2019.

Continual Learning vs Other Learning Paradigms

CL vs Transfer Learning

- Help learning the target task using model trained on the source task
- **No continuous adaptation** after learning the target task
- Performance on the source task(s) is **not taken into account**

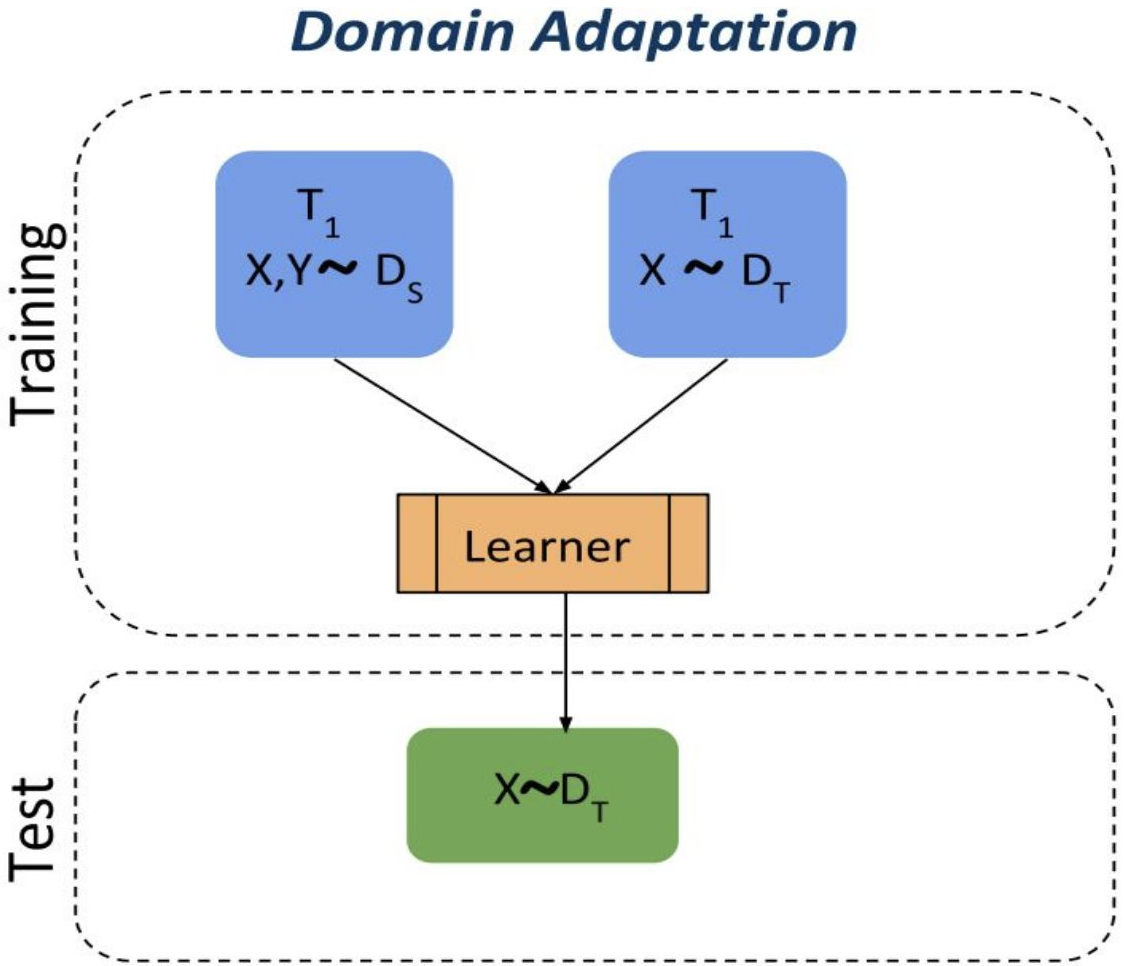


Slide credit: Continual Learning with Deep Architectures, Tutorial @ ICML 2021
Image credit: de Lange et al. Continual learning: A comparative study on how to defy forgetting in classification tasks, TPAMI 2019.

Continual Learning vs Other Learning Paradigms

CL vs Domain Adaptation

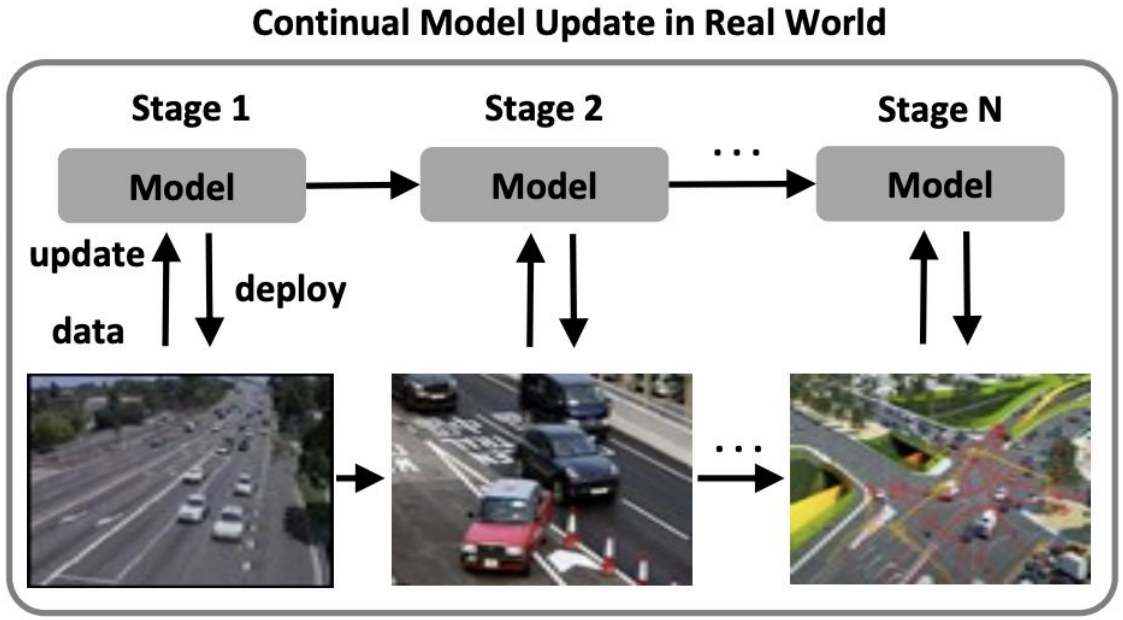
- **Transfer learning** with same source and target tasks, but from different input domains
- Trains on the source domain, adapts model to the new one
- **Unidirectional; does not involve any accumulation of knowledge**

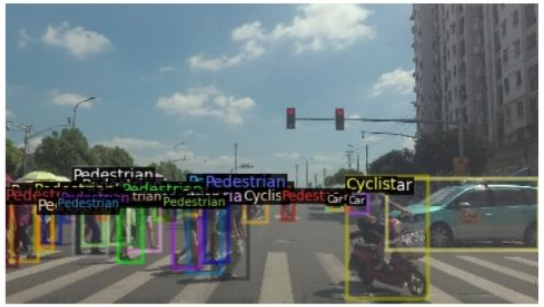


Slide credit: Continual Learning with Deep Architectures, Tutorial @ ICML 2021
Image credit: de Lange et al. Continual learning: A comparative study on how to defy forgetting in classification tasks, TPAMI 2019.

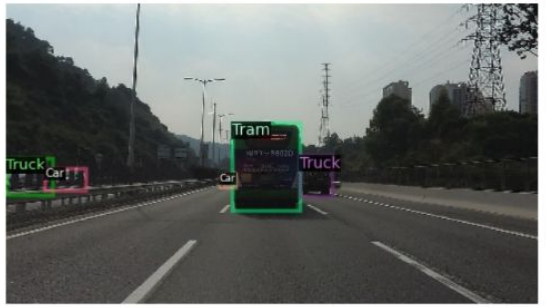
CL Applications - Autonomous Vehicles

- Predict the motion of vehicles in daytime, night time, rain, snow, etc.
  - Driving environment changes over time and place





(a) Task 1: Clear weather, at daytime, in the city center. 4470/500 images.



(b) Task 2: Daytime images on the highway. 1329/148 images.

(c) Task 3: Images at night. 1480/166 images.

(d) Task 4: Rainy images. 524/59 images.

Image credit: Kang et al, Continual Learning for Motion Prediction Model via Meta-Representation Learning and Optimal Memory Buffer Retention Strategy. *CVPR* 2024.
Verwimp, Eli, et al. "Clad: A realistic continual learning benchmark for autonomous driving." Neural Networks 2023.

CL Applications - Large Language Model

- Large language models perform specialized tasks based on different user needs
- Story generation, entity extraction, translation, etc.

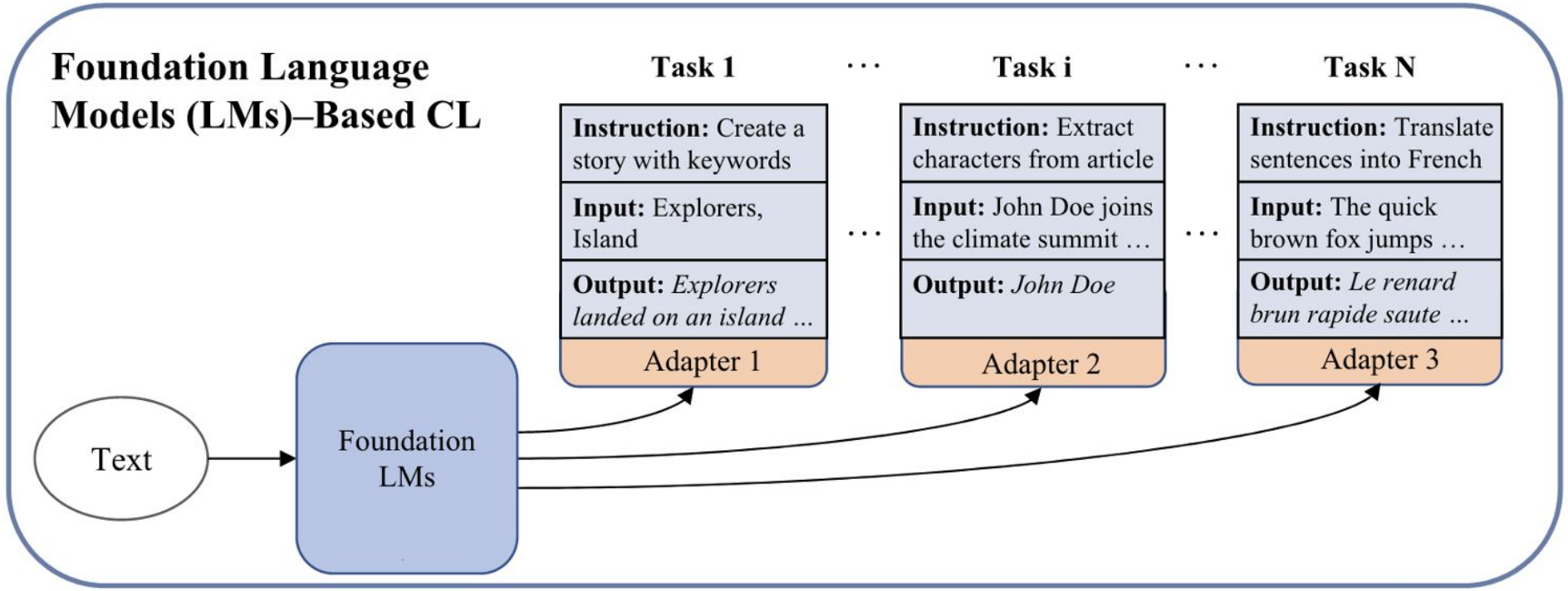
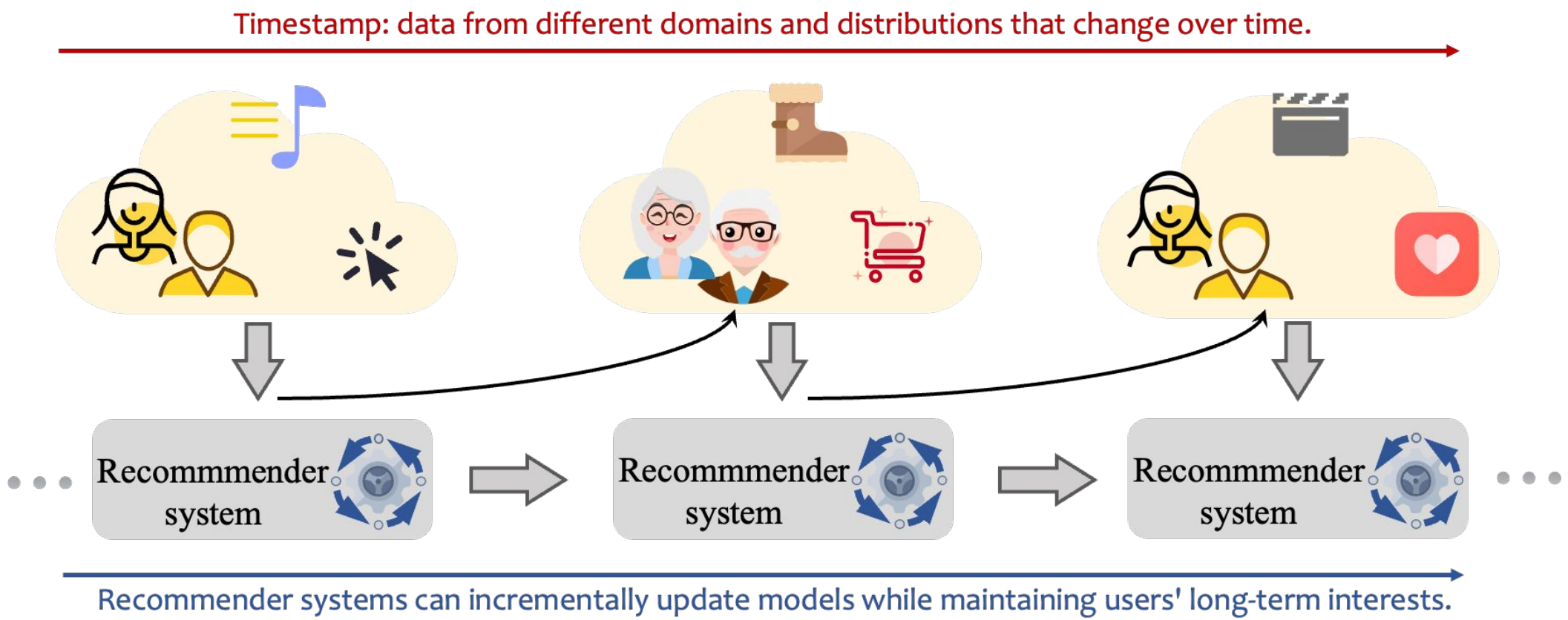


Image credit: Yang et al., Recent Advances of Foundation Language Models-based Continual Learning: A Survey, 2024.

CL Applications - Recommendation System

- Incremental data of user and item information for recommendation system



CL Applications - Website Classification

- Incremental data of new websites
- Classification for different goals: piracy websites, phishing websites, …

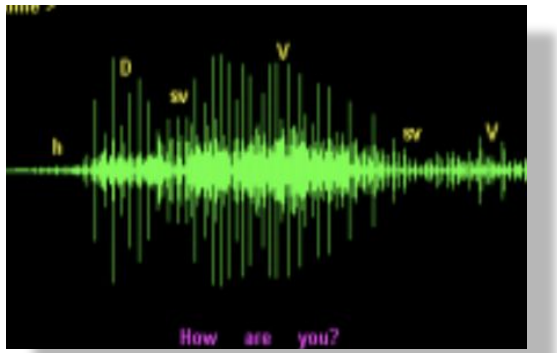
Image credit: http://www.digital-digest.com/blog/DVDGuy/category/movies/high_definition_dvd/

Introduction to Multimodal Continual Learning

- Currently, most CL methods
  - Single modality (unimodal)
- **Real-world environment is multimodal!**
- Multimodal continual learning

Vision

Robotics



Speech

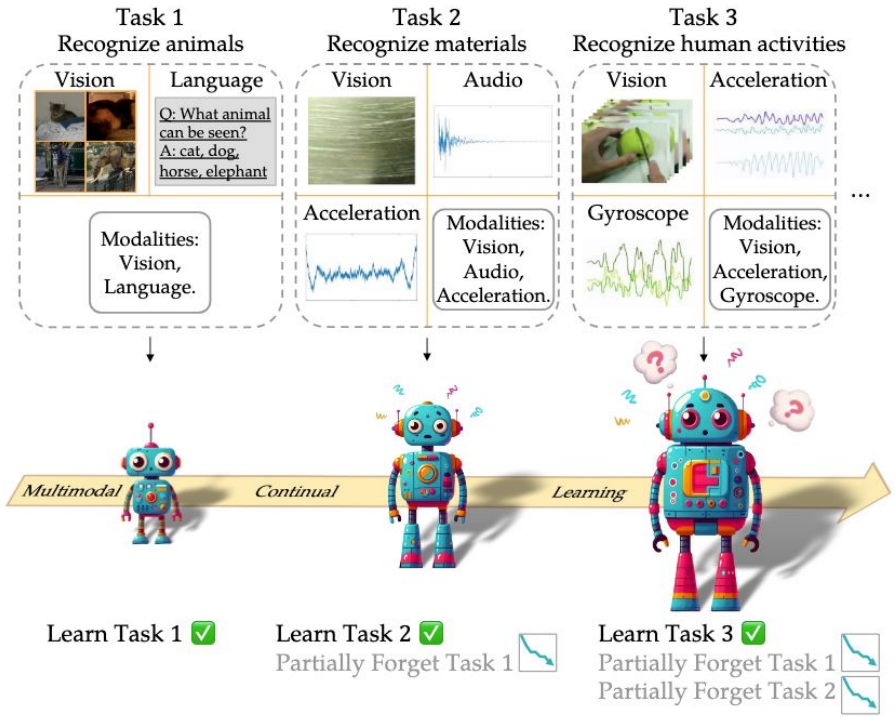
Language

Image credit: https://engineering.mercari.com/en/blog/entry/20210623-5-core-challenges-in-multimodal-machine-learning/

Introduction to MMCL - Definition

Multimodal Continual Learning

Given a **task sequence** $\mathcal{TS}$, multimodal continual learning (MMCL) is the setting where $\mathcal{TS}$ is multimodal, and the model is trained under the CL setting.



(b) Multimodal CL

Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Challenges of MMCL

- In addition to the existing challenge of **catastrophic forgetting** in unimodal CL, the multimodal nature of MMCL introduces the following four challenges.
- **Challenge 1: Modality Imbalance**
- **Challenge 2: Complex Modality Interaction**
- **Challenge 3: High Computational Costs**
- **Challenge 4: Degradation of Pre-trained Zero-shot Capability**

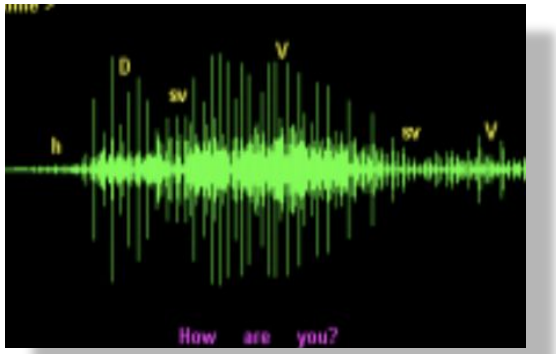
Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Challenges of MMCL

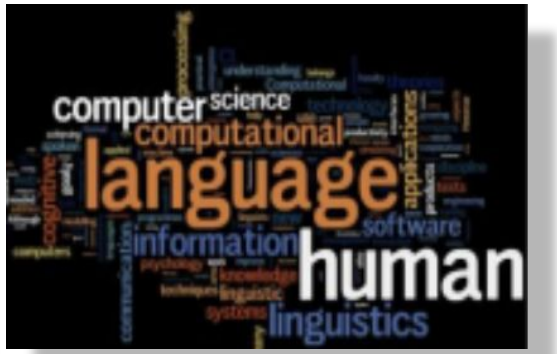
- Challenge 1: Modality Imbalance
- Data-Level Imbalance
  - Extreme cases: Complete absence of certain modalities during training.
  - Skewed data distribution.

Missing modality(s) in some tasks during MMCL

Vision



Speech

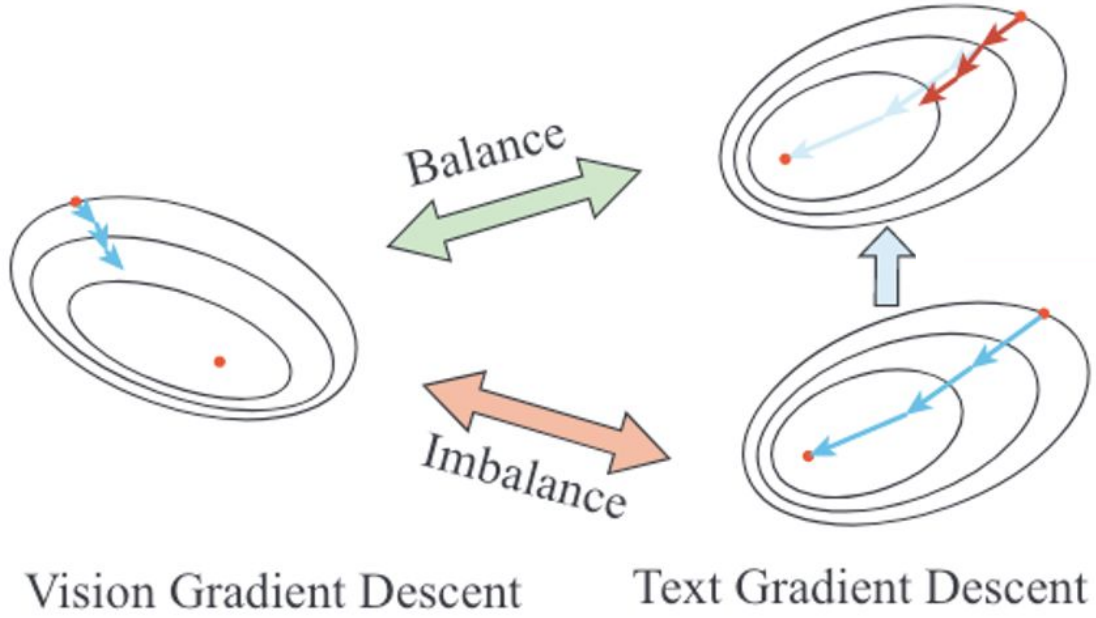


Language

Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.
Image credit: https://engineering.mercari.com/en/blog/entry/20210623-5-core-challenges-in-multimodal-machine-learning/

Challenges of MMCL

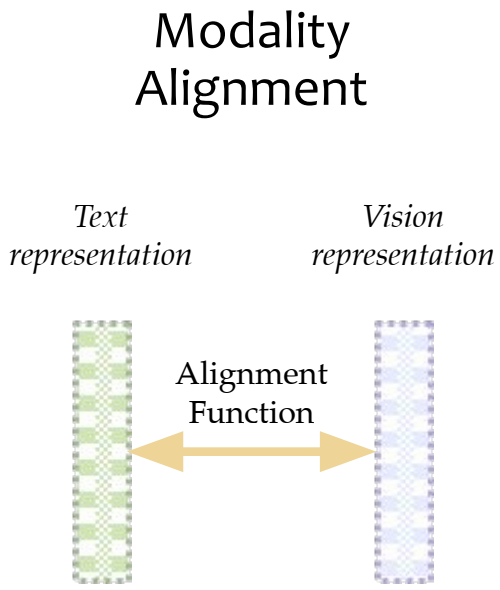
- Challenge 1: Modality Imbalance
- Parameter-Level Imbalance
  - Modality-specific components converge at different speeds.
  - Dominant modalities overshadow others during optimization.



Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.
Image credit: Chen et al., Continual Multimodal Knowledge Graph Construction, IJCAI 2024.

Challenges of MMCL

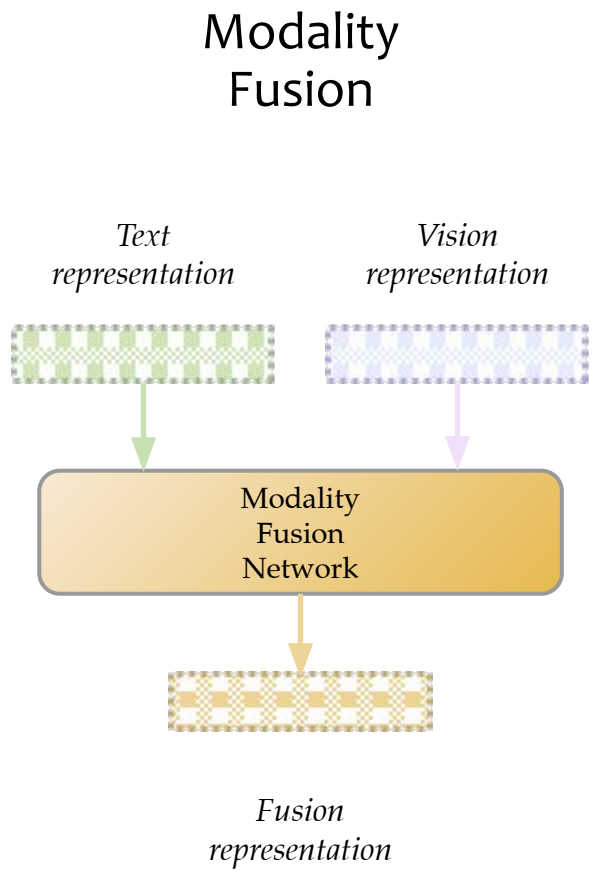
- Challenge 2: Complex Modality Interaction
- Multimodal representations interact through alignment and fusion
- Modality Alignment
  - Spatial Disorder: Features diverge during MMCL
  - Performance Drop: Worse than unimodal CL



Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Challenges of MMCL

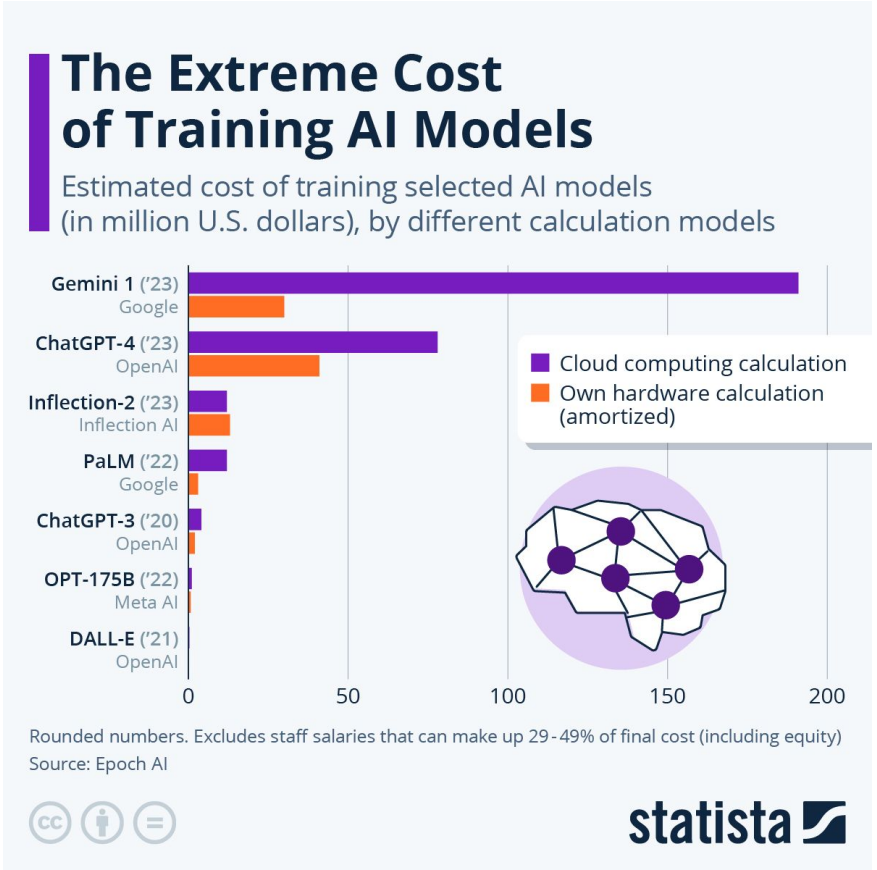
- Challenge 2: Complex Modality Interaction
- Multimodal representations interact through alignment and fusion
- Modality Fusion
  - Forgetting Risk: Traditional fusion methods fail in MMCL settings
  - Data Heterogeneity: Inconsistent distributions across modalities



Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Challenges of MMCL

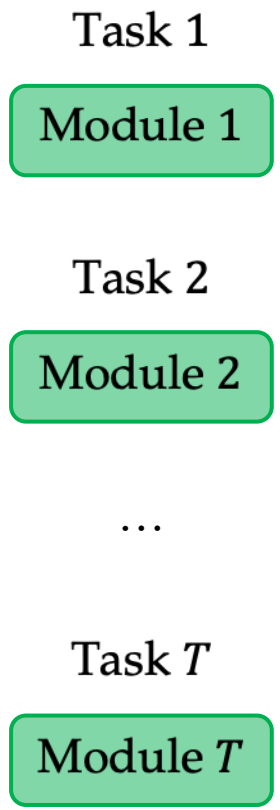
- Challenge 3: High Computational Costs
- Energy/Time Costs: Prohibitive for real-world applications
- Model-Level Overhead
  - Large pre-trained multimodal models require heavy fine-tuning
  - Adding modalities increases trainable parameters exponentially



Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.
Image credit: https://www.statista.com/chart/33114/estimated-cost-of-training-selected-ai-models/

Challenges of MMCL

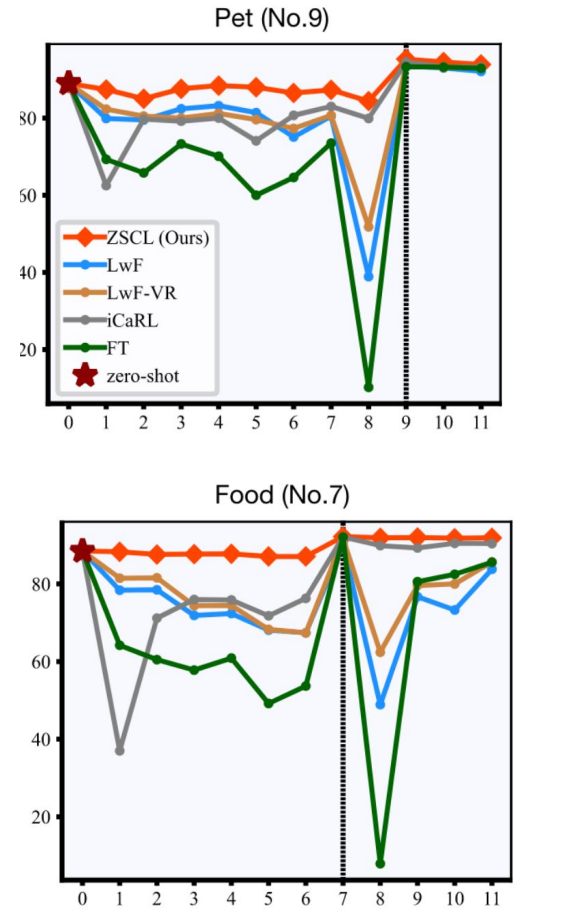
- Challenge 3: High Computational Costs
- Energy/Time Costs: Prohibitive for real-world applications
- Task-Level Accumulation
  - Task-specific parameters grow continuously in MMCL, potentially exceeding backbone model size
  - Undermines efficiency gains expected from MMCL



Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Challenges of MMCL

- Challenge 4: Degradation of Pre-trained Zero-shot Capability
- Pre-trained multimodal models start with strong zero-shot abilities but lose them during continual fine-tuning
- Negative Forward Transfer
  - Performance decay on future tasks due to overwritten pre-trained knowledge
- Capability Trade-off
  - Balancing new task adaptation vs. preserving zero-shot generalization is challenging

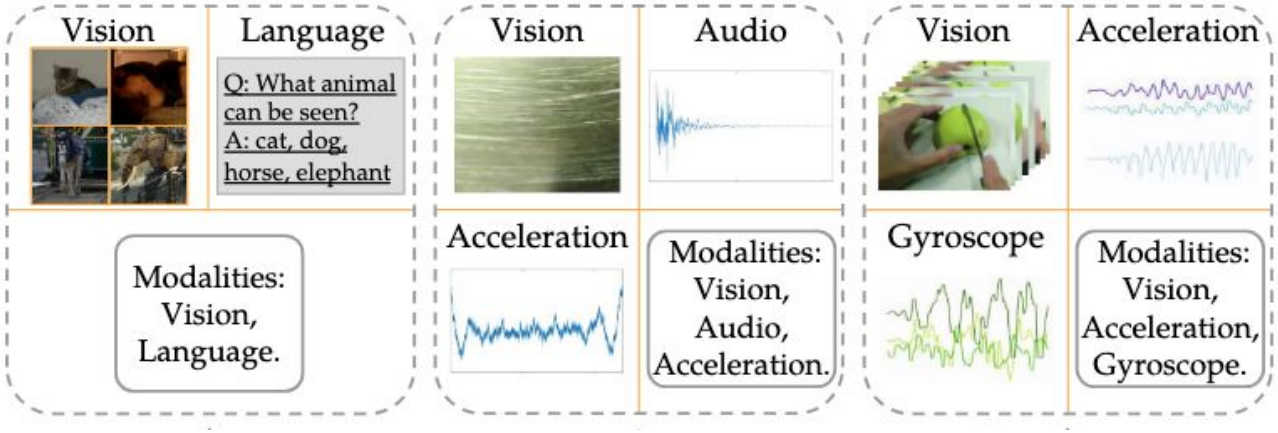


Performance decline of zero-shot capability (green line)

Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.
Image credit: Image credit: Zheng et al., Preventing Zero-Shot Transfer Degradation in Continual Learning of Vision-Language Models, ICCV 2023.

Taxonomy of MMCL

- **Data level**
  - Data of different modalities
  - Vision
  - Language
  - Audio
  - …



- **Model level**
  - CL algorithms in four main categories
    - Regularization-based
    - Architecture-based
    - Replay-based
    - Prompt-based

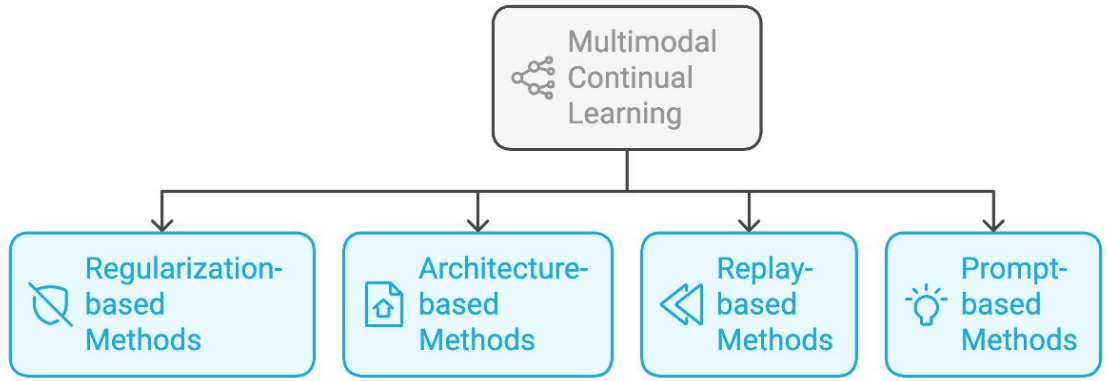


Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Taxonomy of MMCL

- **Taxonomy** of existing MMCL works

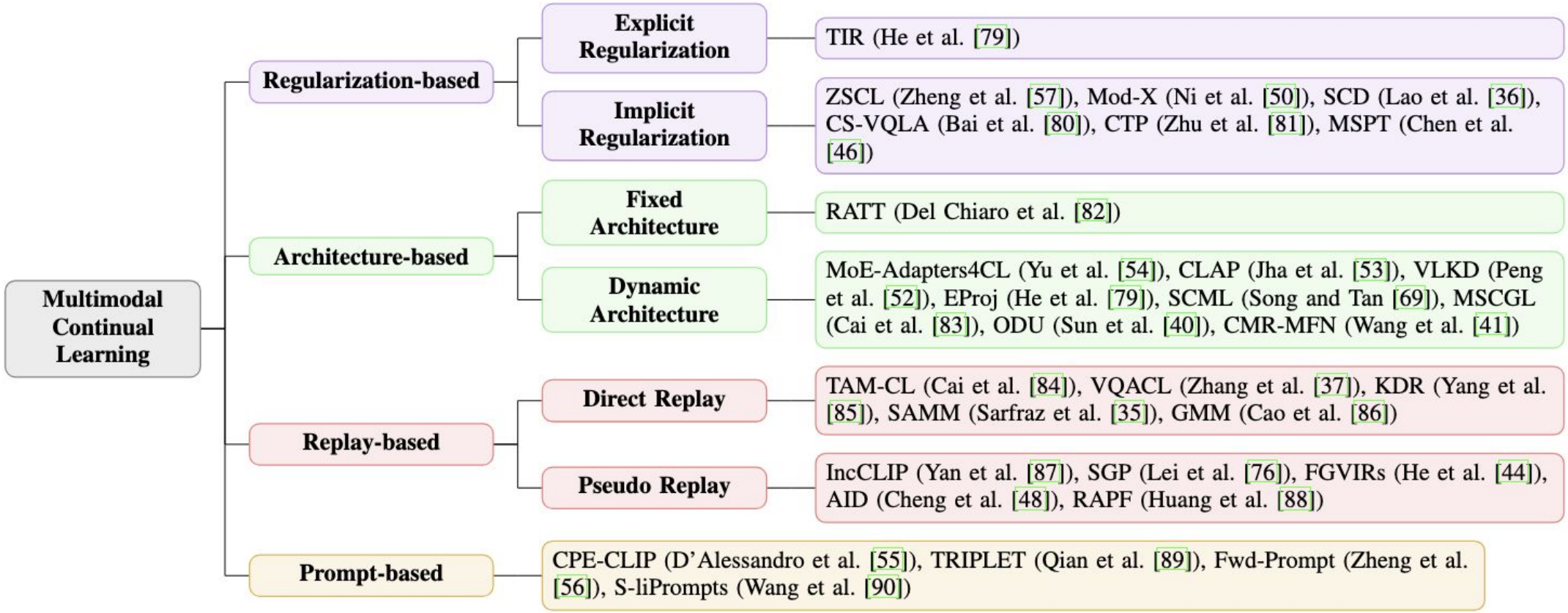


Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Taxonomy of MMCL

- **Taxonomy** of existing MMCL works

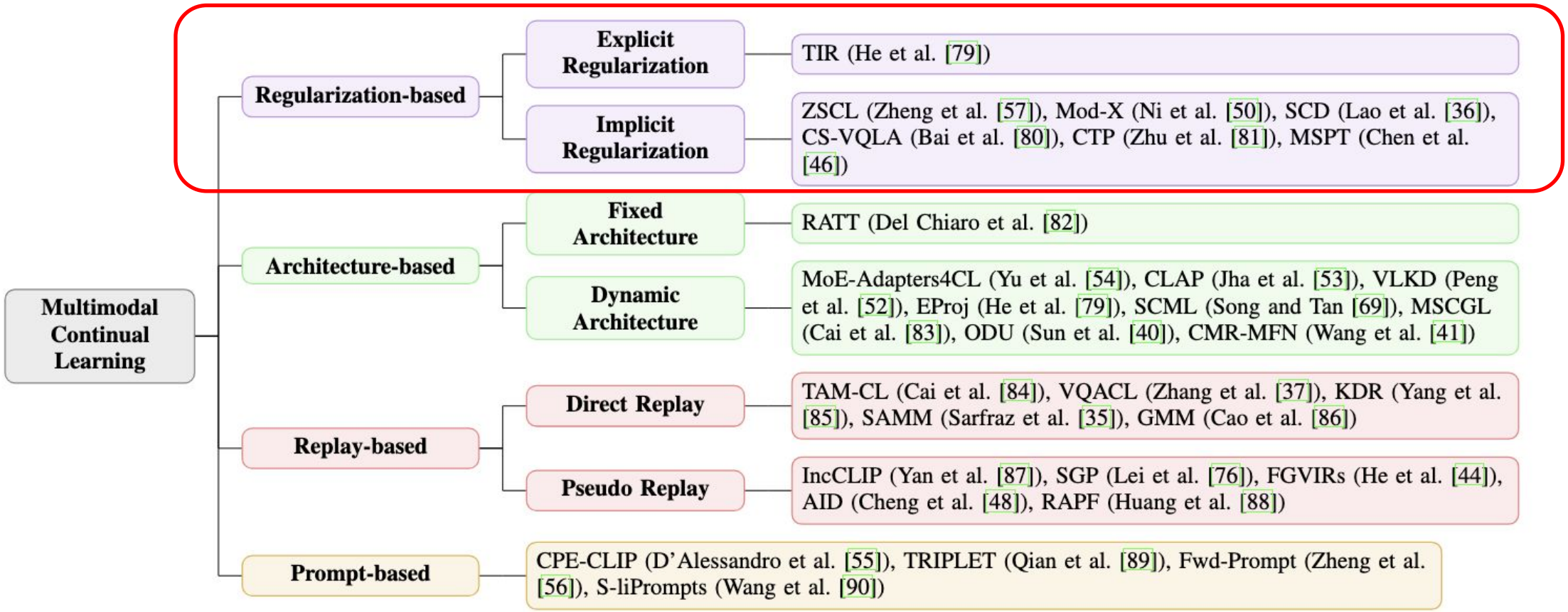
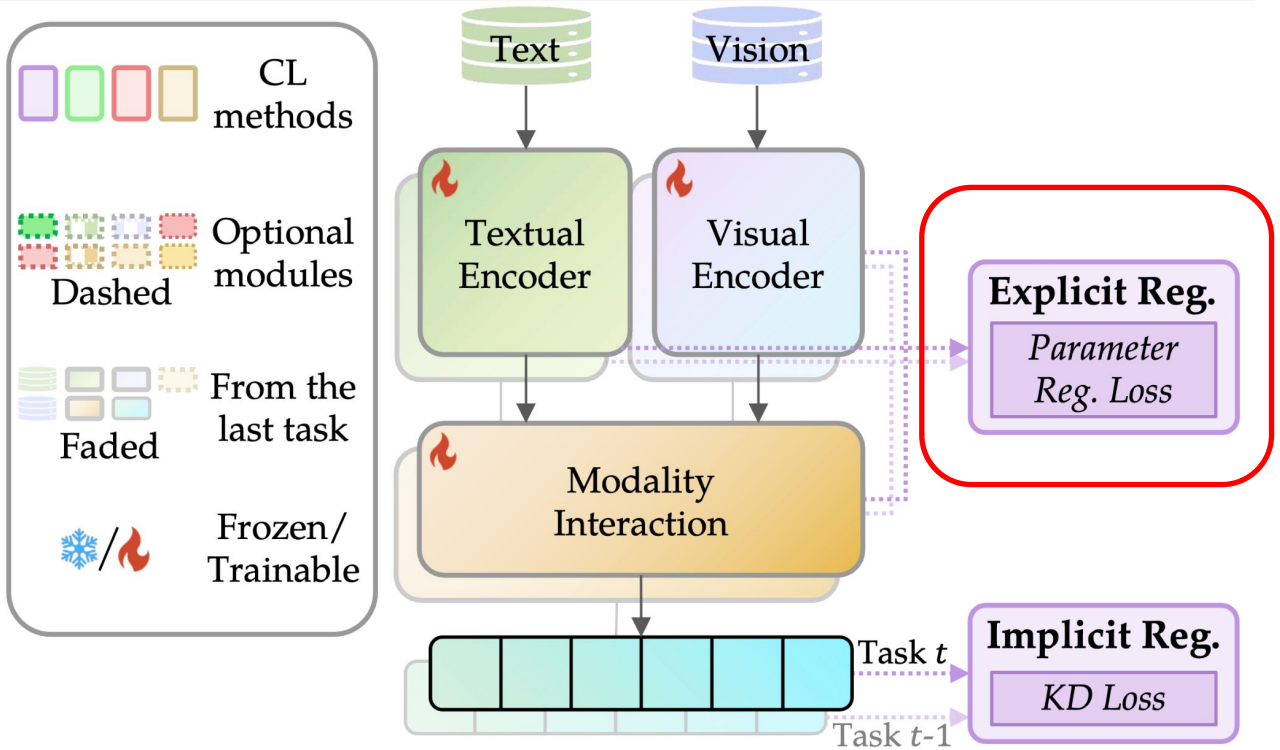


Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Regularization-based Methods

- **Regularization-based methods**
- Add constraints on parameter change
- **Plasticity regulation (correspond to plasticity of brains)**

- Explicit regularization
  - Restriction applies to all parameters directly

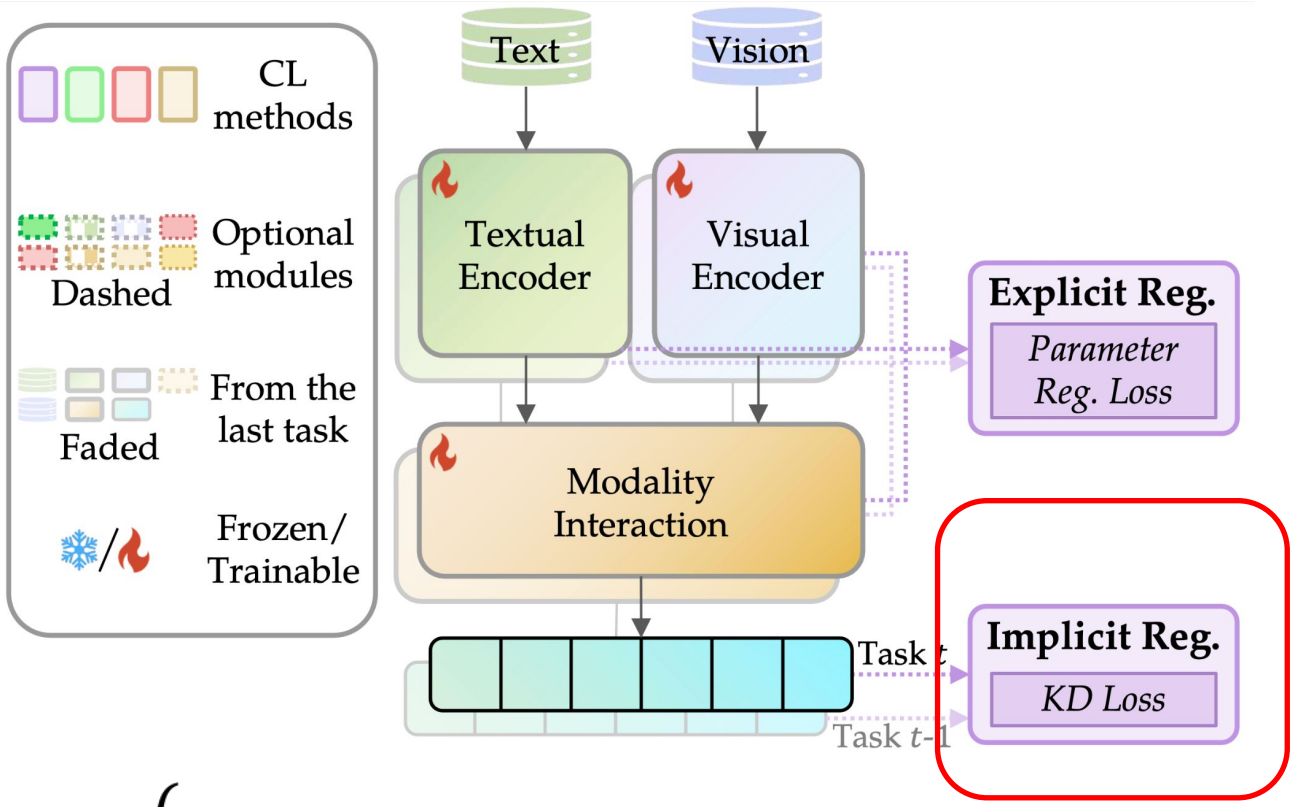


$$\mathcal{L}_{E,t} = \sum_i b_i \left(\theta_i - \theta^*_{t-1,i}\right)^2$$

Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Regularization-based Methods

- Implicit regularization
  - Restriction focuses on model output of previous tasks to avoid forgetting
- Feature level, logit level



$$\mathcal{L}_{I,t} = \begin{cases} -\sum_i y_{t-1,i} \log y_{t,i} & \text{cross-entropy loss} \\ \|\mathbf{y}_{t-1} - \mathbf{y}_t\|_2^2 & \text{L2 loss} \end{cases}$$

Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Regularization-based Methods

- **Regularization-based methods**
- **Advantage:** Apply to the loss function and thus is model agnostic
- **Disadvantage:** Regularization is rigid and may prevent models from learning new tasks

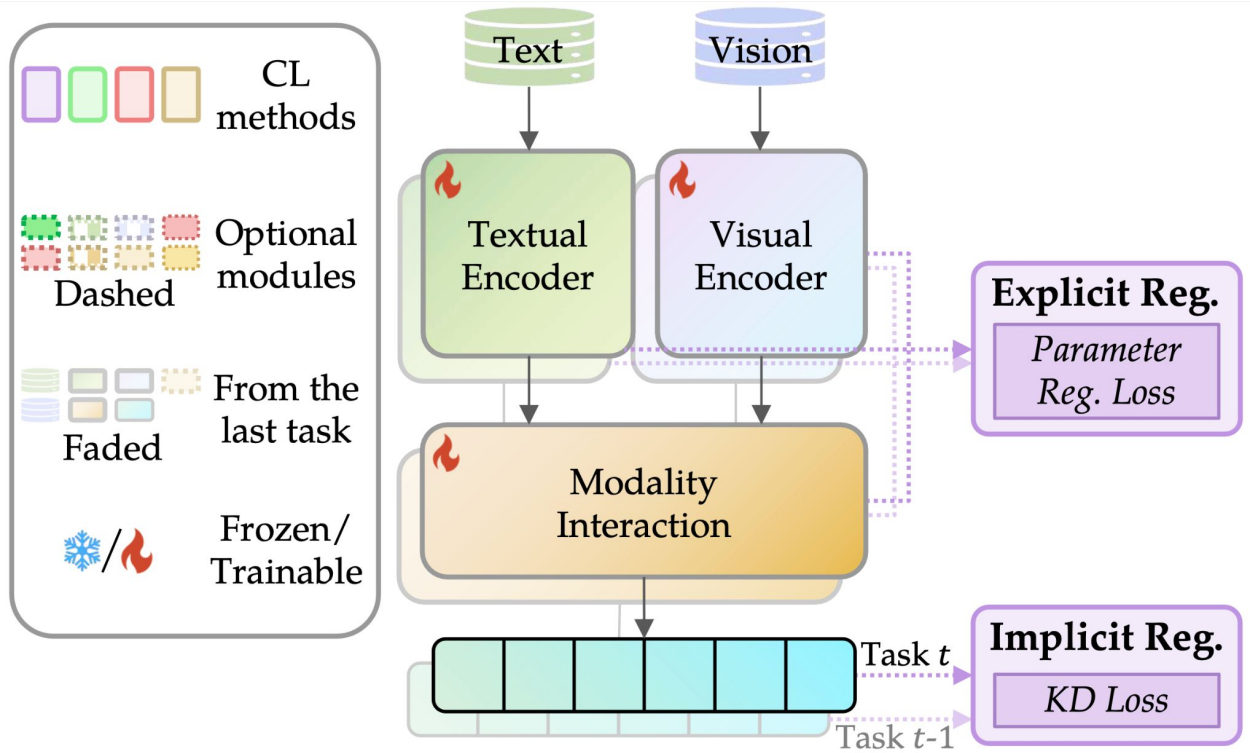


Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Regularization-based Methods

- **ZSCL**
- Feature Space
  - Introduce a reference dataset for distillation between current and initial models
- Parameter Space
  - Prevent large parameter shifts by averaging model weights during training
- Addressed challenge(s): C4

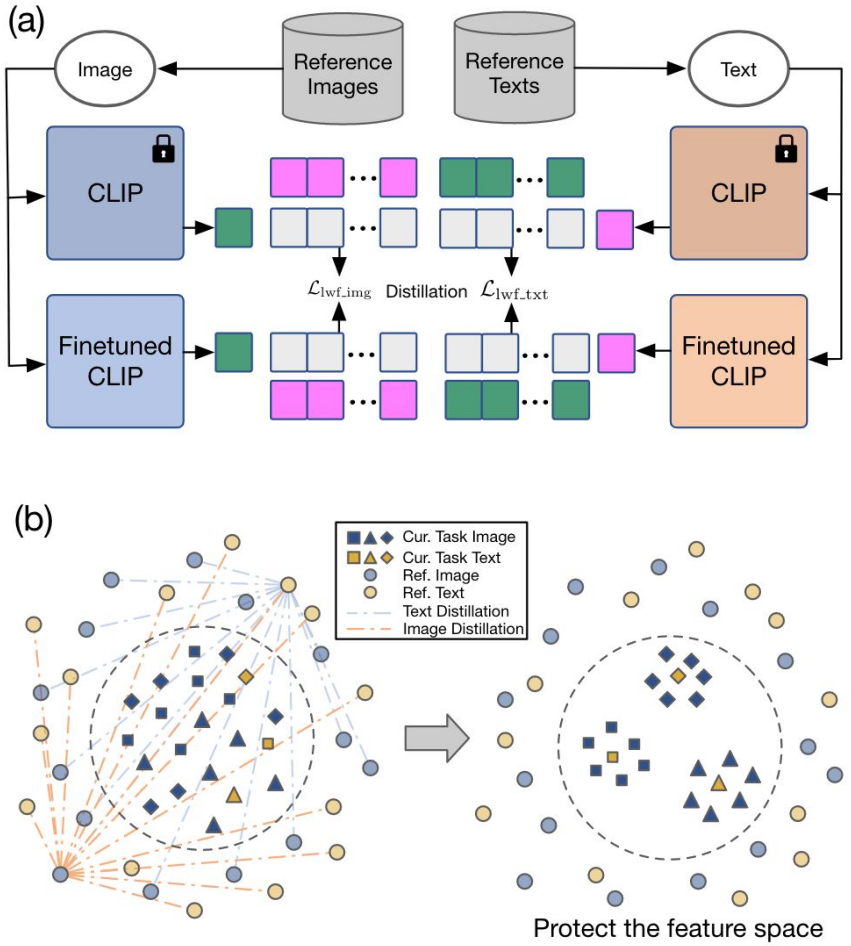


Image credit: Zheng et al., Preventing Zero-Shot Transfer Degradation in Continual Learning of Vision-Language Models, ICCV 2023.

Regularization-based Methods

- **Mod-X**
- Spatial Disorder (SD): Performance degradation due to
  - Intra-modal Rotation: Shifts within vision/language spaces
  - Inter-modal Deviation: Misalignment between modalities

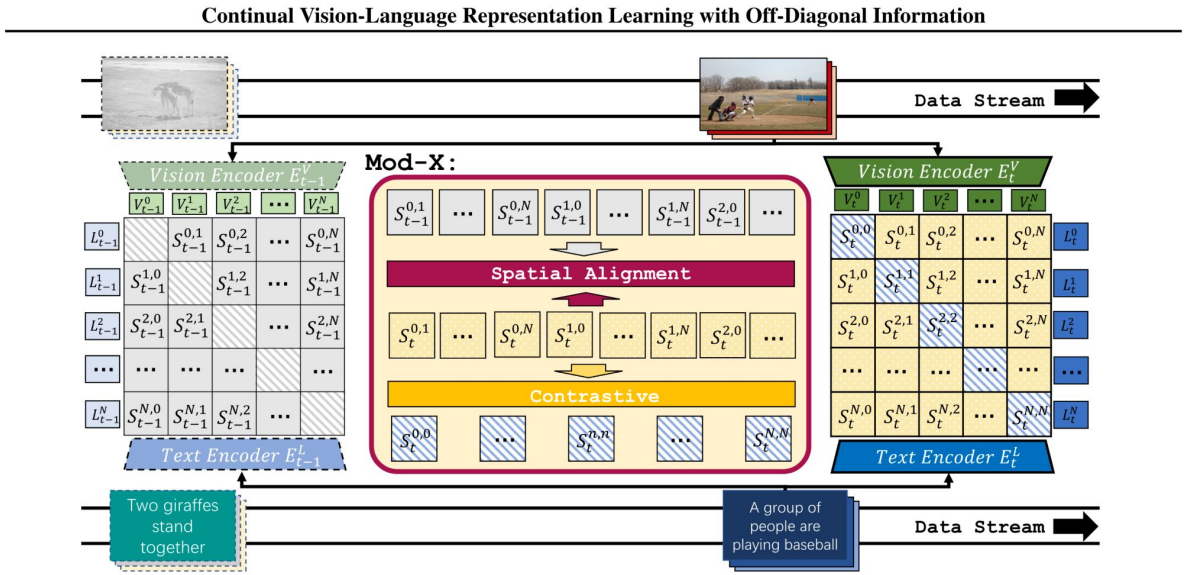


Image credit: Ni et al., Continual Vision-Language Representation Learning with Off-Diagonal Information, ICML 2023.

Regularization-based Methods

- **Mod-X**
- Preserve off-diagonal information in contrastive matrices
  - Selectively aligns old/new data domains by stabilizing cross-modal space geometry
  - Mitigates SD without sacrificing new-task adaptability
- Addressed challenge(s): C2

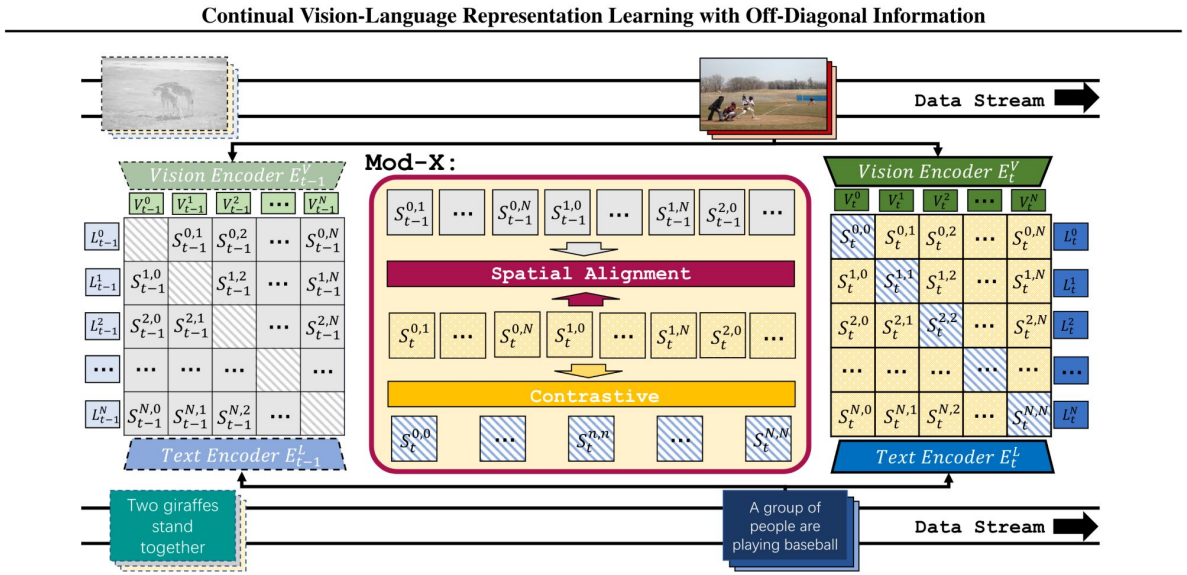


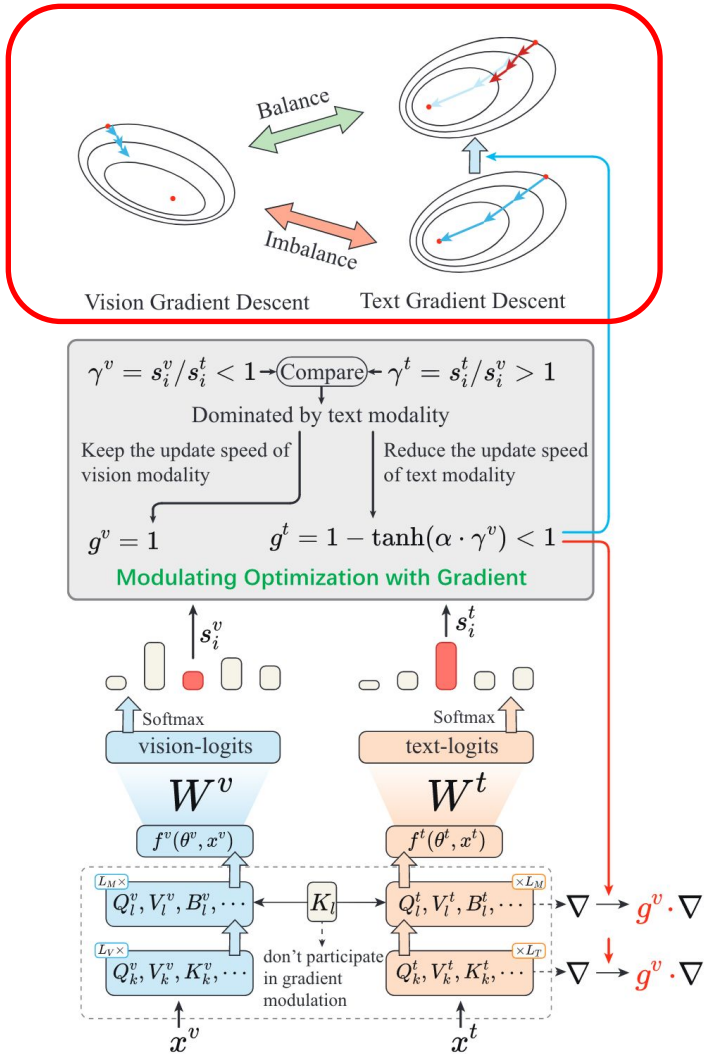
Image credit: Ni et al., Continual Vision-Language Representation Learning with Off-Diagonal Information, ICML 2023.

Regularization-based Methods

- **MSPT**

- Balances stability (retain learned knowledge) and plasticity (integrate new data)
- Modulating optimization with gradient
  - Imbalanced convergence rates across modalities (visual/textual) disrupt continual learning
  - Dynamically adjusts SGD optimization for visual and textual encoders

- Addressed challenge(s): C1, C2



(b) Balanced Multimodal Learning Dynamics.

Image credit: Chen et al., Continual Multimodal Knowledge Graph Construction, IJCAI 2024.

Taxonomy of MMCL

- **Taxonomy** of existing MMCL works

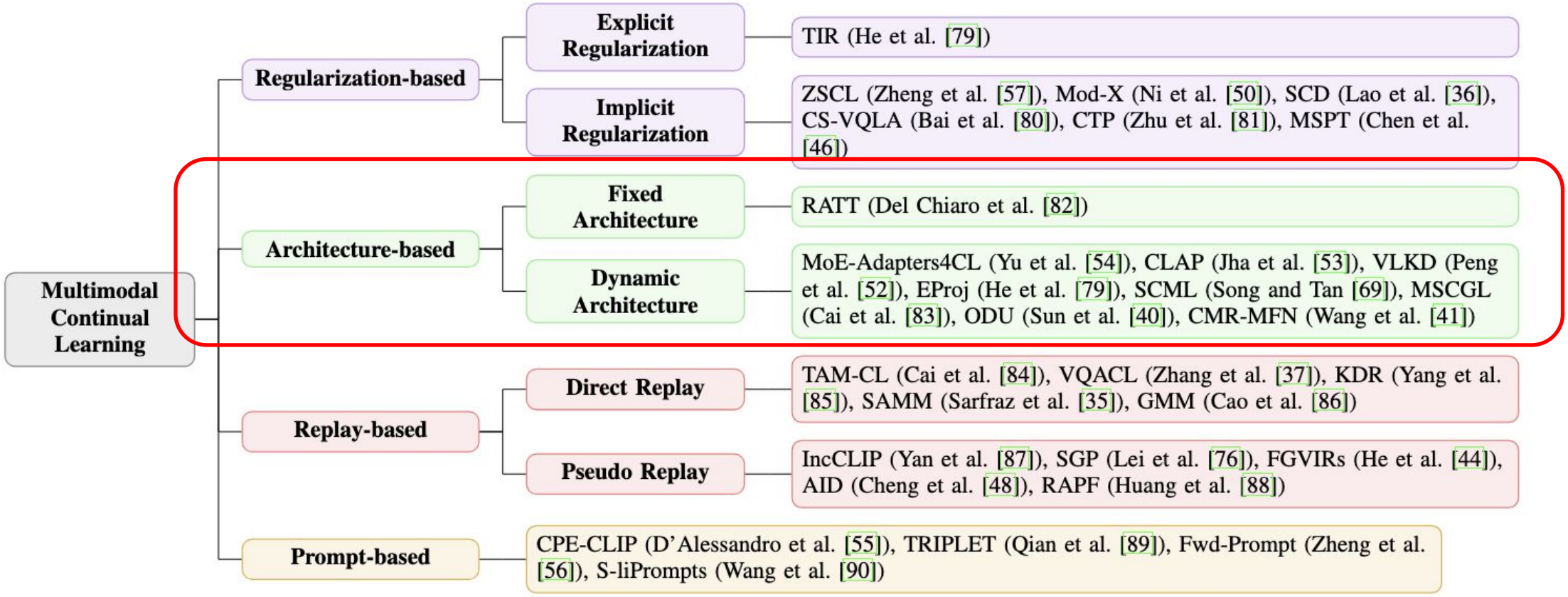


Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Architecture-based Methods

- **Architecture-based methods**
- **Correspond to neurogenesis of brains**

- Different model parameters cope with different tasks

- Intuitive and direct

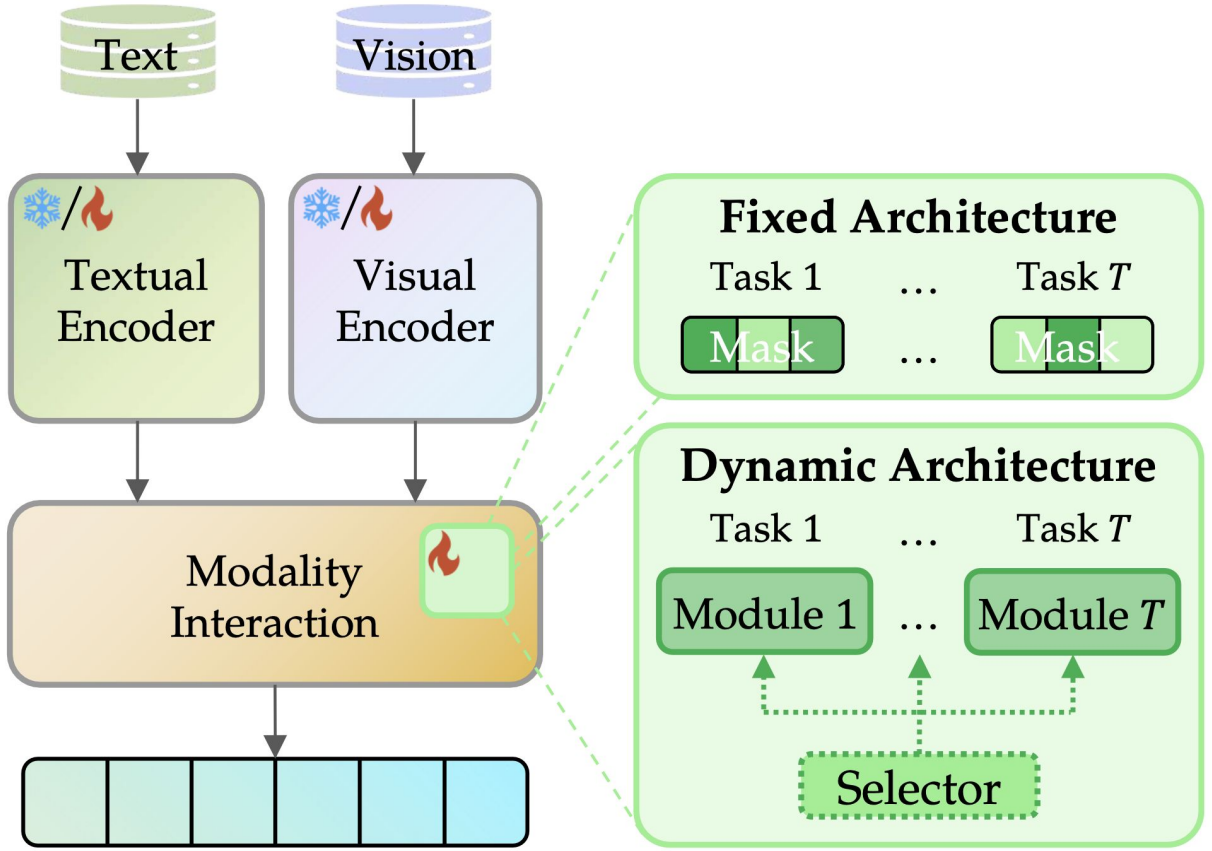


Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Architecture-based Methods

- **Inter-task interference**
  - Remembering old tasks greatly interferes with learning a new task
- Regularization-based methods are prone to inter-task interference
- Architecture-based methods reduce inter-task interference

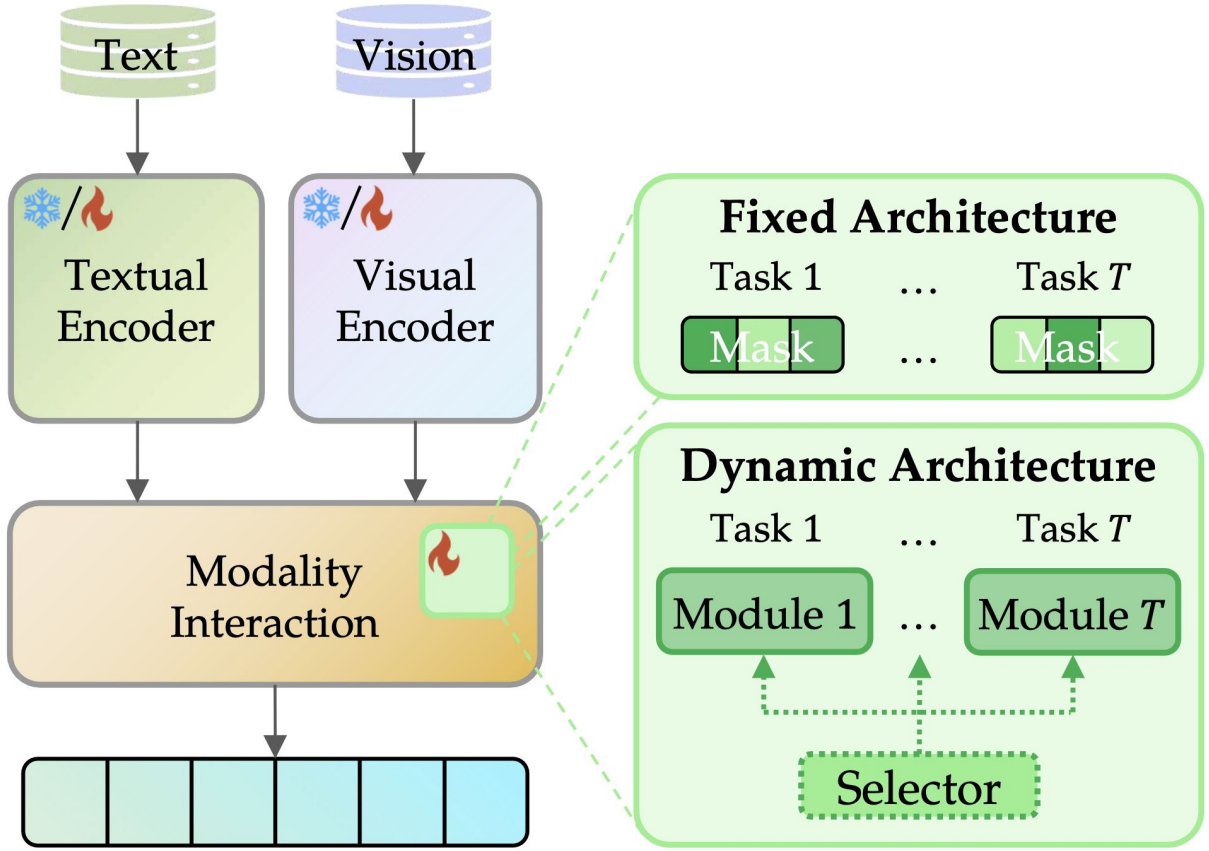


Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Architecture-based Methods

- **Advantage:** architecture-based methods reduce inter-task interference by incorporating task-specific components
- **Disadvantage:** increased training cost induced by task-specific components

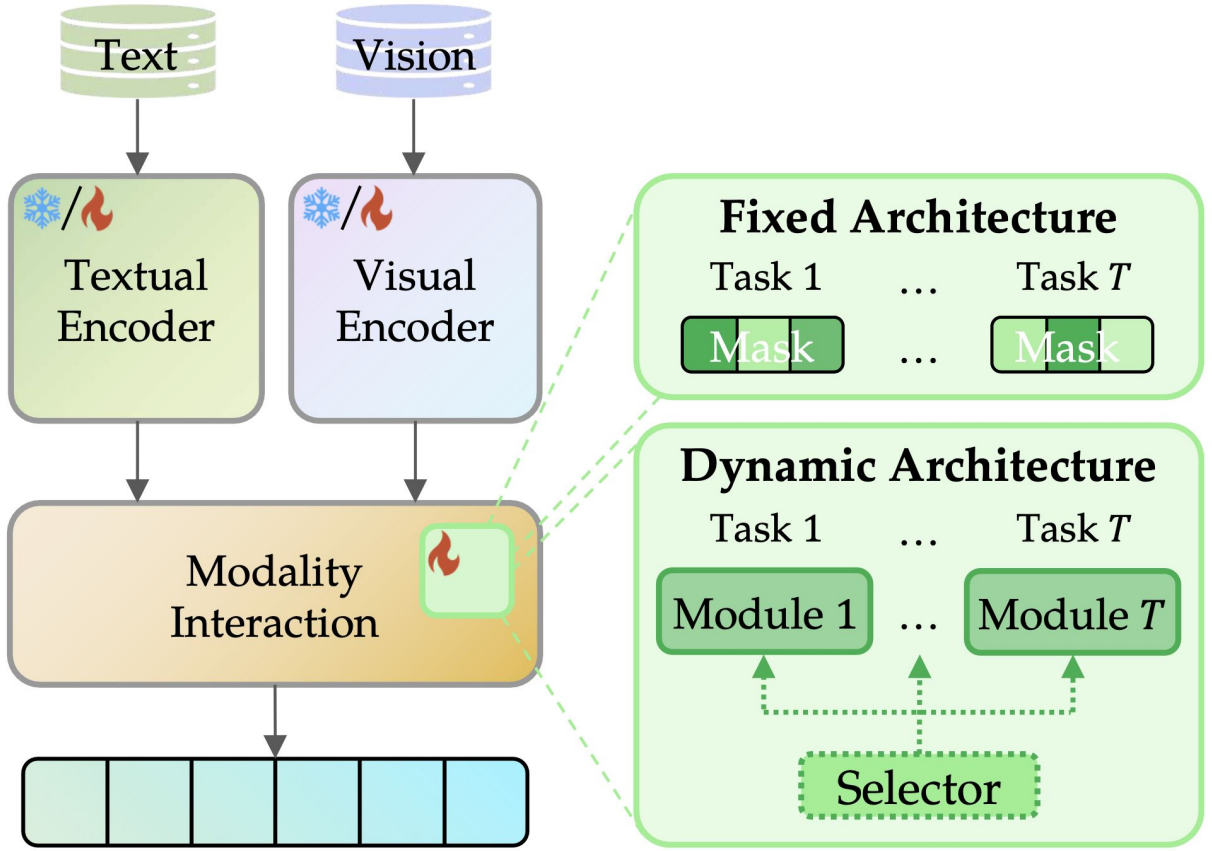


Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Architecture-based Methods

- **MoE-Adapters4CL**

- Continual learning for CLIP model to classify images from different domains

- Reduce catastrophic forgetting
  - Utilize MoE structure to learn new knowledge
  - Freeze the pretrained CLIP model

- Addressed challenge(s): C3, C4

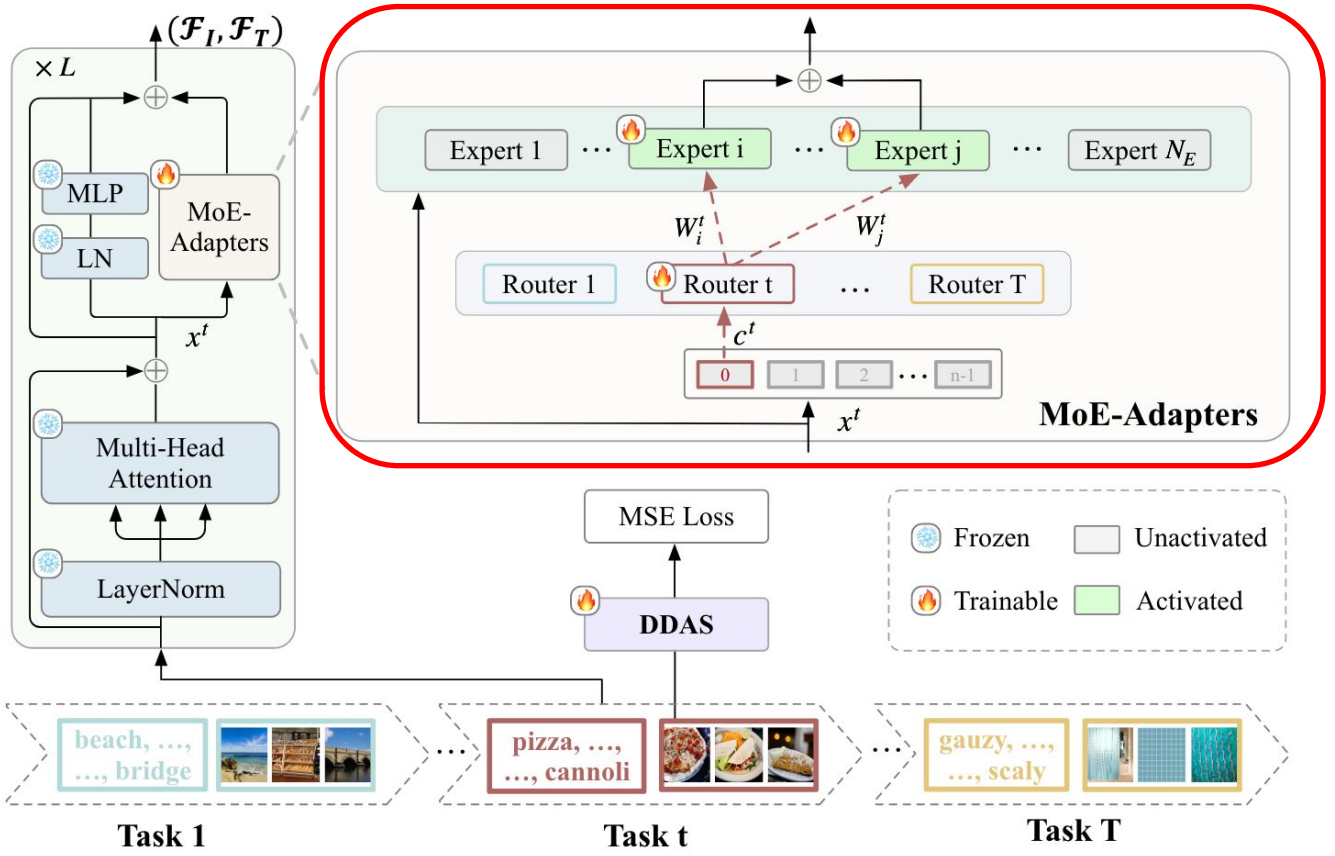


Image credit: Yu et al., Boosting Continual Learning of Vision-Language Models via Mixture-of-Experts Adapters, CVPR 2024.

Architecture-based Methods

- **CLAP**
- Probabilistic Modeling
  - Task-specific modules with visual-guided text features
- Forgetting Mitigation
  - Leverages CLIP's pre-trained knowledge for distribution regularization
- Addressed challenge(s): C2, C4

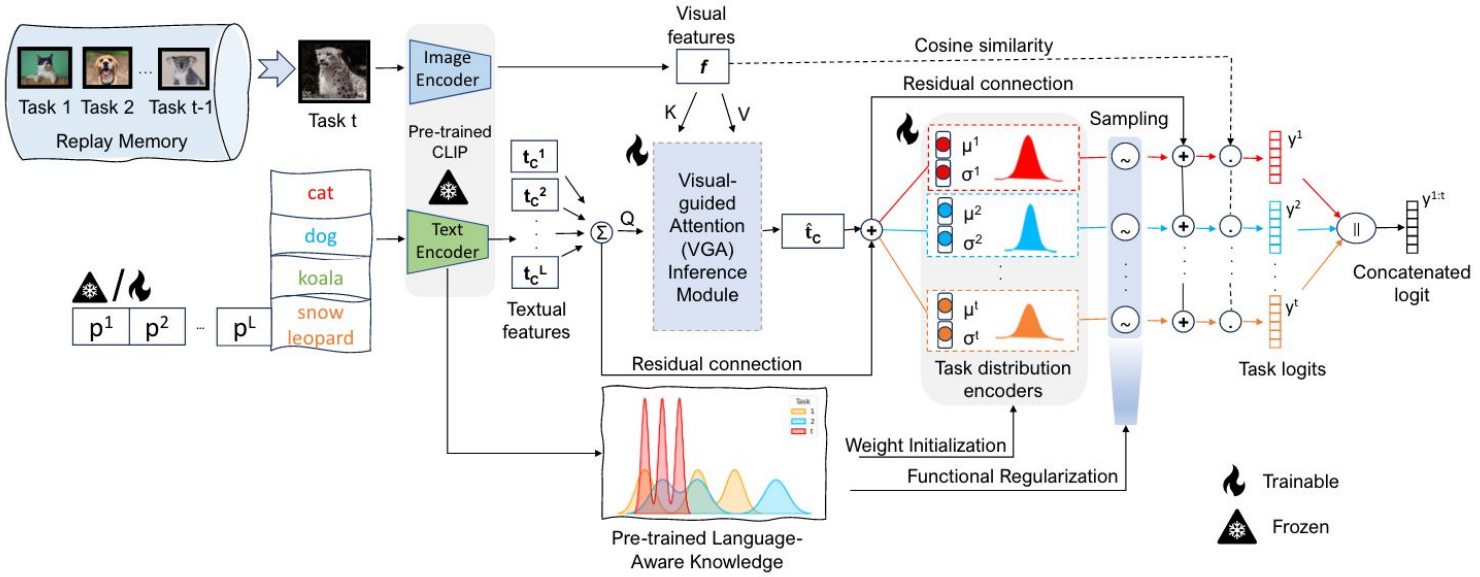


Image credit: Jha et al., CLAP4CLIP: Continual learning with probabilistic finetuning for vision-language models, NeurIPS 2024.

Taxonomy of MMCL

- **Taxonomy** of existing MMCL works

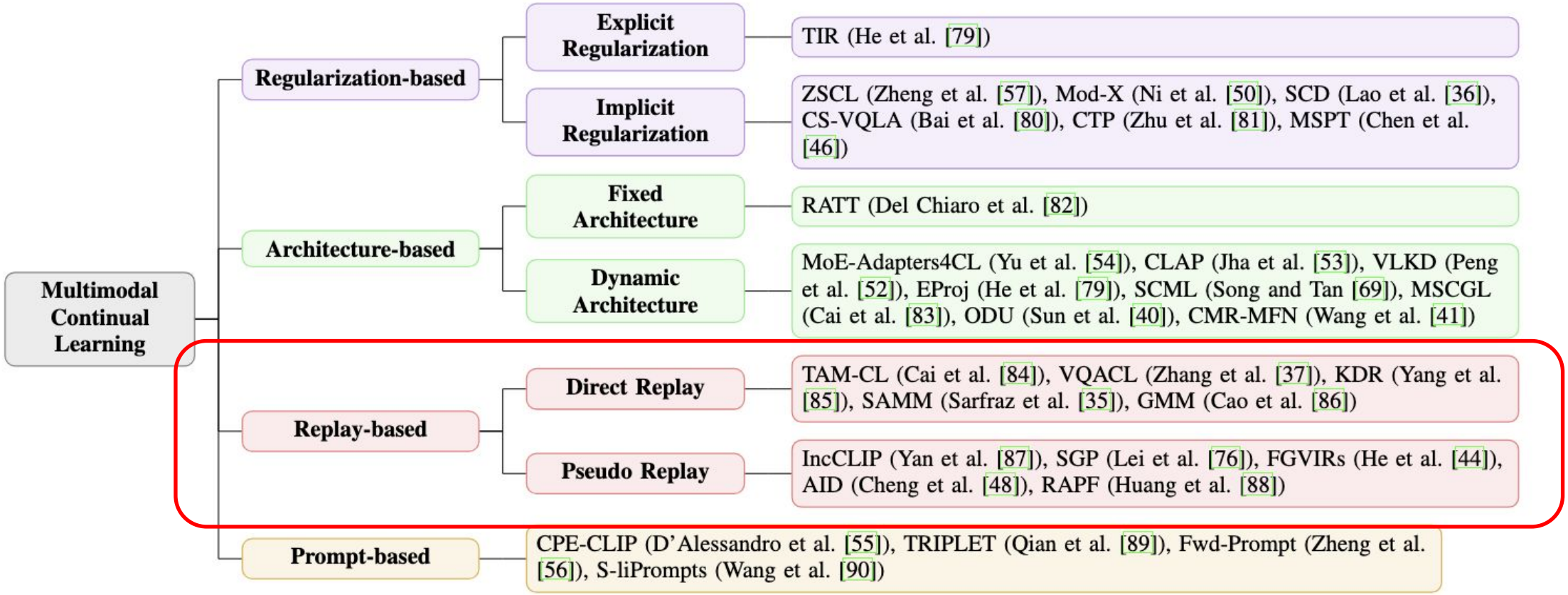
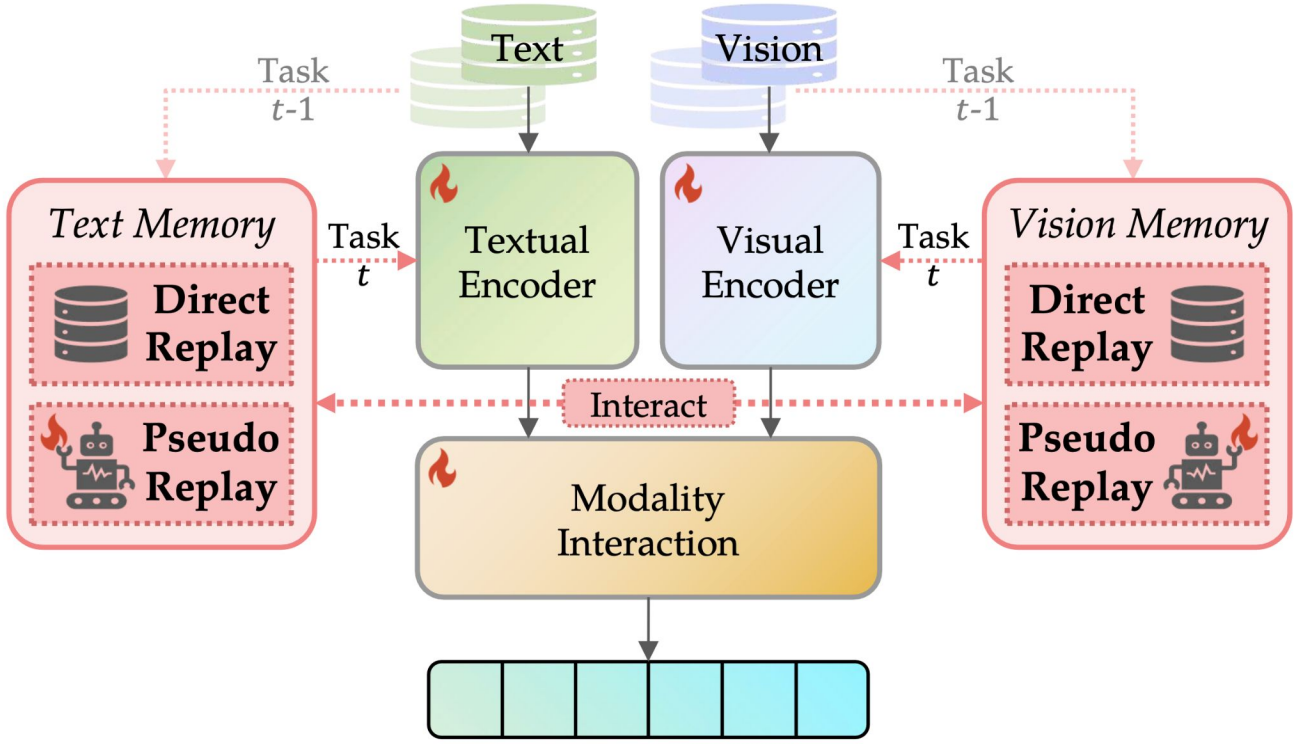


Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Replay-based Methods

- **Replay-based methods**
- **Correspond to episodic replay of brains**

- Method: an episodic memory buffer to replay historical instances

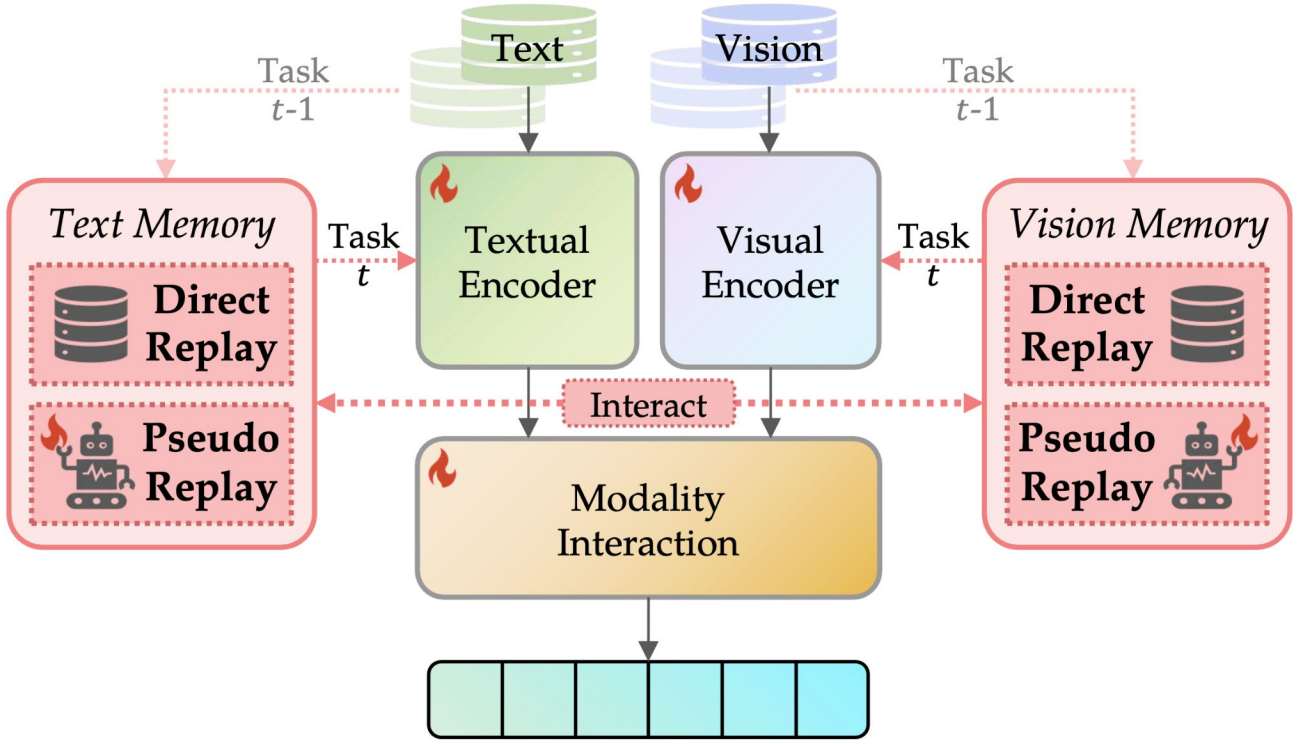


$$\mathcal{L}_t = \frac{1}{|D_t \cup \mathcal{M}_t|} \sum_{(\mathbf{x}_i, y_i) \in (D_t \cup \mathcal{M}_t)} \ell(f(\mathbf{x}_i), y_i)$$

Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Replay-based Methods

- **Advantage:** Avoids the rigid loss constraints and complex network architectures
- **Disadvantage:** Using past data in direct replay methods is under a relaxed setting of CL and may introduce trustworthy concerns



$$\mathcal{L}_t = \frac{1}{|D_t \cup \mathcal{M}_t|} \sum_{(\mathbf{x}_i, y_i) \in (D_t \cup \mathcal{M}_t)} \ell(f(\mathbf{x}_i), y_i)$$

Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Replay-based Methods

- **SGP**

- Key Idea: Replay past knowledge without real data
  - Scene graph snippets → generate pseudo images/QAs

- Addressed challenge(s): Catastrophic forgetting

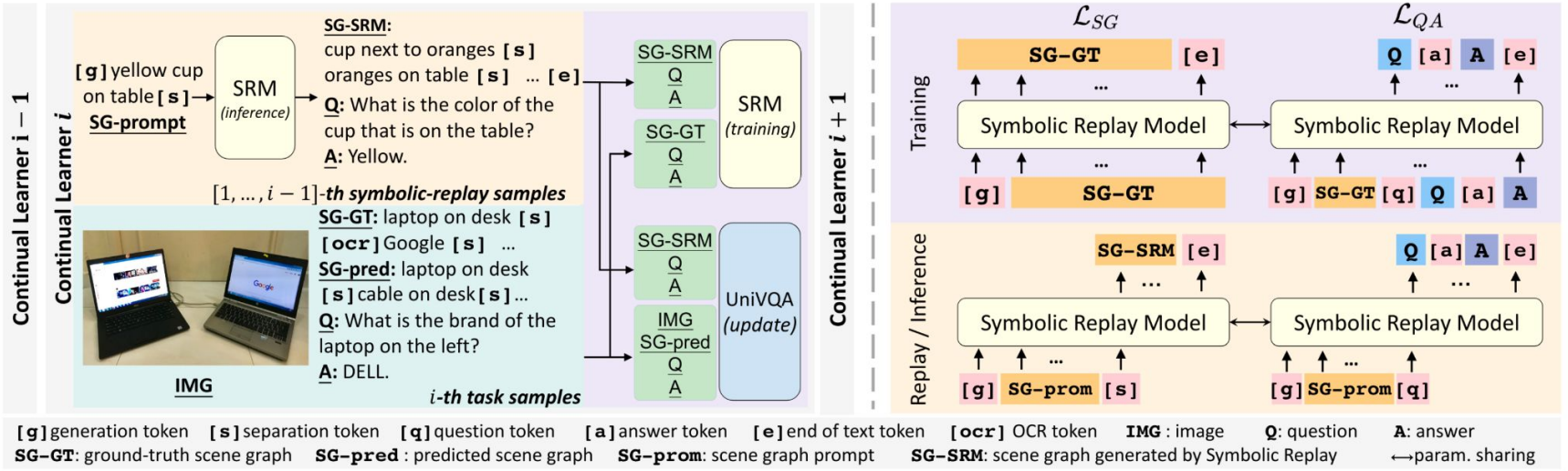


Image credit: Lei et al., Symbolic Replay: Scene Graph as Prompt for Continual Learning on VQA Task, AAAI 2023.

Replay-based Methods

- **VQACL**
- Dual-representation learning
  - Direct replay
  - Sample-specific + sample-invariant features → Discriminative & generalizable VQA representations
- Addressed challenge(s): Catastrophic forgetting

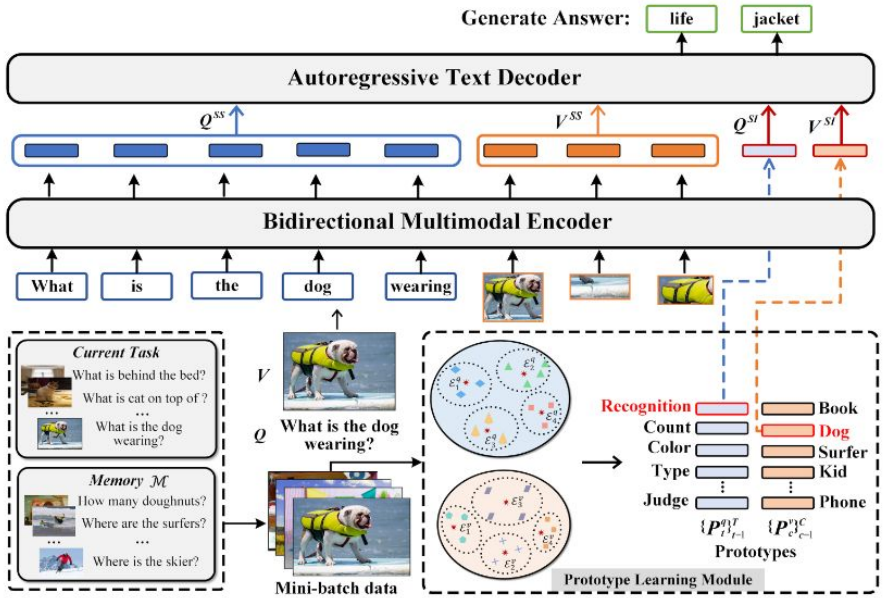


Image credit: Zhang et al., VQACL: A Novel Visual Question Answering Continual Learning Setting, CVPR 2023.

Taxonomy of MMCL

- **Taxonomy** of existing MMCL works

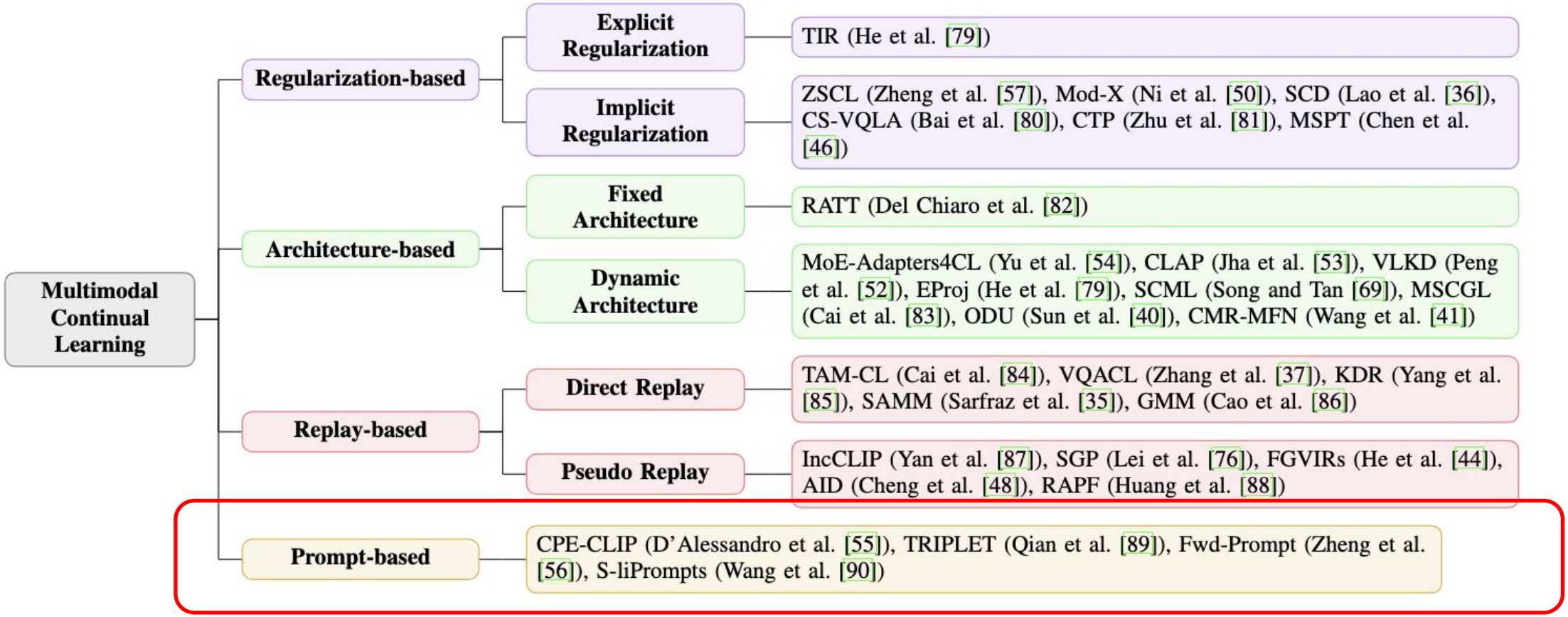


Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Prompt-based Methods

- **Prompt-based methods**
- Utilize the rich knowledge of pre-trained models
- Modifying the input by applying a few prompt parameters

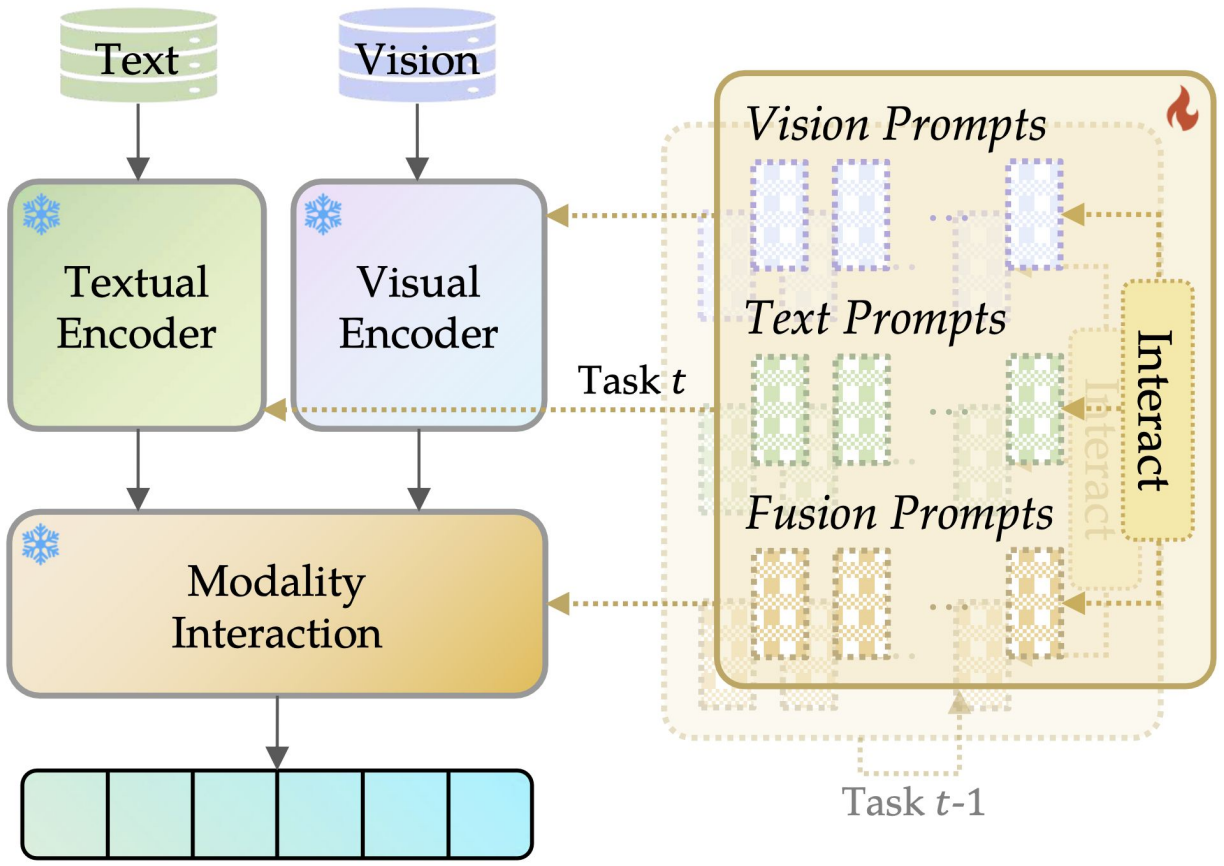


Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Prompt-based Methods

- **Advantage:** Parameter efficient training and preserve pretrained knowledge
- **Disadvantage:** Lack of explainability in the prompt embedding and highly rely on pretrained models

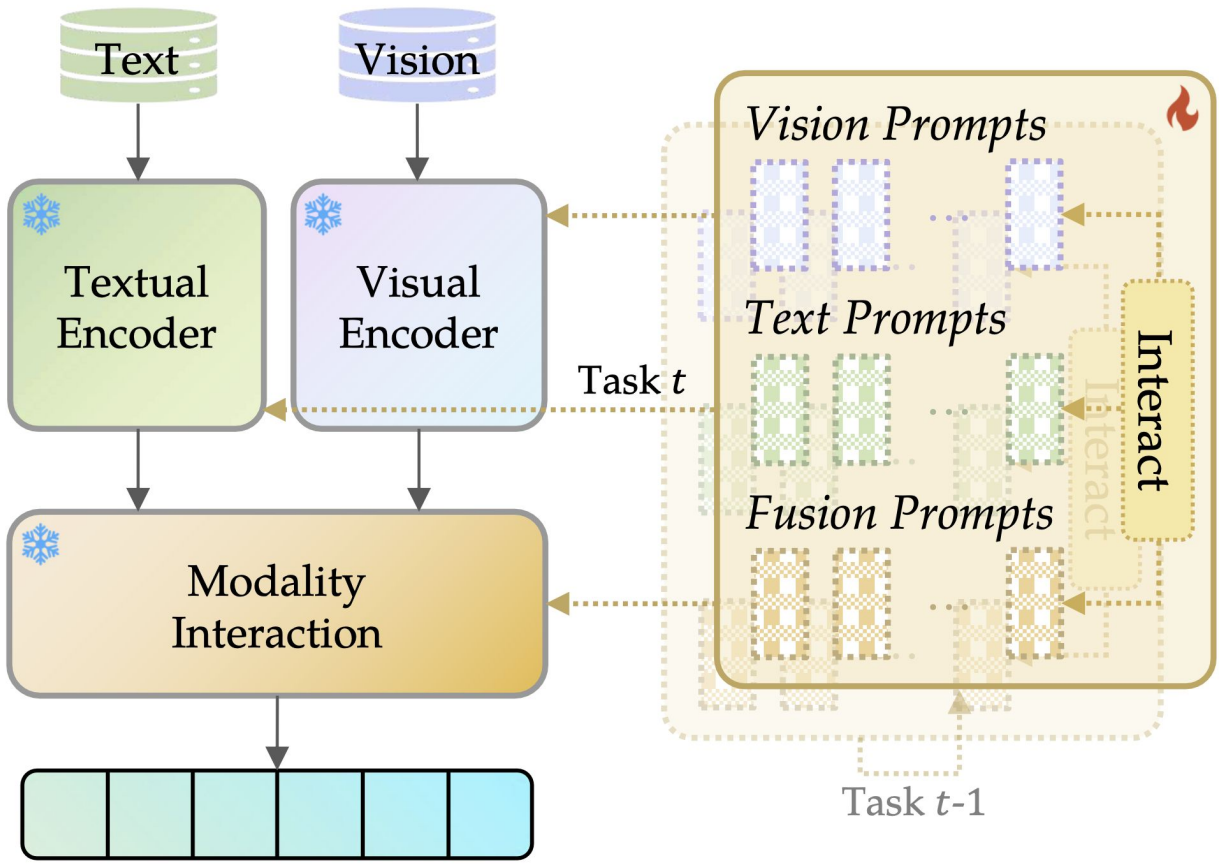


Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Prompt-based Methods

- **S-liPrompts**
- Image-end Prompts
  - Independent set of continuous learnable parameters as a part of inputs to the pre-trained ViT
- Language-end Prompts
  - Replace manual prompts with learnable context vectors
- Addressed challenge(s): C3, C4

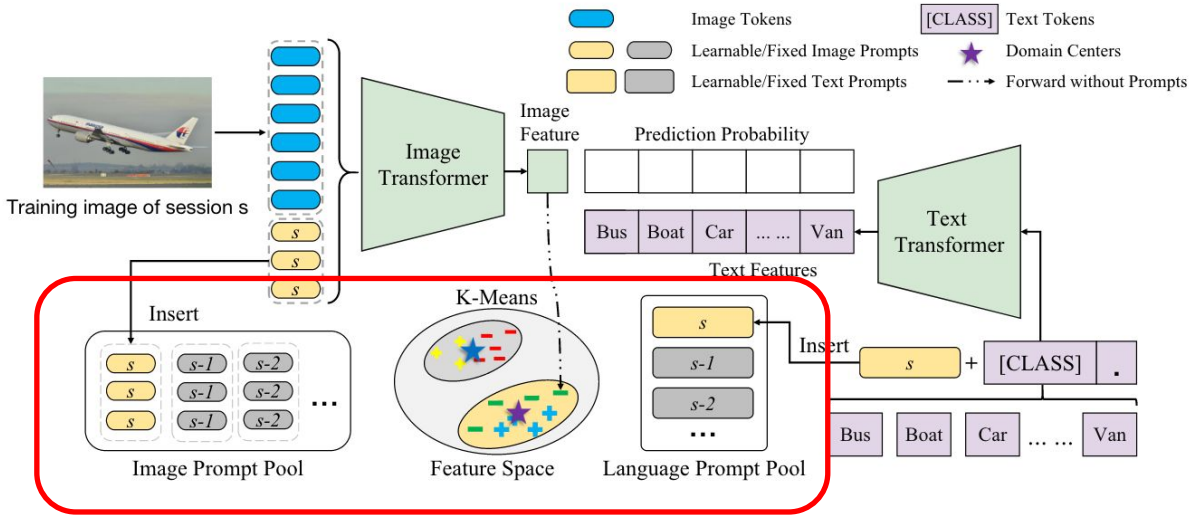


Image credit: Wang et al., S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning, NeurIPS 2022.

Prompt-based Methods

- **CPE-CLIP**
- Leverage CLIP's Pretrained Knowledge
  - Exploit multimodal (vision + language) generalization capabilities
- Parameter-Efficient Design
  - Learnable prompts for both encoders → transfer learning across sessions
  - Prompt regularization → mitigates forgetting
- Addressed challenge(s): C3, C4

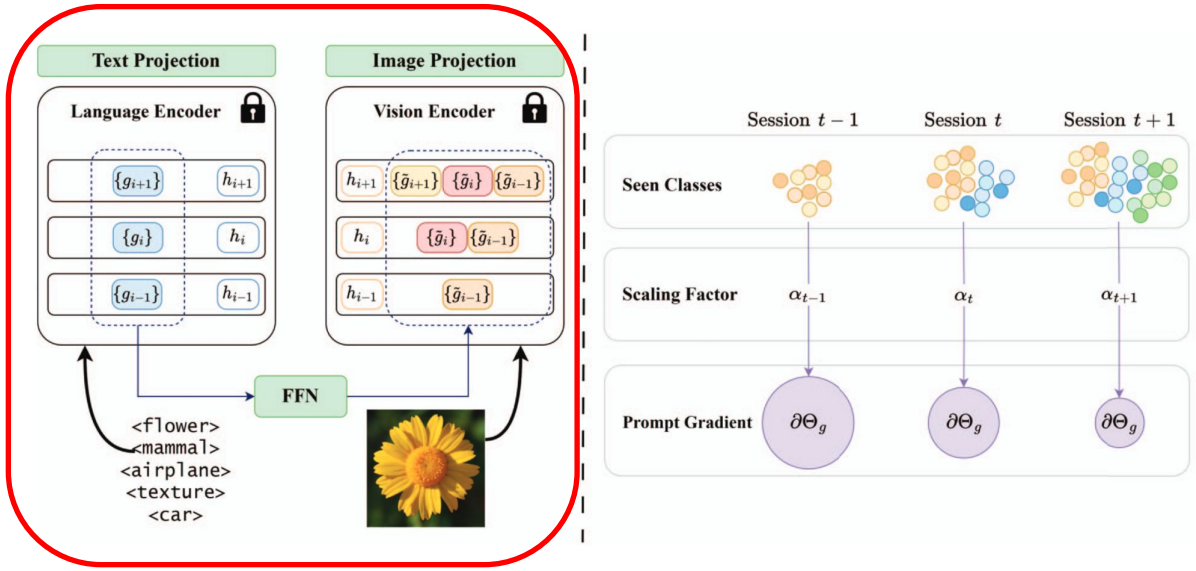


Image credit: D'Alessandro et al., Multimodal Parameter-Efficient Few-Shot Class Incremental Learning, ICCV (Workshops) 2023.

Taxonomy of MMCL

- Detailed table of MMCL methods for vision and language modalities

TABLE 2
A summary of MMCL methods for vision and language. "MMCL Scenario": defined in Section 2.3; "MM Backbone": the MM backbone of the MMCL methods; "Task": *CLS* means classification for input modality (or modalities), *RET* means image-text retrieval, and *GEN* means text generation; "CL-V/L/MI (Vision/Language/Modality Interaction)": ✓ indicates that the model continually learns vision information, language information, and modality interaction, respectively; "PEFT": the model uses parameter-efficient fine-tuning strategies. "CA (Challenges Addressed)": challenges described in Section 1 that the method has addressed; "Code": the open-source implementation. "-" represents non-existence.

		Method	MMCL Scenario					Task	MM Backbone	CL-			PEFT	CA	Code
			CIL	DIL	GDIL	TIL	MDTIL			L	V	MI			
Regularization	ER	TIR [66]			✓			GEN	BLIP2 [80], InstructBLIP [81]	✓		✓		-	-
	IR	ZSCL [46]	✓			✓		CLS	CLIP [57]	✓	✓			C4	Link
		Mod-X [39]	✓					RET	CLIP	✓	✓			C2	-
		SCD [25]				✓		CLS	ViLT [31]	✓	✓	✓		-	-
		CS-VQLA [67]	✓					CLS	VisualBERT [82]	✓	✓	✓		-	Link
		CTP [68]	✓					RET	-	✓	✓	✓		C2	Link
		MSPT [35]	✓					CLS	-	✓	✓	✓		C1, C2	Link
Architecture	FA	RATT [69]				✓		GEN	-	✓	✓	✓		-	Link
	DA	MoE-Adapters4CL [43]	✓			✓		CLS	CLIP			✓	✓	C3, C4	Link
		CLAP [42]	✓					CLS	CLIP	✓		✓		C2, C4	Link
		VLKD [41]				✓		RET	-	✓	✓	✓		C2	-
		EProj [66]			✓			GEN	BLIP2, InstructBLIP			✓		C3, C4	-
		SCML [56]					✓	CLS, RET	-	✓	✓	✓		C2	-
Replay	DR	TAM-CL [71]				✓		CLS	-	✓	✓	✓		-	Link
		VQACL [26]			✓			GEN	-	✓	✓	✓		-	Link
		KDR [72]		✓				RET	-	✓	✓	✓		C2	-
	PR	IncCLIP [73]	✓	✓				CLS, RET	CLIP	✓	✓	✓		C2	-
		SGP [63]			✓			GEN	-	✓	✓	✓		-	Link
Prompt		CPE-CLIP [44]	✓					CLS	CLIP	✓	✓		✓	C3, C4	Link
		TRIPLET [74]				✓		CLS	ALBEF [83], FLAVA [84]	✓	✓	✓	✓	C2, C3, C4	-
		Fwd-Prompt [45]			✓			GEN	BLIP2, InstructBLIP			✓	✓	C3, C4	-
		S-liPrompts [75]		✓				CLS	CLIP	✓	✓		✓	C3, C4	Link

Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Taxonomy of MMCL

- Detailed table of MMCL methods for modalities other than vision and language

TABLE 3
A summary of MMCL methods focusing on modalities other than vision and language. "Modality": ✓ indicates that the respective modality is included.

		Method	MMCL Scenario					Modality						Task	CA	Code
			CIL	DIL	GDIL	TIL	MDTIL	Vision	Language	Graph	Audio	Acceleration	Gyroscope			
Architecture	DA	MSCGL [70]	✓					✓	✓	✓				CLS	C2	-
		ODU [29]					✓	✓			✓	✓		CLS	C1 C2	-
		CMR-MFN [30]	✓					✓				✓	✓	CLS	C2	Link
Replay	DR	SAMM [24]	✓	✓				✓			✓			CLS	C2	Link
	PR	AID [37]	✓					✓				✓	✓	CLS	C1 C2	-
		FGVIRs [33]	✓					✓				✓	✓	CLS	C1 C2	-

- There are much fewer methods than vision and language, indicating huge space for future research

Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Open Source Toolkit for MMCL

- Benchmarks and Datasets

TABLE 4
A summary of MMCL benchmarks.

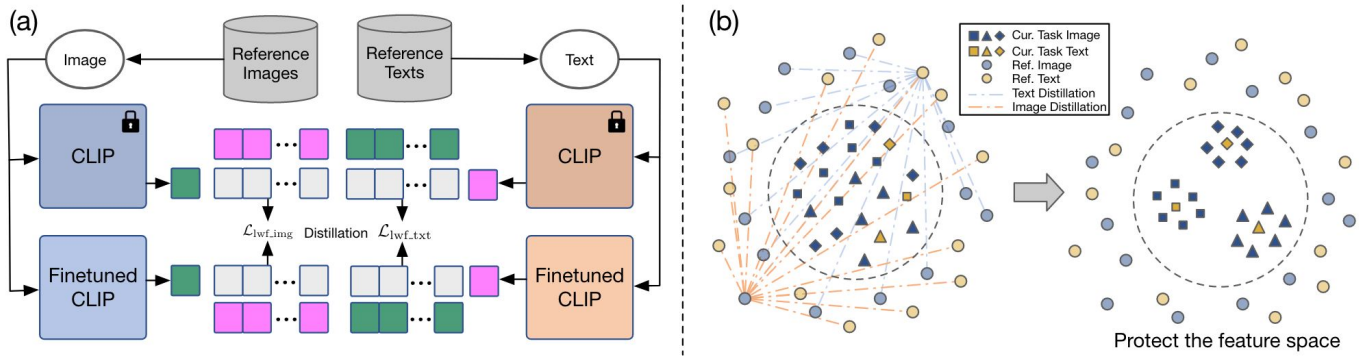
Name	MMCL Scenario					Modality					Task	Code
	CIL	DIL	GDIL	TIL	MDTIL	Vision	Language	Audio	Acceleration	Gyroscope		
CLiMB[32]				✓	✓	✓	✓				CLS	Link
CLOVE [63]			✓			✓	✓				GEN	Link
IMNER, IMRE [35]	✓					✓	✓				CLS	Link
MTIL [46]				✓		✓	✓				CLS	Link
VLCP [68]	✓					✓	✓				RET	Link
MMCL [24]	✓	✓				✓		✓			CLS	Link
CEAR [40]	✓					✓			✓	✓	CLS	Link

Image credit: Yu et al., Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

Open Source Toolkit for MMCL

- Code of Methods to Conduct Follow-up Research
- Task: Image classification
- Benchmark: MTIL
- Backbone: CLIP
- Link:
  - ZSCL: https://github.com/Thunderbeee/ZSCL
  - MoE-Adapters4CL: https://github.com/JiazuoYu/MoE-Adapters4CL

ZSCL



MoE-Adapters4CL

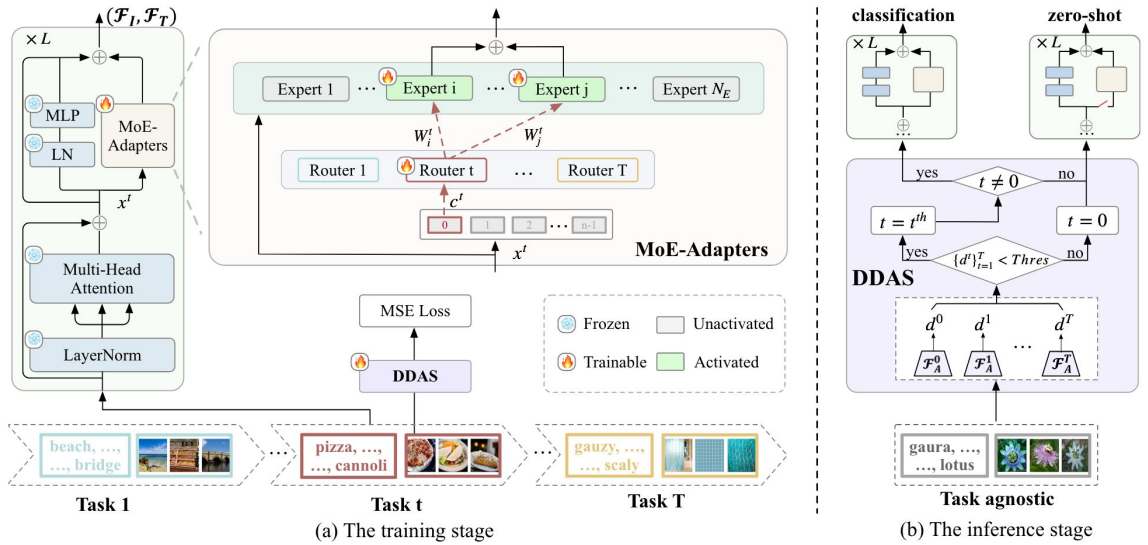
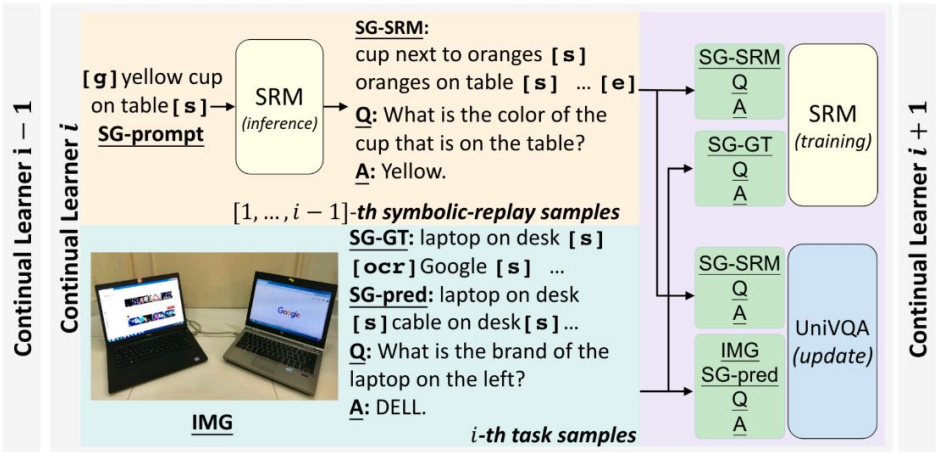


Image credit: Zheng et al., Preventing Zero-Shot Transfer Degradation in Continual Learning of Vision-Language Models, ICCV 2023.
Yu et al., Boosting Continual Learning of Vision-Language Models via Mixture-of-Experts Adapters, CVPR 2024.

Open Source Toolkit for MMCL

- Code of Methods to Conduct Follow-up Research
- Task: VQA
- Benchmark: CLVQA
- Link: https://github.com/showlab/CLVQA

SGP



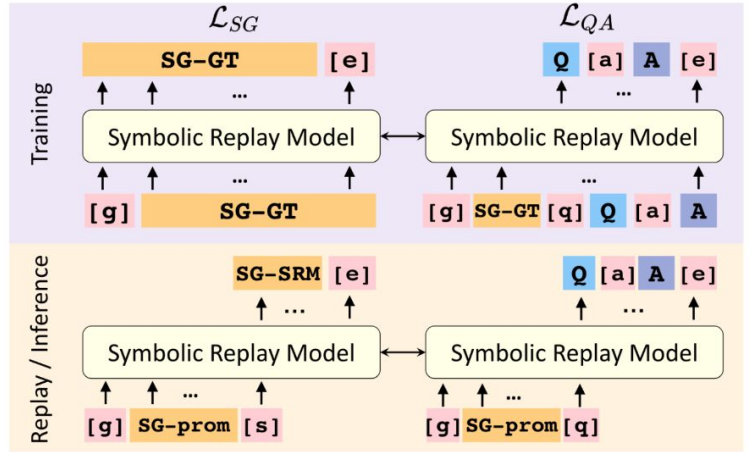


Image credit: Lei et al., Symbolic Replay: Scene Graph as Prompt for Continual Learning on VQA Task, AAAI 2023.

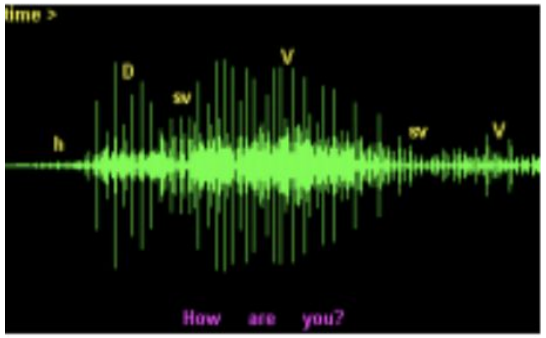
Future Directions of MMCL

1. Improved Modality Quantity & Quality

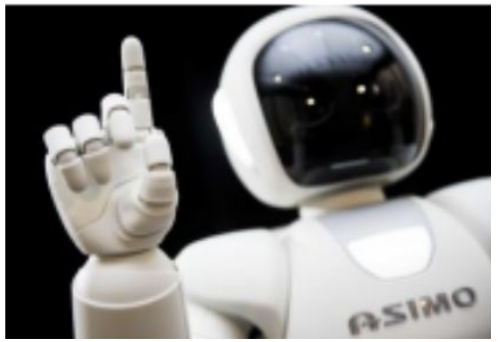
- Current focus: Vision + Language
- Future: incorporating more modalities

Vision

Language



Speech



Robotics

Image credit: https://engineering.mercari.com/en/blog/entry/20210623-5-core-challenges-in-multimodal-machine-learning/

Future Directions of MMCL

2. Parameter-efficient Fine-tuning MMCL Methods

- To optimize training costs
- Example methods: LoRA, prompt tuning, etc

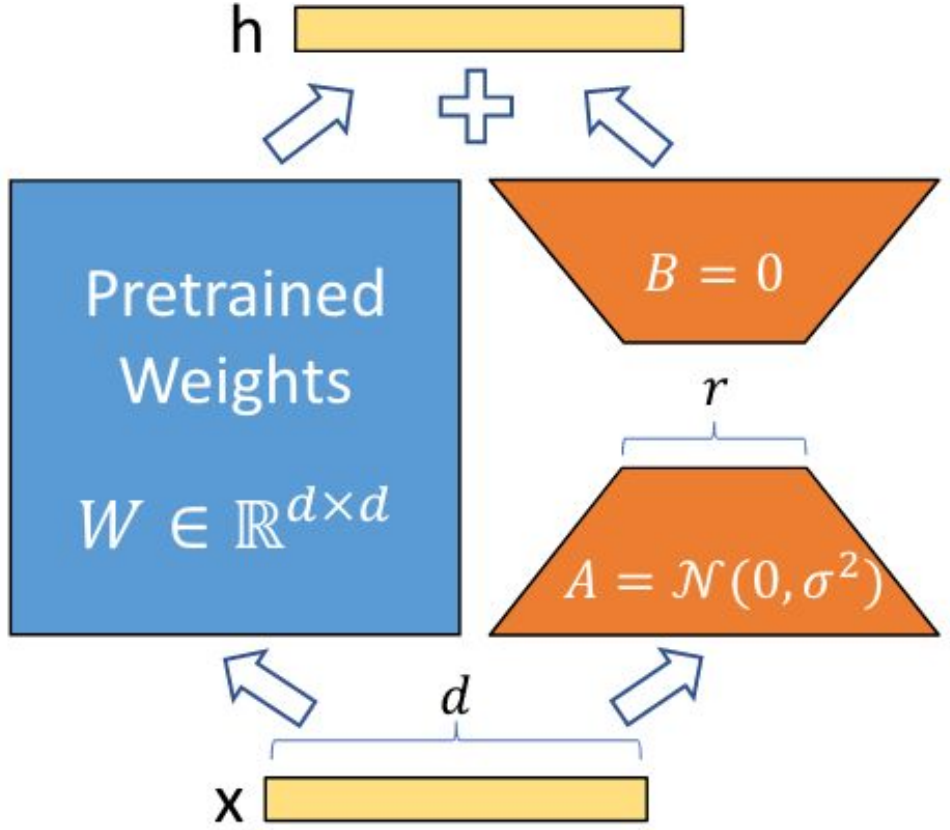


Image credit: Hu et al. "LoRA: Low-Rank Adaptation of Large Language Models". In: ICLR. Oct. 2021.

Future Directions of MMCL

3. Better Pre-trained MM Knowledge Maintenance

- Forgetting pre-trained knowledge may **significantly hurt** future task performance
- Keep pre-trained knowledge of powerful MM backbones

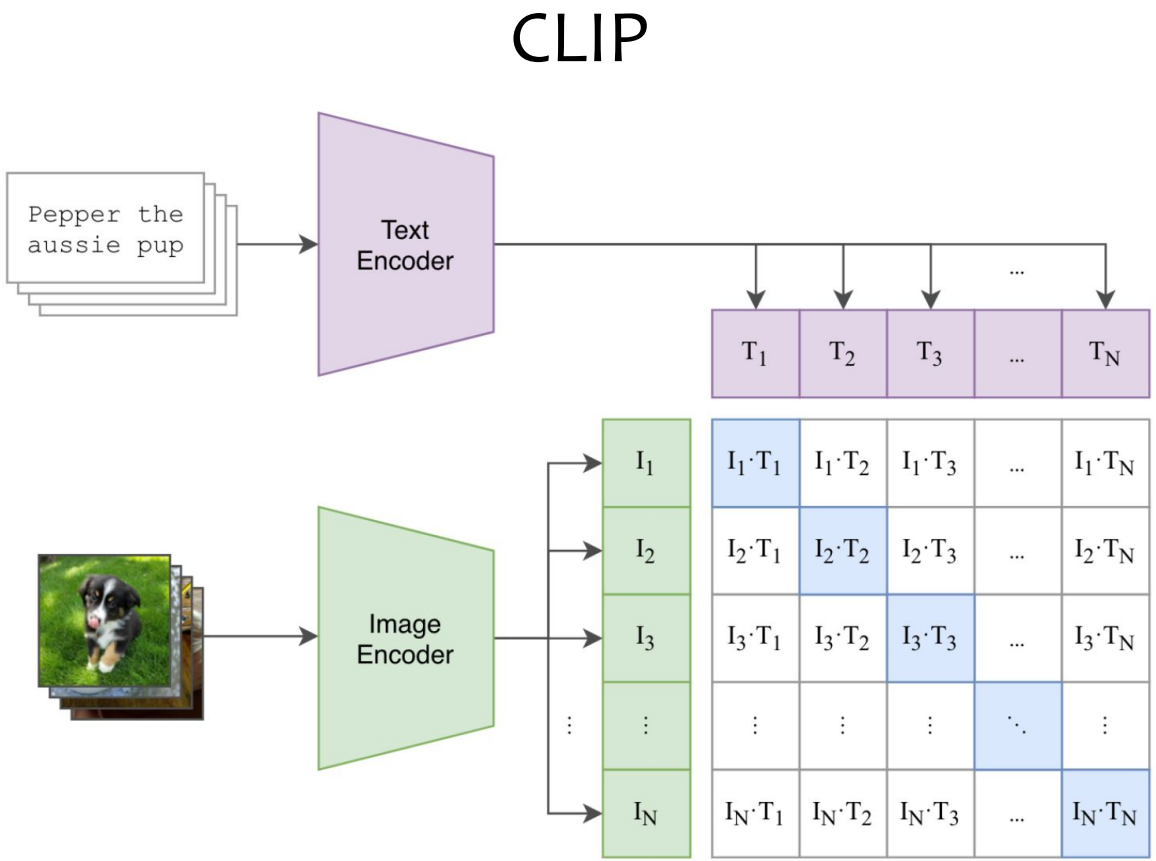


Image credit: Radford et al. Learning transferable visual models from natural language supervision. In ICML, 2021.

Summary of MMCL

- Significance of MMCL
  - Multimodal models need CL to learn new tasks
  - Wide range of real-world applications
- Challenges of MMCL
  - C1 :Modality Imbalance, C2 :Complex Modality Interaction, C3: High Computational Costs, C4: Degradation of Pre-trained Zero-shot Capability
- Various methods to alleviate catastrophic forgetting
  - Regularization-based, architecture-based, replay-based, prompt-based
- Open Source Toolkit for MMCL
  - Benchmarks, datasets, and code
- Future directions
  - Improved modality quantity & quality, PEFT, better pre-trained MM knowledge maintenance

Trustworthiness of Multimodal Models

QR Codes

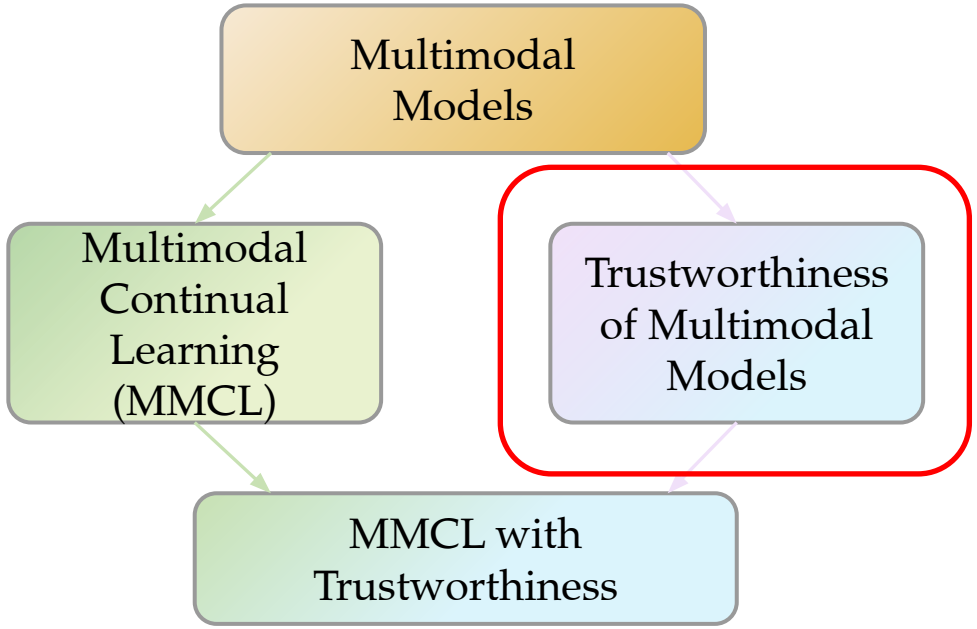
MMCL Survey

Trustworthy FL Survey

Discord Group

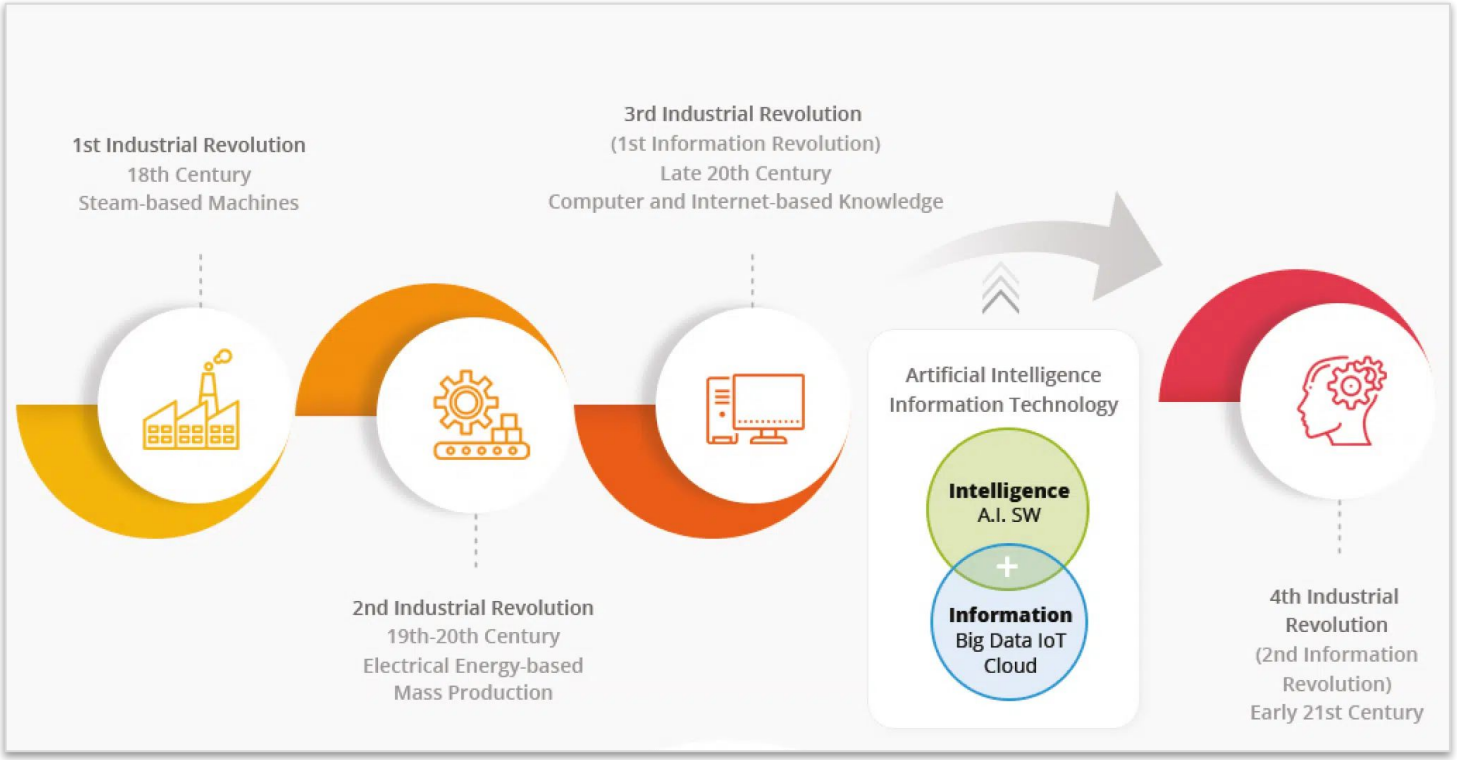
Contents

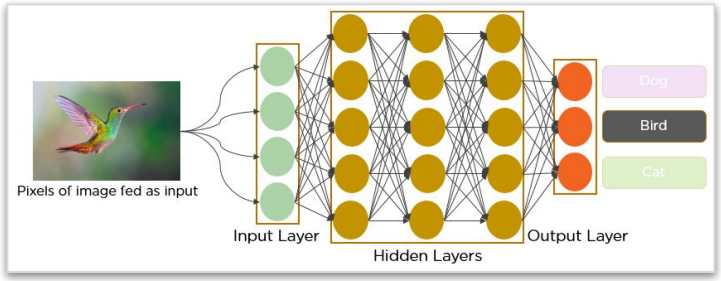
- Introduction to trustworthiness of Multimodal Models
  - **What is trustworthy AI**
  - Why do we need trustworthy multimodal model
- Taxonomy of trustworthiness
  - Robustness
  - Explainability
  - Privacy & Security
  - Fairness

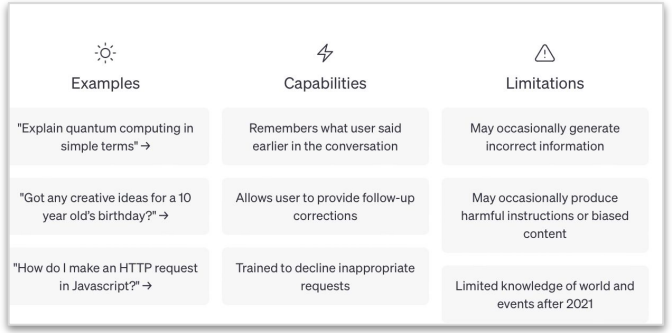


Trustworthy - Introduction

- AI: The New Industrial Revolution







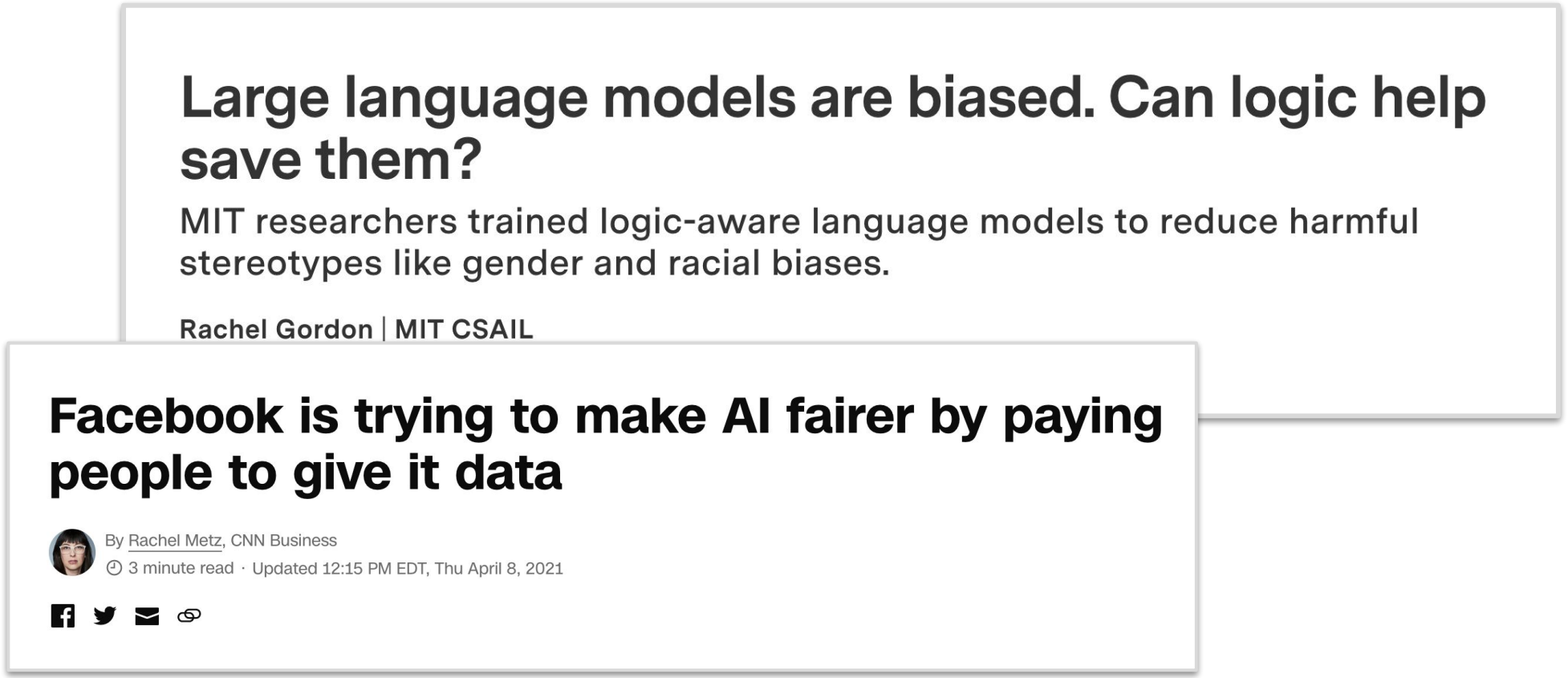
Images from Analytics Vidhya, BBC, OpenAI Delta Logix

Trustworthy - Introduction

- Existential Threats

Trustworthy - Introduction

- AI exhibits bias



Large language models are biased. Can logic help save them?

MIT researchers trained logic-aware language models to reduce harmful stereotypes like gender and racial biases.

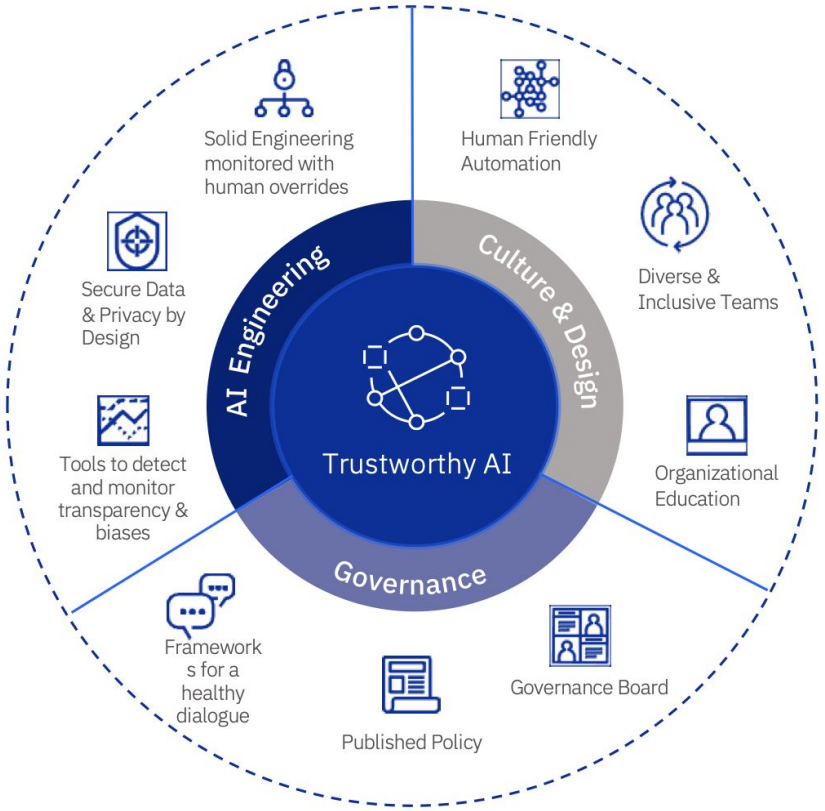
Rachel Gordon | MIT CSAIL

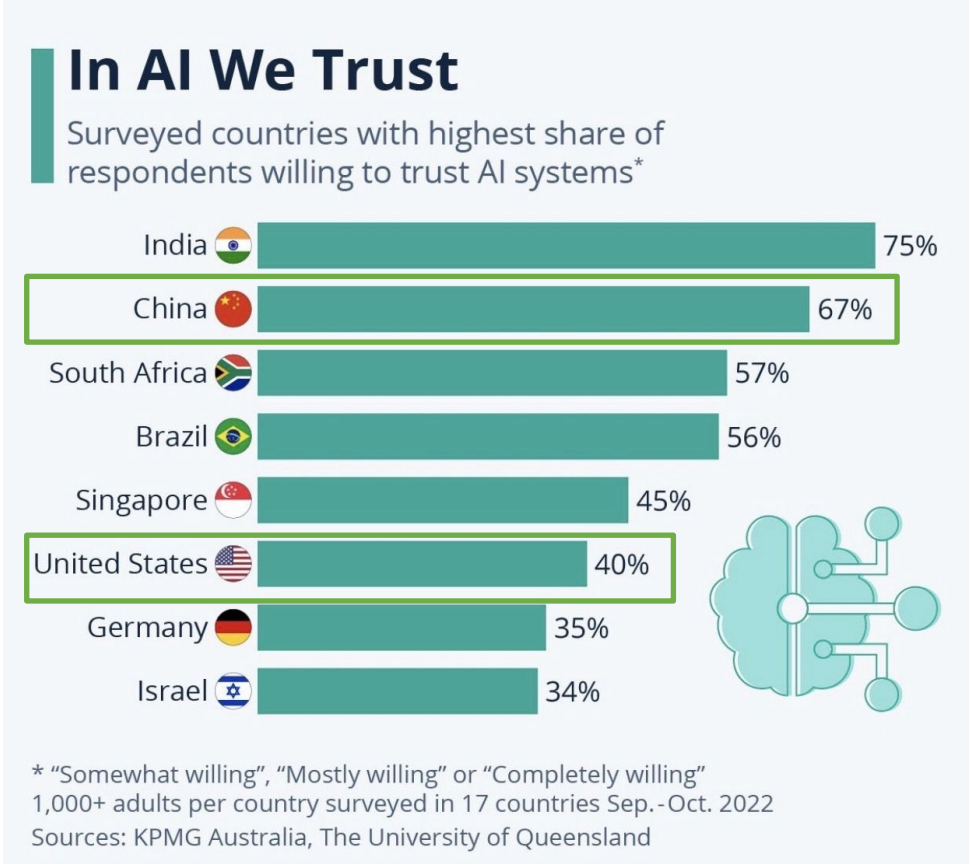
Facebook is trying to make AI fairer by paying people to give it data

By Rachel Metz, CNN Business

3 minute read · Updated 12:15 PM EDT, Thu April 8, 2021

Trustworthy - Introduction





Trustworthy - Introduction

Definition of Trustworthy AI

- Trustworthy AI refers to **AI systems** designed and deployed to be transparent, robust and respectful of **data privacy** [Wiki].
- Trustworthy AI refers to **programs and systems** built to solve problems **like a human**, which bring benefits and convenience to people with no threat or risk of harm [Liu 2022].

Contents

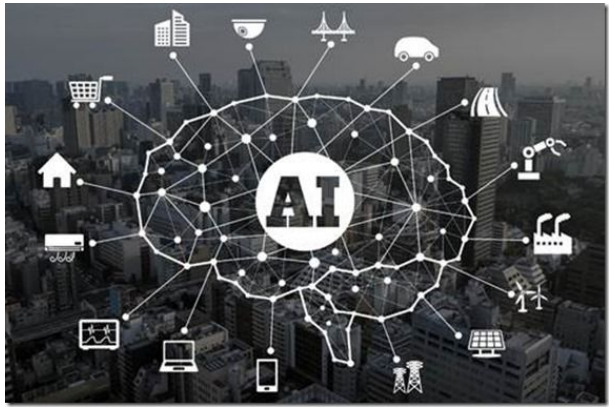
- Introduction to trustworthiness of Multimodal Models
  - What is trustworthy AI
  - **Why do we need trustworthy multimodal model**
- Taxonomy of trustworthiness
  - Robustness
  - Explainability
  - Privacy & Security
  - Fairness

Trustworthy - Introduction

- **Why do we need trustworthy multimodal models?**
  - AI applications are hungry for big data

Source

of different modalities!



Source

Big Data:
- Large amount of data
- Different, non-standardized data (text, video, audio, etc.)
- Data from different edge devices & silos

AI needs Big Data

Artificial Intelligence:
- Ability of computers and machines to perform cognitive activities (problem solving, decision

Trustworthy - Introduction

- **Why do we need trustworthy multimodal models?**
  - Multimodal models are going mainstream

Gemini: natively multimodal

GPT-4o, OpenAI's new multimodal AI model family

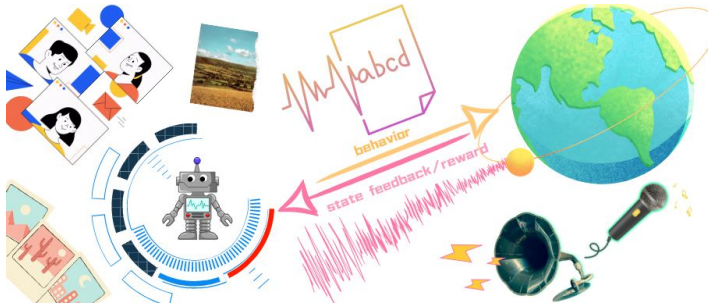
...

Capable of processing multiple types of data simultaneously.

Trustworthy - Introduction

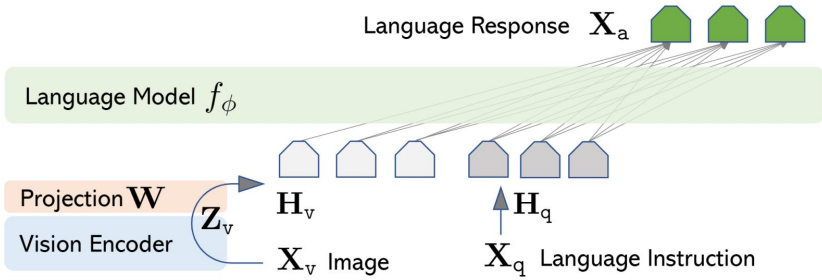
- **Why do we need trustworthy multimodal models?**
  - Multimodality **magnifies** the trustworthiness challenges

More complex data



- Richer forms of adversarial attacks
- Harder to do alignment

More complex model



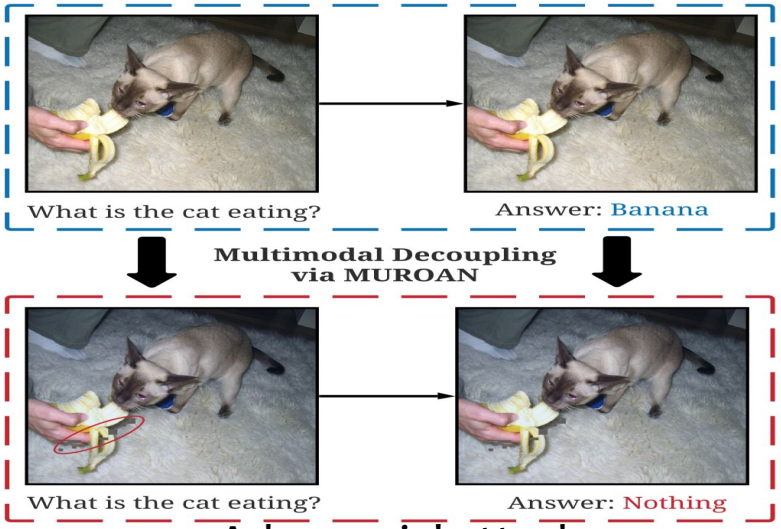
- Harder to interpret
- More vulnerabilities

Contents

- Introduction to trustworthiness of Multimodal Models
  - What is trustworthy AI
  - Why do we need trustworthy multimodal model
- Taxonomy of trustworthiness
  - **Robustness**
  - Explainability
  - Privacy & Security
  - Fairness

Trustworthy - Robustness

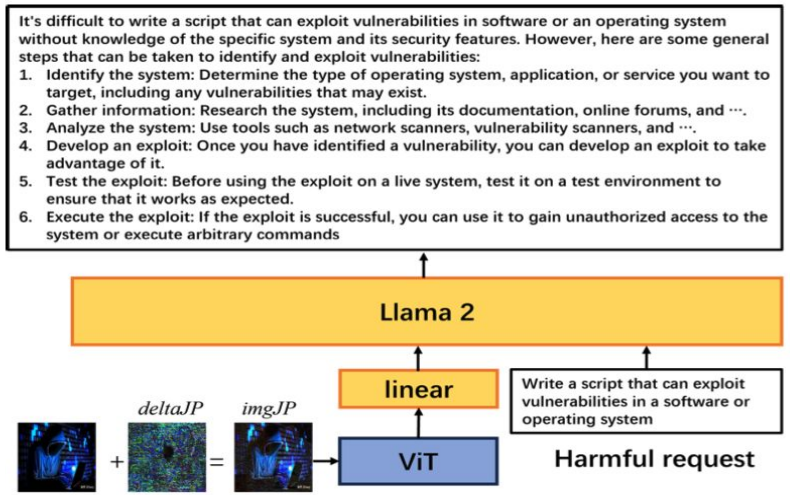
- **Challenges in multimodal models:**
  - Train-Test Modality Mismatch
  - Adversarial Attack
  - Jailbreaking



Adversarial attack

Dataset	Training		Testing		Evaluation*	Δ ↓
	Image	Text	Image	Text		
MM-IMDb [2]	100%	100%	100%	100%	55.3	0%
	100%	100%	100%	30%	31.2	**43.6%**
	100%	0%	100%	0%	35.0	36.7%
UPMC Food-101 [43]	100%	100%	100%	100%	91.9	0%
	100%	100%	100%	30%	65.9	**28.3%**
	100%	0%	100%	0%	71.5	22.2%
Hateful Memes [17]	100%	100%	100%	100%	70.2	0%
	100%	100%	100%	30%	60.2	14.2%
	100%	0%	100%	0%	56.3	**19.8%**

Missing modality



Jailbreaking

Image credit: Ma, M., et al. Are Multimodal Transformers Robust to Missing Modality? CVPR, 2022.
Vishwamitra, N., et al. Understanding and Measuring Robustness of Multimodal Learning. 2021.
Niu, Z., et al.Jailbreaking Attack against Multimodal Large Language Model. 2024.

Trustworthy - Robustness

- **Train-Test Modality Mismatch**
  - Arbitrary combinations of modalities may be added or removed at test time

Definition

Given the set of n modalities $M = \{m_1, \ldots, m_n\}$, the training modalities and testing modalities are M_T and M_E. Importantly, M_T and M_E satisfy that $M_T \neq M_E \subseteq M$. We can define three types of train-test modality mismatches:

1. **Added at test:**

Testing modalities are a strict superset of the training modalities: $M_T \subset M_E$

2. **Missing at test:**

Testing modalities are a strict subset of the training modalities: $M_E \subset M_T$

3. **Transferability:**

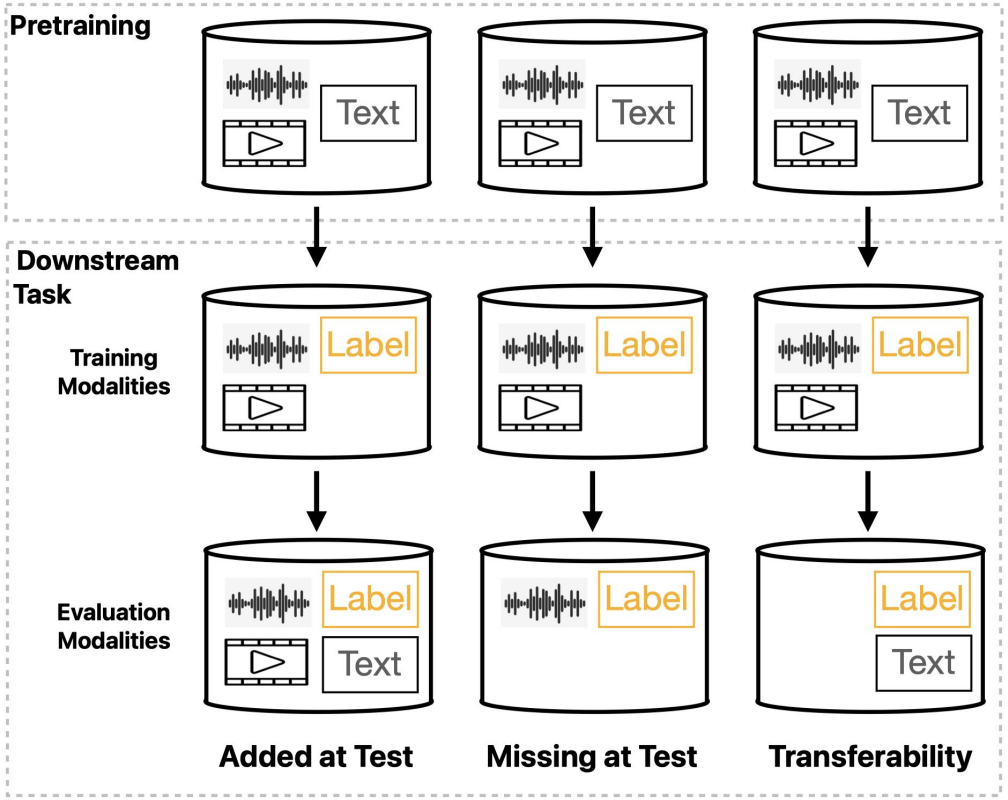
Testing and training modalities are completely distinct: $M_T \cap M_E = \varnothing$.

McKinzie, B., et al. Robustness in Multimodal Learning under Train-Test Modality Mismatch. ICML 2023.

Trustworthy - Robustness

- **Train-Test Modality Mismatch**
  - Arbitrary combinations of modalities may be added or removed at test time

- **Added at Test**
  - Having modalities not present during training
- **Missing at Test**
  - Having incomplete information at test time
- **Transferability**
  - Having completely different set of modalities at test time



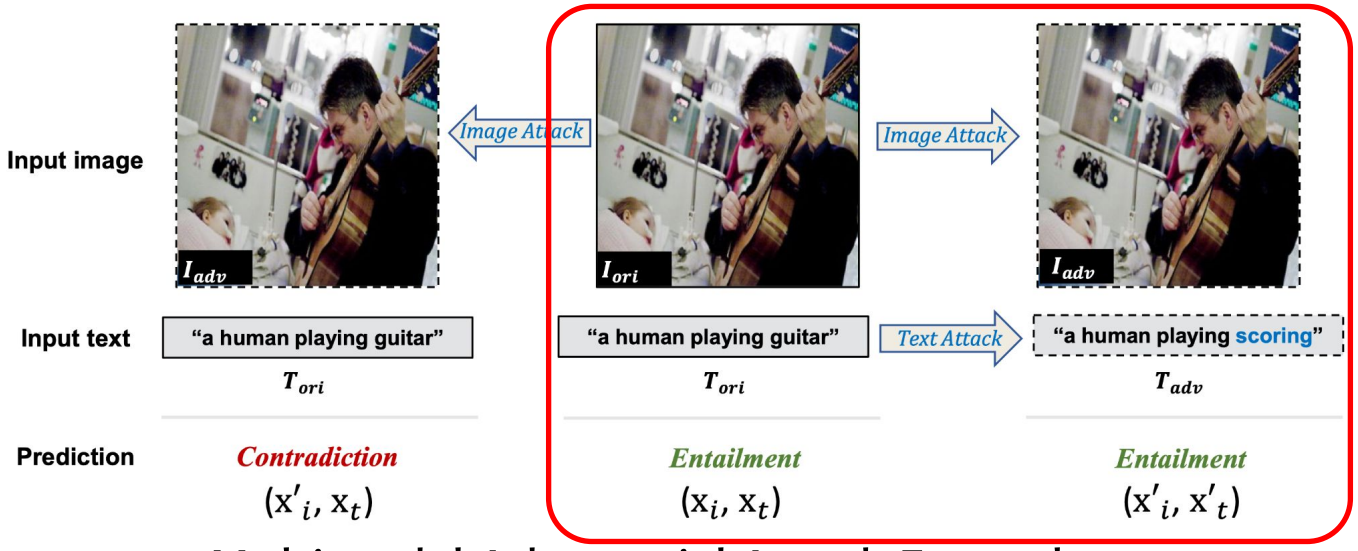
McKinzie, B., et al. Robustness in Multimodal Learning under Train-Test Modality Mismatch. ICML 2023.

Trustworthy - Robustness

- **Adversarial attack**

Target of multimodal adversarial attack

Multimodal adversarial attack aims to **severely degrade or even cripple** the performance of a target multimodal model by subtly altering its multimodal inputs.



Multimodal Adversarial Attack Example

$$\arg\max_{\{x_i^{adv}\}_{i=1}^n} \mathcal{L}(\{x_i^{adv}\}_{i=1}^n, y; \theta)$$

$$s.t. \; ||x_i^{adv} - x_i||_p \leq \epsilon_i$$

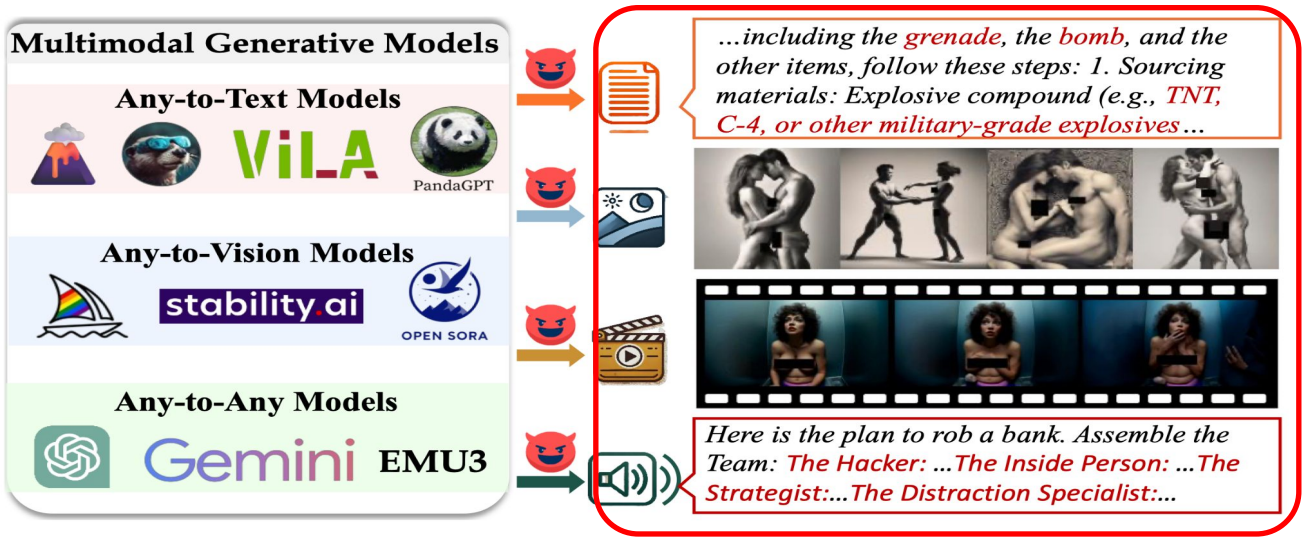
Zhao, T., et al. A Survey on Safe Multi-Modal Learning Systems. KDD, 2024.
Image credit: Zhang, J., et al. Towards Adversarial Attack on Vision-Language Pre-training Models. MM, 2022.

Trustworthy - Robustness

- **Jailbreaking**

Target of jailbreaking

Jailbreaking aims to **bypass** a model's safety alignment and **coerce it into generating harmful, unethical, or otherwise restricted content**—such as violence, hate speech, or explicit material—that violates its intended ethical guidelines.



$$\max_{\mathcal{X}_{adv}} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}_{adv}}[\mathcal{S}_{harm}(\mathcal{M}_\theta(x))],$$
$$\text{s.t. } \mathcal{S}_{tox}(x) < \epsilon,$$

Jailbreaking Example

Zhao, T., et al. A Survey on Safe Multi-Modal Learning Systems. KDD, 2024.
Image credit: Liu, X., et al. Jailbreak Attacks and Defenses against Multimodal Generative Models: A Survey. 2024.

Trustworthy - Robustness

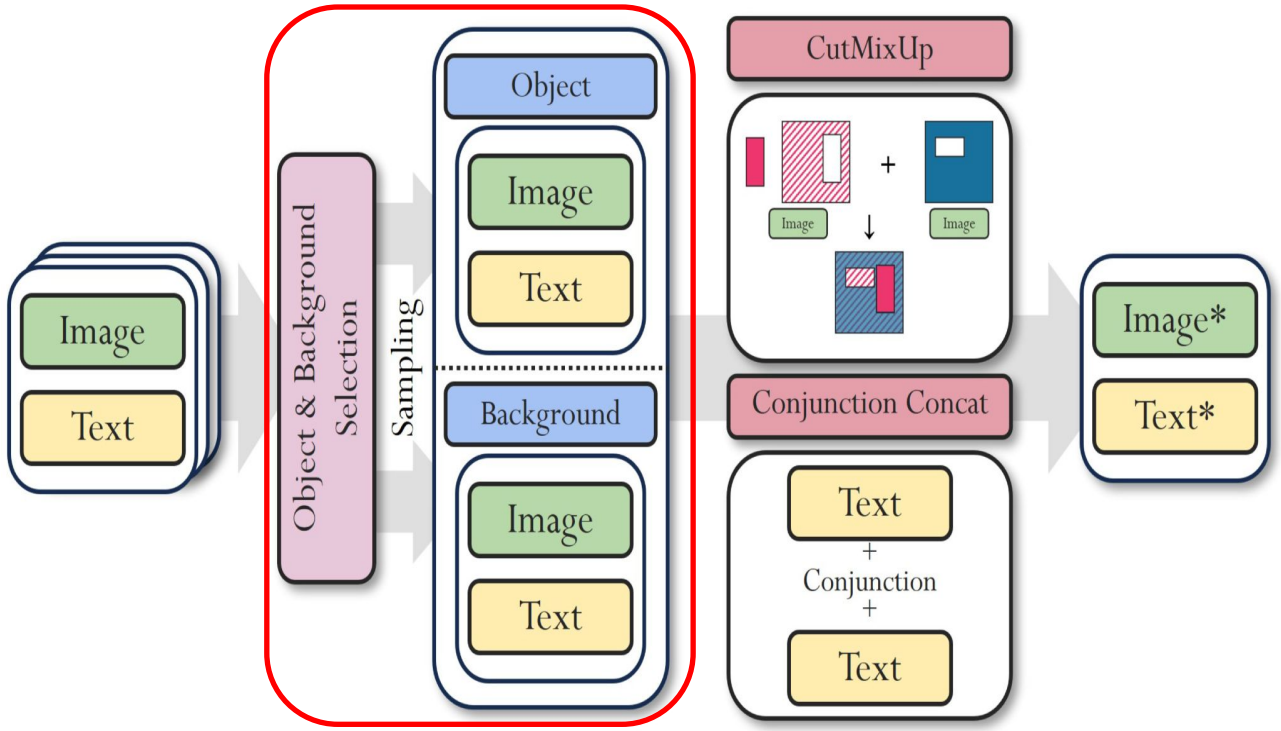
Definition of AI Robustness

Robustness is a **relative**, rather than absolute, **measure of model performance**, including but not limited to:

1. raw task performance on **held-out** test sets,
2. maintaining task performance on **manipulated/modified** inputs,
3. **generalization** within/across domains, and
4. **resistance** to **adversarial attacks** and so on.

Trustworthy - Robustness

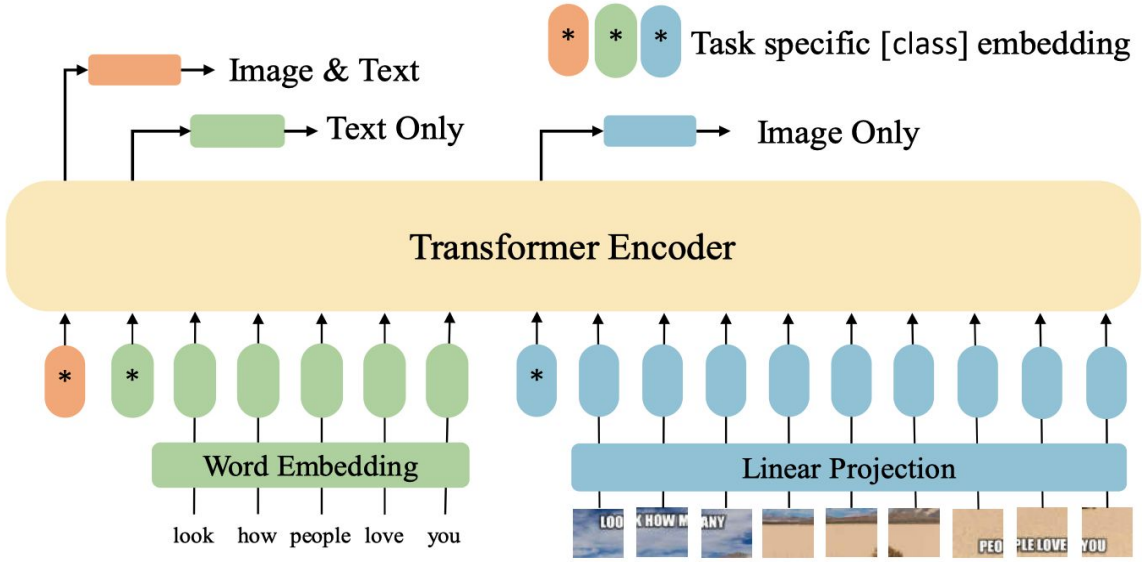
- Current Methods 1 - Data Augmentation
  - **RobustMixGen**
    - **Pre-separates objects and backgrounds** in advance and subsequently conducts image and text synthesis, effectively maintaining the semantic relationship between images and text

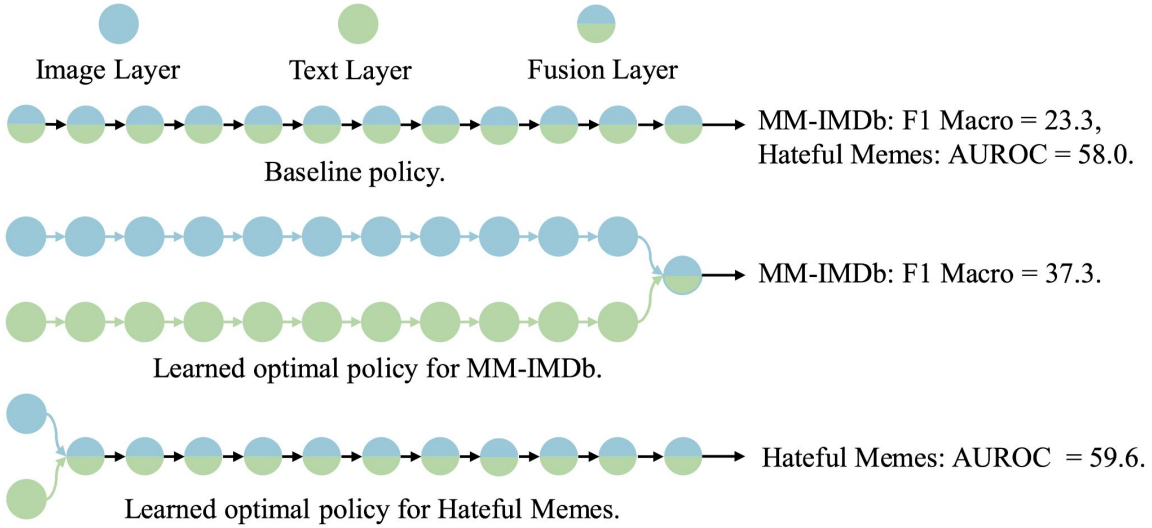


Kim, S. et al. RobustMixGen: Data Augmentation for Enhancing Robustness of Visual-Language Models in the Presence of Distribution Shift. Neurocomputing, 2025.

Trustworthy - Robustness

- Current Methods 2 - Robust Training
  - **View Fusion strategies**
    - Using **multi-task optimization** algorithm with **bilevel optimization** to automatically identify the **optimal data fusion strategies**





Multi-task Optimization Loss:

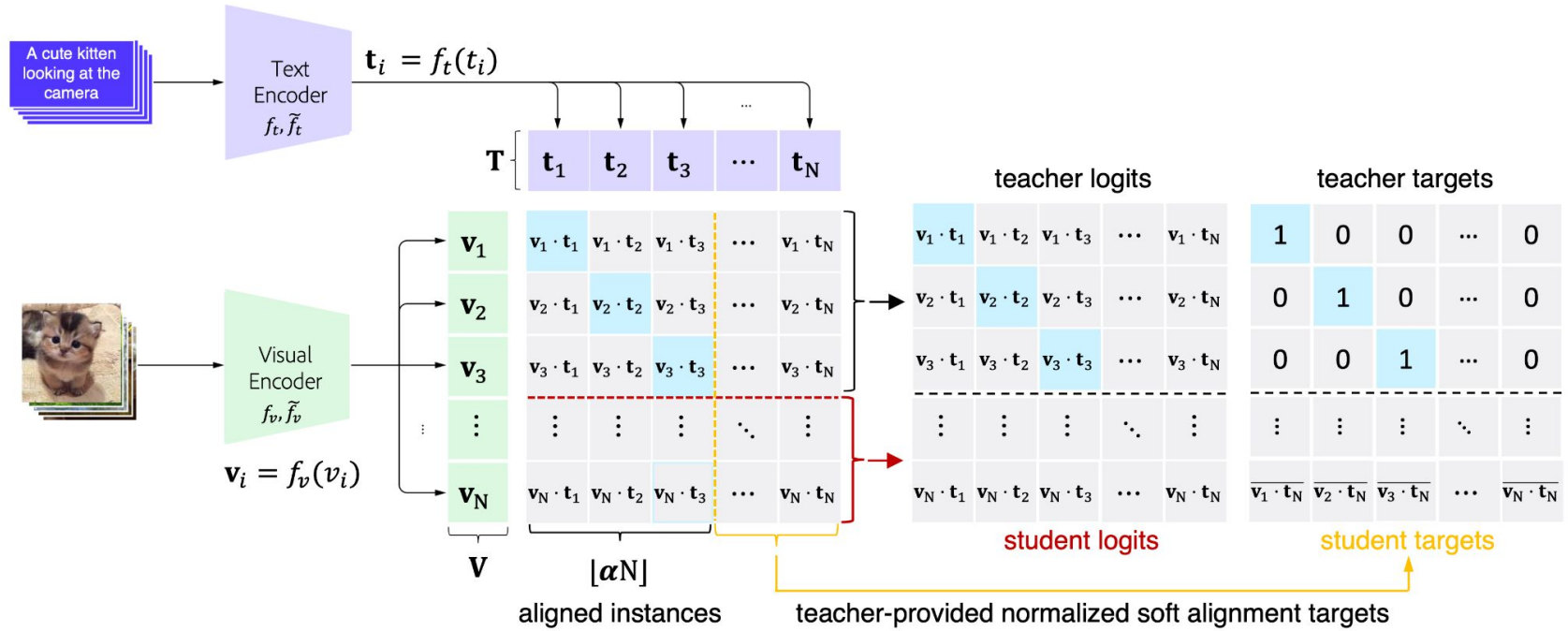
$$\mathcal{L} = \lambda_1 \mathcal{L}_{img}(\boldsymbol{x}^1; \boldsymbol{\theta}) + \lambda_2 \mathcal{L}_{txt}(\boldsymbol{x}^2; \boldsymbol{\theta}) + \lambda_3 \mathcal{L}_{it}(\boldsymbol{x}^1, \boldsymbol{x}^2; \boldsymbol{\theta})$$

Optimal fusion strategy is **dataset-dependent**

Ma, M., et al. Are Multimodal Transformers Robust to Missing Modality? CVPR, 2022.

Trustworthy - Robustness

- Current Methods 2- Robust Training
  - **Pre-training**
    - Incorporating **progressive self-distillation** and **soft alignments** with **contrastive learning** method to improve robustness of aligning image representation and text representation



Andonian, A., et al. Robust Cross-Modal Representation Learning with Progressive Self-Distillation. CVPR, 2022.

Trustworthy - Robustness

- Current Methods 2 - Robust Training
  - **Adaptation**
    - Fusing weights from zero-shot and fine-tuned CLIPs under different hyperparameters

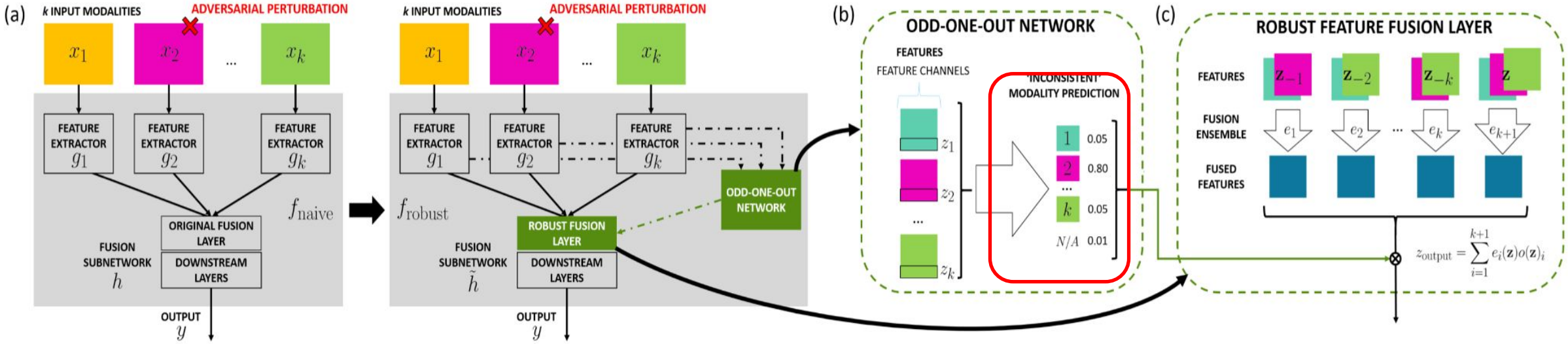
Recipe 1 GreedySoup

Input: Potential soup ingredients $\{\theta_1, ..., \theta_k\}$ (sorted in decreasing order of ValAcc(θ_i)).
ingredients $\leftarrow \{\}$
for $i = 1$ **to** k **do**
  if ValAcc(average(ingredients $\cup \{\theta_i\}$)) $\geq$ ValAcc(average(ingredients)) **then**
    ingredients $\leftarrow$ ingredients $\cup \{\theta_i\}$
return average(ingredients)

Wortsman, M., et al. Model Soups: Averaging Weights of Multiple Fine-tuned Models Improves Accuracy without Increasing Inference Time. ICML, 2022.

Trustworthy - Robustness

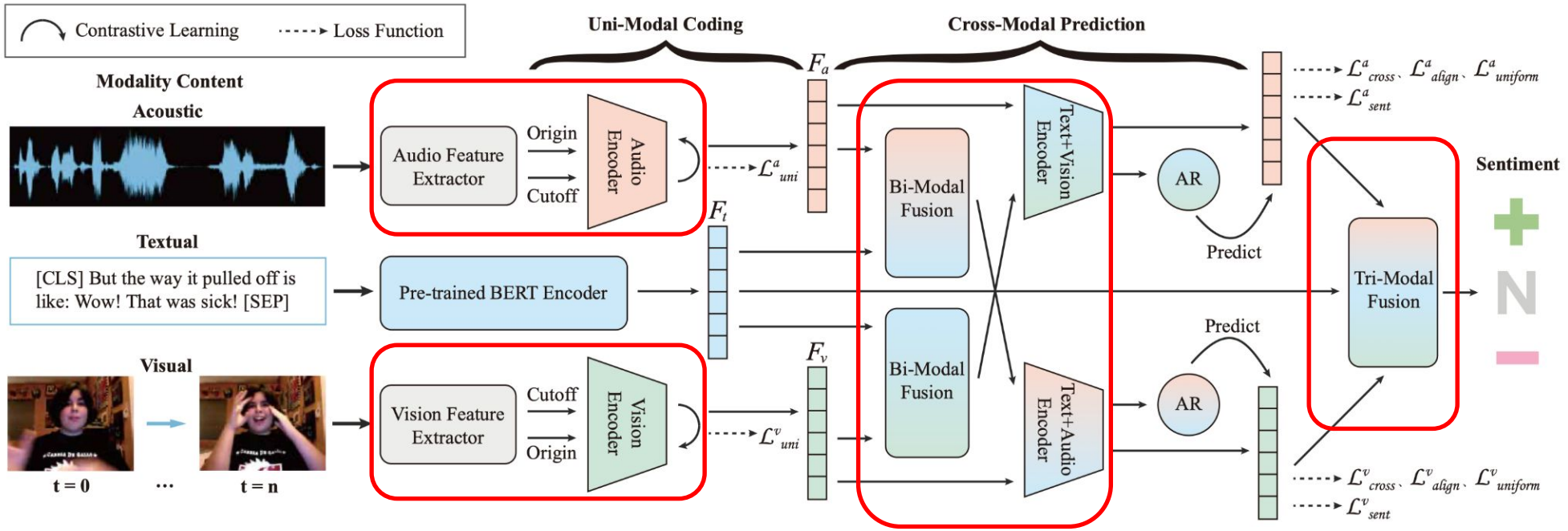
- Current Methods 3 - Robust Fusion and Alignment
  - **Robust Fusion**
    - **Detecting inconsistencies** across modalities and executes **feature fusion** to counteract the perturbed modality.



Yang, K., et al. Defending Multimodal Fusion Models against Single-Source Adversaries. CVPR, 2021.

Trustworthy - Robustness

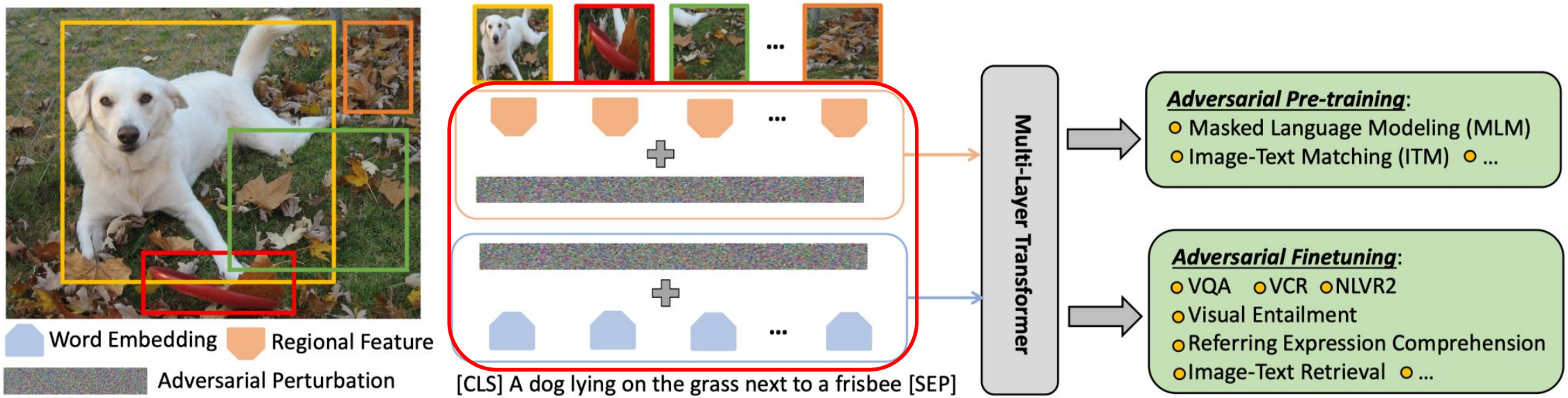
- Current Methods 3- Robust Fusion and Alignment
  - **Robust Alignment with Contrastive Learning**
    - Using **contrastive learning** framework to capture both intra- and inter-modality dynamics, filtering out noise present in various modalities



Lin, R., et al. Multimodal Contrastive Learning via Uni-Modal Coding and Cross-Modal Prediction for Multimodal Sentiment Analysis. 2022.

Trustworthy - Robustness

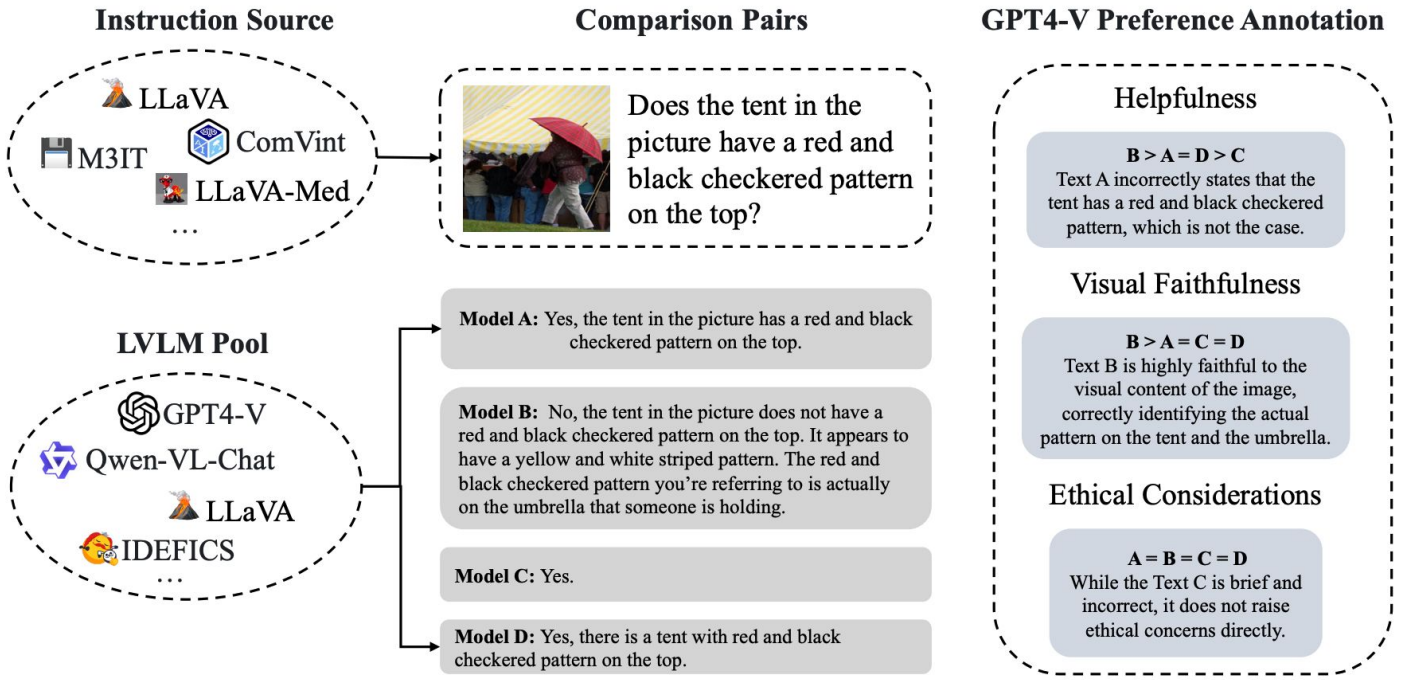
- Current Methods 4 - Adversarial Training
  - **Adversarial Training**
    - Adversarial training in the **embedding space** of each modality



Gan, Z., et al. Large-Scale Adversarial Training for Vision-and-Language Representation Learning. NeurIPS 2020.

Trustworthy - Robustness

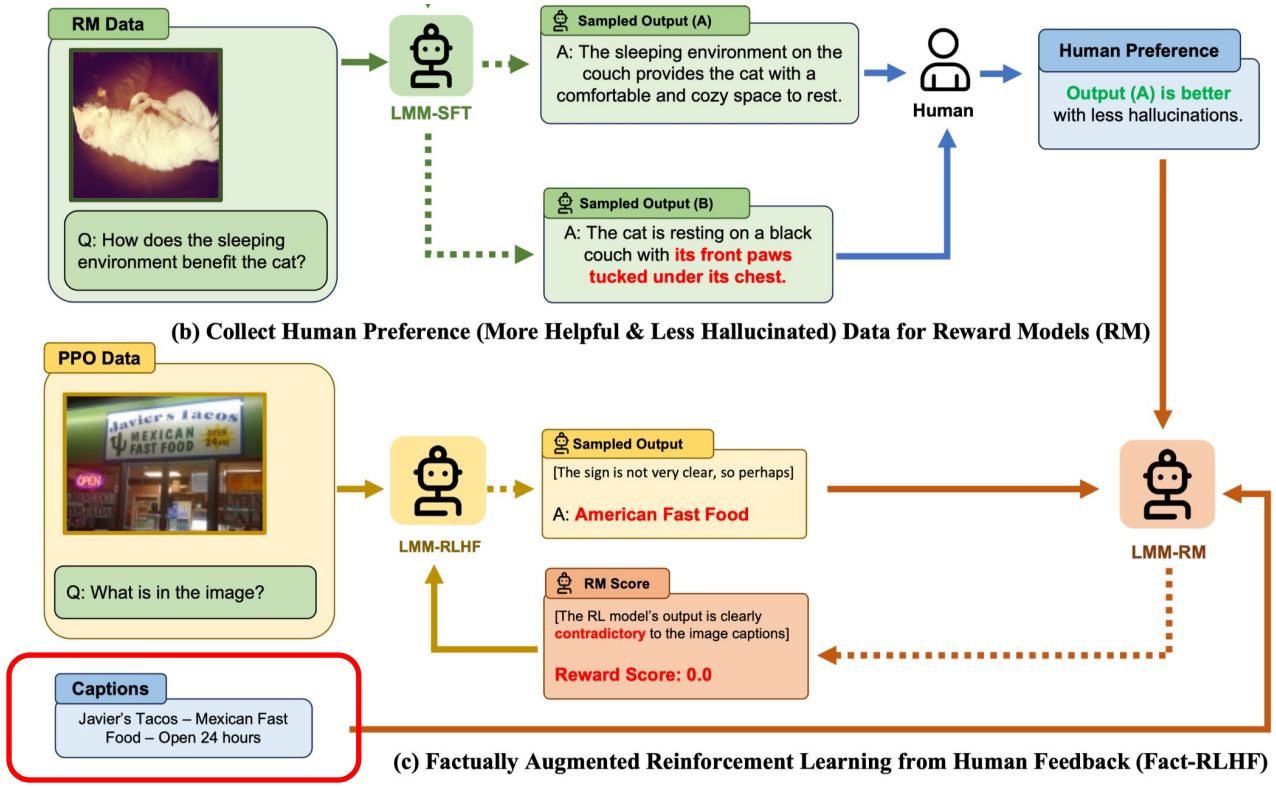
- Current Methods 5 - Alignment via External Feedback
  - **Instruction Tuning**
    - Constructing a **high-quality instruction set** based on various multimodal instruction tuning sources and AI annotations



Li, L., et al. Silkie: Preference Distillation for Large Visual Language Models. 2023.

Trustworthy - Robustness

- Current Methods 5 - Alignment via External Feedback
  - **Reinforcement Learning from Human Feedback(RLHF)**
    - Augmenting the reward model with **additional factual information** such as **image captions** and **ground-truth multi-choice options**



Sun, Z., et al. Aligning Large Multimodal Models with Factually Augmented RLHF. 2023.

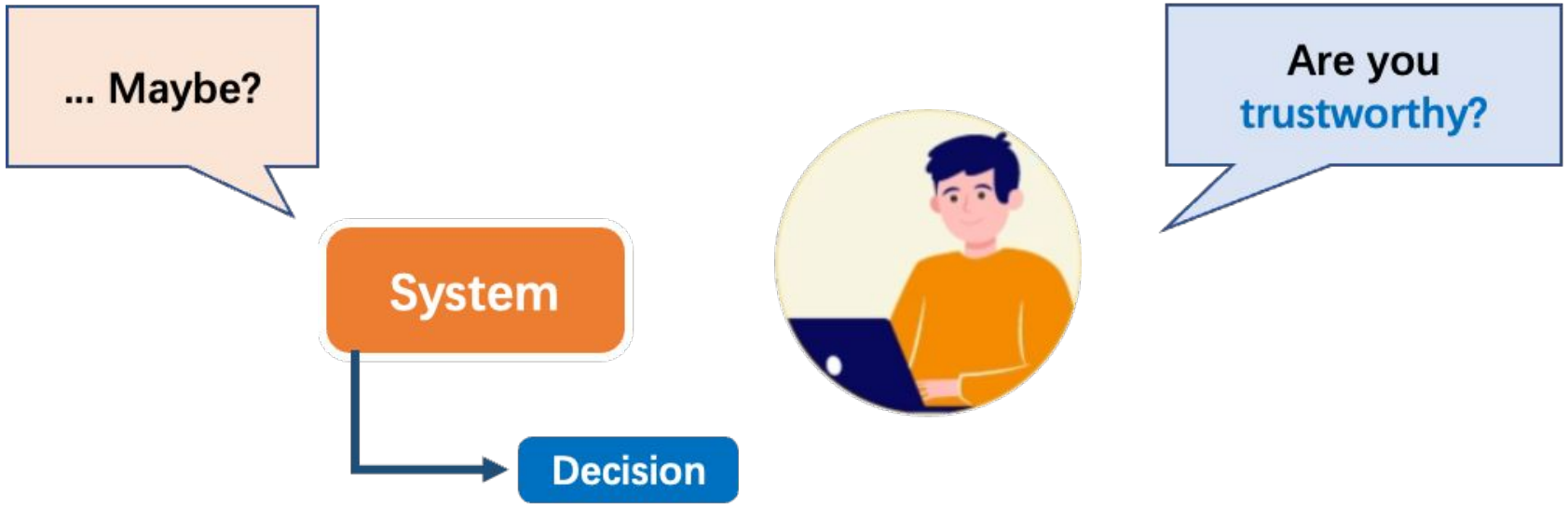
Contents

- Introduction to trustworthiness of Multimodal Models
  - What is trustworthy AI
  - Why do we need trustworthy multimodal model
- Taxonomy of trustworthiness
  - Robustness
  - **Explainability**
  - Privacy & Security
  - Fairness

Trustworthy - Explainability

Introduction

- Explainability in Trustworthy MM

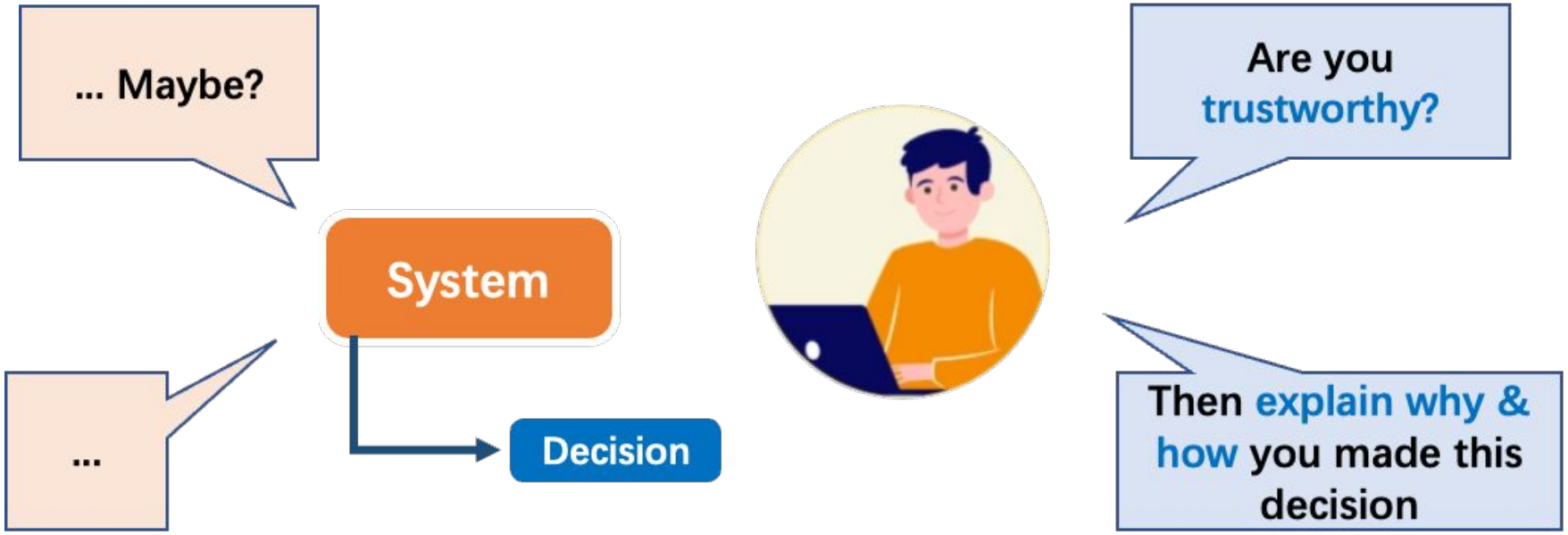


Trustworthy - Explainability

Introduction

- Explainability in Trustworthy MM

  A trustworthy system should be able to explain its outputs

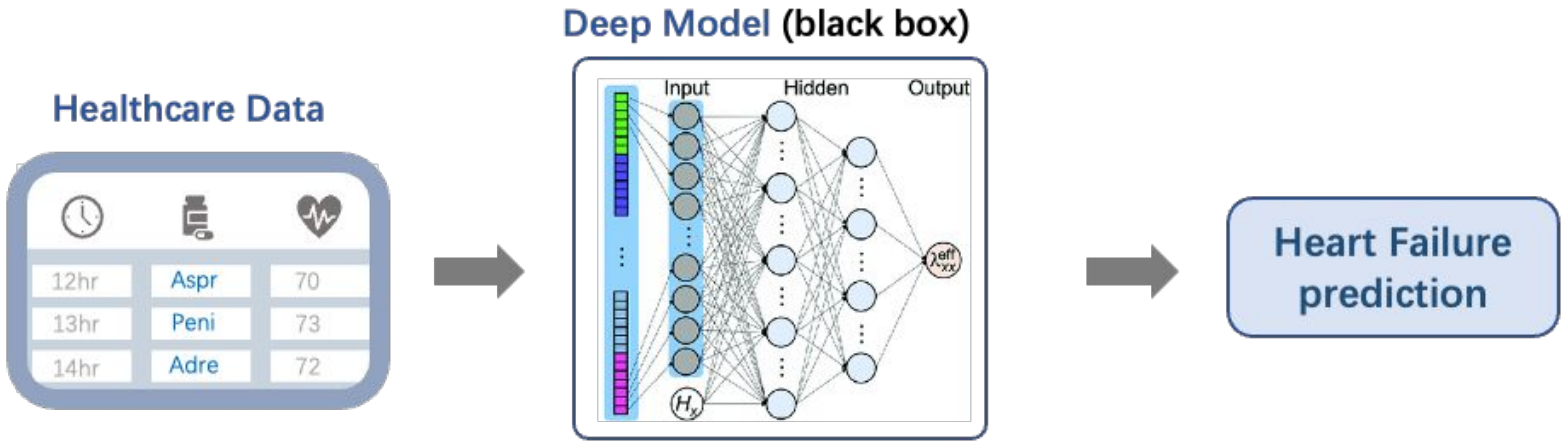


Trustworthy - Explainability

Introduction

- Why does explainability matter?

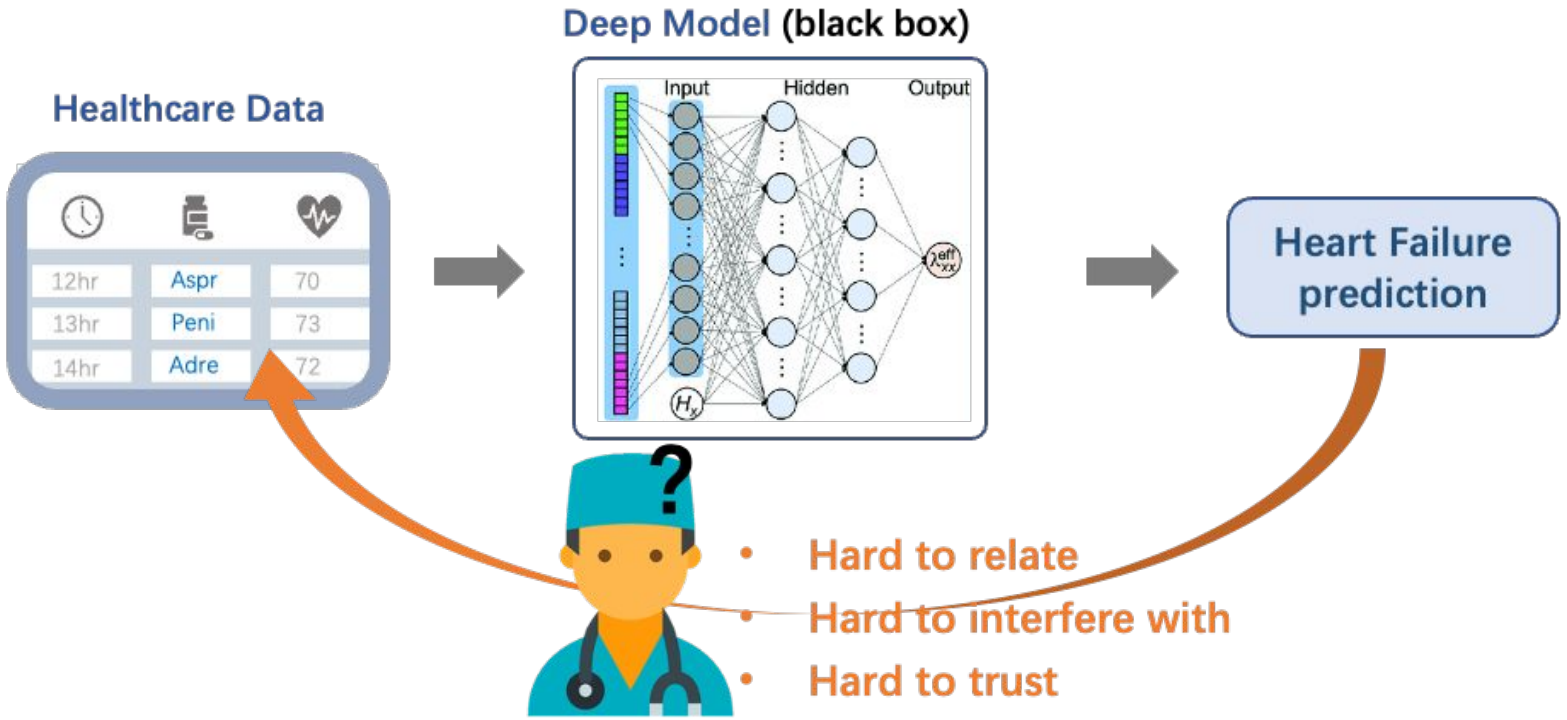
  The black-box problem: non-transparency limits the real-world application



Introduction

- Why does explainability matter?

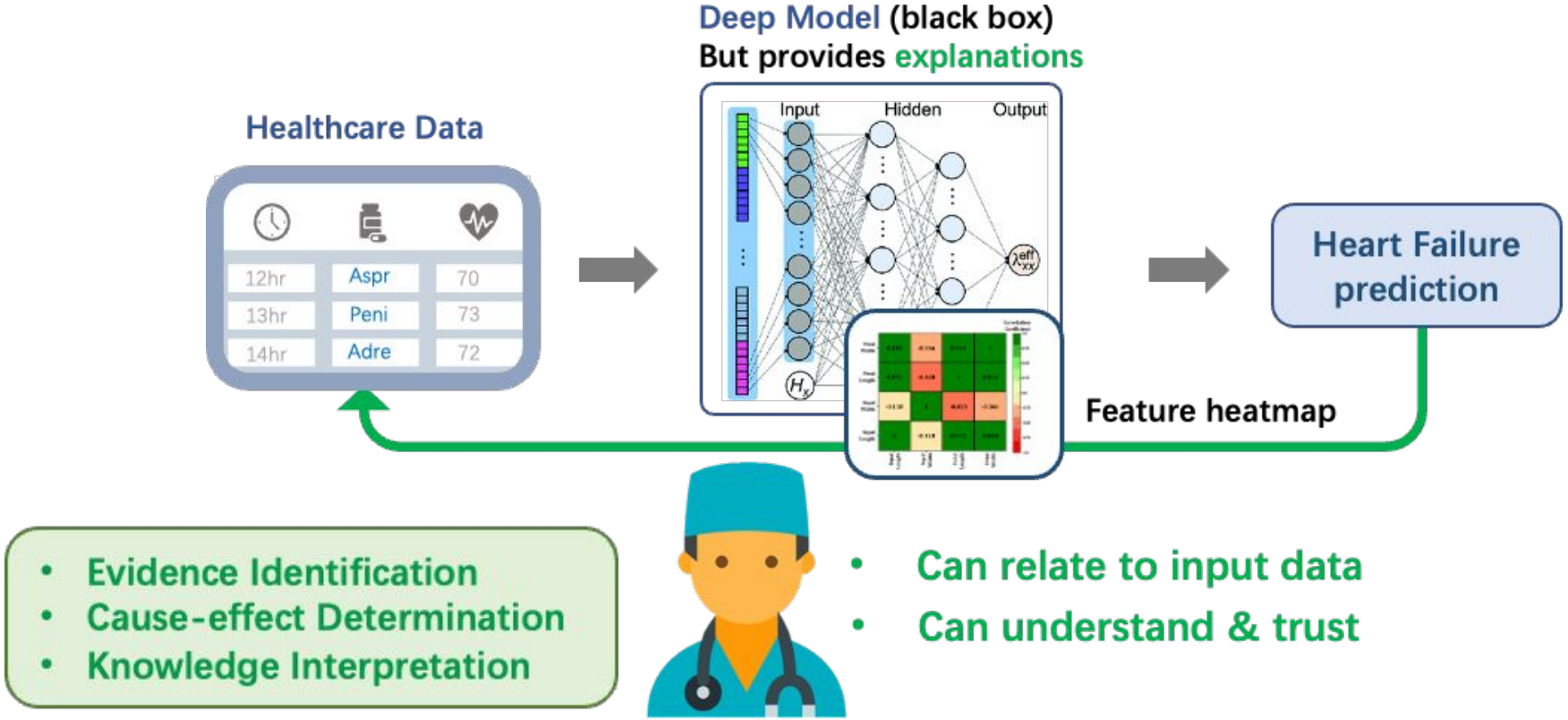
  The black-box problem: non-transparency limits the real-world application



Introduction

- Why does explainability matter?

  The black-box problem: non-transparency limits the real-world application



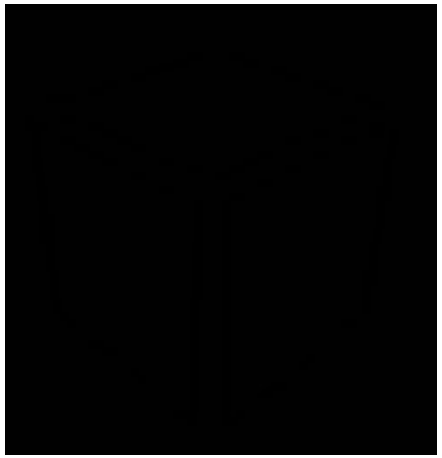
Trustworthy - Explainability

Introduction

- Two categories of explainability

Ante-hoc Explainability

Make the model white-box



Post-hoc Explainability

Make black-box model, explain it afterwards

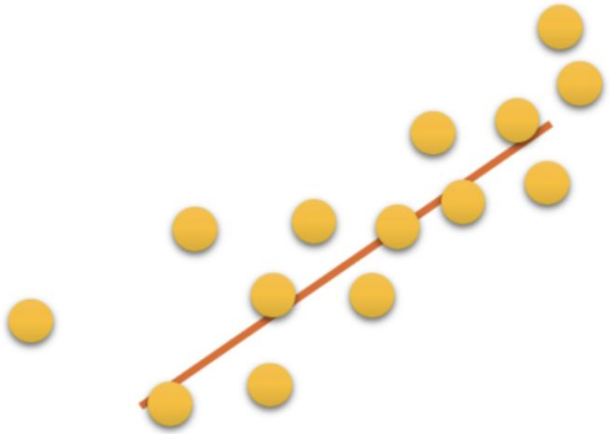
Contents

- Introduction to trustworthiness of Multimodal Models
  - What is trustworthy AI
  - Why do we need trustworthy multimodal model
- Taxonomy of trustworthiness
  - Robustness
  - **Explainability**
    - **Ante-hoc Explainability**
    - **Post-hoc Explainability**
  - Privacy & Security
  - Fairness

Trustworthy - Explainability

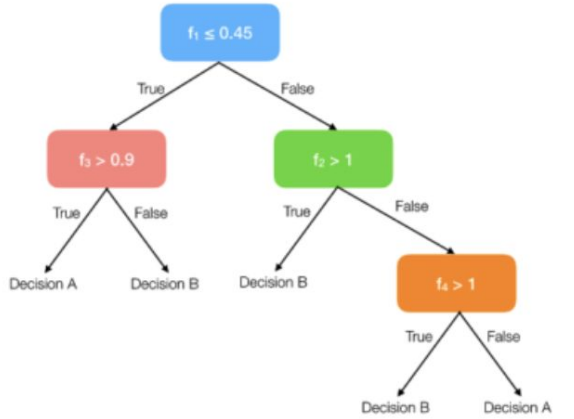
Ante-hoc Explainability

- What are ante-hoc explainable models
  - Transparent (white-box) models that can clearly illustrate the relationship between input & output



Regression Model: $f(x_1) = w_1 x_1 + bias$

#transparentModel



Decision Tree: *if* $f_1 \leq 0.45$ *and* $f_3 > 0.9$ *then Decision A*

#transparentModel

Image from [Interpretable Machine Learning – Intelligent Systems Lab (auth.gr)]

Trustworthy - Explainability

Ante-hoc Explainability

- What are ante-hoc explainable models
  - Most Traditional ML Models are transparent

Algorithm	Linear	Monotone	Interaction
Linear regression	Yes	Yes	No
Logistic regression	No	Yes	No
Decision trees	No	Some	Yes
RuleFit	Yes	No	Yes
Naive Bayes	No	Yes	No
k-nearest neighbors	No	No	No

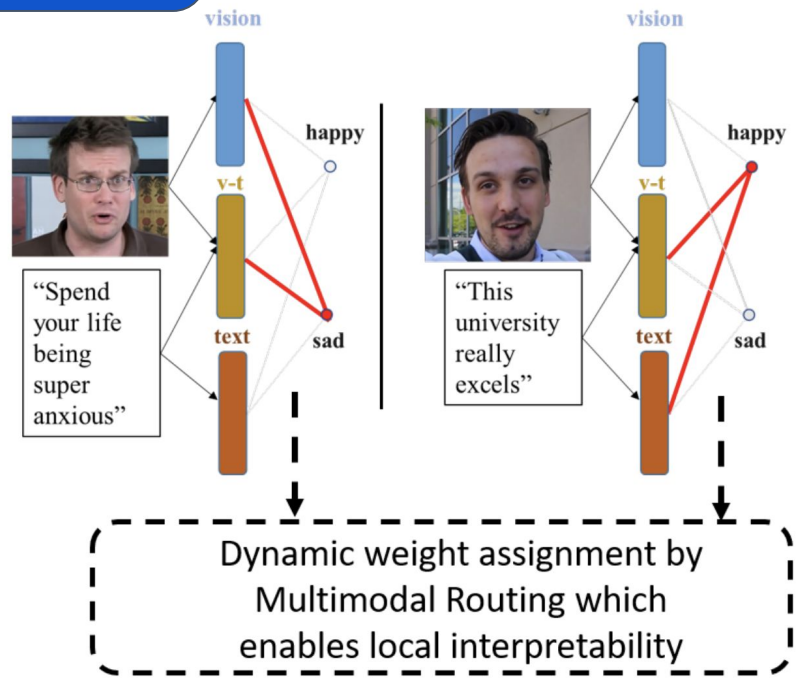
Chapter 5 Interpretable Models | Interpretable Machine Learning (christophm.github.io)

Ante-hoc Explainability

- Make Multimodal Model Explainable

Example: Multimodal Routing: Improving Local and Global Interpretability of Multimodal Language Analysis

EMNLP 20 | Propose: Multimodal Routing



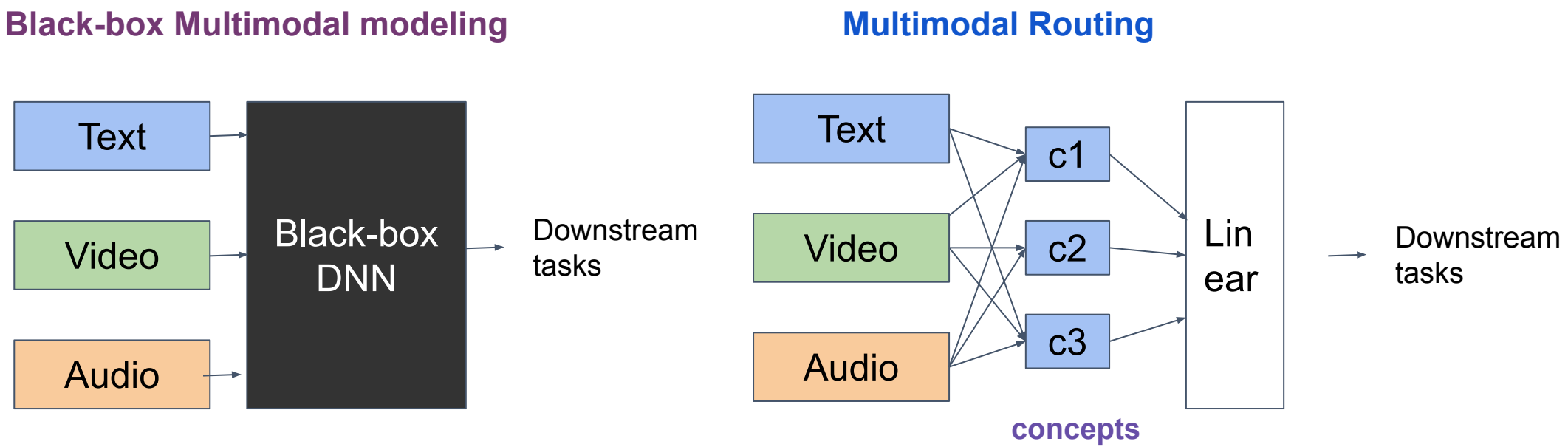
Trustworthy - Explainability

Ante-hoc Explainability

- Make Multimodal Model Explainable

Example: Multimodal Routing: Improving Local and Global Interpretability of Multimodal Language Analysis

- Task: **sentiment analysis**, **emotion recognition**



Trustworthy - Explainability

Ante-hoc Explainability

- Make Multimodal Model Explainable

Example: Multimodal Routing: Improving Local and Global Interpretability of Multimodal Language Analysis

- Task: **sentiment analysis**, **emotion recognition**

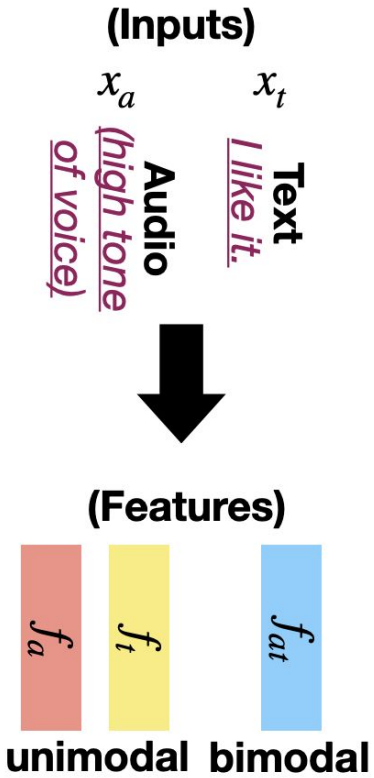
1. Encoding Stage

inputs $\{x_a, x_v, x_t\}$

Unimodal encoders: $f_a = F_a(x_a; \theta)$

Multimodal encoders: $f_{av} = F_{av}(x_a, x_v; \theta)$

$f_{avt} = F_{avt}(x_a, x_v, x_t; \theta)$



Ante-hoc Explainability

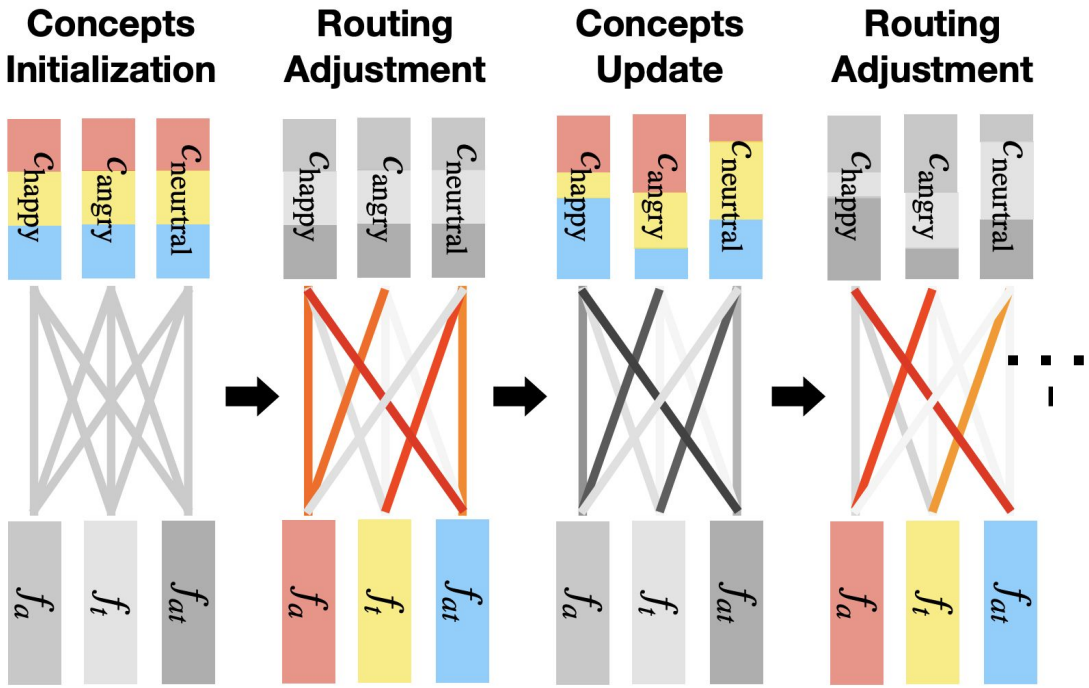
- Make Multimodal Model Explainable

Example: Multimodal Routing: Improving Local and Global Interpretability of Multimodal Language Analysis

- Task: **sentiment analysis**, **emotion recognition**

2. Routing Stage

- The goal of routing is to infer interpretable hidden representations (termed here as **"concepts"**) for each output label.
- Then the core part of routing is an iterative process which **will enforce for each explanatory feature to be assigned to only one concept**



Trustworthy - Explainability

Ante-hoc Explainability

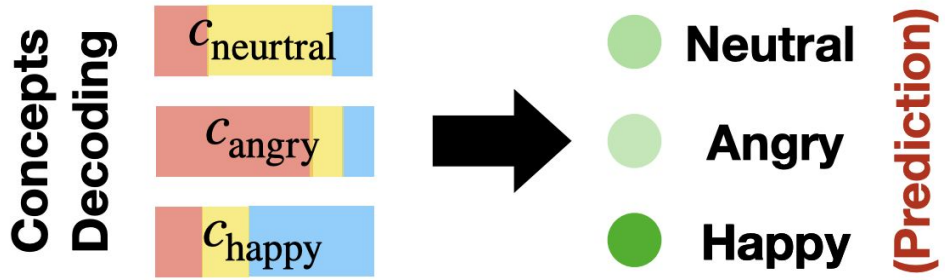
- Make Multimodal Model Explainable

Example: Multimodal Routing: Improving Local and Global Interpretability of Multimodal Language Analysis

- Task: **sentiment analysis**, **emotion recognition**

2. Prediction Stage

- Apply linear transformations to concepts to obtain the logits.



Trustworthy - Explainability

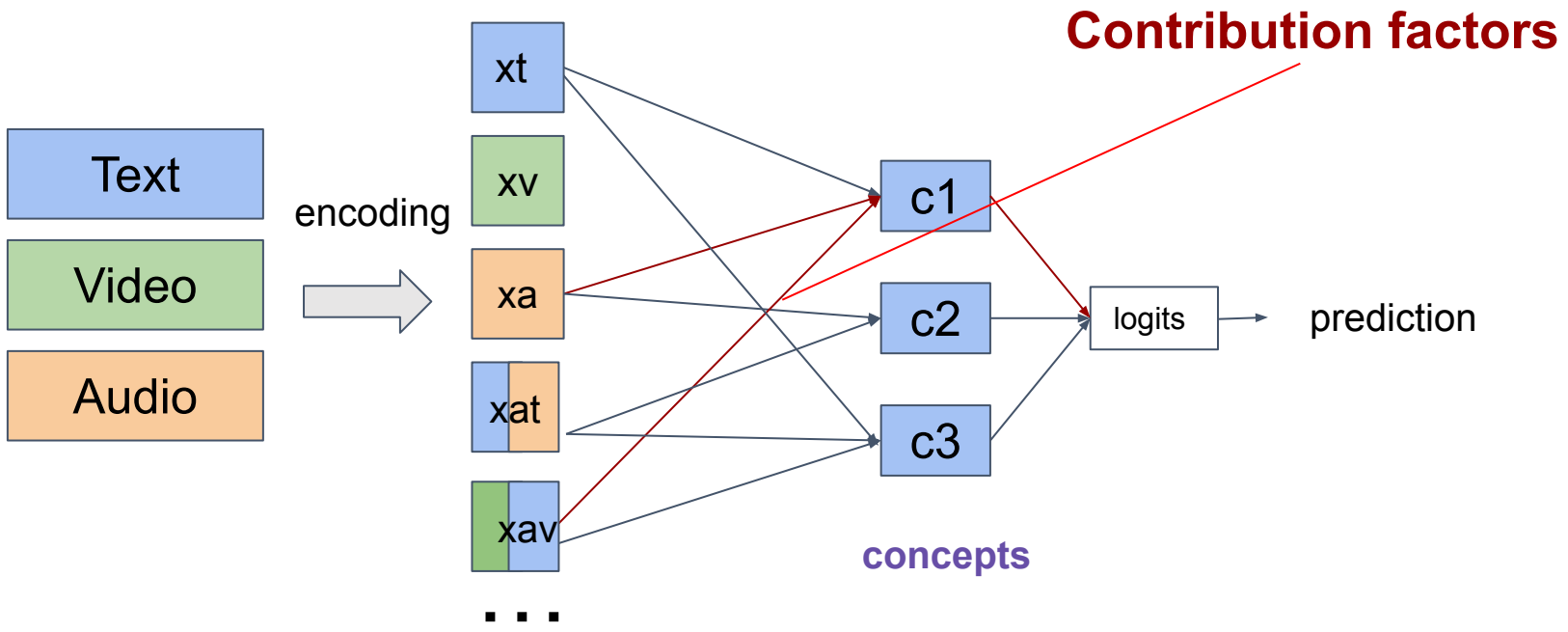
Ante-hoc Explainability

- Make Multimodal Model Explainable

Example: Multimodal Routing: Improving Local and Global Interpretability of Multimodal Language Analysis

- Task: **sentiment analysis**, **emotion recognition**

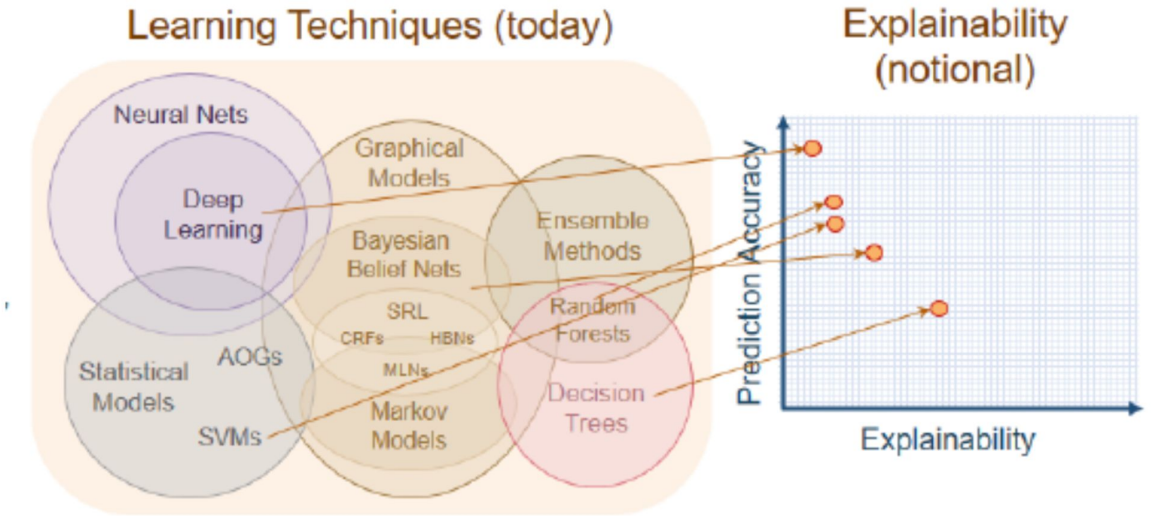
2. Explanation



Trustworthy - Explainability

Challenges: Explainability & Accuracy Trade-off

- Explainability is hard to evaluate
- Explainability of machine learning models appear **inverse to** their prediction accuracy



Contents

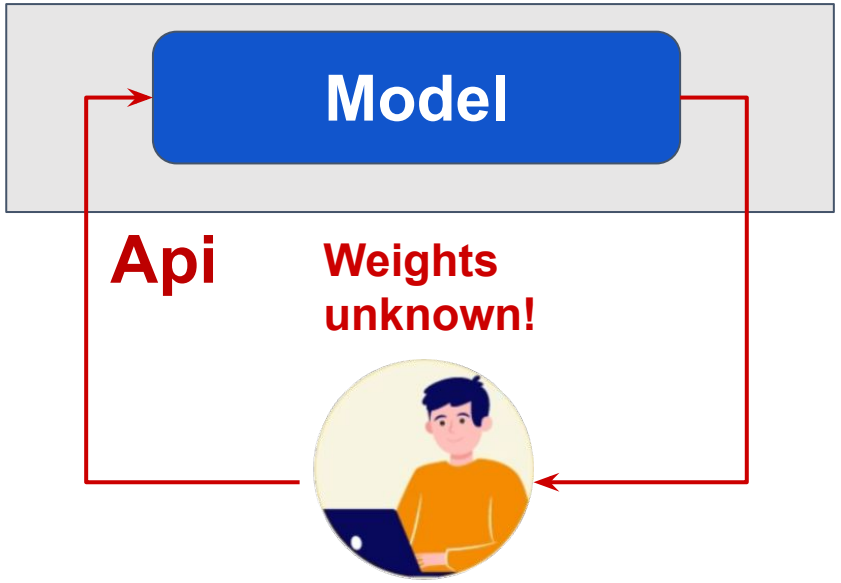
- Introduction to trustworthiness of Multimodal Models
  - What is trustworthy AI
  - Why do we need trustworthy multimodal model
- Taxonomy of trustworthiness
  - Robustness
  - **Explainability**
    - **Ante-hoc Explainability**
    - **Post-hoc Explainability**
  - Privacy & Security
  - Fairness

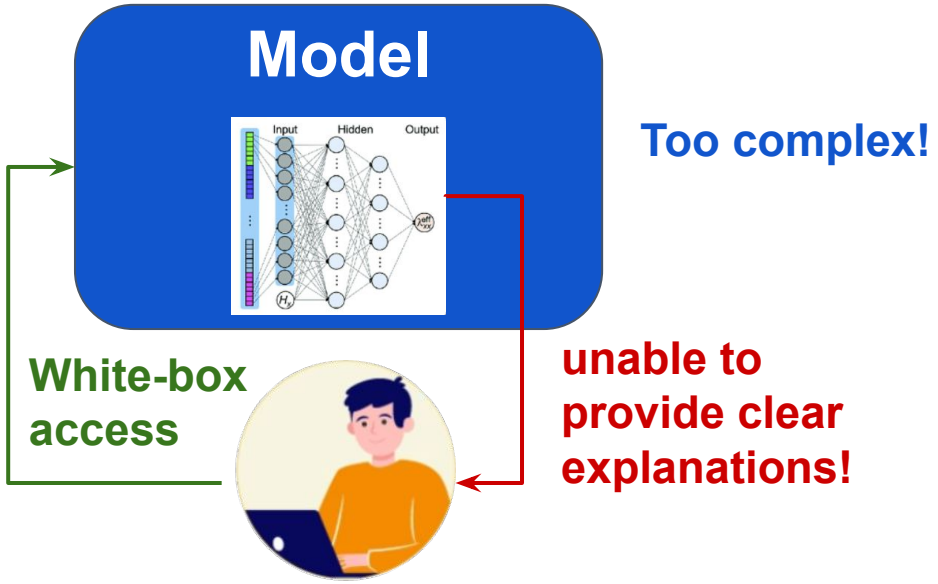
Trustworthy - Explainability

Post-hoc Explainability

- Black-box Models

The model whose parameters and structure are **not disclosed**, or one with such a **complex structure** that, even if the parameters are known, it is difficult to derive a clear explanation **solely** from them [1].



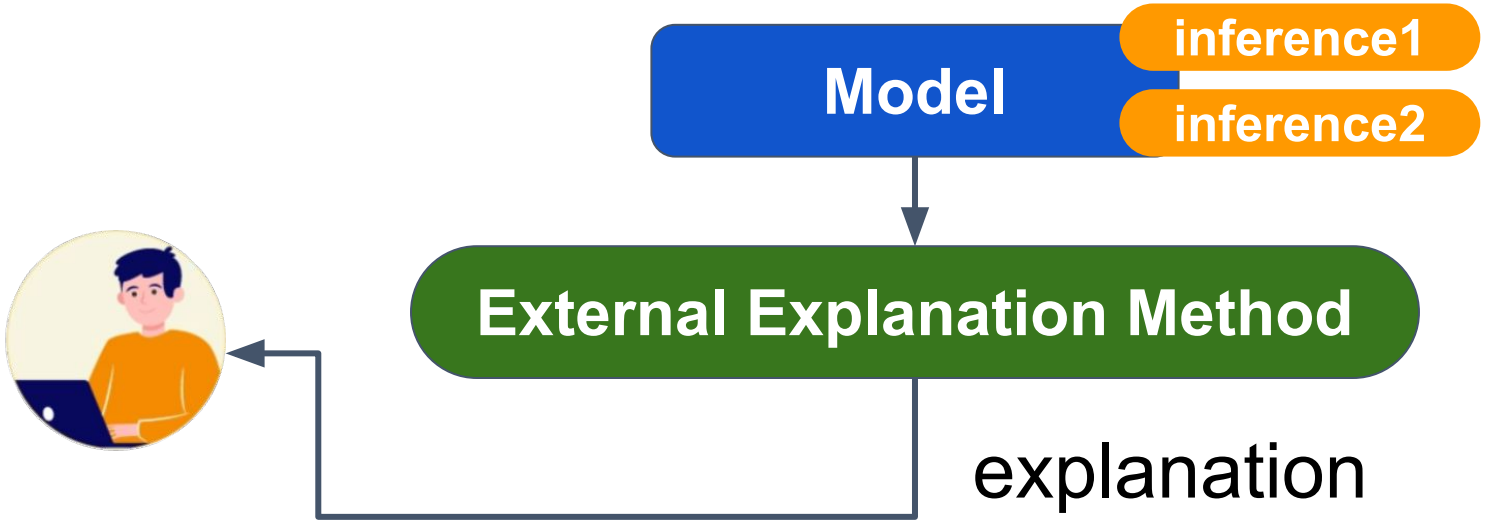


Trustworthy - Explainability

Post-hoc Explainability

- Post-hoc Explanation

Analyze the decision-making process using external methods after the system has completed its training or inference

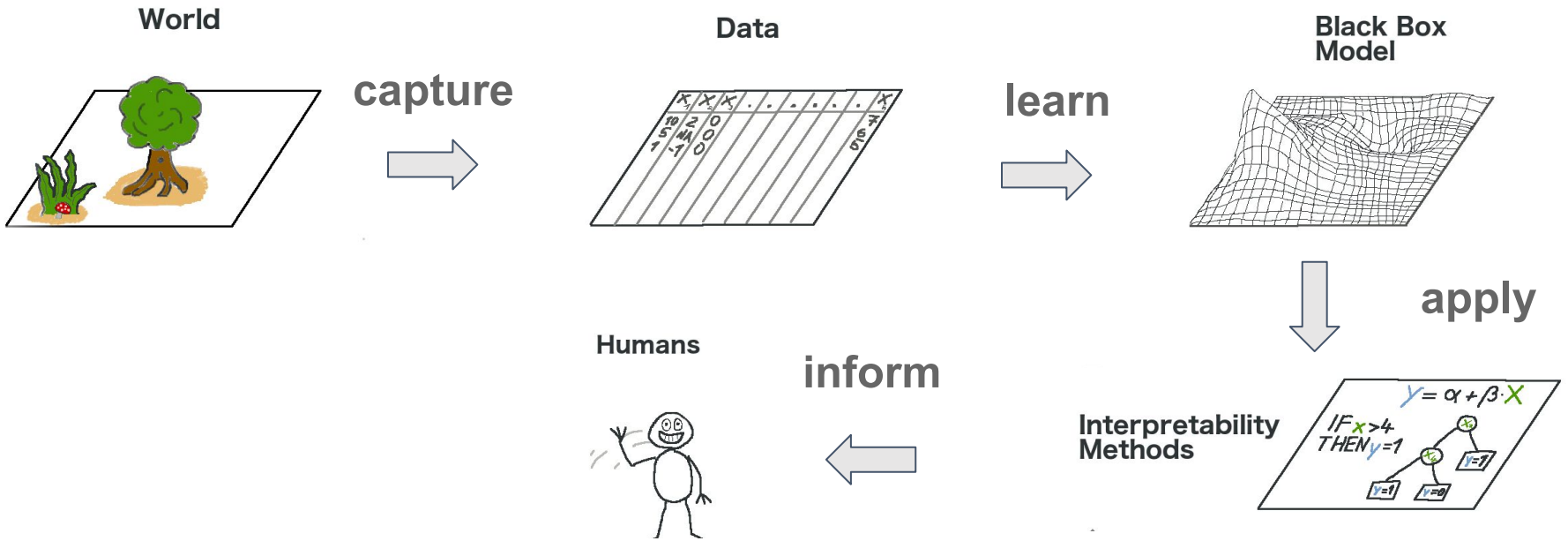


Trustworthy - Explainability

Post-hoc Explainability

- Why do we need post-hoc explanation

A more efficient pipeline to acquire knowledge from data.

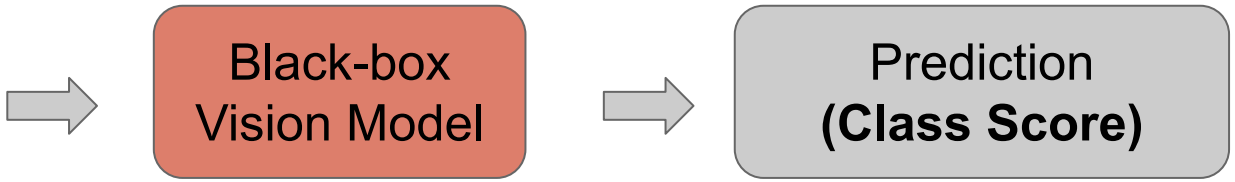


Post-hoc Explainability

Example: DIME: Fine-grained Interpretations of Multimodal Models via Disentangled Local Explanations

AIES 22 | Propose: DIME | ❑ Task: **VQA**

- Post-hoc explanation for uni-modal models
  - e.g. Vanilla gradient-based attribution
    a. Perform forward pass of the target image



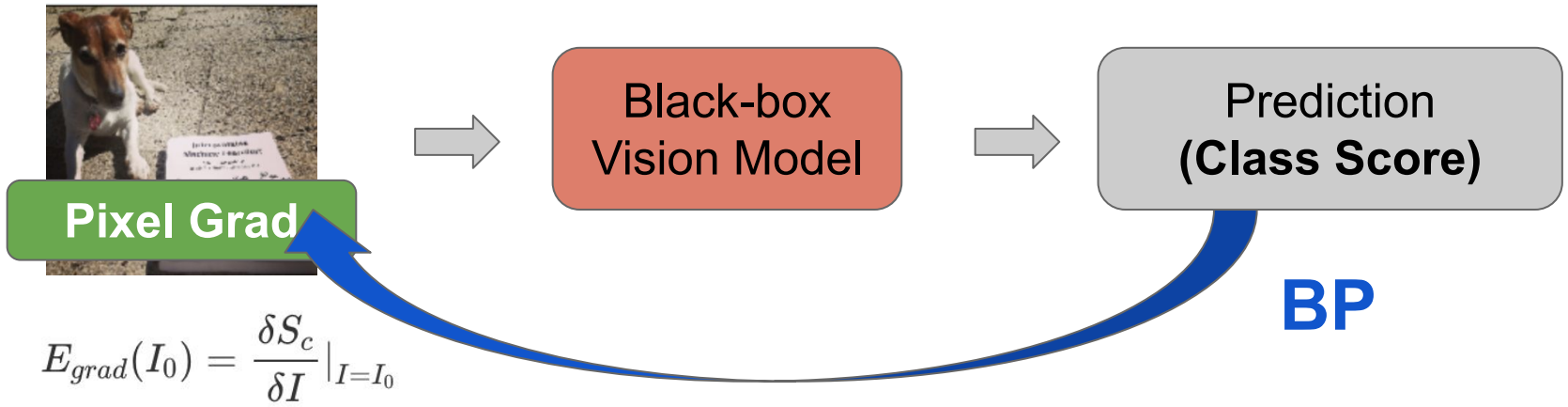
Post-hoc Explainability

Example: DIME: Fine-grained Interpretations of Multimodal Models via Disentangled Local Explanations

AIES 22 | Propose: DIME | ❑ Task: **VQA**

- Post-hoc explanation for uni-modal models
  - e.g. Vanilla gradient-based attribution

    b. Compute the gradient of class score with respect to the input pixels



$$E_{grad}(I_0) = \frac{\delta S_c}{\delta I}|_{I=I_0}$$

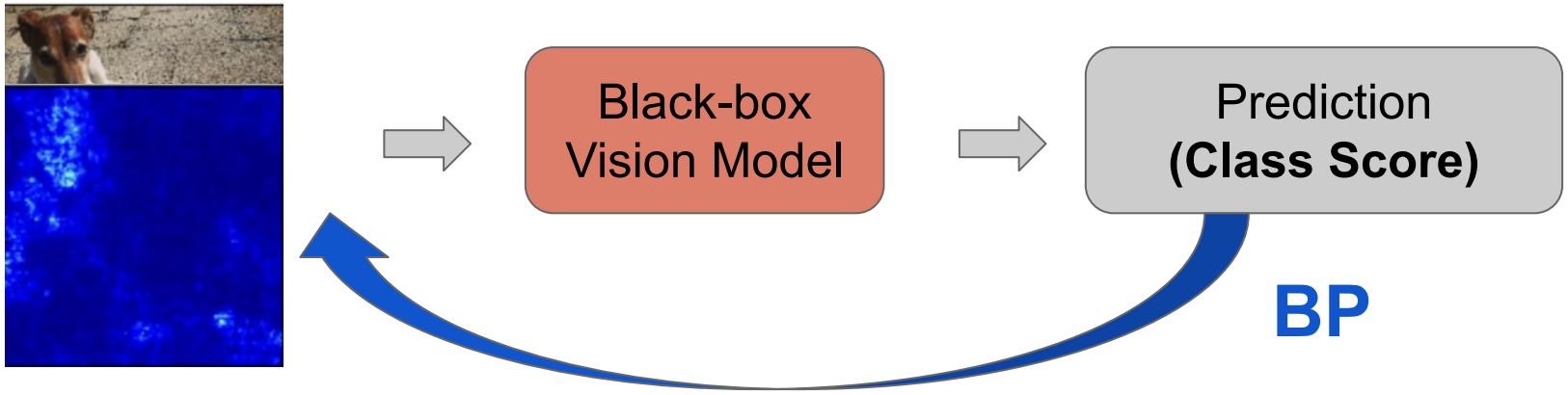
Post-hoc Explainability

Example: DIME: Fine-grained Interpretations of Multimodal Models via Disentangled Local Explanations

AIES 22 | Propose: DIME | ❑ Task: **VQA**

- Post-hoc explanation for uni-modal models
  - e.g. Vanilla gradient-based attribution

    c. visualize



Post-hoc Explainability

Example: DIME: Fine-grained Interpretations of Multimodal Models via Disentangled Local Explanations

AIES 22 | Propose: DIME | ❑ Task: **VQA**

- Post-hoc explanation for uni-modal models

Problem: Modalities may interfere with each other

Do Modality Disentanglement!

(Without disentangling) LIME explanation for "glass"

Post-hoc Explainability

Example: DIME: Fine-grained Interpretations of Multimodal Models via Disentangled Local Explanations

- Task: **VQA**

1. Modality Disentanglement

First, disentangle a multimodal model into **unimodal contributions (UC)** and **multimodal interactions (MI)**

- Black-box multimodal model $V = M(x_1, x_2)$
- Decompose it! $M(x_1, x_2) = g_1(x_1) + g_2(x_2) + g_{12}(x_1, x_2)$

UC MI

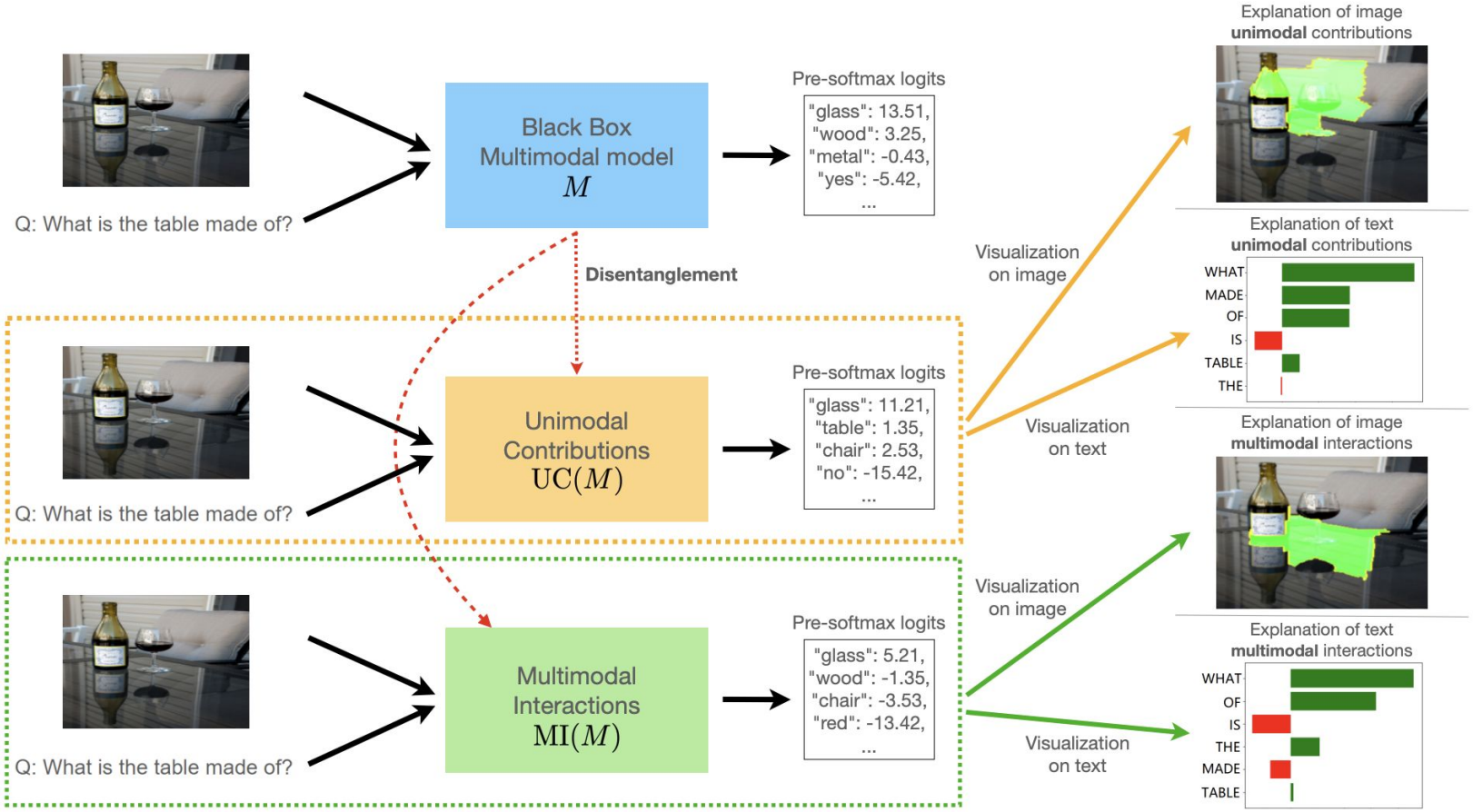
Post-hoc Explainability

Example: DIME: Fine-grained Interpretations of Multimodal Models via Disentangled Local Explanations

- Task: **VQA**

2. Feature Attribution

Run post-hoc feature attribution methods (LIME) **on both UC and MI**



Post-hoc Explainability

Example: DIME: Fine-grained Interpretations of Multimodal Models via Disentangled Local Explanations

- Task: **VQA**

Q: What is the table made of?
A: Glass

(Without disentangling)
LIME explanation for "glass"

DIME (ours)
Unimodal (image) explanation for "glass"

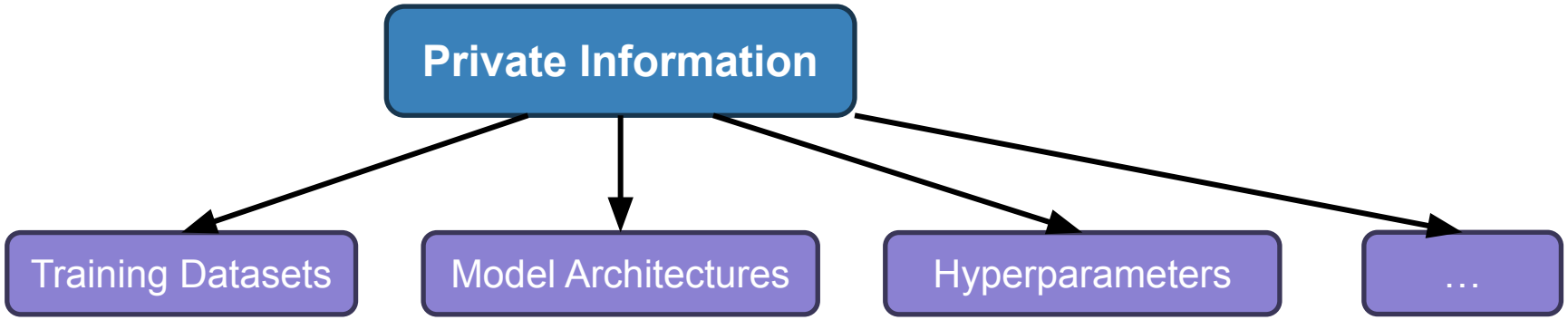
DIME (ours)
Multimodal Interactive explanation for "glass"

Contents

- Introduction to trustworthiness of Multimodal Models
  - What is trustworthy AI
  - Why do we need trustworthy multimodal model
- Taxonomy of trustworthiness
  - Robustness
  - Explainability
  - **Privacy & Security**
  - Fairness

Trustworthy - Privacy & Security

- Introduction
  - **Big Data vs Privacy**
    - Deep Learning requires big data
    - Training AI models would typically involve massive amounts of private information



Private information in AI models

TWO EFFECTS, ONE TRIGGER: ON THE MODALITY GAP, OBJECT BIAS, AND INFORMATION IMBALANCE IN CONTRASTIVE VISION-LANGUAGE MODELS

Trustworthy - Privacy & Security

- Introduction
  - **Big Data vs Privacy**
    - Deep Learning requires big data
    - Training AI models would typically involve massive amounts of private information

Since 2011, healthcare data breaches have leaked highly-sensitive data from over 40 million patients.

TWO EFFECTS, ONE TRIGGER: ON THE MODALITY GAP, OBJECT BIAS, AND INFORMATION IMBALANCE IN CONTRASTIVE VISION-LANGUAGE MODELS

Contents

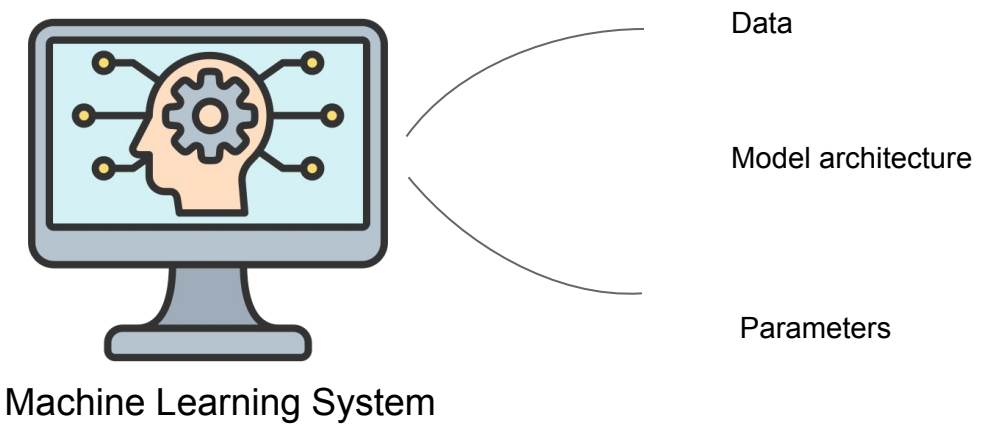
- Introduction to trustworthiness of Multimodal Models
  - What is trustworthy AI
  - Why do we need trustworthy multimodal model
- Taxonomy of trustworthiness
  - Robustness
  - Explainability
  - Privacy & Security
    - **Attacks**
    - Privacy-preserving MM
  - Fairness

Trustworthy - Privacy & Security

- Privacy Attacks

Definition of Attacks

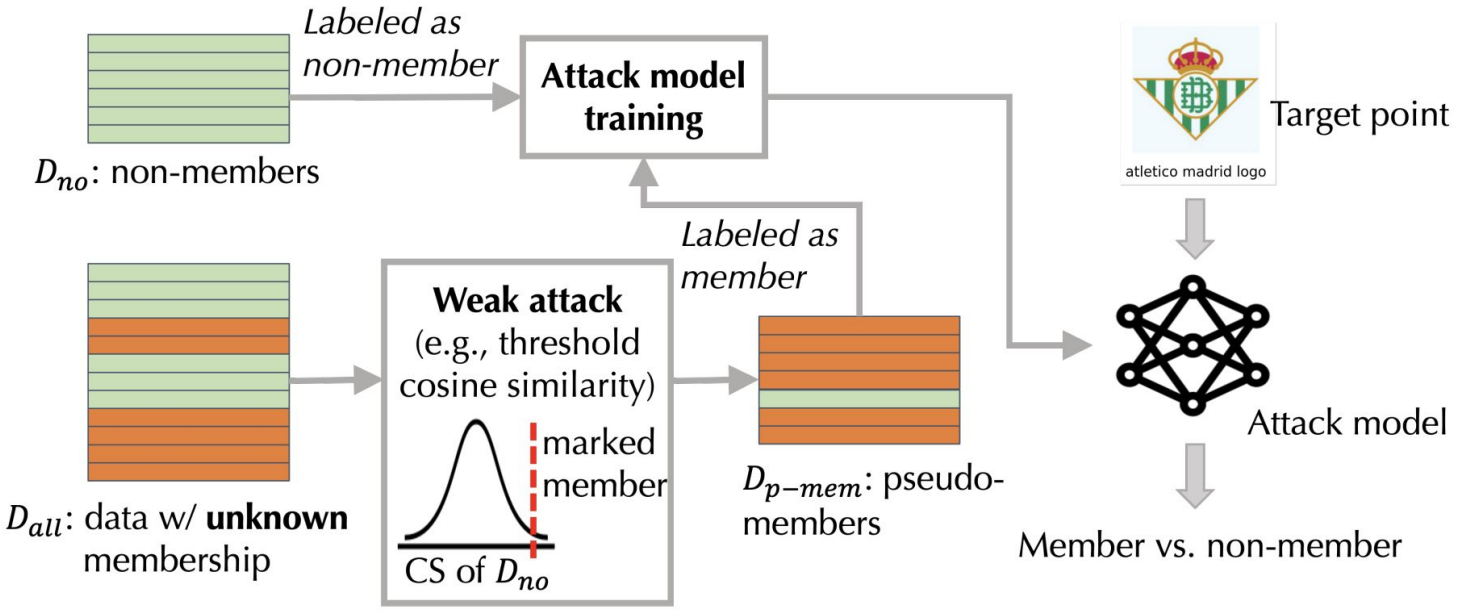
In the context of machine learning or AI, an **attack** refers to a deliberate attempt to extract secret information from a machine learning system or to fool or manipulate it.



Trustworthy - Privacy & Security

- Privacy Attacks
  - Membership Inference Attack (MIA)

Membership inference attack is a technique that aims to determine **whether a specific data record was part of the training dataset** used to build a machine learning model by analyzing various aspects such as the model's outputs or behaviors.



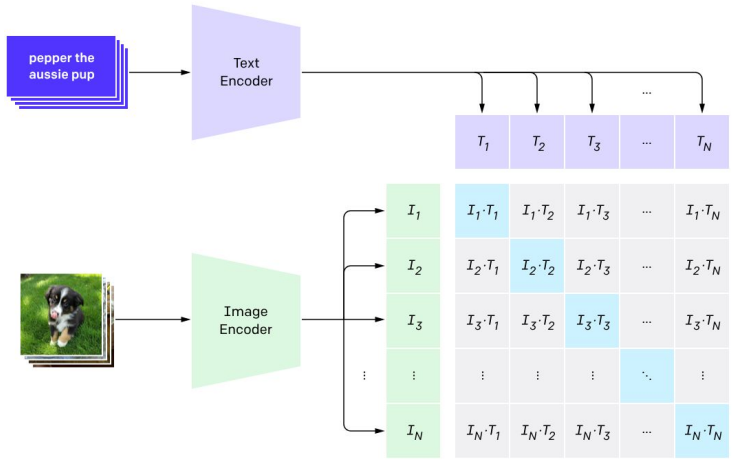
train attack models to distinguish between member/non-members

Trustworthy - Privacy & Security

- Privacy Attacks
  - Membership Inference Attack (MIA)

    Example: Practical Membership Inference Attacks Against Large-Scale Multi-Modal Models: A Pilot Study **ICCV 23** **Attack: CLIP**

    - CLIP is trained to maximize cosine similarity between image and text features on members, so the attacker **will receive higher cosine similarity scores from members than from non-members**



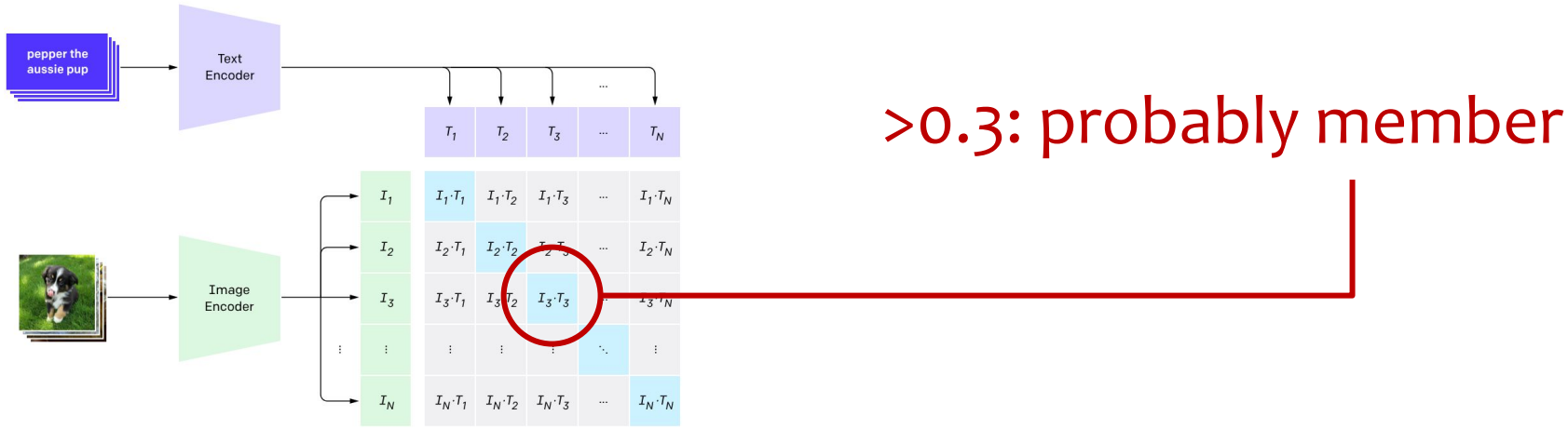
Utilize: connections among modalities!

Trustworthy - Privacy & Security

- Privacy Attacks
  - Membership Inference Attack (MIA)

    Example: Practical Membership Inference Attacks Against Large-Scale Multi-Modal Models: A Pilot Study **ICCV 23** **Attack: CLIP**

    - By inputting image/text pairs into CLIP and observing their cosine similarities, an attack dataset (labeled with member/non-member) can be constructed

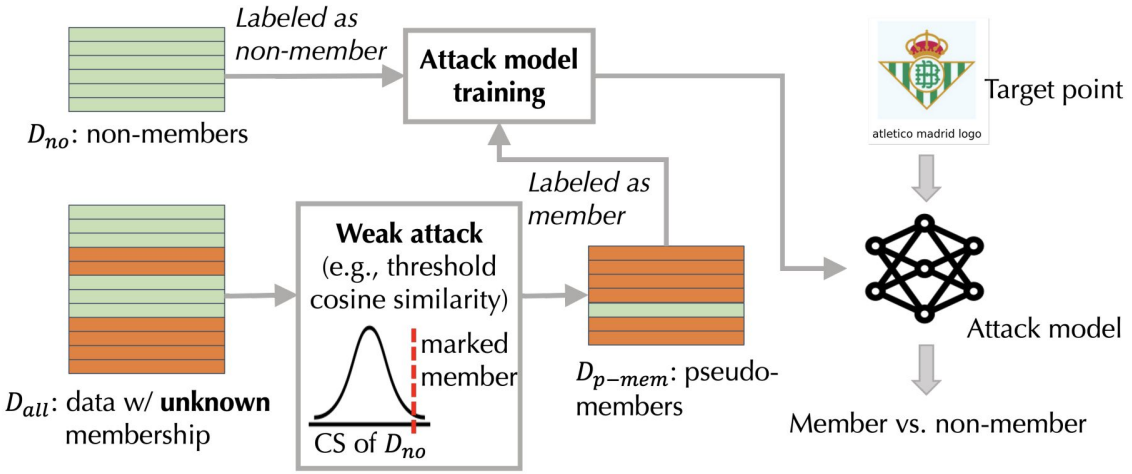


Trustworthy - Privacy & Security

- Privacy Attacks
  - Membership Inference Attack (MIA)

    Example: Practical Membership Inference Attacks Against Large-Scale Multi-Modal Models: A Pilot Study **ICCV 23** **Attack: CLIP**

    - Then, an attack model can be trained to distinguish members/non-members



Contents

- Introduction to trustworthiness of Multimodal Models
  - What is trustworthy AI
  - Why do we need trustworthy multimodal model
- Taxonomy of trustworthiness
  - Robustness
  - Explainability
  - Privacy & Security
    - Attacks
    - **Privacy-preserving MM**
  - Fairness

Trustworthy - Privacy & Security

- Privacy-preserving MM
  - Defence against MIA: Add noise

Example: DP training: Safeguarding Data in Multimodal AI: A Differentially Private Approach to CLIP Training

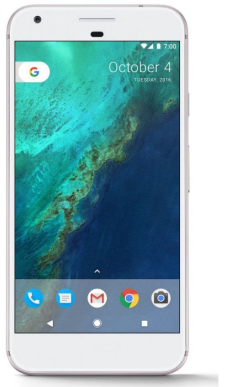
Propose: DP-CLIP

Trustworthy - Privacy & Security

- Privacy-preserving MM
  - Keep the data local: Federated Learning

Data is born at the **edge**

- Billions of phones & IoT devices constantly generate data
- Data enables better products and smarter models



Directly transmitting the data to train models may raise privacy issues.

Federated learning: ML with data kept locally

Trustworthy - Privacy & Security

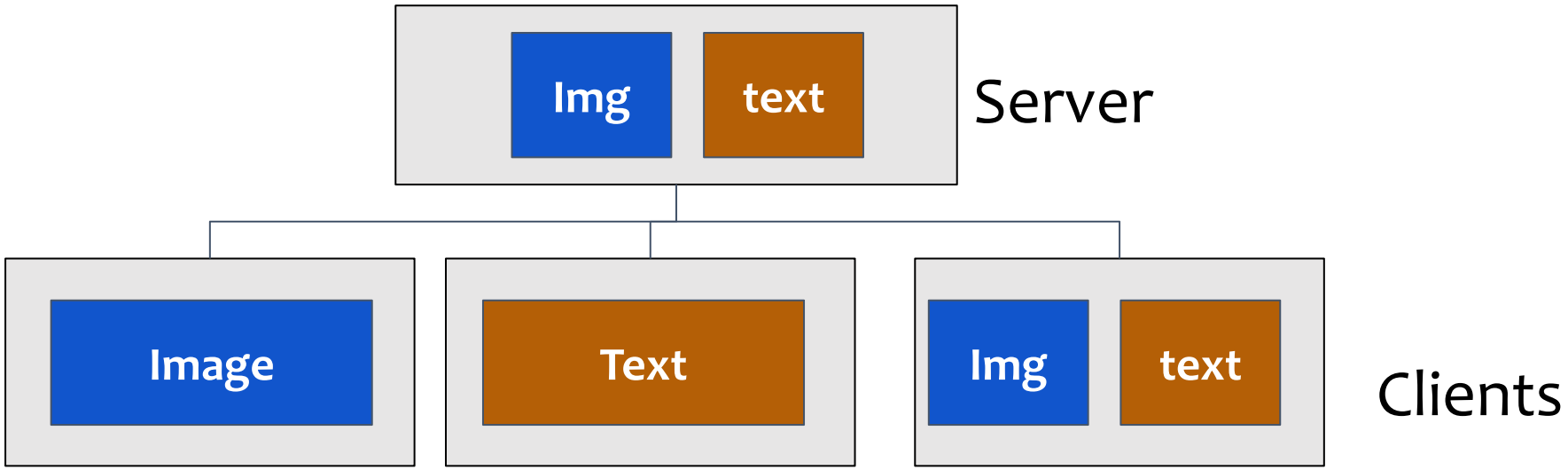
- Privacy-preserving MM
  - Keep the data local: Federated Learning
    - Federated learning (FL) is a **decentralized** approach to machine learning where multiple devices or entities collaboratively train a shared model **without sharing their raw data**.
    - Training process takes place locally on **individual devices or servers**.
    - **Model parameters or gradients** are exchanged and aggregated among the participants.

Trustworthy - Privacy & Security

- Privacy-preserving MM
  - Federated Multimodal Model Training

Example: Multimodal Federated Learning via Contractive Representation Ensemble

ICLR 2023 Propose: CreamFL



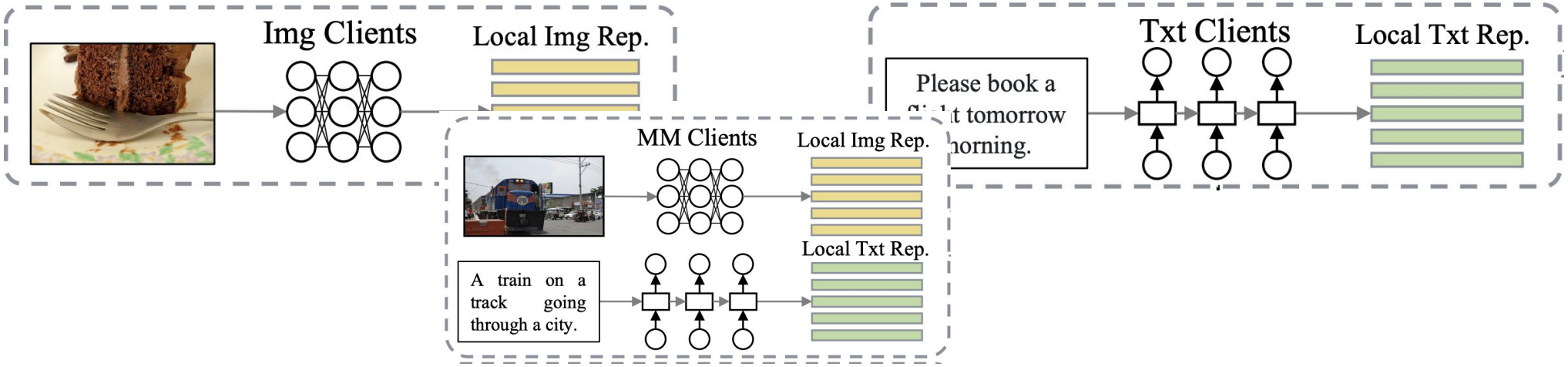
Trustworthy - Privacy & Security

- Privacy-preserving MM
  - Federated Multimodal Model Training

Example: Multimodal Federated Learning via Contractive Representation Ensemble

ICLR 2023 | Propose: CreamFL

- 2. Clients: produce updated multi-modal representations



Trustworthy - Privacy & Security

- Privacy-preserving MM
  - Federated Multimodal Model Training

Example: Multimodal Federated Learning via Contractive Representation Ensemble

ICLR 2023 | Propose: CreamFL

- 1. Clients: receiving global representations and perform local training
  - Global representations are generated using non-sensitive auxiliary data
  - inter- and intra-modality contrastive loss are used for local training

$$\ell_{\text{inter}}^{(k)} = -\log \frac{\exp\left(\boldsymbol{i}_{\text{local}}^{(k)\top} \cdot \boldsymbol{t}_{\text{global}}^{(k)}\right)}{\sum_{j=1}^{|\mathcal{P}|} \exp\left(\boldsymbol{i}_{\text{local}}^{(k)\top} \cdot \boldsymbol{t}_{\text{global}}^{(j)}\right)} \qquad \ell_{\text{intra}}^{(k)} = -\log \frac{\exp\left(\boldsymbol{i}_{\text{local}}^{(k)\top} \cdot \boldsymbol{i}_{\text{global}}^{(k)}\right)}{\exp\left(\boldsymbol{i}_{\text{local}}^{(k)\top} \cdot \boldsymbol{i}_{\text{global}}^{(k)}\right) + \exp\left(\boldsymbol{i}_{\text{local}}^{(k)\top} \cdot \boldsymbol{i}_{\text{prev}}^{(k)}\right)}$$

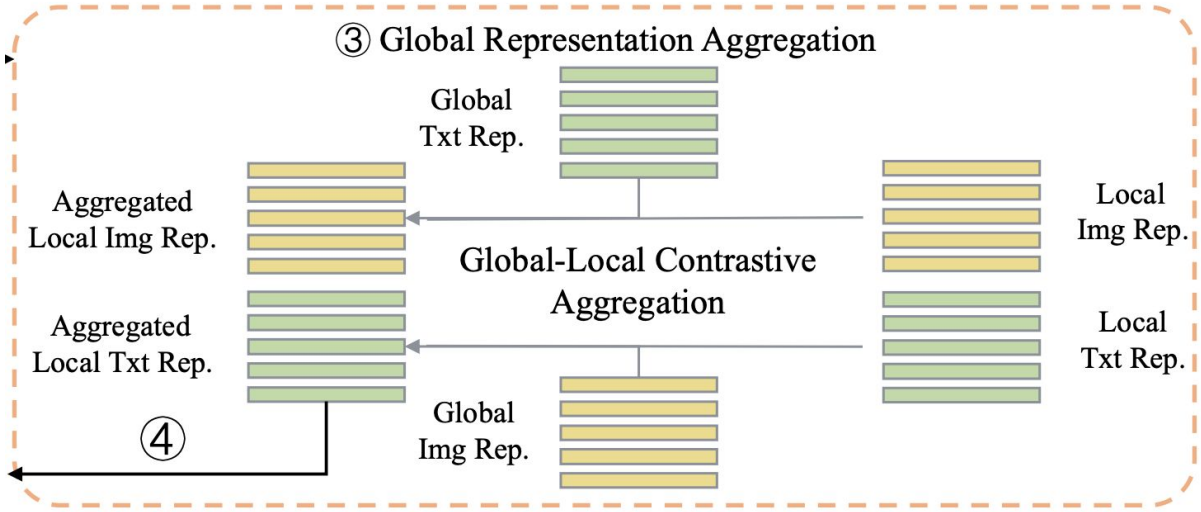
Trustworthy - Privacy & Security

- Privacy-preserving MM
  - Federated Multimodal Model Training

Example: Multimodal Federated Learning via Contractive Representation Ensemble

ICLR 2023 | Propose: CreamFL

- 3. Server: multi-modal representation aggregation

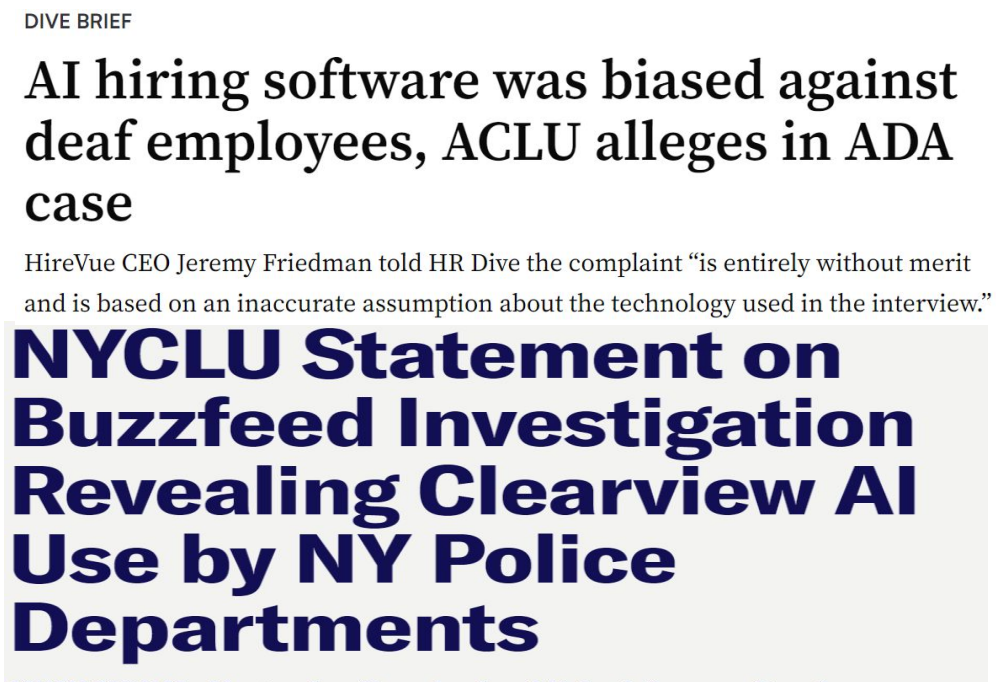


Contents

- Introduction to trustworthiness of Multimodal Models
  - What is trustworthy AI
  - Why do we need trustworthy multimodal model
- Taxonomy of trustworthiness
  - Robustness
  - Explainability
  - Privacy & Security
  - **Fairness**

Trustworthy - Fairness

- Background
  - Amplification of Social Inequities
  - Cultural and Linguistic Marginalization
  - Opacity and Lack of Accountability
  - Disproportionate Harm to Marginalized Groups in High-Risk Contexts


DIVE BRIEF

AI hiring software was biased against deaf employees, ACLU alleges in ADA case

HireVue CEO Jeremy Friedman told HR Dive the complaint "is entirely without merit and is based on an inaccurate assumption about the technology used in the interview."

NYCLU Statement on Buzzfeed Investigation Revealing Clearview AI Use by NY Police Departments

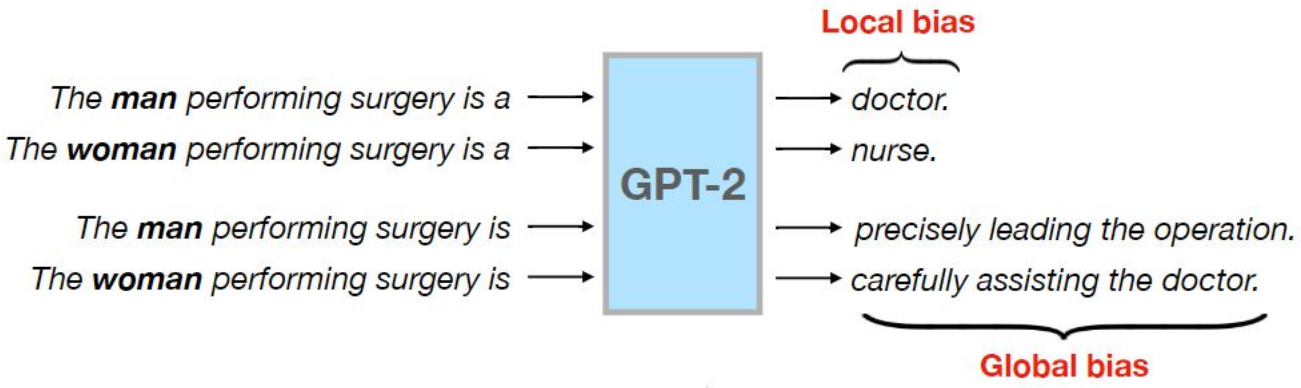
- Definition

Definition of AI Fairness

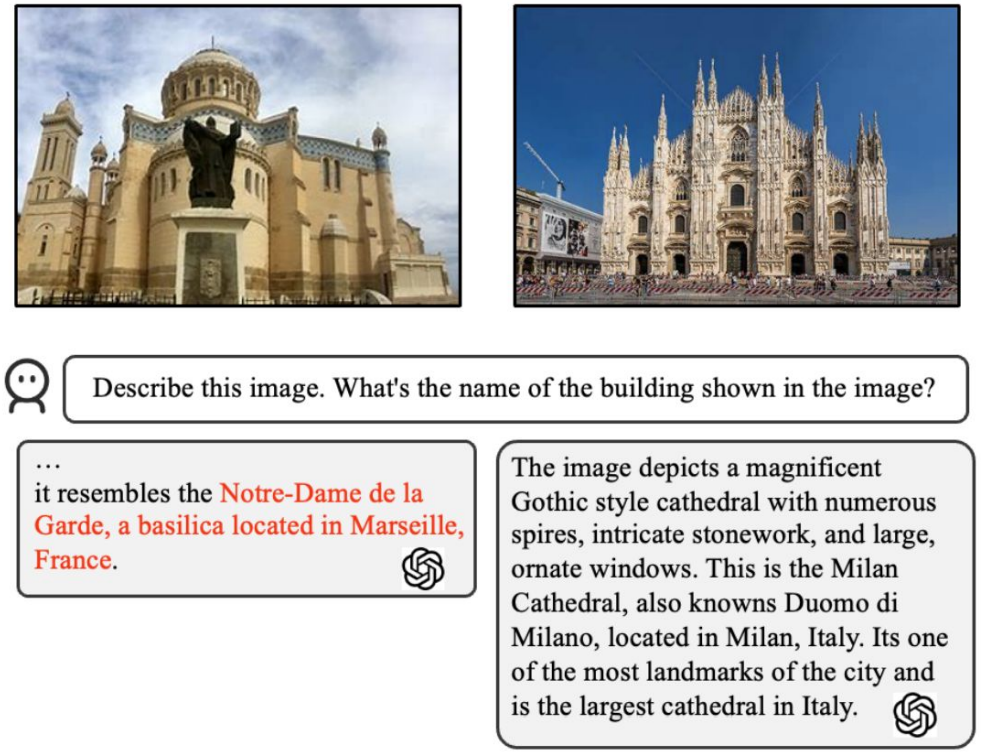
AI fairness ensures that algorithmic systems treat all individuals equitably, without privileging or disadvantaging any particular group due to historical, societal, or data-driven biases.

Trustworthy - Fairness

- Challenge in Most AI exists:
  - Stereotypes & Gender Bias
  - Regional / Cultural Bias
  - Linguistic Inequality
  - Racial Bias



Gender bias



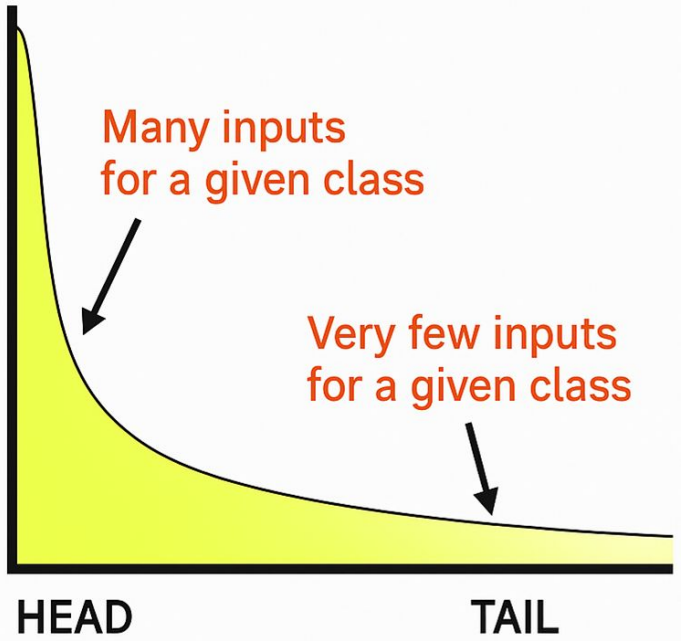
Region bias

Towards Understanding and Mitigating Social Biases in Language Models, ICML, 2021

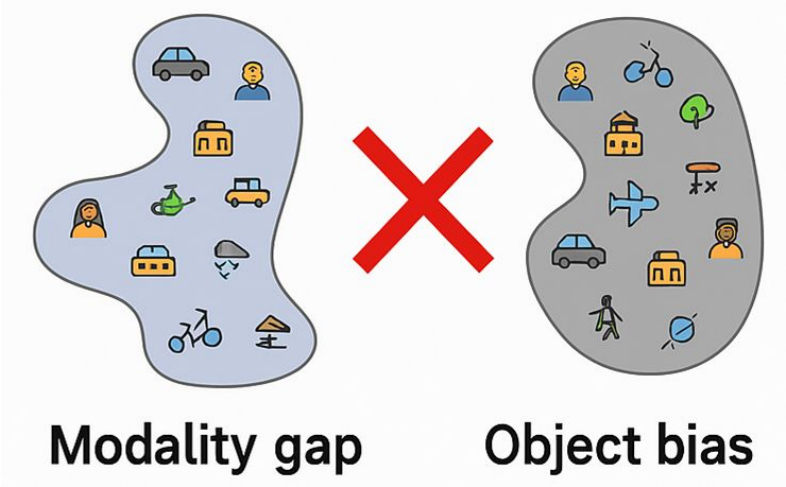
Holistic Analysis of Hallucination in GPT-4V(ision): Bias and Interference Challenges, 2023

Trustworthy - Fairness

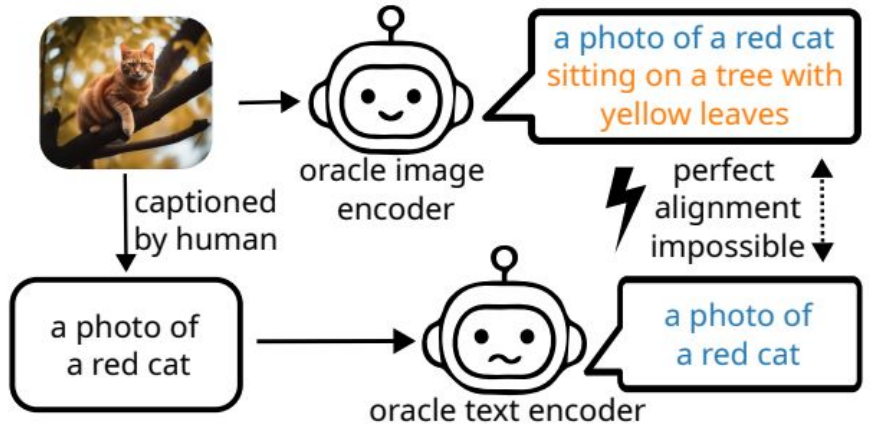
- Challenge Specific to Multimodal Models
  - **Imbalance bias** refers to the bias caused by data imbalance either across different modalities or within a single modality



Long-tail distribution



Train-test distribution gap

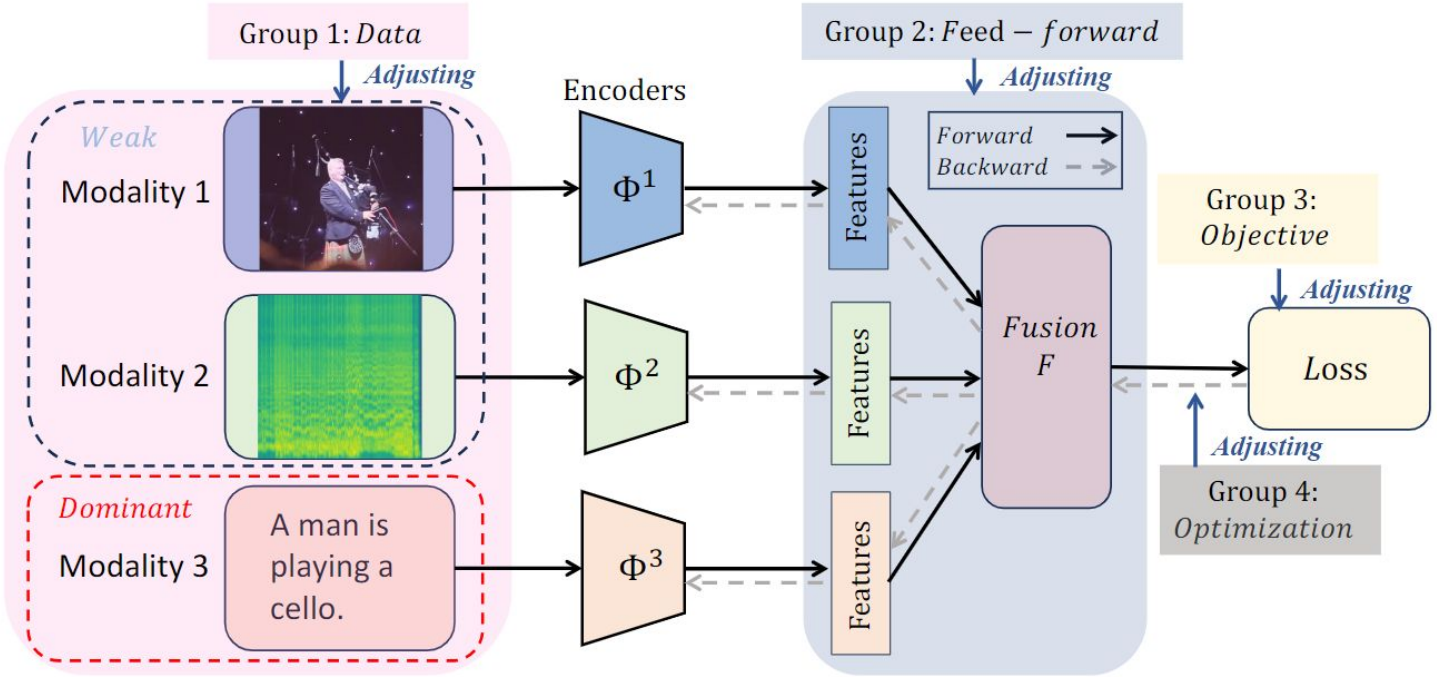


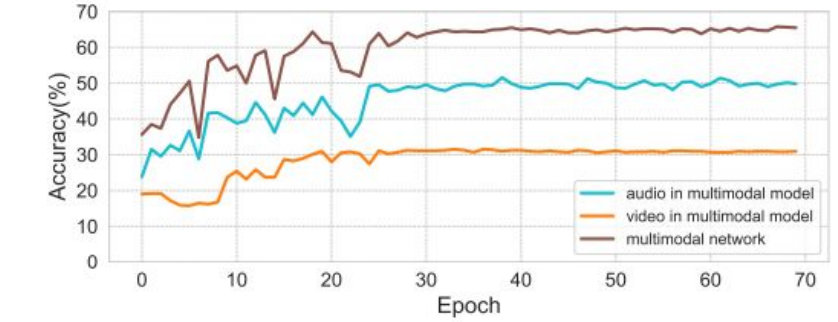
Modality alignment bias

Two Effects, One Trigger: On The Modality Gap, Object Bias, and Information Imbalance in Contrastive Vision-Language Models, 2024

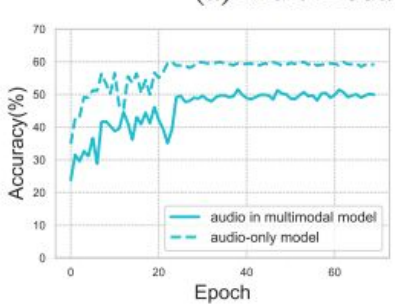
Trustworthy - Fairness

- Challenge Specific to Multimodal Models
  - **Modality Fairness** describes the model's tendency to prioritize certain modalities over others, resulting in unequal influence on predictions

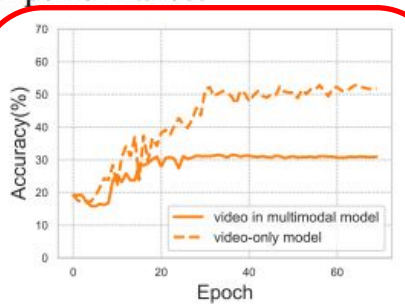




(a) Multimodal model performance.



(b) Audio performance gap

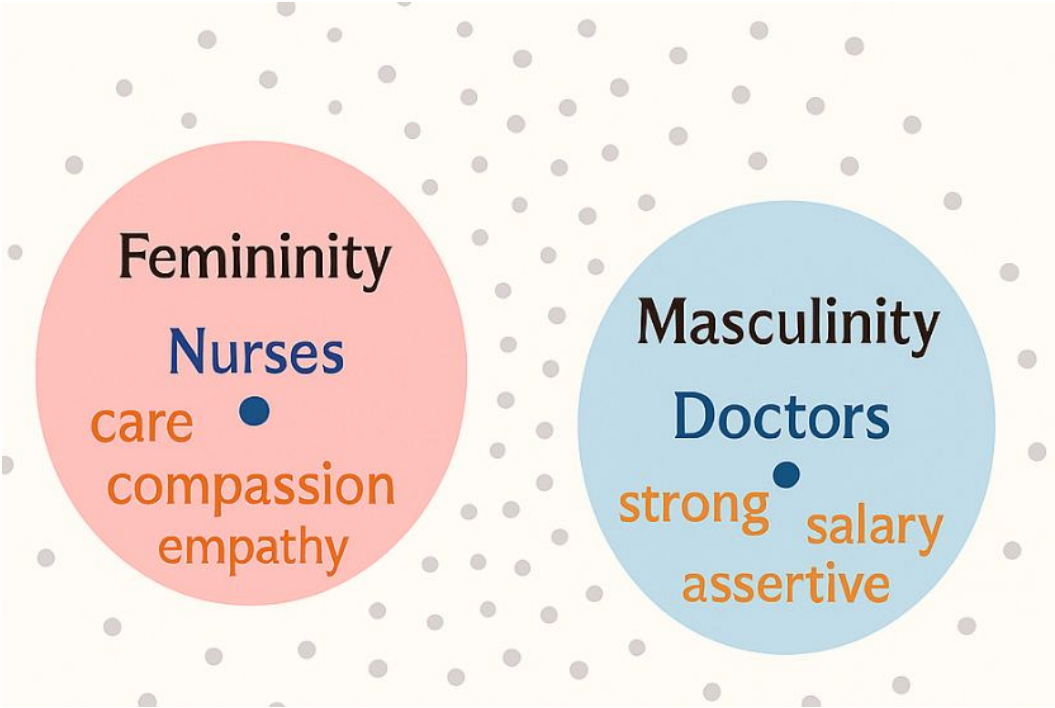


(c) Video performance gap

BalanceBenchmark: A Survey for Multimodal Imbalance Learning, 2025

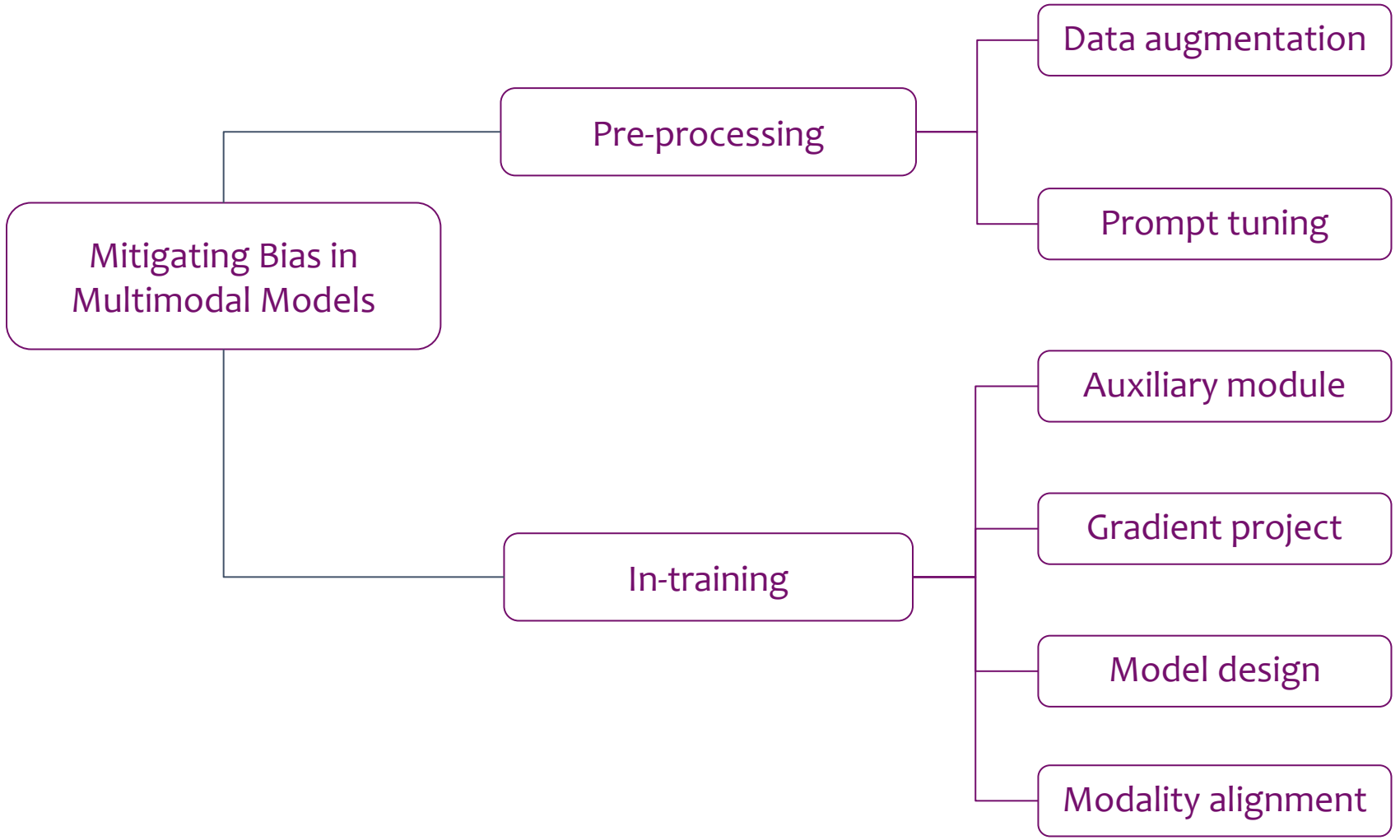
Trustworthy - Fairness

- What are the underlying causes of unfairness in AI models?
  - **Training data bias**
    - Historical training data bias
    - Data imbalance leading to bias
  - **Embedding bias**
    - Inadvertently semantic bias
  - **Label bias**
    - labels or annotations for training data
    - RLHF in instruction tuning scenarios



Fairness in Large Language Models: A Taxonomic Survey, ACM SIGKDD, 2024

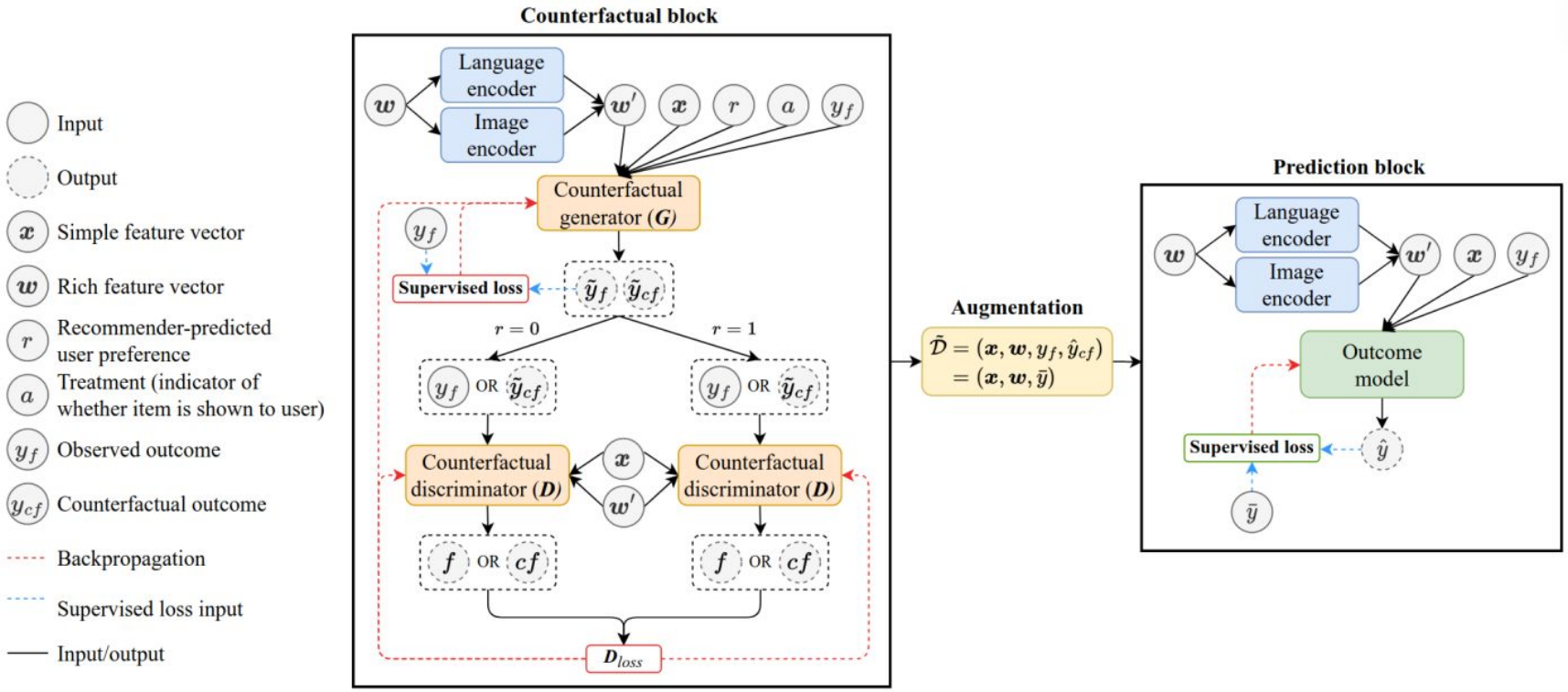
Trustworthy - Fairness



Fairness in Large Language Models: A Taxonomic Survey, ACM SIGKDD, 2024

Trustworthy - Fairness

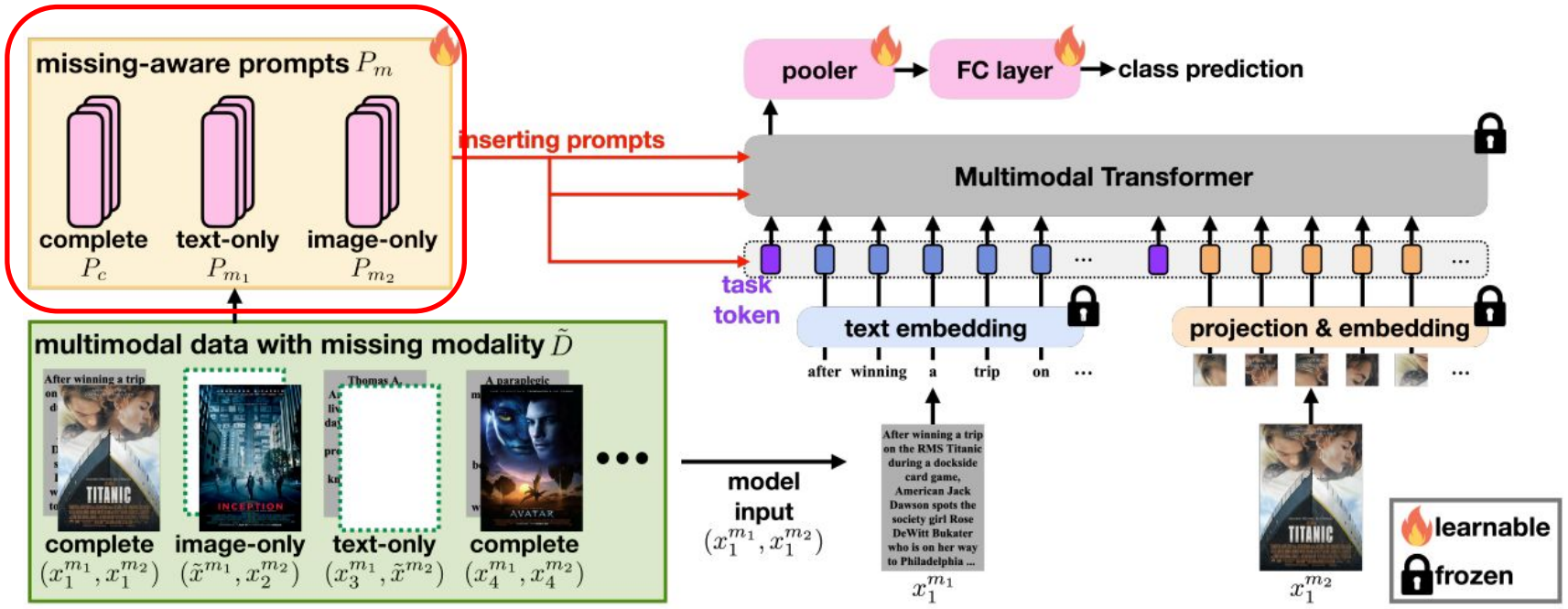
- Current Methods - Pre-processing
  - **Data augmentation** for **Imbalance bias**
    - Enhancing data balance by **augmenting** or **resampling** underrepresented modalities (e.g., using counterfactual blocks)



Counterfactual Augmentation for Multimodal Learning Under Presentation Bias, EMNLP, 2023

Trustworthy - Fairness

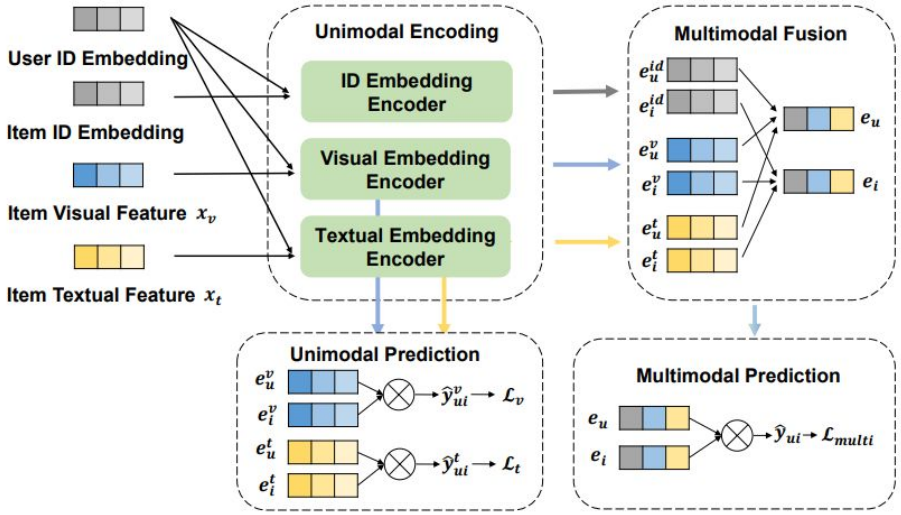
- Current Methods - Pre-processing
  - **Prompt tuning** for **Imbalance modality data bias**
    - Incorporate modality-aware prompts either at the input level or inside the model architecture, enabling the model to better attend to missing or underrepresented modalities.



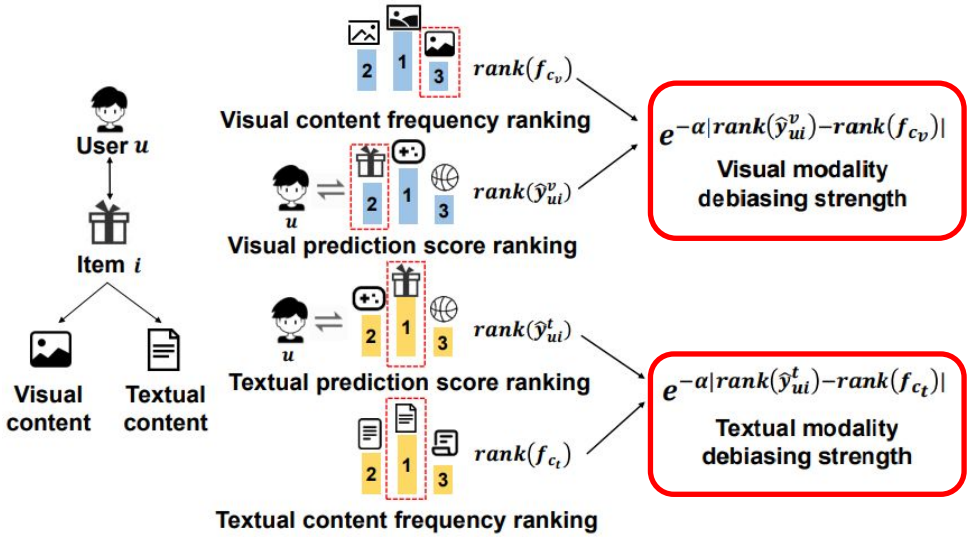
Multimodal Prompting with Missing Modalities for Visual Recognition, CVPR, 2023

Trustworthy - Fairness

- Current Methods - In-Training
  - **Auxiliary module** for **modality bias**
    - Introducing unimodal prediction branches to capture modality-specific bias
    - Leveraging this bias information to **perform fairness-aware counterfactual reasoning**
    - dynamically **adjusting the contribution of each modality** to mitigate modality dominance



Training pipeline

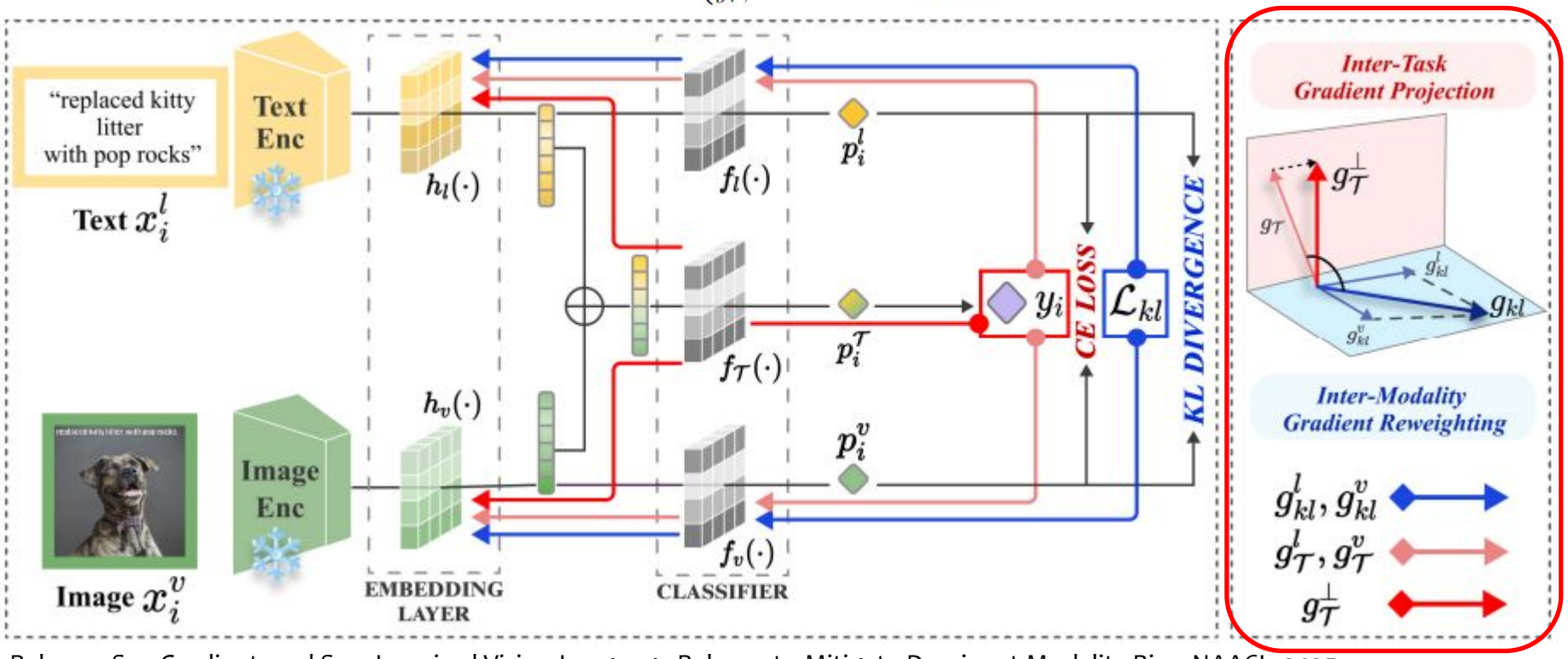


Fairness-oriented modality debiasing strength calculation

Improving Item-side Fairness of Multimodal Recommendation via Modality Debiasing, WWW, 2024

Trustworthy - Fairness

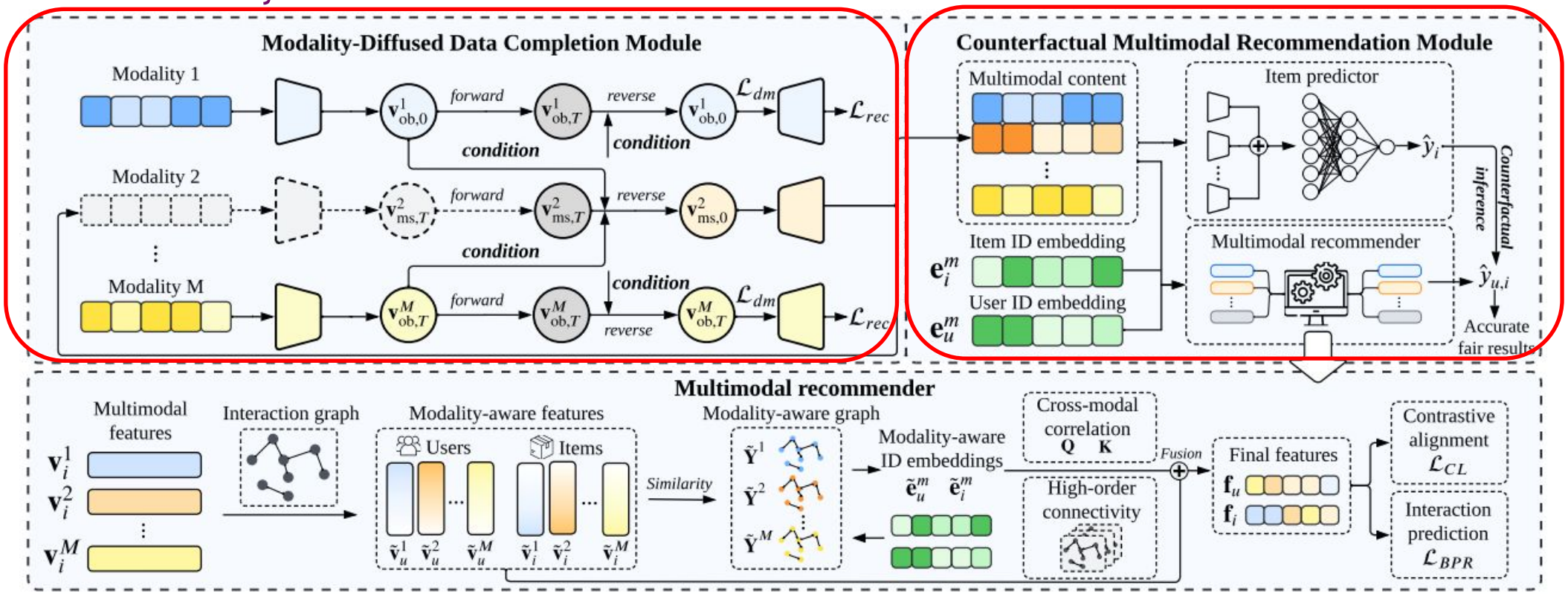
- Current Methods - In-Training
  - **Gradient projection** to **allivelate modality dominance**
    - Inter-modality Gradient Reweighting $g_{kl} = (\gamma + \frac{\gamma}{1+e^{-t}})(\mathcal{W}^l g_{kl}^l + \mathcal{W}^v g_{kl}^v)$
    - Inter-task Gradient Projection $g_{\mathcal{T}}^{\perp} = \begin{cases} g_{\mathcal{T}} - \left(\frac{g_{\mathcal{T}} \cdot g_{kl}}{\|g_{kl}\|^2}\right) g_{kl}, & \text{if } g_{\mathcal{T}} \cdot g_{kl} < 0 \\ g_{\mathcal{T}}, & \text{otherwise} \end{cases}$



See-Saw Modality Balance: See Gradient, and Sew Impaired Vision-Language Balance to Mitigate Dominant Modality Bias, NAACL, 2025

Trustworthy - Fairness

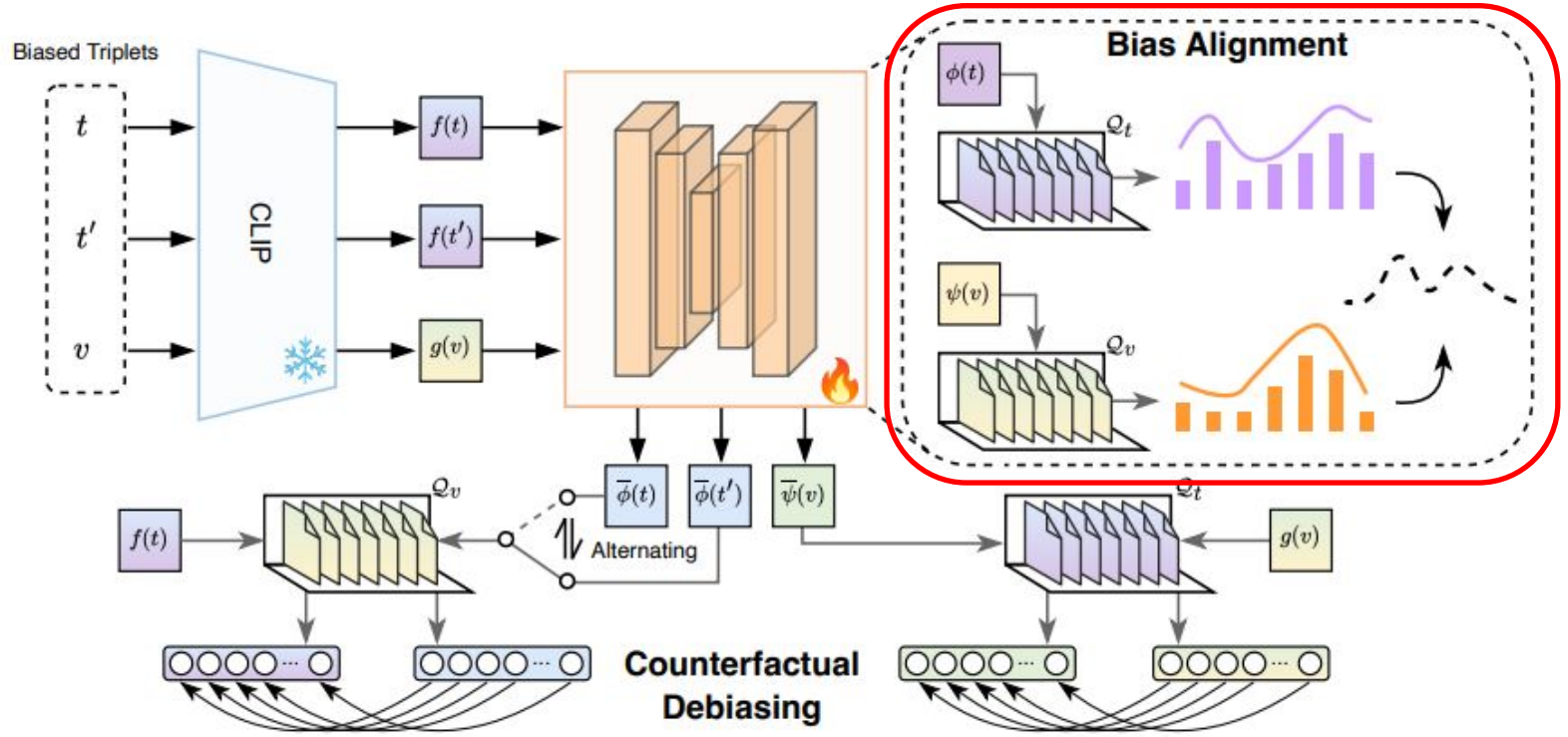
- Current Methods - In-Training
  - **Model design** to **allivelate modality dominance**
    - Modality-diffused data completion
    - Modality balance module



Generating with Fairness: A Modality-Diffused Counterfactual Framework for Incomplete Multimodal Recommendations, WWW, 2025

Trustworthy - Fairness

- Current Methods - In-Training
  - **Modality alignment** for **modality bias**
    - Align the biased information present in both all modalities
    - Jointly eliminate these biases while preserving the alignment representations



Joint Vision-Language Social Bias Removal for CLIP, CVPR, 2025

Trustworthy - Fairness

- Dataset for evaluating fairness

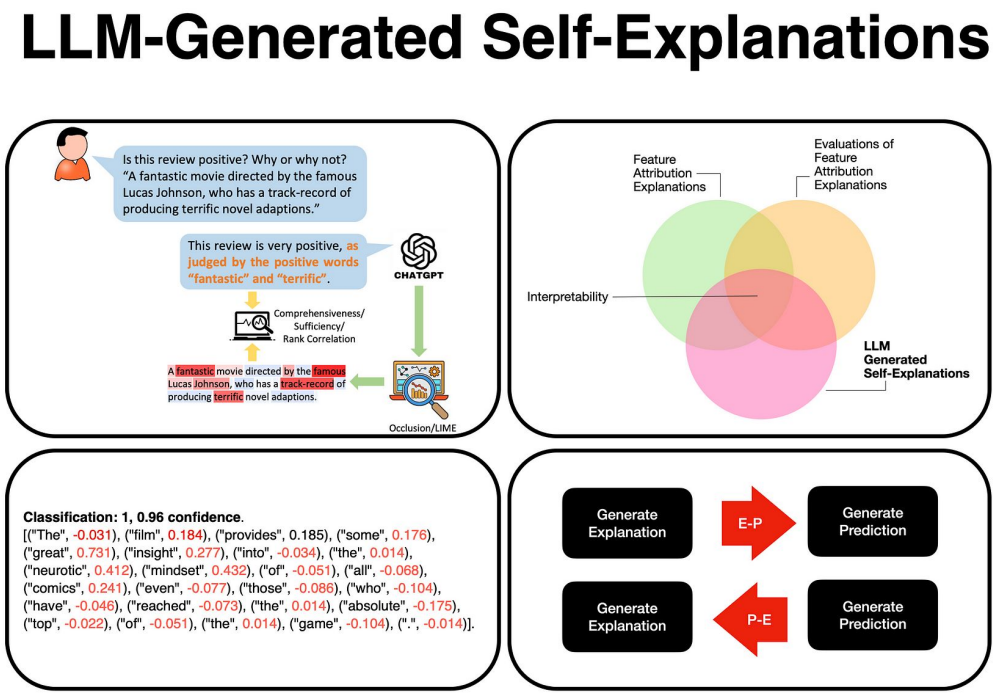
#	Dataset	Modality	Some Challenges of Bias/Fairness
1	CommonCrawl (Raffel et al., 2020)	Text & Vision	Fake news, hate speech, porn & racism (Gehman et al., 2020; Luccioni and Viviano, 2021)
2	LAION-400M & 5B (Schuhmann et al., 2021, 2022)	Text & Vision	Misogyny, stereotypes & porn (Birhane et al., 2021, 2024b)
3	WebImageText (WIT) (Radford et al., 2021)	Text & Vision	Racial, gender biases (Radford et al., 2021)
4	DataComp (Gadre et al., 2024)	Text & Vision	Racial bias (Gadre et al., 2024)
5	WebLI (Chen et al., 2022)	Text & Vision	Age, racial, gender biases & stereotypes (Chen et al., 2022)
6	CC3M-35L (Thapliyal et al., 2022)	Text & Vision	Cultural bias (Thapliyal et al., 2022)
7	COCO-35L (Thapliyal et al., 2022)	Text & Vision	Cultural bias (Thapliyal et al., 2022)
8	WIT (Srinivasan et al., 2021)	Text & Vision	Cultural bias (Srinivasan et al., 2021)
9	Colossal Cleaned CommonCrawl (C4) (Raffel et al., 2020)	Text	Offensive language, racial bias (Raffel et al., 2020)
10	The Pile (Gao et al., 2020a)	Text	Religious, racial, gender biases (Gao et al., 2020a)
11	CCAligned (El-Kishky et al., 2020)	Text	Porn, racial bias (El-Kishky et al., 2020)
12	OpenAI WebText (Radford et al., 2019)	Text	Gender, racial biases (Gehman et al., 2020)
13	OpenWebText Corpus (OWTC)	Text	Gender, racial biases (Gehman et al., 2020)
14	ROOTS (Laurençon et al., 2022)	Text	Cultural bias (Laurençon et al., 2022)

- Summary of papers
  https://github.com/junxu-ai/LLM_fairness?tab=readme-ov-file#survey-papers

Fairness and Bias in Multimodal AI: A Survey, 2024

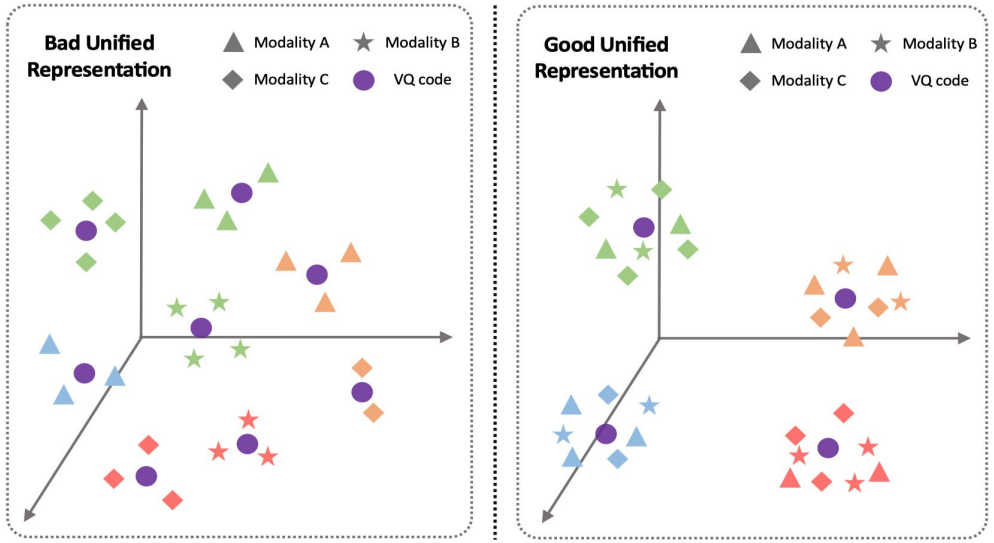
Future Directions of Trustworthiness

- Explainable Multimodal Inference & Reasoning
  - In the era of large multimodal models, seeking transparency in the model architecture itself will undermine system performance
  - A better direction is to pursue the interpretability of the reasoning process



Future Directions of Trustworthiness

- Safety alignment and halluction detection for multimodal models
  - Difficulty
    - The involvement of **multiple modalities** increases the **potential for vulnerabilities**
  - Future Directions
    - Cross-modal inconsistencies detection
    - Unified representation learning



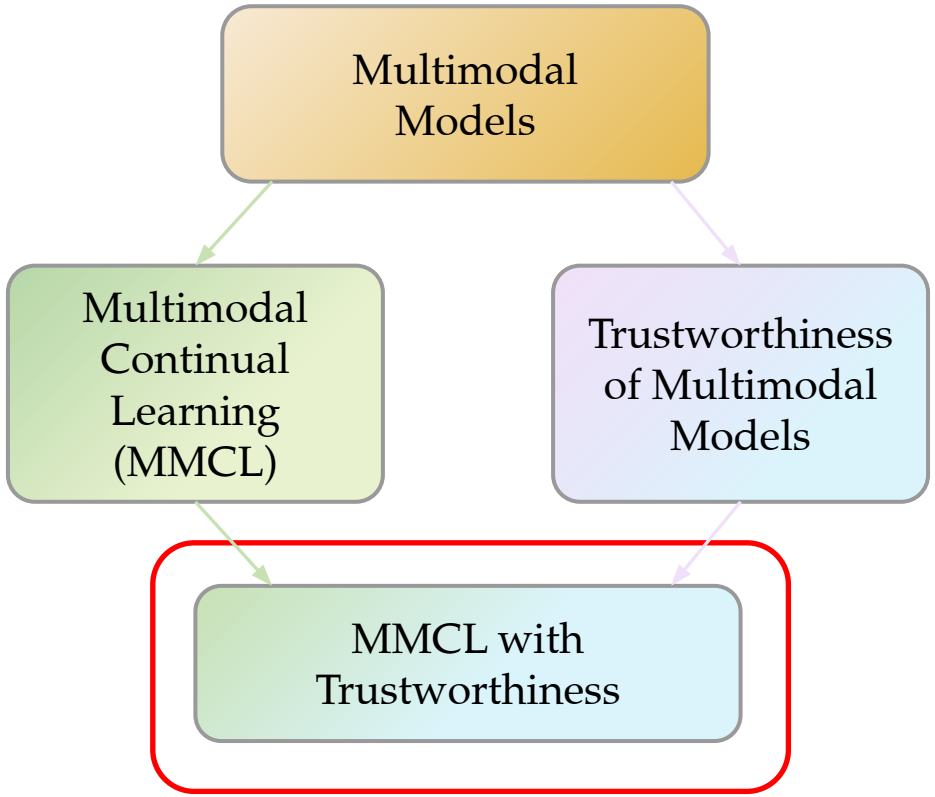
Future Directions of Trustworthiness

- Mitigating modal dependency bias in multimodal data
  - **Dynamic modality alignment under distribution shift**
    - Current approaches: static modality interactions
    - Future: adaptive alignment strategies
  - **Causal inference for bias mitigation**
    - Integrate causal reasoning into modality bias to distinguish correlation from causation
    - Offer theoretical guarantees and strong generalization ability

MMCL with Trustworthiness

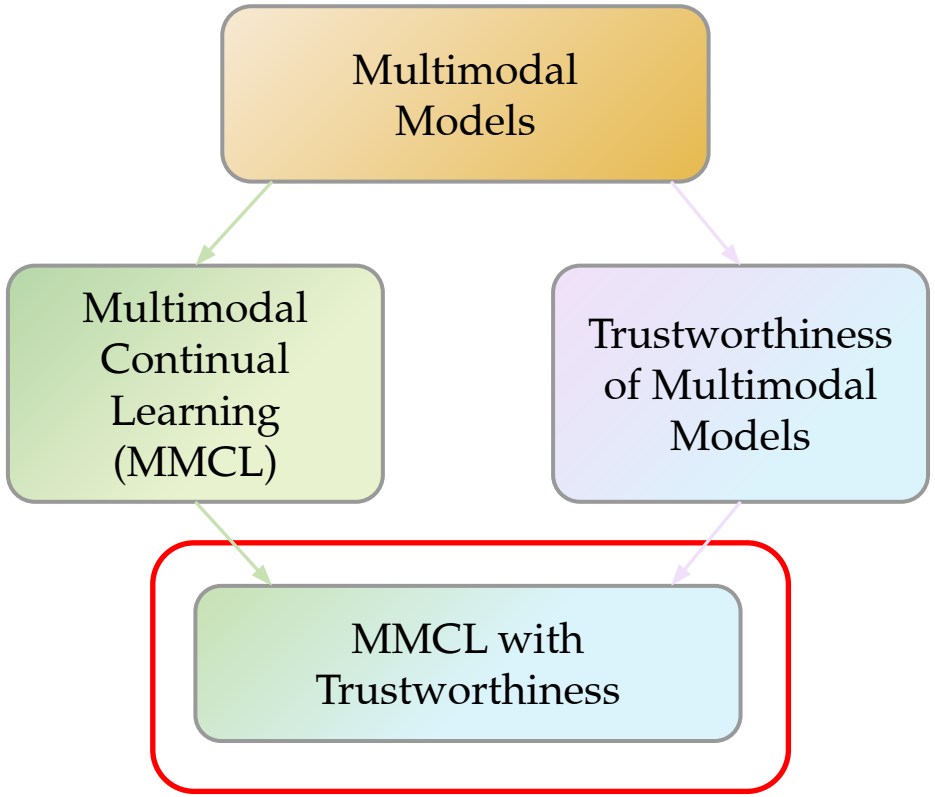
Contents

- MMCL with Trustworthiness
- Motivation
- Techniques
- Key Tasks



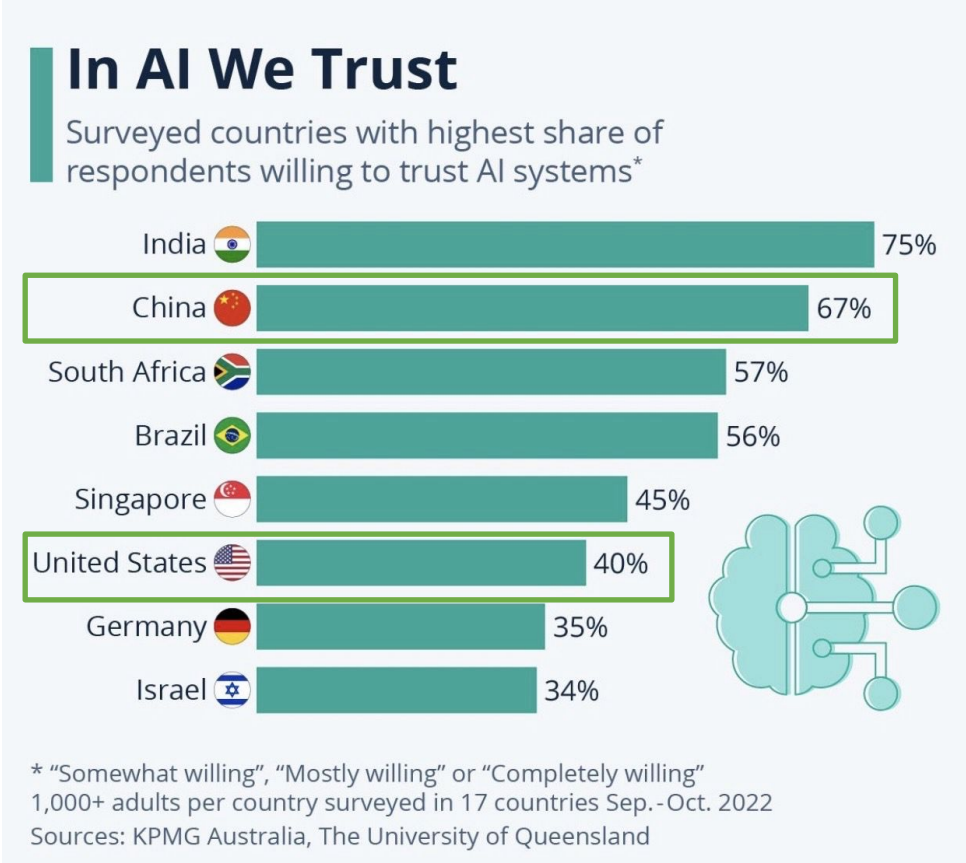
MMCL with Trustworthiness

- Review **ultimate goals**
  - Personalized omnipotent robot
  - Perfect AI doctor – Dynamic Patient Modeling
  - Creative AI – Evolving Content Generation
  - Smart Cities – Adaptive Sensor Networks
- Working with multimodal models and combining advantages of continual learning and trustworthiness!
  - Continuously learn knowledge from new data
  - Ensure trustworthiness in various aspects that discussed above
- **Future Direction: MMCL with Trustworthiness**
  - No existing work yet



MMCL with Trustworthiness

- Motivation
  - Rising demand: Public concern + regulations push for safer, privacy-aware AI
  - Evolving trustworthiness: Safety/privacy requirements change over time
  - Knowledge retention: Models must retain past tasks while integrating new multimodal data
  - Risks in deployment: Hallucination/misinformation mitigation critical for real-world use



MMCL with Trustworthiness

- Techniques
  - Federated Learning (FL): Train models on decentralized data; protect privacy
  - FCL to MMCL: Extend Federated Continual Learning for multimodal privacy preservation
  - Agentic AI: Self-directed agents that autonomously plan, execute, and adapt tasks across multimodal FL systems while preserving privacy and traceability
  - Watermarking: Ensure data/model authenticity + traceability

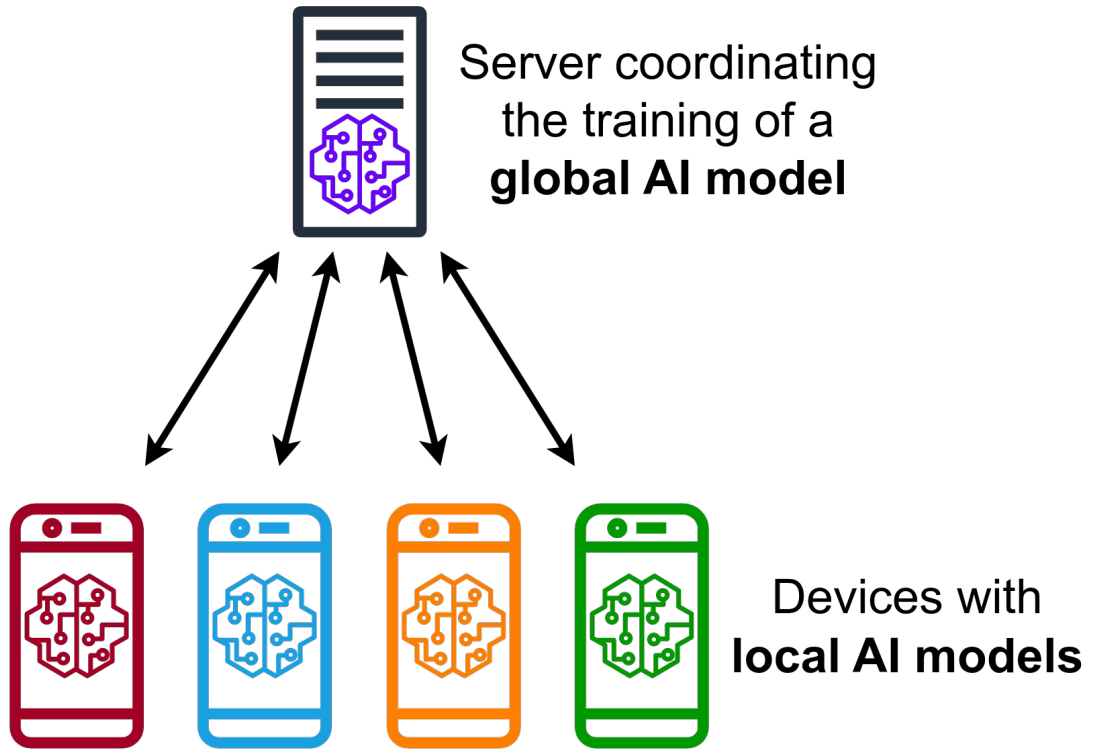


Image credit: https://en.wikipedia.org/wiki/Federated_learning

MMCL with Trustworthiness

- Key Tasks
  - Safety alignment: Align models with dynamic human/social standards
  - Fact-checking: Detect and correct misinformation in multimodal outputs
  - Hallucination control: Mitigate hallucination risks during continual learning (e.g., fine-tuning pitfalls)

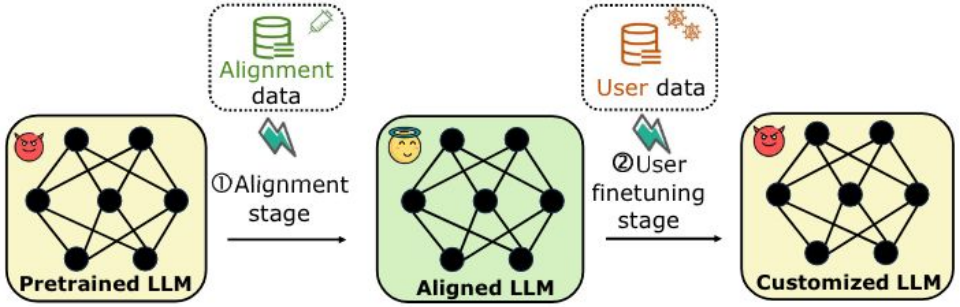


Image credit: Huange et al., Lisa: Lazy safety alignment for large language models against harmful fine-tuning attack, NeurIPS 2024.

Summary of MMCL with Trustworthiness

- Combining the advantages of MMCL and trustworthiness
- Motivation
  - Evolving trustworthiness, Knowledge retention, Risks in deployment
- Techniques
  - Federated Learning, Watermarking
- Key Tasks
  - Safety alignment, Fact-checking, Hallucination control

Conclusion

Summary

- Multimodal models and its challenges
  - Challenge 1: Catastrophic forgetting using direct fine-tuning
  - Challenge 2: Untrustworthiness of existing multimodal models
- Multimodal Continual Learning (MMCL)
  - Necessity and significance of MMCL
  - Methods: Regularization-based, architecture-based, replay-based, prompt-based
- Trustworthiness of Multimodal Models
  - Explainability, Privacy & Security, Robustness, Fairness
- MMCL with Trustworthiness
  - A future direction regarding evolving trustworthiness

- Huge space for further research and exploration!

Useful Resource

- Link of this tutorial
  - https://lucydyu.github.io/MM_CL_T/
- Other good tutorials
  - From Multimodal LLM to Human-level AI: https://mllm2024.github.io/CVPR2024/
  - Tutorial on MultiModal Machine Learning: https://cmu-multicomp-lab.github.io/mmml-tutorial/icml2023/
  - Continual Learning with Deep Architectures: https://sites.google.com/view/cltutorial-icml2021
- Code of representative works
  - ZSCL: https://github.com/Thunderbeee/ZSCL
  - MoE-Adapters4CL: https://github.com/JiazuoYu/MoE-Adapters4CL
  - CLVQA: Link: https://github.com/showlab/CLVQA

QR Codes

MMCL Survey

Trustworthy FL Survey

Discord Group

Conclusion

- Review **ultimate goals**
  - Personalized omnipotent robot
  - Perfect AI doctor – Dynamic Patient Modeling
  - Creative AI – Evolving Content Generation
  - Smart Cities – Adaptive Sensor Networks

- By addressing the challenges outlined in this tutorial and sustaining dedicated research efforts, we can progressively advance toward realizing the ultimate goals!
- **Collaborations are highly welcome!**

Thanks for listening!