



# **Week 14: Multi-modal Transformers: Integrating Vision, Audio, and Text**

**Lecturer: Jordan Hill**



## Learning Objectives

- **Understand** the fundamentals of Multi-modal Transformers encompassing vision, audio, and text.
- **Explore** how transformers integrate multiple data modalities.
- **Discover** applications and implications of multi-modal learning across various domains.
- **Discuss** ethical considerations in multi-modal AI systems.

## Note on NLP Coverage

- **Important Note:** While this course provides an overview of multi-modal transformers, it does not delve deeply into Natural Language Processing (NLP) specifics. Students interested in a comprehensive understanding of NLP beyond simple sentiment analysis etc. are encouraged to explore additional resources or courses dedicated to this field.

**See additional resources provided at the end of the presentation for extension in NLP**

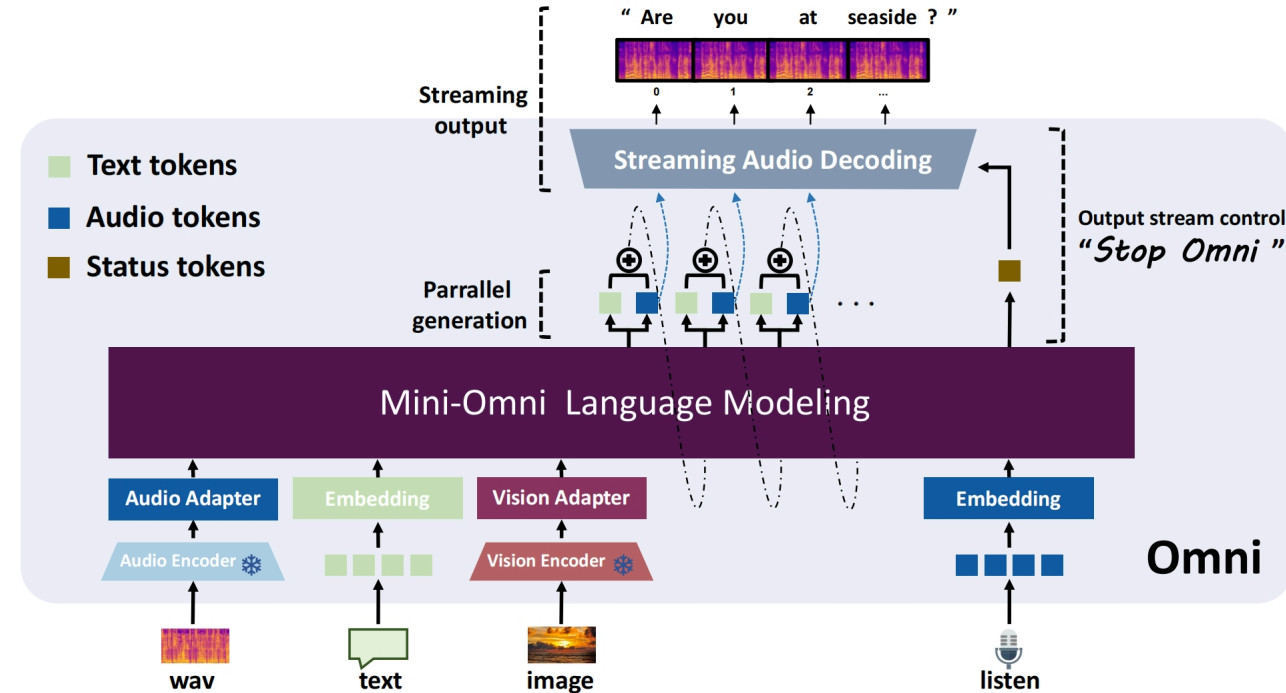
# Introduction to Multi-modal Transformers

- **Definition:** Models that process and integrate multiple data types, such as **text**, **images**, and **audio**.
- **Significance:**
  - Bridges the gap between different modalities for richer understanding.
  - Enhances performance in tasks requiring comprehension of visual, auditory, and textual information.
  - Enables more natural and intuitive human-computer interactions.

# Stages of Multi-modal Transformers Data Processing

## 1. Data Encoding:

- **Text:** Token embeddings using techniques like Word2Vec or BERT.
- **Vision:** Image patches transformed into visual embeddings using CNNs or Vision Transformers.
- **Audio:** Audio signals converted into spectrograms or embeddings using models like Wave2Vec.



# Data Fusion

## 2. Modality Fusion:

- **Fusion Techniques:**

- Early fusion (combining inputs)
- Late fusion (combining outputs)
- Hybrid approaches

- **Transformer Mechanisms:**

- Self-attention and cross-attention layers to learn intra- and inter-modal relationships.

## 3. Prediction:

- The fused representation is used for downstream tasks like classification, generation, or retrieval.

## Benefits of Using Multi-modal Transformers

- **Comprehensive Understanding**: Captures complex patterns and relationships across different modalities.
- **Improved Accuracy**: Combining modalities often leads to better performance than single-modal models.
- **Versatility**: Applicable to tasks in healthcare, entertainment, accessibility, and more.
- **Enhanced User Experience**: Supports more interactive and immersive applications.

## Why Multi-modal? — Applications

### Audio-Visual Speech Recognition

- **Combines** lip movements and audio for robust speech recognition.
- **Use Case**: Understanding speech in noisy environments.

### Video Captioning

- **Generates** textual descriptions of video content by integrating visual frames and audio cues.
- **Use Case**: Improving accessibility for visually impaired users.

## Why Multi-modal? — Applications

### Emotion Recognition

- **Analyzes** facial expressions (vision), tone of voice (audio), and word choice (text).
- **Use Case**: Customer service bots that can detect and respond to user emotions.

### Cross-modal Retrieval

- **Enables** searching for images using text queries or finding audio clips based on images.
- **Use Case**: Multimedia databases and content recommendation systems.



# Challenges in Multi-modal Learning

- **Data Alignment:**
  - Synchronizing modalities with different formats and time scales.
  - Dealing with missing or noisy data in one or more modalities.
- **Computational Complexity:**
  - Increased model size and training time due to multiple data types.
  - Requires significant computational resources and efficient algorithms.
- **Interpretability:**
  - Difficulty in understanding how models make decisions based on complex, fused data.

# Architectures for Multi-modal Transformers

- **Modality-specific Encoders :**
  - Separate encoders process each modality before fusion.
  - Preserves modality-specific features.
- **Joint Encoders :**
  - Inputs from all modalities are combined from the start.
  - May lead to better integration but can lose modality-specific nuances.
- **Cross-modal Attention Mechanisms :**
  - Allows the model to focus on relevant parts of different modalities.
  - Enhances learning of inter-modal relationships.

# Ethical Considerations

- **Privacy:**
  - Handling sensitive information across modalities like voice and facial data.
  - Need for robust data protection measures.
- **Bias and Fairness:**
  - Multi-modal data can reflect societal biases.
  - Important to ensure fairness across different demographic groups.
- **Consent and Transparency:**
  - Users should be informed about how their data is used.
  - Models should be transparent and explainable where possible.

# Workshop: Multi-modal Transformers in Action

## Activity

### 1. Exploring Multi-modal Models :

- Navigate HuggingFace to find models that integrate at least two different modes at the input or output stage; vision, audio, or text.  
(Could be text to text, text to image, image to text, text to audio, audio to text, etc.)

[Hugging Face Models](#)

### 2. Select a model.

### 3. Build a pipeline & test using huggingface transformers and test the model with an example input.

## Extension Activity: Research Your Model's Architecture

- **Research Task:** Investigate the main components and how the model integrates different modalities.
- **Fusion Approach:** Identify and justify whether the model uses early, late, or hybrid fusion.
- **Performance Insights:** Analyze the benefits and limitations of the fusion method.

This activity encourages you to delve deeper into understanding the architecture of your selected model.

## Discussion

### What challenges does multi-modal present as opposed to text-only models?

- **Data Integration Complexity**: Effectively combining and processing diverse data types (text, images, audio) requires sophisticated architectures to handle varying data structures and representations.
- **Alignment and Synchronization**: Ensuring that multi-modal data is properly aligned in time and context can be difficult, especially when modalities have different sampling rates or missing data.
- **Computational Demand**: Multi-modal models often require significantly more computational resources for training and inference due to their increased complexity and the volume of data.

**Anything else?**

# Resources for further skill development

- Audio course: [Hugging Face](#)
- Computer Vision course (transformers): [Hugging Face](#)
- Computer Vision course (multimodal): [Hugging Face](#)

## Summary and Q&A

### Summary

- **Multi-modal Transformers** unlock new possibilities by integrating vision, audio, and text.
- **Applications** are vast, impacting industries like entertainment, healthcare, and accessibility.
- **Challenges** exist but can be addressed with thoughtful design and ethical considerations.

### Discussion Questions

- How can multi-modal models enhance assistive technologies for people with disabilities?
- What are potential risks of multi-modal AI, and how can we mitigate them?