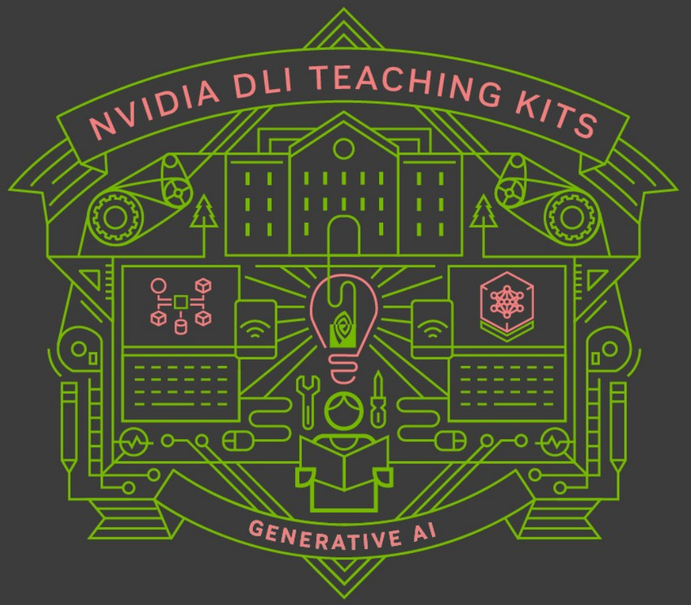
Lecture 5.1 - Introduction to Multimodal Models

Generative AI Teaching Kit



This lecture

- Introduction to Modalities
- Historical Approaches
- Multimodalities in GenAI
- Whisper: Combining Audio and Text with Transformers

Introduction to Modalities

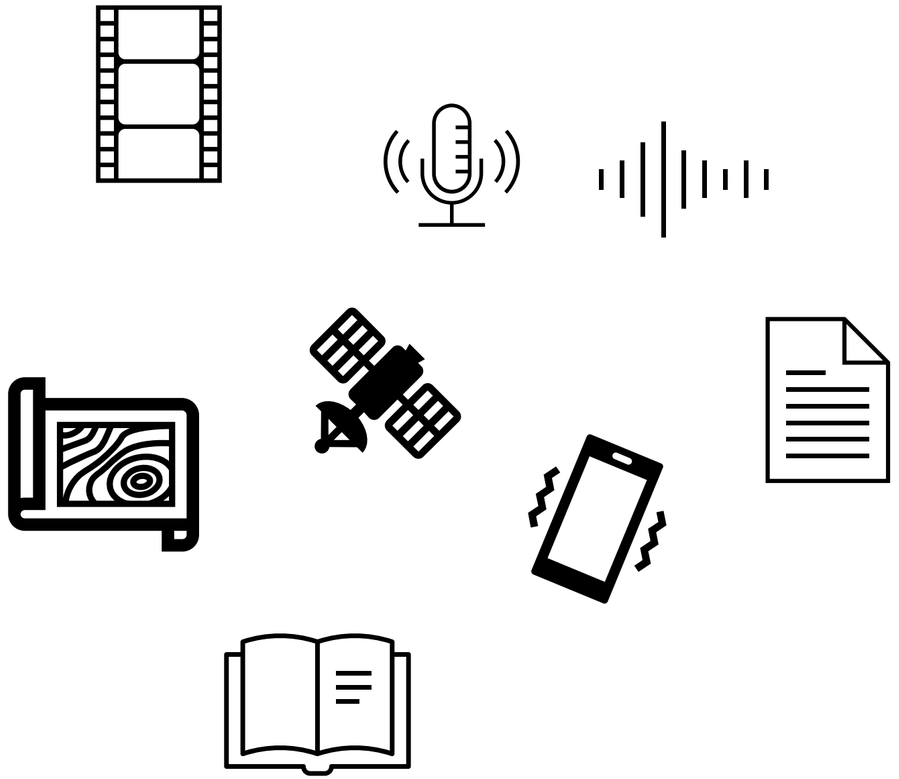
Listen, Read, Speak, Watch

Data Modalities

Modalities are different forms of information or sensory inputs that can be processed

Common modalities include:

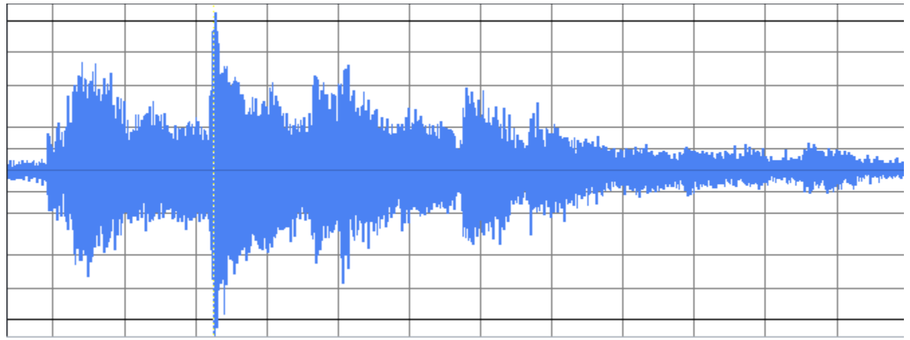
- Visual (images, video, diagrams)
- Audio (speech, music, environmental sounds)
- Text/Language
- Numerical/Structured data
- Time series
- Tactile/haptic feedback
- Spatial/geometric information

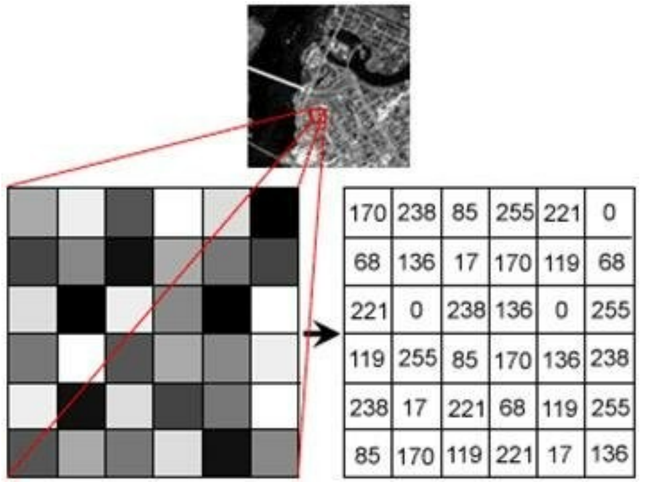


Modalities – Statistical Properties

Each modality has distinct characteristics:

- Different dimensionality (e.g., 2D for images, 1D for audio)
- Varying scales and distributions
- Unique noise patterns and types
- Different sparsity levels
- Temporal dependencies (strong in audio/video, less so in static images)
- Varying degrees of structure (highly structured text vs. raw sensor data)

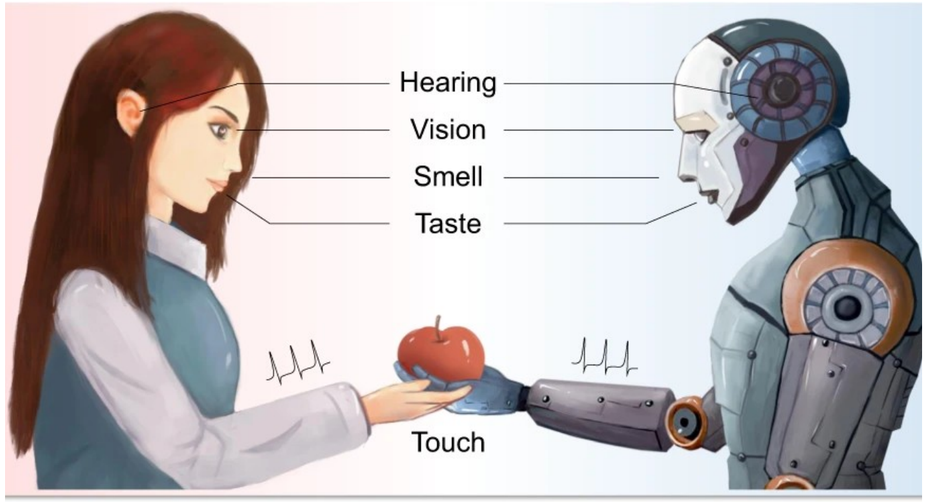




Modalities – Human Processing

Humans naturally integrate multiple modalities:

- Cross-modal integration (e.g., lip reading combines visual and audio)
- Complementary information (text captions helping understand ambiguous images)
- Redundancy for robustness (understanding speech better when seeing the speaker)
- Hierarchical processing (from raw sensory input to abstract concepts)
- Attention mechanisms to focus on relevant modal information



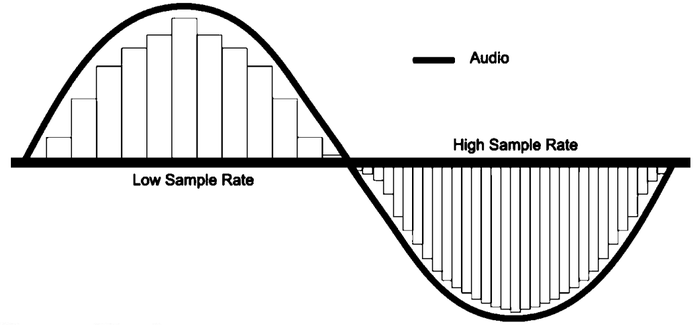
Modality Incompatibilities – Sampling Rates

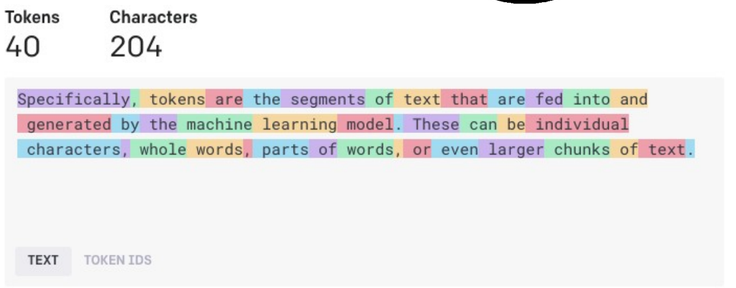
Technical obstacles:

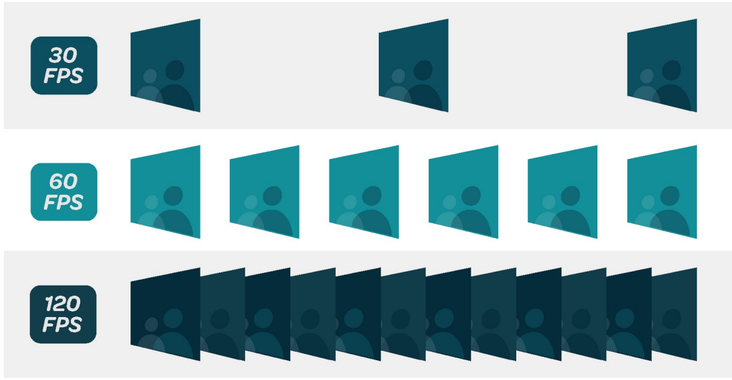
- Images might be 224x224 pixels while audio is 16kHz
- Text comes in discrete tokens vs continuous sensor values
- Video frames at 30fps vs audio at 44.1kHz

Design considerations:

- Need for modality-specific encoders
- Synchronization/resampling strategies
- Managing computational complexity across scales







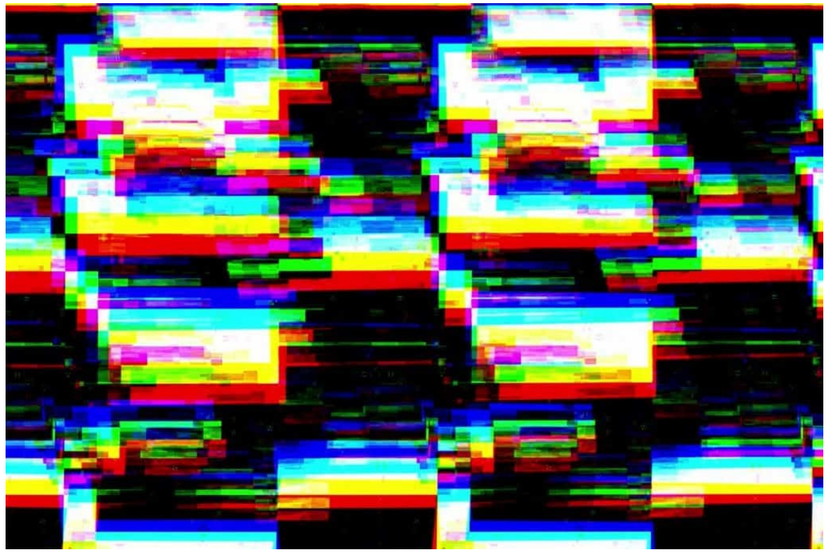
Modality Incompatibilities – Missing or Noisy Data

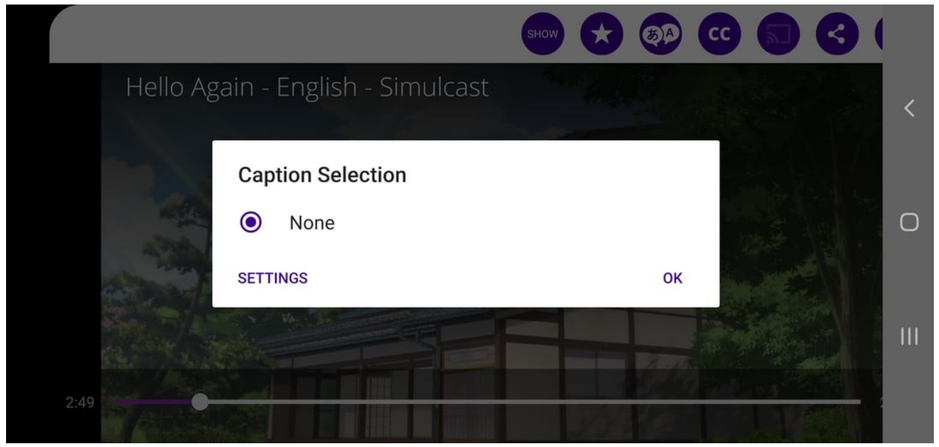
Common scenarios:

- Images without captions
- Videos with corrupted audio
- Partial sensor readings

Key challenges:

- Handling incomplete pairs during training
- Robust inference with missing modalities
- Cross-modal imputation strategies
- Quality assessment across modalities





Modality Incompatibilities – Alignment between modalities

Temporal alignment:

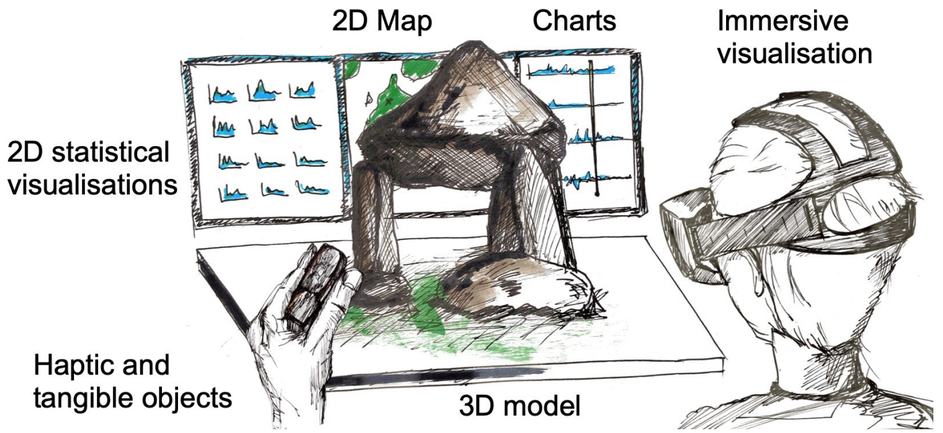
- Synchronizing speech with lip movements
- Matching text descriptions to video segments
- Handling varying speeds/durations

Semantic alignment:

- Connecting abstract concepts across modalities
- Dealing with ambiguous relationships
- Managing cultural/contextual differences

Spatial alignment:

- Object references across image and text
- Mapping 3D scenes to 2D descriptions
- Coordinating multiple viewpoints



Historical Approaches

Previous attempts to build multimodal models

Early deep learning approaches to multimodal models

CNN + LSTM Approaches

- Common architectures:
  - CNN for visual features
  - LSTM for sequential data (text/audio)
  - Early visual-linguistic models (2014-2016)
- Notable examples:
  - Show and Tell (2015): CNN+LSTM for image captioning
  - VQA (2015): CNN features + LSTM for question answering
- Limitations:
  - Sequential bottleneck in LSTMs
  - Limited bidirectional interaction
  - Fixed visual features from pretrained CNNs

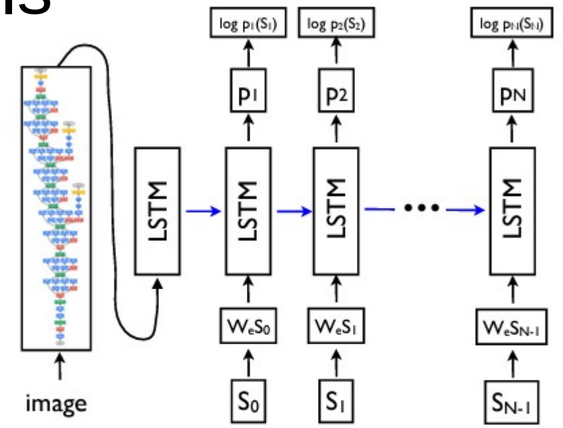


Figure 3. LSTM model combined with a CNN image embedder (as defined in [12]) and word embeddings. The unrolled connections between the LSTM memories are in blue and they correspond to the recurrent connections in Figure 2. All LSTMs share the same parameters.

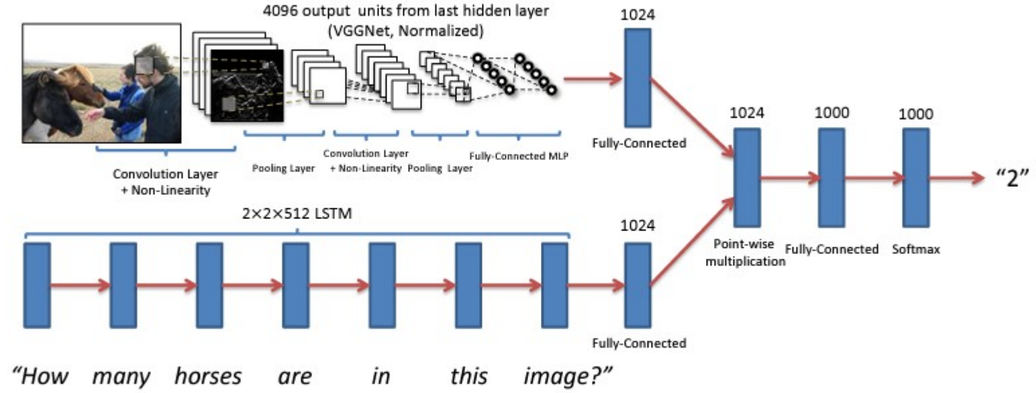


Fig. 8: Our best performing model (deeper LSTM Q + norm I). This model uses a two layer LSTM to encode the questions and the last hidden layer of VGGNet [48] to encode the images. The image features are then ℓ_2 normalized. Both the question and image features are transformed to a common space and fused via element-wise multiplication, which is then passed through a fully connected layer followed by a softmax layer to obtain a distribution over answers.

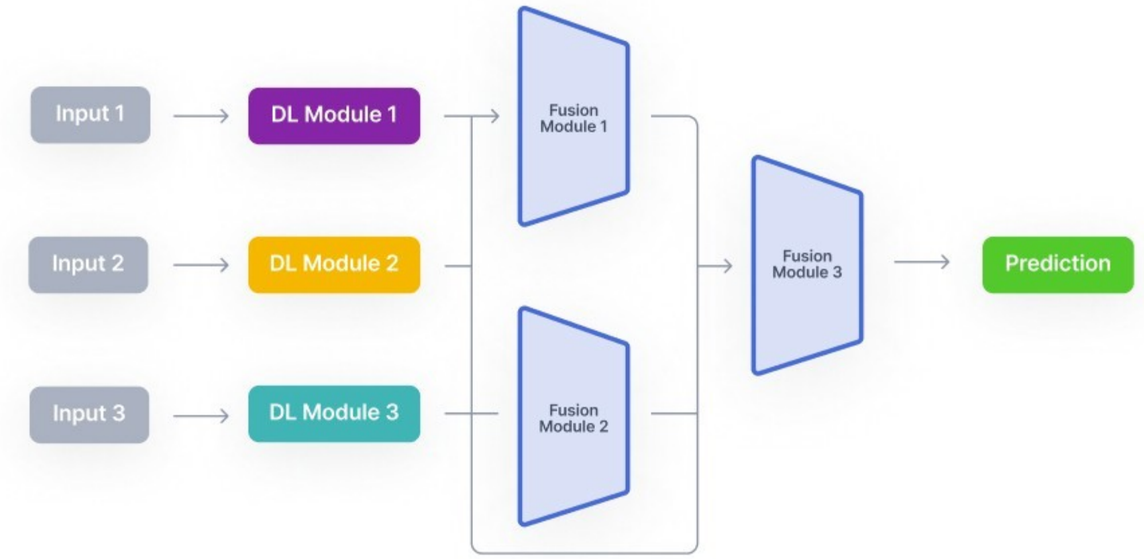
Early deep learning approaches to multimodal models

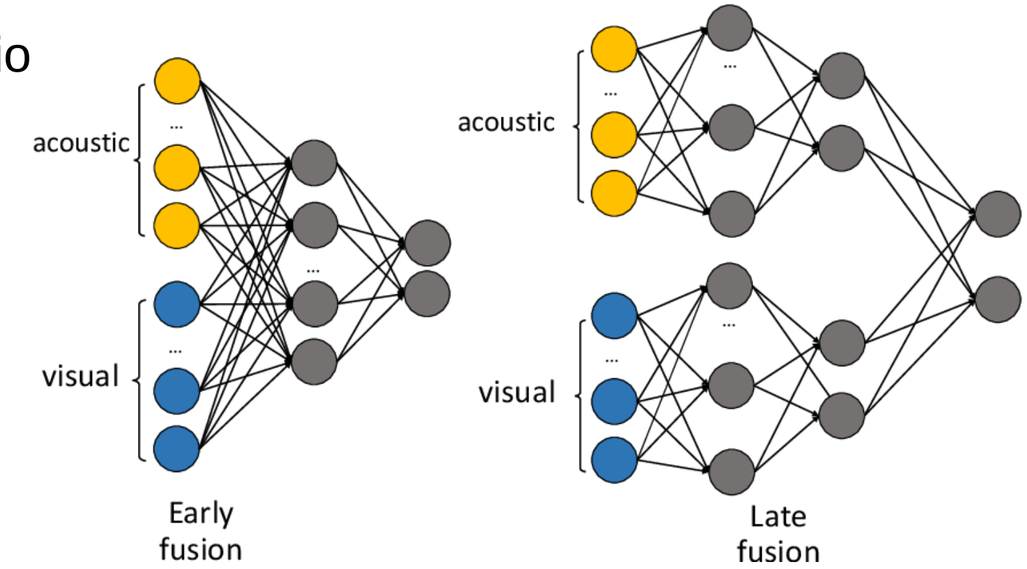
Feature-level fusion:

- Simple concatenation of CNN + LSTM features
- Element-wise operations (add, multiply)
- Basic attention mechanisms

Common architectures:

- Two-stream networks
- Encoder-decoder with feature injectio
- Hierarchical fusion networks





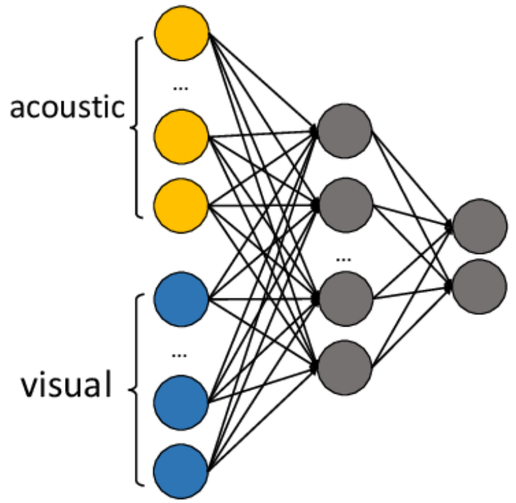
Early deep learning approaches to multimodal models

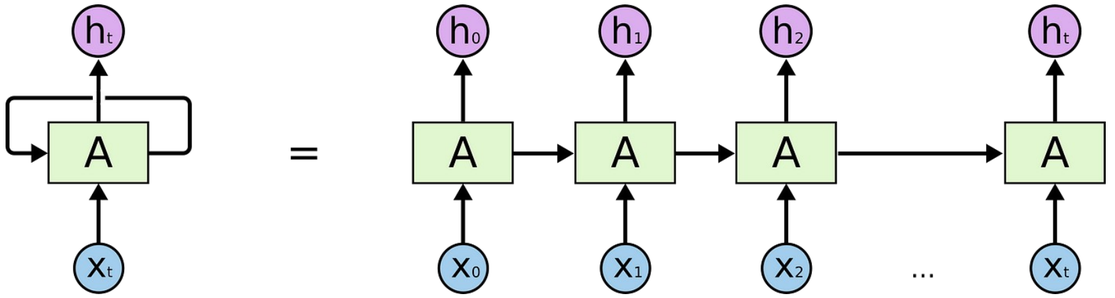
Design patterns:

- Mostly pipeline approaches
- Limited end-to-end training
- Heavy reliance on pretraining
- Modality-specific architectures first

Technical constraints:

- GPU memory limitations
- Computational bottlenecks
- Training instability
- Limited cross-modal interaction





Modern Foundations

Utilizing attention and modern AI tools to enable Multimodal Models

Modern Text Modeling

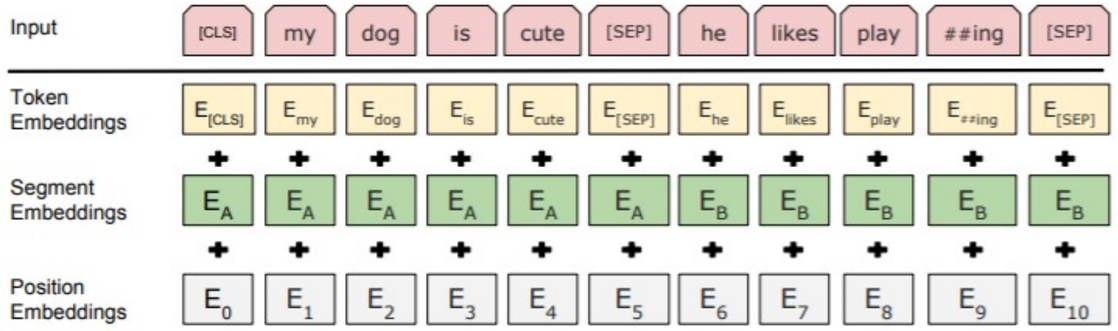
Text Tokenization

- Subword tokenization (BPE, WordPiece, SentencePiece)
- Character-level vs subword vs word-level tradeoffs
- Specialized tokenizers for code/formulas

Implementation considerations:

- Vocabulary size optimization
- Special token handling ([CLS], [SEP], [MASK])
- Length standardization/truncation
- Language-specific considerations
- Handling of emojis/special characters

Iteration	Sequence	Vocabulary
0	a b a b c a b c	{a, b, c}
1	ab ab c ab c	{a, b, c, ab}
2	ab abc abc	{a, b, c, ab, abc}
3	ababc abc	{a, b, c, ab, abc, ababc}
4	ababcabc	{a, b, c, ab, abc, ababc, ababcabc}



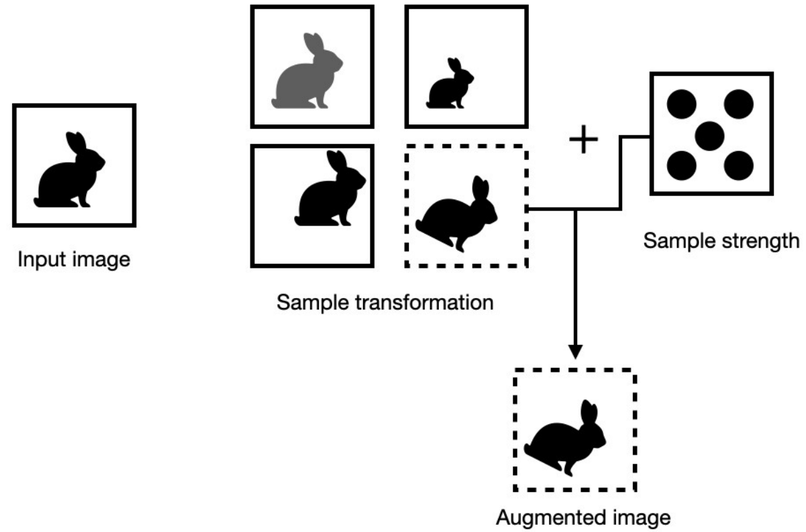
Modern Image Modeling

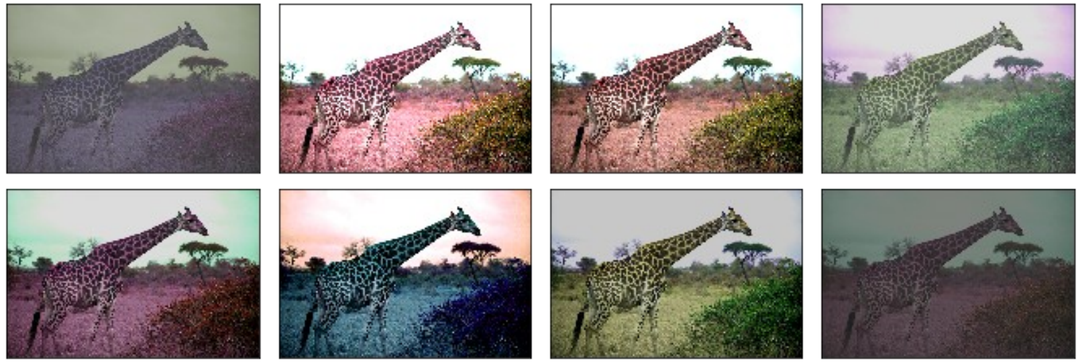
Standard pipeline:

- Resizing/cropping strategies (224x224, 384x384, 512x512)
- Normalization (ImageNet stats commonly used)
- Patch extraction (16x16 or 32x32 patches)
- Positional embeddings

Advanced techniques:

- Random augmentations during training
- Color jittering/normalization
- Aspect ratio handling
- Resolution adaptation
- Region feature extraction





Modern Audio Processing

Spectrogram generation:

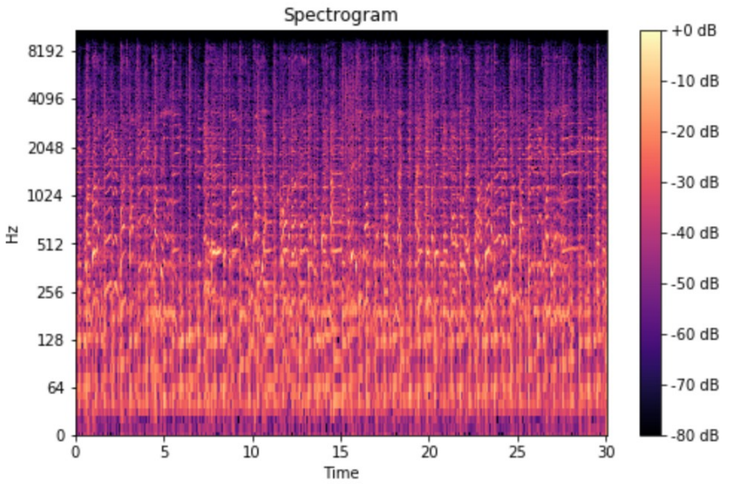
- Short-time Fourier transform (STFT)
- Mel-scale conversion
- Log-scale transformation

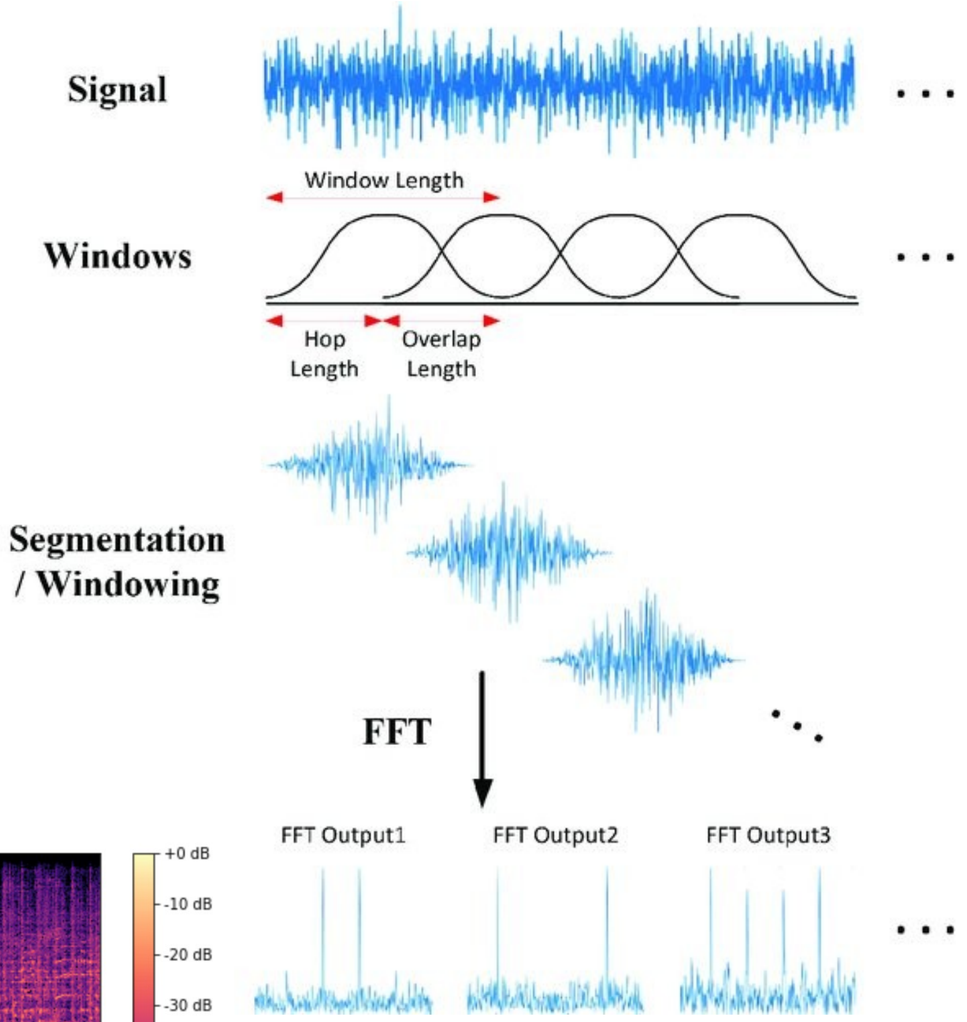
Key parameters:

- Window size/overlap
- Number of mel bands
- Frequency range
- Sample rate considerations

Additional processing:

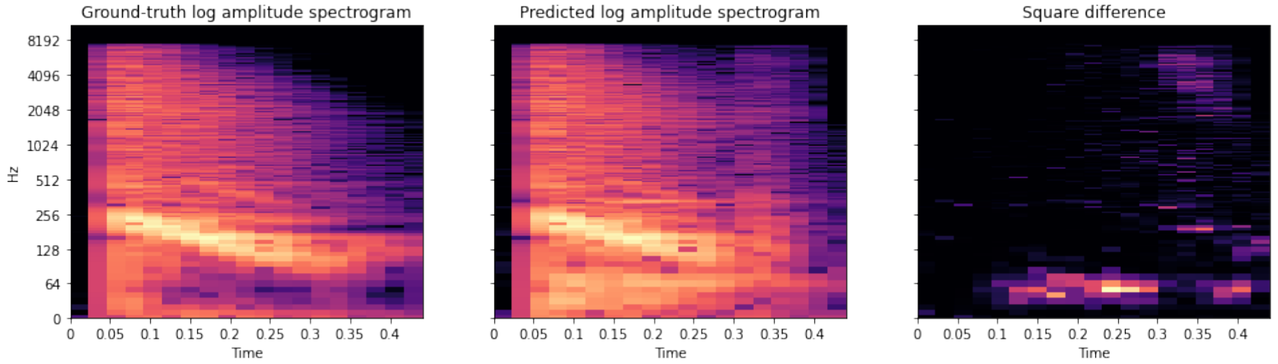
- Normalization strategies
- Time-frequency masking
- Phase information handling
- Feature stacking/splicing





Modern Training Strategies

Training multimodal models involves careful consideration for the losses relative to each mode's input/output relationship

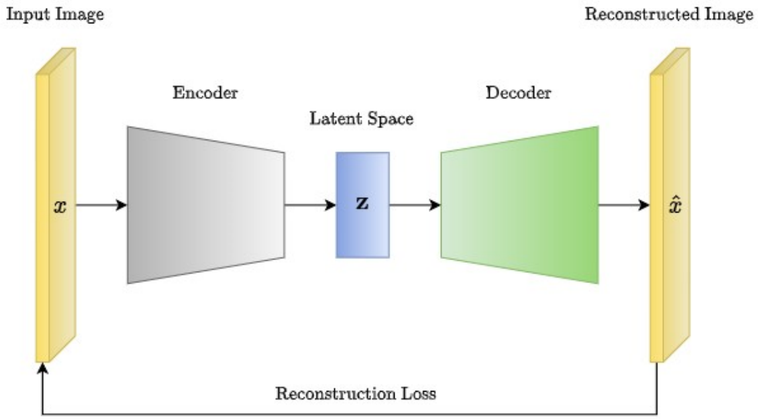


Per modality:

- MSE for continuous features
- L1 for sparse signals
- Perceptual losses for images
- Spectral losses for audio

Cross-modal:

- Cycle consistency losses
- Feature matching losses
- Style-content separation
- Domain adaptation losses



DARTMOUTH ENGINEERING | NVIDIA

Audio and Text Modality - Whisper

Applying Transformers Automatic Speech Recognition

The Whisper Model – Transforming Transcription

In 2022, OpenAI released the **Whisper** family of models

Whisper was one of the first multimodal applications of the transformer architecture with the purpose of audio transcription and translation.

Audio Input processing:

- Mel spectrograms (80 bands)
- 30-second context windows
- 25ms frame length, 10ms stride

Model components:

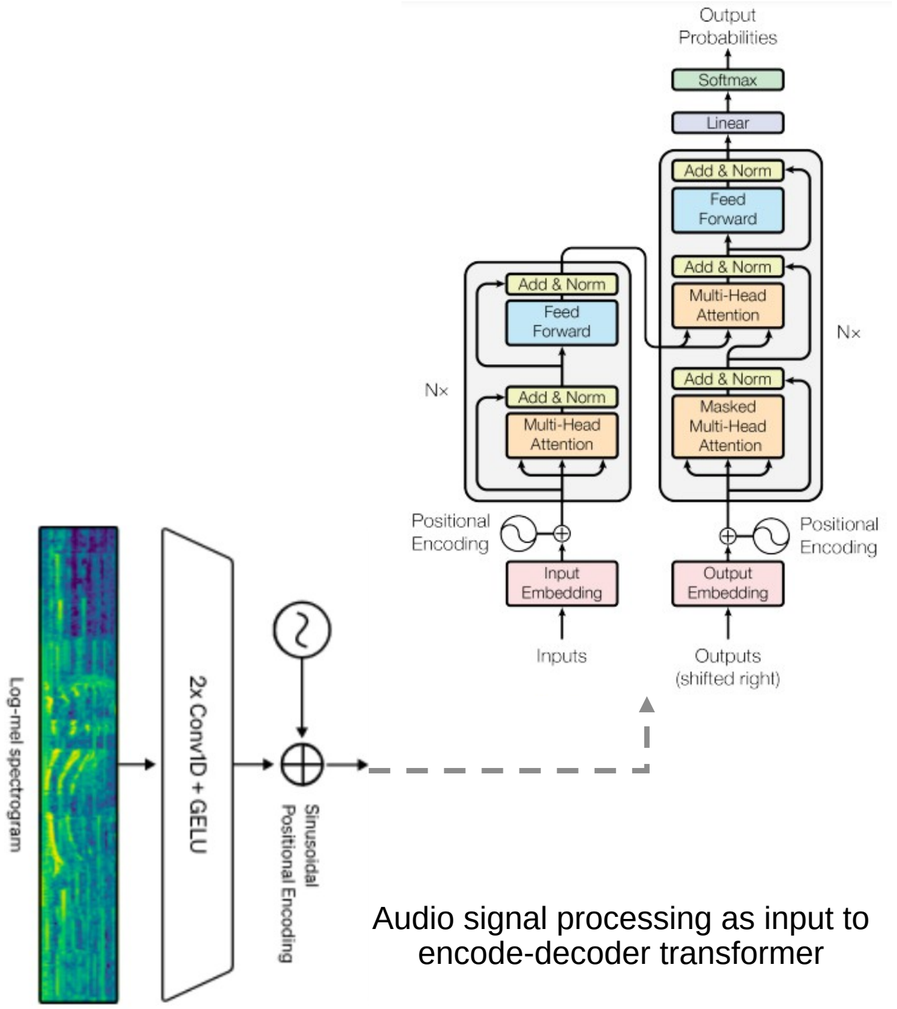
Input:

Convolutional frontend to process the input audio spectrograms

Transformer backbone:

Encoder: 12-24 transformer layers to encode the audio signal

Decoder: Causal masking for autoregressive generation



Audio signal processing as input to encode-decoder transformer

The Whisper Model – Transforming Transcription

Whisper Model Design

Convolutional frontend

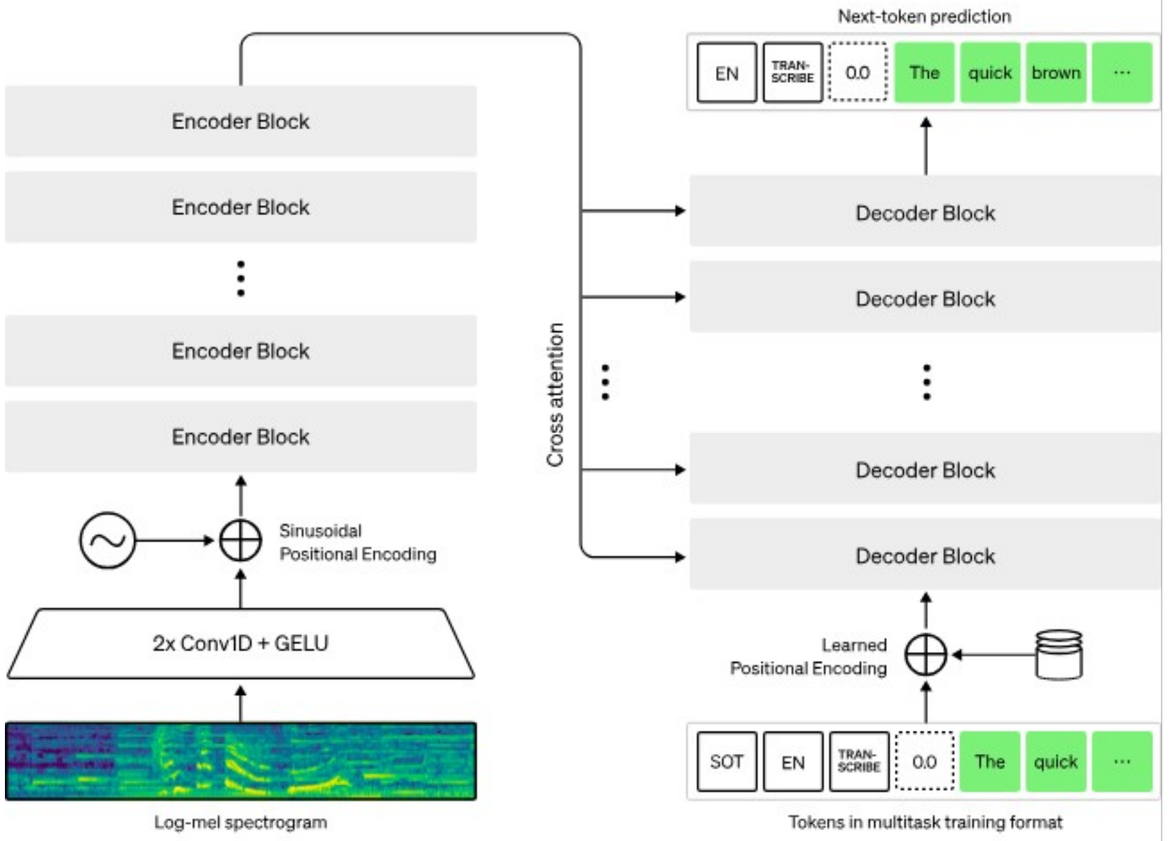
- Down-sampling and feature extraction
- Local pattern recognition

Encoder:

- 12-24 transformer blocks (size dependent)
- Multi-head cross-attention
- Layer normalization strategy

Decoder:

- Causal masking for autoregressive generation
- Cross-attention to audio features
- Vocabulary of 50k tokens (similar to GPT-3)
- Special tokens for tasks/languages



Model	Layers	Width	Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	12	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M

Table 1. Architecture details of the Whisper model family.

Training Whisper on large data and tasks

Dataset composition:

- WebText-like filtering
- YouTube, podcasts, audiobooks
- **98 languages represented**
- Quality tiering system

Training methodology:

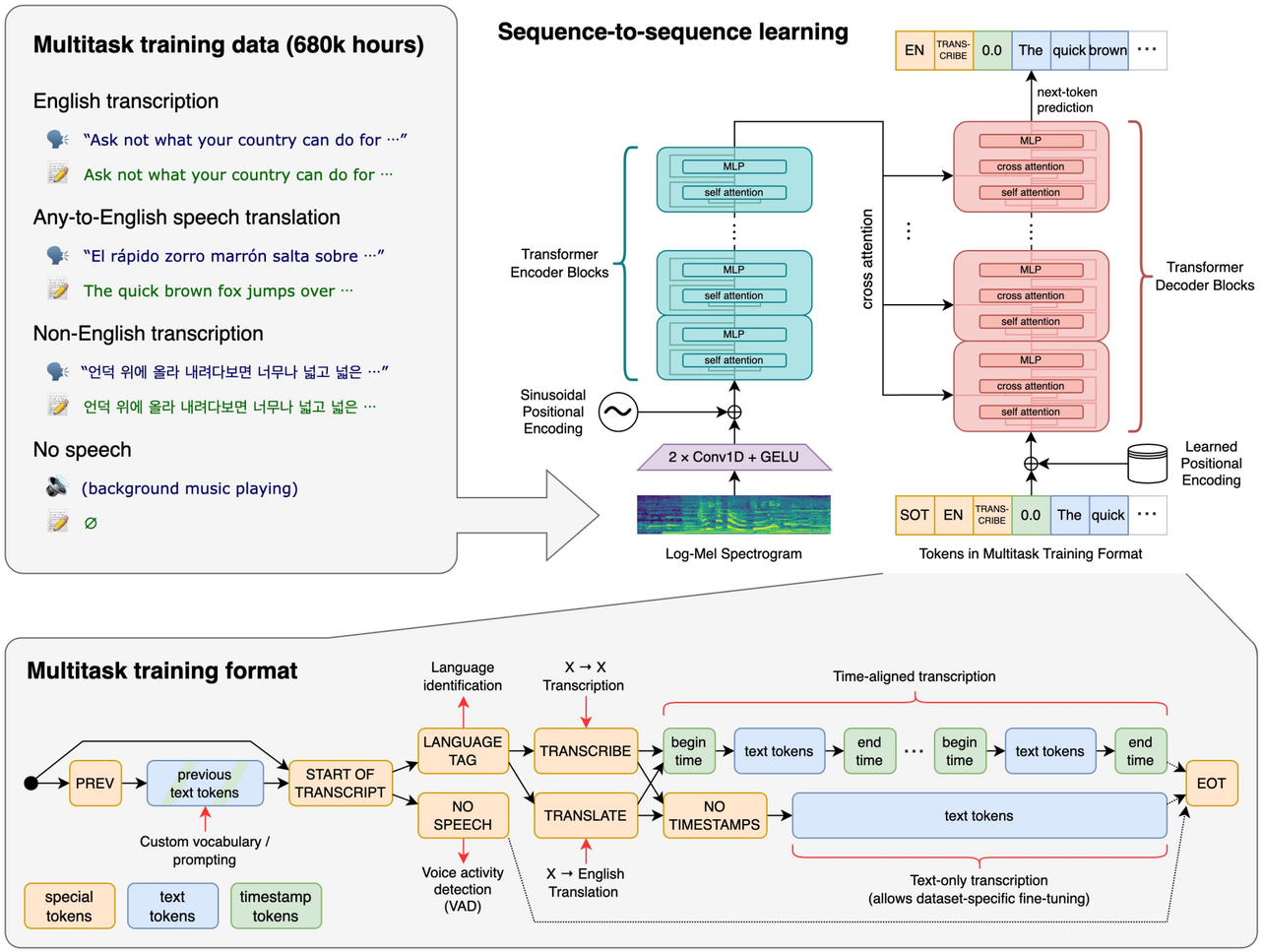
Multi-task setup:

- Speech recognition
- Translation
- Language identification
- Transcription formats (verbatim/clean)

Data augmentation:

- Speed perturbation
- Background noise addition
- Channel dropping
- Time masking

At 680,000 hours of labeled audio, the Whisper dataset is one of the largest ever created in supervised speech



Whisper Performance on Benchmarks

Speech recognition research typically evaluates and compares systems based on the word error rate (WER) metric.

Benchmark datasets:

- LibriSpeech
- Common Voice
- MUST-C
- Fleurs

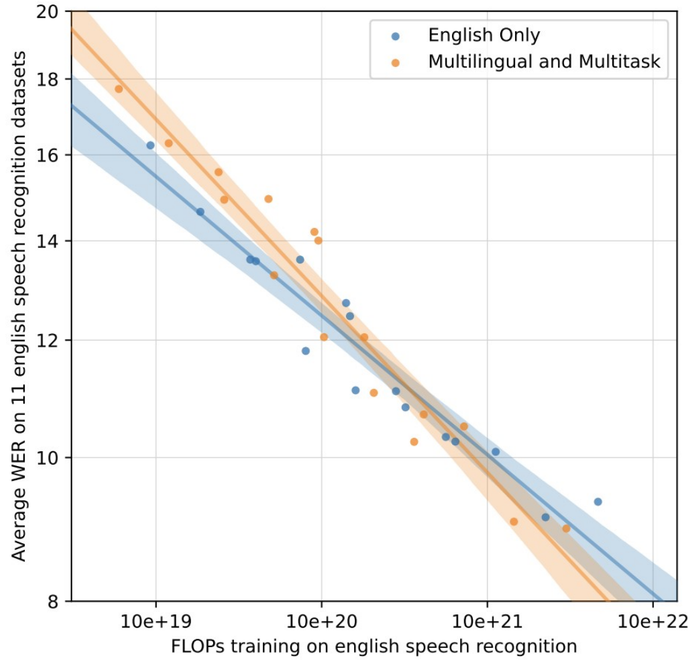
Metric categories:

Transcription accuracy:

- Word Error Rate (WER)
- Character Error Rate (CER)
- Proper noun accuracy

Translation quality:

- BLEU score



Dataset size	English WER (↓)	Multilingual WER (↓)	X→En BLEU (↑)
3405	30.5	92.4	0.2
6811	19.6	72.7	1.7
13621	14.4	56.6	7.9
27243	12.3	45.0	13.9
54486	10.9	36.4	19.2
681070	**9.9**	**29.2**	**24.8**

Table 6. **Performance improves with increasing dataset size.**

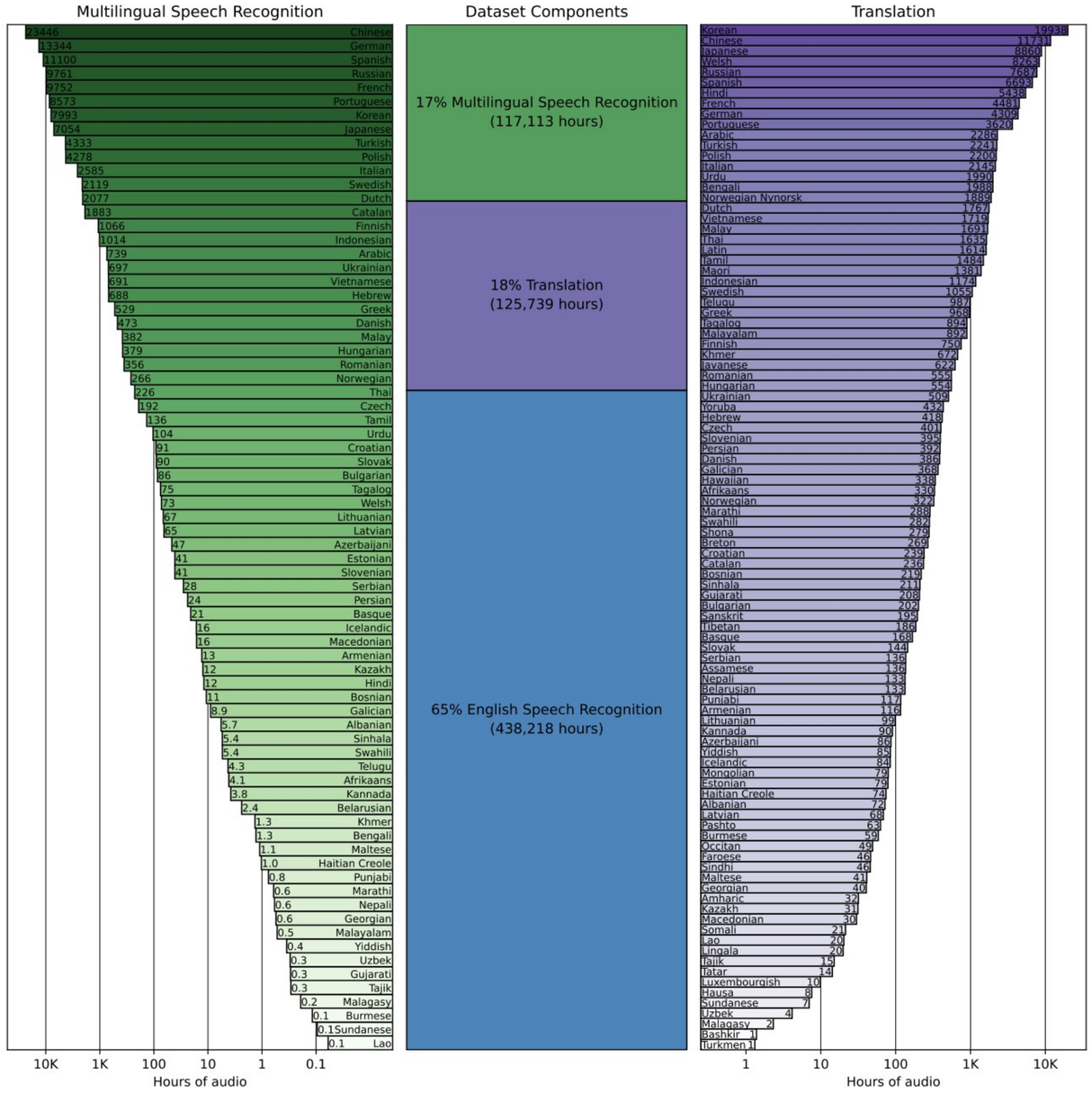
Transformers as Ideal Multimodal Backbones

Architectural Simplicity

- Demonstrated that a standard transformer architecture could handle non-text data
- Minimal audio-specific modifications needed
- Success through scale rather than complex design

Data-Centric Innovation

- Showed the power of large-scale, diverse, noisy training data
- 680,000 hours of multilingual audio
- Proved robustness emerges from data diversity, not architectural complexity



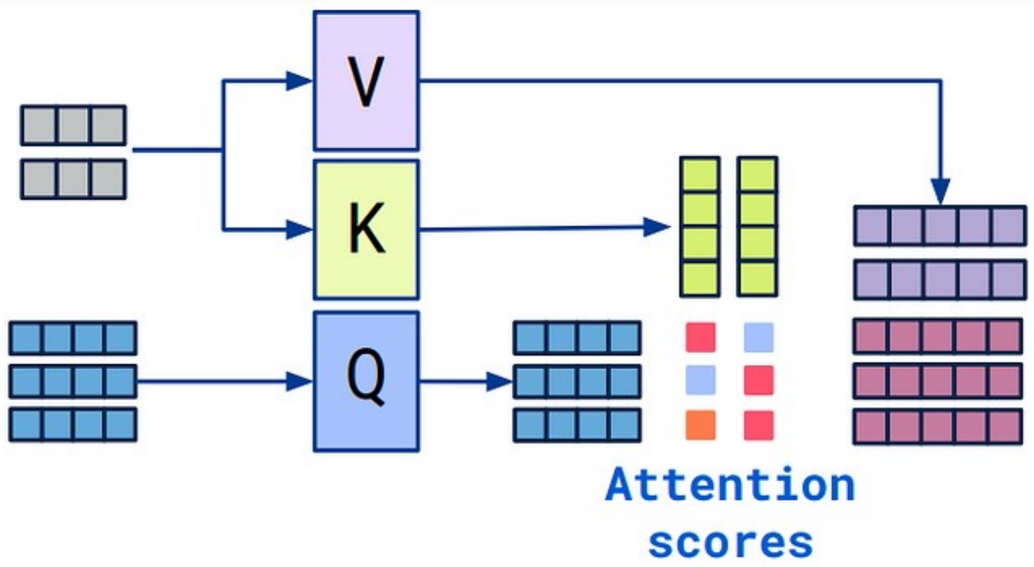
Transformers as Ideal Multimodal Backbones

Cross-Modal Translation

- Direct speech-to-text translation across 99 languages
- Bridged the gap between audio and text modalities
- Demonstrated transformers could learn cross-modal relationships

Foundation for Multimodality

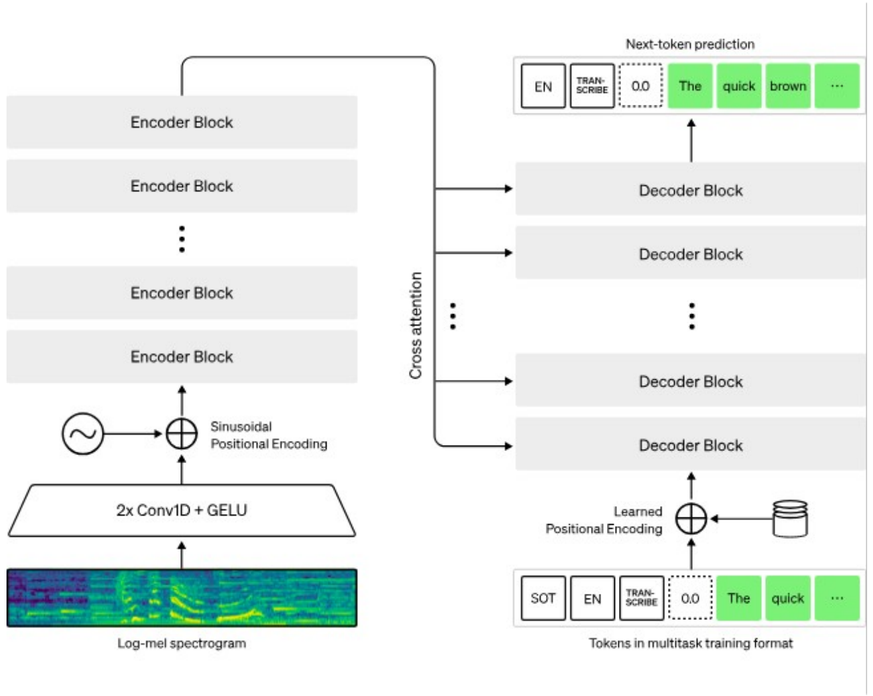
- Proved transformers could effectively process alternative input formats
- Suggested a general pattern: adapt input encoding, keep core architecture
- Hinted at transformer's potential as a universal sequence processor



Wrap Up

Introduction to Multimodal Models

- Today we introduced the concept of modalities of data and how they can be processed with machine learning
- The value of multimodal models arises when comparing the ability of monomodal with how humans process various data streams
- We discussed some of the older approaches to combine modes of data in deep learning with feature fusion
- The Whisper model family of speech-to-text analysis was introduced as one of the first applications of the transformer architecture
- Attention and the transformer approach was discussed as a general framework to combine modal features



In the next class we will explore this transformer approach with other modalities

Thank you!