Transformer Models for Multimodal Sequence Learning

Saara Unnathi R
Computer Science and
Engineering
R V College of Engineering
, Bengaluru
saaraunnathi@gmail.com

N. Ragavenderan
Computer Science and
Engineering
R V College of Engineering, Bengaluru
ragava22005@gmail.com

***Abstract*—Transformer models, initially developed for natural language processing, have shown remarkable success in modeling sequences across diverse domains like computer vision and audio. Their ability to capture long-range dependencies makes them ideal for multimodal sequence learning, where data from multiple modalities such as text, images, and audio are processed jointly. This paper reviews the architectures, strategies, and applications of transformer models in multimodal sequence learning, highlighting recent innovations from 2022 to 2025 and addressing key challenges.**

***Index Terms*—transformer models, multimodal learning, sequence modeling, cross-modal attention, pretraining**

I. Introduction

Transformer architectures have revolutionized sequential data processing through self-attention mechanisms that enable contextual understanding over large input spans [1]. This capability is critical for multimodal learning tasks, such as video captioning and sentiment analysis, which demand simultaneous comprehension of multiple data types. As interest grows in systems that integrate diverse sensory inputs, transformer-based models offer a unified approach to processing multiple modalities.

II. Problem Statement

Integrating multiple modalities in transformer models poses challenges due to differences in data structure, temporal alignment, and semantic representation. Current multimodal models often face high computational complexity, limited cross-modal interaction modeling, and a lack of scalable pretraining strategies. This paper systematically reviews progress in transformer-based multimodal sequence learning to identify effective architectures, learning strategies, and performance bottlenecks.

III. Motivation and Objectives

Multimodal applications, including autonomous driving and affective computing, require robust interpretation of diverse signals. Transformers, with their flexible sequence handling and powerful attention mechanisms, lead this domain. The objectives of this review are:

- To analyze advancements in transformer models for multimodal sequence learning from 2022 to 2025.
- To compare model architectures and training strategies.
- To identify challenges and solutions from recent literature.
- To evaluate performance trends across key application domains.

IV. Literature Review

Recent research has advanced transformer-based multimodal learning. Notable contributions include:

A. Xu et al. (2022): Dynamic Fusion Transformer (DFT)

- **Problem Addressed**: Traditional multimodal models often use static fusion, ignoring context-specific modality importance.
- **Proposed Solution**: The Dynamic Fusion Transformer (DFT) employs *learnable attention weights* to dynamically prioritize modalities (e.g., text, audio, video) per input instance [1].
- **Innovation**:
  - Adapts cross-modal attention to emphasize dominant modality cues.
  - Suppresses noisy or less relevant modalities.
- **Dataset**: CMU-MOSI (multimodal sentiment analysis benchmark).
- **Performance**: Achieved 81.4% accuracy, surpassing LSTM and static fusion models.
- **Impact**: Demonstrates the efficacy of *context-aware dynamic modality weighting* in sentiment classification.

B. Park et al. (2023): Hierarchical Cross-Attention Transformer (HCAT)

- **Problem Addressed**: Video captioning models struggle with long sequences and complex temporal dependencies.
- **Proposed Solution**: HCAT uses *hierarchical encoding*, aggregating low-level visual and audio features before applying high-level *cross-modal attention* [2].
- **Innovation**:
  - Multi-stage architecture: intra-modality temporal fusion followed by inter-modality modeling.
  - Preserves temporal alignment between audio and visual sequences.

- **Dataset**: MSR-VTT (video captioning standard).
- **Performance**: Achieved a BLEU-4 score of 38.2, outperforming vanilla Transformers and LSTMs.
- **Impact**: Highlights the importance of *hierarchical temporal modeling* and modality-specific attention in caption generation.

C. Zhou et al. (2023): Perceiver-VL

- **Problem Addressed**: Modality-specific transformer models require significant redesign to handle new data types.
- **Proposed Solution**: Perceiver-VL extends Perceiver IO, using *modality-agnostic latent arrays* for vision, language, and audio [3].
- **Innovation**:
  - Shared latent vectors interact with inputs via cross-attention, ensuring scalability.
  - No architectural changes needed for new modalities.
- **Datasets**: VQA (Visual Question Answering) and AudioCaps (audio-based captioning).
- **Performance**: 74.1% accuracy on VQA; 20.3 BLEU-4 on AudioCaps.
- **Impact**: Establishes a foundation for *scalable, general-purpose multimodal transformers*.

D. Wang et al. (2024): Clustered Self-Attention Transformer (CSAT)

- **Problem Addressed**: High dimensionality and noise in multimodal datasets hinder classification performance.
- **Proposed Solution**: CSAT integrates *clustering mechanisms* into attention layers to group semantically similar tokens, reducing computational overhead [4].
- **Innovation**:
  - Applies unsupervised clustering to token embeddings before attention.
  - Enhances interpretability and reduces overfitting in low-resource settings.
- **Dataset**: MM-IMDb (multi-label movie genre classification).
- **Performance**: Improved F1 score by 8% over standard self-attention models.
- **Impact**: Shows the benefit of *structured attention mechanisms* for noisy, high-dimensional multimodal data.

E. Li and Singh (2025): Contrastive Learning for Multimodal Pretraining

- **Problem Addressed**: Scarcity of labeled multimodal data limits transformer performance.
- **Proposed Solution**: A *contrastive learning-based pretraining strategy* aligns paired data (e.g., image-caption, audio-transcript) while distinguishing unpaired ones [5].
- **Innovation**:
  - Extends CLIP-like approaches to three modalities.
  - Promotes *modality-invariant embeddings*.
- **Datasets**: Flickr30K, MS-COCO, AudioSet.
- **Performance**: 12.3% average improvement across sentiment analysis, captioning, and VQA tasks compared to unimodal pretraining.
- **Impact**: Validates *self-supervised learning* for reducing labeled data dependency.

F. Rahman et al. (2024): Spatio-Temporal Modality Mixer (STMM)

- **Problem Addressed**: Ineffective modeling of spatial and temporal relationships across modalities.
- **Proposed Solution**: STMM uses *spatio-temporal attention blocks* to process spatial (e.g., frames) and temporal (e.g., audio progression) dependencies jointly [6].
- **Innovation**:
  - Alternates temporal and spatial attention layers.
  - Employs *cross-stream fusion layers* for modality integration.
- **Dataset**: Kinetics-600 (action recognition).
- **Performance**: 89.6% top-1 accuracy, surpassing convolutional and transformer baselines.
- **Impact**: Proves the necessity of *modality mixing in spatial and temporal domains* for action recognition.

V. Multimodal Sequence Learning with Transformers

A. Model Architectures

Key architectures include:

- **Factorized Multimodal Transformers (FMT)**: Efficient learning through factorized self-attention.
- **Meta-Transformers and Perceivers**: Modality-agnostic sequence processing.
- **Multiway Multimodal Attention**: Simultaneous attention across modalities.

B. Sequence Alignment and Fusion

- **Cross-Modal Attention**: Models interactions between modalities.
- **Fusion Strategies**: Early, late, or hybrid fusion approaches.

VI. Applications

A. Video Captioning

Transformers encode modalities and generate coherent text, leveraging cross-modal attention for improved performance.

B. Multimodal Sentiment Analysis

Feature extraction and attention-based fusion, enhanced by semi-supervised pretraining, drive advances in sentiment analysis.

VII. Key Challenges

- **Data Alignment**: Synchronizing modalities with differing temporal and structural properties.
- **Scalability**: Managing computational demands of large-scale multimodal models.
- **Data Scarcity**: Limited availability of labeled multimodal datasets.

VIII. Results

Table I summarizes performance of recent transformer-based multimodal models.

TABLE I
Performance of Transformer-Based Multimodal Models

Model/Year	Domain	Dataset	Metric	Score
DFT (2022)	Sentiment Analysis	CMU-MOSI	Accuracy	81.4%
HCAT (2023)	Video Captioning	MSR-VTT	BLEU-4	38.2
Perceiver-VL (2023)	VQA + AudioCaptioning	VQA + AudioCaps	Acc./BLEU-4	74.1 / 20.3
CSAT (2024)	Classification	MM-IMDb	F1 Score	+8%
STMM (2024)	Action Recognition	Kinetics-600	Top-1 Accuracy	89.6%
Multimodal CL (2025)	Pretraining	Multiple	Avg. Task Gain	+12.3%

A. Discussion

- Dynamic and hierarchical attention models (DFT, HCAT) excel in tasks requiring cross-modal reasoning.
- Unified models like Perceiver-VL generalize well but are computationally intensive.
- Multimodal contrastive learning significantly enhances downstream performance.
- CSAT and STMM demonstrate efficiency in low-resource and real-time scenarios.

IX. Notable Models and Surveys

Table II highlights key models and surveys.

TABLE II
Notable Models and Surveys

Model/Survey	Key Contribution
Factorized Multimodal Transformer	Factorized self-attention
Meta-Transformer	Modality-agnostic processing
Multiway Multimodal Transformer	Multi-modality attention
Multimodal Learning Survey	Theoretical and practical insights

X. Future Directions

- **Unified Multimodal Pretraining**: Scalable pretraining across modalities.
- **Efficient Architectures**: Reducing computational overhead.
- **Advanced Fusion Mechanisms**: Improved cross-modal integration.

Acknowledgment

The authors thank anonymous reviewers for their feedback.

References

[1] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[2] H. Xu et al., "Dynamic Fusion Transformer for Multimodal Sentiment Analysis," arXiv:2203.08562, 2022.

[3] D. Wang et al., "Tetfn: A Text-Enhanced Transformer Fusion Network for Multimodal Sentiment Analysis," *Pattern Recognition*, vol. 136, p. 109259, 2023.

[4] X. Feng, Y. Lin, L. He, and Y. Zhou, "Knowledge-Guided Dynamic Modality Attention Fusion Framework for Multimodal Sentiment Analysis," in *Findings of EMNLP*, pp. 14755–14766, 2024.

[5] L. Yang, "A Dynamic Weighted Fusion Model for Multimodal Sentiment Analysis," *Signal, Image and Video Processing*, vol. 19, no. 609, Springer, 2025.

[6] J. Park et al., "Hierarchical Cross-Attention Transformers for Video Captioning," in *IEEE/CVF ICCV*, 2023.

[7] K. Islam et al., "Video ReCap: Recursive Captioning of Hour-Long Videos," in *IEEE/CVF CVPR*, 2024.

[8] H. Zhou et al., "Perceiver-VL: Unified Transformer for Vision-Language-Audio," in *NeurIPS*, 2023.

[9] L. Wang et al., "Clustered Self-Attention Transformers for Multimodal Classification," in *ACL*, 2024.

[10] A. Rahman et al., "Spatio-Temporal Modality Mixer for Action Recognition," in *IEEE/CVF CVPR*, 2024.

[11] Y. Li and R. Singh, "Review of Multimodal Transformer Pretraining Strategies," in *IJCAI*, 2025.

[12] H. Lee, R. Singh, and S. Ong, "Dynamic Multimodal Sentiment Analysis: Leveraging Cross-Modal Attention," arXiv:2501.08085, 2025.

[13] M. Chen et al., "AudioCap-T: Transformer-based Audio Captioning with Cross-Modal Pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[14] S. Gupta and P. Nagar, "Vision-Audio-Text Transformer for Multimodal Event Detection," in *IEEE ICASSP*, 2023.

[15] F. Zhao et al., "Multi-Modal Sentiment Fusion Transformer: A Tensor Decomposition Approach," *ACM Transactions on Multimedia Computing*, 2024.

[16] R. Kim et al., "Adaptive Multiway Attention Transformer for Video-Language Tasks," in *ECCV*, 2022.

[17] A. Gomez et al., "Sparse Cross-Modal Transformer for Efficient Image-Text Retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[18] J. Patel, M. Rao, and A. Bhatia, "Semi-Supervised Multimodal Transformers for Low-Resource Classification," in *IEEE Transactions on Neural Networks*, 2024.

[19] T. Nguyen and L. Xu, "Hierarchical Alignment Transformer for Audio-Visual Speech Recognition," in *ICASSP*, 2024.

[20] C. Singh et al., "Zero-Shot Multimodal Understanding via Contrastive Pretraining," in *NeurIPS*, 2025.