DeepMLF: Multimodal language model with learnable tokens for deep fusion in sentiment analysis

Efthymios Georgiou, *Graduate Student Member, IEEE,* Vassilis Katsouros, *Member, IEEE,*
Yannis Avrithis, *Senior Member, IEEE,* and Alexandros Potamianos, *Fellow, IEEE*

Abstract—While multimodal fusion has been extensively studied in Multimodal Sentiment Analysis (MSA), the role of fusion depth and multimodal capacity allocation remains underexplored. In this work, we position fusion depth, scalability, and dedicated multimodal capacity as primary factors for effective fusion. We introduce DeepMLF, a novel multimodal language model (LM) with learnable tokens tailored toward deep fusion. DeepMLF leverages an audiovisual encoder and a pretrained decoder LM augmented with multimodal information across its layers. We append learnable tokens to the LM that: 1) capture modality interactions in a controlled fashion and 2) preserve independent information flow for each modality. These fusion tokens gather linguistic information via causal self-attention in LM Blocks and integrate with audiovisual information through cross-attention MM Blocks. Serving as dedicated multimodal capacity, this design enables progressive fusion across multiple layers, providing depth in the fusion process. Our training recipe combines modality-specific losses and language modelling loss, with the decoder LM tasked to predict ground truth polarity. Across three MSA benchmarks with varying dataset characteristics, DeepMLF achieves state-of-the-art performance. Our results confirm that deeper fusion leads to better performance, with optimal fusion depths (5-7) exceeding those of existing approaches. Additionally, our analysis on the number of fusion tokens reveals that small token sets ($\sim$20) achieve optimal performance. We examine the importance of representation learning order (fusion curriculum) through audiovisual encoder initialization experiments. Our ablation studies demonstrate the superiority of the proposed fusion design and gating while providing a holistic examination of DeepMLF's scalability to LLMs, and the impact of each training objective and embedding regularization.

Index Terms—Multimodal Learning, Deep Fusion, Learnable Tokens, Multimodal LM, LLMs, Multimodal Sentiment Analysis (MSA)

✦

1 INTRODUCTION

HUMANS perceive and combine information from different sources and senses to understand and interact with their surroundings. Multimodal signals and representations are also utilized by the human brain when learning concepts. We can therefore claim that multimodality spans the entire human cognitive process. Multimodal Machine Learning (MML) investigates how to develop systems or agents that can process and integrate heterogeneous and interconnected types of data, such as visual, auditory, and textual information. The aim of this field involves the design of systems that understand, reason, and learn from the world through multiple sensory modalities, *e.g.*, verbal and non-verbal communication and scene understanding.

From recognizing emotions through speech and language to generating images from text, the fundamental operation is *multimodal fusion* [1]. Technically, fusion is the problem of learning representations that capture both unimodal information and cross-modal interactions between elements of different modalities. Conceptually, more homogeneous modalities are easier to combine compared to more heterogeneous. Fusion techniques can be broadly categorized into early, late, hybrid and deep fusion methods. Early fusion combines data at earlier stages, late fusion at the final stages, and hybrid fusion combines these schemes. Deep fusion typically involves multiple fusion stages within the architecture.

Recent works in the MML field employ deep fusion schemes to leverage the benefits of multimodality. From the self-supervised approaches of ViLBERT [2] and UNITER [3] to multimodal large language model (LLM) based approaches [4], [5], fusion is performed across several layers, *e.g.*, 24 for UNITER. However, for purely supervised multimodal tasks, such as affective understanding of human-centered video clips, the fusion mechanisms utilized are rather shallow. In particular, they usually involve combining pretrained architectures with shallow fusion mechanisms.

The focus of this work is on *Multimodal Sentiment Analysis* (MSA). MSA involves understanding sentiments by interpreting behavioral signals such as speech, language, facial expressions, and body language [6]. Despite advancements in this area [7], [8], [9], the development of an architecture that employs a deep fusion scheme remains an open

- *Efthymios Georgiou is with the School of ECE, National Technical University of Athens, Athens, Greece, and the Instituie for Speech and Language Processing, Athena Research Center, Athens, Greece*
  E-mail: efthygeo@mail.ntua.gr
- *Vassilis Katsouros is with the Instituie for Speech and Language Processing, Athena Research Center, Athens, Greece*
  E-mail: vsk@athenarc.gr
- *Yannis Avrithis is with the Institute of Advanced Research on Artificial Intelligence (IARAI), Vienna, Austria*
  E-mail: yannis@avrithis.net
- *Alexandros Potamianos is with the School of ECE, National Technical University of Athens, Athens, Greece*
  E-mail: potam@central.ntua.gr

Manuscript received xxx; revised xxx. (Corresponding author: Efthymios Georgiou)

challenge. The predominant body of research focuses on designing increasingly complex architectures [8], [10], [11] and training recipes [12], [13], [14], while efforts focusing on deep fusion schemes [15], [16], [17] remain restricted to fusion schemes involving three layers at most. Moreover, there is limited understanding of optimal capacity allocation for capturing multimodal interactions. In this work, we explore the questions of *when does depth and scale help multimodal fusion* and *how much dedicated capacity is optimal for processing multimodal information* in the context of MSA.

We propose DeepMLF, a novel MSA fusion scheme that focuses on *deepening the fusion process rather than complicating the fusion architecture*. Our approach utilizes a pretrained language model (LM) and augments it with a small set of learnable fusion tokens appended after the language tokens. These fusion tokens serve two key functions: 1) accumulate multimodal information, and 2) maintain linguistic and non-linguistic information flow through the network by design. At the same time, the textual modality is restricted from directly affecting other modalities, providing an inherent bias against language dominance. The acoustic and visual cues are processed via a multimodal encoder, which injects multimodal information into the fusion tokens of the LM through novel cross-attention blocks. This design choice again allows for the accumulation of multimodal information solely in the fusion tokens while retaining the audiovisual information flow through the multimodal encoder. The proposed fusion process can be repeated across multiple decoder LM layers, providing depth and scalability to our approach. The overall learning objective combines task losses for each individual modality and is coupled with language model regularization via language embedding augmentation and language modelling loss.

What sets DeepMLF apart is the deep and scalable fusion framework that, in synergy with the learnable fusion tokens and the training recipe, achieves enhanced fusion benefits. DeepMLF by design allows for adjustable fusion depth and multimodal fusion token allocation. DeepMLF emerges as a strong multimodal LM deep fusion architecture across MSA benchmark datasets. To the best of our knowledge, we are the first to explore both deep fusion configurations and multimodal capacity allocation in the MSA literature.

We validate the effectiveness of DeepMLF through experiments conducted on three widely used MSA benchmark datasets: MOSI [18], MOSEI [19], and SIMS [20]. These datasets were selected to cover different languages, dominance scenarios, and data availability conditions. We compare our method with reproduced state-of-the-art fusion approaches [21] and provide a comprehensive analysis of the DeepMLF components (fusion depth, fusion tokens, and encoder quality). Our contributions are:

1) We introduce DeepMLF, a novel multimodal language model with learnable tokens for deep fusion. DeepMLF leverages pretrained LMs and augments them using a small set of learnable fusion tokens that progressively integrate multimodal information. Our design naturally promotes deep fusion schemes and allows for adjusting the number of fusion tokens for multimodal information processing.

2) We propose a novel cross-attention fusion mechanism (MM Block) which captures interactions between the fusion tokens and the audiovisual information. Non-linguistic features are first processed by a dedicated encoder and then integrated with language representations at various depths through MM Blocks. This design maintains independent information flow for each modality in the network, allowing multimodal information to accumulate exclusively in the learnable tokens alone.

3) Through extensive experiments on MOSI, MOSEI and SIMS, we demonstrate that DeepMLF achieves state-of-the-art performance across different languages and dataset scenarios. Our analysis reveals that deeper fusion schemes (5-7 layers), consistently outperform shallower approaches, and show that a small number of fusion tokens (8-20) achieves optimal results.

4) We examine the importance of representation learning order through audiovisual encoder initializations, demonstrating the benefits of progressive multimodal representation learning. We also scale DeepMLF up to small (1.7B) billion LLMs and provide comprehensive analysis on the interplay of fusion mechanisms, loss terms, and language embedding augmentation, illustrating the synergistic nature of our framework.

The paper is structured as follows: section 2 covers related work and section 3 provides formulation and technical background for our study. Next, section 4 details the DeepMLF architecture, and section 5 outlines our experimental setup. Our experimental analysis lies in section 6 and starts with a comparison of DeepMLF with reproducible state-of-the-art approaches (subsection 6.1). An analysis on the interplay of performance, fusion depth, the number of fusion tokens, the audiovisual encoder intilization, and the impact of language distribution is provided in subsection 6.2. In subsection 6.3 we present ablation results on various alternatives for the fusion mechanism. We also examine the importance of loss terms in the total objective, the impact of embedding regularization method, and gating mechanism variants. Finally, section 7 highlights method limitations, draws conclusions and discusses future research directions.

2 RELATED WORK

This section provides an overview of the literature, beginning with an exploration of works on multimodal fusion, which is the foundamental MML challenge. We then discuss advancements in the MSA field, which is the core of our experimentation.

2.1 Multimodal Fusion

In this section, we provide an analysis of multimodal fusion and, in particular, offer a dual perspective on fusion granularities. First, we outline the basic fusion mechanisms-operations (microscopic view) that may be employed for learning multimodal representations, and then, we discuss deep fusion schemes (macroscopic view) utilized across various tasks.

Fusion Mechanisms: Simpler fusion methods include addition, multiplication, and concatenation, followed by a projection step. Concatenation, in particular, is a popular choice among real-world supervised multimodal setups [13], [22]. More advanced techniques include tensor products [8] between latent modality representations and higher-order polynomial fusion [23]. Gating mechanisms are also widely used, with variations ranging from cell-like approaches such as GMU [24], to attention mechanisms [25], [26]. Specifically, fusion through attention comes in several flavors. Directional (non-symmetric) attention, such as *cross-attention* [2], [27], [28], [29], fuses information from one modality[1](acting as keys, values) to another (acting as queries). *Causal attention* [4], [30] is also non-symmetric and, processes multimodal input (token) information in an autoregressive manner. *Self-attention* [3], [31], processes multimodal information in a bidirectional manner and is considered a symmetric operation. Combinations of these mechanisms [28], [32] can also form larger fusion modules. Besides attention, other strategies employ graph neural networks (GNNs) to create nodes for each latent modality or entities such as speakers and objects, and fuse them via message passing across the constructed nodes [33], [34], [35].

Deep Fusion Schemes: Moving beyond early, late and hybrid schemes, the predominant paradigm in the deep learning era is deep fusion, *i.e.*, fusion across multiple layers. ViLBERT [2] introduces the multimodal co-attention mechanism, applies it across several transformer layers, and shows that some tasks (or datasets) benefit from shallow (two-layer) while others benefit from deeper (six or more layers) fusion. UNITER [3], uses a single encoder transformer and appends visual and textual inputs. Fusion is performed via (bidirectional) self-attention modules across 24 layers. Frozen [4] feeds a pretrained large language model (LLM) with latent image representations followed by language tokens, and trains an image encoder while keeping the LLM frozen. Mutlimodal fusion is performed across the layers of LLMs via causal attention. BLIP-2 [5] and MiniGPT-4 [36] follow similar, yet more sophisticated and better performing approaches than Frozen. LLaMA-Adapter [37] follows a different method and feeds the causal attention mechanism of the decoder with learnable layer-specific fused prompts, across the last layers of a frozen LLM. From a conditional image generation perspective, Unimo-G [38] utilizes a multimodal transformer and conditions the layers of a diffusion model with multimodal information. For an extended review of multimodal LLMs and multimodal transformers, we refer to [39] and [40], respectively.

In a spirit more similar to our work, Ziegler *et al*. [41] propose an encoder-agnostic fusion and condition GPT-2 [42] layers (12) for multimodal language generation. Flamingo [28] inserts cross-modal information across the layers of LLMs, via gated cross-attention followed by a trainable and randomly initialized feed-forward block. Flamingo also feeds the decoder interleaved vision-language inputs. DeepMLF advances this architectural direction with several key innovations. approach differs from Flamingo in several design choices. We introduce a set of learnable fusion tokens that accumulate multimodal information progressively through the network. This design choice is closer to ideas from adapters [37], perceiver [43], and bottleneck fusion [44]. The appended set of learnable tokens gathers linguistic information via causal self-attention (CSA), while a parallel audiovisual encoder processes the non-linguistic signals. The fusion tokens are then integrated with the audiovisual information via cross-attention mechanisms across LM layers, offering fusion depth while maintaining independent information flow for each modality. Our novel gated cross-attention block improves upon the gated cross-attention mechanism of [28], [45], [46] by 1) initializing the feed-forward block from the corresponding LM block and further tuning it[2], 2) restricting cross-modal attention between non-linguistic and fusion tokens alone, to ensure efficient computation and allow for independent information flow, and 3) implementing sigmoid gating for optimal information flow control, as demonstrated in our ablations.

2.2 Multimodal Sentiment Analysis

MSA research mainly focuses on building better fusion schemes and utilizing diverse learning recipes to enhance representation learning for the task at hand. In particular, TFN [8] employs outer product of unimodal representations to capture cross-modal interactions. Poria *et al*. [47] and Gu *et al*. [48] implement multi-level and hierarchical attention to better contextualize information. DHF [15] is the first deep fusion approach for MSA, and serves as an inspiration for this work. Its simple yet deep fusion design across three language levels demonstrates that fusion depth (macroscopic design) is more crucial for performance than complex fusion mechanisms (microscopic design).

Other types of neural structures employed in MSA include neural memory modules [19], top-down fusion [49], capsule networks [10], and GNNs [50]. Tsai *et al*. [16] utilize transformers, where cross-attention blocks act as early fusion and concatenation serves as late fusion. Rahman *et al*. [51] fine-tune a pre-trained BERT [52] model by incorporating a multimodal shifting layer as early fusion, and Zhang *et al*. [11] use language-guided fusion along with a fused hypermodality. In CENet [53] authors exploit a pre-trained language model tailored towards sentiment analysis instead of BERT.

Another line of work utilizes more complex learning recipes such as canonical correlation analysis [9] and cycle-consistency loss [54] across modalities. Coupling different learning recipes with pre-trained models has been a popular choice among researchers. Yu *et al*. [13] introduce a unimodal pseudo-labeling module that backpropagates three additional losses. Hazarika *et al*. [55] augment the learning objective with feature reconstruction loss as well as attracting and repelling objectives. A two-step hierarchical learning recipe based on mutual information maximization is proposed in [12], while Sun *et al*. [9] propose a meta-learning framework that learns each unimodal network and then adapts them for the MSA task. Sun *et al*. [14] propose a transformer architecture leveraging dual-level reconstruction loss and an attraction loss in a Siamese setup

---

1. The term modality here is used loosely.

2. Full fine-tuning and LoRA adapters are integrated in DeepMLF

between complete and incomplete data. NIAT [56] learns a unified joint representation between clean and noisy data by coupling masking-based feature augmentation with an adversarial training strategy. Hu *et al.* [17] employ a text generation encoder-decoder architecture, using T5 [57], and implement a contrastive loss among unimodal encoders. Fusion is performed across the last 3 layers of the transformer's encoder. The decoder generates text sequences, which are in turn decoded into MSA-related info such as polarity. Notably, none of the aforementioned approaches utilizes more than three fusion layers.
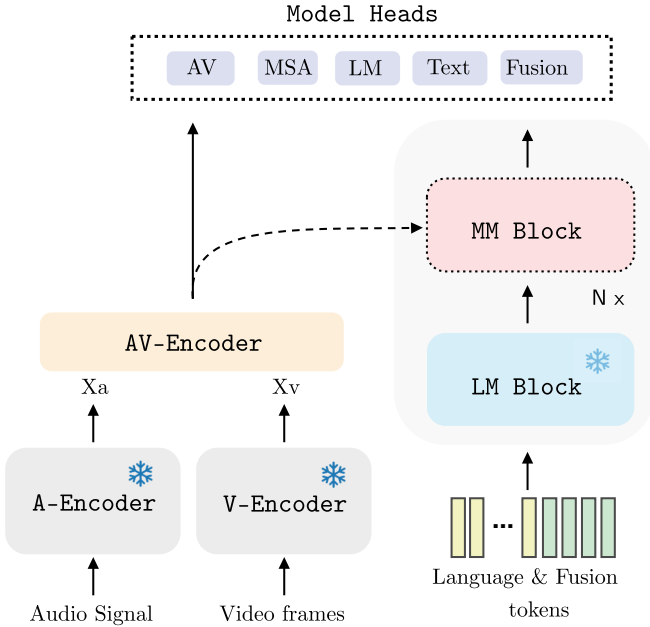
3 BACKGROUND

Fig. 1: *DeepMLF architecture overview.* Audio and Visual features are being processed by a trainable *AV-Encoder* and then fed to the Language Model (LM) for deep multimodal fusion. The LM consists of N layers where the *LM Block* remains frozen and the *MM Block* is trainable. The output of the overall architecture are audiovisual, language and fused tokens which encapsulate audiovisual, linguistic and mutlimodal sentiment information respectively. The *Model Heads* denote the involved objectives in out training recipe.

We first formulate *multimodal sentiment analysis* (MSA) as a multimodal fusion task. We then present the transformer architecture with an abstractive notation suitable for the variants in this paper, and briefly outline the (conditional) language modelling objective.

3.1 Problem formulation and notation

Vectors and matrices are denoted by lowercase and uppercase bold letters respectively, *i.e.*, $\mathbf{x}$ and $\mathbf{X}$. Tensors are represented as $\mathbf{X}$, and sets with calligraphic letters $\mathcal{M}$. Depending on the context subscripts can denote timesteps ($\mathbf{x}_t$) or modalities ($\mathbf{X}_f$). Upperscripts in parenthesis depict different layers ($\mathbf{H}^{(l)}$). MSA is a task which takes as input three

modalities, *i.e.*, language, audio and video, and predicts the sentiment polarity. Each input modality m resides in an input space $\mathcal{X}_m \subseteq \mathbb{R}^{D_m \times L_m}$. Index m denotes the modality from a set (of indices) $\mathcal{M} = \{1, \ldots, M\}$, D_m is the input space dimensionality, and L_m is the per modality (maximum) sequence length. The multimodal input space can be expressed as the cartesian product of unimodal spaces $\mathcal{X}_\mathcal{M} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_M$. Any supervised multimodal task can now be formulated as learning the (neural) mapping parameterized with $\boldsymbol{\theta}$; $f_{\boldsymbol{\theta}} : \mathcal{X}_m \to \mathcal{U}$, where $\mathcal{U} \subset \mathbb{R}$ in the MSA case. Each multimodal input (m-tuple) is represented as the collection of M modalities as $\mathbf{X}_i = [\mathbf{X}_1, \cdots, \mathbf{X}_M]$ along with a scalar label $y_i \in \mathcal{U}$. During the rest of the paper we denote the linguistic, the acoustic and the visual modalities with subscripts t, a, v respectively. The audiovisual tokens are denoted with the av subscript, and the (learnable) fusion tokens with the f subscript.

3.2 Transformer Architecture

Based on the transformer architecture paper [26] we briefly outline its architecture and in particular the pre-norm Encoder-only and Decoder-only [58] design, which are utilized across this paper. Our presentation maintains a level of abstraction so that it can encapsulate transformer variants, and in particular different flavors in the attention mechanism [59], the normalization and the feed-forward components [60].

3.2.1 Encoder Layer

The typical encoder layer design consists of a multihead Self-Attention (SA) module followed by a feed forward (FFW) block [26]. We utilize the pre-norm transformer variant in our experiments. Stacking encoder layers together, forms the Encoder transformer architecture. Formally for an encoder layer l, and a latent input (from the previous layer) $\mathbf{H}^{(l-1)}$ we have

$$\widetilde{\mathbf{H}}^{(l)} = \mathbf{H}^{(l-1)} + \mathrm{SA}(\mathrm{Norm}(\mathbf{H}^{(l-1)})) \tag{1}$$
$$\mathbf{H}^{(l)} = \widetilde{\mathbf{H}}^{(l)} + \mathrm{FFW}(\mathrm{Norm}(\widetilde{\mathbf{H}}^{(l)})) \tag{2}$$

where Norm denotes the normalization layer (LayerNorm, RMSNorm), SA the multihead self-attention layer, $\mathbf{H}^{(l)}$ denotes the output representation of encoder layer l (which is the input of layer $l + 1$). All hidden representations $\mathbf{H}^{(l-1)}, \mathbf{H}^{(l)}$ lie in the same space $\mathbb{R}^{d \times L}$. The AV-Encoder of Figure 1 utilizes a stack of such Encoder Layers.

3.2.2 Decoder Layer

The decoder follows a structure similar to the encoder with one main difference. Self-attention becomes causal self-attention (CSA), *i.e.*, each position can only attend to its previous positions. Formally, given a hidden input representation $\mathbf{H}^{(l-1)}$ to the l-th decoder layer the decoder layer equations are:

$$\widetilde{\mathbf{H}}^{(l)} = \mathbf{H}^{(l-1)} + \mathrm{CSA}(\mathrm{Norm}(\mathbf{H}^{(l-1)})) \tag{3}$$
$$\mathbf{H}^{(l)} = \widetilde{\mathbf{H}}^{(l)} + \mathrm{FFW}(\mathrm{Norm}(\widetilde{\mathbf{H}}^{(l)})) \tag{4}$$

where CSA is the causal (masked) self-attention. The pre-trained decoder LMs utilize a stack of such Decoder Layers, illustrated as LM Block in Figure 1.

3.3 Language Modelling Objective

Autoregressive or causal neural language models, generate text sequentially, predicting one token at a time, conditioned on previously generated/input tokens. These LMs are trained based on minimizing the negative log-likelihood of the conditional probability p_{LM} over a set of (previous) tokens $\mathbf{x}_{<t}$ and some context information $\mathbf{Z}$. Formally expressed:

$$L_{LM} = \sum_{t=1}^{L} -\log p_{LM}(\mathbf{x}_t|\mathbf{x}_{<t}, \mathbf{Z}) \tag{5}$$

During this work the context information $\mathbf{Z}$ denotes audiovisual (multimodal) information from the AV-Encoder (see Figure 1).

4 DEEPMLF

In this section, we describe the proposed multimodal language model framework. First, we introduce the novel multimodal language model architecture and describe its components, *i.e.*, *AV Encoder*, *MLM*, and *Model Heads*. Then we describe the training recipe, and finally include a discussion section on the components of DeepMLF.

4.1 Overview

DeepMLF's main architectural components are the novel *Multimodal LM (MLM)* and the *AV-Encoder*. The *MLM* consists of chained *LM Block* and *MM Block* modules, while *AV-Encoder* is a standard encoder transformer architecture. The overall architecture is illustrated in Figure 1.

The information flow through DeepMLF is described below. The *A-Encoder* and *V-Encoder* handle the feature extraction process (handcrafted or neural) for the acoustic and visual modalities, producing $\mathbf{X}_a$ and $\mathbf{X}_v$ features respectively. These extracted features are fed to *AV-Encoder* which supplies the *MLM*[3] with mutlimodal (audiovisual) information. In the MLM architecture, we append learnable fusion tokens $\mathbf{X}_f$ after the language embeddings $\mathbf{X}_t$ (Figure 1 assumes tokenization is already performed) and feed them to the MLM. The pretrained *LM Block* processes the concatenated input and accumulates linguistic information at the learnable fusion tokens. After the (frozen) LM block, the fusion tokens alone interact with the audiovisual tokens within the *MM Block*, and capture multimodal information. This process can be repeated for each decoder layer, rendering the proposed approach as a deep fusion scheme. Moreover, the audiovisual, linguistic, and fused components maintain their information flow through the network, and are fused in the *Task Head* for the final prediction.

4.2 AV Encoder

The *AV Encoder* fuses acoustic and visual information before feeding it in the MLM decoder. We utilize two separate modality-specific transformer encoders to process unimodal information, and then fuse their representations through a feedforward network (Fusion FFW). The modality-specific

---

3. should not be confused with MLM which is the masked language modelling, *e.g.*, BERT-like approach [52]

encoders process the (projected) acoustic and visual features, and output $\mathbf{Z}_a$ and $\mathbf{Z}_v$ respectively. These representations are concatenated (along the dimension-axis) and processed by the Fusion FFW, which outputs the fused $\mathbf{Z}$. This process is summarized as:

$$\mathbf{Z} = \mathcal{E}_{AV}(\mathbf{X}_a, \mathbf{X}_v) \in \mathbb{R}^{d_{av} \times L} \tag{6}$$

where $\mathbf{Z} = \mathbf{Z}_{av}$ is the multimodal information which is fed to the MLM through the AV Encoder. The AV Encoder is trained in isolation, and its weights are used to initialize the AV Encoder of DeepMLF (see Figure 1).

4.3 Multimodal Language Model (MLM)

Here we describe the proposed Multimodal LM (MLM) architecture and focus on the learnable fusion tokens, the preatrained decoder *LM Block*, and the novel *MM Block* with its gated cross-attention (GCA) and feed-forward projection (FFW).

4.3.1 Learnable Fusion Tokens

We append n_f learnable fusion tokens $\mathbf{X}_f \in \mathbb{R}^{n_f \times d_t}$ to the pretrained LM input. For any language input $\mathbf{X}_t \in \mathbb{R}^{L_t \times d_t}$ (after tokenization and embedding layer) we have

$$\mathbf{H}^{(0)} = [\mathbf{X}_t^{(0)} || \mathbf{X}_f^{(0)}] \tag{7}$$

where $[\cdot || \cdot]$ denotes concatenation, and $\mathbf{H}^{(0)}$ is the input to the first LM Block. This set of fusion tokens appended after the language tokens serves two core functions: 1) gathers the linguistic information, and 2) interacts with the audiovisual information in the MM Block. The hyperparameter n_f is analyzed in our experiments, and its optimal value is typically small, *i.e.*, 12.

4.3.2 LM Block

Transformer based language modelling typically consists of stacked transformer decoder (see subsubsection 3.2.2) layers. Our approach differs in two ways compared to standard decoder LM: 1) LM Blocks process both fusion tokens and language tokens in every layer, and 2) all LM layers are kept frozen to minimize the number of trainable parameters and avoid catastrophic forgetting [28], [41], [46], [61]. Layer l output goes to either the MM Block (when present) or the next LM Block ($l + 1$).

4.3.3 MM Block

The proposed MM Block performs the core fusion operation through gated (multimodal) cross-attention (GCA), followed by a feed-forward projection (FFW). MM Block takes the previous LM Block output $\widehat{\mathbf{H}}^{(l)} = [\widehat{\mathbf{X}}_t^{(l)} || \widehat{\mathbf{X}}_f^{(l)}]$, and the audiovisual context information $\mathbf{Z}$ as input. We split language and fusion tokens, and feed only fusion tokens to the GCA layer to capture multimodal interactions:

$$\widehat{\mathbf{X}}_t^{(l)}, \widehat{\mathbf{X}}_f^{(l)} = \text{split}(\widehat{\mathbf{H}}^{(l)}) \tag{8}$$

$$\overline{\mathbf{X}}_f^{(l)} = \widehat{\mathbf{X}}_f^{(l)} + \sigma(a_1^{(l)}) \odot \text{GCA}((\mathbf{Z}), \text{Norm}(\widehat{\mathbf{X}}_f^{(l)})) \tag{9}$$

The GCA mechanism utilizes audiovisual information $\mathbf{z}$ as keys/values, and fusion tokens as queries. The fused tokens

are then concatenated back with the language representation $\widehat{\mathbf{X}}_t^{(l)}$ and processed via the FFW layer as:

$$\overline{\mathbf{H}}^{(l)} = [\widehat{\mathbf{X}}_t^{(l)} || \overline{\mathbf{X}}_f^{(l)}] \tag{10}$$

$$\mathbf{H}^{(l)} = \overline{\mathbf{H}}^{(l)} + \sigma(a_2^{(l)}) \odot \text{FFW}(\text{Norm}(\overline{\mathbf{H}}^{(l)})) \tag{11}$$

Output $\mathbf{H}^{(l)}$ is fed to the next LM Block. Here, $\sigma(\cdot)$ denotes the sigmoid gating function with learnable per-layer parameters $a_1^{(l)}, a_2^{(l)}$. All hidden representations $\mathbf{H}, \widehat{\mathbf{H}}, \overline{\mathbf{H}}$ lie in $\mathbb{R}^{d \times L}$, with $L = L_t + n_f$.

The proposed scheme performs best when FFW is initialized with parameters from its corresponding LM Block FFW layer. We also experimented with other gating schemes, but found sigmoid most effective.
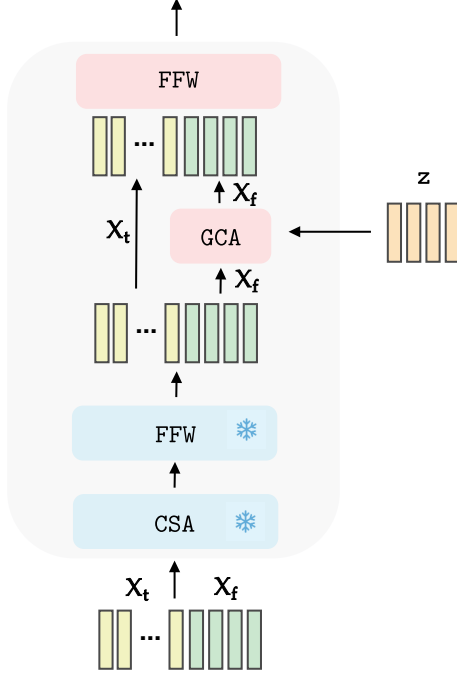

Fig. 2: *MM Block*. The fusion tokens (green) are appended at the LM input. First they accumulate linguistic information though the frozen LM Blocks (CSA and FFW). Then the fusion tokens ($\mathbf{X}_f$) are fed to the GCA module where they are fused with audiovisual information ($\mathbf{z}$) and together with the language tokens ($\mathbf{X}_t$) are fed to the FFW module of the MM Block. This modular design allows for integration of multiple MM Blocks across LM layers, enabling deep fusion capabilities.

4.4 Model Heads

Task Head: A two layer MLP, $g(\cdot)$, operates as the task head, and maps the learned representations to the task space, *i.e.*, sentiment polarity. In particular, it accepts the pooled audiovisual representation $\langle \mathbf{z} \rangle$, final layer's last (assume position K) language embedding $\mathbf{X}_t^{(N)}[K] = \mathbf{x}_{t,K}$, and the mean fusion representation $\langle \mathbf{x}_f \rangle = \langle \mathbf{X}_f^{(N)} \rangle$:

$$y_o = g([\langle \mathbf{z} \rangle || \mathbf{x}_{t,K} || \langle \mathbf{x}_f \rangle]) \tag{12}$$

The mapping $g : \mathbb{R}^{2d+d_{av}} \to \mathcal{U}$ is the late fusion operation.

Modality Heads: Linear mappings W_{av}, W_t, W_f process audiovisual $\langle \mathbf{z} \rangle$, textual $\mathbf{x}_{t,K}$, and fused $\langle \mathbf{x}_f \rangle$ representations for auxiliary task losses, producing sentiment polarity values y_{av}, y_t, y_f respectively.

Multimodal LM Head: A linear layer W_{LM} from the LM hidden space to the LM's vocabulary is adopted as in common language modelling. This layer is transfered from the pretrained LM (GPT2, SmolLM2) into our architecture and kept frozen during DeepMLF's optimization process.

4.5 Training Recipe

This section presents the loss objectives and the regularization technique employed on language embeddings. We train DeepMLF using $L_1(y, \hat{y}) = L_{\text{MAE}}(y, \hat{y}) = ||y - \hat{y}||$ modality specific loss terms, along with a LM loss.

Task Loss: The primary loss of our approach is $L_{msa} = L_1(y, y_o)$, based on the Task Head output.

Auxiliary Loss: For each modality, we also predict the ground truth multimodal polarity as an auxiliary loss term:

$$L_{\text{aux}} = \lambda_{\text{av}} L_1(y, y_{\text{av}}) + \lambda_t L_1(y, y_t) + \lambda_f L_1(y, y_f) \tag{13}$$

where $\lambda_{\text{av}}, \lambda_t, \lambda_f$ are weighting coefficients.

Multimodal LM Loss: Following benefits of LM loss in downstream tasks [61] to avoid overfitting and preserve language regularities, we employ (multimodal) LM loss L_{LM} as in Eq. (5). In DeepMLF's case the conditioning information $\mathbf{c}$ is the AV Encoder output $\mathbf{z}$.

Total Loss: The total training objective is the combined task, auxiliary and reweighted multimodal LM loss.

$$L_{\text{tot}} = L_{msa} + L_{\text{aux}} + \lambda_{\text{LM}} L_{\text{LM}} \tag{14}$$

Language Embedding Regularization: We apply regularization to the pretrained LM language embeddings, which aims at enhancing robustness against overfitting and language dominance [62], [63]. After tokenization, we employ SeqAug [64] on the extracted language embeddings[4].

4.6 Implementation Details

In this section we discuss the implementation details as well as key concepts of DeepMLF.

AV Initialization: We first pretrain the AV encoder separately, then fine-tune it when integrated with the LLM backbone. This approach aligns with other Multimodal LLM works that employ visual or acoustic backbones, and either fine-tune their encoder or use trainable connectors like Perceiver [28], [43] or adapters [65]. For simplicity, we chose to fine-tune our AV encoder.

Learnable Fusion Tokens: Our approach draws inspiration from methods like bottleneck fusion [44], but implements the concept differently. We append learnable tokens to the LM input and process them through two distinct steps: first using causal attention to gather linguistic information, then through cross-attention where they interact with multimodal data. This differs from bottleneck fusion's

4. [5]

single step self-attention-only modality interaction. Therefore, DeepMLF effectively extends the concept across both causal (decoder) and cross-attention layers.

MM Block: Our Multimodal interaction block combines two components: GCA and FFW. The GCA shares similarities with Flamingo, but differs in two key ways. First, we feed only learnable fusion tokens to cross-attention for more controlled fusion, while Flamingo uses the complete interleaved multimodal (language and vision) input. This reduces memory cost to $\mathcal{O}(n_f^2)$ compared to Flamingo's $\mathcal{O}((L_t + L_v)^2)$. Second, we use sigmoid gating instead of tanh, which proved more effective in our experiments. For the FFW component, we initialize it using parameters from its corresponding FFW LM Block. This initialization approach enhances performance without introducing training instabilities [28]. We also integrate LoRA in MM Block FFW layer, to reduce training parameters and be robust against overfitting. This helps especially when using LLM backbones like SmolLM2.

Auxiliary Task Losses: DeepMLF maintains independent information flow for the audiovisual, language and fusion modalities by design. We add an auxiliary task loss for each one of these modalities. This design choice follows evidence that better individual modality representations help better capture cross-modal interactions [66] and improve representation predictability.

Regularization: To be more robust against overfitting and language dominance, we employ two regularization methods: multimodal LM loss and language embedding augmentation (via SeqAug). These techniques prove essential for typical language distributions, where the multimodal LM loss helps preserve language regularities [61] in the training data while preventing catastrophic forgetting, and SeqAug effectively samples from the underlying input language feature distribution [64]. The effectiveness of these regularizers depends on the input data characteristics. Our experiments verify that for standard[6] language distributions, regularization is vital. However, for distributions significantly different from the pretraining corpus, regularization may not provide benefits. Therefore, DeepMLF integrates both fusion and regularization components in its learning approach. Previous work in multimodal learning and MSA has either studied modality imbalance [67], [68], [69], [70], regularization [63], [71], [72], [73], or robustness [74], [75]. We are the first to propose an integrated multimodal fusion framework with regularization by design.

5 EXPERIMENTAL SETUP

We evaluate DeepMLF over three benchmark datasets for MSA to cover different languages, data abundance, and modality imbalance scenarios.

5.1 Benchmark datasets

MOSEI: The CMU-MOSEI [19] dataset, is the largest MSA benchmark containing 66h of multimodal content. MOSEI offers a diverse range of samples containing $23,453$

6. in the sense that they could be similar to the pretraining corpus

manually transcribed and annotated utterance-level video segments from 1000 distinct speakers, and covers 250 topics. The average segment length is 7.28 sec, with segmentation based on punctuation from the *high-quality manual transcriptions*. Each segment is manually annotated in a Likert scale from -3 (strongly negative) to +3 (strongly positive). In our experiments, we observed a 21.98% relative performance gap between text and the other modalities, with audio and visual features showing nearly identical performance levels.

MOSI: CMU-MOSI [18] dataset contains approximately 2.5h of YouTube videos (2-5 minute clips), consisting of 2,199 utterance-level movie review opinion segments from 93 videos and 89 different speakers (41 female, 48 male). Each segment averages 4.2 seconds and includes written transcripts and human annotator sentiment ratings on a likert scale from -3 to +3.

Compared to MOSEI, MOSI has key differences: it uses far fewer speakers (only about 10%), covers a narrower range of topics (movie reviews), employs a smaller vocabulary size, and features more informal language (see also Table 9. Notably, MOSI exhibits the largest performance gap between text and audio (second best performing modality) features (33.94%) among our experiments.

SIMS: CH-SIMS [20] is a Chinese MSA becnhmark, with size comparable to MOSI, containing 2.3h of 60 high-quality videos, spanning movies, TV series, and variety shows. Researchers manually segmented the collected videos into $2,281$ utterance-level monologue video segments, averaging 3.67 sec each. Human annotators transcribed the content, and assigned sentiment polarity scores ranging from -1 (strongly negative) to +1 (strongly positive). SIMS is the most balanced MSA benchmark, showing relative modality performance gap 3.72% in our experiments.

5.2 Multimodal features

Processing raw multimodal content presents significant challenges, including high computational costs and potential copyright restrictions. Instead, researchers commonly use pre-extracted features that offer key advantages: they reduce the heterogeneity gap between modalities (language, audio, video), and leverage information embedded into hand-crafted or neural representations. However, feature extraction pipelines in MSA vary across methods [21], [76], making comparisons across methods difficult. For consistency, we use the feature sets from Mao et.al. [21] across all methods and datasets.

Text modality: DeepMLF is compatible with any decoder LLM model. For MOSEI we utilize the english pretrained `GPT2-large` model and its corresponding `GPT2-base` for the chinese[7] SIMS dataset. For MOSI we utilize a LLM, *i.e.*, SmolLM2-1.7B[8] to verify the scalability and compatibility of DeepMLF. For other competitors and baseline models we follow Mao *et al*. [21] and use BERT [52] embeddings. In particular, we use `bert-base-uncased` for English and `bert-base-chinese` for the Chinese language.

7. https://huggingface.co/uer/gpt2-chinese-cluecorpussmall
8. https://huggingface.co/HuggingFaceTB/SmolLM2-1.7B

Acoustic modality: Acoustic analysis typically relies on extracted sound features. For MOSI and MOSEI datasets, researchers use COVAREP [77] to extract 74 acoustic properties per frame, including pitch and 12 MFCCs. For SIMS, we use Librosa [78] to generate 33 acoustic features per frame.

Video modality: For video analysis, MSA tasks include facial landmarks, eye gaze, and facial action units. MOSI and MOSEI use Facet[9] to extract 35 facial action units linked to emotions and sentiment polarity. SIMS employs Open-Face2.0 [79] to capture 709 features per frame, including 68 facial landmarks and 17 facial action units.

5.3 Evaluation metrics

We evaluate MSA as a regression task using *mean absolute error* (MAE) and *Pearson correlation* (Corr), following standard MSA practices [21], [55], [80]. We also map continuous sentiment predictions into discrete categories and measure classification accuracy (Acc-k). Our evaluation includes binary metrics (Acc-2 and $F1$), as well as 3-class, 5-class and 7-class accuracies, depending on the benchmark requirements.

5.4 Competitors and MSA models

We evaluate our approach against leading MSA models using the M-SENA [21] framework for fair comparison. Our experiments include five state-of-the-art architectures: MulT [16], MISA [55], Self-MM [13], ALMT [11], and TETFN [81]. These models have demonstrated strong performance across MSA datasets and provide a holistic performance overview. We also report other models performance from the literature.

LF-DNN: The *late fusion deep neural network* (LF-DNN) [82] processes each modality separately through neural networks before combining them for final prediction.

TFN: The *tensor fusion network* (TFN) [8] uses LSTM for text processing while averaging acoustic and visual features. It fuses these processed features via Kroenecker product.

MAG-BERT: MAG-BERT [51] enhances BERT with a multimodal adaptation gate to fuse information from audio and visual modalities.

MulT: The *multimodal transformer* (MulT) [16] combines information across modalities through *cross-attention* (CA) blocks. It then processes these fused representations using *self-attention* (SA) mechanisms before concatenating for final prediction.

MISA: MISA [55] processes audio and video using LSTM networks and fine-tunes BERT for text analysis. It embeds modalities into shared and modality-specific spaces to capture mutual information while preserving unique features. MISA fuses these representations in two ways: one branch reconstructs the input, while another uses *self-attention* (SA) for the final multimodal prediction.

9. https://imotions.com/platform

TABLE 1: *DeepMLF configurations across datasets*. (L)LM: language backbone; MM Blocks: the LM layers after which we insert multimodal blocks; MM Layers: the total number of MM Blocks; n_f: the number of trainable fusion tokens; λ_*: loss weights; FFW-FT: fine-tuning method for the FFW layer of each MM Block

	MOSEI	MOSI	SIMS
(L)LM	GPT2-large	SmolLM2-1.7B	GPT2-base
MM Blocks	8-15-22-29-36	12-15-18-21-24	6:12
MM Layers	5	5	7
n_f	12	8	16
λ_f	1.0	0.4	1.0
λ_{av}	1.0	0.8	1.0
λ_t	1.0	0.8	1.0
λ_{LM}	1.0	0.0	1.0
FFW-FT	Full	LoRA	Full

Self-MM: Self-MM [13] uses LSTM networks for audio and visual processing and fine-tunes BERT for text. It utilizes *unimodal label generation module* (ULGM) that creates individual modality labels from multimodal labels and embeddings. For prediction, Self-MM concatenates modality representations and processes them through dual linear layers. The model combines a main task loss with modality-specific losses from pseudolabeling.

ALMT: ALMT [11] processes all modalities using transformers and fine-tunes BERT for text. It introduces an Adaptive Hyper-modality module that applies self-attention to language and cross-attention between text and other modalities. The model guides fused information via language-based cross-attention blocks.

TETFN: TETFN [81] applies text-based multi-head attention to enhance non-linguistic features with textual information. The model uses Vision-Transformer (ViT) for visual processing and combines cross-modal mappings with unimodal label prediction. All modalities are first encoded individually, then paired using text-oriented attention before final sentiment prediction.

5.5 Implementation details

We implement DeepMLF in PyTorch [83], using AdamW [84] optimizer with β_1=0.9, β_2=0.95, batch size 32, warmup for one epoch, cosine annealing, and validation loss early stopping. For non-linguistic modalities (audio and video), we use the features provided in the M-SENA [21] framework. For language, MOSEI uses `GPT2-large`, SIMS the chinese `GPT2-base` and MOSI `SmolLM2-1.7B` LLM, with the latter employing LoRA [85] adaptation (r= 512) in the MM Block's FFW layer. MM Blocks are inserted every k layers backward from the final layer until performance plateaus. We tune n_f in the range $\{8, 12, 16, 20\}$, set loss weights (auxiliary and MLM) to 1.0 for MOSEI and SIMS, while tuning them in $\{0.0, \cdots, 1.0\}$ for MOSI, and adjust learning rates around 10^{-4}. For a detailed configuration we refer to Table 1. All baselines are reproduced using M-SENA and evaluation results are averaged over at least 5 independent runs. All experiments conducted in a single NVIDIA RTX 3090 (22GB).

TABLE 2: *Unimodal feature comparisons*. For each modality we evaluate the performance of pretrained models equipped with a trainable classification head. *Native* denotes the features provided in the original papers and in [21]. For the language models we use different model weights for the english and the chinese languages. ↑ / ↓: higher/lower is better. Red: worse than the baseline; bold: best for each MSA model.

UNI. FEATURES	MOSI		MOSEI		SIMS	
	Acc2↑	Acc7↑	Acc2↑	Acc7↑	Acc2↑	Acc5↑
LANGUAGE						
BERT	**78.96**	**31.83**	**83.39**	**50.31**	**77.35**	**37.71**
GPT2	66.85	24.30	80.76	48.71	76.66	34.61
GPT2-large	73.45	29.88	82.50	49.82	-	-
SmolLM2-1.7B*	75.96	30.98	-	-	-	-
AUDIO	52.16	16.45	65.09	41.36	66.16	23.42
VISION	43.09	15.5	64.41	41.88	74.47	25.46

6 EXPERIMENTAL RESULTS

We evaluate and analyze DeepMLF across multiple dimensions. First, we compare DeepMLF against state-of-the-art reproduced algorithms from the literature. Our analysis then extends to algorithmic dimensions such as fusion depth, number of learnable fusion tokens, encoder initialization, language distribution characteristics, and scaling properties. The evaluation concludes with ablation experiments over components, such as loss terms, regularization strategies, and gating mechanism variants.

6.1 Comparison with the state of the art

Our comparative analysis builds upon results of state-of-the-art models reproduced in the M-SENA framework [21], utilizing their publicly available code and standardized feature sets. The performance results we obtained through reproduction align with those documented in M-SENA[10].

MOSEI: We train DeepMLF with GPT2 backbones of various sizes on MOSEI. As illustrated in Table 4, utilizing the `GPT2-base` model as language backbone, achieves state-of-the-art performance across all metrics examined. Notably Table 2 illustrates that despite generative LMs underperform embedding models as BERT, DeepMLF outperforms all BERT-based competitors[11], highlighting its efficacy as a fusion method. Since MOSEI is the larger MSA dataset, we scale DeepMLF to `GPT2-medium` and `GPT2-large`, further improving the multimodal performance and pushing the limits of state-of-the-art for MOSEI as illustrated in Table 3 and Table 4. Overall, we get large improvements of 1.92% for Acc-2, 2.2% for Acc-5, and 2% for Acc-7, over the previous state-of-the-art models. We further discuss the role of scale and depth in subsection 6.2.

MOSI: For MOSI, GPT2 backbone performance significantly lags behind embedding models such as BERT (see Table 2). We therefore integrate SmolLM2-1.7B LLM

10. https://github.com/thuiar/MMSA/blob/master/results/result-stat.md

11. This result align with literature findings [76], where authors scale up to 13B LLMs to match smaller embedding model performance.

into DeepMLF, to get language-only performance closer to BERT. Despite the performance gap between our language backbone and BERT[12], DeepMLF achieves state-of-the-art multimodal performance on MOSI as illustrated in Table 3. This result highlights that DeepMLF is an efficient multimodal fusion method, since 1) it achieves a significant fusion improvement over the text-only performance, of 12.63% relative Acc-2 improvement, compared to 6.65% relative improvement of Self-MM and TETFN, and 2) it is capable of operating fusion in scenarios with large modality performance gap (33.94%). We present a detailed analysis on the MOSI case in subsection 6.2.

SIMS: For SIMS, we utilize the Chinese GPT2-base model as the language backbone for DeepMLF, though its performance is inferior to BERT (Table 2). Nevertheless, we achieve state-of-the-art results with significant relative improvements of 15.75% MAE, 22.52% Corr, and 3.38% Acc-2 over previous state-of-the-art multimodal approaches. This performance improvement demonstrates that DeepMLF also achieves better fusion in balanced scenarios with narrower modality performance gaps.

6.2 Analysis

This section offers a detailed algorithmic analysis of DeepMLF by studying factors such as fusion depth, number of learnable fusion tokens, encoder initialization, LM size and language data distributions.

6.2.1 The role of fusion depth and size

We conduct experiments on both SIMS and MOSEI in this section. SIMS based experiments cover different fusion depth configurations for `GPT2-base` LM backbone, while MOSEI investigates the role of scaling the backbone itself, *i.e.*, from `GPT2-base` to `GPT2-large`.

Multimodal Fusion Depth: Table 6 illustrates three different fusion setups for the SIMS dataset. The results clearly highlight the existence of an optimal fusion depth configuration, specifically seven MM Blocks in our case. This depth exceeds typical fusion depths from the literature which are primarily limited to three layers. Moreover, we observe that increasing depth beyond this optimal value slightly decreases performance, while decreasing depth significantly harms our results. This finding clearly demonstrates that adequate depth is necessary for efficient fusion.

Across all datasets and backbones, we observe that *it is better to skip adding MM Blocks at the shallow LM layers*. We hypothesize that low-level linguistic features learned in these layers, do not effectively integrate with non-linguistic (audio-visual) signals.

Fusion Scheme Size vs Depth: Table 7 illustrates different fusion configurations for the MOSEI dataset, utilizing three progressively larger MM Block variants (and consequently larger GPT2 backbones). We observe a linear decrease (7, 6, 5) in the optimal fusion depth as the size of the MM Blocks increases. Our optimal configuration employs a fusion scheme of five layers, maintaining deeper

12. We need to scale LLM further to get comparable performance [76]

TABLE 3: *State of the art comparisons for MOSI and MOSEI.* †: results reported in [21]; *: results reproduced; ↑ / ↓: higher/lower is better. Blue: first and second best performing score per metric; bold: best for each MSA model.

Model	MOSI						MOSEI					
	Acc2↑	F1↑	MAE↓	Corr↑	Acc5↑	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑	Acc5↑	Acc7↑
LF-DNN†	79.39	79.45	0.945	0.675	-	-	82.78	82.38	0.558	0.731	-	-
TFN†	78.02	78.09	0.971	0.652	-	-	82.23	81.47	0.573	0.718	-	-
MAG-BERT†	83.41	83.47	0.761	0.772	-	-	84.87	84.85	0.539	0.764	-	-
MulT*	80.26	80.32	0.927	0.689	40.10	34.71	84.07	83.93	0.564	0.731	53.97	52.56
MISA*	82.93	82.95	0.772	0.774	47.55	42.10	84.51	84.47	0.549	0.759	53.57	51.96
TETFN*	84.10	84.14	0.725	0.790	52.77	45.92	85.20	85.18	0.544	0.759	55.54	53.74
Self-MM*	84.22	84.23	0.724	0.791	52.22	45.64	84.26	84.24	0.532	0.765	55.52	53.85
ALMT*	83.90	83.89	0.746	0.784	48.77	43.58	85.23	85.32	0.539	0.766	54.64	53.05
DeepMLF	85.60	85.58	0.692	0.811	53.18	46.27	87.15	87.10	0.499	0.804	57.70	55.88

TABLE 4: *MOSEI DeepMLF performance for different sizes of GPT2 LM backbones.*

DeepMLF	Acc2↑	F1↑	MAE↓	Corr↑	Acc5↑	Acc7↑
base	85.83	85.74	0.529	0.773	56.61	54.65
med	87.00	86.98	0.511	0.795	56.89	55.01
large	87.15	87.10	0.499	0.804	57.70	55.88

TABLE 5: *State of the art comparisons for SIMS.* †: results reported in [21]; *: results reproduced; ↑ / ↓: higher/lower is better. Blue: first and second best performing score per metric; bold: best for each MSA model.

Model	Acc2↑	F1↑	MAE↓	Corr↑
LF-DNN †	76.68	76.48	0.446	0.567
TFN†	77.07	76.94	0.437	0.582
MAG-BERT†	74.44	71.75	0.492	0.399
MulT*	78.56	78.66	0.453	0.564
MISA*	76.54	76.59	0.447	0.563
TETFN*	79.21	79.05	0.419	0.592
Self-MM*	80.04	80.44	0.425	0.595
ALMT*	78.16	78.16	0.433	0.575
DeepMLF	82.75	83.15	0.353	0.729

fusion than competitor approaches, demonstrating the benefits of deep fusion. Furthermore, our best performing model utilizes fewer trainable parameters than ALMT and Self-MM, further highlighting the efficiency of DeepMLF. The last two columns of Table 7 display the relative (absolute) improvement compared to the base model. Consistent with literature, we observe decaying performance improvements (smaller deltas) when using larger LM backbones in MOSEI. This preludes a performance saturation with the increase of trainable parameters, similar to the one observed in SIMS Table 6. For a fixed amount of data, there exists an

TABLE 6: *DeepMLF fusion depth analysis on SIMS.* GPT2: language backbone size; MM Blocks (#): the LM layers after which we insert multimodal blocks (their total number); MM Params: the number of total (mutlimodal) trainable parameters.

GPT2	MM Blocks (#)	MM Params(M)	MAE(↓)	Corr(↑)
base	8:12 (5)	29.75	0.374	0.697
base	6:12 (7)	41.65	0.353	0.729
base	4:12 (9)	53.55	0.358	0.724

TABLE 7: *DeepMLF fusion depth analysis on MOSEI.* GPT2: language backbone size; MM Blocks (#): the LM layers after which we insert multimodal blocks (their total number); MM Par.: the number of total (mutlimodal) trainable parameters in millions (M); Δ(Metric): relative metric improvement (%) from base DeepMLF model.

GPT2	MM Blocks(#)	MM Par.	$|\Delta$Acc2$|$	$|\Delta$MAE$|$
base	4-6-8:12 (7)	41.65	–	–
med.	9-12-15-18-21-24 (6)	63.30	1.36	3.40
large	8-15-22-29-36 (5)	82.35	0.17	2.35

optimal fusion depth, that achieves the best multimodal performance, typically exceeding that of existing competitors.

6.2.2 The impact of learnable fusion tokens (n_f)

The optimal number of learnable fusion tokens (n_f) remains consistently small across all architectural configurations, with values varying according to modality performance gaps in different datasets. Specifically, we observe larger n_f values for datasets with smaller modality performance gaps: SIMS requires 20 tokens, MOSEI 12, and MOSI 8 tokens respectively. *This pattern suggests that datasets with more balanced modality contributions benefit from additional fusion tokens which capture richer multimodal interactions.*

Figure 3 illustrates that n_f transfers across LM backbones, showing consistent behavior as a robust hyperparameter. Increasing n_f beyond its optimal value leads to performance degradation, suggesting that a small set of fusion tokens provides the most effective approach for multimodal information integration.

6.2.3 The impact of encoder initialization

In this experiment we evaluate the impact of the audio-visual encoder initialization scheme on the overall performance of DeepMLF. We conduct experiments on both MOSEI and SIMS datasets, to assess the effect in language-dominated and more balanced setups. We examine three initialization strategies: pretrained[13] encoder and further fine-tuned (*Pre&Tune*), frozen pretrained encoder (*Pre&Fro*), and randomly initialized encoder trained from scratch (*Random&Tune*).

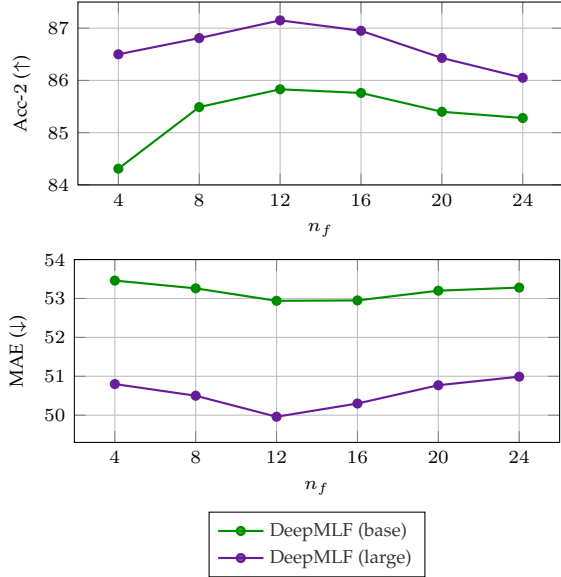13. Pretraining of AV Encoder is performed on each MSA dataset before integrating into DeepMLF.

Fig. 3: Impact of the number n_f of *learnable fusion tokens* on MOSEI. ↑ / ↓: higher/lower is better.

TABLE 8: Assesing the impact of encoder initialization in DeepMLF. *Pre&Tune*: initialize encoder from pretrained and further tune; *Pre&Fro*: intitialize encoder from pretrained and keep frozen; *Random&Tune*: initialize encoder randomly and tune it.

SETTING	MOSEI		SIMS	
	ACC2↑	MAE↓	ACC2↑	MAE↓
Pre&Tune	87.15	0.499	82.75	0.353
Pre&Fro	86.94	0.504	81.99	0.363
Random&Tune	86.59	0.508	80.43	0.382

The *Pre&Fro* setup shows a minor performance degradation compared to *Pre&Tune* for MOSEI and a larger for SIMS. This result shows that further tuning of the audiovisual encoder impacts DeepMLF positively. The *Random&Tune* configuration for MOSEI yields a larger performance degradation compared to *Pre&Fro*. SIMS exhibits a substantial performance degradation under random initialization and training from scratch.

We attribute these outcomes to each dataset's modality characteristics. In the smaller and more balanced SIMS dataset, audiovisual features contribute significantly to multimodal performance, making encoder initialization a crucial factor for achieving optimal performance. For MOSEI, which is more language dominated, the AV Encoder effectively captures task-relevant audiovisual features, resulting in more robust performance across initialization schemes.

These results highlight the importance of a progressive learning approach (fusion curriculum), *i.e.*, first developing unimodal representations (pretrained LM), then building audiovisual representations (AV encoder) and finally learning multimodal connections, rather than jointly learning everything from scratch. This finding on fusion curriculum aligns with broader multimodal learning literature [66] suggesting that progressive representation acquisition benefits multimodal representation learning.

TABLE 9: Language characteristics for MOSI and MOSEI. These factors highlight key differences in language distributions between the two datasets. The final row demonstrates SBERT's ability to discriminate between sentences from each distribution.

FACTORS	MOSI	MOSEI
Sentences	2,199	23,453
Vocabulary Size	3,107	23,026
Sentence Length(s)	4.2	7.3
Topics	Movie Reviews	Diverse
Tone	Informal	Mixed
MOSI vs MOSEI	ACC2	90.0(%)

6.2.4 The impact of language data on DeepMLF

DeepMLF demonstrates effective performance across both English (MOSI, MOSEI) and Chinese (SIMS) languages. This cross-lingual capability, while desirable in language fusion algorithms, is not uniformly observed in MSA models. Notably, BERT-Chinese features alone achieves 77.35 Acc-2 (Table 2), surpassing several multimodal fusion approaches shown in Table 5 and documented by Mao *et al*. [21] (Table 4).

We focus our analysis on the MOSI dataset, where we need to scale up to billion parameter LLMs (SmolLM2-1.7B) to achieve performance comparable, though still inferior, to much smaller embedding models. To understand this behavior, we compare MOSI with MOSEI, both English datasets, by first examining the impact of dataset size. We conduct experiments on MOSEI by randomly sampling 300 subsets, each containing 5-10% of the full data, training a classifier on GPT2-base embeddings, and evaluating on MOSEI's test set. GPT2-base achieves 77% Acc-2 on these MOSEI subsets on average, matching SmolLM2-1.7B's MOSI performance but exceeding GPT2-base's MOSI results by 10%. This GPT2 performance gap between MOSI and MOSEI subsets suggests additional factors affecting MOSI's language distribution, besides size.

These factors, illustrated in Table 9, demonstrate that MOSI's language is characterized as informal and topic-specific. We hypothesize that these factors create a *language distribution shift* compared to GPT2's training dataset, which contains high-quality web pages that have been curated/filtered by humans [42]. This distribution shift explains why GPT2s show degraded performance on MOSI, and why DeepMLF does not improve with the addition of the LM task loss. To validate our hypothesis we utilize SBERT [86] embeddings[14] and train a linear classifier to distinguish between MOSI and MOSEI language samples. This classifier achieves 90% binary accuracy, confirming that the two language distributions are separable, *i.e.*, the distribution shift exists.

Findings: DeepMLF shows the desirable property of multilingual capability. Furthermore, DeepMLF leverages (L)LM backbones that are primarily designed for generation rather than downstream tasks, making it more sensitive to language distribution shifts. To address this, in cases

14. https://huggingface.co/sentence-transformers/all-mpnet-base-v2

TABLE 10: *Fusion block ablation.* The fusion mechanism consists of two main parts. The GCA (gated cross-attention) which is a multimodal cross-attention operation between the audiovisual information **Z** and a combination of fusion ($\mathbf{X}_f$) or language ($\mathbf{X}_t$) tokens. The FFW which is inserted on top of the GCA following standard encoder-decoder design. The colored line illustrates the scheme utilized in DeepMLF. ↑ / ↓: higher/lower is better.

GCA	FFW	ACC2↑	F1↑	MAE↓	CORR↑
$(\mathbf{X}_f, \mathbf{Z})$	✓	82.75	83.15	0.353	0.729
$(\mathbf{X}_t \| \mathbf{X}_f, \mathbf{Z})$	✓	82.20	82.49	0.359	0.721
$(\mathbf{X}_t, \mathbf{Z})$	✓	81.14	81.46	0.365	0.711
$(\mathbf{X}_f, \mathbf{Z})$	✗	80.88	81.32	0.384	0.678

like MOSI, we utilize larger LLM backbones, which better capture language characteristics. Notably, integrating small-billion LLMs (1.7B) enables DeepMLF to outperform all encoder-based (*e.g.* BERT) competitors in multimodal scenarios.

6.3 Ablation study

6.3.1 Fusion Scheme Ablation

In this experiment, we systematically evaluate different configurations of our proposed fusion mechanism. As described in Equation 10, the MM Block consists of two primary modules: the Gated Cross-Attention (GCA) and the Feed-Forward Network (FFW). We modify the GCA module to investigate all four possible input combinations of fusion $\mathbf{X}_f$ and language tokens $\mathbf{X}_t$. Additionally, we examine the impact of the FFW layer by conducting experiments where this component is omitted from the architecture. Our experiments are carried on SIMS.

DeepMLF setup, where fusion tokens $\mathbf{X}_f$ interact with audiovisual features **Z** followed by a FFW layer, achieves the best results. Further injecting linguistic information $\mathbf{X}_t$ into the GCA mechanism slightly degrades performance (second row). In the third row, when we omit the fusion tokens from the GCA and retain only the language tokens, we observe a larger performance drop. Finally, removing the FFW from our setup significantly impacts performance negatively.

These results highlight two crucial factors in the fusion mechanism design. First, the FFW component plays an essential role in the overall fusion process. Second, the design decision to implement controlled interaction specifically between fusion tokens and audiovisual tokens, without additional elements, produces optimal performance.

6.3.2 Loss term ablation

Our ablation study in Table 11 reveals a consistent pattern in the importance of loss terms across both MOSEI and SIMS datasets, consisting of a primary and a secondary tier. The primary tier consists of the fusion loss L_f and language modelling loss L_{LM}, which emerge as the most crucial terms of our objective function. Removing them leads to the largest performance drops (MOSEI: 0.73/0.44 Acc-2, and SIMS: 1.57/1.61 Acc-2, for L_f and L_{LM}). The fusion loss L_f plays a fundamental role by forcing the network to accumulate meaningful task-predictive multimodal information in

TABLE 11: *Loss term ablation on DeepMLF*: audiovisual loss (L_{av}), text loss (L_t), Fusion Token Loss (L_f), and causal language modelling loss (L_{LM}). The table illustrates results as MOSEI/SIMS for both Acc-2 and MAE. ↑ / ↓: higher/lower is better; <mark>Tier-1</mark>, <mark>Tier-2</mark>

L_{av}	L_t	L_f	L_{LM}	ACC-2↑	MAE↓
✓	✓	✓	✓	**87.15/82.75**	**0.499/0.353**
	✓	✓	✓	86.89/81.96	0.502/0.358
✓		✓	✓	86.73/81.66	0.505/0.355
✓	✓		✓	86.42/81.18	0.511/0.367
✓	✓	✓		86.71/81.14	0.512/0.377

TABLE 12: *Language embedding augmentation ablation.* ↑ / ↓: higher/lower is better.

EMBEDDING AUG.	ACC2↑	F1↑	MAE↓	CORR↑
SeqAug	82.75	83.15	0.353	0.729
Noise Injection	81.05	81.44	0.368	0.708
Dropout	80.74	81.12	0.372	0.702

the set of learnable fusion tokens. Meanwhile, the LM loss serves as a critical regularization mechanism, preventing catastrophic forgetting on the pretrained LM.

In the secondary tier our ablation places the audiovisual (L_{av}) and language(L_t) loss terms. Specifically, for both datasets, L_t and L_{av} have comparable effects, *i.e.*, 0.79 vs 1.09 Acc-2 for SIMS, and 0.42 vs 0.26 Acc-2 for MOSEI. Moreover, MOSEI-traned DeepMLF appears to be more robust against loss term removal, exhibiting smaller performance drops compared to SIMS-trained DeepMLF.

6.3.3 Augmentation Ablation

In this ablation, we demonstrate the critical role of language embedding augmentation in DeepMLF's performance by comparing our integrated SeqAug method against conventional approaches such as dropout and noise injection on the SIMS dataset. To ensure fair comparison, we independently tune dropout and noise injection hyperparameters and report the average scores of their best performing configuration in Table 12. Our experiments reveal that embedding regularization is crucial for DeepMLF's training recipe, with SeqAug emerging as significantly more effective than both alternatives (performance drops of 1.36 Acc-2 and 0.1 MAE for noise injection, 1.67 Acc-2 and 0.13 MAE for noise injection). SeqAug proves more suitable for our pretrained LM backbone, since it resamples from the underlying language distribution, *i.e.*, performs a soft permutation in the sequence which acts as an augmentation while preserving semantics. Replacing SeqAug with alternative augmentation methods leads to performance degradation comparable to removing a tier-1 loss term from the total objective (recall drops of 1.32 Acc-2 and 0.2 for tier-1 losses). These results underscore that proper regularization through SeqAug is crucial for DeepMLF's optimal performance.

6.3.4 Gating Mechanism Ablation

In this ablation study, we examine variants of the gating mechanism, with analysis conducted on the MOSEI dataset. We compare our proposed sigmoid gating against two alternatives: 1) tanh gating (as used in Flamingo [28]) and 2)

TABLE 13: *Gating mechanism ablation*. Gating Mech: the gating mechanism studied; Init: the initial gate value; $\uparrow$ / $\downarrow$: higher/lower is better.

GATING	INIT.	RANGE	ACC2$\uparrow$	F1$\uparrow$	MAE$\downarrow$	CORR$\uparrow$
sigmoid	0.5	[0,1]	87.15	87.10	0.499	0.804
tanh	0.0	[-1,1]	87.01	86.98	0.505	0.800
None	1.0	-	86.86	86.84	0.511	0.798

complete gate removal (None), which results in a vanilla encoder-decoder architecture [26].

Our experiments demonstrate that sigmoid gating consistently outperforms both alternative approaches. Furthermore, the presence of any gating mechanism proves superior to non-gating configurations, though even the configuration without gating still achieves state-of-the-art performance. The primary differences between our proposed sigmoid and tanh gating are: 1) initialization strategy (sigmoid initialized at 0.5 versus tanh at zero/closed gate), and 2) gate bounds (sigmoid's bounded positive range versus tanh's $[-1, 1]$ range). The balanced initialization and positive range characteristics of sigmoid gating prove most effective for our MM Block implementation.

7 DISCUSSION

7.1 Limitations

DeepMLF leverages pretrained LMs as language backbones, inheriting certain limitations of these decoder architectures. First, these models are primarily designed for generation rather than predictive tasks, which, as shown in our language distribution analysis, can impact predictive performance, and requires larger LM integration. Second, the autoregressive nature of language modelling has increased inference time compared to encoder-based (non-autoregressive) approaches. However, all existing literature regarding speeding up inference is directly transferable to DeepMLF.

7.2 Conclusion

This work introduces DeepMLF, a multimodal language model framework with learnable tokens for deep fusion. Unlike current MSA research, DeepMLF positions fusion depth, scalability and dedicated multimodal capacity allocation as necessary factors for effective multimodal fusion. Our framework consists of a multimodal encoder that supplies a pretrained LM backbone with audiovisual information. A small set of fusion tokens is appended at the LM input and progressively: 1) gather linguistic information via LM blocks and 2) interact with information from the audiovisual encoder via MM Blocks. These MM Blocks are novel cross-modal cross-attention modules that are inserted after the LM layers (LM Blocks), enabling both deep and scalable fusion by design while accumulating multimodal information in the learnable tokens. DeepMLF is coupled with a learning recipe which consists of modality-specific task losses, a language modelling loss, and an embedding regularization technique which acts as a language regularizer. These elements collectively form a multimodal deep fusion framework, that can be applied at any language-based multimodal scenario.

Our comprehensive experimental analysis reveals that deeper fusion schemes (5-7 layers) consistently outperform shallower approaches, challenging existing approaches in the field. Moreover, we demonstrate that a relatively small multimodal capacity (8-20 fusion tokens) achieves optimal performance, providing important insights for multimodal architecture design. Furthermore, we show that progressively learning multimodal representations consistently outperforms jointly learning all representations at once, highlighting the importance of fusion curriculum.

We evaluate DeepMLF across three MSA benchmarks covering different languages, dataset sizes, language distributions, and modality imbalance levels. Our recipe achieves state-of-the-art results across all datasets examined showcasing its applicability and versatility. Our ablation illustrates that the proposed MM Block design outperforms other alternatives and that removing any loss term component results in performance degradation, highlighting the synergetic nature of our learning recipe. Additionally, we find that the integrated language embedding augmentation consistently works better than existing approaches in the literature.

7.3 Future Work

In future work, we plan to extend DeepMLF to additional tasks, domains, and modalities, including vision-language and audio-language models. A promising research direction is applying our approach, DeepMLF, to self-supervised setups such as multimodal language modelling, where integration with other modalities and larger LLMs could establish a novel architectural paradigm for multimodal learning. Further research will also explore in-depth analysis and utilization of learnable tokens in the generative process, and integration with other generative frameworks, such as diffusion models, where these tokens can serve as multimodal latents.

REFERENCES

[1] P. Maragos, A. Potamianos, and P. Gros, *Multimodal processing and interaction: audio, video, text*. Springer Science & Business Media, 2008, vol. 33.

[2] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Neural Information Processing Systems*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:199453025

[3] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *European conference on computer vision*. Springer, 2020, pp. 104–120.

[4] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 200–212, 2021.

[5] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 19 730–19 742. [Online]. Available: https://proceedings.mlr.press/v202/li23q.html

[6] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[7] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, 2012.

[8] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114.

[9] Y. Sun, S. Mai, and H. Hu, "Learning to learn better unimodal representations via adaptive multimodal meta-learning," *IEEE Transactions on Affective Computing*, 2022.

[10] Y. H. Tsai, M. Ma, M. Yang, R. Salakhutdinov, and L. Morency, "Multimodal routing: Improving local and global interpretability of multimodal language analysis," in *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 2020, pp. 1823–1833.

[11] H. Zhang, Y. Wang, G. Yin, K. Liu, Y. Liu, and T. Yu, "Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023, pp. 756–767.

[12] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 9180–9192.

[13] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 10 790–10 797.

[14] L. Sun, Z. Lian, B. Liu, and J. Tao, "Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, 2023.

[15] E. Georgiou, C. Papaioannou, and A. Potamianos, "Deep hierarchical fusion with application in sentiment analysis," *Interspeech 2019*, 2019.

[16] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 6558–6569.

[17] G. Hu, T.-E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li, "UniMSE: Towards unified multimodal sentiment analysis and emotion recognition," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Dec. 2022, pp. 7837–7851.

[18] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[19] A. Zadeh and P. Pu, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, 2018.

[20] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, and K. Yang, "Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3718–3727.

[21] H. Mao, Z. Yuan, H. Xu, W. Yu, Y. Liu, and K. Gao, "M-SENA: An integrated platform for multimodal sentiment analysis," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 2022, pp. 204–213.

[22] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 8717–8727, 2018.

[23] M. Hou, J. Tang, J. Zhang, W. Kong, and Q. Zhao, "Deep multimodal multilinear fusion with high-order polynomial pooling," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/f56d8183992b6c54c92c16a8519a6e2b-Paper.pdf

[24] J. E. A. Ovalle, T. Solorio, M. Montes-y-Gómez, and F. A. González, "Gated multimodal units for information fusion," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=S12_nquOe

[25] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[27] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[28] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[29] F. Chen, M. Han, H. Zhao, Q. Zhang, J. Shi, S. Xu, and B. Xu, "X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages," *arXiv preprint arXiv:2305.04160*, 2023.

[30] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023.

[31] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, 2020, pp. 121–137.

[32] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[33] D. Gao, K. Li, R. Wang, S. Shan, and X. Chen, "Multi-modal graph neural network for joint reasoning on vision and scene text," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 746–12 756.

[34] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T.-S. Chua, "Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1437–1445.

[35] Y. Wang, M. Yasunaga, H. Ren, S. Wada, and J. Leskovec, "Vqagnn: Reasoning with multimodal knowledge via graph neural networks for visual question answering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 21 582–21 592.

[36] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=1tZbq88f27

[37] R. Zhang, J. Han, C. Liu, P. Gao, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, and Y. Qiao, "Llama-adapter: Efficient fine-tuning of language models with zero-init attention," *arXiv preprint arXiv:2303.16199*, 2023.

[38] W. Li, X. Xu, J. Liu, and X. Xiao, "Unimo-g: Unified image generation through multimodal conditional diffusion," *arXiv preprint arXiv:2401.13388*, 2024.

[39] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *arXiv preprint arXiv:2306.13549*, 2023.

[40] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 113–12 132, 2023.

[41] Z. M. Ziegler, L. Melas-Kyriazi, S. Gehrmann, and A. M. Rush, "Encoder-agnostic adaptation for conditional language generation," *arXiv preprint arXiv:1908.06938*, 2019.

[42] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," in *https://api.semanticscholar.org/CorpusID:160025533*, 2019.

[43] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *International conference on machine learning*. PMLR, 2021, pp. 4651–4664.

[44] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun,

"Attention bottlenecks for multimodal fusion," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 200–14 213, 2021.

[45] Z. Kong, A. Goel, R. Badlani, W. Ping, R. Valle, and B. Catanzaro, "Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities," in *Forty-first International Conference on Machine Learning*, 2024.

[46] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[47] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency, "Multi-level multiple attentions for contextual multimodal sentiment analysis," in *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2017, pp. 1033–1038.

[48] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2018. NIH Public Access, 2018, p. 2225.

[49] G. Paraskevopoulos, E. Georgiou, and A. Potamianos, "Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4573–4577.

[50] A. Joshi, A. Bhat, A. Jain, A. Singh, and A. Modi, "COGMEN: COntextualized GNN based multimodal emotion recognitioN," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2022, pp. 4148–4164.

[51] W. Rahman, M. K. Hasan, S. Lee, A. B. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2359–2369.

[52] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[53] D. Wang, S. Liu, Q. Wang, Y. Tian, L. He, and X. Gao, "Cross-modal enhancement network for multimodal sentiment analysis," *IEEE Transactions on Multimedia*, vol. 25, pp. 4909–4921, 2023.

[54] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6892–6899.

[55] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1122–1131.

[56] Z. Yuan, Y. Liu, H. Xu, and K. Gao, "Noise imitation based adversarial training for robust multimodal sentiment analysis," *IEEE Transactions on Multimedia*, vol. 26, pp. 529–539, 2024.

[57] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[58] P. J. Liu*, M. Saleh*, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer, "Generating wikipedia by summarizing long sequences," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=Hyg0vbWC-

[59] N. Shazeer, "Fast transformer decoding: One write-head is all you need," *arXiv preprint arXiv:1911.02150*, 2019.

[60] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," *ArXiv*, vol. abs/2302.13971, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257219404

[61] A. Chronopoulou, C. Baziotis, and A. Potamianos, "An embarrassingly simple approach for transfer learning from pretrained language models," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio,

Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 2089–2095. [Online]. Available: https://aclanthology.org/N19-1213

[62] D. Gkoumas, Q. Li, C. Lioma, Y. Yu, and D. Song, "What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis," *Information Fusion*, vol. 66, pp. 184–197, 2021.

[63] E. Georgiou, G. Paraskevopoulos, and A. Potamianos, "M3: MultiModal Masking Applied to Sentiment Analysis," in *Proc. Interspeech 2021*, 2021, pp. 2876–2880.

[64] E. Georgiou and A. Potamianos, "Seqaug: Sequential feature resampling as a modality agnostic augmentation method," *arXiv preprint arXiv:2305.01954*, 2023.

[65] S. Hu, L. Zhou, S. Liu, S. Chen, L. Meng, H. Hao, J. Pan, X. Liu, J. Li, S. Sivasankaran *et al.*, "Wavllm: Towards robust and adaptive speech large language model," *arXiv preprint arXiv:2404.00656*, 2024.

[66] C. Du, J. Teng, T. Li, Y. Liu, T. Yuan, Y. Wang, Y. Yuan, and H. Zhao, "On uni-modal feature learning in supervised multi-modal learning," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 8632–8656.

[67] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 695–12 705.

[68] N. Wu, S. Jastrzebski, K. Cho, and K. J. Geras, "Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 2022, pp. 24 043–24 055.

[69] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, "What makes multi-modal learning better than single (provably)," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 944–10 956, 2021.

[70] Y. Huang, J. Lin, C. Zhou, H. Yang, and L. Huang, "Modality competition: What makes joint training of multi-modal network fail in deep learning? (Provably)," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 17–23 Jul 2022, pp. 9226–9259.

[71] Y. Liu, Z. Yuan, H. Mao, Z. Liang, W. Yang, Y. Qiu, T. Cheng, X. Li, H. Xu, and K. Gao, "Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module," in *Proceedings of the 2022 International Conference on Multimodal Interaction*, 2022, pp. 247–258.

[72] Z. Liu, Z. Tang, X. Shi, A. Zhang, M. Li, A. Shrivastava, and A. G. Wilson, "Learning multimodal data augmentation in feature space," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[73] E. Georgiou, Y. Avrithis, and A. Potamianos, "Powmix: A versatile regularizer for multimodal sentiment analysis," *arXiv preprint arXiv:2312.12334*, 2023.

[74] D. Hazarika, Y. Li, B. Cheng, S. Zhao, R. Zimmermann, and S. Poria, "Analyzing modality robustness in multimodal sentiment analysis," *arXiv preprint arXiv:2205.15465*, 2022.

[75] C. Jin, C. Luo, M. Yan, G. Zhao, G. Zhang, and S. Zhang, "Weakening the dominant role of text: Cmosi dataset and multimodal semantic enhancement network," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2023.

[76] Z. Lian, L. Sun, Y. Ren, H. Gu, H. Sun, L. Chen, B. Liu, and J. Tao, "Merbench: A unified evaluation benchmark for multimodal emotion recognition," *arXiv preprint arXiv:2401.03429*, 2024.

[77] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 960–964.

[78] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[79] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[80] Y. H. Tsai, P. P. Liang, A. Zadeh, L. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *ICLR*, 2019.

[81] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, and X. Luo, "Tetfn: A text enhanced transformer fusion network for multimodal sentiment analysis," *Pattern Recognition*, vol. 136, p. 109259, 2023.

[82] E. Cambria, D. Hazarika, S. Poria, A. Hussain, and R. Subramanyam, "Benchmarking multimodal sentiment analysis," in *Computational Linguistics and Intelligent Text Processing: 18th International Conference, CICLing 2017, Budapest, Hungary, April 17–23, 2017, Revised Selected Papers, Part II 18*. Springer, 2018, pp. 166–179.

[83] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[84] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[85] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[86] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084