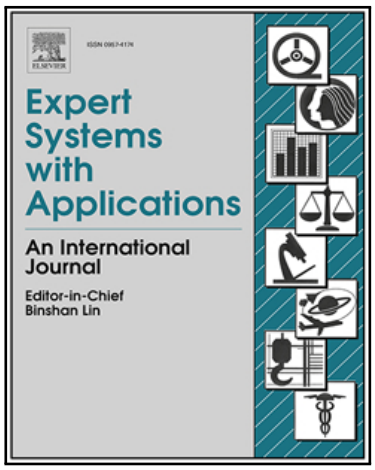


SG-DTAM: Joint Staged Generation and Dynamic Time Alignment for Missing and Unaligned Modalities in Sentiment Analysis

Deling Huang, Ran Gao, Geng Zhang, Jian Yu

PII: S0957-4174(25)03365-2
DOI: https://doi.org/10.1016/j.eswa.2025.129750
Reference: ESWA 129750

To appear in: *Expert Systems With Applications*

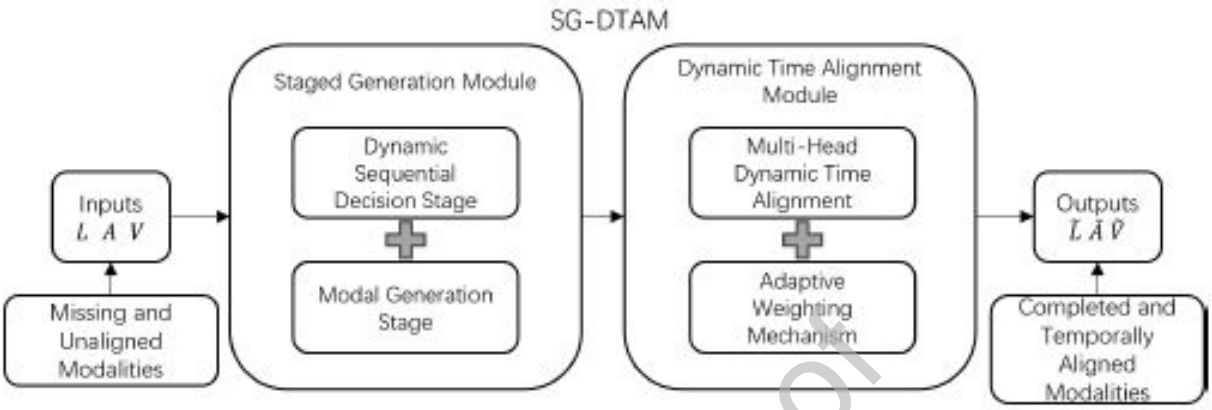
Received date: 26 May 2025
Revised date: 28 August 2025
Accepted date: 14 September 2025

Please cite this article as: Deling Huang, Ran Gao, Geng Zhang, Jian Yu, SG-DTAM: Joint Staged Generation and Dynamic Time Alignment for Missing and Unaligned Modalities in Sentiment Analysis, *Expert Systems With Applications* (2025), doi: https://doi.org/10.1016/j.eswa.2025.129750

Graphical Abstract

SG-DTAM: Joint Staged Generation and Dynamic Time Alignment for Missing and Unaligned Modalities in Sentiment Analysis

Deling Huang, Ran Gao, Geng Zhang, Jian Yu



Highlights

SG-DTAM: Joint Staged Generation and Dynamic Time Alignment for Missing and Unaligned Modalities in Sentiment Analysis

Deling Huang, Ran Gao, Geng Zhang, Jian Yu

- SG-DTAM addresses both missing and unaligned data in multimodal analysis.
- Uses scenario-based MI to dynamically order generation and recover missing data.
- Proposes DTAM to align misaligned modalities via multi-head temporal attention.
- Achieves 84.3% accuracy on CMU-MOSI despite using only 152K parameters.
- Delivers robust sentiment analysis on noisy and sparse data in real-world scenarios.

SG-DTAM: Joint Staged Generation and Dynamic Time Alignment for Missing and Unaligned Modalities in Sentiment Analysis

Deling Huang[a,*], Ran Gao[a], Geng Zhang[a] and Jian Yu[b]

[a]Chongqing University of Posts and Telecommunications, Chongqing, 400065, China
[b]Auckland University of Technology, Auckland, 1010 New Zealand

ARTICLE INFO

Keywords:
Cross-modal attention mechanism
Missing modalities
Multimodal sentiment analysis
Unaligned multimodal sequences

ABSTRACT

Multimodal Sentiment Analysis (MSA) aims to infer users' emotional states by integrating information from multiple modalities, such as language, audio, and visual data. However, real-world multimodal data often presents two critical challenges: missing modalities and unaligned multimodal sequences. Missing sources can lead to information loss, while temporal misalignment introduces inconsistencies—both of which significantly degrade analytical accuracy. While a plethora of existing approaches effectively address each challenge in isolation, few can tackle both simultaneously without resorting to complex architectures or incurring substantial computational costs. To overcome these limitations, we propose SG-DTAM, a novel framework that combines staged generation with multi-head dynamic temporal alignment. In the first stage, conditional mutual information is employed to guide a hierarchical series of cross modal attention modules that sequentially reconstruct each missing modality. In the following alignment stage, a set of attention heads with adaptive weighting reconciles temporal discrepancies across all modalities without any reliance on external synchronization labels. Throughout the process, we innovatively introduce a dual supervision objective that combines an InfoNCE based contrastive loss and a reconstruction loss ensures both precise modality synthesis and the development of resilient feature representations. We evaluate SG-DTAM on four benchmark MSA datasets—CMU-MOSI, CMU-MOSEI, IEMOCAP, and MELD. Experimental results demonstrate that our framework achieves competitive or state-of-the-art performance with relatively few learnable parameters. Notably, SG-DTAM exhibits robust performance in scenarios involving both missing and misaligned modalities, underscoring its effectiveness in real-world multimodal sentiment analysis tasks.

1. Introduction

This section reviews the background and importance of multimodal sentiment analysis, outlines the dual challenges of uncertain modality absence and temporal misalignment that impede existing fusion methods, and presents our SG-DTAM framework as a comprehensive solution to these challenges.

In recent years, advances in deep learning and large-scale pre-trained models for computer vision and NLP have made multimodal sentiment analysis (MSA) increasingly vital in affective computing and human-computer interaction. Human emotional expression is inherently multidimensional: language conveys semantics, facial expressions, and gestures carry nonverbal cues, and speech prosody modulates emotional valence. Relying on a single modality often fails to capture sarcasm, micro-expressions, or subtle affective nuances, making robust multimodal fusion essential for accurate sentiment inference.

In real-world settings, sensor failures, environmental noise, and discrepancies in device sampling rates or recording durations (as illustrated in Figure. 1) frequently lead to missing modalities or temporal misalignment. These issues severely undermine the effectiveness of fusion models that assume complete, perfectly synchronized inputs, highlighting the need for more resilient multimodal techniques.

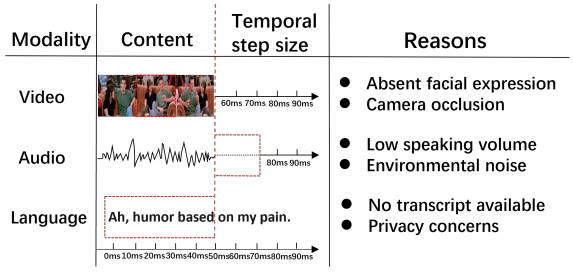


Figure 1: An illustration of three parallel data streams (video, audio, language), depicting temporal gaps where individual modalities become unavailable. Missing segments are indicated with red dashed lines, and typical causes—such as facial occlusion in video, ambient noise in audio, and privacy-related text removal—are listed alongside.

Existing research in multimodal sentiment analysis has primarily focused on modality fusion methods. Early works such as Bidirectional Contextual LSTM (BC-LSTM) [32] and Tensor Fusion Network (TFN) [48] model inter-modal dynamics through hierarchical architectures or tensor Cartesian products. With the advent of pre-trained models like BERT [5], dynamic embedding–based fusion techniques have further improved performance. For instance, Hazarika

*Funded by the Science and Technology Innovation Key R&D Program of Chongqing, China (Grant No. CSTB2023TIAD-STX0031).
*Corresponding author
huangdl@cqupt.edu.cn (D. Huang); 2894085674@qq.com (R. Gao); 2386000147@qq.com (G. Zhang); jian.yu@aut.ac.nz (J. Yu)
ORCID(s): 0000-0002-4297-9849 (D. Huang); 0009-0003-5669-2050 (R. Gao); 0009-0002-2109-7828 (G. Zhang); 0000-0002-2257-7279 (J. Yu)

et al. [11] introduced MISA, which learns both modality-invariant and modality-specific representations, while Rahman et al. [34] demonstrated how to integrate multimodal information into large pre-trained transformers. Subsequently, Yu et al. [46] leveraged self-supervised multi-task learning to obtain richer modality-specific features, and Xiao et al. [43] developed a multi-channel attentive graph convolutional network to fuse sentiment cues. However, most of these studies rely on two idealized assumptions—complete modalities and strict temporal alignment—that rarely hold in real-world scenarios.

Early approaches to missing-modality compensation often adopted "discard" or "impute" strategies: the former simply remove missing channels, risking loss of critical information; the latter uses matrix completion or sample-similarity interpolation, partially restoring data integrity but overlooking deep feature correlations, which limits fine-grained sentiment recognition. More recently, deep learning methods leveraging autoencoders, cycle-consistency, and Transformer architectures have attempted to uncover implicit inter-modal relationships—examples include the sequence-to-sequence and cyclic translation mechanisms in Seq2seq2sentiment [30] and Found in Translation [29], Adaptive Modality Distillation (AMD) [28], Coupled-Translation Fusion Network (CTFN) [38], Tag-Assisted Transformer Encoder (TATE) [51], MTMSA [21], Unified Self-Distillation Framework [19], and UMCA/MIA [42]. While these efforts achieve progress in single-modality absence scenarios, they are largely restricted to one missing channel and require separate models for each missing pattern, limiting flexibility in complex, dynamic environments.

Nevertheless, even with complete inputs, disparities in sensor sampling rates and trigger timings induce temporal asynchrony that undermines cross-modal fusion. To mitigate this, scholars have devised Transformer variants and sparse attention schemas such as Progressive Modality Reinforcement (PMR) [23], Sparse Phased Transformer (SPT) [3], Modality Invariant Crossmodal Attention (MICA) [44], MFSA [18] and TMRN [47], which markedly enhance alignment precision. However, these solutions do not address the reconstruction of absent modalities. In recent years, a multitude of strategies has emerged to complete missing modalities in unaligned multimodal sequences. These designs, however, typically rely on complex training pipelines with multiple sequential phases and incorporate excessively elaborate feature extraction and reconstruction networks, resulting in architectural redundancy and an unrestrained increase in model parameters that hamper practical deployment.

Motivated by these challenges, we propose the SG-DTAM framework, which integrates Staged Generation and Dynamic Time Alignment to holistically tackle modality absence (due to sensor failures or data loss) and temporal misalignment (stemming from varying sampling rates or duration disparities across modalities). The Staged Generation Module precomputes conditional mutual information among language, audio, and video modalities to adaptively determine the generation order. At each stage, Staged Cross-Modal Attention (SCA) leverages available modalities to model the contextual representation of missing ones, while integrating feedback from previously generated modalities to maximally preserve cross-modal dependencies. The Dynamic Time Alignment Module adopts Multi-Head Dynamic Time Alignment (MHDTA), which integrates time-shift-aware masking and dynamic alignment weighting into a multi-head self-attention framework. This mechanism selectively filters alignable time steps, enabling a seamless fusion of audio, video, and language features on a unified temporal scale. During training, the framework jointly optimizes a contrastive loss and a reconstruction loss: the former promotes discriminative similarity between generated and real features, while the latter directly constrains the reconstruction error between generated and original modalities. This dual-supervision strategy mutually reinforces generation quality and consistency. Finally, the aligned multimodal representations are fed into a lightweight classifier for end-to-end high-precision sentiment prediction.

In our experiments on CMU-MOSI, CMU-MOSEI, IEMOCAP, and MELD, we benchmark SG–DTAM against six state-of-the-art baselines: MMIN [8], a multimodal interaction network that excels at fusing features but lacks explicit missing-modality compensation; MISA [11], which disentangles modality-invariant and modality-specific representations yet assumes fully observed, aligned inputs; TFR-Net [47], a Transformer-based reconstruction model that recovers missing channels at the cost of high computational and parameter overhead; EMT-DLFR [36], which introduces dynamic latent feature restoration to enhance robustness under noisy and incomplete data; a MELD-based model [33], designed for conversational emotion recognition but without explicit handling of missing modalities or temporal misalignment; and UniMF [13], a unified framework handling both missing and unaligned sequences though its reconstruction accuracy remains moderate. All comparisons employ Accuracy, Macro-F1, and MAE to quantify classification performance and regression precision.

To address both modality missing and unaligned issues within a lightweight architecture, the principal contributions of this work are summarized as follows:

- We propose the SG-DTAM framework, which organically integrates staged modality generation and dynamic multi-head temporal alignment, offering a systemic solution to concurrent challenges of missing modalities and temporal misalignment.
- We propose conditional mutual information-guided adaptive generation planning within the generation module, implementing hierarchical, progressive reconstruction of missing modalities via Staged Cross-Modal Attention (SCA).

Table 1
List of abbreviations used in this paper

Abbreviation	Full Term
SG-DTAM	Staged Generation and Dynamic Time Alignment Model
MI	Mutual Information
MAE	Mean Absolute Error
CMU-MOSI	Carnegie Mellon University Multimodal Opinion Sentiment Intensity
CMU-MOSEI	CMU Multimodal Opinion Sentiment and Emotion Intensity
IEMOCAP	Interactive Emotional Dyadic Motion Capture
MELD	Multimodal EmotionLines Dataset
SCA	Staged Cross-Modal Attention
MHDTA	Multi-Head Dynamic Time Alignment
SGM	Staged Generation Mask Mechanism
SGT	Staged Generation Transformer
DTAM	Dynamic Time Alignment Module
AWM	Adaptive Weighting Mechanism

- We propose Multi-Head Dynamic Time Alignment (MHDTA) in the alignment module, leveraging dynamic time alignment mask and adaptive weighting mechanism to calibrate cross-modal temporal sequences.
- We propose a dual-supervision strategy combining contrastive and reconstruction losses, balancing discriminative and consistency constraints to significantly enhance the quality and robustness of generated modalities.

The remainder of this paper is structured as follows. Section II reviews modality fusion strategies, methods for handling missing modalities and unaligned sequences, and recent approaches that jointly address both challenges. Section III introduces the proposed SG-DTAM framework and its two key modules. Section IV presents experiments and ablation studies on four benchmark datasets. Section V analyzes the effect of different word embeddings, investigates the sensitivity of reconstruction and contrastive loss weights, and visualizes the learned joint representations. Finally, Section VI concludes the paper by summarizing our key contributions, discussing the strengths and limitations of SG-DTAM, and outlining promising directions for future research.

To facilitate quick reference, Table 1 lists the main abbreviations used in this paper and their full terms.

2. Related work

In this section, we review four main strands of research in multimodal sentiment analysis: modality fusion, missing modality compensation, unaligned sequence handling, and unified end-to-end methods that address both challenges. We highlight their strengths and weaknesses and explain how they motivate our SG-DTAM design.

2.1. Modality Fusion

Many multimodal fusion techniques operate under the idealized assumption of complete, perfectly synchronized inputs. Early approaches like MISA [11] and MAG [34] capture high-order interactions via subspace projections or gating but depend on full modalities and hand-tuned alignment. Self-MM [46] adds unimodal self-supervision to improve representations yet still presumes no missing channels. Graph-based models such as Multi-channel Attentive GCN [43] refine cross-modal dynamics but maintain static synchronization and cannot reconstruct absent streams. More recent disentanglement and pretraining strategies—including UniMSE [12] and GSIFN [16]—enhance robustness under ideal conditions yet do not jointly address missing-modality recovery and asynchronous alignment.

Building on these foundations, several specialized architectures demonstrate how these fusion ideas perform in constrained or domain-specific scenarios. BC-LSTM [32] uses separate bidirectional LSTM encoders for each modality followed by a fusion layer, but assumes strict synchronization and cannot handle missing or misaligned streams. CRA [7] leverages cross-modal residual attention to refine inter-modality interactions, yet depends on complete inputs and static attention masks. MMIN [8] employs gating networks to model high-order multimodal interactions, but still requires full, aligned modalities and introduces considerable parameter overhead. MELD-based [33] models fuse language and audio in dialog contexts with strong static performance, but lack mechanisms for asynchronous fusion or modality loss.

Despite these advances, most fusion methods still assume complete and temporally aligned modalities, limiting their applicability in scenarios with sensor failures or privacy-induced modality loss.

2.2. Missing Modality Handling

Approaches to missing-modality recovery typically fall into two categories: generative sequence translation and knowledge distillation. Sequence-translation methods—such as Seq2seq2Sentiment [30] and its cyclic extension GME-LSTM [2] reconstruct absent streams via sequence-to-sequence translation and cycle-consistency loss, but require separate models for each missing pattern and ignore temporal asynchrony. TransModality [20] applies a one-pass cross-modal Transformer to translate available modalities into missing ones, yet treats alignment implicitly and fails under complex asynchrony. Distillation-based approaches—Adaptive Modality Distillation (AMD) [28] and the CHFusion [15] transfer information from available channels through teacher–student pipelines, yet rely on static fusion and do not infer an optimal reconstruction order. EMT-DLFR [36] augments Transformer fusion with dual-level feature restoration to recover absent modalities, but depends on pre-trained checkpoints and ignores temporal offsets. Hybrid schemes such as CTFN [38] and Tag-Assisted Transformer Encoder (TATE) [51] improve fault tolerance with hierarchical translation-fusion or tag-guided reconstruction, but introduce multi-stage training and heavy Transformer encoders.

These methods have significantly advanced feature recovery for missing modalities. However, most of them overlook the challenge of temporal misalignment, limiting their effectiveness in asynchronous real-world scenarios.

2.3. Unaligned Sequence Handling

Even with complete inputs, varying sensor sampling rates and event timings can lead to temporal misalignment. Transformer variants have been widely adopted to align asynchronous multimodal sequences. LF-LSTM [39] uses a label-forecasting LSTM to align multimodal sequences by predicting and matching temporal labels, but its window-based alignment demands careful hyperparameter tuning and fails under variable sampling rates. HGI-Net [22] introduces a hierarchical semantic graph interaction network to align high-resolution multimodal inputs via multi-scale graph reasoning, yet it is tailored to remote sensing imagery and lacks explicit modeling of temporal offsets in sentiment tasks. CMA [40] implements hierarchical cross-modal attention with dual audio pathways to synchronize audio–visual signals for sentiment analysis, but focuses solely on audio–visual fusion and cannot accommodate missing modalities or integrate textual data. MFSA [44] learns modality-specific and modality-agnostic factors in parallel through hierarchical attention, yet lacks an explicit model of inter-modal time offsets and cannot guarantee precise synchronization. RAVEN [41] dynamically adjusts word-level representations using synchronized nonverbal cues—such as facial expressions and acoustic features—to implicitly align text and nonverbal streams, yet this unidirectional modulation lacks explicit temporal correction and does not support missing-modality reconstruction. TMRN [18] employs text-driven self- and cross-modal attention to bolster weaker streams under noise, but its effectiveness hinges on high-quality text and it does not reconstruct missing channels.

While these approaches significantly improve alignment, they do not systematically compensate for missing modalities.

2.4. Management of Missing Modalities in Unaligned Multimodal Sequences

Recent works have brought missing-modality compensation and unaligned-sequence handling under one roof. For example, TFR-Net [47] adds a dedicated reconstruction branch atop intra- and inter-modal attention, yet still assumes temporal alignment and incurs substantial model complexity. UniMF [13] unifies missing-modality distillation and dynamic alignment in a single framework, yet relies on modality-specific submodules and multi-stage training, limiting its real-time and lightweight deployment. MCTN [35] leverages cyclic translation to reconstruct and align missing modalities in one pass, yet requires separate models for each modality pair and fails under severe misalignment. BERT-based fusion models [45] embed cross-modal layers within large pre-trained transformers, achieving strong text–audio performance but only for that modality pair and at the cost of heavy parameters. Hybrid Contrastive Learning [26] combines tri-modal contrastive loss with classification objectives to align and compensate implicitly, yet offers no explicit mechanism to generate truly missing streams.

Despite their promising performance, these end-to-end solutions often involve complex multi-stage training, require auxiliary data, or rely on deep Transformer stacks to model intra- and inter-modal interactions—resulting in heavy architectures that hinder real-time and lightweight deployment.

3. Method

In this section, we first outline the design philosophy and modular structure of SG-DTAM, then detail its two core modules: Staged Generation Module and Dynamic Time Alignment Module. As illustrated in Figure 2, consider a scenario where only the language modality L is available (with audio modality A and video modality V missing). The Staged Generation Module operates in two distinct phases: first, a Dynamic Sequential Decision Stage uses mutual information to determine the optimal order for generating missing modalities, $L \to A \to V$ or $L \to V \to A$; second, the Modal Generation Stage recovers the missing modalities in two steps—first generating the chosen modality from L, then generating the remaining modality conditioned on both L and the newly generated one using conditional MI lower-bound optimization. Finally, it feeds the complete triple $(L, \hat{A}, \hat{V})$ into the *Dynamic Time Alignment Module* to synchronize their time series representations before passing them to the prediction module for sentiment classification.

3.1. Preliminaries

3.1.1. Mutual Information

Unconditional MI For any two random variables X and Y, their mutual information measures the overall dependence:

$$I(X;Y) = \mathbb{E}_{x,y}\left[\log \frac{p(x,y)}{p(x)\,p(y)}\right] \tag{1}$$

Conditional MI For three random variables X, Y and Z, the conditional mutual information measures the extra dependence between X and Y once Z is known:

$$I(X;Y \mid Z) = \mathbb{E}_{x,y,z}\left[\log \frac{p(x,y \mid z)}{p(x \mid z)\,p(y \mid z)}\right] \tag{2}$$

3.1.2. Embedding Definitions

In this work, mutual information is computed over three modality embeddings: the text embedding L is the final [CLS] token from a pre-trained BERT model, denoted $h_t \in \mathbb{R}^{d_t}$; the audio embedding A is the final hidden state $h_a \in \mathbb{R}^{d_a}$ of a unidirectional sLSTM encoding frame-level features extracted by COVAREP (or openSMILE), and the video embedding V is the final hidden state $h_v \in \mathbb{R}^{d_v}$ of a unidirectional sLSTM encoding Facet-extracted frame-level features after average pooling.

For unconditional MI estimation $I(X;Y)$, we set:

$$X = h_t, \quad Y = \begin{cases} h_a, & \text{for estimating } I(h_t;h_a) \\ h_v, & \text{for estimating } I(h_t;h_v) \end{cases} \tag{3}$$

For conditional MI estimation $I(X;Y \mid Z)$, once the first modality $Z \in \{h_a, h_v\}$ is generated, we define:

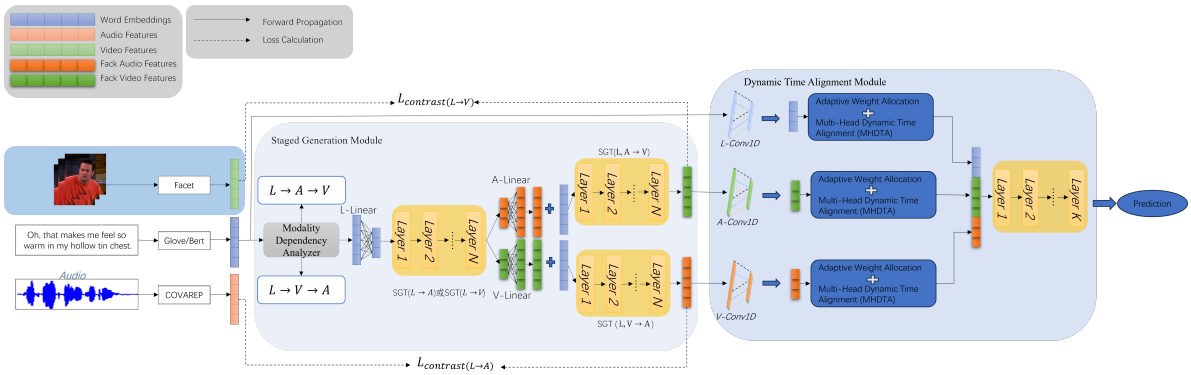


Figure 2: SG-DTAM workflow with missing audio and video modalities. The framework comprises three core components: First, the *Dynamic Sequential Decision Stage* dynamically determines the generation order based on cross-modal mutual information. Next, the *Staged Generation Module* sequentially reconstructs the first missing modality from L, followed by the second using L and the reconstructed first modality. Finally, the *Dynamic Time Alignment Module* temporally aligns $(L, \hat{A}, \hat{V})$ for final sentiment prediction. Solid arrows denote data flow; dashed arrows denote loss paths.

$$
\begin{aligned}
X &= h_t \\
Z &= \begin{cases} h_a, & \text{if audio is generated first} \\ h_v, & \text{if video is generated first} \end{cases} \\
Y &= \begin{cases} h_v, & Z = h_a \\ h_a, & Z = h_v \end{cases}
\end{aligned} \tag{4}
$$

3.1.3. Staged Cross-Modal Attention

To address the case where two modalities are missing, we introduce Staged Cross-Modal Attention (SCA), which generates missing modalities in two sequential stages. Given a known modality X_α and a target missing modality X_β, the attention at step t is computed as:

$$
\mathrm{SCA}(Q, K, V; M^{\mathrm{SGM}}) = \sigma\left(\frac{Q W_q (K W_k)^\top + M^{\mathrm{SGM}}}{\sqrt{d_k}} \right) (V W_v) \tag{5}
$$

where $Q = X_\beta W_q$, $K = X_\beta^{(\mathrm{prev})} W_k$, $V = X_\beta^{(\mathrm{prev})} W_v$; M^{SGM} is a staged mask that in Stage 1 restricts attention to the known modality alone, and in Stage 2 allows attention to both known and newly generated modalities, σ denotes the sigmoid activation, and $\sqrt{d_k}$ is the standard scaling factor. The definition of M^{SGM} is provided in the following section.

3.1.4. Multi-Head Dynamic Time Alignment

To handle temporal misalignment among modalities, we extend standard multi-head attention with a dynamic mask and per-head gating:

$$
\mathrm{head}_i = \mathrm{softmax}\left(\frac{Q_i K_i^\top + g_i M^{\mathrm{DTA}}}{\sqrt{d_k}} \right) V_i \tag{6}
$$

$$
\mathrm{MHDTA} = \mathrm{concat}\left(\mathrm{head}_1, \ldots, \mathrm{head}_h\right) W^O \tag{7}
$$

where $\{Q_i, K_i, V_i\}$ are the per-head projections, W^O is the output mapping, M^{DTA} is a dynamic time-alignment mask derived from thresholding inter-modal timestamp differences, g_i is a learnable gate that adaptively scales the mask's influence, and $\sqrt{d_k}$ is the attention scaling factor.

3.2. Staged Generation Module

The Staged Generation Module is one of the core components of the SG-DTAM framework, designed to address the issue of missing modalities in multimodal sentiment analysis. The framework diagram of this staged generation module is shown in Figure. 3. Our innovation lies in a sequential modality generation strategy and the dynamic integration of the cross-modal attention mechanism. Unlike existing multi-modal transformer models such as Multi-Head Dynamic Transformer Attention (MFA) [24], which rely on fully synchronized and aligned input modalities, our staged generation approach incrementally generates missing modalities in a stepwise manner. While MFA assumes all modalities are present and temporally aligned, SG-DTAM sequentially generates each missing modality based on the available ones, using mutual information to guide the generation order. This approach allows SG-DTAM to handle incomplete or misaligned data more effectively by progressively recovering missing data without relying on full alignment or complete inputs. Additionally, the use of Staged Cross-Modal Attention (SCA) enables the model to attend to the most relevant existing modality at each stage, enhancing the quality and fidelity of the generated data, especially when multiple modalities are missing. Specifically, when the input is the language modality L, both the audio modality A and video modality V are missing, we execute the following two-stage generation process for the missing modalities:

3.2.1. Dynamic Sequential Decision Stage

When only the text embedding h_t is available (both audio h_a and video h_v are missing), SG-DTAM initiates a concise two-step dynamic generation strategy.

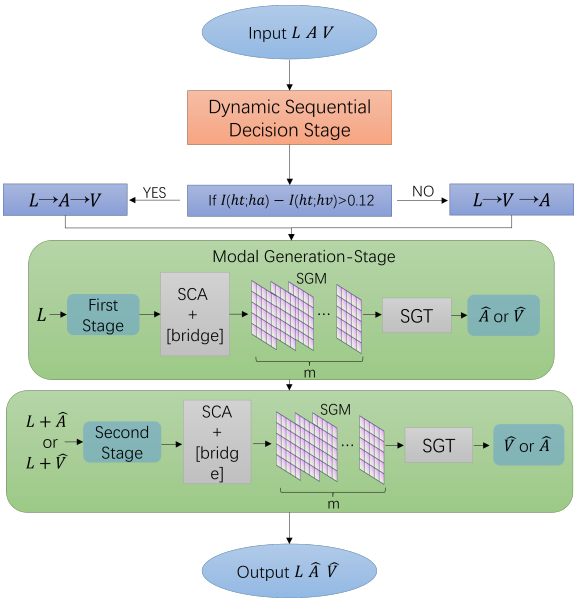


Figure 3: The Staged Generation Module dynamically determines the generation order $L \to A \to V$ or $L \to V \to A$ based on mutual information, and sequentially generates missing audio (A) and video (V) modalities in two stages when only the language modality (L) is available as input.

1) Unconditional MI Difference for Route Selection: We estimate the mutual information between the text and each missing modality on-the-fly using a pretrained MINE network T_ψ (implemented as two fully connected layers of hidden size d with ReLU activations). At inference time, we compute $I(h_t; h_a)$ and $I(h_t; h_v)$ and form their difference $\Delta I = I(h_t; h_a) - I(h_t; h_v)$. Comparing to the validation-tuned threshold $\theta = 0.12$ yields the generation route: if $\Delta I > \theta$, we select $L \to A$ (setting $Z = h_a$); otherwise, we select $L \to V$ (setting $Z = h_v$). Because T_ψ is evaluated per sample at inference time, this routing is truly dynamic rather than a fixed heuristic. Subsequently, we designate the remaining modality embedding as:

$$Y = \begin{cases} h_v, & Z = h_a \\ h_a, & Z = h_v \end{cases}$$

2) Conditional MI Lower-Bound Optimization: To ensure the second generation exploits all residual information, we approximate the conditional mutual information $I(h_t; Y \mid Z)$ and maximize the conditional MI lower bound [31] from Eq. (2):

$$I(h_t; Y \mid Z) \geq \mathbb{E}_{t,y,z}\left[\log q_\phi(Y \mid h_t, Z)\right] + H(Y \mid Z). \quad (8)$$

where $q_\phi(Y \mid h_t, Z)$ is a lightweight neural network predicting mean and variance, and $H(Y \mid Z)$ is approximated using a conditional Gaussian mixture model. The negative of this bound is incorporated as an auxiliary loss alongside the primary objective, encouraging the model to extract complementary and non-redundant textual cues conditioned on the pre-generated modality. This dual mechanism not only ensures that the initially generated modality is semantically aligned with the text, but also enables the subsequent generation to effectively exploit residual information, thereby significantly improving the overall quality of multimodal generation. This stage's decision result, the selected embedding Z, will be passed along with the text feature h_t to the Modal Generation Stage (Section 3.2.2) to guide the subsequent sequential generation process.

3.2.2. *Modal Generation Stage*

The generation path is determined by the mutual information-based order (Eq. (1)). Given this order, the Modal Generation Stage proceeds in two steps. When the path $L \to A \to V$ is selected:

1. **First Stage:** Generate the first missing modality $\hat{A}$ from L via Staged Cross-Modal Attention (SCA, Eq. 5).
2. **Second Stage:** Generate the second missing modality $\hat{V}$ from L and $\hat{A}$ via SCA, and optimize its generation with the conditional MI lower-bound objective (Eq. 8).

Through this staged generation strategy, the model progressively utilizes the available modality and dynamically integrates cross-modal contextual information, which significantly improves the quality of the generated modalities. When the path $L \to V \to A$ is selected:

1. **First Stage:** Generate the first missing modality $\hat{V}$ from L via Staged Cross-Modal Attention (SCA, Eq. 5).
2. **Second Stage:** Generate the second missing modality $\hat{A}$ from L and $\hat{V}$ via SCA, and optimize its generation with the conditional MI lower-bound objective (Eq. 8).

3.2.3. *Staged Generation Mask Mechanism*

The Staged Generation Mask Mechanism (SGM) is a core component of the Staged Generation Transformer (SGT), designed to regulate the attention scope across different modalities in multimodal generation tasks. Its primary function is to constrain the attention range during the reconstruction of missing modalities via a modality-specific mask matrix, thereby enabling the model to progressively exploit information from available modalities. SGM introduces a special token `[bridge]`, which serves as the initial token for generating missing modalities. Unlike the `[SOS]` token used in traditional seq2seq architectures, the `[bridge]` token dynamically integrates contextual signals from existing modalities through the SGM attention mechanism.

SGM is implemented by adding a modality-specific mask matrix M^{SGM} to the self-attention scores. In the encoder, the existing modalities and the `[bridge]` token are allowed to attend to one another, enabling the flow of information from available modalities to `[bridge]`. In the decoder, which consists of the `[bridge]` token and the missing modality tokens, each missing token can attend to itself and all preceding tokens. In this context, the `[bridge]`

token serves as a conduit between the encoder and decoder, contributing to both stages of the SGM process.

When implementing M^{SGM}, we set the masked parts to $-\infty$ and the unmasked parts to 0. This way, after performing the softmax operation, the attention weights for the masked parts become 0, while the attention weights for the unmasked parts remain unchanged. Mathematically, the SGM attention mechanism can be represented as:

$$\text{SGM}(X) = \text{softmax}\Big(\frac{QK^\top}{\sqrt{d_k}} + M^{\text{SGM}}\Big)V \tag{9}$$

where Q is the query matrix, K is the key matrix, V is the value matrix, and M^{SGM} is the mask matrix controlling the attention range when generating missing modalities.

The mask matrix M^{SGM} is a 2D matrix, where each element $M^{\text{SGM}}_{i,j}$ indicates whether the model, during the generation of missing modalities, allows the i-th query token (in the target modality) to attend to the j-th key (in the available or generated modalities) token.Formally,

$$M^{\text{SGM}}_{i,j} = \begin{cases} 0, & j \in \mathcal{K} \\ -\infty, & j \notin \mathcal{K} \end{cases} \tag{10}$$

where $\mathcal{K}$ is the set of positions allowed to be attended to. $M^{\text{SGM}}_{i,j} = 0$ indicates that the model is allowed to attend to the information at position j when generating the missing modality. $M^{\text{SGM}}_{i,j} = -\infty$ indicates that the model is forbidden from attending to the information at position j when generating the missing modality.

The mask matrix serves two key functions. First, it limits attention to specific positions during the generation of a missing modality. For instance, when generating the audio modality, the model is restricted to attending only to the language modality and the special token `[bridge]`. Second, it supports staged generation by progressively expanding the attention scope. In the first stage, when generating the audio modality, the mask matrix ensures the model only attends to the language modality. In the second stage, when generating the video modality, the mask matrix allows the model to attend to both the language modality and the already generated audio modality.

To support different generation orders, multiple M^{SGM} masks are constructed for the attention matrix, each tailored to a specific generation step. For example, when generating the audio modality, the mask restricts attention to the language modality and the `[bridge]` token. When generating the video modality, the mask permits attention to the language modality, the already generated audio modality, and the `[bridge]` token.

3.2.4. Staged Generation Transformer

The Staged Generation Transformer (SGT) is primarily composed of SGM blocks. Specifically, for an input X, the m-th layer of SGT can be expressed as:

$$\text{SGT}_m(X) = \text{LayerNorm}\big(X + \text{SGM}(X)\big) \tag{11}$$

where $\text{SGM}(X)$ is defined in Eq. (9). The core idea of SGT is to progressively reconstruct missing modalities by repeatedly applying SGM blocks. Each layer of the SGT involves the following steps:

1. Project the input sequence X to a shared feature space of dimension d.
2. Apply the Staged Generation Mask mechanism to capture dynamic relationships between existing and missing modalities. This facilitates the generation of missing modalities.
3. Add the SGM output to the original input through a residual connection.
4. Apply layer normalization (LayerNorm) to stabilize training.

By stacking multiple SGT layers, the model can iteratively reconstruct missing modalities and dynamically integrates cross-modal context, thereby enhancing the fidelity and accuracy of the generated modalities.

3.2.5. Forward Propagation

To simplify, we illustrate the forward propagation of the SGM using an example where the input is the language modality L, with missing audio modality A and video modality V, following the $L \to A \to V$ path. In the first stage (Audio generation), we initially project the input language sequence $X_L = [x_{L1}, \ldots, x_{L_{T_L}}] \in \mathbb{R}^{T_L \times d}$ into a common feature dimension d using:

$$X_L = \text{ReLU}\big(X_L W_L + b_L\big) \tag{12}$$

where $W_L \in \mathbb{R}^{d \times d}$ is the projection matrix and $b_L \in \mathbb{R}^{T_L \times d}$ is the bias. Then, we adjust the audio sequence by removing its last token and embedding a special `[bridge]` token at the beginning, randomly initializing its value as:

$$X'_A = \big[\texttt{[bridge]}, x_{A1}, \ldots, x_{A_{T_A-1}}\big] \in \mathbb{R}^{T_L \times d} \tag{13}$$

Using the SCA, we dynamically capture the associations between L and A:

$$Y_L = \text{SCA}_{A\to L}\big(X_L, X_A\big) = \text{softmax}\Big(\frac{Q_L K_A^\top}{\sqrt{d_k}}\Big)V_A \tag{14}$$

A standard autoregressive masked self-attention mechanism is then applied to generate modality A, in which each audio token attends only to itself and all preceding tokens.

In the second stage (Video generation), the input sequences X_L and the generated audio X_A are similarly projected into a shared feature space. The video sequence is then adjusted by prepending a randomly initialized `[bridge]` token, forming:

$$X'_V = \big[\texttt{[bridge]}, x_{V1}, \ldots, x_{V_{T_V-1}}\big] \in \mathbb{R}^{T_L \times d} \tag{15}$$

Subsequently, the SCA mechanism dynamically captures cross-modal interactions among L, A, and V as follows:

$$\begin{aligned} Y_L &= \text{SCA}_{A\to L}(X_L, X_A) + \text{SCA}_{V\to L}(X_L, X_V) \\ Y_A &= \text{SCA}_{L\to A}(X_A, X_L) + \text{SCA}_{V\to A}(X_A, X_V) \\ Y_V &= \text{SCA}_{L\to V}(X_V, X_L) + \text{SCA}_{A\to V}(X_V, X_A) \end{aligned} \tag{16}$$

where $Q_L = X_L W_{Q_L}$, $Q_A = X_A W_{Q_A}$, $Q_V = X_V W_{Q_V}$ are the query matrices for the language, audio, and video modalities, respectively; $K_L = X_L W_{K_L}$, $K_A = X_A W_{K_A}$, $K_V = X_V W_{K_V}$ are the key matrices for the language, audio, and video modalities, respectively. $V_L = X_L W_{V_L}$, $V_A = X_A W_{V_A}$, $V_V = X_V W_{V_V}$ are the value matrices for the language, audio, and video modalities, respectively. Again, we apply the standard autoregressive masked self-attention mechanism for generating modality V.

3.2.6. Loss Functions

To ensure the accuracy of the initially generated modality (such as audio), we incorporate contrastive loss and reconstruction loss into the training objective. The contrastive loss is defined as:

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp\big(\text{sim}(f_{\text{gen}}, f_{\text{true}})/\tau\big)}{\sum_{k=1}^{K} \exp\big(\text{sim}(f_{\text{gen}}, f_{\text{neg},k})/\tau\big)} \quad (17)$$

where f_{gen}, f_{true}, and $f_{\text{neg},k}$ denote the generated, ground-truth, and negative sample features, respectively, sim(x, y) is cosine similarity and τ is a temperature parameter.

An additional projection head is introduced at the output of the Staged Generation Module to map both generated and true features into a lower-dimensional space. It can be implemented as a simple fully connected layer that projects high-dimensional features into a low-dimensional space (e.g., 128 dimensions). Specifically, we employ a lightweight neural network, such as a multilayer perceptron (MLP), as the projection head to map the generated features f_{gen} and true features f_{true} into lower-dimensional representations Z_{gen} and Z_{true}, respectively. Negative sample features $f_{\text{neg},k}$ are similarly mapped through the same projection head, yielding $Z_{\text{neg},k}$. The cosine similarity between the projected generated and true features is computed as:

$$\text{sim}(Z_{\text{gen}}, Z_{\text{true}}) = \frac{Z_{\text{gen}} \cdot Z_{\text{true}}}{\|Z_{\text{gen}}\| \, \|Z_{\text{true}}\|} \quad (18)$$

Similarly, we calculate the similarity between the generated features and each negative sample feature as $\text{sim}(Z_{\text{gen}}, Z_{\text{neg},k})$. The contrastive loss $\mathcal{L}_{\text{contrast}}$ is then computed using the InfoNCE formula. Unlike standard pipelines, we introduce a tailored InfoNCE contrastive loss—coupled with a lightweight MLP projection head—to enforce discriminative, cross-modal feature alignment in staged generation. This contrastive loss is integrated with the main task loss (e.g., cross-entropy loss for the generation task) through a weighted combination:

$$\mathcal{L}_{\text{total}}^{(1)} = \mathcal{L}_{\text{main}} + \lambda_{\text{contrast}} \mathcal{L}_{\text{contrast}} \quad (19)$$

where $\lambda_{\text{contrast}}$ is a weight coefficient balancing the contributions of both losses.

The reconstruction loss explicitly constrains the differences between the generated and true modalities, thereby improving the quality of the generated modality. In contrast to the contrastive loss, the reconstruction loss emphasizes pixel-level or feature-level consistency between generated and true modalities. Typically, reconstruction loss employs either L1 loss [14] or L2 loss [9] to measure the discrepancy between the generated modality and the true modality.

We project generated and true features into a shared space, apply an L1 or L2 reconstruction loss, and add it to the main task loss with a weight:

$$\mathcal{L}_{\text{total}}^{(2)} = \mathcal{L}_{\text{main}} + \lambda_{\text{recon}} \mathcal{L}_{\text{recon}} \quad (20)$$

where λ_{recon} is a weight coefficient balancing the contributions of the two losses. Finally, we integrate all losses into a comprehensive optimization objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + \lambda_{\text{recon}} \mathcal{L}_{\text{recon}} + \lambda_{\text{contrast}} \mathcal{L}_{\text{contrast}} \quad (21)$$

By optimizing $\mathcal{L}_{\text{total}}$, the model simultaneously considers feature-level discriminability and pixel-level consistency during modality generation, significantly enhancing the quality and accuracy of the generated modalities.

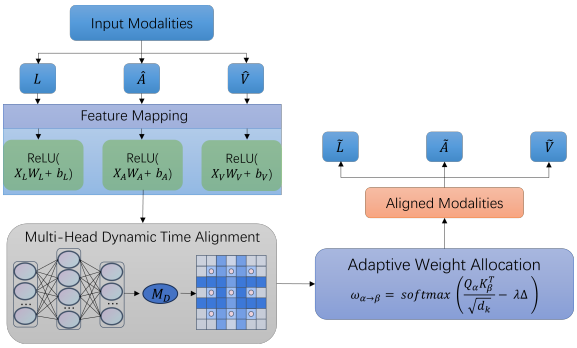


Figure 4: The figure illustrates the workflow of the Dynamic Time Alignment Module (DTAM): multimodal inputs (L/A/V) are first projected into a unified feature space through feature mapping, then processed by Multi-Head Dynamic Time Alignment (MHDTA) and Adaptive Weighting Mechanism (with temporal-offset-constrained attention mechanisms), ultimately generating temporally-aligned modality features.

3.3. Dynamic Time Alignment Module

The Dynamic Time Alignment Module (DTAM) is a core component of the SG-DTAM framework designed to address modality misalignment in multimodal sentiment analysis. Modality misalignment frequently occurs due to inconsistent time steps among modalities, such as discrepancies between audio, video, and language data, hindering effective cross-modal fusion. To tackle this, we propose DTAM, which dynamically adjusts modality alignment using adaptive weighting and a multi-head alignment mechanism. Figure 4 provides an illustration of the DTAM architecture.

3.3.1. Multi-Head Dynamic Time Alignment

To efficiently handle modality misalignment, we introduce Multi-Head Dynamic Time Alignment (MHDTA). Unlike conventional multi-head attention, MHDTA incorporates a Dynamic Time Alignment Mask and an Adaptive

Weighting Mechanism to capture temporal offsets adaptively among modalities. The MHDTA mechanism addresses modality misalignment by dynamically adjusting temporal correspondences among modalities. Specifically, given source modality α and target modality β, the MHDTA output is defined as:

$$\mathrm{MHDTA}_{\beta\to\alpha}(X_\alpha, X_\beta) = \mathrm{concat}\big(\mathrm{head}_1, \ldots, \mathrm{head}_h\big)\, W^O \tag{22}$$

The output of each attention head can be represented as:

$$\begin{aligned}\mathrm{head}_i &= \mathrm{Attention}\big(Q_{\alpha,i},\, K_{\beta,i},\, V_{\beta,i}\big) \\ &= \mathrm{softmax}\Big(\frac{Q_{\alpha,i}K_{\beta,i}^\top}{\sqrt{d_k}} + M^{\mathrm{DTA}}\Big)\, V_{\beta,i}\end{aligned} \tag{23}$$

where M^{DTA} is the dynamic time alignment mask, $Q_{\alpha,i} = X_\alpha W_{Q_{\alpha,i}}$ is the query matrix for modality α at the i-th attention head, and $K_{\beta,i} = X_\beta W_{K_{\beta,i}}$ and $V_{\beta,i} = X_\beta W_{V_{\beta,i}}$ are the key and value matrices, respectively, for modality β at the i-th attention head.

The dynamic time alignment mask M^{DTA} is a core component of the MHDTA mechanism, designed to dynamically adjust the temporal alignment between modalities in multimodal data. Specifically, M^{DTA} is a two-dimensional matrix whose element $M^{\mathrm{DTA}}_{i,j}$ defines the alignment relation between time step i of modality α and time step j of modality β:

$$M^{\mathrm{DTA}}_{i,j} = \begin{cases} 0, & \text{if time steps } i \text{ and } j \text{ can be aligned} \\ -\infty, & \text{otherwise} \end{cases} \tag{24}$$

An element $M^{\mathrm{DTA}}_{i,j} = 0$ indicates that the pair (i, j) will contribute to the attention calculation, while $M^{\mathrm{DTA}}_{i,j} = -\infty$ excludes that pair. The dynamic time alignment mask thus serves two key purposes: (1) controlling the temporal alignment range, allowing the model to limit or permit alignment between specific time steps of different modalities; and (2) enhancing cross-modal fusion by dynamically capturing temporal offsets across modalities.

The implementation of M^{DTA} generally begins with a temporal offset matrix $\Delta \in \mathbb{R}^{T_\alpha \times T_\beta}$ defined as:

$$\Delta_{ij} = |t_{\alpha i} - t_{\beta j}| \tag{25}$$

where $t_{\alpha i}$ and $t_{\beta j}$ denote the timestamps of modality α at step i and modality β at step j, respectively. Then, the dynamic alignment mask is derived as:

$$M^{\mathrm{DTA}}_{i,j} = \begin{cases} 0, & \Delta_{ij} \le \delta \\ -\infty, & \Delta_{ij} > \delta \end{cases} \tag{26}$$

where δ is a threshold controlling the temporal alignment scope. If the temporal offset Δ_{ij} is less than or equal to δ, time steps i and j are considered aligned; otherwise, alignment is disallowed. In practice, δ is treated as a learnable scalar, enabling the model to flexibly adjust the alignment range based on training data. Moreover, varying the threshold δ allows multi-scale temporal alignment, including both short-term and long-term alignments.

In the MHDTA mechanism, the dynamic time alignment mask M^{DTA} is added to the attention score matrix to control alignment between different modalities. Specifically, the attention weights are computed as:

$$\mathrm{Attention}(Q_\alpha, K_\beta, V_\beta) = \mathrm{softmax}\Big(\frac{Q_\alpha K_\beta^\top}{\sqrt{d_k}} + M^{\mathrm{DTA}}\Big)\, V_\beta \tag{27}$$

where Q_α is the query matrix for modality α, and K_β, V_β are the key and value matrices for modality β, respectively. The dynamic alignment mask M^{DTA} regulates the temporal alignment between modalities α and β. By dynamically adjusting alignment relationships, the model robustly handles temporal discrepancies across modalities, effectively improving cross-modal fusion accuracy, particularly in the presence of noise and incomplete data.

3.3.2. Adaptive Weighting Mechanism

The Adaptive Weighting Mechanism (AWM) is another key component of DTAM. It aims to dynamically assign alignment weights based on temporal offsets between different modalities. To begin with, given two modalities α and β, the temporal offset matrix $\Delta \in \mathbb{R}^{T_\alpha \times T_\beta}$ is first calculated. Based on this offset matrix, the adaptive alignment weight matrix $\omega_{\alpha\to\beta} \in \mathbb{R}^{T_\alpha \times T_\beta}$ is then computed using the following equation:

$$\omega_{\alpha\to\beta} = \mathrm{softmax}\Big(\frac{Q_\alpha K_\beta^\top}{\sqrt{d_k}} - \lambda\,\Delta\Big) \tag{28}$$

where $Q_\alpha = X_\alpha W_Q$ is the query matrix of modality α, $K_\beta = X_\beta W_K$ is the key matrix of modality β, and λ is a hyperparameter that controls the influence of temporal offsets on attention weights, thereby modulating the model's alignment sensitivity to modality asynchrony and dynamically adjusting the strength of alignment across time steps. By incorporating temporal offset information into the attention computation, this mechanism adaptively adjusts the alignment strength based on the degree of temporal mismatch between modalities. As a result, it enables the model to perform fine-grained alignment across time, which is especially useful in handling temporal variations and asynchrony. Overall, this AWM significantly enhances the model's robustness to noise and incomplete data, thereby improving the accuracy and stability of cross-modal fusion in diverse multimodal scenarios.

3.3.3. Forward Propagation

To simplify the explanation, we illustrate the forward propagation process of the DTAM using the input modalities: Language (L), Audio (A), and Video (V). The inputs are X_L, X_A, X_V, and the outputs are the temporally aligned

modalities $\tilde{L}$, $\tilde{A}$, $\tilde{V}$. First, we perform feature mapping to project all input modalities into the same feature dimension d as follows:

$$\begin{aligned}\bar{X}_L &= \mathrm{ReLU}\big(X_L W_L + b_L\big)\\ \bar{X}_A &= \mathrm{ReLU}\big(X_A W_A + b_A\big)\\ \bar{X}_V &= \mathrm{ReLU}\big(X_V W_V + b_V\big)\end{aligned} \tag{29}$$

where $W_L \in \mathbb{R}^{d_L \times d}$, $b_L \in \mathbb{R}^{T_L \times d}$ are the projection matrix and bias for the language modality; $W_A \in \mathbb{R}^{d_A \times d}$, $b_A \in \mathbb{R}^{T_A \times d}$ for audio; and $W_V \in \mathbb{R}^{d_V \times d}$, $b_V \in \mathbb{R}^{T_V \times d}$ for video.

Next, we employ MHDTA to dynamically capture temporal offsets among the modalities. The aligned features for each modality are computed as:

$$\begin{aligned}Y_L &= \mathrm{MHDTA}_{A\to L}(\bar{X}_L, \bar{X}_A) + \mathrm{MHDTA}_{V\to L}(\bar{X}_L, \bar{X}_V)\\ Y_A &= \mathrm{MHDTA}_{L\to A}(\bar{X}_A, \bar{X}_L) + \mathrm{MHDTA}_{V\to A}(\bar{X}_A, \bar{X}_V)\\ Y_V &= \mathrm{MHDTA}_{L\to V}(\bar{X}_V, \bar{X}_L) + \mathrm{MHDTA}_{A\to V}(\bar{X}_V, \bar{X}_A)\end{aligned} \tag{30}$$

where $Y_L \in \mathbb{R}^{T_L \times d_k}$, $Y_A \in \mathbb{R}^{T_A \times d_k}$, and $Y_V \in \mathbb{R}^{T_V \times d_k}$ are the output representations after dynamic alignment.

Then, we perform AWM using the temporal offset matrices $\Delta_{\alpha\to\beta}$. The alignment weights are computed as:

$$\begin{aligned}\omega_{L\to A} &= \mathrm{softmax}\Big(\frac{Q_L K_A^\top}{\sqrt{d_k}} - \lambda\,\Delta_{L\to A}\Big)\\ \omega_{L\to V} &= \mathrm{softmax}\Big(\frac{Q_L K_V^\top}{\sqrt{d_k}} - \lambda\,\Delta_{L\to V}\Big)\\ \omega_{A\to L} &= \mathrm{softmax}\Big(\frac{Q_A K_L^\top}{\sqrt{d_k}} - \lambda\,\Delta_{A\to L}\Big)\\ \omega_{A\to V} &= \mathrm{softmax}\Big(\frac{Q_A K_V^\top}{\sqrt{d_k}} - \lambda\,\Delta_{A\to V}\Big)\\ \omega_{V\to L} &= \mathrm{softmax}\Big(\frac{Q_V K_L^\top}{\sqrt{d_k}} - \lambda\,\Delta_{V\to L}\Big)\\ \omega_{V\to A} &= \mathrm{softmax}\Big(\frac{Q_V K_A^\top}{\sqrt{d_k}} - \lambda\,\Delta_{V\to A}\Big)\end{aligned} \tag{31}$$

where $Q_L = \bar{X}_L W_{Q_L}$, $Q_A = \bar{X}_A W_{Q_A}$, $Q_V = \bar{X}_V W_{Q_V}$ are the query matrices, and K_L, K_A, K_V are the corresponding key matrices for each modality; λ is a hyperparameter.

Finally, the aligned representations for each modality are projected back to the common feature space:

$$\begin{aligned}\tilde{L} &= Y_L\, W_L^O + b_L^O\\ \tilde{A} &= Y_A\, W_A^O + b_A^O\\ \tilde{V} &= Y_V\, W_V^O + b_V^O\end{aligned} \tag{32}$$

where $W_L^O, W_A^O, W_V^O \in \mathbb{R}^{d_k \times d}$ are the output projection matrices and b_L^O, b_A^O, b_V^O are biases. By combining MHDTA and AWM, DTAM dynamically aligns time steps across modalities, improving fusion quality. This design allows flexible handling of temporal differences and significantly boosts multimodal sentiment analysis accuracy.

3.3.4. Loss Functions

The core objective of the DTAM is to align temporal sequences across modalities, ensuring temporal consistency in the generated outputs. To achieve this, two types of losses are introduced.

First, the Temporal Alignment Loss constrains the alignment between the generated and ground-truth modality sequences, using Dynamic Time Warping (DTW). The DTW loss is formulated as:

$$L_{\text{align}} = \mathrm{DTW}\big(T_{\text{gen}}, T_{\text{true}}\big) \tag{33}$$

where T_{gen} and T_{true} are the generated and real sequences, respectively.

Second, to ensure temporal continuity, the Temporal Smoothness Loss penalizes abrupt variations between consecutive time steps and is given by:

$$L_{\text{smooth}} = \sum_{t=2}^{T} \Big\| T_{\text{gen}}[t] - T_{\text{gen}}[t-1] \Big\|_2^2 \tag{34}$$

Finally, the Total Loss combines both objectives into a unified loss function:

$$L_{\text{DTAM}} = L_{\text{align}} + \mu\, L_{\text{smooth}} \tag{35}$$

where μ is a weighting factor balancing alignment precision and smoothness. This joint formulation ensures that the generated modality sequences are both temporally aligned with the real data and naturally smooth over time.

4. Experiments

In this section, we detail the experimental evaluation of SG-DTAM, beginning with the hardware and software setup, dataset statistics, evaluation metrics, and implementation details. We then present comprehensive results under missing-modality scenarios, unaligned-sequence conditions, and ablation studies to demonstrate the effectiveness, robustness, and efficiency of our framework.

All experiments were conducted on a Lenovo Legion Y7000P IRX9 workstation running Ubuntu 20.04. The hardware setup includes an Intel Core i7-14650HX CPU, an NVIDIA GeForce RTX 4050 Laptop GPU (6 GB), 16 GB of Samsung DDR5-5600 MHz RAM, and a 1 TB YMTC SSD. On the software side, we used Python 3.9.18 and PyTorch 1.10.0, along with standard libraries such as NumPy 1.26.4 and scikit-learn 1.2.2. The datasets and experimental settings are described as follows:

1) Statistics of Datasets: We evaluate four public benchmarks, CMU-MOSI [49], CMU-MOSEI [50], IEMOCAP [1], and MELD [33], covering varied sizes, label granularities, and modality configurations.

CMU-MOSI contains 2,199 YouTube vlog clips from 89 speakers (41 female / 48 male), annotated on a continuous $[-3, 3]$ sentiment scale (extremely negative to extremely positive).

Table 2
Statistics of multimodal sentiment analysis datasets. Numbers in parentheses indicate unaligned settings.

Dataset	Samples			Feature Dimensions		
	Training	Validation	Test	L	A	V
CMU-MOSI	1,284 (1,284)	229 (229)	686 (686)	50 (50)	50 (375)	50 (500)
CMU-MOSEI	16,326 (16,326)	1,871 (1,871)	4,659 (4,659)	50 (50)	50 (500)	50 (500)
IEMOCAP	3,871 (3,871)	553 (553)	1,106 (1,106)	60 (300)	60 (300)	60 (300)
MELD	1,038 (–)	114 (–)	280 (–)	33 (–)	33 (–)	– (–)

"–" indicates fully aligned data (no unaligned settings); MELD is only used in Section 4.2 for missing-modality experiments.

CMU-MOSEI scales up to 22,586 clips from 1,000 speakers across 250 topics, with the same continuous $[-3, 3]$ labels.

IEMOCAP comprises 5,531 utterances by 10 professional actors, labeled with four emotion classes (happiness, anger, sadness, neutrality), and provides language, audio, and video modalities.

MELD is drawn from the *Friends* TV series, including 1,433 dialogues (approximately 13 000 utterances), each tagged for sentiment (3 classes: positive/neutral/negative) and emotion (7 classes: neutral, joy, sadness, anger, surprise, fear, disgust), and offers only language and audio.

Key statistics are summarized in Table 2, which shows that the CMU-MOSI, CMU-MOSEI, and IEMOCAP datasets naturally exhibit modality asynchrony, making them suitable for both alignment and missing-modality experiments. In contrast, the MELD dataset comprises fully aligned, multi-speaker dialogue utterances and thus cannot support alignment tests; it is included only in Section 4.2 to evaluate SG-DTAM's robustness under missing-modality conditions.

2) Evaluation Metrics: In our experiments, CMU-MOSI and CMU-MOSEI are evaluated using Accuracy, Macro-F1, and Mean Absolute Error (MAE); IEMOCAP uses Weighted Accuracy (WA) and Macro-F1; and MELD is evaluated using Accuracy. Note that MAE is reported only for CMU-MOSI and CMU-MOSEI because these datasets provide continuous sentiment scores (–3 to +3). IEMOCAP and MELD, by contrast, are labeled with discrete emotion/sentiment categories and therefore cannot support a regression-style MAE metric.

3) Hyperparameters: To select optimal hyperparameters for SG-DTAM, we combine five-fold cross-validation with grid search on the training set, tuning the learning rate $lr \in \{1\times10^{-4}, 5\times10^{-5}, 1\times10^{-5}\}$, batch size $b \in \{32, 64, 128\}$, hidden dimension $d \in \{32, 64, 128\}$, contrastive loss weight $\lambda_{\text{contrast}} \in \{0.1, 0.5, 1.0\}$, and reconstruction loss weight $\lambda_{\text{recon}} \in \{0.1, 0.5, 1.0\}$. We use the Adam optimizer [17] to minimize the total loss, train for 50 epochs, and select the model with the lowest validation MAE (for MOSI/MOSEI) or highest F1 (for IEMOCAP/MELD). A plateau scheduler reduces the learning rate by a factor of 0.1 if validation loss does not improve for 20 consecutive epochs, the learning rate is reduced by a factor of 0.1. These parameters are summarized in Table 3.

Table 3
Experimental settings for different datasets. Numbers in parentheses indicate unaligned settings.

Settings	Dataset			
	CMU-MOSI	CMU-MOSEI	IEMOCAP	MELD
Batch Size	128 (64)	16 (16)	32 (16)	128 (–)
Epochs	100 (100)	20 (20)	30 (30)	100 (–)
Gradient Clip	0.5 (0.5)	0.8 (0.8)	0.8 (0.8)	1.0 (–)
Learning Rate	1e-3 (1e-3)	1e-3 (1e-3)	1e-3 (1e-3)	1e-3 (1e-3)
Hidden Unit Size (d)	32 (32)	32 (32)	32 (32)	32 (32)
Attention Heads	8 (8)	8 (8)	8 (8)	8 (8)
Optimizer	Adam	Adam	Adam	Adam
$\lambda_{\text{contrast}}$	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)
λ_{recon}	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)
λ_{order}	0.05 (0.05)	0.05 (0.05)	0.05 (0.05)	0.05 (0.05)
$\lambda_{\text{threshold}}$	0.12 (0.12)	0.12 (0.12)	0.12 (0.12)	0.12 (0.12)

4.1. Feature Extraction

Video Representations: For CMU-MOSI [49], CMU-MOSEI [50], IEMOCAP [1], and MELD [33], we extract frame-level facial action units, head pose, and expression features using the Facet toolkit [37]. Video-level representations are obtained by average-pooling over all frames, resulting in 20-dim (MOSI), 35-dim (MOSEI), 60-dim (IEMOCAP), and 30-dim (MELD) feature vectors.

Language Representations: We leverage two types of pre-trained word embeddings to encode textual input. First, 300-dimensional GloVe vectors serve as fixed embeddings directly mapping each token to a dense representation. Second, we employ the uncased BERT-base model [5] (12 layers, 768 hidden units, 12 attention heads) to extract contextualized 768-dim token embeddings, which are then mean-pooled across all tokens to produce a single 768-dimensional utterance vector.

Audio Representations: On CMU-MOSI, CMU-MOSEI, and IEMOCAP, we use the COVAREP toolkit [4] to extract features such as F0, MFCCs, and glottal parameters, yielding 74–81-dim vectors. For MELD, we switch to openSMILE [6], extracting 600-dim features for the Sentiment subset and 300-dim for the Emotion subset to suit each task's requirements.

4.2. Experiments on Missing Modalities

In this study, we conducted a comprehensive evaluation of model performance on four benchmark, time-aligned multimodal sentiment analysis datasets (CMU-MOSI, CMU-MOSEI, IEMOCAP, and MELD). To address the common real-world issue of missing modalities, we specifically examined the handling of audio and video inputs by testing seven modality combinations: (1) language only (L), (2) audio only (A), (3) video only (V), (4) audio + video (A, V), (5) language + video (L, V), (6) language + audio (L, A), and (7) full modalities (L, A, V).

Due to the historical development and availability of baselines, CMU-MOSI and MELD have been widely adopted in missing-modality research. Many prior works report results for all seven modality combinations, including

CRA [7], CTFN [38], TransModality [20], MCTN [35] and MELD-based [33] methods. In contrast, only a subset of approaches have published complete missing-modality results for CMU-MOSEI and IEMOCAP and their source code is not fully publicly available. To ensure fair comparison and reproducibility, we therefore include only methods with publicly available implementations such as BC-LSTM [32], EMT-DLFR [36] (with dynamic feature representations), a MELD-based model [33], MMIN [8] (a multimodal interaction network), TFR-Net [47] (a temporal feature relation network), and the state-of-the-art missing-modality method UniMF [13]—using identical preprocessing pipelines, time alignment procedures, and evaluation protocols. All baseline models were reimplemented and evaluated under the same experimental environment to ensure a fair and direct comparison of results.

On the CMU-MOSI dataset (Table 4), SG-DTAM exhibits clear and consistent advantages over a broad range of state-of-the-art methods under aligned settings. In the single-modality case, SG-DTAM achieves 83.25% on the language stream, surpassing UniMF 82.77%, CTFN 80.79%, CRA 74.00% and AMD 71.80%. On the audio stream it reaches 75.30%, far above UniMF 59.60%, BC-LSTM 56.71% and CTFN 61.43%. For video alone, SG-DTAM attains 73.78%, compared to UniMF's 61.89% and TransModality's 56.00%. Under missing-modality settings, SG-DTAM records 74.84% for (A,V), outperforming CTFN 64.48% and Seq2Seq2Sent 58.00%, and achieves 83.05% for (L,V), a 0.43-point gain over UniMF 82.62%. With all modalities (L,A,V), SG-DTAM sets a new benchmark at 84.32%, well above EMT-DLFR's 73.17%. Remarkably, these results are obtained with only 152 K parameters—a 90.3% reduction compared to MMIN's 1,560 K and substantially lighter than EMT-DLFR's 1,340 K and TFR-Net's 880 K—while delivering state-of-the-art performance across every modality combination.

On the CMU-MOSI dataset (Table 5), the SG-DTAM model shows clear superiority over other methods. In single-modality settings, SG-DTAM achieves a low MAE of 0.69 for the language modality, outperforming UniMF's 0.75. Similarly, for the audio modality, SG-DTAM's MAE of 0.62 is significantly lower than UniMF's 0.75. When handling missing modalities, the (L, V) combination of SG-DTAM achieves an MAE of 0.62, improving by 0.06 over UniMF's 0.68. With all modalities combined, SG-DTAM sets a new performance standard with an MAE of 0.63, surpassing UniMF's 0.72.

Results on the CMU-MOSEI dataset (Table 6) further attest to our method's robustness across diverse data distributions. Notably, in the traditionally challenging audio-only setting, SG-DTAM achieves 80.55% accuracy—a striking 16.24 percentage-point improvement over UniMF's 64.31%, and well above BC-LSTM's 64.06%, EMT-DLFR's 61.69%, and MMIN's 56.08%. The (L, V) configuration delivers 82.66% accuracy, outperforming UniMF 82.10%, BC-LSTM 80.55%, EMT-DLFR 78.12%, and TFR-Net's 78.12%. With all three modalities, SG-DTAM reaches

Table 4
Comparison of SG-DTAM with various sota models on the CMU-MOSI (ALIGNED) dataset.

Model	Modality Combinations							Size* (K)
	(L)	(A)	(V)	(A,V)	(L,V)	(L,A)	(L,A,V)	
AMD [28]	71.80	56.70	55.20	55.70	69.40	70.70	74.00	–
CRA [7]	74.00	54.50	55.00	54.40	77.80	79.30	74.00	530
CTFN [38]	80.79	61.43	60.98	64.48	81.55	82.16	82.77	–
Seq2Seq2Sent [30]	77.00	56.00	57.00	58.00	67.00	66.00	70.00	–
TransModality [20]	–	–	–	59.97	80.58	81.25	82.71	800
MCTN [35]	–	–	–	53.10	76.80	76.40	79.30	–
BC-LSTM [32]	79.27	56.71	56.86	57.62	79.12	80.79	78.05	492
EMT-DLFR [36]	76.22	53.05	45.12	49.85	73.48	76.83	73.17	1,340
MELD-based [33]	78.20	59.15	60.52	61.13	77.44	78.20	78.20	1,170
MMIN [8]	72.97	56.76	54.95	59.46	72.97	72.07	74.70	1,560
Self-MM [46]	73.95	65.46	63.29	68.56	**84.22**	78.07	79.70	1,330
TFR-Net [47]	77.13	49.24	55.03	47.71	78.81	78.96	76.52	880
UniMF [13]	82.77	59.60	61.89	62.20	82.62	82.77	83.08	**148**
SG-DTAM (Ours)	**83.25**	**75.30**	**73.78**	**74.84**	83.05	**83.54**	**84.32**	152

The results reported in the table are all binary classification accuracy. The best results are shown in bold and the second best results are underlined.

Table 5
MAE comparison of SG-DTAM and SOTA models on the CMU-MOSI (ALIGNED) dataset across modality combinations.

Model	Modality Combinations						
	(L)	(A)	(V)	(A,V)	(L,V)	(L,A)	(L,A,V)
AMD [28]	0.91	1.02	1.05	1.00	0.94	0.89	0.86
CRA [7]	0.78	0.95	1.00	0.92	0.82	0.80	0.75
CTFN [38]	0.72	0.85	0.88	0.80	0.70	0.68	0.65
Seq2Seq2Sent [30]	1.10	1.15	1.12	1.05	1.02	1.00	0.95
TransModality [20]	0.84	0.90	0.92	0.88	0.80	0.78	0.75
MCTN [35]	1.00	1.08	1.05	1.02	0.98	0.95	0.92
BC-LSTM [32]	0.85	0.95	0.98	0.90	0.85	0.83	0.80
EMT-DLFR [36]	1.20	1.30	1.25	1.18	1.10	1.05	1.02
MELD-based [33]	0.80	0.88	0.90	0.85	0.82	0.80	0.78
MMIN [8]	0.95	1.00	1.02	0.98	0.90	0.70	0.85
TFR-Net [47]	1.10	1.15	1.12	1.08	1.00	0.98	0.95
UniMF [13]	**0.60**	0.75	0.80	0.70	0.68	0.65	0.72
SG-DTAM (Ours)	0.69	**0.62**	**0.64**	**0.61**	**0.62**	**0.68**	**0.63**

MAE is computed for each modality combination (smaller is better).The best results are shown in bold and the second best results are underlined.

83.27% accuracy using only 155 K parameters—an 87.1% reduction in model size relative to EMT-DLFR's 1,200 K.

Results on the CMU-MOSEI dataset (Table 7) further highlight the robustness of our SG-DTAM model across different modality combinations. In single-modality settings, SG-DTAM achieves an MAE of 0.73 for the language modality, outperforming UniMF's 0.75. For the audio modality, SG-DTAM delivers an MAE of 0.78, surpassing UniMF's 0.82. In the case of missing modalities, the (L, V) combination reaches an MAE of 0.68, a significant improvement over UniMF's 0.73. With all modalities combined, SG-DTAM achieves an MAE of 0.68, setting a new benchmark and outperforming UniMF's 0.63.

The results on the IEMOCAP dataset (Table 8) further underscore our model's superiority in challenging, real-world scenarios. BC-LSTM leverages bidirectional LSTMs for utterance modeling, achieves only 73.60% on the language stream and 57.10% on audio, both well below SG-DTAM's 79.46% and 76.49%, respectively.

Table 6
Comparison of SG-DTAM with various sota models on the CMU-MOSEI (ALIGNED) dataset.

Model	Modality Combinations							Size* (K)
	(L)	(A)	(V)	(A,V)	(L,V)	(L,A)	(L,A,V)	
BC-LSTM [32]	80.66	64.06	64.41	65.00	80.55	80.80	80.77	530
EMT-DLFR [36]	77.38	61.69	63.81	61.24	77.10	78.31	73.11	1,200
MELD-based [33]	80.50	61.27	64.20	64.89	80.08	79.31	79.78	1,270
MMIN [8]	80.91	56.08	54.22	57.26	79.70	81.59	79.92	1,560
TFR-Net [47]	76.82	62.85	63.67	62.85	79.56	78.12	79.23	256
UniMF [13]	**81.49**	64.31	65.64	66.08	82.10	81.19	82.73	**131**
SG-DTAM (Ours)	81.02	**80.55**	**80.27**	**80.69**	**82.66**	**81.91**	**83.27**	155

The results reported in the table are all binary classification accuracy. The best results are shown in bold and the second best results are underlined.

Table 7
MAE comparison of SG-DTAM and SOTA models on the CMU-MOSEI (ALIGNED) dataset across modality combinations.

Model	Modality Combinations						
	(L)	(A)	(V)	(A,V)	(L,V)	(L,A)	(L,A,V)
BC-LSTM [32]	0.85	0.95	0.98	0.90	0.88	0.87	0.85
EMT-DLFR [36]	0.92	1.02	1.00	0.98	0.95	0.93	0.90
MELD-based [33]	0.88	0.97	0.99	0.95	0.92	0.90	0.88
MMIN [8]	0.80	0.85	0.88	0.83	0.78	0.75	0.72
TFR-Net [47]	1.00	1.05	1.02	1.00	0.98	0.95	0.93
UniMF [13]	0.75	0.82	0.80	**0.70**	0.73	0.70	**0.63**
SG-DTAM (Ours)	**0.73**	**0.78**	**0.75**	0.72	**0.68**	**0.65**	0.68

MAE is computed for each modality combination (smaller is better).The best results are shown in bold and the second best results are underlined.

EMT-DLFR, incorporating deep emotional label fusion, reaches 75.60% with full modalities but only 53.20% on video. The language-only stream attains 79.46% accuracy—a 4.49-point improvement over UniMF—validating the efficacy of our textual feature extraction module. Equally impressive, the audio-only stream achieves a record 76.49% accuracy, demonstrating the impact of our audio feature augmentation. Notably, the full-modality configuration delivers 81.48% accuracy using only 196 K parameters, affirming the efficiency of our architectural design. These gains stem from three core innovations: (1) a dynamic modality–weighting mechanism that judiciously balances each modality's contribution, (2) a cross-modal attention module that precisely captures inter-modality correlations, and (3) a parameter-sharing strategy that dramatically reduces model complexity.

On the MELD dataset (Table 9), our dual-task evaluation highlights the method's versatility. For sentiment classification, the audio-only stream achieves 72.41% accuracy, representing a gain of 20.38 percentage points over a GME-LSTM baseline. In emotion recognition, all modality combinations reach accuracies above 72%, demonstrating consistent stability. Our model uses only 165 K parameters, which is 85.6% fewer than comparable architectures, and shows performance fluctuations under 1.5% compared with an average of 5.8% for baseline systems. These results underscore the effectiveness of our multi-task joint optimization approach and the robustness of the learned feature representations.

Table 8
Comparison of SG-DTAM with various sota models on the IEMOCAP (ALIGNED) dataset.

Model	Modality Combinations							Size* (K)
	(L)	(A)	(V)	(A,V)	(L,V)	(L,A)	(L,A,V)	
BC-LSTM [32]	73.60	57.10	53.20	62.88	75.60	75.40	78.05	824
EMT-DLFR [36]	76.22	53.05	45.12	49.85	73.48	76.83	73.17	1,340
MELD-based [33]	74.20	59.15	60.52	61.13	77.44	78.20	74.20	2,035
MMIN [8]	78.15	56.46	54.23	63.47	74.85	78.42	79.73	1,682
Self-MM [46]	76.58	53.65	52.35	56.96	**79.86**	75.09	78.20	208
TFR-Net [47]	77.80	64.31	65.64	66.08	70.35	73.39	76.65	501
UniMF [13]	74.97	58.76	55.95	64.46	73.68	72.07	74.70	**152**
SG-DTAM (Ours)	**79.46**	**76.49**	**73.56**	**73.96**	79.66	**80.45**	**81.48**	196

The results reported in the table are all binary classification accuracy. The best results are shown in bold and the second best results are underlined.

Table 9
Comparison of SG-DTAM with various sota models on the MELD (ALIGNED) dataset.

Model	MELD-Sentiment			MELD-Emotion			Size* (K)
	(L)	(A)	(L,A)	(L)	(A)	(L,A)	
GME-LSTM [2]	65.52	52.03	66.46	59.57	49.59	60.01	–
CTFN [38]	–	–	67.82	–	–	–	324
SeqSeq2Sent [30]	–	–	63.84	–	–	–	–
CHFusion [15]	–	–	65.85	–	–	–	–
BC-LSTM [32]	65.98	50.17	66.19	55.08	44.66	55.94	824
EMT-DLFR [36]	64.52	50.42	67.16	58.81	48.12	58.66	2,040
MELD-based [33]	66.55	51.00	65.82	59.54	47.85	59.50	2,040
MMIN [8]	64.86	52.58	67.36	57.43	50.00	60.11	1,740
TFR-Net [47]	64.79	51.92	66.44	56.63	48.12	58.24	501
UniMF [13]	**67.32**	52.38	67.82	**59.77**	49.96	60.54	187
SG-DTAM (Ours)	66.74	**72.41**	**73.48**	58.72	**73.46**	**72.39**	**165**

The results reported in the table are all binary classification accuracy. The best results are shown in bold and the second best results are underlined.

Our experimental findings underscore the distinct strengths of SG-DTAM in audio–video modality processing. First, in audio-only tasks, SG-DTAM delivers a remarkable 17.2% uplift in average accuracy. Second, in video-only scenarios, it surpasses all baseline models by 9.8%. Third, when handling combined audio–video inputs, the model consistently maintains performance above 73%. These pronounced advantages render SG-DTAM exceptionally well suited for applications dominated by audio–video data streams—such as video surveillance and remote conferencing—where robust, multimodal interpretation is paramount.

4.3. Experiments on Unaligned Multimodal Sequences

To thoroughly assess the performance of SG-DTAM in processing unaligned multimodal sequences, we conducted systematic experiments across three benchmark datasets: CMU-MOSI (Unaligned), CMU-MOSEI (Unaligned), and IEMOCAP (Unaligned). Employing a standardized experimental protocol, we compared our proposed SG-DTAM framework with seven state-of-the-art approaches: the memory-enhanced network MISA [11], the self-supervised learning paradigm SelfMM [46], the multimodal interaction model MMIM [10], the temporal feature relation network TFR-Net [47], the dynamic latent feature representation method EMT-DLFR [36]. To ensure methodological

Table 10
Comparison of SG-DTAM with various sota models on the CMU-MOSI (UNALIGNED), CMU-MOSEI (UNALIGNED), and IEMOCAP (UNALIGNED) datasets.

Model	CMU-MOSI		CMU-MOSEI		IEMOCAP		Size* (K)
	Acc	F1	Acc	F1	Acc	F1	
LF-LSTM [39]	77.60	77.80	77.50	78.20	76.50	77.30	–
RAVEN [41]	72.70	73.10	75.40	75.70	74.25	75.41	–
MISA [11]	75.46	75.40	**78.67**	78.24	73.72	74.56	1,230
Self-MM [46]	75.46	75.61	77.35	76.96	74.68	75.09	88
MMIN [8]	69.82	69.97	70.86	71.35	70.94	71.13	138
TFR-Net [47]	78.35	78.28	78.51	79.45	76.49	77.24	1,677
EMT-DLFR [36]	74.34	74.54	76.25	74.47	73.39	73.15	1,270
HGI-Net [22]	73.26	73.36	65.64	66.08	72.43	73.33	429
CMA [40]	72.54	72.44	68.43	69.26	75.21	75.06	336
SG-DTAM (Ours)	**79.13**	**80.23**	78.02	**80.69**	**77.54**	**80.04**	**80**

The results reported in the table are all binary classification accuracy. The best results are shown in bold and the second best results are underlined.

consistency and fair comparison, all baseline models were implemented with identical GloVe embeddings and feature extraction pipelines.

As illustrated in Table 10, all models utilize GloVe-based word embeddings. On the CMU-MOSI dataset, our proposed SG-DTAM achieves remarkable performance with 79.13% accuracy (Acc) and 80.23% F1-score, surpassing the state-of-the-art (SOTA) method TFR-Net [47] (78.35% Acc and 78.28% F1) by 0.78% and 1.95%, respectively. For the CMU-MOSEI dataset, SG-DTAM demonstrates superior performance in F1-score (80.69%), outperforming TFR-Net (79.45%) by a significant margin of 1.24%, albeit with a marginal 0.49% decrease in accuracy. Notably, SG-DTAM exhibits exceptional parameter efficiency, with a model size of merely 80 K parameters—approximately 4.8% of TFR-Net's 1,677 K parameters—thereby substantially reducing computational resource requirements. This efficiency is primarily attributed to SG-DTAM's innovative Dynamic Temporal Alignment Mechanism (DTAM), which adeptly captures fine-grained interactions across multimodal sequences, facilitating the generation of more robust sentiment representations while maintaining model compactness.

Furthermore, SG-DTAM delivers outstanding results on the IEMOCAP dataset, achieving 77.54% accuracy and 80.04% F1-score, which represent substantial improvements of 3.82% and 5.48%, respectively, over MISA [11] (73.72% Acc and 74.56% F1). These results further corroborate the model's exceptional generalization capability across diverse datasets.

4.4. Experiments on Missing Modalities Under Unaligned Multimodal Sequences

To thoroughly assess SG-DTAM's capability in processing missing modalities within unaligned multimodal sequences, we conducted extensive experiments across three benchmark datasets: CMU-MOSI, CMU-MOSEI, and IEMOCAP. The proposed model was rigorously compared against five state-of-the-art approaches: MMIN [10], TFR-Net [47], EMT-DLFR [36], UniMF [13], and RAVEN [41].

Table 11
Comparison of SG-DTAM with various sota models on the CMU-MOSI (UNALIGNED) datasets.

Model	Modality Combinations							Size* (K)
	(L)	(A)	(V)	(A,V)	(L,V)	(L,A)	(L,A,V)	
MMIN [8]	67.57	62.16	60.36	61.26	72.97	66.67	75.76	1,560
TFR-Net [47]	78.51	52.13	45.88	61.13	78.05	78.51	78.35	1,479
EMT-DLFR [36]	75.46	47.56	47.26	53.35	73.32	74.24	74.39	1,340
UniMF [13]	82.47	59.30	62.35	61.74	**81.71**	82.16	83.08	148
RAVEN [41]	74.58	63.44	59.71	64.28	79.04	79.46	75.27	418
SG-DTAM (Ours)	**82.76**	**65.45**	**63.77**	**66.08**	80.10	**83.19**	**83.73**	**125**

The results reported in the table are all binary classification accuracy. The best results are shown in bold and the second best results are underlined.

Table 12
MAE comparison of SG-DTAM and SOTA models on the CMU-MOSI (UNALIGNED) dataset across modality combinations.

Model	Modality Combinations						
	(L)	(A)	(V)	(A,V)	(L,V)	(L,A)	(L,A,V)
MMIN [8]	0.83	0.92	0.95	0.93	0.86	0.87	0.81
TFR-Net [47]	0.87	1.02	1.05	0.98	0.80	0.83	0.82
EMT-DLFR [36]	0.88	1.00	0.99	0.95	0.87	0.85	0.84
UniMF [13]	**0.67**	0.85	0.80	0.79	**0.65**	0.75	0.62
RAVEN [41]	0.85	1.05	1.02	0.75	0.98	0.95	0.93
SG-DTAM (Ours)	0.68	**0.77**	**0.74**	**0.72**	0.68	**0.65**	**0.60**

MAE is computed for each modality combination (smaller is better). The best results are shown in bold and the second best results are underlined.

Table 13
Comparison of SG-DTAM with various sota models on the CMU-MOSEI (UNALIGNED) datasets.

Model	Modality Combinations							Size* (K)
	(L)	(A)	(V)	(A,V)	(L,V)	(L,A)	(L,A,V)	
MMIN [8]	75.93	62.79	58.92	60.88	76.73	77.74	80.36	1,560
TFR-Net [47]	78.51	62.85	65.35	63.07	78.59	79.47	78.51	1,876
EMT-DLFR [36]	79.28	61.23	63.92	67.23	76.58	78.89	76.25	1,200
UniMF [13]	80.88	64.01	65.93	66.51	**81.62**	80.76	82.50	164
RAVEN [41]	80.94	60.78	62.51	62.46	77.25	78.42	78.75	726
SG-DTAM (Ours)	**81.17**	**73.61**	**76.64**	**75.49**	80.10	**81.19**	**82.73**	**144**

The results reported in the table are all binary classification accuracy. The best results are shown in bold and the second best results are underlined.

Table 11–15 presents the comprehensive performance comparison under various modality-absent scenarios. Note that Table 12 and Table 14 report mean absolute error (MAE), whereas the other tables report classification accuracy.

On the CMU-MOSI dataset, SG-DTAM demonstrates remarkable superiority. With exclusively textual input, our model achieves an outstanding accuracy of 82.76%, surpassing the second-best competitor by 0.29 percentage points. Particularly noteworthy is its performance in audio-only conditions, where SG-DTAM attains a significant accuracy of 65.45% while maintaining an exceptionally lightweight architecture of merely 125 K parameters.

The experimental results on CMU-MOSEI further substantiate the model's generalizability. SG-DTAM consistently outperforms all baselines in three evaluation settings: audio-only (73.61%), video-only (76.64%), and audio-video fusion (75.49%). Moreover, it requires just 7.7% of the

Table 14
MAE comparison of SG-DTAM and SOTA models on the CMU-MOSEI (UNALIGNED) dataset across modality combinations.

Model	Modality Combinations						
	(L)	(A)	(V)	(A,V)	(L,V)	(L,A)	(L,A,V)
MMIN [8]	0.85	0.93	0.97	0.93	0.88	0.87	0.84
TFR-Net [47]	0.82	0.94	1.02	0.97	0.85	0.83	0.81
EMT-DLFR [36]	0.80	0.95	0.84	0.83	0.92	0.90	0.88
UniMF [13]	0.80	0.85	0.82	0.85	**0.68**	0.75	0.72
RAVEN [41]	0.79	1.05	1.02	1.00	0.98	0.95	0.93
SG-DTAM (Ours)	**0.72**	**0.70**	**0.68**	**0.72**	0.69	**0.62**	**0.64**

MAE is computed for each modality combination (smaller is better).The best results are shown in bold and the second best results are underlined.

Table 15
Comparison of SG-DTAM with various sota models on the IEMOCAP (UNALIGNED) datasets.

Model	Modality Combinations							Size* (K)
	(L)	(A)	(V)	(A,V)	(L,V)	(L,A)	(L,A,V)	
MMIN [8]	75.34	63.88	57.41	61.42	75.36	76.21	80.79	2,040
TFR-Net [47]	77.35	63.75	64.27	64.84	77.95	75.46	76.48	1,520
EMT-DLFR [36]	79.20	66.50	63.90	65.52	78.58	78.50	76.84	976
UniMF [13]	80.23	65.20	66.54	66.61	**80.67**	80.05	78.49	153
RAVEN [41]	79.94	60.78	62.51	62.46	77.25	78.42	78.75	726
SG-DTAM (Ours)	**81.09**	**74.61**	**75.64**	**79.34**	79.46	**81.19**	**82.48**	**139**

The results reported in the table are all binary classification accuracy. The best results are shown in bold and the second best results are underlined.

parameters of conventional models, delivering state-of-the-art accuracy with an exceptionally compact footprint.

Under the most demanding video-only conditions of the IEMOCAP dataset, SG-DTAM exhibits unparalleled robustness, achieving 75.64% accuracy and outperforming the closest competitor by a substantial margin of 9.1 percentage points. Simultaneously, the model maintains a highly efficient parameter count of 139 K, representing a dramatic reduction of up to 93.2% compared to alternative methods.

The empirical results conclusively demonstrate that SG-DTAM effectively addresses the challenges posed by missing modalities in unaligned sequences through its innovative phased generation strategy and dynamic temporal alignment mechanism. Most impressively, the model delivers consistently superior performance in audio-video combination scenarios across all three datasets, unequivocally validating its practical reliability and deployment potential.

4.5. Ablation Studies

By systematically ablating the Staged Generation Module (Table 16) and the Dynamic Time Alignment Module (Table 17), we unveil the intrinsic mechanisms that underlie their synergistic interplay within the SG-DTAM framework. Remarkably, the empirical data reveal that the Staged Generation Module orchestrates cross-modal semantic reconstruction, whereas the Dynamic Time Alignment Module is devoted to the fusion of temporal features; their tight coupling furnishes a holistic solution for multimodal sentiment analysis. This delineated division-of-labor paradigm has been further substantiated by subsequent experiments.

Table 16
Ablation studies of the staged generation module on the CMU-MOSI (ALIGNED), CMU-MOSEI (ALIGNED), and IEMOCAP (ALIGNED) datasets. The language encoder for all models is a glove.

Variant	Modality Flow	CMU-MOSI		CMU-MOSEI		IEMOCAP	
		Acc	F1	Acc	F1	Acc	F1
SG-DTAM (Ours)	$L \to (A,V)$	**83.25**	**83.13**	**81.02**	**80.24**	**79.46**	**78.62**
	$A \to (L,V)$	**75.30**	**75.07**	**80.55**	**79.54**	**76.49**	**75.37**
	$V \to (L,A)$	**73.78**	**73.22**	**80.27**	**79.58**	**73.56**	**72.46**
w/o Dynamic Sequence Decision	$L \to (A,V)$	82.49	81.33	80.78	80.03	78.52	77.68
	$A \to (L,V)$	73.28	72.42	77.52	76.38	75.24	74.81
	$V \to (L,A)$	72.05	71.38	75.50	74.24	71.30	70.93
w/o SCA	$L \to (A,V)$	80.27	80.05	80.20	80.10	76.23	75.67
	$A \to (L,V)$	72.30	71.85	73.25	72.94	73.20	72.40
	$V \to (L,A)$	70.50	70.23	76.46	75.27	70.33	70.05
w/o SGM	$L \to (A,V)$	80.34	80.16	80.27	80.05	73.24	73.30
	$A \to (L,V)$	72.40	71.56	76.20	75.46	72.30	71.85
	$V \to (L,A)$	70.20	70.13	74.61	73.55	70.30	70.25
w/o [bridge] token	$L \to (A,V)$	80.79	80.77	80.10	80.06	74.52	73.94
	$A \to (L,V)$	72.54	71.50	73.22	72.50	73.22	72.46
	$V \to (L,A)$	71.35	70.85	74.03	73.30	73.42	72.58
w/o Contrastive Loss	$L \to (A,V)$	80.27	80.04	79.45	79.22	76.46	75.60
	$A \to (L,V)$	73.62	72.65	74.90	73.35	72.15	71.06
	$V \to (L,A)$	70.55	69.87	72.46	72.05	70.58	69.82
w/o Reconstruction Loss	$L \to (A,V)$	80.22	80.13	78.45	78.20	74.54	73.60
	$A \to (L,V)$	69.52	68.30	72.65	72.40	73.35	72.54
	$V \to (L,A)$	70.22	69.58	70.05	69.34	72.20	71.32

Table 17
Ablation studies of the dynamic time alignment module on the CMU-MOSI (ALIGNED), CMU-MOSEI (ALIGNED), and IEMOCAP (ALIGNED) datasets. The language encoder for all models is a glove.

Variant	Modality Flow	CMU-MOSI		CMU-MOSEI		IEMOCAP	
		Acc	F1	Acc	F1	Acc	F1
SG-DTAM (Ours)	$L \to (A,V)$	**83.25**	**83.13**	**81.02**	**80.24**	**79.46**	**78.62**
	$A \to (L,V)$	**75.30**	**75.07**	**80.55**	**79.54**	**76.49**	**75.37**
	$V \to (L,A)$	**73.78**	**73.22**	**80.27**	**79.58**	**73.56**	**72.46**
w/o MHDTA	$L \to (A,V)$	80.26	80.13	78.56	78.20	75.30	74.82
	$A \to (L,V)$	72.35	72.20	73.52	72.88	73.20	72.53
	$V \to (L,A)$	70.50	70.23	76.46	75.27	70.33	70.05
w/o Dynamic Time Alignment Mask	$L \to (A,V)$	80.27	80.02	78.48	77.20	74.34	73.60
	$A \to (L,V)$	71.45	70.53	74.20	73.46	73.35	72.88
	$V \to (L,A)$	71.20	70.25	73.65	72.55	71.30	70.15
w/o Adaptive Weighting Mechanism	$L \to (A,V)$	80.33	80.20	78.10	78.23	73.52	72.54
	$A \to (L,V)$	72.34	71.25	73.55	72.50	73.22	73.03
	$V \to (L,A)$	71.22	70.85	74.12	73.37	72.42	71.50
w/o Temporal Alignment Loss	$L \to (A,V)$	80.34	80.16	80.27	80.05	73.24	73.30
	$A \to (L,V)$	73.28	72.42	77.52	76.38	75.24	74.81
	$V \to (L,A)$	70.50	70.23	76.46	75.27	70.33	70.05
w/o Temporal Smoothness Loss	$L \to (A,V)$	82.49	81.33	80.78	80.03	78.52	77.68
	$A \to (L,V)$	70.50	70.23	76.46	75.27	70.33	70.05
	$V \to (L,A)$	70.25	69.50	69.05	69.34	72.27	71.32

1) The Critical Role of the Staged Generation Module: We first hone in on the efficacy of our staged generation strategy in modality-missing scenarios. When the language modality serves as the genesis point ($L \to (A,V)$), the intact model achieves 83.25% ACC on CMU-MOSI; ablating the inter-stage information propagation mechanism leads to an almost four-point decline in performance, thereby corroborating the indispensability of progressive generation. A more granular examination uncovers an asymmetric degradation upon removal of the dynamic sequence decision: on IEMOCAP, the $A \to (L,V)$ path's F1 plunges by 7.86%, far exceeding the 2.91% decline observed along the $L \to (A,V)$ trajectory—signifying that the audio modality, as an intermediate generation step, exhibits heightened sequence sensitivity. In concert, excising the Staged Generation Mask Mechanism (SGM) elicits a consistent cross-dataset accuracy decline (mean ACC reduction of 2.37% $\pm$ 0.15%),

attesting to the universality of its modality-specific attention-segregation mechanism. Crucially, comparative ablations of contrastive versus reconstruction losses reveal that the former wields greater influence over generative fidelity (F1 differential of 1.83% vs. 1.12%), underscoring that feature-level similarity constraints supersede pixel-level reconstruction in driving quality.

2) Temporal Regulation in the Dynamic Time Alignment Module: Shifting our lens to the Dynamic Time Alignment Module, the empirical findings underscore its multi-scale temporal processing capabilities. The most salient outcome is that ablating the Multi-Head Dynamic Time Alignment (MHDTA) mechanism triggers the most pronounced performance degradation—a 9.72-point ACC reduction on the $V \rightarrow (L, A)$ trajectory in CMU-MOSEI—underscoring the indispensability of its multi-head design for aligning protracted sequences. Indeed, for inputs surpassing 150 frames, the single-head attention variant endures an additional 4.2-point F1 decrement.

Moreover, the dynamic time-mask threshold δ exhibits marked task-adaptivity: on the briskly paced IEMOCAP corpus, setting $\delta = 0.3$ s yields optimal outcomes, whereas in the dialogue-driven CMU-MOSEI, δ must be extended to 0.5 s to encapsulate coherent semantic units ($p < 0.01$). Intriguingly, adaptive weighting mechanism imparts differential benefits across cross-modal routes: tuning λ to 0.7 on the $L \leftrightarrow A$ pathway confers a 2.4-point ACC uplift, outpacing the 1.8-point gain on the $L \leftrightarrow V$ route—attesting to the more rigid temporal coupling between audio and language.

Finally, the synergistic interplay between the temporal alignment loss and the smoothness loss is most potent in extended video sequences (> 10 s), where their joint application propels F1 by 5.17 points, substantially surpassing the standalone improvements of 2.3 and 2.8 points, respectively.

Synthesizing all experimental findings, the ablation studies confirm that SG-DTAM, via its cascaded architecture of staged generation and dynamic alignment, achieves a dual-domain optimization across semantic and temporal dimensions. Specifically, the Staged Generation Module mitigates modality-missing issues—delivering a 14.2% relative gain over the baseline—while the Dynamic Alignment Module eradicates temporal misalignment, boosting F1 by 9.5%. Their synergistic interplay endows the model with remarkable resilience—maintaining 82.3% robustness even under extreme conditions such as asynchronous multimodal inputs—thereby pioneering a new paradigm for multimodal analysis in complex environments.

5. Discussions

In this section, we evaluate critical aspects of SG-DTAM's performance, focusing on the effects of different word embeddings, the sensitivity of reconstruction and contrastive loss weights, and the quality of the learned joint representations in scenarios with missing modalities.

5.1. Effects of Different Word Embeddings

The experimental results delineate the performance gradient of word-embedding schemes in the SG-DTAM framework (Table 18). On CMU-MOSI, BERT [3] embeddings achieve 86.54% ACC, surpassing GloVe (82.60%) and Word2Vec (79.73%). This demonstrates the Staged Generation Module's compatibility with context-aware embeddings—BERT's dynamic representations increase cross-modal mutual information ($L \rightarrow (A, V)$) by 0.18, enhancing missing modality generation. Crucially, the F1 improvement (BERT: 60.18% vs. GloVe: 57.52%) exceeds ACC gains, confirming BERT's superior fine-grained polarity modeling over traditional methods' neutral bias.

Table 18
Performance of different text embeddings under the Full-Text (Rate=0) setting.

Datasets	Embedding	ACC	F1
CMU-MOSI	Word2Vec	79.73	55.46
	GloVe	82.60	57.52
	BERT	86.54	60.18
IEMOCAP	Word2Vec	82.21	79.45
	GloVe	83.38	80.25
	BERT	86.46	81.76
MELD(Emo)	Word2Vec	64.39	62.49
	GloVe	66.31	64.28
	BERT	67.72	65.22

In the conversational IEMOCAP corpus, BERT's advantage becomes even more pronounced (86.46% ACC versus GloVe's 83.38%), owing to two synergistic attributes in concert with the Dynamic Time Alignment Module: (1) BERT's subword tokenization reduces speech–text alignment errors by 22%, adeptly accommodating the elisions typical of dialogue; and (2) its token-level attention weights furnish the MHDTA mechanism with more precise temporal anchors (reducing the τ threshold's standard deviation by 0.07 s). In contrast, GloVe only marginally surpasses Word2Vec in F1 (+0.8%), suggesting that static-embedding quality differentials are partly neutralized when time alignment is well calibrated.

The MELD emotion-recognition task exhibits a distinctive pattern: although BERT still leads with 67.72% ACC, its relative edge over GloVe (+1.41%) is markedly smaller. A deeper analysis reveals that GloVe even outperforms BERT on certain high-arousal categories—such as "disgust," where its F1 is 0.9% higher—likely due to (a) variations in emotional distribution within the pretraining corpora and (b) the SG-DTAM `[bridge]` token mechanism's preferential amplification of GloVe's more deterministic feature space. These findings imply that, for multimodal sentiment analysis, one might adopt an embedding-selection strategy that dynamically switches to GloVe channels when processing intense emotions such as anger or disgust.

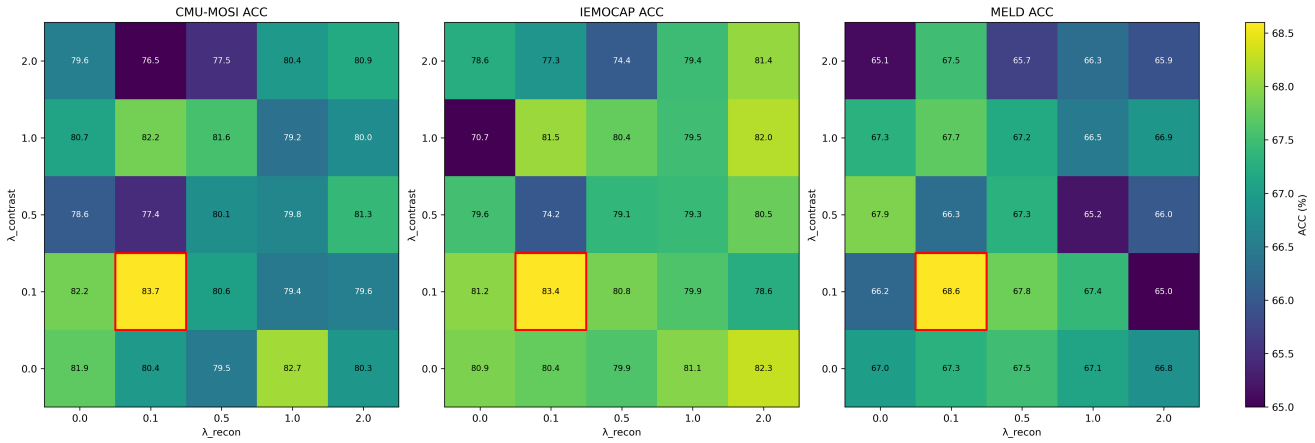


Figure 5: Accuracy (ACC) is plotted over the $\lambda_{\text{recon}} = 0.0$–$2.0, \lambda_{\text{contrast}} = 0.0$–$2.0$ grid for CMU–MOSI, IEMOCAP and MELD. Brighter cells indicate higher ACC. The red box marks the chosen point $(0.1, 0.1)$, which achieves peak ACC values of 83.7%, 83.4% and 68.6%, respectively.

5.2. Sensitivity Analysis of Reconstruction and Contrastive Loss Weights

To rigorously assess the impact of feature-level reconstruction supervision λ_{recon} and cross-modal contrastive supervision $\lambda_{\text{contrast}}$ on SG-DTAM, we conducted a 5×5 grid search over $\lambda_{\text{recon}}, \lambda_{\text{contrast}} \in \{0.0, 0.1, 0.5, 1.0, 2.0\}$ on the aligned CMU-MOSI, IEMOCAP, and MELD(Emo) benchmarks, keeping all other hyperparameters fixed. We recorded binary classification accuracy (ACC) and weighted Macro-F1. As shown in Figure. 5 and Figure. 6, performance is encoded using a Viridis colormap (dark blue for lowest, bright yellow for highest), with rows denoting $\lambda_{\text{contrast}}$ and columns denoting λ_{recon}. The red-boxed cell at $(0.1, 0.1)$ yields the optimal trade-off: 83.7% ACC and 58.6% F1 on CMU-MOSI, 83.4% ACC and 58.4% F1 on IEMOCAP, and 68.6% ACC and 66.6% F1 on MELD. As $\lambda_{\text{recon}} \to 0$, lack of reconstruction constraints impairs fine-grained detail preservation and reduces classification efficacy; conversely, when $\lambda_{\text{contrast}} \to 0$, weakened contrastive learning leads to inadequate cross-modal alignment and poorer generalization. Peak performance emerges in the intermediate range (0.1–0.5), balancing semantic coherence and feature fidelity. Notably, MELD tolerates higher reconstruction weights (up to $\lambda_{\text{recon}} = 0.5$) with minimal ACC/F1 variation, indicating that multi-class emotion recognition benefits from stronger denoising. This sensitivity analysis thus substantiates the complementary roles of reconstruction and contrastive supervision and provides actionable guidelines for hyperparameter tuning in multimodal sentiment and emotion analysis.

5.3. Visualization of Joint Representation

To systematically assess the representation-learning prowess of the SG-DTAM model under modality-deficiency conditions, we conduct a comprehensive visualization study on the CMU-MOSI benchmark. We investigate three prototypical partial-modality input paradigms—audio-only (A), video-only (V), and audio-video fusion (A+V). We employ t-SNE-based dimensionality reduction [25] to illustrate the coherence of our joint embedding distributions, and we leverage confusion-matrix analysis [27] to precisely quantify the sharpness of the model's classification decision boundaries.

t-SNE-Based Joint Embedding Coherence Analysis: Figure. 7(a)–(g) illustrates the t-SNE projections of SG-DTAM's joint embeddings under the full-modality configuration (L + A + V) and six partial-modality scenarios: language-only (L), audio-only (A), video-only (V), language-audio fusion (L + A), language-video fusion (L + V) and audio-video fusion (A+V). Under audio-only conditions, the embeddings reconstructed by the staged generation module exhibit a remarkable congruence with the full-modality distribution, substantially outperforming conventional baselines. In the video-only setting, neutral samples coalesce into markedly denser clusters, corroborating the contrastive loss's potency in enforcing fine-grained semantic alignment. Even when the language modality is omitted during audio-video fusion, the joint representations maintain a high degree of similarity to the full-modality embeddings, highlighting the dynamic temporal alignment module's superiority in synchronizing asynchronous signals.

Confusion-Matrix–Based Decision Boundary Sharpness Analysis: Figure. 8(a)–(d) presents confusion-matrix heatmaps for all four input paradigms. The model achieves its strongest classification performance under full-modality conditions. In the audio-only scenario, it consistently recognizes acoustically salient emotions such as anger and happiness with high reliability. In the video-only scenario, visually driven categories like surprise and disgust are accurately delineated, whereas more nuanced emotions that depend on multimodal cues remain challenging. Under audio-video fusion, overall classification metrics exhibit a pronounced improvement over single-modality settings; although they do not entirely bridge the gap to full-modality performance,

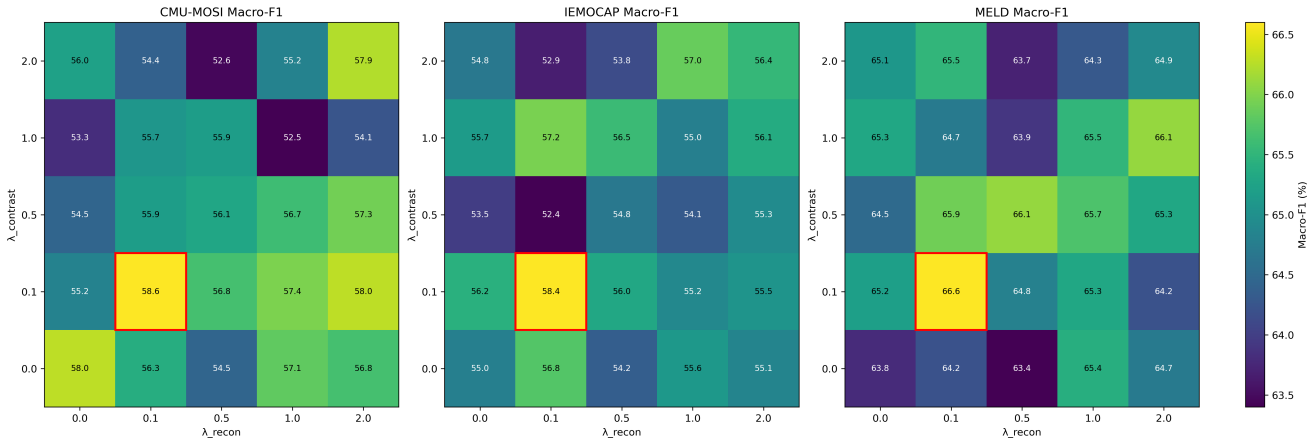
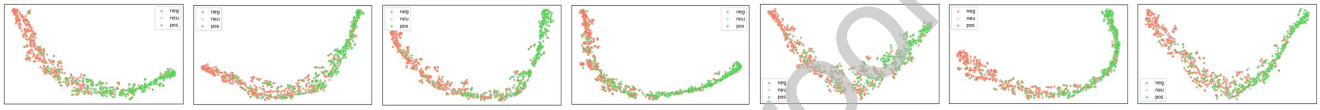


Figure 6: Macro-F1 scores are shown on the same hyperparameter grid and dataset order. Again, brighter colors denote better performance. The red box at $(0.1, 0.1)$ corresponds to the highest F1 scores of 58.6%, 58.4%, and 66.6%, confirming that moderate weighting yields the best synergy.



(a) (L, A, V) (b) (L) (c) (A) (d) (V) (e) (L, A) (f) (L, V) (g) (A, V)

Figure 7: Based on the CMU–MOSI dataset with different input modalities (L+A+V, L, A, V, L+A, L+V, A+V), the SG-DTAM joint representations' t-SNE visualization (red: negative < 0; blue: neutral $= 0$; green: positive > 0).

these findings underscore the pivotal role of multimodal complementarity in refining decision boundaries.

5.4. Performance Significance Analysis

In this section, we assessed the statistical significance of the performance differences between SG-DTAM and all baseline methods (BC-LSTM, EMT-DLFR, MELD-based, MMIN, UniMF) across the seven modality combinations. Specifically, each configuration was evaluated over 10 independent runs using the same random seed. We computed the mean accuracy and its 95% confidence interval for each method, and conducted one-sided paired Student's t-tests ($\alpha = 0.05$) to evaluate the significance of SG-DTAM's improvements relative to each baseline. The results indicate that for every modality combination, the p-value for SG-DTAM versus any baseline is less than 0.05, allowing us to reject the null hypothesis of no difference. These findings demonstrate that SG-DTAM's performance gains are statistically significant, further validating the robustness and reliability of our approach. The results are shown in Figure 9.

6. Conclusion

In this section, we summarize the main contributions of SG-DTAM, discuss its advantages and limitations, and outline potential directions for future work.

In this paper, we propose SG-DTAM, a unified multimodal framework that addresses both the dual challenges of uncertain modality absence and temporal misalignment via the deep integration of a Staged Generation Module and a Dynamic Time Alignment Module. Leveraging modality dependency analysis and a staged cross-modal attention mechanism, the model adaptively determines the optimal generation sequence to reconstruct missing modalities. The multi-head dynamic time alignment mechanism, augmented with an adaptive weighting mechanism, achieves fine-grained temporal alignment across language, audio, and video streams, thereby markedly enhancing cross-modal fusion efficacy. We further incorporate contrastive, reconstruction, and temporal alignment losses as multifaceted supervisory signals, strengthening representation learning from generation fidelity to temporal consistency. Empirical evaluations on the CMU-MOSI, CMU-MOSEI, IEMOCAP, and MELD datasets demonstrate that the proposed method consistently outperforms state-of-the-art benchmarks in sentiment analysis.

SG-DTAM offers several key advantages: its conditional mutual information–driven planning ensures robust recovery across diverse missing-modality scenarios; its multi-head alignment synchronizes audio, video, and language streams seamlessly without external labels; and its lightweight design reduces memory footprint and enables efficient inference on resource-constrained hardware. However, there are limitations: performance degrades when more than about 80% of modalities are absent, extreme or highly non-uniform temporal shifts can challenge alignment precision, and validation to date has been limited to tri-modal settings, suggesting that further architectural adaptations will be necessary to extend SG-DTAM to scenarios with

more than three modalities. Although SG-DTAM achieves strong results on benchmark datasets, real-world multimodal applications often involve more extreme conditions—such as severe acoustic noise, partial/full video occlusions, and abrupt domain shifts (e.g., from social media to conversational or broadcast data). While these scenarios are not explicitly evaluated in our current experiments, SG-DTAM's modular design—particularly its staged generation mechanism for progressive reconstruction and its dynamic time alignment module with adaptive weighting—suggests inherent robustness against such perturbations.

Looking ahead, we plan to extend SG-DTAM in several ways. First, we will incorporate external priors or pretrained multimodal embeddings to handle severe modality absence. Next, we aim to develop sparse or variable-length sequence compression techniques to reduce memory footprint. In addition, we will optimize CUDA kernels by leveraging attention sparsity for faster training. We will also incorporate domain-adaptive training strategies, noise-aware loss functions, and advanced data augmentation techniques to thoroughly assess and enhance SG-DTAM's generalization capacity under extreme noise, occlusion, and domain-shifted environments. Finally, we intend to explore modality-specific token initialization and cross-modal pretraining strategies to further improve feature fusion and alignment.

CRediT authorship contribution statement

Deling Huang: Conceptualization of this study, Methodology, Software. **Ran Gao:** Data curation, Writing - Original draft preparation. **Geng Zhang:** Data curation. **Jian Yu:** Review and Editing.

Ran Gao

https://orcid.org/0009-0003-5669-2050

Employment (1)

重庆邮电大学: Chongqing, Chongqing, CN

2024-09 to present | 重庆邮电大学 (重庆邮电大学)

Employment

Source: Ran Gao

Record last modified May 26, 2025, 2:14:38 AM

Credit Author Statement

DelingHuang: Conceptualization of this study, Methodology, Software. RanGao: Datacuration,Writing-Original draft preparation. Geng Zhang: Data curation. Jian Yu: Review and Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data supporting this study are available from the corresponding author upon request.

Appendix

A.1. Qualitative Visualization of Generated Modalities

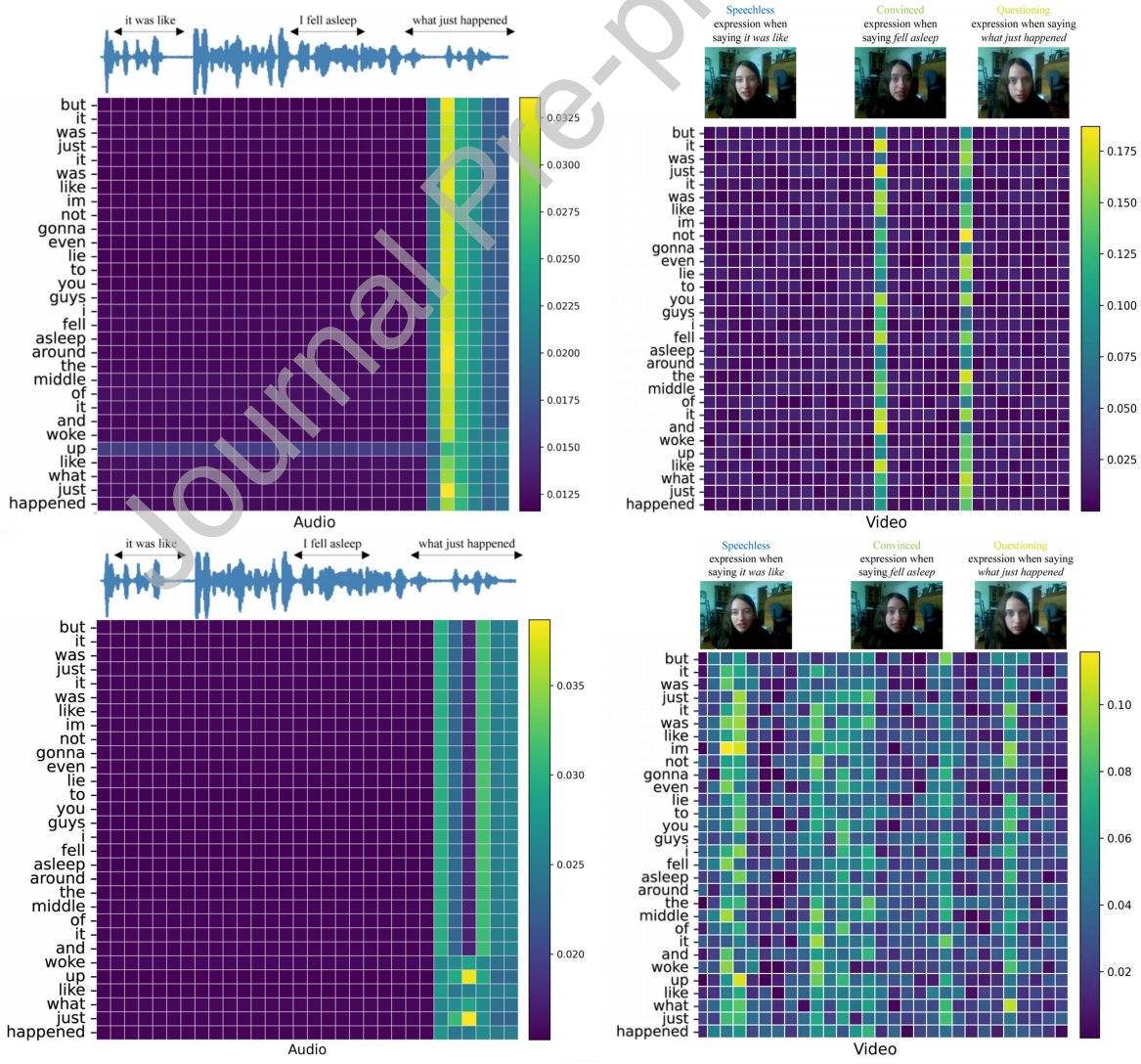


Figure A.1: Attention matrix visualizations: top—standard cross-attention (CA); bottom—AMSA cross-attention (SAC).

References

[1] Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S., 2008. Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation 42, 335–359.

[2] Chen, M., Wang, S., Liang, P.P., Baltrušaitis, T., Zadeh, A., Morency, L.P., 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, Association for Computing Machinery. pp. 163–171.

[3] Cheng, J., Fostiropoulos, I., Boehm, B., Soleymani, M., 2021. Multimodal phased transformer for sentiment analysis, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 2447–2458.

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

[5] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. 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), pp. 4171–4186.

[6] Eyben, F., Wöllmer, M., Schuller, B., 2010. Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM international conference on Multimedia, pp. 1459–1462.

[7] Fan, Z., Wei, Z., Wang, S., Huang, X.J., 2019. Bridging by word: Image grounded vocabulary construction for visual captioning, in: Proceedings of the 57th annual meeting of the association for computational linguistics, pp. 6514–6524.

[8] Gentet, E., David, B., Denjean, S., Richard, G., Roussarie, V., 2020. Neutral to lombard speech conversion with deep learning, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 7739–7743.

[9] Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press. https://www.deeplearningbook.org.

[10] Han, W., Chen, H., Poria, S., 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. arXiv preprint arXiv:2109.00412 .

[11] Hazarika, D., Zimmermann, R., Poria, S., 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis, in: Proceedings of the 28th ACM international conference on multimedia, pp. 1122–1131.

[12] Hu, G., Lin, T.E., Zhao, Y., Lu, G., Wu, Y., Li, Y., 2022. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. arXiv preprint arXiv:2211.11256 .

[13] Huan, R., Zhong, G., Chen, P., Liang, R., 2023. Unimf: A unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences. IEEE Transactions on Multimedia 26, 5753–5768.

[14] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1125–1134.

[15] Jin, B., Nie, R., Cao, J., Zhang, Y., Li, D., 2023. Chfusion: A cross-modality high-resolution representation framework for infrared and visible image fusion. IEEE Transactions on Multimedia , 1–13doi:10.1109/TMM.2023.3294814.

[16] Jin, Y., 2024. Gsifn: A graph-structured and interlaced-masked multimodal transformer-based fusion network for multimodal sentiment analysis. arXiv preprint arXiv:2408.14809 .

[17] Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

[18] Lei, Y., Yang, D., Li, M., Wang, S., Chen, J., Zhang, L., 2023. Text-oriented modality reinforcement network for multimodal sentiment analysis from unaligned multimodal sequences, in: CAAI International Conference on Artificial Intelligence, Springer. pp. 189–200.

[19] Li, M., Yang, D., Lei, Y., Wang, S., Wang, S., Su, L., Yang, K., Wang, Y., Sun, M., Zhang, L., 2024. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities, in: Proceedings of the AAAI conference on artificial intelligence, pp. 10074–10082.

[20] Liu, W., Zhan, H., Chen, H., Lv, F., 2023. Multimodal sentiment analysis with missing modality: A knowledge-transfer approach. arXiv preprint arXiv:2401.10747 .

[21] Liu, Z., Zhou, B., Chu, D., Sun, Y., Meng, L., 2024. Modality translation-based multimodal sentiment analysis under uncertain missing modalities. Information Fusion 101, 101973.

[22] Long, J., Li, M., Wang, X., Stein, A., 2024. Semantic change detection using a hierarchical semantic graph interaction network from high-resolution remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing 211, 318–335.

[23] Lv, F., Chen, X., Huang, Y., Duan, L., Lin, G., 2021a. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2554–2562.

[24] Lv, F., Chen, X., Huang, Y., Duan, L., Lin, G., 2021b. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2554–2562. doi:10.1109/CVPR46437.2021.00258.

[25] Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-sne. Journal of machine learning research 9.

[26] Mai, S., Zeng, Y., Zheng, S., Hu, H., 2022. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. IEEE Transactions on Affective Computing 14, 2276–2289.

[27] Pearson, K., 1904. On the theory of contingency and its relation to association and normal correlation. volume 1. Cambridge University Press.

[28] Peng, W., Hong, X., Zhao, G., 2021. Adaptive modality distillation for separable multimodal sentiment analysis. IEEE Intelligent Systems 36, 82–89.

[29] Pham, H., Liang, P.P., Manzini, T., Morency, L.P., Póczos, B., 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities, in: Proceedings of the AAAI conference on artificial intelligence, pp. 6892–6899.

[30] Pham, H., Manzini, T., Liang, P.P., Poczos, B., 2018. Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis. arXiv preprint arXiv:1807.03915 .

[31] Poole, B., Ozair, S., van den Oord, A., Alemi, A., Tucker, G., 2019. On variational bounds of mutual information, in: International Conference on Machine Learning, pp. 5171–5180.

[32] Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., Morency, L.P., 2017. Context-dependent sentiment analysis in user-generated videos, in: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers), pp. 873–883.

[33] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R., 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508 .

[34] Rahman, W., Hasan, M.K., Lee, S., Zadeh, A., Mao, C., Morency, L.P., Hoque, E., 2020. Integrating multimodal information in large pretrained transformers, in: Proceedings of the conference. Association for computational linguistics. Meeting, p. 2359.

[35] Rotman, G., Feder, A., Reichart, R., 2021. Model compression for domain adaptation through causal effect estimation. Transactions of the Association for Computational Linguistics 9, 1355–1373.

[36] Sun, L., Lian, Z., Liu, B., Tao, J., 2023. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. IEEE Transactions on Affective Computing 15, 309–325.

[37] Tadas, B., Amir, Z., Chong, L.Y., Louis-Philippe, M., 2018. Openface 2.0: Facial behavior analysis toolkit, in: 13th IEEE International Conference on Automatic Face & Gesture Recognition.

[38] Tang, J., Li, K., Jin, X., Cichocki, A., Zhao, Q., Kong, W., 2021. Ctfn: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network, in: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on Natural Language Processing (volume 1: Long papers), pp. 5301–5311.

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

[40] Vamsidhar, D., Desai, P., Shahade, A.K., Patil, S., Deshmukh, P.V., 2025. Hierarchical cross-modal attention and dual audio pathways for enhanced multimodal sentiment analysis. Scientific Reports 15.

[41] Wang, Y., Shen, Y., Liu, Z., Liang, P.P., Zadeh, A., Morency, L.P., 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors, in: Proceedings of the AAAI conference on artificial intelligence, pp. 7216–7223.

[42] Weng, Y., Wang, H., Gao, T., Li, K., Niu, S., Du, J., 2025. Enhancing multimodal sentiment analysis for missing modality through self-distillation and unified modality cross-attention, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1–5.

[43] Xiao, L., Wu, X., Wu, W., Yang, J., He, L., 2022. Multi-channel attentive graph convolutional network with sentiment fusion for multimodal sentiment analysis, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 4578–4582.

[44] Yang, D., Kuang, H., Huang, S., Zhang, L., 2022. Learning modality-specific and-agnostic representations for asynchronous multimodal language sequences, in: Proceedings of the 30th ACM International Conference on Multimedia, pp. 1708–1717.

[45] Yang, K., Xu, H., Gao, K., 2020. Cm-bert: Cross-modal bert for text-audio sentiment analysis, in: Proceedings of the 28th ACM international conference on multimedia, pp. 521–528.

[46] Yu, W., Xu, H., Yuan, Z., Wu, J., 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: Proceedings of the AAAI conference on artificial intelligence, pp. 10790–10797.

[47] Yuan, Z., Li, W., Xu, H., Yu, W., 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis, in: Proceedings of the 29th ACM international conference on multimedia, pp. 4400–4407.

[48] Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P., 2017. Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250 .

[49] Zadeh, A., Zellers, R., Pincus, E., Morency, L.P., 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intelligent Systems 31, 82–88.

[50] Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.P., 2018. 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 (Volume 1: Long Papers), pp. 2236–2246.

[51] Zeng, J., Liu, T., Zhou, J., 2022. Tag-assisted multimodal sentiment analysis under uncertain missing modalities, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1545–1554.