entropy

MDPI

Article

DTA-CBF: Decoupled-Then-Aligned Context-Calibrated Bottleneck Fusion for Multimodal Sentiment Analysis

Yunan Zhang, Bo Zhang * and Wenlong Tan

College of Mathematics and Informatics, College of Software Engineering, South China Agricultural University, Guangzhou 510642, China; zhangyn912@stu.scau.edu.cn (Y.Z.); twl2024@stu.scau.edu.cn (W.T.)
* Correspondence: zb@scau.edu.cn

Abstract

Multimodal sentiment analysis (MSA) combines linguistic, acoustic, and visual signals, but effective prediction depends on selecting complementary cues while limiting redundant or conflicting information. Direct alignment and fusion can propagate modality-specific variation into the representation used for sentiment prediction. We propose decoupled-then-aligned context-calibrated bottleneck fusion (DTA-CBF), a framework that organizes multimodal information before fusion. Modality decoupling and semantic regularization (MDSR) separates modality-specific and modality-common components and regularizes them through orthogonality, prototype-level consistency, and homogeneous alignment based on maximum mean discrepancy. Text-guided cross-modal alignment (TGCA) uses text as a semantic query to retrieve complementary acoustic and visual evidence. Context-calibrated bottleneck fusion (CBF) progressively reduces the number of bottleneck tokens and calibrates the compact representation with global modality context. Experiments on CMU-MOSI, CMU-MOSEI, and CH-SIMS show competitive performance, particularly on the two English benchmarks, and ablations support the contributions of its three modules. On CMU-MOSI, the CBF fusion module uses 228.4 million multiply–add operations (MAdds), compared with 356.4 million MAdds for concatenation followed by self-attention, while achieving higher correlation and classification scores. These results support DTA-CBF as an architecture-level approach to organizing multimodal information and controlling fusion cost.

Keywords: multimodal sentiment analysis; affective computing; representation decoupling; cross-modal alignment; bottleneck fusion

Check for updates

Academic Editor: Friedhelm Schwenker

Received: 26 October 2026
Revised: 21 November 2026
Accepted: 26 November 2026
Published: 28 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Multimodal sentiment analysis (MSA) combines linguistic, acoustic, and visual signals to infer affective states from utterances [1]. Surveys describe multimodal learning as the integration of complementary evidence across heterogeneous sources [2–4], while datasets such as CMU-MOSI [5], CMU-MOSEI [6], and CH-SIMS [7] support utterance-level prediction. The challenge is not simply to collect more signals. Each modality contains sentiment-bearing cues alongside redundant, noisy, or conflicting variation. Effective prediction therefore depends on how information is separated, selected, and combined. Text typically provides explicit semantics, whereas audio and vision contribute prosodic and behavioral evidence that can complement, but sometimes conflict with, the linguistic signal.

The first challenge lies in modality heterogeneity. In video-based MSA, an utterance is commonly decomposed into textual transcripts, audio signals, and visual frames, and each

modality is encoded independently before multimodal learning. These modalities differ substantially in sampling rate, feature distribution, semantic granularity, and noise patterns. For example, text is discrete and semantically dense, while audio and vision are continuous, redundant, and sensitive to environmental or behavioral variations. Zadeh et al. [8] introduced tensor fusion to model high-order cross-modal interactions, and Liu et al. [9] reduced the cost of this process with low-rank multimodal fusion. Tsai et al. [10] explored factorized multimodal representations, and Hazarika et al. [11] separated modality-invariant and modality-specific subspaces in MSA. Recent work has refined this direction through disentangled representation learning [12], decoupled distillation [13], bias purification [14], disentanglement translation [15], and triple-disentangled affective modeling [16]. Distribution matching in multimodal learning can be formulated using maximum mean discrepancy (MMD) [17] or optimal-transport objectives [18–20]. These studies indicate that modality heterogeneity is not only a feature-space mismatch problem, but also an entanglement problem: raw representations contain both private modality cues and shared affective semantics. If these factors are not separated before interaction, direct alignment or fusion may confuse modality-specific noise with common sentiment evidence.

The second challenge is cross-modal alignment and interaction. Since text, audio, and visual streams are often unaligned in time and semantics, researchers have designed temporal alignment, semantic alignment, and cross-modal attention mechanisms to capture correspondences among modalities. Tsai et al. [21] proposed MulT to learn directional attention between unaligned multimodal sequences. Tang et al. [22] modeled discriminative multi-relation representations, Zhang et al. [23] introduced semantic graph refinement, and Wang et al. [24] developed hierarchical cross-modal fusion. Text-enhanced models introduce nonverbal information into pretrained language representations because language usually provides the main semantic scaffold for sentiment analysis. Rahman et al. [25] injected multimodal adaptation gates into BERT, while Yu et al. [26] generated unimodal supervision from multimodal labels. Han et al. [27] preserved task-relevant information by maximizing mutual information. Later text-centered methods further used textual semantics to guide multimodal interaction, including TETFN [28], DLF [29], and the cross-modal enhancement network [30]. Despite their effectiveness, many alignment methods operate directly on raw or insufficiently disentangled modality features. Attention can therefore transfer both useful correspondences and irrelevant acoustic-visual variation. The design question is not only how modalities interact, but which nonverbal evidence should be transferred to the sentiment-bearing text stream.

The third challenge is multimodal fusion. After alignment, a model must integrate long multimodal sequences into a compact representation for prediction. Direct concatenation and self-attention over all modality tokens can model interactions, but dense attention has high computational cost and may preserve redundant information. Latent-token architectures such as Perceiver IO process high-dimensional inputs through compact arrays [31], while attention bottlenecks restrict cross-modal exchange to a limited set of tokens [32]. These designs expose a representation-budget trade-off: reducing the number of tokens can lower interaction cost, but places greater demands on selecting informative inputs and retaining global context. Thus, compact fusion depends on both the quality of the representations entering the bottleneck and the way its limited capacity is used.

These challenges motivate DTA-CBF, a decoupled-then-aligned context-calibrated bottleneck fusion framework for MSA. Its information flow strategy has three stages: separate modality factors, selectively transfer nonverbal evidence, and integrate the resulting representations through a compact bottleneck. Modality decoupling and semantic regularization (MDSR) uses orthogonality, prototype-level regularization, and MMD-based homogeneous alignment to organize modality-specific and modality-common features. Text-guided cross-

modal alignment (TGCA) uses text as a semantic query for acoustic and visual evidence. Context-calibrated bottleneck fusion (CBF) progressively reduces bottleneck tokens and calibrates them with global modality context. Together, these modules provide an architectural approach to limiting uncontrolled interaction while retaining complementary cues for sentiment prediction.

We evaluate DTA-CBF on CMU-MOSI [5], CMU-MOSEI [6], and CH-SIMS [7], using regression, correlation, binary classification, and multi-class metrics. Module ablations, loss and fusion analyses, hyperparameter sensitivity tests, and qualitative cases examine its behavior. The main contributions are as follows:

1. We propose DTA-CBF, a framework integrating modality factorization, text-guided cross-modal interaction, and context-calibrated bottleneck fusion to organize complementary sentiment information before prediction.
2. We develop a prototype regularization objective that couples cross-modal assignment consistency with prototype-based feature reconstruction to coordinate modality-specific representations.
3. We extend hierarchical bottleneck fusion with global-context calibration, using learnable residual corrections to refine compressed bottleneck representations at each fusion level.

2. Related Work

2.1. Multimodal Representation Learning

Decoupled representation learning aims to separate modality-invariant or common semantics from modality-specific characteristics. MISA projects each modality into invariant and specific subspaces to reduce heterogeneity while preserving complementary cues [11]. Yang et al. [12] explored disentangled representation learning for multimodal emotion recognition, and Li et al. [13] used decoupled distilling to reduce cross-modal interference. Bias-purification methods further address spurious modality dependence [14], while triple-disentangled affective modeling separates multiple affective factors more explicitly [16]. DecAlign extends this idea by explicitly decoupling modality-unique and modality-common representations [33]. It then aligns modality-specific features through Gaussian mixture prototypes and multi-marginal optimal transport, and aligns common features through latent distribution matching and MMD regularization. These studies motivate separating shared affective semantics from modality-specific variations before cross-modal interaction.

DTA-CBF builds on this factorization and uses prototype-based cross-modal reconstruction to regularize modality-specific representations before text-guided interaction and compact fusion.

2.2. Multimodal Alignment

Alignment is a central issue in MSA because multimodal sequences are not naturally synchronized in either time or semantic space. Audio and visual features are sampled at different rates from text, and signals observed at the same timestamp may not express the same affective meaning. Factorized representation learning attempts to reduce cross-modal inconsistency by decomposing multimodal factors [10]. Modality translation provides another route, especially when some modalities are missing or uncertain [34]. MulT is a representative transformer-based method that learns directional pairwise attention between unaligned multimodal sequences [21]. Tang et al. [22] improved cross-modal correspondence with multi-relation representations, and Zhang et al. [23] used semantic graph refinement for alignment.

Text-centered semantic alignment uses language as an anchor for learning latent correspondences across modalities. For example, MAG-BERT injects nonverbal cues into a pretrained language backbone [25], and TETFN exploits textual semantics for multimodal interaction [28]. DLF further preserves text-centered sentiment semantics through disentangled language-focused modeling [29]. Missing-modality studies also emphasize the need to preserve textual semantic stability while recovering complementary nonverbal cues [35]. However, when alignment is applied to raw or weakly disentangled features, attention may capture modality-specific artifacts together with shared affective information. This motivates text-guided alignment after representation decoupling.

2.3. Multimodal Fusion and Interaction

Multimodal fusion integrates information from different modalities to obtain task-oriented representations. Early fusion directly combines unimodal features, as in tensor fusion [8] and low-rank fusion [9]. Recent MSA models further explore cross-modal enhancement [30], fusion using a multilayer perceptron (MLP) [36], hierarchical fusion [24], and unified multimodal representation learning [37]. Token-level interaction methods use cross-modal attention to exchange information across unaligned sequences [21]. Language-centered methods inject nonverbal cues into pretrained textual representations [25]. Semi-IIN explicitly separates intra-modal and inter-modal interaction with masked attention and uses a gating mechanism to control the two information flows [38].

Recent studies also introduce latent tokens, information hubs, or bottleneck mechanisms to constrain multimodal communication. Perceiver IO uses latent arrays to attend to high-dimensional inputs [31], and attention bottlenecks route cross-modal information through a limited set of shared tokens [32]. DashFusion combines dual-stream alignment with hierarchical bottleneck fusion, progressively compressing tokens while exchanging information with the modality streams [39]. Complementary approaches target task-relevant information directly, including mutual-information maximization in MMIM [27]. Balanced multimodal learning further addresses the risk that dominant modalities overshadow weaker ones [40], while unified representations can improve cross-modal generalization [37]. Together, these studies highlight two linked design choices: what information enters cross-modal interaction and how much capacity is allocated to represent it. DTA-CBF extends DashFusion's hierarchical bottleneck fusion with global-context calibration to refine the compressed multimodal representation.

3. Method

The overall framework of DTA-CBF is shown in Figure 1. Given text, audio, and visual inputs, the model first obtains unimodal sequence representations through modality-specific encoders. The encoded features are then processed by MDSR, which produces decoupled and re-projected modality representations. TGCA further aligns the audio-visual context to the text stream and uses a feed-forward network (FFN) to produce the aligned multimodal text representation H_{tav}. Finally, CBF integrates H_{tav} with the decoupled modality streams, and the resulting fused representation is fed into an MLP regression head.

3.1. Problem Formulation

Given an utterance represented by language, acoustic, and visual streams, multimodal sentiment analysis aims to predict its sentiment polarity or intensity. In this work, each sample consists of a textual sequence, an acoustic sequence, a visual sequence, and a continuous sentiment annotation. On CMU-MOSI [5], CMU-MOSEI [6], and CH-SIMS [7], the task is optimized as sentiment regression, while the predicted scores can also be converted into binary or multi-class sentiment labels for evaluation.

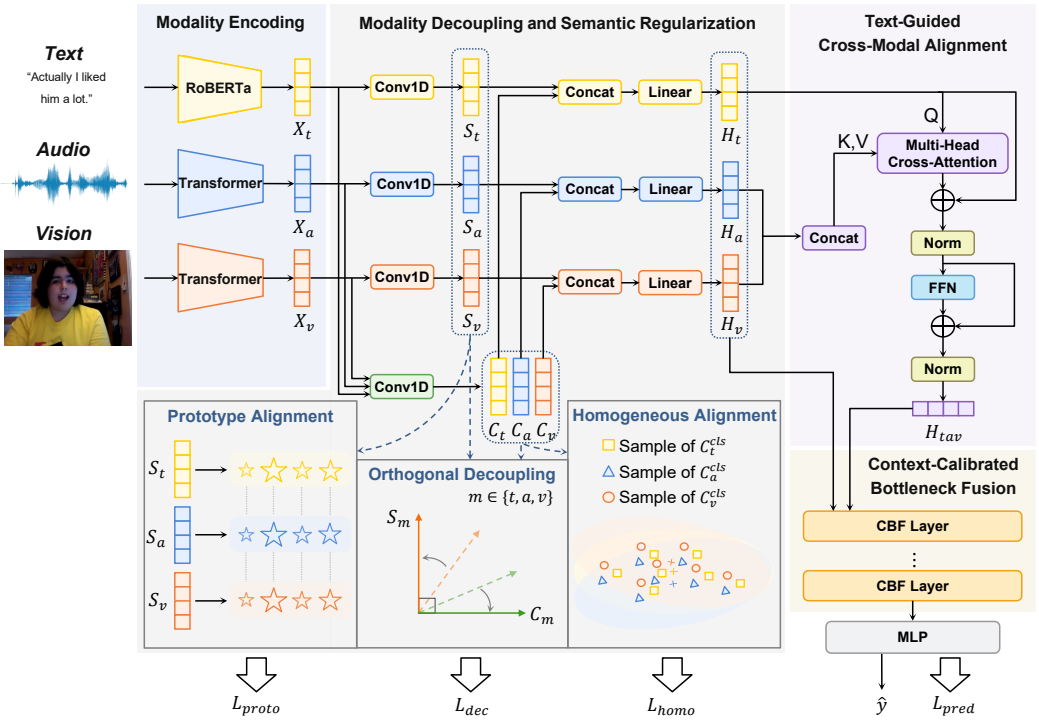


Figure 1. The overall architecture of DTA-CBF.

3.2. Modality Encoding

Regarding the multimodal inputs—which consist of text, audio, and video data—the raw inputs are first converted into modality-specific feature sequences. For all three datasets, we use the unaligned multimodal feature release provided by MMSA, which contains raw text, acoustic sequences, visual sequences, and labels. The stored acoustic/visual feature dimensions are 5/20 for CMU-MOSI, 74/35 for CMU-MOSEI, and 33/709 for CH-SIMS. The acoustic and visual arrays are fixed inputs rather than trainable feature extractors; no additional global feature standardization is performed. The acoustic and visual sequences are then processed by two separate Transformer encoders [41]. Before entering the encoders, each sequence is linearly projected, combined with positional embeddings, and prepended with a learnable classification token. The two encoders capture temporal dependencies and contextual relationships within the audio and visual sequences, respectively. Padding masks are used to prevent padded frames from participating in attention computation and pooling.

For text encoding, we use the roberta-large checkpoint [42] with its corresponding byte-level byte-pair encoding tokenizer on CMU-MOSI and CMU-MOSEI. For CH-SIMS, we use the hfl/chinese-roberta-wwm-ext checkpoint with its corresponding Chinese WordPiece tokenizer [43]. Across all three datasets, the maximum input length is 128 tokens, including special tokens, and all text encoder parameters are fine-tuned. The contextual token representations are projected to the same feature dimension as the nonverbal modalities. Padding tokens are also masked during textual encoding and pooling.

The resulting modality-specific features are denoted as $X_m \in \mathbb{R}^{T_m \times d_m}$, where $m \in \{t, a, v\}$. Here, T_m denotes the sequence length and d_m denotes the feature dimension of modality m.

3.3. Modality Decoupling and Semantic Regularization

Although the modality encoders map text, audio, and vision into sequence-level representations, each encoded feature still contains mixed information. Part of the representation reflects modality-specific cues, such as lexical structure, prosodic variation, or facial motion, while another part corresponds to sentiment semantics that can be shared across modalities. Like MISA [11], MDSR distinguishes modality-specific from shared representations. Following DecAlign [33], MDSR uses a cosine constraint to separate these components and matches the distributions of shared features across modalities.

Given the encoded representation X_m of modality $m \in \{t, a, v\}$, MDSR decomposes it into a modality-specific component S_m and a modality-common component C_m. The specific branch is implemented with modality-dependent projection layers, so that each modality can preserve its own discriminative characteristics. The common branch is shared across modalities, encouraging text, audio, and vision to extract comparable semantic factors. This process can be written as

$$S_m = E_m^s(X_m), \qquad C_m = E^c(X_m), \tag{1}$$

where E_m^s denotes the modality-specific encoder and E^c denotes the shared common encoder. In implementation, these encoders are lightweight one-dimensional convolutional projections. The modality encoders set masked positions to zero. Both projections act independently at each position and have no bias, so padding contributes nothing to the cosine similarity below.

To encourage the two components to capture different information, we impose an orthogonality constraint between S_m and C_m. Specifically, the decoupling loss minimizes the absolute cosine similarity between the flattened specific and common representations:

$$\mathcal{L}_{dec} = \frac{1}{3} \sum_{m \in \{t,a,v\}} |\cos(\mathrm{vec}(S_m), \mathrm{vec}(C_m))|. \tag{2}$$

This constraint encourages separation between the specific and common representations.

Prototype learning organizes samples around representative points in a learned feature space [44]. We use learnable prototypes to regularize modality-specific features through assignment consistency and cross-modal reconstruction. In multimodal learning, Fan et al. [45] used prototypes for modal rebalancing, and Hu et al. [46] used decoupled prototype completion for incomplete multimodal emotion recognition. For each modality, a learnable prototype matrix $P_m = [p_{m,1}; \ldots; p_{m,K}]$ is maintained. Let z_m denote the first classification-level token selected from the modality-specific component S_m. The assignment weight between z_m and the k-th prototype is computed according to their Euclidean distance and normalized with a softmax operation:

$$q_{m,k} = \frac{\exp(-\|z_m - p_{m,k}\|_2)}{\sum_{j=1}^{K} \exp(-\|z_m - p_{m,j}\|_2)}. \tag{3}$$

The assignment vector q_m describes how a sample is distributed over the prototype space of modality m. To make the modality-specific sentiment structures comparable, we minimize the discrepancy among the assignment distributions of text, audio, and vision:

$$\mathcal{L}_{assign} = \frac{1}{3}\left(\|q_t - q_a\|_2^2 + \|q_t - q_v\|_2^2 + \|q_a - q_v\|_2^2\right). \tag{4}$$

In addition, cross-modal prototype reconstruction is used to strengthen semantic correspondence among prototype spaces. For a mini-batch, let Z_m and Q_m collect the selected global modality-specific representations and their prototype assignment vectors, respectively. We apply the source modality's assignment weights to the target modality's prototypes, then

compare the reconstruction with the original source feature. We average the mean squared error over all ordered modality pairs:

$$\mathcal{L}_{rec} = \frac{1}{6} \sum_{\substack{m,n \in \{t,a,v\} \\ m \neq n}} \mathrm{MSE}(Z_m, Q_m P_n), \tag{5}$$

where $Q_m P_n$ denotes the features reconstructed by using the prototype assignments of modality m over the prototype set of modality n. The prototype regularization loss is defined as the combination of assignment consistency and reconstruction consistency:

$$\mathcal{L}_{proto} = \mathcal{L}_{assign} + \frac{1}{2}\mathcal{L}_{rec}. \tag{6}$$

The prototype objective includes no explicit constraint to prevent assignment collapse.

The homogeneous alignment loss uses maximum mean discrepancy [17] to measure the distance between distributions of modality-common features in a reproducing kernel Hilbert space. Since the three modalities have different sequence lengths, MMD is not computed on the full token sequences. We use the first classification-level common token of each sample as its global common representation. Each global common representation is L2-normalized before MMD computation; c_m denotes the collection of these normalized representations for modality m in a mini-batch. For each modality pair, we use the biased mini-batch estimator of squared MMD, retaining the diagonal entries of both within-modality kernel matrices. We use a single radial basis function (RBF) kernel $k(x, y) = \exp(-\|x - y\|_2^2/(2b))$ with $b = 1.85$. The homogeneous alignment loss is the sum over the three modality pairs:

$$\mathcal{L}_{homo} = \widehat{\mathrm{MMD}}^2(c_t, c_a) + \widehat{\mathrm{MMD}}^2(c_t, c_v) + \widehat{\mathrm{MMD}}^2(c_a, c_v). \tag{7}$$

The normalization and bandwidth were selected by validation MAE in a CMU-MOSI seed-1 sweep and held fixed across subsequent runs. The same normalization and kernel are used for the MMD training loss and representation evaluation.

The decoupled modality representation is obtained by concatenating and re-projecting the specific and common components:

$$H_m = W_o[S_m; C_m], \tag{8}$$

where H_m is the output representation of MDSR for modality m. The module therefore produces decoupled feature sequences H_t, H_a, and H_v, together with the auxiliary losses $\mathcal{L}_{dec}$, $\mathcal{L}_{proto}$, and $\mathcal{L}_{homo}$.

3.4. Text-Guided Cross-Modal Alignment

The streams produced by MDSR are still not directly synchronized at the token or frame level. Audio and visual sequences are continuous and usually longer than the text sequence, and their local cues may only be useful when they correspond to sentiment-bearing words or phrases. Textual representations usually provide explicit sentiment semantics, whereas acoustic and visual streams offer complementary but noisier nonverbal evidence. Jin et al. [47] further showed that reducing excessive text dominance can improve multimodal semantic enhancement. We therefore design a text-guided cross-modal alignment (TGCA) module. TGCA uses text as the semantic anchor, but limits audio-visual transfer to evidence that is relevant to the textual sentiment context.

Let H_t, H_a, and H_v denote the decoupled representations produced by MDSR. We first concatenate the decoupled acoustic and visual sequences along the temporal dimension to build a unified audio-visual context. In the text-centered alignment notation, the text sequence is used as the query, and the audio-visual context is used as the key and value to

obtain the audio-visual-to-text representation $H_{av\to t}$. In this way, each text token can attend to nonverbal segments that provide complementary evidence, without forcing all audio and visual frames to be equally fused into the final representation. The alignment process is followed by residual normalization and a feed-forward block:

$$H_{av} = [H_a; H_v], \qquad H_{av\to t} = \mathrm{CA}(H_t, H_{av}), \tag{9}$$

$$Z_t = \mathrm{LN}\big(H_t + H_{av\to t}\big), \qquad H_{tav} = \mathrm{LN}\big(Z_t + \mathrm{FFN}(Z_t)\big). \tag{10}$$

The output H_{tav} keeps the same sequence length as the text stream, but each textual position has been enhanced with selected acoustic and visual evidence. This text-length aligned representation provides a compact multimodal sequence for the subsequent bottleneck fusion stage.

3.5. Context-Calibrated Bottleneck Fusion

The aligned multimodal text sequence H_{tav} still needs to be integrated with the three decoupled modality streams for sentiment prediction. Directly applying dense attention over all modality tokens may introduce redundant interactions and increase computational cost. CBF retains the progressive token compression and bidirectional exchange between bottleneck tokens and modality streams used in DashFusion [39]. After the bottleneck tokens attend to the three streams, we add a residual correction derived from their global modality context before updating the streams.

The internal structure of one CBF layer is shown in Figure 2. At the l-th level, the layer receives the previous multimodal bottleneck representation H_{tav}^{l-1} and the three modality streams H_t^{l-1}, H_a^{l-1}, and H_v^{l-1}. A Transformer layer first updates the full input sequence; its leading tokens are then retained as bottleneck tokens B^l, which are updated into modality-aware bottleneck tokens $\tilde{B}^l$ through the Multi-CA block. The updated bottleneck representation is then calibrated by the global context vector g^l extracted from the classification-level modality features. Finally, the calibrated bottleneck tokens guide the CA block to update the text, acoustic, and visual streams. The left part of the figure gives a detailed view of the Multi-CA block, where the bottleneck tokens act as queries and the three modality streams provide keys and values.

CBF takes H_{tav} as the initial multimodal sequence and uses H_t, H_a, and H_v as the modality streams to be fused. We set $H_{tav}^0 = H_{tav}$ and $H_m^0 = H_m$, where $m \in \{t, a, v\}$. The first layer therefore receives the full aligned sequence. Let L denote the number of hierarchy levels and p the maximum number of tokens retained after the first Transformer layer. In our implementation, p is set to 8. At the l-th level, a Transformer layer first redistributes the information in the previous bottleneck representation, and only the first n_l tokens are retained:

$$B^l = \text{Transformer-layer}(H_{tav}^{l-1})[0 : n_l], \tag{11}$$

where B^l denotes the bottleneck tokens at level l, and H_{tav}^{l-1} is the multimodal bottleneck representation from the previous level. The actual retained length is defined by

$$n_0 = T_t, \qquad n_l = \min\left\{n_{l-1}, \max\left(1, \left\lfloor \frac{p}{2^{l-1}} \right\rfloor\right)\right\}. \tag{12}$$

Here, T_t denotes the text sequence length after padding. CBF computes attention over all input tokens. If the input sequence is shorter than the current retention limit, all its tokens are retained. By reducing the token retention limit layer by layer, the module performs progressive compression along the sequence dimension.

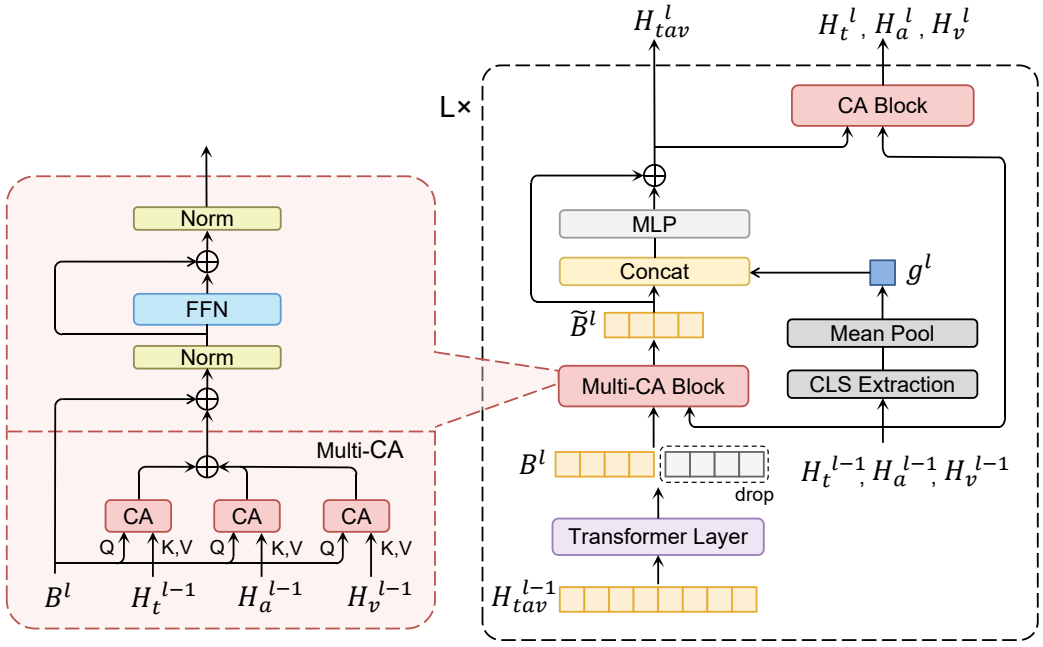


Figure 2. The architecture of the context-calibrated bottleneck fusion (CBF) layer.

The retained bottleneck tokens then act as queries to collect information from the text, acoustic, and visual streams through the Multi-CA block:

$$\hat{B}^l = \mathrm{LN}\big(B^l + \text{Multi-CA}\big(B^l, \{H_m^{l-1}\}_{m\in\{t,a,v\}}\big)\big), \tag{13}$$

$$\tilde{B}^l = \mathrm{LN}\big(\hat{B}^l + \mathrm{FFN}(\hat{B}^l)\big), \tag{14}$$

$$\text{Multi-CA}\big(B^l, \{H_m^{l-1}\}_{m\in\{t,a,v\}}\big) = \sum_{m\in\{t,a,v\}} \mathrm{CA}(B^l, H_m^{l-1}). \tag{15}$$

Here, $\hat{B}^l$ and $\tilde{B}^l$ denote the bottleneck tokens after multimodal cross-attention and feed-forward updating, respectively. Both updates include residual addition followed by layer normalization. H_m^{l-1} represents the modality feature of modality m from the previous level.

To prevent compact bottleneck tokens from losing utterance-level cues during progressive compression, we insert a lightweight context calibration after bottleneck updating. The global context is obtained from the classification-level tokens of the three modality streams:

$$g^l = \frac{1}{3}\big(H_t^{l-1}[0] + H_a^{l-1}[0] + H_v^{l-1}[0]\big). \tag{16}$$

The context vector is broadcast to the length of $\tilde{B}^l$ and used to refine the bottleneck representation through a residual correction:

$$H_{tav}^l = \tilde{B}^l + \gamma^l \Phi^l\big([\mathrm{LN}(\tilde{B}^l); \mathrm{LN}(g^l)]\big), \tag{17}$$

where H_{tav}^l denotes the calibrated multimodal bottleneck representation at level l, $\Phi^l(\cdot)$ is a lightweight feed-forward calibration function, and γ^l is a learnable residual scale that adaptively controls the strength of context correction. This design allows CBF to refine the bottleneck representation while preserving the original multimodal aggregation pathway.

The calibrated bottleneck tokens are then used to update each modality stream through cross-attention:

$$Z_m^l = \mathrm{LN}\big(H_m^{l-1} + \mathrm{CA}(H_m^{l-1}, H_{tav}^l)\big), \tag{18}$$

$$H_m^l = \mathrm{LN}\big(Z_m^l + \mathrm{FFN}(Z_m^l)\big), \qquad m \in \{t,a,v\}. \tag{19}$$

This bidirectional process allows the bottleneck tokens to summarize multimodal evidence and then return the compact interaction information to each modality stream. At the final

hierarchy level, the first calibrated bottleneck token and the classification-level tokens of the three modality streams are concatenated to form the fused representation:

$$h_f = [H_{tav}^L[0]; H_t^L[0]; H_a^L[0]; H_v^L[0]]. \tag{20}$$

The fused feature h_f is then passed to the regression head for sentiment prediction.

3.6. Regression Head and Training Objective

The fused representation h_f contains the calibrated bottleneck summary and the classification-level representations of the three modality streams. We feed h_f into a multilayer perceptron to predict the sentiment score:

$$\hat{y} = \mathrm{MLP}(h_f), \tag{21}$$

where $\hat{y}$ denotes the predicted sentiment intensity. The MLP maps the fused feature into a scalar output through stacked linear layers with nonlinear activation.

Since the sentiment annotations in CMU-MOSI, CMU-MOSEI, and CH-SIMS are continuous values, the main task is optimized as regression. Given N training samples, the prediction loss is computed by mean squared error:

$$\mathcal{L}_{pred} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{22}$$

where y_i and $\hat{y}_i$ denote the ground-truth sentiment score and predicted score of the i-th sample, respectively.

In addition to the regression objective, DTA-CBF uses the auxiliary losses introduced in MDSR to regularize the representation space. The decoupling loss $\mathcal{L}_{dec}$ encourages modality-specific and modality-common components to encode non-redundant information. The prototype loss $\mathcal{L}_{proto}$ preserves cross-modal semantic consistency in modality-specific prototype spaces, while the homogeneous alignment loss $\mathcal{L}_{homo}$ reduces the distribution discrepancy among modality-common representations. The overall training objective is formulated as

$$\mathcal{L}_{all} = \mathcal{L}_{pred} + \alpha\mathcal{L}_{dec} + \beta\mathcal{L}_{proto} + \lambda\mathcal{L}_{homo}, \tag{23}$$

where α, β, and λ control the relative contributions of the three constraint losses. During inference, only the predicted score $\hat{y}$ produced by the regression head is used for evaluation.

4. Experiments

4.1. Datasets

We evaluate the proposed framework on CMU-MOSI [5], CMU-MOSEI [6], and CH-SIMS [7]. The split settings of these datasets are reported in Table 1, and a brief introduction to each benchmark is provided below.

CMU-MOSI is an English utterance-level benchmark built from online opinion videos. It contains 2199 annotated video segments from 93 videos and 89 speakers. The sentiment label is a real-valued intensity score from −3 to 3, where negative values indicate negative sentiment and positive values indicate positive sentiment.

CMU-MOSEI comprises 22,856 annotated video clips collected from YouTube. It extends CMU-MOSI by increasing the number of samples, utterances, speakers, and topics. Each utterance is assigned a sentiment intensity score ranging from −3 to +3, as in CMU-MOSI.

CH-SIMS is a Chinese benchmark for multimodal sentiment analysis. It contains 2281 video segments collected from movies, television series, and variety shows. The dataset provides both multimodal and unimodal annotations; in this paper, only the multimodal sentiment labels are used. Its sentiment score ranges from -1 to 1.

Table 1. Dataset statistics in CMU-MOSI, CMU-MOSEI, and CH-SIMS.

Dataset	Train	Valid	Test	All
CMU-MOSI	1284	229	686	2199
CMU-MOSEI	16,326	1871	4659	22,856
CH-SIMS	1368	456	457	2281

4.2. Evaluation Metrics

For our experiments, the evaluation metrics are defined as follows. For regression, we report mean absolute error (MAE) and Pearson correlation (Corr) on CMU-MOSI [5], CMU-MOSEI [6], and CH-SIMS [7]. For classification, we report binary accuracy (Acc-2) and weighted F1 score on all three datasets. F1 is weighted by the number of samples in each ground-truth class. We report classification and regression results averaged over three runs with random seeds 1, 11, and 111. Lower MAE is better, while higher values are better for the remaining metrics.

For CMU-MOSI and CMU-MOSEI, we additionally report seven-class accuracy (Acc-7). Acc-2 and F1 follow the Non0 setting, excluding samples with zero ground-truth labels. The remaining labels and predictions are classified as positive when greater than zero and negative otherwise. Acc-7 is computed by clipping predictions and labels to $[-3, 3]$ and rounding to the nearest integer, with half-integer ties rounded to the nearest even integer. MAE and Corr are computed from the original, unclipped continuous values.

For CH-SIMS, we additionally report three-class accuracy (Acc-3) and five-class accuracy (Acc-5). Predictions and labels are clipped to $[-1, 1]$ before calculating all metrics, including MAE and Corr. Acc-2 and F1 use $[-1, 0]$ and $(0, 1]$ as the non-positive and positive classes, respectively. Acc-3 uses $[-1, -0.1]$, $(-0.1, 0.1]$, and $(0.1, 1]$ for the negative, neutral, and positive classes, respectively. Acc-5 uses the ordered intervals $[-1, -0.7]$, $(-0.7, -0.1]$, $(-0.1, 0.1]$, $(0.1, 0.7]$, and $(0.7, 1]$. These intervals are applied identically to predictions and ground-truth labels. Zero-label samples are retained; zero values belong to the non-positive class for Acc-2 and the neutral class for Acc-3 and Acc-5.

4.3. Implementation Details

The implementation uses Python 3.9.25, PyTorch 1.13.1 (built with CUDA 11.7), and Transformers 4.30.0. We optimize the model using AdamW. The learning rate is set separately for each dataset, and the feature dimensionality of each modality is unified to 128 to ensure consistent inputs. Model training is conducted on a single NVIDIA RTX 4090 GPU with 24 GB of memory. Additional implementation details are shown in Table 2.

Table 2. Detailed configuration of hyperparameters for three MSA datasets.

Hyperparameters	CMU-MOSI	CMU-MOSEI	CH-SIMS
Batch size	16	16	16
Learning rate	1×10^{-5}	1×10^{-5}	5×10^{-5}
Epochs	100	25	100
Optimizer	AdamW	AdamW	AdamW
Weight decay	0.001	0.001	0.001
Warm-up epochs	10	1	10

Table 2. *Cont.*

Hyperparameters	CMU-MOSI	CMU-MOSEI	CH-SIMS
Feature dimension d	128	128	128
Transformer layers [A, V]	[2, 2]	[4, 4]	[2, 2]
Attention heads [A, V]	[8, 8]	[8, 8]	[8, 8]
Text-guided interaction heads	8	8	8
Fusion depth	2	3	2
First-layer token retention limit	8	8	8
Number of prototypes K	8	8	8
Dropout [T, A, V]	[0.1, 0.1, 0.1]	[0.1, 0.1, 0.1]	[0.1, 0.1, 0.1]
Decoupling loss weight α	0.1	0.4	0.1
Prototype loss weight β	0.05	0.1	0.1
Homogeneous alignment loss weight λ	0.05	0.1	0.1

4.4. Baselines

To evaluate the effectiveness of our model, we selected several state-of-the-art baselines, as stated in Tables 3 and 4. The details of the competitors are listed as follows.

TFN. The Tensor Fusion Network [8] explicitly models unimodal, bimodal, and trimodal interactions by constructing a high-dimensional tensor with outer-product operations.

LMF. Low-rank Multimodal Fusion [9] reduces the computational cost of tensor fusion by decomposing high-order fusion tensors into low-rank modality-specific factors.

MulT. The Multimodal Transformer [21] uses directional pairwise cross-modal attention to model interactions among unaligned multimodal sequences and learn adaptive cross-modal correspondences.

MAG-BERT. MAG-BERT [25] introduces multimodal adaptation gates into a BERT-based language backbone, allowing nonverbal information to refine textual representations during fusion.

MISA. MISA [11] projects each modality into modality-invariant and modality-specific subspaces, and performs fusion on the disentangled representations to reduce cross-modal heterogeneity.

Self-MM. Self-MM [26] generates unimodal supervision signals from multimodal labels and jointly optimizes multimodal and unimodal representation learning.

MMIM. MMIM [27] maximizes mutual information at different fusion levels to preserve task-relevant information during multimodal representation learning.

ConFEDE. ConFEDE [48] decomposes each modality into similarity and dissimilarity features, and uses text similarity as the anchor for contrastive learning.

ICCN. ICCN [49] constructs text–audio and text–video outer-product features, and uses deep canonical correlation analysis to align them before sentiment prediction.

DashFusion. DashFusion [39] combines dual-stream alignment and hierarchical bottleneck fusion. It aligns multimodal features at temporal and semantic levels, and progressively integrates them through compressed bottleneck tokens.

TCDN. TCDN [50] performs trimodal collaborative disentanglement to suppress irrelevant modality-specific noise and enhance inter-/intra-modal representation learning.

Semi-IIN. Semi-IIN [38] separately captures intra-modal and inter-modal interactions with masked attention, and uses a gating mechanism to dynamically select useful interaction patterns. It further incorporates self-training to exploit unlabeled data.

DecAlign. DecAlign [33] decouples modality-unique and modality-common representations and applies hierarchical alignment through prototype-guided optimal transport and semantic distribution matching.

Table 3. Performance comparison on CMU-MOSI and CMU-MOSEI. † results are from [39], and ‡ results are from [38]. For DTA-CBF and the reproduced baselines, results are reported as mean ± standard deviation over three runs. ↑ indicates that higher values are better, whereas ↓ indicates that lower values are better. Bold values indicate the best result for each metric within each dataset.

Method	CMU-MOSI				
	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑	Acc-7 ↑
TFN †	0.901	0.698	80.80	80.70	34.90
LMF †	0.917	0.695	82.50	82.40	33.20
MulT †	0.871	0.698	83.00	82.80	40.00
MAG-BERT †	0.790	0.766	83.54	83.58	41.43
MISA †	0.783	0.761	83.40	83.60	42.30
Self-MM †	0.708	0.796	85.46	85.43	46.67
MMIM †	0.714	0.794	85.11	85.24	46.20
ConFEDE ‡	0.742	0.784	85.52	85.52	42.27
ICCN ‡	0.862	0.714	83.00	83.00	39.00
TCDN ‡	0.697	0.805	87.10	87.20	-
Semi-IIN ‡	0.679	0.822	87.04	87.00	46.50
DashFusion	0.714 ± 0.0688	0.812 ± 0.0225	85.98 ± 1.10	85.99 ± 1.11	46.02 ± 1.93
DecAlign	0.684 ± 0.0151	0.829 ± 0.0050	86.64 ± 0.87	86.65 ± 0.87	46.84 ± 0.75
DTA-CBF	**0.646** ± 0.0307	**0.843** ± 0.0146	**87.40** ± 0.35	**87.37** ± 0.32	**47.86** ± 1.21
Method	CMU-MOSEI				
	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑	Acc-7 ↑
TFN †	0.593	0.700	82.50	82.10	50.20
LMF †	0.623	0.677	82.00	82.10	48.00
MulT †	0.559	0.733	84.63	84.52	52.84
MAG-BERT †	0.583	0.741	83.86	83.88	50.41
MISA †	0.555	0.756	85.50	85.30	52.20
Self-MM †	0.531	0.765	85.15	84.90	53.87
MMIM †	0.538	0.763	85.24	85.19	53.23
ConFEDE ‡	0.522	0.780	85.82	85.83	54.86
ICCN ‡	0.565	0.704	84.20	84.20	51.60
TCDN ‡	0.521	0.782	87.50	87.20	-
Semi-IIN ‡	0.497	0.804	87.70	87.65	**55.89**
DashFusion	0.516 ± 0.0089	0.794 ± 0.0010	87.02 ± 0.55	87.07 ± 0.48	54.07 ± 1.34
DecAlign	0.529 ± 0.0063	0.778 ± 0.0108	86.76 ± 0.92	86.58 ± 0.58	54.30 ± 1.18
DTA-CBF	**0.496** ± 0.0055	**0.812** ± 0.0011	**88.10** ± 0.23	**88.10** ± 0.21	55.68 ± 0.73

Table 4. Performance comparison on CH-SIMS. † results are from [39]. For DTA-CBF and the reproduced baselines, results are reported as mean ± standard deviation over three runs. ↑ indicates that higher values are better, whereas ↓ indicates that lower values are better. Bold values indicate the best result for each metric.

Method	**CH-SIMS**					
	MAE ↓	**Corr ↑**	**Acc-2 ↑**	**F1 ↑**	**Acc-3 ↑**	**Acc-5 ↑**
TFN †	0.432	0.591	78.38	78.62	65.12	39.30
LMF †	0.441	0.576	77.77	77.88	64.48	40.53
MulT †	0.446	0.579	78.24	78.54	64.38	40.06
MISA †	0.447	0.563	76.54	76.59	-	-
MAG-BERT †	0.492	0.399	74.44	71.75	-	-
Self-MM †	0.428	0.585	78.56	78.60	64.68	41.69
DashFusion	0.418 ± 0.0053	0.586 ± 0.0070	77.97 ± 0.55	77.90 ± 1.18	64.77 ± 1.09	43.40 ± 0.50
DecAlign	0.411 ± 0.0031	**0.622** ± 0.0031	78.63 ± 0.88	78.68 ± 0.91	**67.76** ± 0.45	42.74 ± 0.91
DTA-CBF	**0.404** ± 0.0036	0.612 ± 0.0055	**79.65** ± 0.66	**79.75** ± 0.69	66.96 ± 0.44	**45.66** ± 0.77

4.5. Performance Comparison

Tables 3 and 4 present the comparison results of DTA-CBF and representative baseline methods on CMU-MOSI, CMU-MOSEI, and CH-SIMS. These baselines cover tensor-based fusion, cross-modal attention, pretrained-language-model-based fusion, representation disentanglement, mutual information maximization, interaction modeling, and hierarchical bottleneck fusion.

DashFusion and DecAlign are reproduced to cover alignment with bottleneck fusion and modality disentanglement with hierarchical alignment, respectively. We use publicly available implementations and the same labeled training, validation, and test splits as DTA-CBF on all three datasets. All three models share the preprocessed audio and visual features, labels, evaluation scripts, and text-encoding settings described in Section 3.2. Text is re-tokenized from raw inputs and dynamically padded within each batch. Checkpoints are selected by the lowest validation MAE for test evaluation, while optimization hyperparameters remain model-specific. Results for other baselines are taken from the sources identified in the table captions and serve as cross-protocol references.

On CMU-MOSI, DTA-CBF achieves the lowest reported MAE and the highest reported Corr, Acc-2, F1, and Acc-7 among the listed results. Under the controlled reproduction protocol, it reduces mean MAE from 0.684 to 0.646 and increases mean Corr from 0.829 to 0.843 compared with DecAlign. Relative to DecAlign, the strongest reproduced baseline on all five metrics, mean Acc-2, F1, and Acc-7 increase by 0.76, 0.72, and 1.02 percentage points, respectively. These comparisons show improvements in the reported mean scores across both regression and classification metrics.

On CMU-MOSEI, DTA-CBF achieves the highest reported Corr, Acc-2, and F1 and the lowest reported MAE among the listed results, whereas Semi-IIN reports the highest Acc-7. Under the controlled reproduction protocol, DTA-CBF reduces mean MAE by 0.020 and increases mean Corr by 0.018 compared with DashFusion. Mean Acc-2 and F1 improve by 1.08 and 1.03 percentage points, respectively, over DashFusion, while mean Acc-7 improves by 1.38 percentage points over DecAlign, the strongest reproduced baseline on this metric. Relative to the published Semi-IIN results, DTA-CBF has an MAE that is 0.001 lower and a Corr that is 0.008 higher; Acc-2 and F1 are higher by 0.40 and 0.45 percentage points, respectively; and Acc-7 is lower by 0.21 percentage points. This comparison with Semi-IIN is a cross-protocol reference, so these differences are descriptive rather than a controlled estimate of comparative performance.

On CH-SIMS, under the controlled reproduction protocol, DTA-CBF achieves the lowest mean MAE of 0.404. It also achieves the highest mean Acc-2, F1, and Acc-5 of 79.65%, 79.75%, and 45.66%, respectively, whereas DecAlign obtains higher mean Corr and Acc-3. Relative to DecAlign, DTA-CBF reduces MAE by 0.007 and improves Acc-2, F1, and Acc-5 by 1.02, 1.07, and 2.92 percentage points, respectively; its Corr is lower by 0.010, and its Acc-3 is lower by 0.80 percentage points. Published-paper results in Table 4 provide cross-protocol context. These are descriptive comparisons of reported means and do not establish statistical significance.

To assess the statistical significance of the observed differences, Table 5 reports two-sided paired prediction-swap permutation tests based on three independent runs and 10,000 permutations. DTA-CBF is compared with DecAlign on CMU-MOSI and CH-SIMS and with DashFusion on CMU-MOSEI. Δ is the DTA-CBF score minus the comparator score, calculated from the means in Tables 3 and 4; negative values indicate lower MAE, whereas positive values indicate higher scores for the other metrics. The table also reports the unadjusted permutation p-value (p_{perm}) and the Holm-adjusted value (p_{Holm}), with correction applied jointly across all 16 comparisons.

Table 5. Paired significance tests against reproduced comparators. Classification Δ values are in percentage points.

Dataset	Comparator	Metric	Δ	p_{perm}	p_{Holm}
CMU-MOSI	DecAlign	MAE	−0.038	0.0012	0.0168
		Corr	+0.014	0.0034	0.0238
		Acc-2	+0.76	0.0028	0.0225
		F1	+0.72	0.0025	0.0225
		Acc-7	+1.02	0.0021	0.0210
CMU-MOSEI	DashFusion	MAE	−0.020	0.0001	0.0016
		Corr	+0.018	0.0016	0.0192
		Acc-2	+1.08	0.0019	0.0209
		F1	+1.03	0.0014	0.0182
		Acc-7	+1.61	0.0003	0.0045
CH-SIMS	DecAlign	MAE	−0.007	0.0710	0.1420
		Corr	−0.010	0.1360	0.1420
		Acc-2	+1.02	0.0051	0.0255
		F1	+1.07	0.0065	0.0260
		Acc-3	−0.80	0.0448	0.1344
		Acc-5	+2.92	0.0042	0.0252

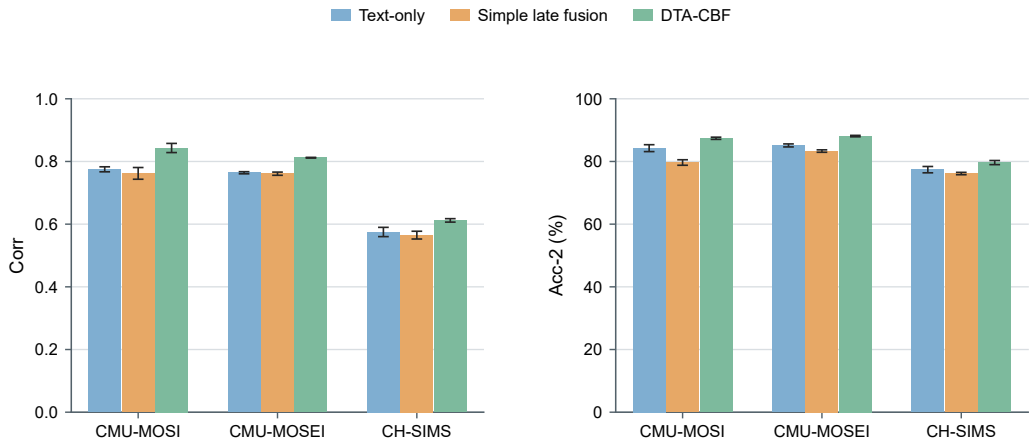
After Holm correction, DTA-CBF shows statistically significant improvements over the selected reproduced comparator on all five metrics for CMU-MOSI and CMU-MOSEI. On CH-SIMS, significant improvements are observed in Acc-2, F1, and Acc-5, whereas the differences in MAE, Corr, and Acc-3 are not statistically significant. These results indicate consistent gains on the two English-language datasets, while the improvements on CH-SIMS vary by metric.

4.6. Ablation Study

To assess the contribution of multimodal modeling under a shared text backbone, we compare DTA-CBF with Text-only and Simple late fusion. All three models use the same text encoder and training protocol for each dataset. Text-only predicts sentiment from text representations alone, whereas Simple late fusion equally averages the predictions of separately trained text, acoustic, and visual branches. As shown in Figure 3, DTA-CBF achieves higher mean Corr and Acc-2 than both controls across all three datasets. The gains over Text-only support the value of incorporating acoustic and visual information under the same text backbone. The advantage over Simple late fusion further suggests that explicitly modeling cross-modal relationships helps capture complementary sentiment cues beyond independent unimodal predictions.

We next evaluate the contribution of each component on CMU-MOSI. We compare the full DTA-CBF with three ablated variants, each of which removes one major module while keeping the remaining training settings unchanged. This design directly examines whether modality decoupling, text-guided cross-modal alignment, and context-calibrated bottleneck fusion provide complementary gains.

In w/o MDSR, the decoupling and re-projection operations are bypassed, and the encoded modality sequences are passed directly to TGCA and CBF. The auxiliary losses $\mathcal{L}_{dec}$, $\mathcal{L}_{proto}$, and $\mathcal{L}_{homo}$ are also removed. In w/o TGCA, the alignment module is bypassed by setting $H_{tav} = H_t$, while retaining H_t, H_a, and H_v as the three modality inputs to CBF. In w/o CBF, MDSR and TGCA are retained, and the first tokens of the four fusion inputs are directly concatenated as $h_f = [H_{tav}[0]; H_t[0]; H_a[0]; H_v[0]]$. This $4d$-dimensional representation is passed to the unchanged regression head.



(**a**) Correlation (**b**) Binary classification

Figure 3. Comparison using the same text encoder. Bars and error bars represent mean ± standard deviation.

Table 6 shows that removing any one of the three modules consistently degrades performance, indicating that MDSR, TGCA, and CBF each contribute to predictive performance. Among the ablated variants, w/o MDSR causes the largest overall degradation. Compared with the full DTA-CBF, MAE increases from 0.646 to 0.690, Corr decreases from 0.843 to 0.814, and Acc-2, F1, and Acc-7 decrease by 1.48, 1.48, and 1.79 percentage points, respectively. These results support the predictive utility of MDSR before cross-modal interaction, but do not directly establish semantic disentanglement.

Table 6. Ablation study on CMU-MOSI. Bold values indicate the best result for each metric. ↑ indicates that higher values are better, whereas ↓ indicates that lower values are better.

Method	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑	Acc-7 ↑
DTA-CBF	**0.646**	**0.843**	**87.40**	**87.37**	**47.86**
w/o MDSR	0.690	0.814	85.92	85.89	46.07
w/o TGCA	0.659	0.830	86.84	86.83	47.28
w/o CBF	0.677	0.821	86.38	86.35	46.55

Removing TGCA produces a smaller but still consistent performance drop. The w/o TGCA variant increases MAE from 0.646 to 0.659 and reduces Corr from 0.843 to 0.830. Acc-2, F1, and Acc-7 decrease by 0.56, 0.54, and 0.58 percentage points, respectively. Since this variant retains the decoupled representations but removes the text-guided retrieval of acoustic and visual cues, the results indicate that TGCA improves the use of nonverbal information by anchoring cross-modal alignment to textual sentiment semantics. The relatively moderate degradation also suggests that the representations learned by MDSR already contain useful sentiment structure, while TGCA further refines how this structure is connected across modalities.

The w/o CBF variant also shows clear degradation and performs worse than w/o TGCA on all reported metrics. Without context-calibrated bottleneck fusion, MAE increases to 0.677 and Corr decreases to 0.821. Acc-2, F1, and Acc-7 decrease by 1.02, 1.02, and 1.31 percentage points, respectively. This pattern indicates that replacing CBF with direct concatenation weakens both sentiment intensity estimation and compact multimodal integration.

Table 7 compares all eight combinations of the three auxiliary losses in MDSR on CMU-MOSI. When all auxiliary losses are removed, the model relies only on the main regression objective and obtains the weakest result in this table. Nevertheless, it remains stronger than the w/o MDSR variant in Table 6, indicating that the decoupling structure

itself is still useful. With only the orthogonality constraint active, MAE decreases from 0.674 to 0.658, Corr increases from 0.823 to 0.829, and the classification metrics also improve. Prototype alignment and homogeneous alignment also improve all five metrics when used individually; their respective MAEs are 0.661 and 0.669.

Table 7. Factorial ablation of three auxiliary losses on CMU-MOSI. Checkmarks indicate active losses. Bold values indicate the best result for each metric. ↑ indicates that higher values are better, whereas ↓ indicates that lower values are better.

$\mathcal{L}_{dec}$	$\mathcal{L}_{proto}$	$\mathcal{L}_{homo}$	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑	Acc-7 ↑
			0.674	0.823	86.33	86.31	46.46
✓			0.658	0.829	86.64	86.63	46.98
	✓		0.661	0.829	86.49	86.54	47.03
		✓	0.669	0.824	86.43	86.38	46.65
✓	✓		0.650	0.840	87.04	87.01	47.57
✓		✓	0.654	0.836	86.84	86.80	47.52
	✓	✓	0.660	0.832	86.73	86.73	47.23
✓	✓	✓	**0.646**	**0.843**	**87.40**	**87.37**	**47.86**

Adding homogeneous alignment to the orthogonality constraint further improves prediction, reducing MAE to 0.654 and increasing Acc-7 to 47.52%. Pairing orthogonality with prototype alignment instead yields the best scores among the three two-loss combinations, with MAE reduced to 0.650 and F1 increased to 87.01%. Without the orthogonality constraint, adding homogeneous alignment reduces MAE by 0.005 when prototype alignment is disabled, while it increases Corr by 0.003 and Acc-2 by 0.24 percentage points when prototype alignment is enabled. With orthogonality active, homogeneous alignment reduces MAE by 0.004 in both cases.

The complete objective achieves the best MAE, Corr, Acc-2, F1, and Acc-7 scores. Compared with the model without auxiliary losses, it reduces MAE by 0.028 and improves F1 by 1.06 percentage points. Each loss thus improves prediction even when the other two are already active.

To examine the effect of homogeneous alignment on the learned common representations, we compare MMD^2 between models trained with and without this loss on CMU-MOSI. As shown in Table 8, enabling homogeneous alignment reduces MMD^2 for all three modality pairs, with the total decreasing from 2.0829 to 0.7550. These results support its role in reducing cross-modal distribution discrepancy under the specified kernel.

Table 8. Pairwise and total squared maximum mean discrepancy (MMD^2) on CMU-MOSI, averaged over three runs. Total MMD^2 is the sum of the three pairwise values.

Homo	Text–Audio	Text–Vision	Audio–Vision	Total
Disabled	0.7210	0.6902	0.6716	2.0829
Enabled	0.2090	0.2389	0.3071	0.7550

To compare the effectiveness of different fusion mechanisms, we replaced the proposed CBF module with several commonly used fusion strategies while keeping the remaining components unchanged. All variants receive the four streams H_{tav}, H_t, H_a, and H_v and use the same regression head. Concat is identical to w/o CBF in Table 6, and the same results are reported in both tables. The computation cost is measured by fusion-only multiply–add operations (MAdds) per utterance. For this calculation, text, acoustic, and visual sequence lengths at the fusion input are fixed at 50, 376, and 501 across all variants. Each multiplication and its associated accumulation count as one MAdd. The count includes matrix multiplications in linear layers and attention, excluding layer normalization, softmax, dropout,

elementwise additions, and activations. In Table 9, the simplest concatenation strategy incurs no additional MAdds, but it obtains the weakest results among the compared fusion strategies. This indicates that although the decoupled and aligned representations already contain useful sentiment cues, directly stacking multimodal features is still insufficient for modeling cross-modal interactions.

Table 9. Effects of different fusion mechanisms on CMU-MOSI. MAdds denotes multiply–add operations, and M denotes million. Bold values indicate the best result for each predictive metric and the lowest MAdds. ↑ indicates that higher values are better, whereas ↓ indicates that lower values are better.

Method	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑	Acc-7 ↑	MAdds
Concat	0.677	0.821	86.38	86.35	46.55	**0**
Concat&SA	**0.642**	0.840	87.09	87.03	47.48	356.4 M
CA	0.670	0.829	86.48	86.50	47.08	139.1 M
BF	0.657	0.833	86.84	86.82	47.33	435.3 M
HBF	0.655	0.832	86.94	86.84	47.47	227.8 M
CBF	0.646	**0.843**	**87.40**	**87.37**	**47.86**	228.4 M

Applying self-attention to the concatenated multimodal sequence substantially improves all metrics over direct concatenation. It achieves the lowest MAE, showing that dense global token interaction can strengthen multimodal regression. However, this improvement is obtained with 356.4 million MAdds because all tokens participate in self-attention. CA introduces explicit cross-modal interaction with a lower computational cost than Concat&SA, but it does not achieve consistent advantages over bottleneck-based fusion. In particular, CA has a higher MAE and lower Corr, Acc-2, and F1 than BF and HBF, indicating that direct cross-attention may be less effective in preserving compact multimodal representations.

BF further uses bottleneck tokens to mediate information exchange and improves MAE, Corr, Acc-2, and F1 over CA, but repeatedly inserting bottleneck tokens into multiple modality streams leads to the largest computational cost among all compared methods. HBF reduces the cost by progressively compressing bottleneck tokens and further improves most metrics over BF, especially Acc-7. By contrast, the proposed CBF achieves the best Corr, Acc-2, F1, and Acc-7 scores while requiring only 228.4 million MAdds, which is close to HBF and much lower than Concat&SA and BF. Although Concat&SA obtains a slightly lower MAE, its computational cost is substantially higher. These results indicate that context-calibrated bottleneck fusion provides a better balance between sentiment prediction performance and fusion computation cost.

Both HBF and CBF restrict cross-modal exchange to a progressively shorter sequence of bottleneck tokens. CBF adds global modality summaries to the retained tokens, which may help preserve utterance-level context during compression. These summaries may still miss brief, localized sentiment cues.

4.7. Computational Efficiency Analysis

To complement predictive performance with a practical assessment of computational cost, we profile all models according to the reproduction protocols described above, including the text encoder and preprocessing used for each dataset. All models are evaluated on the same NVIDIA RTX 4090 GPU with a batch size of 8. We count floating-point operations (FLOPs) for one forward pass of the complete model on a single sample. We report per-sample inference latency as the mean ± standard deviation across five complete test-set passes. Training time is averaged over all completed epochs in each full training run. GPU memory (GPU Mem.) is the maximum allocated memory during the full training run.

Table 10 summarizes the computational profiles of the complete models. The three models have similar parameter counts and prediction FLOPs within each dataset. DTA-CBF

has the lowest measured inference latency on CMU-MOSI, whereas DecAlign is fastest on CMU-MOSEI and CH-SIMS. DTA-CBF nevertheless has the shortest measured training time per epoch on all three datasets. Its peak training memory is lower than that of DashFusion on every dataset and lower than DecAlign on CMU-MOSI and CMU-MOSEI, but is slightly higher than DecAlign on CH-SIMS. During training, DashFusion encodes additional positive and negative examples for its contrastive loss, which requires storing more activations for backpropagation. It also recomputes pairwise similarities across the training set every two epochs, adding to training time particularly on the larger CMU-MOSEI dataset.

Table 10. Computational profiles of the complete models. Params, FLOPs, GPU Mem., Infer. Time, and Train Time are reported in M, G, GB, ms/sample, and min/epoch, respectively.

Model	Params	FLOPs	GPU Mem.	Infer. Time	Train Time
CMU-MOSI					
DashFusion	357.87	80.59	18.870	3.17 ± 0.06	1.11
DecAlign	356.22	80.30	7.969	2.61 ± 0.03	0.56
DTA-CBF	357.98	80.62	7.421	2.52 ± 0.30	0.38
CMU-MOSEI					
DashFusion	359.09	82.17	23.362	3.40 ± 0.06	157.83
DecAlign	356.36	80.32	8.654	2.47 ± 0.05	6.70
DTA-CBF	359.24	82.20	8.350	3.00 ± 0.07	5.55
CH-SIMS					
DashFusion	104.88	23.26	5.455	2.05 ± 0.08	0.64
DecAlign	102.67	22.54	1.911	1.53 ± 0.04	0.44
DTA-CBF	104.98	23.28	1.985	1.90 ± 0.03	0.30

4.8. Hyperparameter Sensitivity

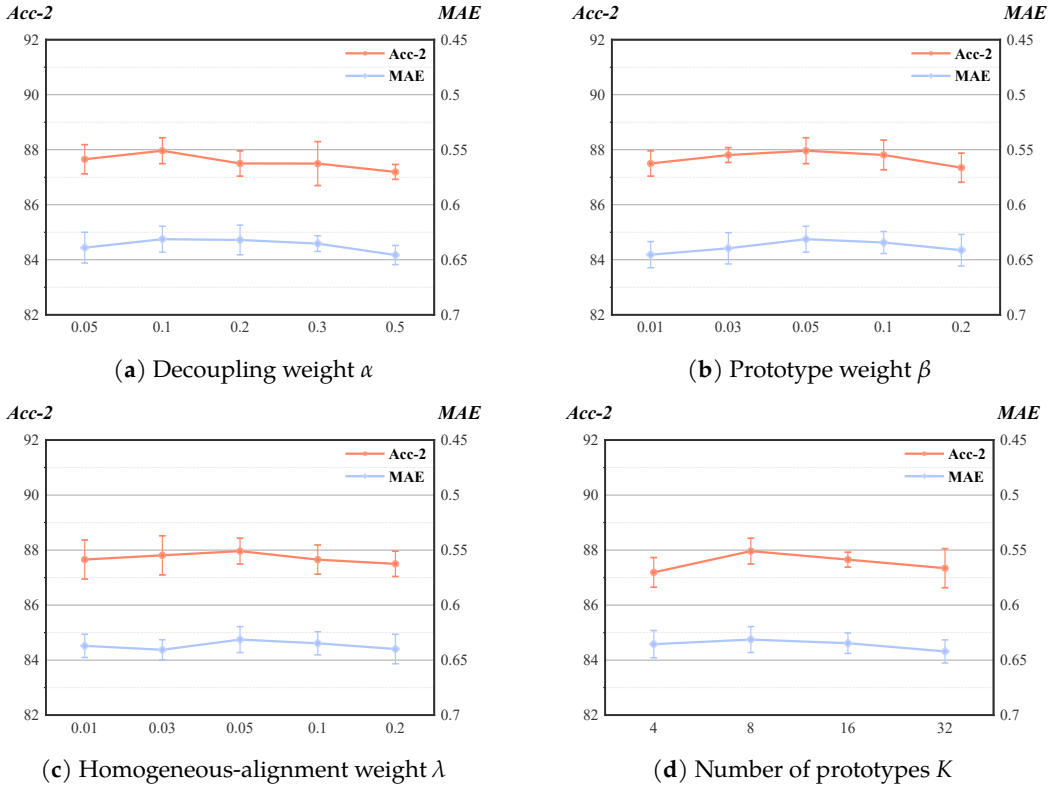
Figure 4 summarizes the validation-set sensitivity analysis. In each panel, one hyperparameter is varied while the remaining settings are fixed at the values used in the final configuration.

Moderate weights may balance the contributions of the decoupling, prototype, and homogeneous-alignment terms. Small weights may limit the influence of the corresponding constraints, whereas excessively large weights may cause regularization to compete with the supervised prediction objective. Performance also remains broadly stable across the tested prototype counts. A small prototype space may be too coarse to capture the diversity of multimodal sentiment patterns. Further increasing the prototype count does not yield consistent performance gains and may fragment the representations. Overall, the metrics vary only modestly across the tested ranges, indicating that the proposed method is relatively insensitive to the precise values of these hyperparameters.

4.9. Case Study and Visualization

To examine how the proposed modules affect sentiment prediction, we present qualitative cases and a TGCA attention visualization on CMU-MOSI.

Figure 5 presents three representative test cases, including two successful predictions and one polarity error. In Case 1, the utterance "It was really boring to me" expresses a clear negative sentiment, whereas the acoustic cue is calm and the visual cue contains a slight smile. Such weak or partially inconsistent nonverbal evidence may pull the prediction toward neutrality. The full DTA-CBF predicts −2.26, which is close to the ground-truth score of −2.40. In contrast, removing MDSR produces a much weaker negative prediction of −1.72, while removing TGCA gives −2.05. The larger prediction shift after removing MDSR suggests that the module helps handle conflicting modality cues in this example.



(**a**) Decoupling weight α (**b**) Prototype weight β

(**c**) Homogeneous-alignment weight λ (**d**) Number of prototypes K

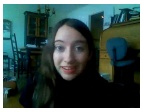
Figure 4. Hyperparameter sensitivity analysis of DTA-CBF on CMU-MOSI. Each curve reports the validation-set mean across three independent runs, and error bars denote the standard deviation.

Case 2 presents a different pattern. The text "I give Shrek Forever After directed by Mike Mitchell a grade of B minus" conveys a mild positive evaluation rather than a strong positive opinion. The speaker also uses an emphasized tone around the grade and shows a slight smile, providing weak but useful nonverbal support for the positive polarity. The full model predicts 1.16 for a ground-truth score of 1.00, whereas the w/o TGCA variant decreases to 0.52 and moves closer to neutrality. In this example, retaining TGCA brings the prediction closer to the ground-truth score, consistent with a benefit from text-guided cross-modal interaction.

Case 3 illustrates a failure when an utterance combines praise and criticism of different aspects of the same person. The speaker praises the actor's appearance but rejects a positive assessment of his acting ability with the final word "no". DTA-CBF predicts +1.81 for a ground-truth score of −1.00, reversing the sentiment polarity. The w/o MDSR and w/o TGCA variants also predict positive scores of +1.67 and +1.84, respectively. These predictions illustrate a shared difficulty in interpreting mixed evaluations and a concluding negation.

Figure 6 shows that TGCA assigns nonverbal evidence in a text-guided and selective manner rather than treating the entire utterance uniformly. In this utterance, the speaker uses a calm tone for "I'm sorry", stretches "and" and "I", and lowers the tone around "disappointed to say this" and "didn't like it". Attention is diffuse during the calm opening and more concentrated during the later negative judgment. These weights show which acoustic and visual segments the text queries attend to. On their own, they cannot quantify each modality's contribution to the final prediction or explain why a prediction fails.

The prediction shifts in Cases 1 and 2 are consistent with the distinct roles of MDSR and TGCA. MDSR retains modality-specific features while aligning the distributions of common features; TGCA then uses text queries to select acoustic and visual cues. This makes retrieval dependent on the text representation. However, the polarity error in Case 3 cannot be attributed to text-guided retrieval alone.

modality	content	Ground Truth	DTA-CBF	w/o MDSR	w/o TGCA
text	It was really boring to me	−2.40	−2.26	−1.72	−2.05
audio	Calm tone				
visual	 Slight smile				

modality	content	Ground Truth	DTA-CBF	w/o MDSR	w/o TGCA
text	I give Shrek Forever After directed by Mike Mitchell a grade of B minus	1.00	1.16	0.78	0.52
audio	Emphasized tone				
visual	 Slight smile				

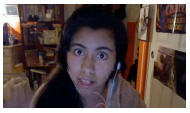
modality	content	Ground Truth	DTA-CBF	w/o MDSR	w/o TGCA
text	I love him as an actor based on his looks. As a actual talented actor, no.	−1.00	1.81	1.67	1.84
audio	Calm tone				
visual	 Wide eyes				

Figure 5. Qualitative case analysis on CMU-MOSI test samples.

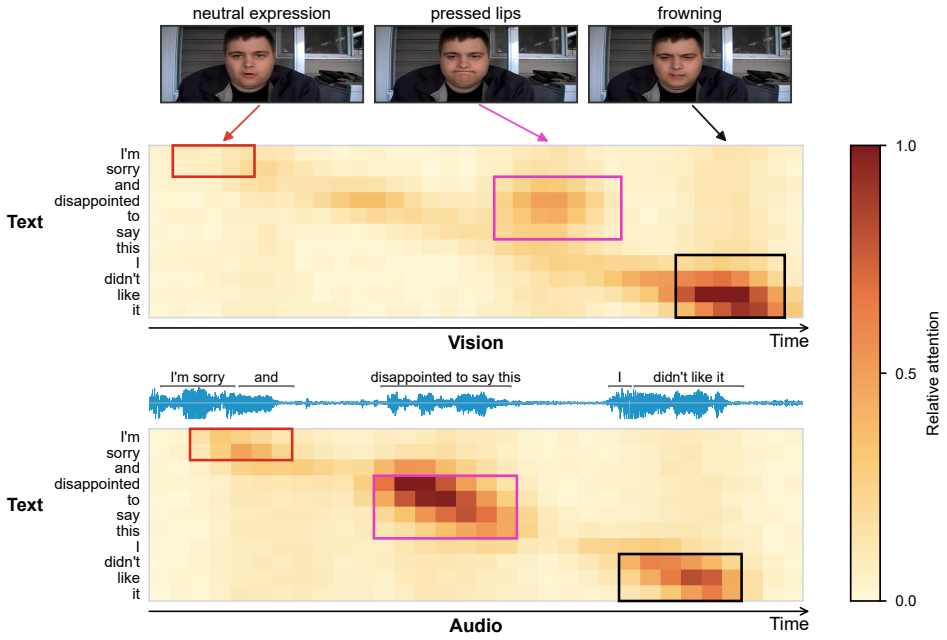


Figure 6. TGCA attention visualization on CMU-MOSI.

5. Conclusions

This paper presented DTA-CBF as an architecture for organizing multimodal information before sentiment prediction. MDSR separates modality-specific and modality-common factors and regularizes their representations. TGCA uses text to guide the transfer of acoustic and visual evidence. CBF then progressively compresses and integrates the aligned sequence and modality streams, with global context calibration.

Experiments on CMU-MOSI, CMU-MOSEI, and CH-SIMS show competitive performance, particularly on the two English benchmarks, while DecAlign obtains higher Corr and Acc-3 on CH-SIMS. Published results use different experimental protocols and are interpreted with those differences in mind. Ablations identify MDSR as the largest contributor, with TGCA and CBF providing additional gains. On CMU-MOSI, the fusion analysis reports 228.4 million MAdds for CBF, compared with 356.4 million MAdds for concatenation followed by self-attention, while CBF yields higher correlation and classification scores.

Several limitations remain. Text-guided retrieval may favor nonverbal cues consistent with the text representation, creating a risk of modality dominance. Ambiguous wording, sarcasm, or transcription errors may therefore limit the use of corrective acoustic and visual evidence. The gains over Text-only support the aggregate value of nonverbal information, but do not establish balanced contributions from the three modalities. Lower MMD^2 indicates closer distributions of modality-common representations under the chosen kernel, but does not by itself establish semantic disentanglement. With RoBERTa-large, DTA-CBF has a large parameter count, which may limit deployment on devices with limited memory. The present evaluation does not cover missing-modality or real-time settings. Future work will examine modality contributions through controlled perturbations and shared/private information through representation probes, alongside evaluation under missing modalities and real-time constraints.

Author Contributions: Conceptualization, Y.Z. and B.Z.; methodology, Y.Z.; software, Y.Z.; validation, Y.Z. and W.T.; formal analysis, Y.Z. and W.T.; investigation, Y.Z.; data curation, Y.Z.; visualization, Y.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, B.Z. and W.T.; supervision, B.Z.; project administration, B.Z.; funding acquisition, B.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (NSFC) for Key Program under Grant 62237001.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study (CMU-MOSI, CMU-MOSEI, and CH-SIMS) are publicly available through the MMSA framework at https://github.com/thuiar/MMSA (accessed on 25 November 2026). No new datasets were generated during this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Das, R.; Singh, T.D. Multimodal sentiment analysis: A survey of methods, trends, and challenges. *ACM Comput. Surv.* **2023**, *55*, 270. [CrossRef]
2. Liang, P.P.; Zadeh, A.; Morency, L.P. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Comput. Surv.* **2024**, *56*, 264. [CrossRef]
3. Baltrusaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [CrossRef] [PubMed]
4. Xu, P.; Zhu, X.; Clifton, D.A. Multimodal learning with transformers: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12113–12132. [CrossRef] [PubMed]

5. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv* **2016**, arXiv:1606.06259. [CrossRef]
6. Zadeh, A.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.P. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 2236–2246. [CrossRef]
7. Yu, W.; Xu, H.; Yuan, Z.; Wu, J. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 3718–3727. [CrossRef]
8. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 1103–1114. [CrossRef]
9. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.; Morency, L.-P. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 2247–2256. [CrossRef]
10. Tsai, Y.H.H.; Liang, P.P.; Zadeh, A.; Morency, L.P.; Salakhutdinov, R. Learning factorized multimodal representations. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
11. Hazarika, D.; Zimmermann, R.; Poria, S. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the ACM International Conference on Multimedia*; ACM: New York, NY, USA, 2020; pp. 1122–1131. [CrossRef]
12. Yang, D.; Huang, S.; Kuang, H.; Du, Y.; Zhang, L. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the ACM International Conference on Multimedia*; ACM: New York, NY, USA, 2022; pp. 1642–1651. [CrossRef]
13. Li, Y.; Wang, Y.; Cui, Z. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2023; pp. 6631–6640. [CrossRef]
14. Yang, D.; Li, M.; Xiao, D.; Liu, Y.; Yang, K.; Chen, Z.; Wang, Y.; Zhai, P.; Li, K.; Zhang, L. Towards multimodal sentiment analysis debiasing via bias purification. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2024; pp. 464–481. [CrossRef]
15. Zeng, Y.; Yan, W.; Mai, S.; Hu, H. Disentanglement translation network for multimodal sentiment analysis. *Inf. Fusion* **2024**, *102*, 102031. [CrossRef]
16. Zhou, Y.; Liang, X.; Chen, H.; Zhao, Y.; Chen, X.; Yu, L. Triple disentangled representation learning for multimodal affective analysis. *Inf. Fusion* **2025**, *114*, 102663. [CrossRef]
17. Gretton, A.; Borgwardt, K.M.; Rasch, M.J.; Scholkopf, B.; Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.* **2012**, *13*, 723–773.
18. Villani, C. *Optimal Transport: Old and New*; Springer: Berlin/Heidelberg, Germany, 2009. [CrossRef]
19. Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2013; pp. 2292–2300.
20. Peyre, G.; Cuturi, M. Computational optimal transport with applications to data sciences. *Found. Trends Mach. Learn.* **2019**, *11*, 355–607. [CrossRef]
21. Tsai, Y.-H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.-P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 6558–6569. [CrossRef] [PubMed]
22. Tang, Z.; Xiao, Q.; Zhou, X.; Li, Y.; Chen, C.; Li, K. Learning discriminative multi-relation representations for multimodal sentiment analysis. *Inf. Sci.* **2023**, *641*, 119125. [CrossRef]
23. Zhang, X.; Cui, W.; Hu, B.; Li, Y. A multi-level alignment and cross-modal unified semantic graph refinement network for conversational emotion recognition. *IEEE Trans. Affect. Comput.* **2024**, *15*, 1553–1566. [CrossRef]
24. Wang, L.; Peng, J.; Zheng, C.; Zhao, T.; Zhu, L. A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning. *Inf. Process. Manag.* **2024**, *61*, 103675. [CrossRef]
25. Rahman, W.; Hasan, M.K.; Lee, S.; Bagher Zadeh, A.; Mao, C.; Morency, L.-P.; Hoque, E. Integrating multimodal information in large pretrained transformers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 2359–2369. [CrossRef] [PubMed]
26. Yu, W.; Xu, H.; Yuan, Z.; Wu, J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2021; pp. 10790–10797. [CrossRef]
27. Han, W.; Chen, H.; Poria, S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 9180–9192. [CrossRef]
28. Wang, D.; Guo, X.; Tian, Y.; Liu, J.; He, L.; Luo, X. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognit.* **2023**, *136*, 109259. [CrossRef]

29. Wang, P.; Zhou, Q.; Wu, Y.; Chen, T.; Hu, J. DLF: Disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2025; Volume 39, pp. 21180–21188. [CrossRef]
30. Wang, D.; Liu, S.; Wang, Q.; Tian, Y.; He, L.; Gao, X. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Trans. Multimed.* **2022**, *25*, 4909–4921. [CrossRef]
31. Jaegle, A.; Borgeaud, S.; Alayrac, J.-B.; Doersch, C.; Ionescu, C.; Ding, D.; Koppula, S.; Zoran, D.; Brock, A.; Shelhamer, E.; et al. Perceiver IO: A general architecture for structured inputs and outputs. In Proceedings of the International Conference on Learning Representations, Virtual Event, 25–29 April 2022.
32. Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; Sun, C. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; pp. 14200–14213.
33. Qian, C.; Xing, S.; Li, S.; Zhao, Y.; Tu, Z. DecAlign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In Proceedings of the International Conference on Learning Representations, Rio de Janeiro, Brazil, 23–27 April 2026.
34. Liu, Z.; Zhou, B.; Chu, D.; Sun, Y.; Meng, L. Modality translation-based multimodal sentiment analysis under uncertain missing modalities. *Inf. Fusion* **2024**, *101*, 101973. [CrossRef]
35. Li, M.; Yang, D.; Lei, Y.; Wang, S.; Wang, S.; Su, L.; Yang, K.; Wang, Y.; Sun, M.; Zhang, L. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2024; Volume 38, pp. 10074–10082. [CrossRef]
36. Sun, H.; Wang, H.; Liu, J.; Chen, Y.W.; Lin, L. CubeMLP: An MLP-based model for multimodal sentiment analysis and depression estimation. In *Proceedings of the ACM International Conference on Multimedia*; ACM: New York, NY, USA, 2022; pp. 3722–3729. [CrossRef]
37. Xia, Y.; Huang, H.; Zhu, J.; Zhao, Z. Achieving cross modal generalization with multimodal unified representation. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36. [CrossRef]
38. Lin, J.; Wang, Y.; Xu, Y.; Liu, Q. Semi-IIN: Semi-supervised intra-inter modal interaction learning network for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2025; Volume 39, pp. 1411–1419. [CrossRef]
39. Wen, Y.; Li, Q.; Zhou, Y.; Gao, Y.; Wen, Z.; Tao, J.; Li, Y. DashFusion: Dual-stream alignment with hierarchical bottleneck fusion for multimodal sentiment analysis. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 17941–17952. [CrossRef] [PubMed]
40. Wei, Y.; Li, S.; Feng, R.; Hu, D. Diagnosing and re-learning for balanced multimodal learning. In *Computer Vision–ECCV 2024. Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2025; Volume 15122, pp. 71–86. [CrossRef]
41. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 5998–6008.
42. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv* **2019**, arXiv:1907.11692. [CrossRef]
43. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Wang, S.; Hu, G. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 657–668. [CrossRef]
44. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4077–4087.
45. Fan, Y.; Xu, W.; Wang, H.; Wang, J.; Guo, S. PMR: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2023; pp. 20029–20038. [CrossRef]
46. Hu, Z.; Zheng, W.; Zong, Y.; Wei, M.; Jiang, X.; Shi, M. A novel decoupled prototype completion network for incomplete multimodal emotion recognition. In *Proceedings of the IEEE International Conference on Multimedia and Expo*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6. [CrossRef]
47. Jin, C.; Luo, C.; Yan, M.; Zhao, G.; Zhang, G.; Zhang, S. Weakening the dominant role of text: CMOSI dataset and multimodal semantic enhancement network. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 222–236. [CrossRef] [PubMed]
48. Yang, J.; Yu, Y.; Niu, D.; Guo, W.; Xu, Y. ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 7617–7630. [CrossRef]
49. Sun, Z.; Sarma, P.K.; Sethares, W.; Liang, Y. Learning Relationships between Text, Audio, and Video via Deep Canonical Correlation for Multimodal Language Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2020; Volume 34, pp. 8992–8999. [CrossRef]
50. Chen, C.; Hong, H.; Guo, J.; Song, B. Inter-intra modal representation augmentation with trimodal collaborative disentanglement network for multimodal sentiment analysis. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 1476–1488. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.