Contextual Augmented Global Contrast for Multimodal Intent Recognition

Kaili Sun[1], Zhiwen Xie[2], Mang Ye[1]*, Huyin Zhang[1]*
[1]School of Computer Science, Wuhan University, Wuhan, China
[2]School of Computer Science, Central China Normal University, Wuhan, China
{kailisun, yemang, zhy2536}@whu.edu.cn, zwxie@ccnu.edu.cn

Abstract

Multimodal intent recognition (MIR) aims to perceive the human intent polarity via language, visual, and acoustic modalities. The inherent intent ambiguity makes it challenging to recognize in multimodal scenarios. Existing MIR methods tend to model the individual video independently, ignoring global contextual information across videos. This learning manner inevitably introduces perception biases, exacerbated by the inconsistencies of the multimodal representation, amplifying the intent uncertainty. This challenge motivates us to explore effective global context modeling. Thus, we propose a context-augmented global contrast (CAGC) method to capture rich global context features by mining both intra-and cross-video context interactions for MIR. Concretely, we design a context-augmented transformer module to extract global context dependencies across videos. To further alleviate error accumulation and interference, we develop a cross-video bank that retrieves effective video sources by considering both intentional tendency and video similarity. Furthermore, we introduce a global context-guided contrastive learning scheme, designed to mitigate inconsistencies arising from global context and individual modalities in different feature spaces. This scheme incorporates global cues as the supervision to capture robust the multimodal intent representation. Experiments demonstrate CAGC obtains superior performance than state-of-the-art MIR methods. We also generalize our approach to a closely related task, multimodal sentiment analysis, achieving the comparable performance.

1. Introduction

Intent recognition (IR) techniques [6, 38] have exhibited remarkable performance in text [6, 26, 63] and visual intent [1, 23, 42]. However, these techniques are primarily tailored for single-modal scenarios and not effectively tackle the challenges of real-world multimodal language.

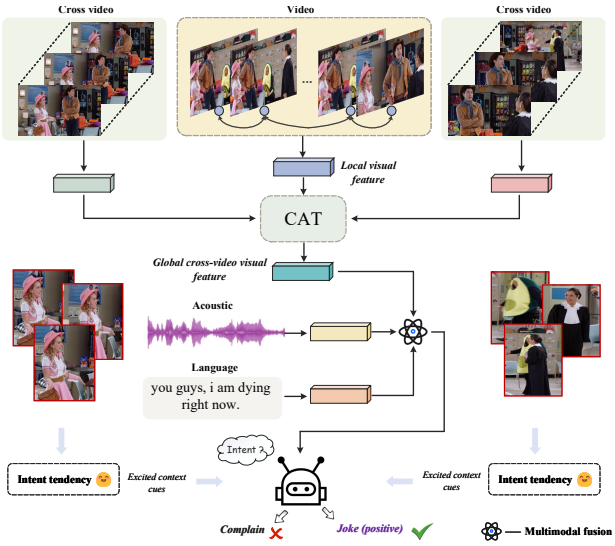


Figure 1. Illustration of multimodal intent recognition for the video sample. Absorbing global context information from videos that share highly similar scenes, as depicted cross-video in the figure, is essential to mitigate perception biases caused by the intentional ambiguity and improve the precise intent recognition.

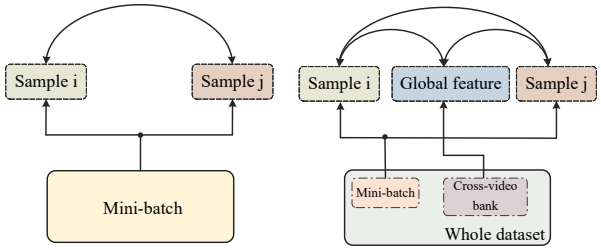
Multimodal intent recognition (MIR) [61], an emerging research area, aims to solve this challenge by aligning intent cues from both the natural language and the non-verbal data. Related cross-modality reasoning studies [15, 40, 43] have been explored to tackle the modality discrepancies in MIR. Nevertheless, these methods struggle to handle the distributional discrepancies arising from the heterogeneity of modality signals. As an alternative, Hazarika *et al*. [15] propose the projection of two complementary distinct subspaces to reduce the distributional discrepancies. However, these methods model the individual video independently, neglecting the contextual information across the videos. Models are inevitably prone to introducing the perception biases in the intent understanding, intensifying the uncertainty in the intent recognition.

Inspired by the global inter-video representation learning in object segmentation [31], the approaches [13, 27, 29,

*Corresponding Authors

30, 47, 49] based on inter-video modeling have achieved success on various visual tasks. Specifically, the intra- and inter-video feature associations are focused on visual correspondence learning in multiple studies [27, 47]. In a broad scenario, the cross-video relation has been extensively studied in visual representation learning [51], audio-visual video parsing [30], medical surgical scenes [22] and group video captioning [29]. Although these cross-video techniques have made significant progress, the extracted cross-video contextual feature still remains coarse due to the lack of meticulous filtering in the video sources. Consequently, the model becomes vulnerable to interference from irrelevant videos during the learning process. In MIR, the refining cross-video context dependency is essential to mitigate biases in intent understanding. Specifically, it not only provides effective and precise global context information but also reveals the robust consistent and discriminative feature cues in the dataset. For example, as illustrated in Fig. 1, we might infer that the speaker is complaining about something based on the text *"you guys, I am dying right now"* when acoustic and visual information fail to provide clear intent cues. If the model incorporates global contextual dependencies from cross-videos to comprehend the context of the current video scene, it is more likely to interpret the intent as "joke" rather than "complaint" in this example. This is because the speaker is situated in an excited rather than an angry context. By exploiting refining cross-video context prior knowledge, we can gain a deep context learning of the current video, thereby reducing biases in intent understanding and alleviating the intent ambiguity. Therefore, *1) how to capture refined cross-video context dependency is the first challenge for MIR.*

MIR also benefits from robust cross-modality alignment. Existing methods for cross-modality alignment primarily focusing on developing various attention-based algorithms [19, 50, 52]. To alleviate alignment bias arising from the simple concatenation of each modality, enhancements such as multimodal co-attention [52], a cross-attention multimodal encoder [19], and transformer-based multimodal token fusion [50] have been designed to boost multimodal alignment. However, these algorithms primarily emphasize fusion on target-relevant features, neglecting potentially valuable information from irrelevant ones. In response to this limitation, contrastive learning [4, 17, 48] is introduced to mine valuable discriminative features form irrelevant information for multimodality alignment [18, 34]. Although these methods have made progress, they ignore modality-specific information, resulting in limited discriminative ability. As an alternative, Yang *et al.* [53] decompose each modality into similar and dissimilarity features based on modality-specific information, leveraging contrastive learning between samples to enhance both consistency and inconsistency learning. The aforementioned



(a) Within Mini-batch Contrast (b) Our Designed Contrast

Figure 2. Different contrastive learning schemes. (a) Exploring the alignment between different modalities at the sample-level or within mini-batch (*e.g.*,[18, 53]). (b) Our designed GCCL algorithm aims to improve alignment across the entire dataset by introducing global features as supervisory signals.

methods are supervised by local modal alignment schemes, as illustrated in Fig. 2(a), which are insufficient for addressing modality discrepancies. Therefore, *2) how to incorporate with the intra-and cross- video contexts globally to reduce the modality discrepancies is the second challenge.*

To address the above two challenges, we propose a contextual augmented global contrast (CAGC) method. CAGC includes two main components: a *context-augmented transformer* (CAT) module and a *global context-guided contrastive learning* (GCCL) scheme, as depicted in Fig. 2(b). Our main idea is to explore rich and comprehensive contextual features to address the uncertainty in intent recognition. CAT aims to learn refined global context dependent features by simultaneously mining contextual relations from both intra-and cross-video to mitigate biases in intent understanding. To ensure effective cross-video sources, we further design a cross-video bank that considers both intentional tendency and similarity between videos. The bank can help the model avoid and mitigate the accumulation of errors from irrelevant videos, ensuring more precise cross-video contextual feature learning. GCCL aims to capture robust consistent and discriminative cross-modality feature and reduce the modality discrepancies. This scheme incorporates global context information as supervision to improve the cross-modality alignment. The contributions of this work are summarized as:

- We propose a context-augmented global contrast (CAGC) learning method to mine rich global contextual cues from both intra-and cross-video to enhance intent understanding for MIR.
- We design a context-augmented transformer module to learn cross-video context dependent features to reduce biases in intent understanding. Furthermore, we introduce a global context-guided contrastive learning scheme to capture the robust consistency and discriminative cross-modality feature and reduce the modality discrepancies.
- We conducted extensive experiments on public benchmark MIR dataset and obtain superior performance than state-of-the-arts. Furthermore, we also achieve a com-

parable performance on a widely used multimodal sentiment dataset. Validation results verify the effectiveness of CAGC.

2. Related work

Single-Modality Intent Recognition. Traditional intent recognition comprises both text [6, 26, 63] and visual intent recognition [1, 23, 42]. For text intent, an open-text recognition platform has been developed to discover the intent [59]. Zhou *et al*. [62] propose a k-nearest neighbors method for the out-of-domain intent classification, while Sadat *et al*. [36] study the intent behind questions and introduce a new dataset. Text intent analysis is also applied to the suicide risk assessment to delve into human psychological states [20]. Moreover, the text intent reasoning also plays a pivotal role in enhancing the text revision [6, 8, 12], the dialog systems [32, 60], and the text semantic matching [9, 46, 63]. As multimedia technology advances, the exploration of visual intent becomes an inevitable necessity. Joo *et al*. [23] emphasize the importance of recognizing intent in images, particularly for the political decision-making, and introduce the communicative visual intent. Simultaneously, collaboration between visual and text intent contributes to the advancement of intent recognition. The research [54] delves into hierarchical relations between visual content and text intent labels, aiming to enhance the global understanding of visual intent. Additionally, Wang *et al*. [45] introduce the prototype learning to address the semantic ambiguity in visual intent perception. Despite the impressive achievements in text and visual intent understanding mentioned above, they predominantly focus on single-modality studies and cannot effectively address challenges in complicated multimodal scenarios.

Multimodal Intent Recognition. Multimodal intent recognition aims to mine intent in scenarios involving multiple modalities. The understanding of intent through the integration of text and acoustic modalities has been explored in speech interaction systems [37]. Exploring intent from both visual and text modalities can enrich the varied expressions in the image caption task [1]. Moreover, intents derived from both visual and text modalities contribute to advancements in video summaries and recommendations [33]. Despite the exploration of intent in different modalities, addressing the challenge of multimodal scenes involving visual, textual, and acoustic modalities remains an existing research gap. To bridge this gap, Zhang *et al*. [61] propose the novel MIR task, aiming to recognize the intent by integrating visual, textual, and acoustic modalities. Several cross-modal reasoning methods have been utilized to tackle intent understanding in MIR. Early work involves the design of a multimodal transformer that facilitates cross-modal interaction through cross-modal attention [43]. Later works delve into specific challenges such as modality mismatch and distributional modality discrepancies. Rahman *et al*. [40] propose a multimodal adaptation gate framework, capable of incorporating nonverbal data during fine-tuning. Hazarika *et al*. [15] project each modality into two distinct modality-invariant and modality-specific subspaces to reduce distributional modality discrepancies.

Our work is fundamentally different from these available works. We not only focus on feature modeling within the individual videos but also emphasize the importance of learning the global contextual information across the videos. The global cross-video contextual information implies robust cross-modality consistent and discriminative semantic feature cues in the dataset. This is suitable for MIR due to the inconsistency in multimodal intent expressions. CAGC is effective in capturing the rich global contextual features through the intra-and cross-video learning manner, thereby enhancing the intent understanding.

Contrastive Representation Learning. Contrastive learning (CL) is a widely used learning method that involves contrasting positive pairs against negative ones [3, 5, 21]. CL is initially proposed to address the absence of supervised signals in self-supervised learning [4, 17, 25]. The further study proposes the supervised CL [24], which combines with class information to learn class distribution of samples. Recent studies have extended CL to multimodal tasks for improved the modality alignment. Hu *et al*. [18] propose a inter-modal CL to minimize intra-class variance and maximize inter-class variance. CL is also introduced to align the global acoustic information and multimodal fusion features to enhance the global sentiment understanding [28]. Further study, Yang *et al*. [53] design a unified contrastive learning scheme at both intra-sample and inter-sample levels to enhance the multimodal representation. Although the aforementioned CL methods improve the performance of multimodal tasks from various perspectives, they are limited by the sample-level contrast within a mini-batch and ignore the global contrast of the entire training dataset. This leads to fewer robust discriminative features, making the difficult in the multimodal alignment. In contrast, we design a GCCL scheme that utilizes global contextual cues as the contrasting guidance to reduce the modality discrepancies and enhance the modality alignment.

3. Proposed Method

3.1. Overall Architecture

The framework of our CAGC is illustrated in Fig. 3. It mainly consists of four parts: the multimodality encoder, the cross-video context learning, the multimodal intent decoder, and the global context-guided contrastive learning. Considering each modality originally resides in different feature spaces, we create the unified modality representations through a multimodality encoder that is introduced

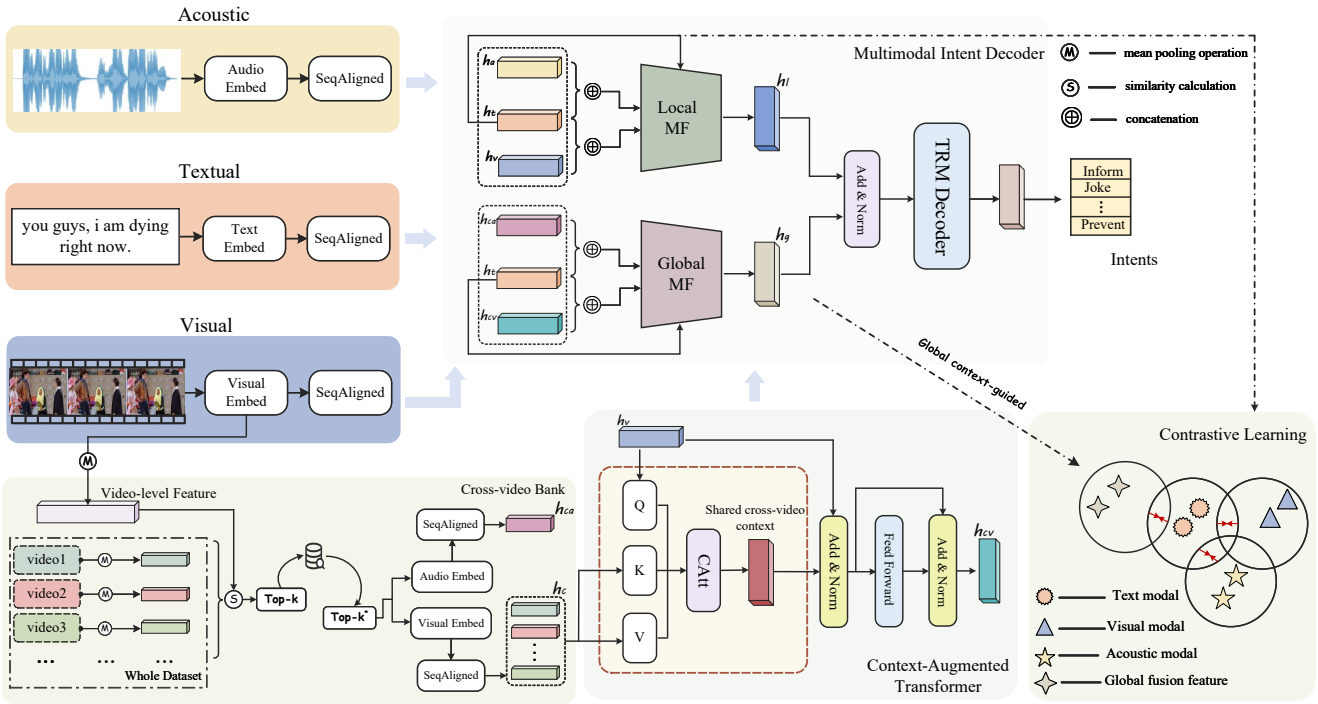


Figure 3. The architecture of CAGC. CAGC mainly consists of a Context-Augmented Transformer module, a cross-video bank, a multimodal intent decoder and a global context-guided contrastive learning scheme. The CAT module integrates scene information from cross-videos with the current video to obtain refined global context features from effective cross-videos based on the cross-video bank. Following this, the multimodal intent decoder employs both local and global multimodal fusion for intent decoding. Simultaneously, the global fused context feature is introduced as supervisory signal in GCCL to enhance the modality alignment.

in Sec. 3.2. To obtain the refined long-range context dependent features, we design a cross-video context learning, which includes a cross-video bank and a context-augmented transformer module in Sec. 3.3. Then, we fuse features from both local and global perspectives to capture the comprehensive multimodal representation and conduct the decoding of intents in Sec. 3.4. Furthermore, to reduce the modality discrepancies and enhance the cross-modality alignment, we propose a global context-guided contrastive learning scheme in Sec. 3.5. Finally, the training objective for the robust MIR is presented in Sec. 3.6. Below, we present the detail of the four components of CAGC.

3.2. Multimodality Encoder

Given a source input m=$\{L, V, A\}$, where L, V, and A denote different modalities: language, visual, and acoustic. We encode the input m into respective shallow modality feature representations. Following prior work [61], we utilize pre-trained BERT [7] to initialize an input text and extract the token embedding from the output of the final layer. For acoustic and visual inputs, we employ pre-trained wav2vec [2] and ResNet [16] models to obtain signal-level acoustic and frame-level visual feature representations. Specifically, formulas are as follows:

$$m_t = TextEmbed(L), \tag{1}$$
$$m_a = AudioEmbed(A), \tag{2}$$
$$m_v = VideoEmbed(V), \tag{3}$$

where $m_t \in \mathbb{R}^{l_t \times d_t}$, $m_a \in \mathbb{R}^{l_a \times d_a}$, and $m_v \in \mathbb{R}^{l_v \times d_v}$ are the extracted feature representations of each modality. $l_{t,a,v}$ and $d_{t,a,v}$ are respective sequence lengths and feature dimensions. To form a unified multimodality feature representation, the Connectionist Temporal Classification (CTC) [10] module is utilized to align on the word-level sequence. The formula is as follows:

$$h_t, h_a, h_v = SeqAligned(m_t, m_a, m_v), \tag{4}$$

where $h_t \in \mathbb{R}^{l \times d}$, $h_a \in \mathbb{R}^{l \times d}$, and $h_v \in \mathbb{R}^{l \times d}$ are unified modality feature representations for language, acoustic, and visual modalities. l and d respectively indicates unit-length and unit-dimension. $SeqAligned(\cdot)$ indicates the CTC module.

3.3. Cross-video Context Learning

Cross-video context learning includes the cross-video bank and the context-augmented transformer module. The

cross-video bank aims to retrieve effective video sources and alleviate the interference of irrelevant videos. The context-augmented transformer module aims to capture refine global context dependencies based on the bank.

Cross-video Bank Construction. The construction of the cross-video bank consists of two stages: the initial establishment process in stage1 and the denoising process in stage2. The bank is denoted as $\mathcal{C}$ in stage1. It stores each video along with videos that share highly similar scene information, which are screened across the entire training dataset. The scene similarity is defined based on the video-level feature, which are regarded as latent scene information. Specifically, given a video v_i that consists of a series of frames v_i=$\{v_i^{fram(j)}\}_{j=1}^{F}$. We define:

$$\overline{v}_i = \frac{1}{F}\sum_{j=1}^{F}(v_i^{fram(j)}), \tag{5}$$

where $\overline{v}_i$ indicates the video-level feature. F denotes the number of frames.

Then cross-video set $\Omega(\overline{v}_i)$ is obtained according to the scene similar score, which is defined as $\mathcal{L}_2$ distance. The formula is as follows:

$$v_c^i \in \Omega(v_i) := \left[\sum_{t=1}^{T}(\overline{v}_c^t - \overline{v}_i^t)^2\right]_k, \tag{6}$$

where $[\cdot]_k$ denotes top-k videos with lowest scores. T indicates the dimension of $\overline{v}_i$. $\Omega(v_i)$=$\{v_c^{i(1)}, v_c^{i(2)}, ..., v_c^{i(k)}\}$ represents the cross-video set corresponding to v_i.

The denoising process for cross-video bank $\mathcal{C}$ is conducted during stage 2. It aims to alleviate error accumulation and interference. Specifically, we obtain videos with consistent intentional tendencies via adopting a voting principle to select videos from set $\Omega(v_i)$ that have same intents. The final cross-video set is denoted as $\Omega^*(v_i)$=$\{v_c^{i(1)}, v_c^{i(2)}, ..., v_c^{i(k^*)}\}(1 \le k^* \le k)$, which may share similar intent tendencies with video v_i. Then the $\Omega^*(v_i)$ is sequentially stored in bank $\mathcal{C}$.

Context-augmented Transformer. The context-augmented transformer is designed to capture global long-range context dependent features to alleviate the ambiguity in the intent understanding. Compared to the typical transformer [44], the context-augmented attention mechanism is designed based on the cross-video bank $\mathcal{C}$, as described in Algorithm 1. It utilizes an enhanced learning manner to progressively strengthen long-range context learning. The formula is as follows:

$$h_m = CAtt(h_v, h_c, h_c), \tag{7}$$

where $CAtt(\cdot)$ indicates the context-augmented attention mechanism. h_c=$\{h_c^i\}_{i=1}^{N}$, $h_c^i \in \{h_c^{i(1)}, h_c^{i(2)}, ..., h_c^{i(k^*)}\}$ denotes the visual feature representation of cross-video set $\Omega^*(v_i)(1 \le i \le N)$. N indicates the batch size.

Algorithm 1 CAtt

Input: Matrix Q: visual feature h_v;
Matrices K, V: cross-video visual feature h_c;
Output: Long-range context dependent feature h_m;
Set v_c=$\{v_c^i\}_{i=1}^{N}$ is the cross-video set for the video within a mini-batch.
Set h_c=$\{h_c^i\}_{i=1}^{N}$ is the visual feature set for the cross-video $\Omega^*(v_i)$.
Initialize the Q=h_v, K, V=h_c;
Initialize the long-range dependent context feature $\mathcal{M}$;
for $j \leftarrow 1, ..., k^*$ **do**
  Update K_c=$h_c^{(1:j+1)} \in \{h_c^{(1)}, h_c^{(2)}, ..., h_c^{(j)}\}$;
  Update K, V=K_c;
  Obtain the context-augmented feature $\mathcal{M}_j$=$\sigma(\frac{QK^T}{\sqrt{k}})V$;
  Apply $\mathcal{M}_j$=$mean(\mathcal{M}_j)$;
  Store the context feature $\mathcal{M}_j$ to $\mathcal{M}$;
end for
return Update $h_m \leftarrow \mathcal{M}$;

$$h_{mv} = f_n(h_m + h_v), \tag{8}$$

$$h_{cv} = f_n(f(h_{mv}) + h_{mv}), \tag{9}$$

where h_{cv} is the captured long-range context dependent feature. $f_n(\cdot)$ indicates the normalization layer and $f(\cdot)$ represents the linear fusion layer.

3.4. Multimodal Intent Decoder

We perform multimodality fusion from local and global perspectives to capture rich and diverse multimodal feature representations. In the local multimodality fusion, we integrate the text modality feature with the visual and acoustic modality feature. The formulas are as follows:

$$h_{lf} = (f_w([h_t, h_v]) \circ f(h_v) + f_w([h_t, h_a]) \circ f(h_a)), \tag{10}$$

$$h_l = f'(h_t, h_{lf}) \circ h_{lf}, \tag{11}$$

where h_l indicates the local multimodality fusion feature. $f_w(\cdot)$ represents a gate function. $f'(\cdot)$ denotes a non-linear function. $\circ$ represents element-wise multiplication.

In the global multimodality fusion, the text modality feature is integrated with the cross-video visual and acoustic modality features. The formulas are as follows:

$$h_{gf} = (f_w([h_t, h_{cv}]) \circ f(h_{cv}) + f_w([h_t, h_{ca}]) \circ f(h_{ca})), \tag{12}$$

$$h_g = f'(h_t, h_{gf}) \circ h_{gf}, \tag{13}$$

where h_g is the global multimodality fusion feature. h_{ca} denotes the acoustic feature that is calculated as h_{cv} in Eq.9.

Finally, an intent decoder is performed on both local and global multimodality fusion features. The formula is as follows:

$$h_{out} = TRM(f_n(h_l + h_g)), \tag{14}$$

where $TRM(\cdot)$ indicates an intent decoder.

3.5. Global Context-guided Contrastive Learning

Contrastive learning has made remarkable advancements in representation learning by considering samples from various perspectives [11, 35, 39]. The idea of contrastive learning is to pull an anchor and its positive instances closer while pushing the anchor and negative instances apart in the feature space. Existing methods mostly focus on inter-class separation, by pulling fusion features of a specific class to be close and pushing fusion features away from the different ones. They either operate between individual samples or within a designed mini-batche, which belong to the local contrast learning strategy and fail to capture the global contrast across the entire training dataset. In our work, we not only perform the local inter-modality but also the global context-guided contrasts to learn robust discriminative features for strong inter-class separation and intra-class compactness.

The text modality has been proven greater significance than visual and acoustic modalities [14]. Therefore, we take text modality as an anchor while the other two modalities are treated as its augmented versions. For the local contrast, we align the text modality feature with other two modality features. For the global contrast, we align the text modality feature with the global multimodality fusion feature.

For the local contrast, positive and negative sample pairs are constructed from videos within mini-batch. In detail, given a text modality feature h_t as an anchor, its positive sample set $\mathcal{P}_l$ includes: 1) acoustic modality features from other videos with the same intent label; 2) visual modality features from other videos with the same intent label; Its negative sample $\mathcal{N}_l$ consists of: 1) acoustic modality features from other videos with different intent label; 2) visual modality features from other videos with different intent label. We can formulate the local contrast that employs the InfoNCE [17] loss as:

$$\mathcal{L}_{ta} = -\log \frac{\sum_{h_a^+ \in \mathcal{P}_l} \exp(h_t \cdot h_a^+ / \tau)}{\sum_{h_a^\pm \in \mathcal{P}_l \cup \mathcal{N}_l} \exp(h_t \cdot h_a^\pm / \tau)}, \tag{15}$$

$$\mathcal{L}_{tv} = -\log \frac{\sum_{h_v^+ \in \mathcal{P}_l} \exp(h_t \cdot h_v^+ / \tau)}{\sum_{h_v^\pm \in \mathcal{P}_l \cup \mathcal{N}_l} \exp(h_t \cdot h_v^\pm / \tau)}. \tag{16}$$

For the global contrast, positive and negative sample pairs are constructed from cross-videos across the entire dataset. In detail, given a text modality feature h_t as an anchor, its positive sample set $\mathcal{P}_g$ includes global multimodal fusion features from cross-videos with the same intent label. Its negative sample $\mathcal{N}_g$ contains global multimadal features from cross-videos with the different intent label. Similarly, our global contrast can be formulated as follows:

$$\mathcal{L}_{g} = -\log \frac{\sum_{h_g^+ \in \mathcal{P}_g} \exp(h_t \cdot h_g^+ / \tau)}{\sum_{h_g^\pm \in \mathcal{P}_g \cup \mathcal{N}_g} \exp(h_t \cdot h_g^\pm / \tau)}. \tag{17}$$

3.6. Training Objective

The overall training objective of our modal are:

$$q = f_c(h_{out}), \tag{18}$$

$$\mathcal{L}_{task} = -\sum_{i=1}^{N} \left(p(i) \circ \log(q(i))\right), \tag{19}$$

where $p(i)$ is the distribution of ground-truth intents. $\mathcal{L}_{task}$ indicates a loss for the intent prediction. $f_c(\cdot)$ represents a linear function for classification.

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \alpha(\mathcal{L}_{ta} + \mathcal{L}_{tv}) + \beta\mathcal{L}_{g}, \tag{20}$$

where $\mathcal{L}_{total}$ indicates the overall training loss. α and β are weight parameters corresponding to a local contrast loss $\mathcal{L}_l$ and a global contrast loss $\mathcal{L}_g$.

4. Experiment

Datasets. We evaluate CAGC on two multimodal benchmarks: MIntRec [61] for the intent recognition and CMU-MOSI [56] for the sentiment analysis. **MIntRec** comprises 2,224 high-quality samples, with 1,334, 445, and 445 samples allocated for training, validation, and testing sets, respectively. The dataset is structured for two intent categories: coarse-grained with binary intent labels and fine-grained with twenty intent labels. Coarse-grained intents encompass expressing emotions or attitudes and achieving goals. Finer-grained intents further elaborate on the coarse-grained categories, consisting of 11 express emotions or attitudes and 9 achieve goals intentions, respectively. **CMU-MOSI** comprises 2,199 short monologue video samples, distributed with 1,284, 229, and 686 samples assigned to the training, validation, and testing sets, respectively. In CMU-MOSI, each sample is annotated with a sentiment score ranging from -3 to 3, encompassing highly negative, negative, weakly negative, neutral, weakly positive, positive, and highly positive sentiments.

Evaluation Metric. For MIntRec, we follow the standard protocol from the previous work [61] to report the results. We evaluate the results via the following metrics: Accuracy (ACC), F1-score (F1), Precision (P), and Recall (R) for the intent recognition. For CMU-MOSI, we follow the previous works [15, 43] to evaluate CAGC by using the metrics: Mean-absolute Error (MAE), Pearson Correlation (Corr), Binary Accuracy (ACC-2), 7-class Accuracy (ACC-7), and F1-score (F1).

Implementation Details. On the MIntRec dataset, we represent the text feature as 768 dimensions, visual feature as 256 dimensions, and acoustic feature as 768 dimensions. On the CMU-MOSI dataset, we represent the text feature as 768 dimensions, visual feature as 47 dimensions, and acoustic feature as 74 dimensions. Specifically, we utilized the Adam optimizer with a learning rate of 2e-5, a dropout rate

Table 1. Comparative experimental results of the intent analysis models on the public MIntRec test-std set. Underlined results indicate the best performance among previous methods. The top 1 results are highlighted by **bold**. ↑ indicates higher is better.

Method	Twenty-class				Binary			
	ACC ↑	F1 ↑	P ↑	R ↑	ACC ↑	F1 ↑	P ↑	R ↑
MAG-BERT [40]	72.65	68.64	69.08	69.28	89.24	89.10	89.10	89.13
MulT [43]	72.52	69.25	70.25	69.24	89.19	89.07	89.02	89.18
MISA [15]	72.29	69.32	70.85	69.24	89.21	89.06	89.12	89.06
CAGC (Ours)	**73.39**	**70.09**	**71.21**	**70.39**	**90.11**	**90.01**	**89.92**	**90.14**

Table 2. Comparative experimental results of the sentiment analysis models on the public CMU-MOSI test-std set.

Method	CMU-MOSI				
	MAE ↓	Corr ↑	ACC-2 ↑	F1-Score ↑	ACC-7 ↑
TFN [57]	0.970	0.633	73.9	73.4	32.1
MFN [58]	0.965	0.632	77.4	77.3	34.1
ICCN [41]	0.862	0.714	83.0	83.0	39.0
MulT [43]	0.871	0.698	83.0	82.8	40.0
MISA [15]	0.817	0.748	82.10	82.0	41.4
HyCon [34]	0.713	0.790	85.20	85.1	46.60
Self-MM [55]	0.708	0.796	85.46	85.43	46.67
ConFEDE [53]	0.742	0.784	85.52	85.52	42.27
CAGC (Ours)	0.775	0.774	**85.70**	**85.60**	44.80

of 0.5, and α, β to be $\{0.02, 0.02\}$ for the training loss. The temperature value is fixed at 0.7 and the number of cross-videos for top-k^* is set to 7. Furthermore, the mini-batch size is set to 16, the maximum length of sentences in the text modality is set to 30, and the number of frames in the video is set to 230. The maximum length of audio is set to 180, while the hidden dimension and attention heads are set to 768 and 8, respectively.

4.1. Comparison to State-of-the-Art Models

Results on the MIntRec Dataset. We compare CAGC with the state-of-the-art MIR methods on the MIntRec dataset, including MAG-BERT [40], MulT [43], and MISA [15]. Tab. 1 illustrates the experimental results. Comparing with the existing MIR methods, our model achieves the superior performance in the metrics. It is worth noting that CAGC demonstrates the significant improvement in ACC, with the performance improvements of 1.1% over MISA [15] (73.39% vs. 72.29%) and 0.92% over MulT [43] (90.11% vs. 89.19%) in twenty-class and binary-class respectively. Likewise, other metrics also exhibit different degrees of improvement, such as the F1 score increasing by 0.77% from 70.09% to 69.32%. Compare with these state-of-the-art methods that focus only on learning frame-to-frame context interactions within a video, the proposed CAGC efficiently extracts the refined cross-video context feature and integrates it with each modality, capturing the global context fusion feature. Further, a global contrastive learning scheme was designed, which leverages global context features as supervisory signals to learn more robust multimodal features with enhanced consistency and discriminability from a global perspective. Due to these advantages, our method outperforms MAG-BERT [40] and achieves superior performance.

Table 3. Abalation study of the proposed model on the public MIntRec test-std set.

Method	Twenty-class				Binary			
	ACC ↑	F1 ↑	P ↑	R ↑	ACC ↑	F1 ↑	P ↑	R ↑
CAGC (Ours)	**73.39**	**70.09**	**71.21**	**70.39**	**90.11**	**90.01**	**89.92**	**90.14**
w/o CAT	71.31	67.08	68.28	67.78	87.84	87.71	87.69	87.83
w/o $\mathcal{L}_g$	72.22	68.07	68.30	69.09	88.25	88.09	88.13	88.11
w/o $\mathcal{L}_l$	72.31	68.33	68.82	69.28	89.12	88.97	89.04	88.98
w/o $\mathcal{L}_g$&$\mathcal{L}_l$	71.66	67.12	67.64	67.79	87.82	87.65	87.71	87.66

Results on the CMU-MOSI dataset. To further validate the performance of CAGC, we evaluate the effectiveness on the CMU-MOSI dataset and report the results in Tab. 2. Compare with the state-of-the-art method ConFEDE [53], the CAGC achieves the top performance in ACC-2 and F1 score. Specifically, the ACC-2 improves from 85.52% to 85.70% and F1 improves from 85.43% to 85.60%. Compared to methods [34, 53] that emphasize contrastive learning within the mini-batch, our designed global context-guided contrastive learning scheme indeed improves the performance of the model. The experimental results demonstrate the remarkable effectiveness of our proposed method compared to existing methods. This can be attributed to the previous methods being confused by confounding factors in multimodal representations, resulting in biased modal alignment and limited learning capability.

4.2. Ablation Studies

We evaluate the effects of CAGC's key components, including CAT, local contrast loss $\mathcal{L}_l$, and global contrast loss $\mathcal{L}_g$ in GCCL. The results are illustrated in Tab. 3. The incorporation of the cross-video context feature extractor CAT into CAGC results in a significant improvement, with the ACC increasing to 73.39%. This phenomenon indicates that refined cross-video visual interaction enhances contextual connections in the relevant visual content, assisting the model in capturing the global context for precise intent prediction.

Furthermore, we observe that eliminating global contrast $\mathcal{L}_g$ from the GCCL resulted in a decrease of 1.17% in ACC for the twenty-class setting. Likewise, the removal of local

Figure 4. Two cases E1 and E2 from the test set. E1 is a case of an explicit intent. E2 is a case with ambiguous intent.

contrast $\mathcal{L}_l$ while maintaining global contrast $\mathcal{L}_g$ in GCCL resulted in a decrease of 1.08% in ACC. This observation indicates the crucial role of the global contrast in enhancing the robust consistency and discriminative features by introducing global contextual features as supervision. Simultaneously, we consider that global contrast $\mathcal{L}_g$ and local contrast $\mathcal{L}_l$ in GCCL complement each other. Utilizing solely the global context CAT module, CAGC achieves an ACC of 71.66%. This is due to the absence of supervision from both global and local contrast, leading to the extraction of less discriminative multimodal features for intent recognition. However, utilizing both the CAT module and the GCCL scheme enables CAGC to achieve optimal performance.

4.3. Case Study

In Fig. 4, we present some examples where CAGC adjusted the understanding of the current scene properly by taking into account cross-video information. These examples demonstrate that CAGC can successfully incorporate cross-video information to complement the current scene. Specifically, two examples are given in Fig. 4, where blue words with underlines potentially express intent polarity, and red boxes are marked as the visual feature of the speaker. From the Fig. 4, we can find that in E1, the textual, visual and audio modalities all provide strong guidance for intentional polarity. Specifically, words "best" and "love" are appeared in the textual modality. The visual modality tends to be positive because of the laughing and open lips on facial features. And aucoustic signals are loud and exciting. It is sufficient to determine the positive intent polarity for the sample as "Praise".

In E2, the textual modality explicitly expresses a strong positive intent polarity such as words "good" and "luck", while the facial visual feature and the acoustic signal are gentle and do not provide a strong intent guidance. In this situation, the model is influenced by the textual modality strongly and tends to infer this sample as positive intent with labels such as "Agree", "Praise" and "Comfort" *etc*. However, in fact, the true label for this sample is "Taunt", which belongs to the negative intent category. Therefore, in such samples with ambiguous intent, it is really hard to determine the true intent polarity. It's evidently insufficient for the model to rely merely on its own video features without considering the context where the speaker is situated. To explore the contextual environment for the speaker, our proposed method first establishes connections between the current video and cross-videos that share strong similar scenes. Subsequently, it captures global cross-video context interactions to to aid the model in accurately inferring intentions. Specifically, we can find that in E2, cross-video and the current video share highly similar scenes, and their intent is "Oppose", which belongs to the negative intent tendency. This information provides a clue to the model that indicates the speaker is probably in a negative contextual situation. It makes the model prone to infer this sample as a negative intent, such as "Taunt."

5. Conclusion

We propose a context-augmented global contrast (CAGC) network for MIR. CAGC is innovative in two aspects: its CAT component utilizes global contextual dependencies across videos to enhance comprehensive context understanding, thereby alleviating intent ambiguity. To ensure the precision of global context, we further develop a cross-video bank that simultaneously considers intent tendencies and inter-video similarity to ensure effective cross-video source, mitigating error accumulation and interference. GCCL component relieves modality discrepancies and enhances cross-modality alignment by introducing global contextual cues as the supervision. Our proposed method is mainly evaluated on the MIntRec. The results demonstrate that the proposed method significantly outperforms state-of-the-art MIR methods. We also conduct an extensive experiments on CMU-MOSI to evaluate the effectiveness of CAGC, and our approach achieves comparable performances to state-of-the-art methods.

Acknowledgement. This work is supported by National Natural Science Foundation of China under Grant (62176188, 62361166629, 62272354).

References

[1] Jyoti Aneja, Harsh Agrawal, Dhruv Batra, and Alexander G. Schwing. Sequential latent spaces for modeling the intention during diverse image captioning. In *ICCV*, pages 4260–4269, 2019. 1, 3

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, 2020. 4

[3] Hritam Basak and Zhaozheng Yin. Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In *CVPR*, pages 19786–19797, 2023. 3

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 2, 3

[5] Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *CVPR*, pages 19016–19026, 2023. 3

[6] Ruining Chong, Cunliang Kong, Liu Wu, Zhenghao Liu, Ziye Jin, Liner Yang, Yange Fan, Hanghang Fan, and Erhong Yang. Leveraging prefix transfer for multi-intent text revision. In *ACL*, pages 1219–1228, 2023. 1, 3

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019. 4

[8] Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. Understanding iterative revision from human-written text. In *ACL*, pages 3573–3590, 2022. 3

[9] Zheren Fu, Zhendong Mao, Yan Song, and Yongdong Zhang. Learning semantic relationship among instances for image-text matching. In *CVPR*, pages 15159–15168, 2023. 3

[10] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, page 369–376, 2006. 4

[11] Marah Halawa, Olaf Hellwich, and Pia Bideau. Action-based contrastive learning for trajectory prediction. In *ECCV*, pages 143–159, 2022. 6

[12] Skyler Hallinan, Alisa Liu, Yejin Choi, and Maarten Sap. Detoxifying text with marco: Controllable revision with experts and anti-experts. In *ACL*, pages 228–242, 2023. 3

[13] Mingfei Han, Yali Wang, Xiaojun Chang, and Yu Qiao. Mining inter-video proposal relations for video object detection. In *ECCV*, pages 431–446, 2020. 1

[14] Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *EMNLP*, pages 9180–9192, 2021. 6

[15] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In *ACMMM*, pages 1122–1131, 2020. 1, 3, 6, 7

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735, 2020. 2, 3, 6

[18] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. In *EMNLP*, pages 7837–7851, 2022. 2, 3

[19] Zhizhang Hu, Xinliang Zhu, Son Tran, René Vidal, and Arnab Dhua. Provla: Compositional image search with progressive vision-language alignment and multimodal fusion. In *ICCV*, pages 2772–2777, 2023. 2

[20] Shaoxiong Ji. Towards intention understanding in suicidal risk assessment with natural language processing. In *EMNLP*, pages 4028–4038, 2022. 3

[21] Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David A. Clifton, and Jie Chen. Expectation-maximization contrastive learning for compact video-and-language representations. In *NeurIPS*, 2022. 3

[22] Yueming Jin, Yang Yu, Cheng Chen, Zixu Zhao, Pheng-Ann Heng, and Danail Stoyanov. Exploring intra- and inter-video relation for surgical semantic scene segmentation. *IEEE Transactions on Medical Imaging*, 41(11):2991–3002, 2022. 2

[23] Jungseock Joo, Weixin Li, Francis F. Steen, and Song-Chun Zhu. Visual persuasion: Inferring communicative intents of images. In *CVPR*, pages 216–223, 2014. 1, 3

[24] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020. 3

[25] Klemen Kotar, Gabriel Ilharco, Ludwig Schmidt, Kiana Ehsani, and Roozbeh Mottaghi. Contrasting contrastive self-supervised representation learning pipelines. In *ICCV*, pages 9929–9939, 2021. 3

[26] Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. Integrating text and image: Determining multimodal document intent in instagram posts. In *EMNLP-IJCNLP*, pages 4621–4631, 2019. 1, 3

[27] Liulei Li, Tianfei Zhou, Wenguan Wang, Lu Yang, Jianwu Li, and Yi Yang. Locality-aware inter-and intra-video reconstruction for self-supervised correspondence learning. In *CVPR*, pages 8709–8720, 2022. 1, 2

[28] Ziming Li, Yan Zhou, Weibo Zhang, Yaxin Liu, Chuanpeng Yang, Zheng Lian, and Songlin Hu. AMOA: global acoustic feature enhanced modal-order-aware network for multimodal sentiment analysis. In *COLING*, pages 7136–7146, 2022. 3

[29] Wang Lin, Tao Jin, Ye Wang, Wenwen Pan, Linjun Li, Xize Cheng, and Zhou Zhao. Exploring group video captioning with efficient relational approximation. In *ICCV*, pages 15281–15290, 2023. 1, 2

[30] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Exploring cross-video and cross-

modality signals for weakly-supervised audio-visual video parsing. In *NIPS*, pages 11449–11461, 2021. 2

[31] Xiankai Lu, Wenguan Wang, Jianbing Shen, Yu-Wing Tai, David J. Crandall, and Steven C. H. Hoi. Learning video object segmentation from unlabeled videos. In *CVPR*, pages 8957–8967, 2020. 1

[32] Zhiyuan Ma, Jianjun Li, Zezheng Zhang, Guohui Li, and Yongjing Cheng. Intention reasoning network for multi-domain end-to-end task-oriented dialogue. In *EMNLP*, pages 2273–2285, 2021. 3

[33] Adyasha Maharana, Quan Hung Tran, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, and Mohit Bansal. Multimodal intent discovery from livestream videos. In *NAACL*, pages 476–489, 2022. 3

[34] Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3):2276–2289, 2023. 2, 7

[35] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *CVPR*, pages 5261–5270, 2023. 6

[36] Maryam Sadat Mirzaei, Kourosh Meshgi, and Satoshi Sekine. What is the real intention behind this question? dataset collection and intention classification. In *ACL*, pages 13606–13622, 2023. 3

[37] Yishuang Ning, Jia Jia, Zhiyong Wu, Runnan Li, Yongsheng An, Yanfeng Wang, and Helen M. Meng. Multi-task deep learning for user intention understanding in speech interaction systems. In *AAAI*, pages 161–167, 2017. 3

[38] Yawen Ouyang, Jiasheng Ye, Yu Chen, Xinyu Dai, Shujian Huang, and Jiajun Chen. Energy-based unknown intent detection with data manipulation. In *ACL/IJCNLP*, pages 2852–2861, 2021. 1

[39] Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S. Melo, Ana Paiva, and Danica Kragic. Geometric multimodal contrastive representation learning. In *ICML*, pages 17782–17800, 2022. 6

[40] Wasifur Rahman, Md. Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Mohammed E. Hoque. Integrating multimodal information in large pretrained transformers. In *ACL*, pages 2359–2369, 2020. 1, 3, 7

[41] Zhongkai Sun, Prathusha Kameswara Sarma, William A. Sethares, and Yingyu Liang. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *AAAI*, pages 8992–8999, 2020. 7

[42] Xiaoou Tang, Ke Liu, Jingyu Cui, Fang Wen, and Xiaogang Wang. Intentsearch: Capturing user intention for one-click internet image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1342–1353, 2012. 1, 3

[43] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, pages 6558–6569, 2019. 1, 3, 6, 7

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 5

[45] Binglu Wang, Kang Yang, Yongqiang Zhao, Teng Long, and Xuelong Li. Prototype-based intent perception. *IEEE Transactions on Multimedia*, pages 1–12, 2023. 3

[46] Hongwei Wang and Dong Yu. Going beyond sentence embeddings: A token-level matching algorithm for calculating semantic textual similarity. In *ACL*, pages 563–570, 2023. 3

[47] Ning Wang, Wengang Zhou, and Houqiang Li. Contrastive transformation for self-supervised correspondence learning. In *AAAI*, pages 10174–10182, 2021. 2

[48] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, pages 9929–9939, 2020. 2

[49] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Changxin Gao, Yingya Zhang, Deli Zhao, and Nong Sang. Molo: Motion-augmented long-short contrastive learning for few-shot action recognition. In *CVPR*, pages 18011–18021, 2023. 2

[50] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *CVPR*, pages 12176–12185, 2022. 2

[51] Haiping Wu and Xiaolong Wang. Contrastive learning of image representations with cross-video cycle-consistency. In *ICCV*, pages 10129–10139, 2021. 2

[52] Yang Wu, Pengwei Zhan, Yunjian Zhang, LiMing Wang, and Zhen Xu. Multimodal fusion with co-attention networks for fake news detection. In *ACL/IJCNLP*, pages 2560–2569, 2021. 2

[53] Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL*, pages 7617–7630. Association for Computational Linguistics, 2023. 2, 3, 7

[54] Mang Ye, Qinghongya Shi, Kehua Su, and Bo Du. Cross-modality pyramid alignment for visual intention understanding. *IEEE Transactions on Image Processing*, 32:2190–2201, 2023. 3

[55] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *AAAI*, pages 10790–10797, 2021. 7

[56] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016. 6

[57] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, pages 1103–1114, 2017. 7

[58] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *AAAI*, pages 5634–5641, 2018. 7

[59] Hanlei Zhang, Xiaoteng Li, Hua Xu, Panpan Zhang, Kang Zhao, and Kai Gao. TEXTOIR: an integrated and visualized platform for text open intent recognition. In *ACL*, pages 167–174, 2021. 3

[60] Hanlei Zhang, Hua Xu, and Ting-En Lin. Deep open intent classification with adaptive decision boundary. In *AAAI*, pages 14374–14382, 2021. 3

[61] Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. Mintrec: A new dataset for multimodal intent recognition. In *ACMMM*, pages 1688–1697, 2022. 1, 3, 4, 6

[62] Yunhua Zhou, Peiju Liu, and Xipeng Qiu. Knn-contrastive learning for out-of-domain intent classification. In *ACL*, pages 5129–5141, 2022. 3

[63] Yicheng Zou, Hongwei Liu, Tao Gui, Junzhe Wang, Qi Zhang, Meng Tang, Haixiang Li, and Daniel Wang. Divide and conquer: Text semantic matching with disentangled keywords and intents. In *ACL*, pages 3622–3632, 2022. 1, 3