Received: 17 May 2022 | Revised: 29 August 2022 | Accepted: 8 January 2023

IET Computer Vision

DOI: 10.1049/cvi2.12173

ORIGINAL RESEARCH

MCR: Multilayer cross-fusion with reconstructor for multimodal abstractive summarisation

Jingshu Yuan[1,2] | Jing Yun[1,2] | Bofei Zheng[1,2] | Lei Jiao[1,2] | Limin Liu[1]

[1]College of Data Science and Application, Inner Mongolia University of Technology, Huhhot, China

[2]Inner Mongolia Autonomous Region Engineering & Technology Research Center of Big Data Based Software Service, Huhhot, China

Correspondence

Jing Yun, College of Data Science and Application, Inner Mongolia University of Technology, Huhhot, China.
Email: yunjing_zoe@163.com

Funding information

Natural Science Foundation of Inner Mongolia, Grant/Award Number: 2020MS06025; National Natural Science Foundation of China, Grant/Award Number: 62062055; Department of Science and Technology of Inner Mongolia, Grant/Award Number: 2019GG372

Abstract

Multimodal abstractive summarisation (MAS) aims to generate a textual summary from multimodal data collection, such as video-text pairs. Despite the success of recent work, the existing methods lack a thorough analysis for consistency across multimodal data. Besides, previous work relies on the fusion method to extract multimodal semantics, neglecting the constraints for complementary semantics of each modality. To address those issues, a multilayer cross-fusion model with the reconstructor for the MAS task is proposed. Their model could thoroughly conduct cross-fusion for each modality via layers of cross-modal transformer blocks, resulting in cross-modal fusion representations with consistency across modalities. Then the reconstructor is employed to reproduce source modalities based on cross-modal fusion representations. The reconstruction process constrains the fusion representations with the complementary semantics of each modality. Comprehensive comparison and ablation experiments on the open domain multimodal dataset How2 are proposed. The results empirically verify the effectiveness of the multilayer cross-fusion with the reconstructor structure on the proposed model.

KEYWORDS

computer vision, multimedia systems, natural language processing

1 | INTRODUCTION

Content providers release massive multimodal news or broadcasts in various formats for users every day, among which video is becoming a popular data form. This leads to a problem of information overload in which people's energy is wasted on tons of multimodal information. The traditional summarisation model outlines the salient parts of the documents or videos with a brief text. However, it cannot analyse and represent the semantics of multimodal data. Hence, multimodal abstractive summarisation (MAS) emerges as the requirement. MAS models, which use brief text descriptions to summarise the salient parts of multimodal data, like video-text pairs, are significant for helping users quickly understand relevant information.

Existing approaches have obtained promising results, and current methods can be roughly divided into two categories. As shown in Figure 1a, some works [1, 2] applied specific encoders to process source inputs and obtain single-modal embeddings. Then they used a summary generation decoder to fuse multimodal information. However, those methods only performed one-stage multimodal fusion in the summary generation process, lacking interactions to model the consistency (the problem is shown by a yellow dash line in Figure 2).

To solve this issue, Liu et al. [3] proposed a single layer co-attention among multi-encoders to extract the multimodal semantics before the decoder, as shown in Figure 1b. These approaches only adopt a shallow fusion approach to model the semantics for multimodal fusion representation. Consistency is still not fully explored, despite its importance in bridging the semantic gap between different modalities. For example, the text words about snowboard posture should be associated with the video frames showing skaters' movements to build consistency among modalities (shown as green dash line in Figure 2).

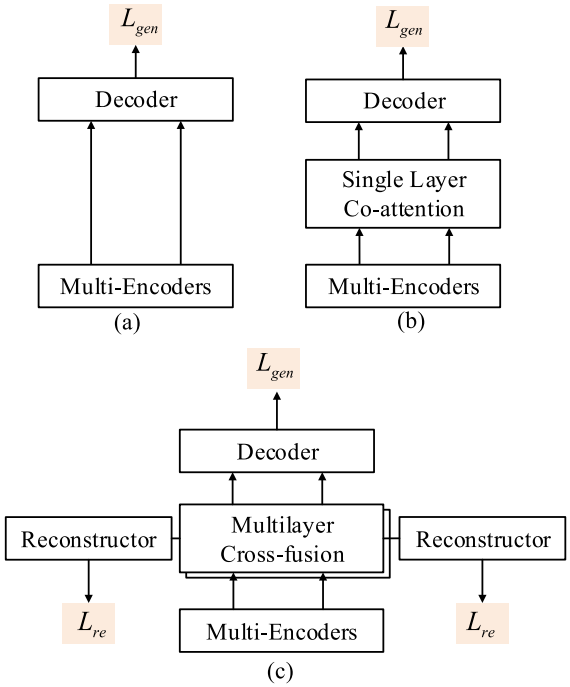


FIGURE 1 (a) Some previous work uses one joint decoder to fuse features from source modalities. (b) Some work used co-attention, based on the scaled dot-product attention, to establish multimodal semantics before the summary decoder. (c) Our proposed MCR model uses multilayer cross-fusion with the reconstructor to jointly learn consistency and complementary semantics of multimodal data. $\mathcal{L}_{re}$ is the loss function of the reconstructor; $\mathcal{L}_{gen}$ is the loss function of the summary decoder.

Besides considering the complementary in each modality, another challenge is that previous works lack the necessary constraint on the complementary for multimodal representations. On the one hand, each modality's complementary information could offer an accurate and comprehensive summary of informative semantics. On the other hand, the previous approaches tend to train the MAS model with uniform loss, lacking the adaptive constraint for complementary information during the multimodal fusion process. This might be since they exclusively relied on the multimodal fusion process (from input modalities to fusion representation) to extract semantics for summary generation, with no information from fusion representation to inputs supplied. However, the reconstruction flow from latent representation to source data has demonstrated its effectiveness in improving generation performance [4, 5].

To address those two challenges, we propose a novel model named MCR for the video-containing MAS task. As shown in Figure 1c, the motivation of our MCR model is the multilayer cross-fusion module with the reconstructor structure. The cross-fusion module aims to capture the consistency between modalities, and the reconstructor is proposed to constrain the complementary semantics of each modality effectively. Our proposed MCR consists of four major modules, as illustrated in Figure 3: Multimodal encoders extract low-level features from source modalities. The Cross-fusion module iteratively conducts the cross-fusion via layers of cross-modal transformer blocks [6] to capture the consistency between modal-

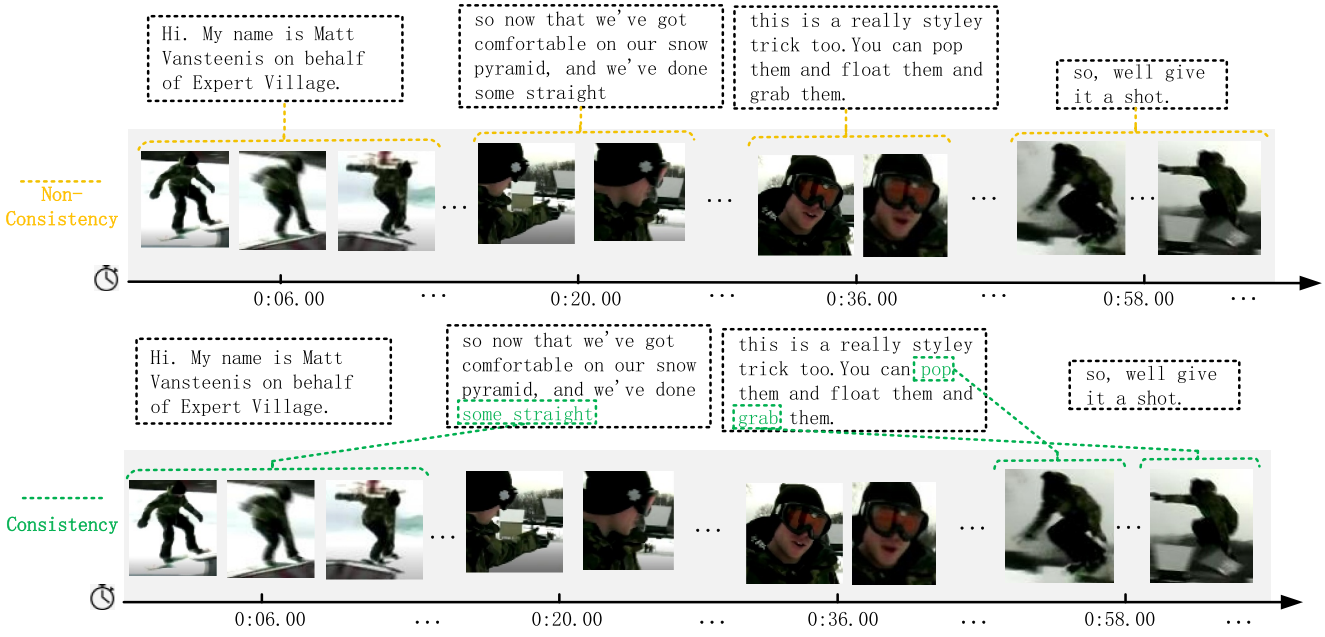


FIGURE 2 Two kinds of alignment examples among modalities. (1) Non-consistency (shown by the yellow dashed line) indicates that the video frames and spoken transcripts occur nearby. However, non-consistency focuses on the temporal match while lacking relevant semantics: The character's current behaviour is unrelated to the 'snowboard'. This phenomenon leads to a lack of meaningful information in multimodal fusion representations. (2) Consistency (shown by the green dashed line) may have a time interval. It connects text modalities (words like 'some straight', 'pop', and 'grab') with more relevant video frames (video frames showing the movement of skaters). Thus, the consistency enriches the fusion representation with relevant semantics, which helps model the salient information for the summary.

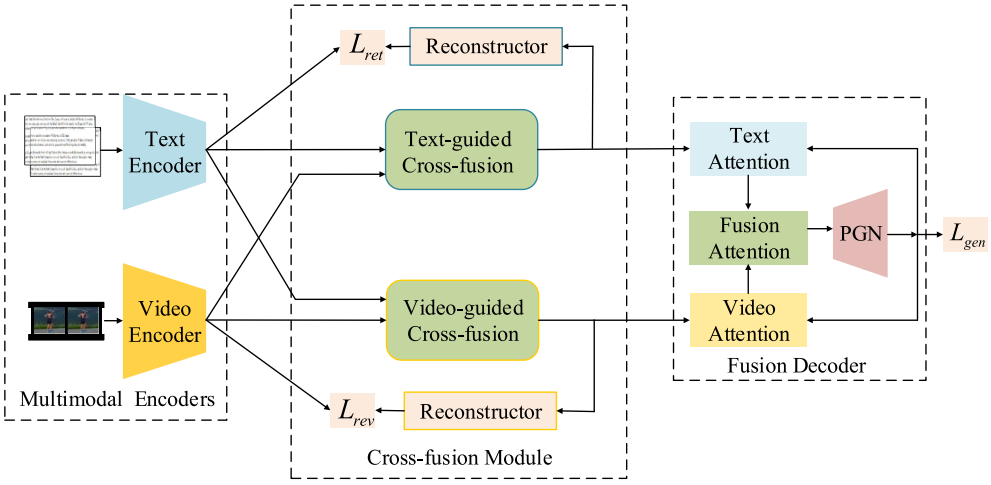


FIGURE 3 The overall architecture for MCR consists of four parts: multimodal encoders, cross-fusion module, reconstructor, and fusion decoder (FD). $\mathcal{L}_{ret}$ is the loss function produced by the text reconstructor, $\mathcal{L}_{rev}$ is the loss function produced by the video reconstructor.

ities, yielding cross-modal fusion representation for each modality. Reconstructors take the cross-modal fusion representation as input to reproduce each source modality accompanied by the reconstruction loss, constraining the cross-modal fusion representation with the complementary semantics of source modalities. The Fusion Decoder (FD) generates a textual summary by incorporating the fusion information. During the backpropagation, we design a separation constraint strategy that combines the reconstruction and generation losses and applies different loss functions to constrain the representations yielded by modules.

We conduct experiments on the large-scale multimodal summarisation dataset, namely How2 [7]. Experimental results show that our proposed MCR achieves competitive performance compared with most baselines. Thorough ablation experiments demonstrate the superiority of the cross-fusion module and reconstruction constraint in improving generation performance. Visualisation analysis indicates that our MCR can effectively preserve intra-modal complementary and inter-modal consistency to enrich the generated summary.

To summarise, the contributions of this work are threefold:

1. We propose a model named MCR for the video-containing MAS task. We propose multilayer cross-fusion implemented by layers of cross-modal transformer blocks. Each modality will have a thorough interaction with another to model the consistency between modalities.
2. We employ a feature-level reconstructor to utilise the information from multimodal fusion representation to reconstruct source modalities. The reconstructor could enhance the connection between fusion representations and each modality, which helps constrain the cross-fusion to preserve more complementary semantics in source modalities.
3. Experimental results show that MCR is competitive and outperforms the comparison baselines under evaluation metrics.

2 | RELATED WORK

The multimodal summarisation task is an essential branch of automatic summarisation that summarises text, images, and videos. However, generating compelling summaries from a combination of modalities is still challenging. Unlike some cross-modal captioning tasks like image-caption [8, 9] or video captioning [10, 11], which focus on detailed descriptions of images or video clips, the multimodal summarisation task requires the model to understand and consolidate the semantics from inter- and intra-modalities. Many studies have been conducted to achieve this goal, including multimodal representation learning and summary generation.

Depending on the sampling method for summary generation, we can divide previous approaches into extractive and abstractive approaches. Those approaches widely adopted the encoder-decoder architecture as their base structure, with a multimodal fusion module as the vital element for analysing source modalities. Therefore, the video-language fusion strategies are also introduced in this section.

2.1 | Multimodal extraction summarisation

Multimodal extractive summarisation selects original content from multimedia inputs as output summaries, such as crucial sentences. Some of these works mainly focus on summarising texts and images. Chen and Zhuge [12] first encoded documents and images with a multimodal Recurrent Neural Network (RNN), then calculated the summary probability of sentences by using text coverage, text redundancy, and image set coverage. Miao et al. [13] proposed an extractive model to integrate multimodal new data and extract keywords as the short title, with a reinforcement training method. Some works utilised a collection of visual, audio, and text modalities sourced from web blogs and multimedia streams to generate summaries [14-17]. Some works have been dedicated to generating multimodal output instead of only text or

images outputs. For example, Jangra et al. [18] proposed a summary generation task that receives text-image-video data as multi-source inputs. The model adopted the Integer Linear Programming framework to extract the salient sentences and key-frames in a video. Besides, Zhu et al. [19] presented an unsupervised graph-based MAS model covering both single-modal and multimodal summarisation goals.

The abstractive approach is more flexible and produces less redundant information than the extraction approach. Leveraging a powerful summary generation model with an effective multimodal process strategy for source modalities is key to constructing MAS models [20].

2.2 | Multimodal abstractive summarisation

MAS methods directly generate a compressed description that does not appear in the source inputs. Early MAS works focussed on summarising texts and images. Lei et al. [21] first introduced the hierarchical attention mechanism to fuse the images and document content in decoding the summary words. Li et al. [22] proposes an aspect-aware multimodal summarisation model for e-commerce products. To improve user satisfaction when reading documents, some works [23, 24] propose generating a compressed description and some selected images for multimodal news documents; however, Zhu et al. [25] found a problem with modality bias when generating multimodal summarisation, so they proposed a multimodal objective function that considers text summary generation and image selection.

Videos contain rich content and have a temporal nature where events are represented chronologically, which is crucial for videos containing the MAS model. Sanabria et al. [7] first released a large-scale multimodal dataset for summarising open-domain videos, namely How2. The dataset provides multi-source information, including video, audio, text transcription, and a human-generated summary. At the same time, the video-containing summarisation task is more challenging due to its complex temporal and spatial features. Based on the How2, Palaskar et al. [1] first proposed to utilise the visual and textual information of video clips into the summary generation process. Khullar and Arora [26] incorporated audio to generate a summary of video content with visual and textual modalities. Liu et al. [3] conducted multistage fusion to interact multi-source modalities together and applied the forget gate module to resist the noise flows from multimodal semantics. Shang et al. [27] introduced a novel short-term order-sensitive attention mechanism to leverage the time clue inside video frames. Liu et al. [2] proposed reducing model parameters with a decoder-only multimodal transformer which combines the source inputs and target summary in the shared feature space.

Despite the success of recent work, most MAS models adopt a shallow fusion approach to model the semantics of multimodal fusion representation. Comprehensive multimodal fusion representation with both consistency and complementary semantics has not yet been fully explored.

2.3 | Video-language fusion strategies

The fusion strategies of video-language (VL) have attracted much research attention recently since the popularity of videos [28]. The transformer [29] model has the superior ability to parallelly model relationships between long sequences. Most VL fusion models broadly use the transformer model in their multimodal fusion process. Those VL models aim to learn an excellent multimodal fusion representation, and they are usually trained by specially self-supervised tasks. Many downstream tasks, such as Video-QA [30], Video Action Recognition [31] and Video Retrieval [32] have benefited from the fine-tuned multimodal representation yielded by VL fusion models. Based on their model structure, we can roughly divide different models into two categories: For the single-stream transformers, like Videobert [33], Clipbert [21] and VLM [34], data of different source modalities is individually processed with a single transformer to capture the intra- and inter-modality features; In two-stream transformers like CBT [35], Vilbert [36] and Univl [37], each modality will be sent to the feature extractors. Then, through the multimodal fusion process in multimodal transformers, we can build the cross-modal relationship from one to another modality.

While few VL fusion models have utilised MAS tasks in their system pipelines, we propose a novel MCR model for the video-containing MAS task to model the comprehensive representation containing consistency and complementary semantics. Inspired by the multimodal fusion model, MulT [6], we design the cross-fusion module implemented by the cross-modal transformer blocks to process text and video source inputs and model the consistency semantics among modalities. Moreover, we use the reconstructor to capture complementary information in source modalities. The backflows from multimodal fusion representations to source inputs can be learnt together as the constraints.

3 | MCR: THE PROPOSED MODEL

In this section, we will explain our MCR in detail. We propose a novel MCR model with the multilayer cross-fusion and reconstructor module, receiving video and text modalities as source inputs. Specifically, the cross-fusion module aims to conduct thorough cross-fusion interactions between modalities to model consistency in multimodal fusion representation. The reconstructor imposes a reconstruction backflow from multimodal fusion representation to the semantic information of source modalities. Thus, we can obtain a reconstruction constraint for the cross-fusion module, which can embed more complementary semantics into multimodal fusion representations.

As illustrated in Figure 3, the proposed MCR consists of four modules, specifically the multimodal encoders, the cross-fusion module, which can be divided into the text-guided fusion branch and the video-guided fusion branch, the reconstructors for text and video reconstruction, and the FD. The cross-modal transformer inside is illustrated in Figure 4,

and the reconstructor inside is shown in Figure 5. Moreover, we apply the separation constraint strategy to offer adaptive constraints on different modules during training, which is illustrated in Figure 6.

3.1 | Problem formula

Our MAS model MCR takes a video and an audio textual transcription as multimodal inputs. The goal of the model is to generate a textual summary that describes the most salient part of the source inputs. Formally, let $T = \{t_1, \ldots, t_n\}$ and $V = \{v_1, \ldots, v_m\}$ represent the textual modality of ground-truth text content and the visual modality of video, respectively, where t_i is the text token and v_i is the feature vector extracted by a pretrained model. The output summary is denoted as a collection of summary sentences $S = \{s_1, \ldots, s_l\}$, where s_i demonstrated the sentence. We can regard the distribution sampling process in the vocabulary dictionary as the summary generation for the given multimodal source inputs that include video and text pairs, which can be presented as

$$\arg\max_{\theta} = (S|T, V; \theta) \tag{1}$$

where θ is the set of trainable parameters.

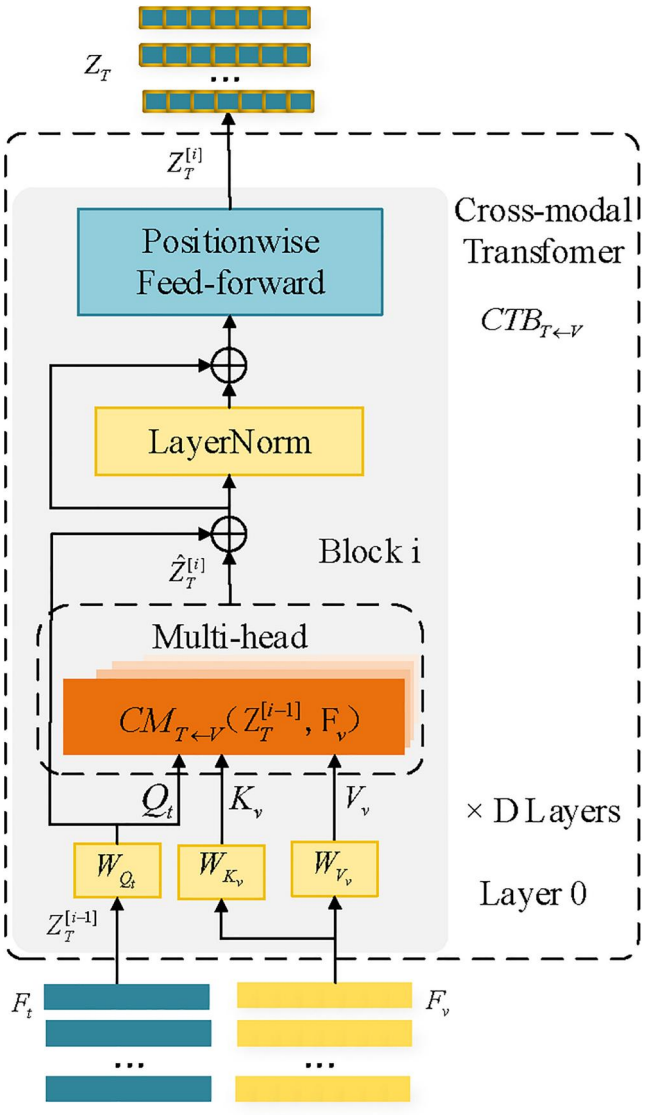


FIGURE 4 Detail of the text-guided cross-fusion $CTB_{T\leftarrow V}$, which is a deep stack of several cross-modal transformer blocks. At every level of the cross-modal transformer block $CTB_{T\leftarrow V}$ during the cross-fusion process, the low-level semantics from the video feature sequence F_v (source modality) are mapped into a set of key-value pairs to thoroughly interact with the text feature sequence F_t (target modality). As a result, the cross-modal transformer could align the correlation semantics across modalities and learn the consistency for fusion representation layer by layer.

3.2 | Multimodal encoders

We need to extract semantic information from the source inputs in textual and video modalities.

3.2.1 | Text encoder

For the textual modality input $T = \{t_1, \ldots, t_n\}$, where n represents the length of the text sequence. We first learn the map for each word x_i that transforms each token into text feature space. Then to get the feature sequence from single word embedding, we apply a Bi-RNN-based text encoder *Enc* to

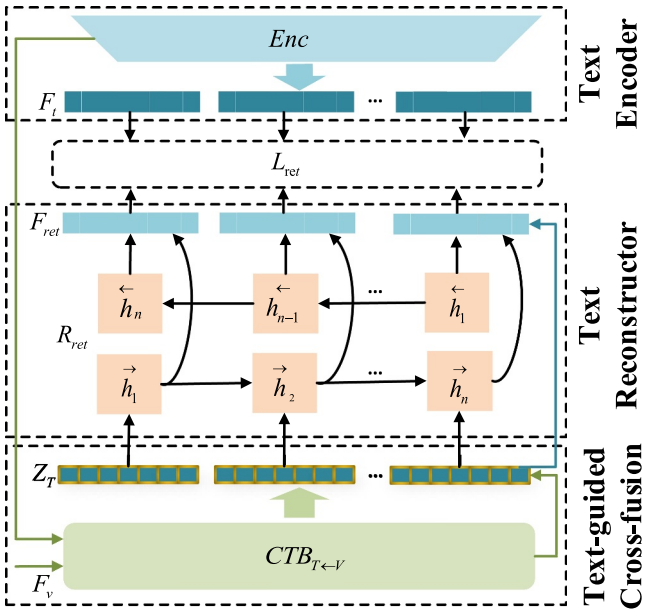


FIGURE 5 An illustration of the proposed multilayer cross-fusion with the reconstructor. The text-guided cross-fusion $CTB_{T\leftarrow V}$ relies on the forward flow from textual and video modalities to the fusion representation (green solid lines) in which the $CTB_{T\leftarrow V}$ generates Z_T with the source modalities features. The text reconstructor R_{ret} exploiting the backward flow from fusion representation (blue solid lines) to textual modality takes the fusion representation Z_T as input and reproduces the reconstructed text feature F_{ret}.

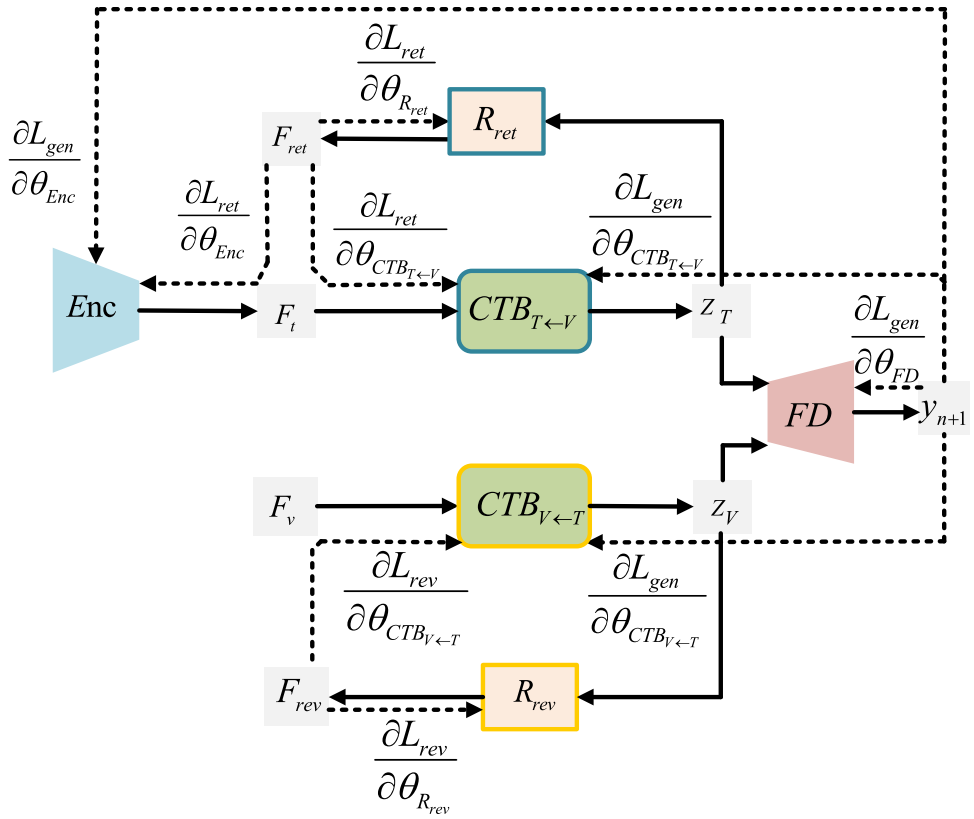


FIGURE 6 The mechanics of the proposed separation constraint strategy: Solid lines show the forward propagation flows, and dashed lines show the backpropagation flows of gradients.

capture the contextual semantics between the sentences (Equation 2). We use Long Short-Term Memory (LSTM) as the Bi-RNN cell.

$$F_t = Enc(x_t, h_{t-1}) \tag{2}$$

where h_t denotes the hidden state of tth step in *Enc*. So the text will be transformed into a feature sequence $F_t \in \mathbb{R}^{n \times d_t}$, where d_t represents the dimension of textual modality space.

3.2.2 | Video encoder

For each video modality, following Ref. [1], we directly take the 2048-dimensional feature representation as source data. Those representations are extracted for every 16 non-overlapping frames by a 3D ResNeXt-101 model [38] which is pre-trained on the Kinetics dataset [39]. Hence, each video sample will be transformed into a feature sequence $F_v \in \mathbb{R}^{m \times d_v}$, where m represents the length of the video clip and d_v represents the dimension of video modality space.

3.3 | Cross-fusion module

Given the F_t and F_v, the thorough analysis of the correlation between modalities would help model the consistency. So our cross-fusion module aims to conduct comprehensive interactions across input modalities and model consistency for multimodal fusion representation. Inspired by Ref. [6], we apply the cross-modal transformer blocks to learn the alignment with another iteratively. At each cross-fusion iteration in the cross-modal transformer block, the cross-modal attention maps the guided modality into key-value pairs to interact with the target modality, which is represented as a query.

We conduct the cross-fusion process for input modalities to preserve more details and essential consistency between multimodal data. Specifically, the cross-fusion module can be divided into the text-guided cross-fusion branch ($CTB_{T\leftarrow V}$) and video-guided cross-fusion branch ($CTB_{V\leftarrow T}$), which are detailed as follows.

3.3.1 | Text-guided cross-fusion

The text-guided cross-fusion $CTB_{T\leftarrow V}$ generates the cross-modal fusion computation under the guidance of a low-level text feature sequence F_t. As shown in Figure 4, $CTB_{T\leftarrow V}$ consists of D layers of cross-modal attention blocks and computes feed-forwardly for $i = \{1, \ldots, D\}$ layers iteratively.

Firstly we need to convert a text feature sequence F_t into queries $Q_t = F_t W_{Q_t}$ (or $Q_t = Z_T^{i-1} W_{Q_t}$ when $i > 1$) and keys as

$K_v = F_v W_{K_v}$, and values as $V_v = F_v W_{V_v}$. Then, we adopt the cross-modal attention mechanism from source K_v/V_v pairs to target feature Q_t and the resulting latent adaption $head_h$ (Equation 3).

$$\begin{aligned} head_h &= CM(Q_t, K_v, V_v) \\ &= softmax\left(\frac{Q_t K_v^T}{\sqrt{d_k}}\right) V_v \\ &= softmax\left(\frac{X W_{Q_t} W_{K_v}^T X^T}{\sqrt{d_k}}\right) V_v \end{aligned} \tag{3}$$

where d_k is the dimension of keys. Therefore, the multi-head cross-modal attention can be calculated as Equation (4) in which we concatenate all the heads and linearly project it to the multimodal dimension space.

$$\begin{aligned} \widehat{Z}_T^{[i]} &= MCM(Q_t, K_v, V_v) \\ &= (head_1 \oplus, \ldots, \oplus head_h) W_o \end{aligned} \tag{4}$$

where W_o is learnable parameters, h represents the number of heads.

To avoid the vanishing gradient problem [40], we added a residual connection between the cross-modal attention computation $\widehat{Z}_T^{[i]}$ and Q_t. The addition result will pass through a position-wise feed-forward network to achieve one fusion iteration (Equation 5).

$$Z_T^{[i]} = FFN\left(\widehat{Z}_T^{[i]}\right) + \widehat{Z}_T^{[i]} \tag{5}$$

where $\widehat{Z}_T^{[i]}$ means a text-guided cross-modal attention computation at layer i.

After D layers of fusion iteration, $CTB_{T \leftarrow V}$ generates the cross-modal fusion representation $Z_T \in \mathbb{R}^{n \times d}$ that is aware of video information under the guidance of text information, where n is the length of the fusion sequence, and d denotes the multimodal space dimension at each time step.

3.3.2 | Video-guided cross-fusion

The video-guided cross-fusion, $CTB_{V \leftarrow T}$ generates the most relevant text information to low-level video feature sequence. The difference between the $CTB_{V \leftarrow T}$ and $CTB_{T \leftarrow V}$ is that they flow in opposite directions. We transform the low-level video features as queries: $Q_v = F_v W_{Q_v}$ (or $Q_v = Z_V^{i-1} W_{Q_v}$ when $i > 1$) and the text as $K_t = F_t W_{K_t}$, and the value as $V_t = F_t W_{V_t}$. The calculation process of the cross-modal attention map from text to video in the ith layer can be presented as follows,

$$\begin{aligned} \widehat{Z}_T^{[i]} &= MCM(Q_v, K_t, V_t) \\ &= (head_1 \oplus head_2, \ldots, \oplus head_h) W_o' \\ &= \left(softmax\left(\frac{Q_v K_t^T}{\sqrt{d_k}}\right) V_t \oplus, \ldots\right) W_o' \end{aligned} \tag{6}$$

$$Z_V^{[i]} = FFN\left(\widehat{Z}_V^{[i]}\right) + \widehat{Z}_V^{[i]} \tag{7}$$

After D layers of fusion iteration, $CTB_{V \leftarrow T}$ generates the cross-modal fusion representations $Z_V \in \mathbb{R}^{m \times d}$ that are aware of text information under the guidance of video information, where m is the length of video frames.

3.4 | Reconstructor

As shown in Figure 5, two kinds of reconstructors are built beside the cross-fusion module, which are expected to reproduce the video and text modalities from the cross-modal fusion representations yielded by the cross-fusion module. Yet, the multimodal inputs have the nature of complexity, so it would be impractical to reconstruct the source modalities directly. The complexity may come from continuous video frames and abstract textual representations.

Therefore, we apply two feature-level reconstructors, R_{ret} and R_{rev}. They take the cross-modal fusion representations Z_T and Z_V as inputs, aiming to reproduce the sequential text and video feature sequences generated by the multimodal encoders. The benefits of such a structure are two-fold. First, the cross-fusion module is promoted to grasp more complementary semantics with the constraints of source modalities. Second, we can also apply reconstruction loss to conduct adaptive constraints for other modules with the generation loss. In practice, the reconstructor is realised by Bi-RNN, which replaces the cell with LSTM.

To ensure the reconstruction can retain most information in text modality, we propose to produce the feature sequence token by token. Starting from Z_T, we apply a text reconstructor R_{ret} to receive the cross-modal fusion representation Z_T and reproduce the feature sequence F_{ret}. The reconstructed feature sequence F_{ret} is expected to be close to the text feature sequence F_t. Hence, the text reconstruction loss is implemented by the Euclidean distance $\psi(\cdot)$ between the reconstructed text feature sequence F_{ret} and source text input F_t (Equation 8).

$$\mathcal{L}_{ret} = \frac{1}{N} \sum_{i=1}^{N} \psi(F_{ret}, F_t) \tag{8}$$

where N represents the total number of our training samples.

Starting from Z_V, we apply a video reconstructor R_{rev} to receive the cross-fusion representation Z_V and

reproduce the video feature sequence F_{rev} frame by frame. The reconstructed feature sequence F_{rev} is expected to be close to the sequential video frame representation F_v. So we use loss $\mathcal{L}_{rev}$ video as the video reconstruction loss (Equation 9).

$$\mathcal{L}_{rev} = \frac{1}{N}\sum_{i=1}^{N}\psi(F_{rev}, F_v) \tag{9}$$

3.5 | Fusion decoder

After the cross-fusion iteration in the CR module, which is constrained by feature-level reconstructors, the features have been mapped to the multimodal space in the form of fusion representations Z_T and Z_V. Now we need to combine the multimodal semantics into the summary generation process. The FD hierarchical read the multimodal fusion information of consistency and complementary in multimodal data and output the text summary. Following Ref. [23], we equip our model with the multimodal hierarchical attention layer to fuse the text and visual context. Moreover, we apply the pointer network [41] to handle the summary generation process.

In each decode step, the hidden state h_t will first perform two attention-based multi-source fusion operations on Z_T and Z_V (Equations 10 and 11). Over the two context vectors, C_T and C_V, fusion attention is constructed to learn the multimodal context vector C_F that is computed (Equation 12).

$$C_T = TextAttention(Z_T, h_t) \tag{10}$$

$$C_V = VideoAttention(Z_V, h_t) \tag{11}$$

$$\begin{aligned} C_F &= FusionAttention(C_T, C_V, h_t) \\ &= softmax(W_t(W_1 C_T + W_2 h_t)C_T) \\ &\quad + softmax(W_v(W_3 C_V + W_4 h_t)C_V) \end{aligned} \tag{12}$$

The fusion representation C_F of hierarchical multimodal fusion accompanied by the decoder hidden state h_t will be sent to calculate the summary distribution (Equation 13).

$$y_{t+1} = FD(C_F, y_t, h_t) \tag{13}$$

We use the negative log-likelihood as the generation loss to keep the relationship between reference y_t and generated summary constant, which is shown as follows:

$$\mathcal{L}_{gen} = -\sum_{t=1}^{l_y}\log P_v(y_t) \tag{14}$$

where l_y presents the length of target summary.

3.6 | Optimisation algorithm: Separation constraint strategy

We develop a separation constraint strategy that conducts different loss functions on MCR to conduct adaptive constraints, especially for the complementary semantics of multimodal fusion representations Z_T and Z_V.

Figure 6 illustrates the mechanics of our separation constraint strategy at training. Let θ_{Enc}, $\theta_{CTB_{T\leftarrow V}}$, $\theta_{CTB_{V\leftarrow T}}$, $\theta_{R_{ret}}$, $\theta_{R_{rev}}$, θ_{FD} respectively denote the parameters of the text encoder *Enc*, text-guide cross-fusion $CTB_{T\leftarrow V}$, video-guide cross-fusion $CTB_{V\leftarrow T}$, text reconstructor R_{ret}, video reconstructor R_{rev}, and fusion decoder *FD*. We analyse the data flow from each module. The data streams outputted by *Enc* and $CTB_{T\leftarrow V}$ flow into both the text reconstruction and summary generation processes. Hence, the first two components are trained on both the text reconstruction loss $\mathcal{L}_{ret}$ and generation loss $\mathcal{L}_{gen}$. The process can be denoted as

$$\min_{\theta_{Enc},\theta_{CTB_{T\leftarrow V}}} \left(\alpha\mathcal{L}_{ret} + \mathcal{L}_{gen}\right) \tag{15}$$

Next, the video-guided cross-fusion $CTB_{V\leftarrow T}$ is trained as follows:

$$\min_{\theta_{CTB_{V\leftarrow T}}} \left(\beta\mathcal{L}_{rev} + \mathcal{L}_{gen}\right) \tag{16}$$

where α and β are trade-off hyperparameters to balance the reconstruction and generation losses.

For $\theta_{R_{ret}}$, $\theta_{R_{rev}}$ and θ_{FD}, we apply $\mathcal{L}_{ret}$, $\mathcal{L}_{rev}$ and $\mathcal{L}_{gen}$ as loss constraint, respectively.

4 | EXPERIMENTAL SETUP

To evaluate the performance of our MCR model in MAS, we conduct comprehensive experiments evaluating textual and video modalities separately and jointly on the How2 dataset. We also use multiple text evaluation metrics to show the quality of the generated summaries.

4.1 | How2 dataset

We evaluate the proposed MCR model on the How2 dataset [7]. The How2 dataset is a large-scale multimodal dataset. It consists of 79,114 How2 instructional videos with an average length of 1.5 min and a total of 2000 h, accompanied by corresponding ground-truth English transcripts with an average length of 291 words, crowdsourced Portuguese translations of transcripts, and user-generated summary with an average length of 33 words. Besides, the videos belong to the open domain and range widely between different topics, such as cooking, sports etc. The statistics are shown in Table 1.

4.2 | Baselines

We compare our model with the following baseline models of single or multiple modalities in an abstractive method:

4.2.1 | Text only

S2S [42] is a standard RNN model with a global attention map for the sequence-sequence generation task; **PG** [41] is a classical text-only summarisation model with sufficient performance in solving the out-of-vocabulary issue; **FT** is an encoder-decoder transformer model for text sequence.

4.2.2 | Video only

VideoRNN [1] is a backbone model for video-only abstractive summarisation tasks; **MT** [44] is a Transformer-based encoder-decoder architecture receiving visual feature sequence of a video clip for end-to-end dense video captions; **VinVL** [45] is the object detection model to provide object-centric representations of images for the VL task.

4.2.3 | Video + text

HA [1] is a baseline multiencoder-decoder model with a hierarchical attention strategy in the decoder part; **MFFG** [3] is a multi-stage fusion architecture to fuse multi-source data, which suppresses the flow of multimodal noise via a forget gate module; **D-MmT** [2] is a decoder-only multimodal transformer framework for video-containing the MAS task.

TABLE 1 The statistics of the How2 dataset

Datasets	Train	Val	Test
Videos	73,993	2965	2156
Hours	1766.6	71.4	51.7

4.3 | Implement details

For all models, we set the word embedding dimension and the hidden dimension to 128. For the text source input, the encoding step is set to 200, while the minimum decoding step is 10, and the maximum step is 30. In the cross-fusion module, the stacked cross-modal transformer layers number is set to four. The vocabulary is built based on the How2 dataset and do not use pre-trained word embeddings, and the vocabulary size is limited to 50k. We separately set the trade-off hyperparameters α and β to 0.55 and 0.1, respectively. We train the model for 30 epochs with a batch size of 4–8 and an initial learning rate of 1.5×10^{-4} and use a warmup step of 3000.

We conduct all the experiments on the platform of Ubuntu 16.04 with Tesla P100 Graphics Processing Units and 32G memory size. During training, our model is optimised using the Adam optimiser [46], and we also apply gradient clipping with a range of [−2, 2]. All the parameters in our model are initialised by Gaussian distribution. During generating for the test, we use beam search with a beam size of three. We decode until the end of a sequence token is reached.

4.4 | Evaluation metrics

We use the following evaluation metrics to measure the similarity between the generated summary and reference: ROUGE-1,2,L [47] and BLEU-1,2,3,4 [48] are used to calculate the recall and precision of the longest common sequence overlaps; METEOR [49] computes the F-Score based on matches

TABLE 2 Results on the How2 test set

Modality	Method	B-1	B-2	B-3	B-4	R-1	R-2	R-L	METEOR	CIDEr
Text-only	S2S [42]	55.25	45.63	39.92	35.84	58.62	40.61	53.82	27.64	2.349
	PG [41]	55.35	45.62	39.82	35.76	57.28	39.51	52.82	26.81	2.134
	FT [43]	56.63	46.71	40.82	36.67	59.05	41.06	54.38	27.72	2.296
Video-only	VideoRNN [1]	44.15	32.92	26.93	22.74	46.58	26.26	41.53	19.92	1.149
	MT [44]	49.61	38.47	32.91	27.45	51.92	32.03	46.81	22.92	1.461
	VinVL [45]	52.43	40.87	35.00	30.67	54.32	35.67	49.76	24.83	1.780
Text + video	HA (RNN) [1]	57.24	47.71	41.83	37.52	60.37	42.51	55.70	28.84	2.476
	HA (Tran) [1]	58.61	48.35	43.34	38.12	60.28	43.16	55.94	28.95	2.512
	MFFG (RNN) [3]	59.10	50.40	45.10	41.10	62.30	46.10	58.20	30.10	2.690
	MFFG (Tran) [3]	60.00	50.90	45.30	41.30	61.60	45.10	57.40	29.90	2.671
	D-MmT [2]	60.32	50.75	45.43	41.28	61.60	45.12	58.26	30.12	2.670
	Ours	**60.60**	**51.23**	**45.82**	**41.30**	**61.67**	**45.15**	**59.00**	**30.13**	**2.700**

Note: To compare the performance of our model with other baselines, we point out our model results in bold.

Abbreviations: B, BLEU; R, ROUGE; RNN, Recurrent Neural Network.

strings and using stemming, exact matches and semantic similarity to resolve correspondences on the word level; CIDEr [50] is a consensus-based assessment metric that capture the concepts of grammaticality, saliency, significance, and accuracy.

5 | RESULTS AND ANALYSIS

5.1 | Model performance

5.1.1 | Effect of multilayer cross-fusion with reconstructor structure

1) Table 2 shows the results of comparative experiments. We can find that text-only or video-only models are disadvantaged under most evaluation metrics compared with multimodal models. The reason may be that multimodal models can build more or less correlated representations from different modalities and semantic spaces, which can help capture the comprehensive multimodal semantic in improving the quality of the generated summary. This study also demonstrates that complementary modalities boost MAS tasks' generation performance more than single-modality input.
2) The proposed MCR performs better than baseline methods in most automatic evaluation metrics. In particular, compared to the HA (Tran), our model made significant progress in ROUGE-L by 3.06 points, BLEU-1 by 1.99 points, and ROUGE-1 by 1.39 points. Compared to baseline MFFG (Tran), our model improves BLEU-1 by 0.6 points, BLEU-3 by 0.52 points, and ROUGE-L by 1.6 points. Compared to the strong baseline, D-MmT, our model still has a competitive advantage and improves ROUGE-L by 0.74 points, a slight improvement in BLEU-1 by 0.28 points, and ROUGE-1 by 0.07 points, which indicates the superiority of our proposed multilayer cross-fusion with reconstructor structure.

5.1.2 | Effect of the separation constraint strategy

As mentioned in Section 3.6, we apply the separation constraint strategy to adaptively constrain the backpropagation of different modules. To study the effect of the separation constraint strategy, we conduct two other loss constraint strategies on the whole model as comparison experiments.

- We only use the generation loss $\mathcal{L}_{gen}$ as the constraint loss signal to train the MCR model.
- We train the model under the sum of reconstruction losses $\mathcal{L}_{ret}, \mathcal{L}_{rev}$ and generation loss $\mathcal{L}_{gen}$.

Table 3 shows this performance variation, and we can find that

1) Compared with the constraint strategies that add reconstruction constraints $\mathcal{L}_{ret}$ and $\mathcal{L}_{rev}$ $\{2, 3\}$, the weakest performance is the only generation loss constraint $\{1\}$, which may be because only focussing on the generation process makes the model lack the consistent semantics in multimodal fusion representations.
2) When we train the model together with reconstruction losses $\{2\}$, compared with $\{1\}$, the model performance improves BLEU-1 by 0.7 points and ROUGE-L by 1.25 points.
3) The separation constraint strategy $\{3\}$ is better than the other two under all the evaluation metrics. The separation constraint strategy $\{3\}$ help the MCR model improve by a large step compared to $\{1\}$ in ROUGE-1 by 1.79 points; compared to the strategy $\{2\}$ still improves in ROUGE-L by 0.75 points. This demonstrates the effectiveness of our proposed separation constraint strategy that can make the cross-fusion process focus more on complementary meanings in each source modality.

In addition, to evaluate the robustness of our model within the separation constraint strategy, we try to weigh the summation of the reconstruction and the generation losses, specifically α and β in Equations (15) and (16). Since our model performs best when $\alpha = 0.55$ and $\beta = 0.1$, we varied the values of α and β from 0.1 to one at the step of 0.15. As can be seen from Figure 7, our model may fluctuate on ROUGE-L by about 2.1 points. This indicates that our model is relatively stable to the trade-off hyperparameter when they vary around the optimal values.

5.2 | Ablations

To further verify the role cross-fusion module and reconstructors, we conduct ablation experiments to validate the impact of the various modules of proposed MCR model. We divide the cross-fusion module into text-guide cross-fusion ($CTB_{T\leftarrow V}$) and video-guide cross-fusion ($CTB_{V\leftarrow T}$) in the cross-fusion module, the following text and video reconstructor (R_{ret}, R_{rev}), and the final *FD*. We retrain our approach by ablating one or more of them.

TABLE 3 Comparison analysis on different loss constraint strategies

No	Constraint strategy	B-1	B-2	B-3	B-4	R-1	R-2	R-L	METEOR	CIDEr
1	$\mathcal{L}_{gen}$	58.81	50.62	45.32	41.00	60.00	44.87	57.00	29.95	2578
2	$\mathcal{L}_{ret} + \mathcal{L}_{rev} + \mathcal{L}_{gen}$	59.51	50.75	45.28	41.10	61.54	45.10	58.25	30.10	2.600
3	Separation constraint strategy	**60.60**	**51.23**	**45.82**	**41.30**	**61.67**	**45.15**	**59.00**	**30.13**	**2.700**

Note: We conduct two kinds of loss constraints on whole MCR model$\{1, 2\}$, and the separation constraint strategy$\{3\}$. To show the performance of different loss strategies, we point out the results of using our proposed separation constraint strategy in bold.

- We retrain only $CTB_{T\leftarrow V}$ and only $CTB_{V\leftarrow T}$ and replace *FD* with a standard decoder to handle single-modal feature representations.
- We add the reconstructors R_{ret} and R_{rev} to the above $CTB_{T\leftarrow V}$ and $CTB_{V\leftarrow T}$ models separately.
- We retain $CTB_{T\leftarrow V}$, $CTB_{V\leftarrow T}$, *FD* and remove all the reconstructors of the full model.

Table 4 lists the results on the How2 dataset. We can observe that:

1) The $CTB_{V\leftarrow T}$'s performance is weaker than the text-guided fusion module $CTB_{T\leftarrow V}$, while the performances of the $CTB_{T\leftarrow V}$ models exceed the performances of most the single-modality models.
2) Compared with using only $CTB_{T\leftarrow V}$ or $CTB_{V\leftarrow T}$, using $CTB_{T\leftarrow V}$ and $CTB_{V\leftarrow T}$ together with *FD* {5}further improves the model effect. It shows that the cross-fusion module can effectively model the multimodal semantics among modalities and improve summary generation performance via multimodal fusion representations.
3) When we add the reconstructors R_{ret}, R_{rev} to the model structure, all the performances improve with the reconstruction constraints $\{2, 4, 5\}$. This further verifies the ability of reconstruction constraints to retain the complementary semantics of source modalities for better generation performance.

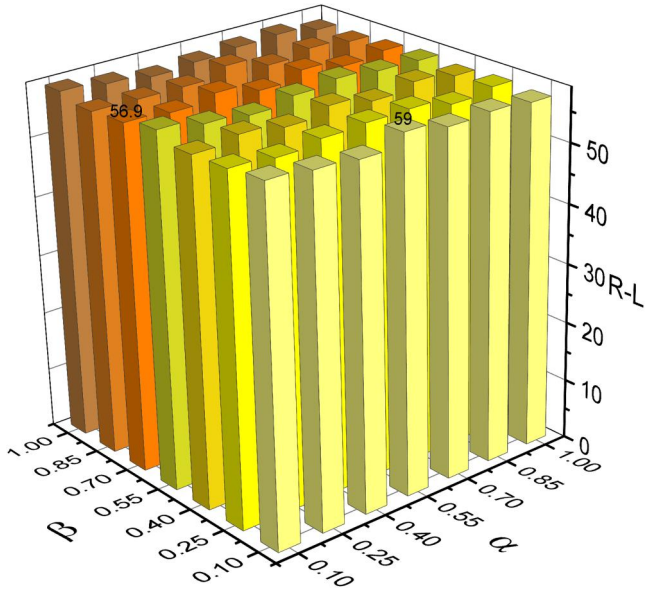


FIGURE 7 Sensitivity analysis for α and β on the How2 dataset. R-L is ROUGE-L. To show the model's sensitivity to the evaluation metric, we have marked the highest and lowest scores above the bars.

5.3 | Visualisation of cross-fusion module

To understand what MCR have learnt while generating the multimodal fusion representations, we take a visualisation study to display the source modalities with the highest attention in each cross-fusion process step. We visualise the attention weights in the cross-modal transformer block to empirically inspect what kind of alignment our model attends.

As shown in Figure 8, we observed that the cross-fusion module had learnt the consistency between source transcripts and video clips, which attaches significant words with video frames with consistency visual features (like stronger posture movemen change). For example, (1) the MCR attends the video frames, which show the consistency actions with keywords in ground-truth summaries, such as 'separate garlic'. For the text modality, MCR also has the same effect (e.g. retain the word 'punch', 'break apart', 'clove'). (2) as shown in the lightest colour block, we can infer that MCR has built consistency among source modalities (e.g. 'punch', 'break apart', and 'clove' with the frames shown on the top separately). This study demonstrates that: our MCR can capture consistency with salient semantics. In addition, MCR can retain more meaningful features in source inputs to help model the semantic relationship, which may source from the reconstruction constraint.

5.4 | Qualitative analysis

To better know when one modality and our cross-modal fusion embeddings help enhance the performance of another modality we dive deeper into qualitative studies on two different

TABLE 4 Ablation analysis on the How2 test set

No	Method	B-1	B-2	B-3	B-4	R-1	R-2	R-L	METEOR	CIDEr
1	$CTB_{T\leftarrow V}$	58.72	49.25	43.68	39.56	60.75	44.05	56.35	29.34	2.541
2	$CTB_{T\leftarrow V} + R_{ret}$	59.12	50.14	44.70	40.78	61.25	44.87	57.29	29.56	2.550
3	$CTB_{V\leftarrow T}$	58.00	48.01	42.11	37.98	59.97	42.21	55.92	28.89	2.512
4	$CTB_{V\leftarrow T} + R_{rev}$	58.15	48.13	42.34	38.12	60.12	42.73	56.00	29.05	2.578
5	$CTB_{V\leftarrow T} + CTB_{T\leftarrow V} + FD$	59.30	49.78	44.89	39.89	60.80	44.67	57.60	29.74	2.600
6	$CTB_{T\leftarrow V} + R_{ret} + CTB_{V\leftarrow T} + R_{rev} + FD(full)$	**60.60**	**51.23**	**45.82**	**41.30**	**61.67**	**45.15**	**59.00**	**30.13**	**2.700**

Note: To validate the impact of the various modules, we point out the results of the complete model with all modules in bold.

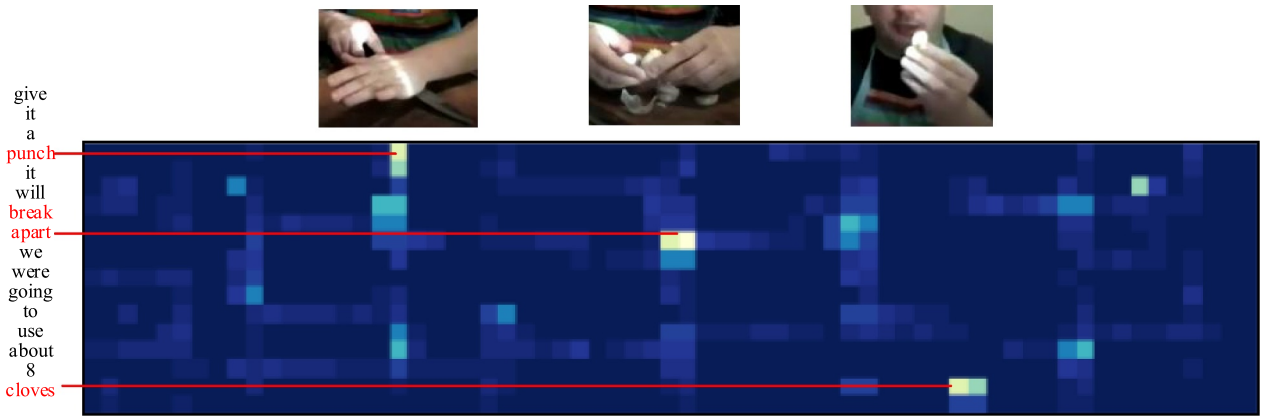
Abbreviations: $CTB_{T\leftarrow V}$, text-guided cross-fusion; $CTB_{V\leftarrow T}$, video-guided cross-fusion; FD, fusion decoder; R_{ret}, text reconstructor; R_{rev}, video reconstructor.

situations. Furthermore, we list some examples to compare the results from multimodal models.

Figure 9 presents the qualitative examples from MCR and some unimodal baselines. We first pick an example where the video-only model produces a better summary than the text-only model. Then we select another example with a vice-versa result with the text modality. In the left example, the text-only model can only focus on the part of the transcripts, so the generated summary could not capture the key behaviours in the video. This may be due to the characters' colloquial expressions, making it more challenging to extract critical information. It can be seen from the MCR results that our model can capture proper nouns (e.g. 'boardslide 180 melon') related to human actions in the original transcripts. In contrast, the only-video model can hardly recognise them. This is because MCR can enhance the modelling of relevant keywords with the guidance of visual information. In the right example, the video-only model's generation performance is not good: it could hardly identify the character's purpose only by visual features. We can only recognise salient information mainly through text modality because the video frames have little relation to the keywords. From the results of MCR, we can find that our multimodal model effectively retains the advantages of text modalities and achieves similar or even richer summary results (e.g. 'scratch').

The above comparisons show that unimodal data is not well represented in some videos. Hence it is difficult to capture salient information for uni-modal models, leading to poor generation performance in some videos. In contrast, our model can significantly retain the dominant modality's features and effectively capture the semantic consistency among multimodal alignment, generating better cross-fusion representations. Such representations can help generate a more comprehensive and accurate summary, thus compensating for the shortcomings of unimodal models under certain conditions.

Figure 10 presents the qualitative examples of How2. As can be seen, the ground truth reference contains more informative details about the visual information and text content. This poses more challenges to modelling the crucial semantics of textual and video modalities. However, we can find that MCR can generate a more accurate and fluent summary for video clips and contains more rich information for the presented examples. In the first example, MFFG either includes some unimportant details (e.g. 'snap' and 'kick returner') or misses the critical points (e.g. '3-point stance' and 'play centre'



GT:Learn how to separate garlic to make bruschetta with expert cooking tips in this free traditional Italian recipe video clip.

FIGURE 8 An example of visualising the attention weights in the cross-modal attention block of the cross-fusion module. We gather the attention weights and utilise the colour blocks in the heatmap to represent the attention degree of the text modality (source transcript) to the video modality. Due to space limitations, we only represent part of the map. GT is the ground truth summary in the How2 dataset.

Modality	Method	Rouge-L	Output
-	GroundTruth	-	Learn how to boardslide 180 melon on a snowboard from our expert on snowboard tricks in this free extreme sports video on doing snowboard tricks at the park
Text-Only	FT	27.9	In this video, we're going to learn how to snowboard the terrain park here at beautiful Mt. Brighton in Brighton, Michigan.
Video-Only	MT	48.4	In this free extreme sports video on completing snowboard tricks
Text+Video	MCR	66.6	Learn how to boardslide 180 melon on a snowboard from our expert on snowboard tips.

Modality	Method	Rouge-L	Output
-	GroundTruth	-	In billiards, you can automatically lose if you sink the balls in the wrong order or scratch on the 8 ball, get expert tips and advice on pool and 8 ball in this free video
Text-Only	FT	31.7	Learn how to automatically lose the game in 8 ball.
Video-Only	MT	22.2	how to to talk in in the expert video .
Text+Video	MCR	60.0	when you scratch on the 8 ball, you may automatically lose the game in 8 ball, get expert tips on 8 ball in this video

FIGURE 9 Qualitative examples generated by MCR and uni-modal baselines. We also show the transcripts below each video frame.

…What we want to talk about right now is the importance of our defensive backs and coming up to stop the run play. Obviously, the main thing to the 6-2 defense …

GT: learn how to fire off from a 3 - point stance and how to play center in this free video clip on football . get football tips from a coach and improve your playing .

MFFG: learn some great tips on how to do a snap as a play as a kick returner in this free video clip on sports .

MCR: learn some football tips on how to fire off from a 3 - point stance and play center in this free video clip .

…We are working on pad drills today. We are going to start with some basic pad drills and the one that we are going to do is called our straight punch.…

GT: in kickboxing , fire the hips and waist to perform the moving straight lead and uppercut pad drill combination . practice this powerful fighting combo with tips from a martial arts instructor in this free kickboxing video .

MFFG: kickboxers use kickboxing techniques such as a straight cut in this free martial arts training video featuring a black belt instructor .

MCR: moving the hips and waist to perform the straight lead and uppercut kickboxing combination . practice those kickboxing techniques with tips in this free instructional video.

FIGURE 10 Qualitative examples generated by MCR and multimodal baselines. We also show the source data in the dataset, including the raw input text, the reference summary, and representative video frames picked in the video clip.

about football tips). These generation errors could be attributed to the lack of consistency critical information to recognise the consistency among modalities. In the second example, MCR can generate the summary with 'hips and waist' by using the visual information, which could be cues for locating key actions and help users understand video content more easily. In contrast, MFFG fails in this case and misses important subtle information (e.g. 'fire the hips and waist' and 'straight lead and uppercut').

5.5 | Computational analysis

To study the computational trade-off that comes into effect with the introduction of the cross-fusion and reconstruction modules. We compare the training time and parameter size.

As shown in Figure 11, we can observe the following: (1) Regarding training time, HA takes the shortest time to peak and has the fewest model parameters. This is due to its simple structure: only one fusion layer is used on the decoding module. In comparison, our MCR takes the longest time to achieve better results. This result may be because MCR has to take more time to balance the components and whole model performance when applying the separation constraint strategy to optimise the summary generation process. (2) Regarding parameter resources, our model has a medium number of parameters. The MFFG is the largest of these, probably because of its multi-step fusion mechanism. The MFFG is the largest of these, probably due to its multi-step fusion mechanism, which includes many fusion modules and forget gates. Although our MCR adds a reconstruction module, only a simple sequence model can achieve the effect of constraints. So this does not exacerbate the model parameters.

6 | CONCLUSION

This paper proposes a novel MCR model for the video-containing the MAS task. We design a cross-fusion module that applies layers of cross-modal transformer blocks to

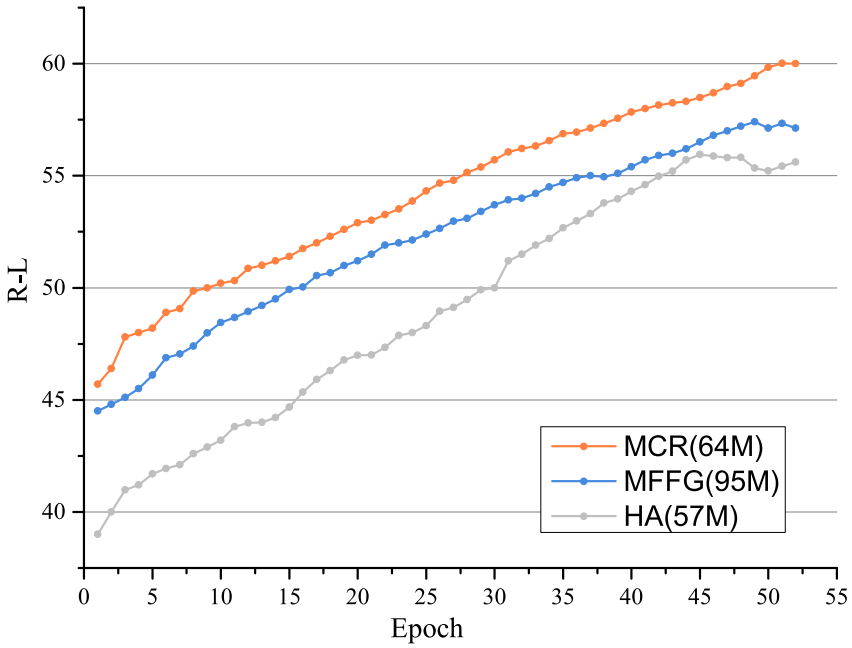


FIGURE 11 The validation set performance of the comparison of model parameter size and training time on the How2 dataset.

conduct a comprehensive interaction between modalities and model the consistency. We introduce a feature-level reconstructor to constrain the cross-fusion process with more complementary semantics in each modality. In addition, we introduce a separation constraint strategy that makes full use of reconstruction and generation loss to offer adaptive constraints on different modules of MCR. Experimental results on the How2 dataset confirm the superiority of our approach's generation performance. Detailed information may be lost since the dataset only contains pre-trained feature sequences to represent the video content. So we will extend our work to the MAS dataset, including more subtle information to improve generation performance further.

AUTHOR CONTRIBUTIONS

Jingshu Yuan: Conceptualisation; Methodology; Software; Writing – original draft; Writing – review & editing. **Jing Yun**: Conceptualisation; Formal analysis; Methodology. **Bofei Zheng**: Investigation; Validation. **Lei Jiao**: Investigation; Validation. **Limin Liu**: Conceptualisation; Formal analysis; Methodology.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grants 62062055, in part by the Department of Science and Technology of Inner Mongolia under Grants 2019GG372, and in part by the Natural Science Foundation of Inner Mongolia under Grants 2020MS06025.

CONFLICT OF INTEREST

The authors declared that they have no conflicts of interest to this work.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Jingshu Yuan https://orcid.org/0000-0001-7937-917X

REFERENCES

1. Palaskar, S., et al.: Multimodal abstractive summarization for how2 videos. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 6587–6596 (2019)
2. Liu, N., et al.: D-MmT: a concise decoder-only multi-modal transformer for abstractive summarization in videos. Neurocomputing 456, 179–189 (2021). https://doi.org/10.1016/j.neucom.2021.04.072
3. Liu, N., et al.: Multistage fusion with forget gate for multimodal summarization in open-domain videos. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1834–1845 (2020)
4. Han, Z., et al.: Y2seq2seq: cross-modal representation learning for 3D shape and text by joint reconstruction and prediction of view and word sequences. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 126–133 (2019)
5. Wang, B., et al.: Reconstruction network for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7622–7631 (2018)
6. Hubert Tsai, Y.-H., et al.: Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the Conference. Association for Computational Linguistics. Meeting, vol. 2019, p. 6558. NIH Public Access (2019)
7. Sanabria, R., et al.: How2: a large-scale dataset for multimodal language understanding. arXiv preprint arXiv:1811.00347 (2018)
8. Chen, J., Zhuge, H.: News image captioning based on text summarization using image as query. In: Semantics Knowledge and Grid (2019)
9. Li, X., et al.: Without detection: two-step clustering features with local–global attention for image captioning. IET Comput. Vis. 16(3), 280–294 (2022). https://doi.org/10.1049/cvi2.12087
10. Iashin, V., Rahtu, E.: Multi-modal dense video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 958–959 (2020)

11. Tang, M., et al.: Clip4caption: clip for video caption. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 4858–4862 (2021)
12. Chen, J., Zhuge, H.: Extractive text-image summarization using multi-modal RNN. In: Semantics Knowledge and Grid (2018)
13. Miao, L., et al.: Multi-modal product title compression. Inf. Process. Manag. 57, 102123-1–102123-12 (2020)
14. Li, H., et al.: Multi-modal summarization for asynchronous collection of text, image, audio and video. In: Empirical Methods in Natural Language Processing (2017)
15. Li, H., et al.: Read, watch, listen, and summarize: multi-modal summarization for asynchronous text, image, audio and video. IEEE Trans. Knowl. Data Eng. 31(5), 996–1009 (2019). https://doi.org/10.1109/tkde.2018.2848260
16. Psallidas, T., et al.: Multimodal summarization of user-generated videos. Appl. Sci. 11(11), 5260 (2021). https://doi.org/10.3390/app11115260
17. Li, M., et al.: Timeline summarization based on event graph compression via time-aware optimal transport. In: Empirical Methods in Natural Language Processing (2021)
18. Jangra, A., et al.: Text-image-video summary generation using joint integer linear programming. In: European Conference on Information Retrieval (2020)
19. Zhu, J., et al.: Graph-based multimodal ranking models for multimodal summarization. In: ACM Transactions on Asian and Low-Resource Language Information Processing (2021)
20. Yu, T., et al.: Vision guided generative pre-trained language models for multimodal abstractive summarization. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3995–4007 (2021)
21. Lei, J., et al.: Less is more: ClipBERT for video-and-language learning via sparse sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7331–7341 (2021)
22. Li, H., et al.: Multimodal sentence summarization via multimodal selective encoding. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 5655–5667 (2020)
23. Zhu, J., et al.: MSMO: multimodal summarization with multimodal output. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4154–4164 (2018)
24. Chen, J., Zhuge, H.: Abstractive text-image summarization using multi-modal attentional hierarchical RNN. In: Empirical Methods in Natural Language Processing (2018)
25. Zhu, J., et al.: Multimodal summarization with guidance of multimodal reference. Proc. AAAI Conf. Artif. Intell. 34(05), 9749–9756 (2020). https://doi.org/10.1609/aaai.v34i05.6525
26. Khullar, A., Arora, U.: MAST: multimodal abstractive summarization with trimodal hierarchical attention. arXiv preprint arXiv:2010.08021 (2020)
27. Shang, X., et al.: Multimodal video summarization via time-aware transformers. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 1756–1765 (2021)
28. Li, J., et al.: Align before fuse: vision and language representation learning with momentum distillation. Adv. Neural Inf. Process. Syst. 34, 9694–9705 (2021)
29. Han, K., et al.: A survey on vision transformer. IEEE Trans. Pattern Anal. Mach. Intell. 45(1), 87–110 (2022). https://doi.org/10.1109/tpami.2022.3152247
30. Jiang, J., et al.: Divide and conquer: question-guided spatio-temporal contextual attention for video question answering. Proc. AAAI Conf. Artif. Intell. 34(07), 11101–11108 (2020). https://doi.org/10.1609/aaai.v34i07.6766
31. Girdhar, R., et al.: Video action transformer network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 244–253 (2019)
32. Dzabraev, M., et al.: MDMMT: multidomain multimodal transformer for video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3354–3363 (2021)
33. Sun, C., et al.: VideoBERT: a joint model for video and language representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7464–7473 (2019)
34. Hu, X., et al.: VLM: task-agnostic video-language model pre-training for video understanding. arXiv preprint arXiv:2105.09996 (2021)
35. Sun, C., et al.: Learning video representations using contrastive bidirectional transformer. arXiv preprint arXiv:1906.05743 (2019)
36. Lu, J., et al.: ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Adv. Neural Inf. Process. Syst. 32, 13–23 (2019)
37. Luo, H., et al.: UniVL: a unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353 (2020)
38. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6546–6555 (2018)
39. Kay, W., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
40. Tan, H.H., Lim, K.H.: Vanishing gradient mitigation with deep learning neural network optimization. In: 2019 7th International Conference on Smart Computing & Communications (ICSCC), pp. 1–4. IEEE (2019)
41. See, A., Liu, P.J., Manning, C.D.: Get to the point: summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368 (2017)
42. Luong, M.-T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
43. Vaswani, A., et al.: "Attention is all you need." In: Advances in Neural Information Processing Systems, vol. 30, pp. 6000–6010 (2017)
44. Zhou, L., et al.: End-to-end dense video captioning with masked transformer. In: Computer Vision and Pattern Recognition (2018)
45. Zhang, P., et al.: VinVL: revisiting visual representations in vision-language models. In: Computer Vision and Pattern Recognition (2021)
46. Kingma, D.P., Jimmy, B.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
47. Lin, C.-Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)
48. Kishore, P., et al.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
49. Denkowski, M., Alon, L.: Meteor 1.3: automatic metric for reliable optimization and evaluation of machine translation systems. In: Workshop on Statistical Machine Translation (2011)
50. Vedantam, R., Zitnick, C.L., Parikh, D.: CIDEr: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566–4575 (2015)

How to cite this article: Yuan, J., et al.: MCR: multilayer cross-fusion with reconstructor for multimodal abstractive summarisation. IET Comput. Vis. 1–15 (2023). https://doi.org/10.1049/cvi2.12173