Vision-Text cross-modal fusion for accurate video captioning

Kaouther Ouenniche, Ruxandra Tapu, Titus Zaharia

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 September 2023, accepted 7 October 2023, date of publication 12 October 2023, date of current version 24 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3324052

RESEARCH ARTICLE

Vision-Text Cross-Modal Fusion for Accurate Video Captioning

KAOUTHER OUENNICHE, RUXANDRA TAPU, (Member, IEEE), AND TITUS ZAHARIA, (Member, IEEE)

Institut Polytechnique de Paris, Télécom SudParis, Laboratoire SAMOVAR, 91011 Evry, France

Corresponding authors: Kaouther Ouenniche (kaouther.ouenniche@telecom-sudparis.eu), Ruxandra Tapu (ruxandra.tapu@telecom-sudparis.eu), and Titus Zaharia (titus.zaharia@telecom-sudparis.eu)

This work has been carried out within the framework of the Artificial Intelligence for Television (AITV) Joint Laboratory established between Telecom SudParis and France Televisions.

ABSTRACT In this paper, we introduce a novel end-to-end multimodal video captioning framework based on cross-modal fusion of visual and textual data. The proposed approach integrates a modality-attention module, which captures the visual-textual inter-model relationships using cross-correlation. Further, we integrate temporal attention into the features obtained from a 3D CNN to learn the contextual information in the video using task-oriented training. In addition, we incorporate an auxiliary task that employs a contrastive loss function to enhance the model's generalization capability and foster a deeper understanding of the inter-modal relationships and underlying semantics. The task involves comparing the multimodal representation of the video-transcript with the caption representation, facilitating improved performance and knowledge transfer within the model. Finally, a transformer architecture is used to effectively capture and encode the interdependencies between the text and video information using attention mechanisms. During the decoding phase, the transformer allows the model to attend to relevant elements in the encoded features, effectively capturing long-range dependencies and ultimately generating semantically meaningful captions. The experimental evaluation, carried out on the MSRVTT benchmark, validates the proposed methodology, which achieves BLEU4, ROUGE, and METEOR scores of 0.4408, 0.6291 and 0.3082, respectively. When compared to the state-of-the-art methods, the proposed approach shows superior performance, with gains in performance ranging from 1.21% to 1.52% across the three metrics considered.

INDEX TERMS Multimodal video captioning, multimodal learning, cross correlation, transformers, contrastive learning.

I. INTRODUCTION

Video is a highly popular media, well-suited to capturing dynamic events and engaging both our visual and auditory senses. Today, thanks to the facility of acquisition and transmission, video content is omnipresent on various social media platforms. However, the vast amount of video data available on the Internet would be of limited use without appropriate tools that can make it possible to access/identify/retrieve it effectively. Within this framework, one important challenge is to describe and summarize the semantics of a given video in natural language.

The associate editor coordinating the review of this manuscript and approving it for publication was Zhaoqing Pan.

In this paper, we notably tackle the task of video captioning, which consists of automatically generating textual descriptions of the video content. Video captioning methodologies have the potential to facilitate the way we consume and interact with videos across a wide range of applications including accessibility, education, healthcare, and security. The numerous benefits of video captioning have led to a growing interest in the research community. This task is however highly challenging, as it requires jointly understanding the visual content, dialogue and actions within the video in order to generate grammatically correct and semantically meaningful sentences. To improve the accuracy of the video captioning process, multimodal learning techniques can be used to incorporate information from multiple modalities such as text, audio and video.

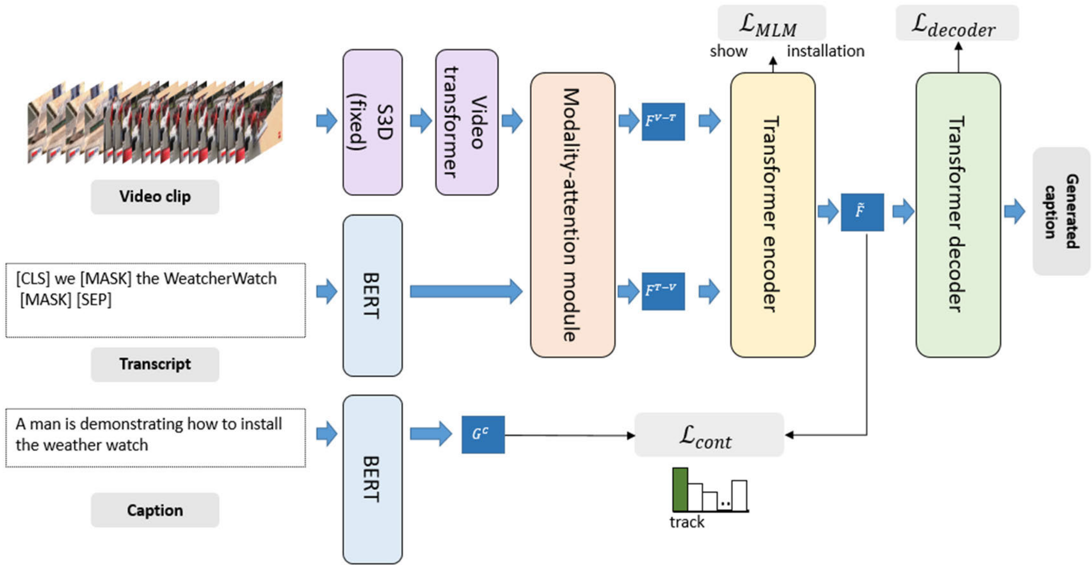


FIGURE 1. Overview of the proposed multi-modal architecture.

Most of the time, a video captioning model includes two main components: (1) an encoder to represent the input features and (2) a decoder whose task consists of generating captions, while ensuring the coherence of the sentence. Early deep-learning-based approaches use 2D Convolutional Neural Networks (CNNs) or 3D CNNs [1] at the encoder level, and Recurrent Neural Networks (RNNs) for the decoder [2], [3]. With the broad success of transformer architectures [4] in the Natural Language Processing (NLP) domain, recent years have featured a trend toward the use of attention mechanisms in both encoder and decoder. Most models [5], [6], [7] rely on large datasets to learn more discriminative multimodal features, with instructional datasets such as HowTo100M [10] and Cooking312k [11] being advantageous due to their aligned speech and transcripts available. Several approaches [5], [6], [12], [13] pre-train sequence-to-sequence models on such unlabeled data using denoising auto-encoders. In this case, the input is purposely modified by masking random words/frames to create artificial noise. This process enables the model to learn robust representations by reconstructing the original input from the corrupted version.

The human perceptual system is inherently multimodal, being capable of integrating information from various sensorial modalities, including vision, hearing, touch, taste, and smell [15]. Replicating this sophisticated system and efficiently harnessing the rich information present in videos remains an open issue of research. The question is how to encompass a diverse range of spatio-temporal elements, including temporally varying visual appearances, motion information, audio features, overlaid text and speech information.

In this work, we propose a novel architecture for multi-modal video captioning. Given an open-domain video and its associated transcript, the objective is to generate a meaningful video description, close to the human judgment and perception. The scope of the videos ranges from general content such as arbitrary YouTube videos or specific ones such as instructional videos with fine-grained activities.

The proposed framework is illustrated in Figure 1. It includes a modality-attention module that uses cross-correlation to simultaneously learn text-guided video features and video-guided textual features. Furthermore, we introduce an auxiliary task based on contrastive learning between video-transcript and caption embeddings, which improves the representation of inter-modal relationships (Figure 1). The objective of the contrastive loss function is to align the visual-transcript representation with the correct caption representation, while simultaneously distinguishing them from other captions within the batch. The alignment ensures that the model focuses on capturing relevant features by discerning meaningful multimodal interactions from irrelevant or unrelated information. Furthermore, the integration of a contrastive loss enhances the model's generalization capability while mitigating overfitting.

It is important to note that the caption information is strictly used during the training phase. It serves as a form of supervision to guide the model in learning grounded multimodal representations. During inference, the model just uses the knowledge and patterns learned during the training phase to generate the predicted caption.

In view of the success of attention mechanisms, we leverage the transformer architecture for both the encoder and decoder.

To summarize, the main contributions of the paper are the following:

1) A novel multimodal video captioning framework that incorporates various attention mechanisms to learn inter-modal and intra-modal representations.
2) A modality-attention strategy that employs cross-correlation to reduce the gap between the visual and textual modality
3) The use of a contrastive loss between the video-transcript multimodal representation and the caption embedding to improve the understanding of inter-modal relationships and underlying semantics.
4) A comprehensive experimental evaluation on the challenging MSRVTT dataset, which shows that the proposed method outperforms state-of-the-art approaches, reaching a 0.4408 BLEU4 score.
5) A fine-level investigation of the impact of each modality in the framework.

The rest of the paper is organized as follows. In Section II, we present the state-of-the-art video captioning methods. Two families of approaches are here identified, including video-based techniques that rely solely on visual cues and multimodal methods that integrate multiple modalities. Section III provides a comprehensive overview of the proposed methodology, detailing the key steps involved. In section IV, we provide the different training strategies used to optimize the model. Section V presents the experimental setup, dataset, and ablation studies conducted. Finally, Section VI concludes the paper by summarizing our main findings, and highlighting potential directions of future work.

II. RELATED WORK

Early video captioning methods [17], [18] follow rule-based methods. The principle here consists of detecting subjects, verbs, and objects (also known as SVO-triplets) from the videos, which are then combined into sentence templates.

More recently, video captioning has been reformulated as a machine translation task [19], [20], [21], leading to the development of the encoder-decoder paradigm that is commonly used today. Within this framework, the encoder processes a set of video features and accumulates its hidden states. The resulting output state is then passed to a decoder, which generates a natural language caption based on the encoded information. Such an approach makes it possible to model complex video features, and thus generate captions that are semantically more meaningful than those obtained by rule-based methods. Moreover, the encoder-decoder paradigm can be trained in an end-to-end fashion, allowing the simultaneous optimization of both encoder and decoder. This leads to improved performance when applied to the video captioning task. We can identify two families of approaches that exploit the encoder-decoder paradigm: the visual-based techniques and the multimodal methods.

A. VISUAL-BASED APPROACHES

Visual-based approaches in video captioning focus primarily on extracting relevant visual information from video frames. Such approaches leverage computer vision techniques to analyze the visual content of the video and identify important elements such as objects, scenes and actions, together with their corresponding spatial and temporal relationships. In early works, the visual encoder is implemented as a 2D CNN applied to video frames. Thus, Venugopalan et al. [22] propose a framework where CNN features from each frame are averaged and provided as input to the decoder at every time step. Zhang et al. [23] introduce the GMNet model, incorporating a guidance module within the encoder-decoder model for video caption generation. GMNet facilitates word generation by considering both preceding and subsequent words in the caption. The model utilizes a soft attention mechanism and leverages InceptionV4 [24] to extract semantic features from the video.

To capture temporal dynamics within the video, the 2D-CNN architecture has been later extended to 3D-CNN [1]. Xu et al. [25] introduce a two-module model for video captioning. A proposal module extracts features using 3D convolutional layers (C3D), while a so-called segment proposal network (SPN) is used for obtaining temporal segments. The model maps the visual representation onto a common vector space, while the syntactic representation relies on the Part-of-Speech (POS) tagging structures of the video description.

Hemalatha and Sekhar [28] introduce a video captioning approach that incorporates domain-specific decoders through the use of a domain classifier. The model utilizes ResNet152 for extracting 2D-CNN features and a 3D-CNN for extracting temporal features. To obtain a video representation, both the 2D-CNN and 3D-CNN features are aggregated using VLAD [29].

For sentence generation, many existing approaches rely on recurrent neural networks (RNNs) such as LSTM [26] and GRU [27] to generate the caption. Yao et al. [30], Donahue et al. [31] and Venugopalan et al. [32] use the LSTM architecture for yielding variable-length video descriptions. Guo et al. [33] further incorporate attention mechanisms within the LSTM model to refine the captions. Similarly, Zhang et al. [34] introduce a hierarchical decoder with temporal or spatial attention. The model implements a teacher-recommended learning system to leverage external language models and incorporate linguistic information.

Overall, visual-based approaches primarily focus on leveraging visual cues to generate accurate and descriptive captions. They are particularly effective in scenarios where the video content is predominantly visual and lacks significant audio or textual cues. However, in many applications videos contain multiple modalities such as visual, audio, and textual information (e.g., subtitles). Such modalities contribute to the overall meaning of the video and need to be jointly considered to generate meaningful captions [35].

The multimodal approaches have gained popularity in the video captioning task as they provide a more comprehensive understanding of the video.

B. MULTIMODAL APPROACHES

Currently, various methods adopt multimodal learning in video captioning tasks. Hessel et al. [35] use both automatic speech recognition (ASR) and video features to perform video captioning and claim that most of the enhancement in performance is attributable to the use of ASR. Similarly, Shi et al. [36] train their video captioning model on both visual and ASR inputs and demonstrate the benefits of adding textual input to the overall understanding of the video. Inspired by such results, we also consider in our work both visual and textual modalities.

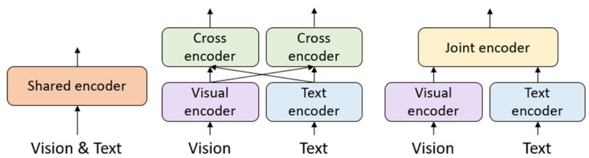
However, multimodal video captioning also presents several challenges. One major challenge concerns the alignment between different modalities, as the content of the visual and textual channels may not always be perfectly synchronized [10], [37], [9]. Furthermore, the size and complexity of multimodal datasets can raise challenges for training models that are both accurate and efficient [38], [39]. To tackle such issues, several works [9], [11], [13], [35] use instructional videos [10], [37], where the synchronization between visual content and subtitles is more favorable for video captioning task. While such videos are useful for training, they have a specific structure that may not be representative of real-world scenarios [14]. This makes it difficult to generalize the model on unseen data. Furthermore, real-world speech tends to be less structured, with key actions or events in the video not always corresponding to the same segments in the input transcript. To address the visually misaligned narrations, various approaches have employed contrastive learning between video and transcript. For instance, MIL-NCE [9] leverages weak and noisy training signals in instructional videos by combining multiple instance learning with contrastive learning. Meanwhile, VideoCLIP [40] constructs temporally overlapped pairs of video and text clips of varying lengths, aiming to enhance the quality and quantity of the pre-training dataset.

Traditionally, most existing methods have applied the contrastive loss to the outputs of visual and text encoders, typically before the multimodal fusion stage [41]. The primary aim of this loss is to establish alignment between the video and transcript during the pre-training phase. In our approach, we tackle the alignment challenge differently by incorporating the modality attention module. This module is specifically designed to bridge the gap between video and text modalities before feeding them into a transformer encoder. Utilizing cross-correlation, the modality attention module generates text-conditioned visual features and video-conditioned textual features, facilitating more effective alignment. Our model is extensively evaluated on a diverse range of YouTube videos using the MSRVTT dataset. Conversely, the contrastive loss serves the purpose of aligning the multimodal representation of the input with its corresponding caption. In contrast to previous state-of-the-art models, we apply this loss to the output of the multimodal transformer.

1) ARCHITECTURE

In view of the success of transformers in several domains, recent methods use this architecture at both encoder and decoder levels. Under this framework, the state of the art encoder architectures can be classified into three main families.

The *share-type* paradigm, illustrated in Figure 2(a), includes Unicoder-VL [21], VL-BERT [43], UNITER [44], VideoBERT [11], and VideoAsMT [12]. In this case, the textual and visual modalities are fed into a single encoder that generates a unified representation. While computationally efficient, this approach suffers from modality entanglement due to the vast differences among various modalities [45]. This challenge stems from the fact that several modalities may interfere, particularly when there are numerous modalities and tasks involved [46]. It is challenging for a foundational model with a single module to find a good balance between the advantages of modality collaboration and the impact of modality entanglement across various modalities.



(a) Share-type (b) Cross-type (c) Joint-type

FIGURE 2. The three main paradigms of video-text training.

The *cross-type* paradigm, illustrated in Figure 2(b), includes models like ViLBERT [47] and LXMERT [48]. Within this framework, multiple separate encoders are used to accommodate the different interactions between modalities. In contrast to the single-stream input in the share-type, the two-stream input allows for interactions between different modalities at various representation depths. At the same time, the cross-type approach can be more computationally demanding due to the use of several cross-encoders.

Finally, the *joint-type* paradigm, illustrated in Figure 2(c), is used by models such as SwinBert [7], UniVL [6], MV-GPT [5] and GIT [8]. This paradigm utilizes a two-stream input, similar to the cross-type architecture, allowing for effective capture of intra-modal features. However, in contrast to the cross-type, the joint-type architecture incorporates a single encoder to capture inter-modal dependencies. This approach offers a good trade-off between computational efficiency and the capacity to capture modality-specific features and interactions. For this reason, in our work, we have adopted the joint-type encoding paradigm.

Concerning the decoder, several architectures can be considered. One common approach is to use RNNs, which generate the caption word by word, in a sequential manner. This method has the advantage of being able to capture long-term dependencies between words, but can suffer from slow convergence and difficulty in modeling complex relationships between visual and textual components [4]. More recently, several studies [5], [6], [7], [11] have explored transformer-based models for video captioning, which show promising results due to their ability to capture long-range dependencies and relationships between different modalities. We follow this line of work and use the attention mechanism to sequentially generate the caption. We use both the encoder's hidden states and the previously generated words in the caption as supervisory signals for the attention mechanism.

2) TRAINING STRATEGIES

In recent years, vision-language pre-training has gained considerable popularity within the research community [45], [53], [54], [55]. This approach involves an initial phase where multimodal models are pre-trained on extensive datasets in an unsupervised manner, followed by subsequent fine-tuning for specific downstream tasks (e.g. video captioning, action recognition, video question answering). Typically, the considered datasets comprise videos along with their associated transcripts, a resource that is abundantly available. These methods learn multimodal representations by formulating proxy tasks such as masked language modeling [5], [6], or vision-language matching [45], [53].

The paradigm of pre-training followed by fine-tuning for multimodal models is undeniably effective and has yielded remarkable results across various applications [41], [56]. However, it is essential to acknowledge that this approach comes with substantial resource requirements, primarily in terms of hardware, rendering it unfeasible for small-scale setups. This is particularly the case when considering multimodal models with billions of parameters, such as the GIT model [8], which has over 5 billion parameters and is pre-trained on 10.5 billion samples. Additional statistics for similar models can be found in Section V-D, TABLE 3. The resource-intensive demands penalize the adoption and deployment of such approaches in the case of applications where the computational ressources are limited/constrained, most often for economical reasons.

Within this context, let us note that pre-training undoubtedly enhances the model's performances. Thus, comparing pre-trained models with models learnt from scratch is not entirely equitable. In our case, due to hardware constraints, we opt for an alternative strategy by forgoing pre-training altogether. Despite this, we demonstrate that competitive results can still be achieved. The proposed approach leverages the available resources efficiently, focusing on task-specific training without the need for massive pre-training datasets or extensive computational power. This resource-aware approach not only makes multimodal modeling accessible to a wider range of users and applications but also highlights the potential for effective multimodal model development in resource-constrained environments.

III. MODEL ARCHITECTURE

Figure 1 illustrates the synoptic scheme of the proposed approach, which comprises three fundamental elements: (1) the modality attention module, (2) the joint encoder and (3) the decoder. As a preprocessing step, we start by extracting the visual and textual embeddings.

A. FEATURE EXTRACTION

The feature extraction process concerns the two components involved, which are the visual and textual (with both transcript and caption) data.

In order to acquire the visual representations, we utilize a uniform sampling approach to divide the video into N fixed-length, non-overlapping clips of 16 frames each. The clips are then processed with the help of the S3D network [1], which is designed to learn robust video representations. Prior to use, the S3D model has been pre-trained on HowTo100M [10] with the MIL-NCE technique [9]. This technique is widely adopted in multimodal learning [6], [40], [57] for its robust handling of diverse and noisy data sources. The feature activations before the final fully connected layer are extracted and we apply average pooling to generate a $d_v = 1024$-dimensional vector (the v subscript stands here for visual). Subsequently, a feed forward network that includes a linear projection, followed by the GeLU [49] activation function and layer normalization, is used to yield the final feature vector (of the same size d_v). The resulting visual features are represented as a $N \times d_v$ matrix, denoted by V. Let us underline that the S3D model is used only as a backbone for feature extraction and its weights are frozen.

A video transformer is further employed to effectively capture the dependencies between frames in video clips and learn the inherent temporal dynamics of video objects, actions, and scenes. This approach enables us to learn grounded visual features that are specifically optimized for the task of video captioning, without being restricted to pre-extracted features from external models. In addition, using a pre-extracted feature-based model with a transformer architecture can significantly reduce the computational cost of training, as the S3D model can be pre-trained on large-scale video datasets and the transformer can be fine-tuned on a smaller dataset dedicated to video captioning.

To take into account the dynamic dependencies between clips, we employ temporal attention on the feature vector V. Our approach is motivated by the observation that video data often contains redundant information, and only a limited number of clips contain discriminative information that is relevant to the video captioning task. For this reason, a multi-head temporal attention mechanism is applied on the visual descriptor V. For each attention head $h \in \{1, \ldots, H^v\}$ (where H^v denotes the number of visual attention heads), we first compute the associated $Query_h^V, Key_h^V$ *and* $Value_h^V$ components defined as:

$$Query_h^v = VW_{query,h}^v; \quad Key_h^v = VW_{key,h}^v \tag{1}$$

$$Value_h^v = VW_{value,h}^v \tag{2}$$

where $W^{v}_{query,h}$, $W^{v}_{key,h}$ and $W^{v}_{value,h}$ are three learnable matrices of size $\left(d_v \times \frac{d_v}{H_v}\right)$.

The visual temporal attention for a given attention head is computed as:

$$Attention^{v}_{h} = softmax(\frac{Query^{v}_{h}(Key^{v}_{h})^{T}}{\sqrt{d_v/H_v}})Value^{v}_{h} \tag{3}$$

where superscript T denotes the matrix transpose operator.

The attention heads are then concatenated under the form of a $(N \times d_v)$ matrix, denoted by $Attention^{v}$ and globally gathering the visual representation. An additional projection is considered in order to obtain the final visual representation, denoted by F^{v} and defined as:

$$F^{v} = Attention^{v} W^{v}_{convert} \tag{4}$$

where $W^{v}_{convert}$ is a learnable matrix of size $(d_v \times d)$. This final operation performs the dimensionality conversion of the visual feature to a common dimension d that will also be used for the textual representation.

Concerning the textual data, we consider the audio transcript (if the audio channel includes speech) as well as the video captions.

To obtain the audio transcript from the input video, we utilize the Whisper model [50], which is an ASR algorithm that exhibits human-level robustness in English speech recognition, even in the presence of background noise and reverberation.

Whatever the source (audio transcript or caption), the textual data undergo a tokenization process using WordPieces [51], which segments the text into sub-words using a vocabulary of $S_{voc} = 30{,}000$ tokens. The tokenized sequences are fed into the BERT-based uncased model [52], following previous state-of-the-art methods [56], to perform the embedding. As recommended in [49], the first token in the input sequence is represented as a dedicated [CLS] token, and the final one is represented by a so-called [SEP] token. To achieve equal length for all the tokenized text sequences, we expand the sentence using padding, with the help of a dedicated [PAD] token. Let us denote by M the length of the padded tokenized sequences, which correspond to the maximal number of tokens that are allowed to appear in a given sentence. Let us also mention that a random masking of the tokens can also be considered. In this case, the input token is replaced by a dedicated token, denoted by [MASK].

The BERT approach also employs a self-attention mechanism, yielding in output a $(M \times d_{BERT})$ feature matrix, with $d_{BERT} = 768$, corresponding to the activations of the last BERT layer.

The text embedding approach is applied to both transcript and caption data. Similarly to the visual component, the transcript feature matrix is finally converted into a $(M \times d)$ matrix denoted by F^{t}, with d being the common dimension considered also for the visual representation. The caption feature matrix, denoted by F^{c}, does not require projection onto a space of common dimension (see its utilization in section IV) and thus remains of size $(M \times d_{BERT})$.

Let us finally note that the BERT encoder is fine-tuned separately for the transcript and the caption data.

B. MODALITY ATTENTION MODULE

Modeling visual and textual dynamics within and across modalities is a highly intricate task. To overcome such a challenge, we have developed a modality-attention module (Figure 3) that effectively captures both intra and inter-modal relationships between the visual and audio transcript modalities. It is designed to bridge the gap between features F^{v} and F^{t}, which are generated from separate models trained on different tasks.

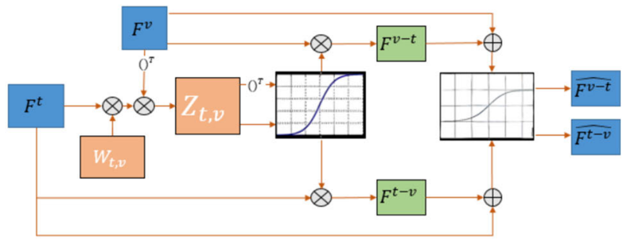


FIGURE 3. The modality attention module.

As we use real-life videos as input, in a majority of cases the feature vectors are not well-aligned. Figure 4 illustrates some video examples with their respective transcripts. We observe that people tend to speak in a disorganized manner, and the key actions or events in the video do not necessarily correspond to the same segment of the input text.

The objective is to create an embedding space that makes semantically related visual-textual pairs of features appear closer together than unrelated pairs. This will enhance the alignment between F^{v} and F^{t}, and enable better modeling of the interactions between visual and audio transcript data. To this purpose, we consider the cross-correlation matrix $Z_{t,v}$, defined as:

$$Z_{t,v} = F^{t} W_{t,v} {F^{v}}^{T} \tag{5}$$

where $W_{t,v}$ is a $(d \times d)$ learnable matrix and T denotes the transpose operator.

A high coefficient in the correlation matrix $Z_{t,v}$ indicates a strong relationship between the corresponding visual and textual features. To create cross-correlation visual-transcript (resp. transcript-visual) weights, we apply the column-wise softmax operator over $Z_{t,v}$ (resp. $Z^{T}_{t,v}$), as described in the equations (6) and (7):

$$F^{t-v} = F^{t} softmax(Z_{t,v}) \tag{6}$$
$$F^{v-t} = F^{v} softmax(Z^{T}_{t,v}) \tag{7}$$

This approach enables us to develop more distinctive and mutually constrained modality representations.

To avoid information loss during the cross-correlation phase, we have considered a dense skip connection technique. This means that we add the reweighted features F^{t-v} and

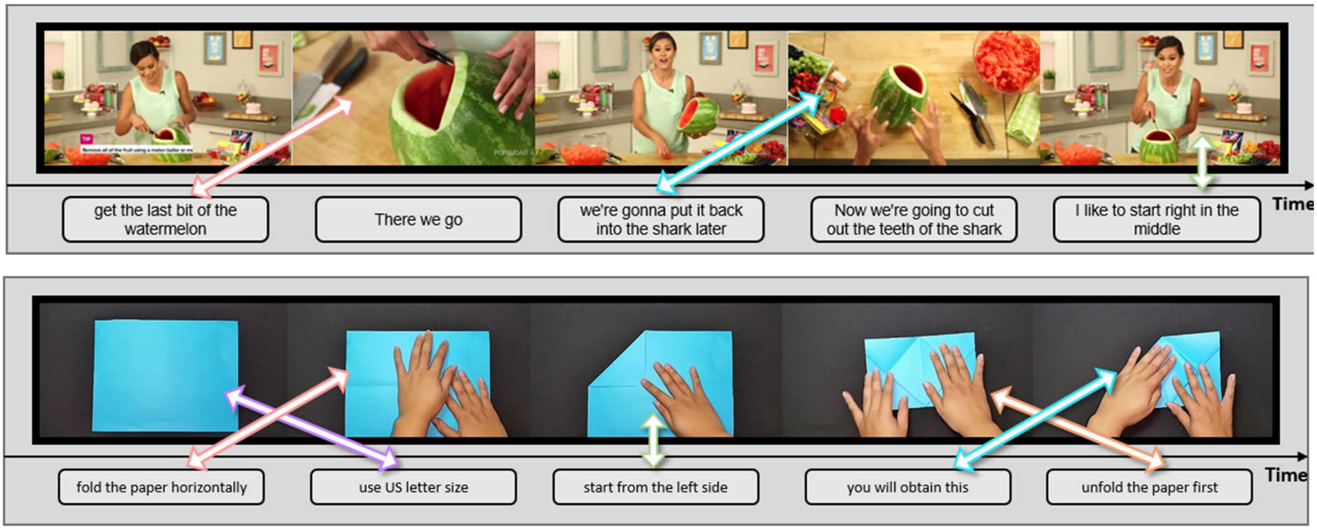


FIGURE 4. Video samples from the MSRVTT dataset for which the transcript and video data are not well-aligned.

F^{v-t} to the original representation of each modality, and regularize the result with the help of a *tanh* function:

$$\hat{F}^{t-v} = tanh(F^{t-v} + F^{t}) \tag{8}$$
$$\hat{F}^{v-t} = tanh(F^{v-t} + F^{v}) \tag{9}$$

The modality attention module addresses the alignment issue between modalities. In (5), the cross-correlation matrix encodes the relationships between video and text features learned by the model through the trainable parameter $W_{t,v}$. Applying softmax to the matrix $Z_{t,v}$ enhances the discriminative power of the features. The model assigns higher weights to visual features when they exhibit strong correlations with textual features, and vice versa. This process potentially improves alignment between modalities. Specifically, it makes it possible to capture and emphasize the most salient correspondences between textual and visual elements. The resulting outcome, seen in (6) and (7), is used to reweight the input features based on their correlation with the other modality. Finally, the skip connection technique in (8) and (9) enforces the preservation of modality-specific information while adding non-linearity to the model.

The obtained features are further exploited in the joint transformer encoder, as described in the following section.

C. TRANSFORMER ENCODER

In order to make the video and text fully interact, we design a transformer-based encoder. The transcript and visual features are first concatenated into a single global descriptor $F = [\hat{F}^{t-v} \mid \hat{F}^{v-t}]$, which is a matrix of size $(M + N) \times d$. The transformer architecture does not include any recurrent connections, which means that the order of the input tokens (or of video clips for the visual component) is lost during the process. To overcome this limitation, a position embedding technique is integrated. It consists of a trainable look-up table, where the embedding of each position in the input sequence is learned during training. To this purpose, we have followed the approach suggested in [52], described in the following equation:

$$E_{pos} = W_{pos}(pos_0, \ldots, pos_{M+N}) \tag{10}$$

where W_{pos} of size $(M + N) \times d$ is a lookup table, mapping the position index of each token pos_i onto its corresponding vector representation.

In addition, a modality embedding is integrated, in order to differentiate between the visual and textual modalities:

$$E_{mod} = W_{mod}(\underbrace{0, \ldots, 0}_{M}, \underbrace{1, \ldots, 1}_{N}) \tag{11}$$

where W_{mod} of size $2 \times d$ is a lookup table, mapping the type of each modality (text: 0; video: 1) onto a vector representation.

The input to the encoder is defined as the sum of all these three features:

$$\mathcal{F}_0 = F + E_{pos} + E_{mod} \tag{12}$$

Our encoder comprises a number of L^{enc} self-attention layers. Each layer l consists of Multi-head Self-Attention (MSA), layer normalization (LN) and Feed Forward Network (FFN). The considered layers, for $l \in \{0, 1, \ldots, L^{enc} - 1\}$, are recursively computed as illustrated in Figure 5 and as described formally in the following equations:

$$\mathcal{F}_l' = MSA\left(LN\left(\mathcal{F}_{l-1}\right)\right) + \mathcal{F}_{l-1} \tag{13}$$
$$\mathcal{F}_l = FFN\left(LN\left(\mathcal{F}_l'\right)\right) + \mathcal{F}_l' \tag{14}$$

The FFN consists of two linear projections separated by a GELU non-linearity [49].

To enhance the model performance, we employ a multi-head attention mechanism, which splits the input into H^{enc} heads, allowing the model to attend to diverse parts of the input simultaneously. For each attention head

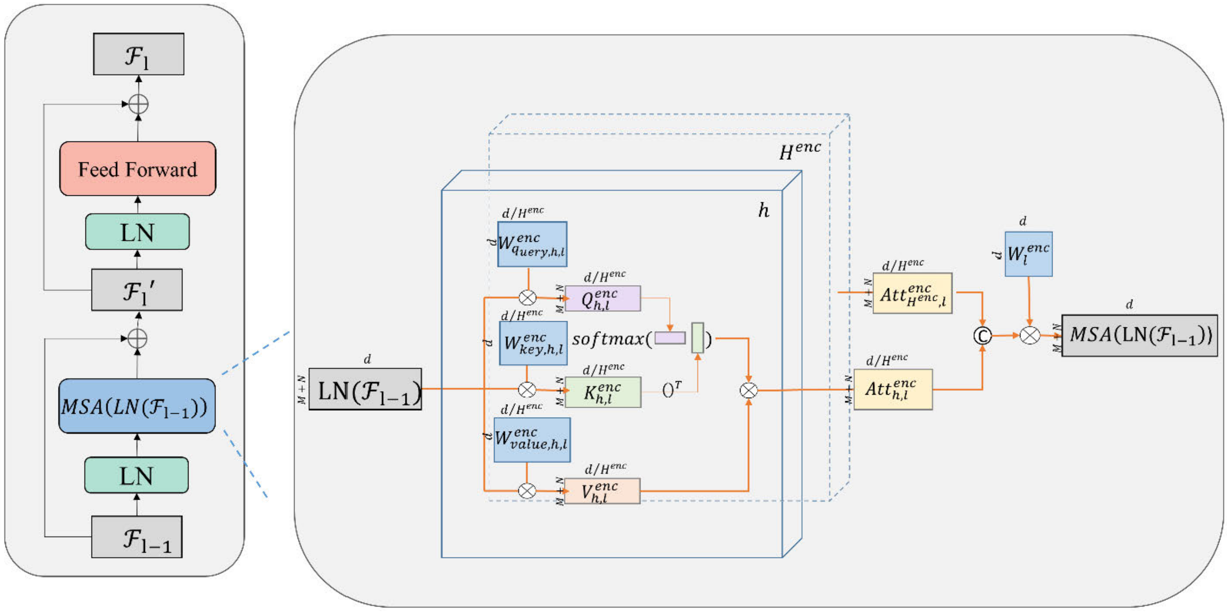


FIGURE 5. Overview of the encoder architecture. (left) Encoder block. (right) multi-head self-attention mechanism.

$h \in \{1, \ldots, H^{enc}\}$, we compute the attention sub-layers of the encoder as follows:

$$Att^{enc}_{h,l}\left(Q^{enc}_{h,l}, K^{enc}_{h,l}, V^{enc}_{h,l}\right) = softmax\left(\frac{Q^{enc}_{h,l} K^{encT}_{h,l}}{\sqrt{d}/H^{enc}}\right) V^{enc}_{h,l} \quad (15)$$

Here, the queries $Q^{enc}_{h,l} = LN(\mathcal{F}_l) W^{enc}_{query,h,l}$, keys $K^{enc}_{h,l} = LN(\mathcal{F}_l) W^{enc}_{key,h,l}$, and values $V^{enc}_{h,l} = LN(\mathcal{F}_l) W^{enc}_{value,h,l}$ represent linear projections of the multimodal input $\mathcal{F}_l$ and d/H^{enc} is a scaling factor used to address the vanishing gradient issue.

Finally, the MSA is computed in (16) as follows:

$$MSA(LN(\mathcal{F}_l)) = Concat(Att^{enc}_{1,l}, \ldots, Att^{enc}_{H^{enc},l}) W^{enc}_l \quad (16)$$

where W^{enc}_l represents the learnable linear projection matrix.

The outputs of the various heads are concatenated and passed through a linear layer to obtain the final output $\mathcal{F}^{enc} = \mathcal{F}^{enc}_{L^{enc}-1}$ of size $(M+N) \times d$.

D. TRANSFORMER DECODER

The objective of the decoder is to generate a caption $C = C(x_v, x_t)$ given the input video x_v and transcript x_t by maximizing the conditional probability $p(C|x_v, x_t)$. The caption C is represented as an ordered sequence of tokens $C = \left(c_1, c_2, \ldots, c_{L_C}\right)$. The joint probability can be recursively decomposed as follows:

$$\begin{aligned} &p\left(C \mid x_v, x_t\right) \\ &\quad = p\left(c_1 \mid x_v, x_t\right) \times p\left(c_2 \mid c_1, x_v, x_t\right) \times \cdots \\ &\qquad \times p\left(c_{L_C} \mid c_{L_C-1}, \ldots, c_1, x_v, x_t\right) \end{aligned} \quad (17)$$

During training, the decoder generates one token at a time, conditionally to the previously generated tokens. However, by adopting such an approach, the errors can propagate and accumulate over time. In order to overcome this difficulty, we use the teacher-forcing technique [53], where the ground truth caption is forced to be provided until a certain token, selected in a random manner. Solely beyond this token, the model is allowed to generate its own ones. This technique stabilizes the training and limits the propagation of errors notably made in the early stages of the decoding process.

Formally, let $y^{C,n} = \left(t^{C,1}, \ldots, t^{C,n}\right)$ denote the sequence of decoded tokens up to token n. This sequence is iteratively providing new inputs $Y^{C,n}$ to the decoder, as described in the following equation:

$$\begin{aligned} &\forall n \in \{1, 2, \ldots, L_C\}, \\ &Y^{C,n} = dp(LN(emb(pad(y^{C,n})) + E^C_{pos}) \end{aligned} \quad (18)$$

where dp is the dropout layer, LN is the layer normalization, pad is the padding operator necessary to complete the $y^{C,n}$ sequence up to length L_C, emb is the embedding layer and E^C_{pos} is the positional embedding of the caption.

The transformer decoder consists in L^{dec} identical layers. Each layer l includes of a Masked-Multi-head Attention (MMA), layer normalization (LN), Multi-head Cross-Attention (MCA) and a Feed Forward Network.

The first layer is initialized as:

$$Y^c_0 = \left(Y^{c,1}, \ldots, Y^{c,L_C}\right) \quad (19)$$

The subsequent layers, for $l \in \{1, \ldots, L^{dec}-1\}$, are recursively computed as illustrated in Figure 5 and as described formally in the following equations:

$$Y'^c_l = MMA\left(LN\left(Y^c_{l-1}\right)\right) + Y^c_{l-1} \quad (20)$$

$$Y''^c_l = MCA\left(LN\left(Y'^c_l\right), LN(\mathcal{F}^{enc})\right) + Y'^c_l \quad (21)$$

$$Y^c_l = FFN\left(LN\left(Y''^c_l\right)\right) + Y''^c_l \quad (22)$$

The Masked Multi-head Attention mechanism represents a modification of the self-attention mechanism, consisting in a masking procedure whose goal is to prevent the decoder from attending future positions during training. This ensures the autoregressive property of the decoder, which is forced to get access solely to the tokens that precede the current position. The masking is achieved by setting the attention scores of future positions to a very large negative value. This ensures that the softmax operation applied to the attention scores assigns a probability close to zero to the future positions, thus effectively blocking their influence on the current position's representation. Formally, for each masked attention head $h \in \{1, \ldots, H^{dec}\}$ and for each layer l, we compute the masked attention ($MAtt$) as:

$$MAtt_{h,l}^{dec}\left(Q_{h,l}^{MMA}, K_{h,l}^{MMA}, V_{h,l}^{MMA}\right) = softmax(\frac{Q_{h,l}^{MMA} {K_{h,l}^{MMA}}^T}{\sqrt{d}/H^{dec}} + \Lambda)V_{h,l}^{MMA} \quad (23)$$

where the queries $Q_{h,l}^{MMA} = LN(\mathrm{Y}_l^c)W_{query,h,l}^{MMA}$, the keys $K_{h,l}^{MMA} = \mathrm{LN}(\mathrm{Y}_l^c)W_{key,h,l}^{MMA}$, and the values $V_{h,l}^{MMA} = LN(Y_l^c)W_{value,h,l}^{MMA}$ represent linear projections of the decoder input Y_l^c. Here, Λ is the masking matrix of size $L_C \times L_C$. It is constructed such that the upper triangular portion (including the main diagonal) is filled with negative infinity values, and the lower triangular portion is filled with zeros.

We employ multi-head masked attention, and we concatenate the outputs of different heads as follows:

$$MMA(LN(\mathrm{Y}_l^c)) = Concat(MAtt_{1,l}^{dec}, \ldots, MAtt_{H^{dec},l}^{dec})W_l^{MMA} \quad (24)$$

where W_l^{MMA} represents the learnable linear projection matrix.

Let us underline that during inference, the masked multi-head-attention is similar to the self-attention as the model does not have access to future positions.

The second attention sub-layer is a Multi-Head Cross Attention (MCA), illustrated in Figure 7 and computed as follows:

$$Q_{h,l}^{MCA} = LN(\mathrm{Y}_l'^c)W_{query,h,l}^{MCA}; \quad (25)$$

$$K_{h,l}^{MCA} = LN(\mathcal{F}^{enc})W_{key,h,l}^{MCA} \quad (26)$$

$$V_{h,l}^{MCA} = LN(\mathcal{F}^{enc})W_{value,h,l}^{MCA} \quad (27)$$

where $W_{query,h,l}^{MCA}$, $W_{key,h,l}^{MCA}$, $W_{value,h,l}^{MCA}$ of size $d \times d/H^{dec}$ are learnable matrices. The cross attention $CAtt_{h,l}^{dec}$ is computed as follows:

$$CAtt_{h,l}^{dec}\left(Q_{h,l}^{MCA}, K_{h,l}^{MCA}, V_{h,l}^{MCA}\right) = softmax(\frac{Q_{h,l}^{MCA} {K_{h,l}^{MCA}}^T}{\sqrt{d}/\mathrm{H}^{dec}})V_{h,l}^{MCA} \quad (28)$$

The outputs of the different cross attention heads are then concatenated and projected using a learnable matrix W_l^{MCA}

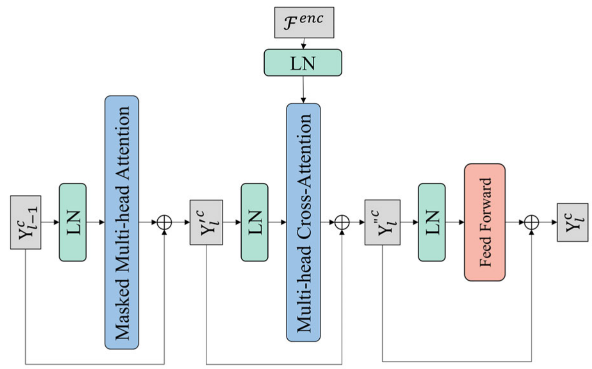


FIGURE 6. Overview of the decoder architecture.

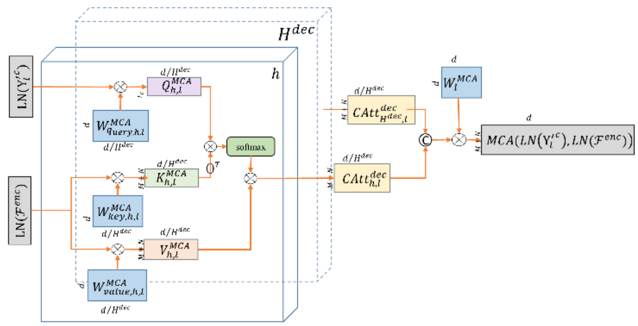


FIGURE 7. Multi-head cross attention process.

as follows:

$$MCA(LN\left(\mathrm{Y}_l'^c\right), LN(\mathcal{F}^{enc})) = Concat(CAtt_{1,l}^{dec}, \ldots, CAtt_{H^{dec},l}^{dec})W_l^{MCA} \quad (29)$$

The output of the final layer $Y_{L^{dec}}^c$ is used to determine the decoded token n as follows:

$$t^{C,n} = argmax(softmax(Y_{L^{dec}}^c W^{dec}) \quad (30)$$

using the learnable matrix W^{dec} and use the softmax function to compute the probability of the token.

During inference, the model does not have access to the ground truth. Using the predicted output from the previous time step can lead to a compounding error problem, where even small errors in the prediction can accumulate and result in poor performance. Therefore, we use the beam search decoding strategy to mitigate this problem. It is a heuristic algorithm that generates output sequences by keeping only the K most probable candidates at each step. Formally, at each time step n the decoder computes the probability distribution over the entire vocabulary for the next token as $p(c_n \mid c_{n-1}, \ldots, c_1, x_v, x_t)$. Then we select the K candidates with the highest probabilities. For each candidate, the process is continued until an end token is generated or the maximum length is reached. Among all the generated candidates, the caption with the highest global probability is selected as output.

IV. TRAINING OBJECTIVES

Three training objectives are considered to optimize the model: (1) masked language modeling, (2) contrastive learning and (3) caption generation.

A. MASKED LANGUAGE MODELING

Similarly to BERT, we also randomly replace 15% of the tokens in the sentence with the special token [MASK] and then generate the masked tokens given the known tokens and video input. The Masked Language Modeling (MLM) loss function is defined as the cross-entropy loss between the predicted probability distribution over the vocabulary and the true distribution for each masked token as seen in (31):

$$\mathcal{L}_{MLM} = -\sum_{i=1}^{S_{mask}}\sum_{j=1}^{S_{voc}} y_{ij}\log(p_{ij}) \tag{31}$$

Here, S_{mask} is the number of masked tokens, S_{voc} is the size of the vocabulary, y_{ij} is the true probability of the j-th token for the i-th masked position and p_{ij} is the predicted probability of the j-th token for the i-th masked position.

B. CONTRASTIVE LEARNING

Our goal is to create a system that can match a video x_v and transcript x_t to their correct caption C by calculating the dot product of their embeddings. We want to assign to incorrect captions a large distance, meaning that the dot product between their corresponding embeddings should be small.

Formally, we start by extracting the multimodal representation of the video, with visual and transcript components. As suggested in [52], we consider as a global representation of the multimodal input the embedding $\mathcal{F}^{enc}_{[CLS]}$ of the [CLS] token, which appears on the first position of the feature matrix $\mathcal{F}^{enc} = \{\mathcal{F}^{enc}_1 = \mathcal{F}^{enc}_{[CLS]}, \mathcal{F}^{enc}_2, \ldots, \mathcal{F}^{enc}_{M+N}\}$. The global video-transcript representation is then computed as:

$$\mathcal{F}_{\text{global}} = dp(\mathcal{F}_{[CLS]}W_{global} + b_{global}) \tag{32}$$

where W_{global} of size $d \times d$ and b_{global} of size d are learned during training. We denote by $f(x_v, x_t)$ the function that associates a pair of video x_v and transcript x_t to their global representation $\mathcal{F}_{\text{global}}$.

Similarly, we extract the global representation of the caption embedding F^C_{CLS} (*cf.* Section A) and project it as follows:

$$F^c_{global} = dp\left(F^C_{CLS}W^c_{global} + b^c_{global}\right) \tag{33}$$

where matrix W^c_{global} of size $d_{BERT} \times d$ and vector b^c_{global} of size d are learned during training. Let us denote by $g(C)$ the function that associates the caption C to its global representation F^c_{global}.

The contrastive loss is then computed as:

$$\mathcal{L}_{Cont} = \max_{f,g}\sum_{i=1}^{batch_{size}} \log \times \left(\frac{e^{f(x_{v_i},x_{t_i})^T.g(c_i)}}{e^{f(x_{v_i},x_{t_i})^T g(c_i)} + \sum_{(x_{v_j},x_{c_j},c_j)\in N_i} e^{f(v_j,t_j)^T.g(c_j)}}\right) \tag{34}$$

Here, given a positive triplet of index i in the batch (x_{v_i}, x_{t_i}, c_i) of (video, transcript, caption), we construct the negative set N_i of negative triplet by concatenating incorrect captions c_j within the training batch to the (video, transcript) pair (x_{v_i}, x_{t_i}) as (x_{v_i}, x_{t_i}, c_j) with $c_j \neq c_i$.

C. CAPTION GENERATION

The decoder loss measures the difference between the predicted and the ground truth captions using cross-entropy as follows:

$$\mathcal{L}_{\text{decoder}} = -\sum_{n=1}^{L_C} \log P\left(c_n \mid c_1, \ldots, c_{n-1}, x_t, x_v\right) \tag{35}$$

The final loss function considered for our model is simply defined as the sum of all these three components:

$$\mathcal{L}_{\text{model}} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{Cont}} + \mathcal{L}_{\text{decoder}} \tag{36}$$

V. EXPERIMENTS AND RESULTS

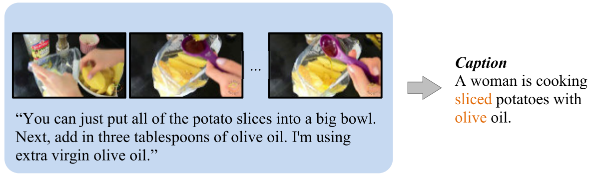
The experimental evaluation has been carried out on the publicly available MSRVTT dataset [16], described in the following section.

A. DATASET

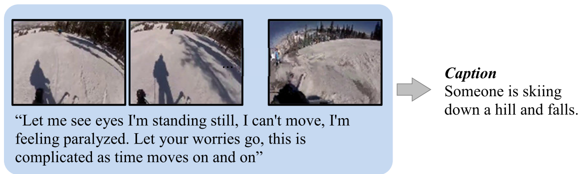
The MSRVTT (*Microsoft Research Video to Text*) dataset is widely used for benchmarking video captioning methods. It spans over 20 domains, including sports, news, education, and how-to videos. The dataset comprises 10,000 video clips, with an average length of 20 seconds, and 200,000 natural language descriptions, which have been collected from crowd-workers, ensuring diverse and human-like language expressions. The videos have been crawled from YouTube, contributed by internet users, and thus correspond to real-life situations.

The MSRVTT dataset raises several challenges, such as recognizing objects, actions, and scenes, as well as understanding the context and generating semantically meaningful captions. Additionally, it is worth noting that the MSRVTT dataset comprises videos with both visual and audio modalities, which adds an extra level of complexity to the task of generating captions. Nevertheless, around 20% of the videos in the dataset have no audio channel, while others have non-English audio, making the task even more challenging with sparse modalities.

In order to study the effect of each modality on the performance of the model, we have manually annotated two distinct subsets. The first subset, labeled as "vision and text" (534 samples), encompasses videos where both the visual features and transcript information contribute to the video captioning task. For example in Figure 8(a), the transcript helps identify specific ingredients such as oil type, difficult to



(a)



(b)

FIGURE 8. (a) Sample requiring both transcript and visual modalities for caption generation. (b) Sample requiring visual cues only.

infer solely from visual modality. The second subset, called "vision only" (663 samples), comprises videos where the task can be accomplished solely through visual cues. Some of these videos include silent or non-English speaking videos, where the transcript modality cannot be provided. Similarly, videos featuring sports or other activities that emphasize visual actions can be categorized in this subset. An example is illustrated in Figure 8(b), where the transcript represents the lyrics of music in the video and is not correlated with the caption. We study the performance of our model on these two subsets to better understand the role of the transcript information in video captioning.

B. IMPLEMENTATION DETAILS

In the pre-processing stage, the videos are divided into $N = 48$ uniformly sampled clips. The clips are then processed with the help of the S3D model. Next, the transformer encoder is applied, with 6 layers to capture the sequential information in the 3D feature. Each block consists of $H_v = 12$ attention heads and a hidden size of $d_v = 1024$.

Regarding the transcript, we utilize the Whisper ASR model to extract the speech from the video. Our initial findings indicate that the quality of the ASR model has a notable influence on the overall performance. We apply the Whisper model on the entire video rather than on individual clips, as people commonly mention key objects or actions before or after they are shown in the video (Figure 4). We set the maximum number of tokens in a given phrase to $M = 48$.

The model includes a 2-layer transformer encoder and a 3-layer transformer decoder, both consisting of 12 attention heads and a hidden size of $d = 768$. To accelerate the training process, we initialize the encoder and decoder weights with the pre-trained weights proposed by the model in [6]. The training process is conducted using 2 NVIDIA GeForce RTX 2080 GPUs over a period of 20 epochs, taking 4 days to complete. We use a linear learning rate schedule with a warm-up strategy, employing an initial learning rate of 1e-5. To overcome the limited GPU memory, we use the gradient accumulation technique [45] with 16 steps in conjunction with a batch size of 256. This technique effectively increases the batch size and allows us to update the model's parameters with fewer samples, without sacrificing the accuracy of the gradient estimation. The final model is selected according to the best performance obtained on the validation set.

C. ABLATION STUDY

We have conducted an ablation study in order to determine the significance of each component within the framework. The study compares several combinations to evaluate their relative performances.

The following methods are considered for evaluation:

1) **text only**, which used only text as input, trained with the transformer encoder and decoder.
2) **video only**, which used only video as input, also trained with the transformer encoder and decoder.
3) **video-text**, which used both video and text as input but did not employ the modality-attention module.
4) **MAM**, which adds the modality attention module (MAM) to the former.
5) **MAM + init**, which uses the initialization of the encoder and decoder weights from the model in [6].
6) **MAM + Cont**, which is trained with all objectives from scratch on MSRVTT, including the contrastive loss with the caption as input.
7) **MAM + Cont + init**, which initializes the encoder and decoder weights using those of [6] and includes both modality-attention and contrastive loss objective techniques. We denote this complete architecture by **CapVT**.

The following evaluation metrics are retained to evaluate the performance of the models: BLEU (1)-(4) [59], METEOR [60] and ROUGE [61]. BLEU evaluates the quality of generated text based on the n-gram (1 to 4) overlap with the reference text. ROUGE measures the overlap of n-grams and word sequences between the generated and reference captions. METEOR considers both n-gram overlap and semantic similarity between the generated and reference text. All scores range between 0 and 1, with higher values indicating better performances.

The results obtained, summarized in TABLE 1, demonstrate that the complete CapVT model (MAM + Cont + init) outperforms other models with a BLEU4 score of 0.4408, indicating the importance of our training choices.

Pre-training the model on external large datasets can be beneficial, but this process is often computationally expensive and requires significant hardware resources. To address this issue, we have used transfer-learning techniques to initialize the weights of our transformer encoder and decoder, which allowed us to leverage the knowledge learned from a larger dataset while reducing the computational load. This is observed with an improvement in performance of 1.2% and 2.17% for BLEU4 when comparing MAM to MAM + Init and MAM + Cont to CapVT, respectively. The study also highlights the importance of the contrastive loss objective

TABLE 1. Ablation study.

Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE
Text only	0.6682	0.4684	0.33	0.2299	0.2087	0.4850
Video only	0.7643	0.6225	0.4878	0.3666	0.2637	0.5825
Video-text	0.7723	0.6456	0.5073	0.3782	0.2674	0.5881
MAM	0.7752	0.6507	0.5250	0.3972	0.2754	0.5987
MAM+Init	0.7909	0.6612	0.5320	0.4104	0.2803	0.6001
MAM+Cont	0.7979	0.6676	0.5373	0.4191	0.2893	0.6087
MAM+Cont+Init	**0.8417**	**0.6784**	**0.5792**	**0.4408**	**0.3082**	**0.6291**

in improving the caption quality, as removing it leads to a significant drop in performance (2.19% in terms of BLEU4). Additionally, incorporating textual information is crucial for generating accurate captions, as evidenced by the lower score achieved when only the video modality is considered. Finally, the results indicate that the visual modality is more informative than the textual one as we achieve better results when feeding only the visual modality as compared to feeding only the textual modality.

As part of our study, we have also investigated how the quality of the generated captions is affected by different input modalities. For this purpose, we selected the first three baselines: text-only model, video-only model and video-text model. We have deliberately excluded the other models that employ additional strategies such as modality attention or contrastive loss. Our primary objective here is to solely examine the impact of the input modality on the model's performance.

We have assessed the performance of each baseline on videos that require only the visual modality to generate captions and those that require both visual and textual modalities. However, it was not feasible to label videos that require only textual modality in the MSRVTT dataset as certain information, such as key objects/persons can only be perceived through visual cues and not through text. We have randomly selected 1179 test samples and manually labeled them as either "vision-only" (663 samples) or "vision and text" (534 samples) to evaluate the performance of each baseline model on these different types of videos. TABLE 2 shows the performance comparison in terms of BLEU4 of the three models trained with different input modalities. The evaluation has been performed on the two subsets of samples that require either only visual modality or both visual and textual modalities to generate captions. The following conclusions can be drawn: (1) The effectiveness of video captioning models is heavily influenced by the input modalities. The different performances obtained on the two subsets underscore the significance of the dataset's modality composition. (2) The text-only model struggles to generate captions from visual content alone with a low score of 0.1671. (3) The video-only model performs well in a vision-only context, and may benefit from leveraging textual cues when available.

TABLE 2. Performance comparison (BLEU4) across models using different input modalities on two subsets.

Model	Vision-Only subset	Vision-text subset
Text-only model	0.1671	0.3247
Video-only model	0.3241	0.3051
Video-text model	0.3556	0.4192

Thus, adding the textual modality as input improves the performance with 11% on the vision-text subset. (4) The video-text model consistently outperforms the models relying on a single modality on the two subsets. This observation underscores the significance of multimodal approaches in video captioning. The ability to seamlessly integrate visual and textual information results in enhanced caption quality, making the model versatile and well-suited for real-world applications where both modalities are accessible.

D. COMPARISON WITH STATE OF THE ART

To facilitate a meaningful comparison between our work and previous state-of-the-art models in the context of video captioning, we have examined key statistics pertaining to these models. Specifically, we have compiled comprehensive data encompassing model size (number of parameters), the scale of pre-training samples, hardware infrastructure employed (GPU/TPU), and the training duration. The detailed findings of this analysis are presented in TABLE 3, drawing from information extracted from the respective authors' publications and the survey introduced in [41].

For a fair comparison, we have retained models that are comparable to ours, including OA-BTG [21], VideoAsMT [12], SwinBert [7], and OpenBook [3]. Additionally, we have retained the UniVL model [6], as we leveraged its weights to initialize the encoder and decoder parameters in our own approach. This approach aims to deliver a comprehensive and equitable evaluation of our method in relation to its peers, thus establishing a clear understanding of its performance within a defined resource context.

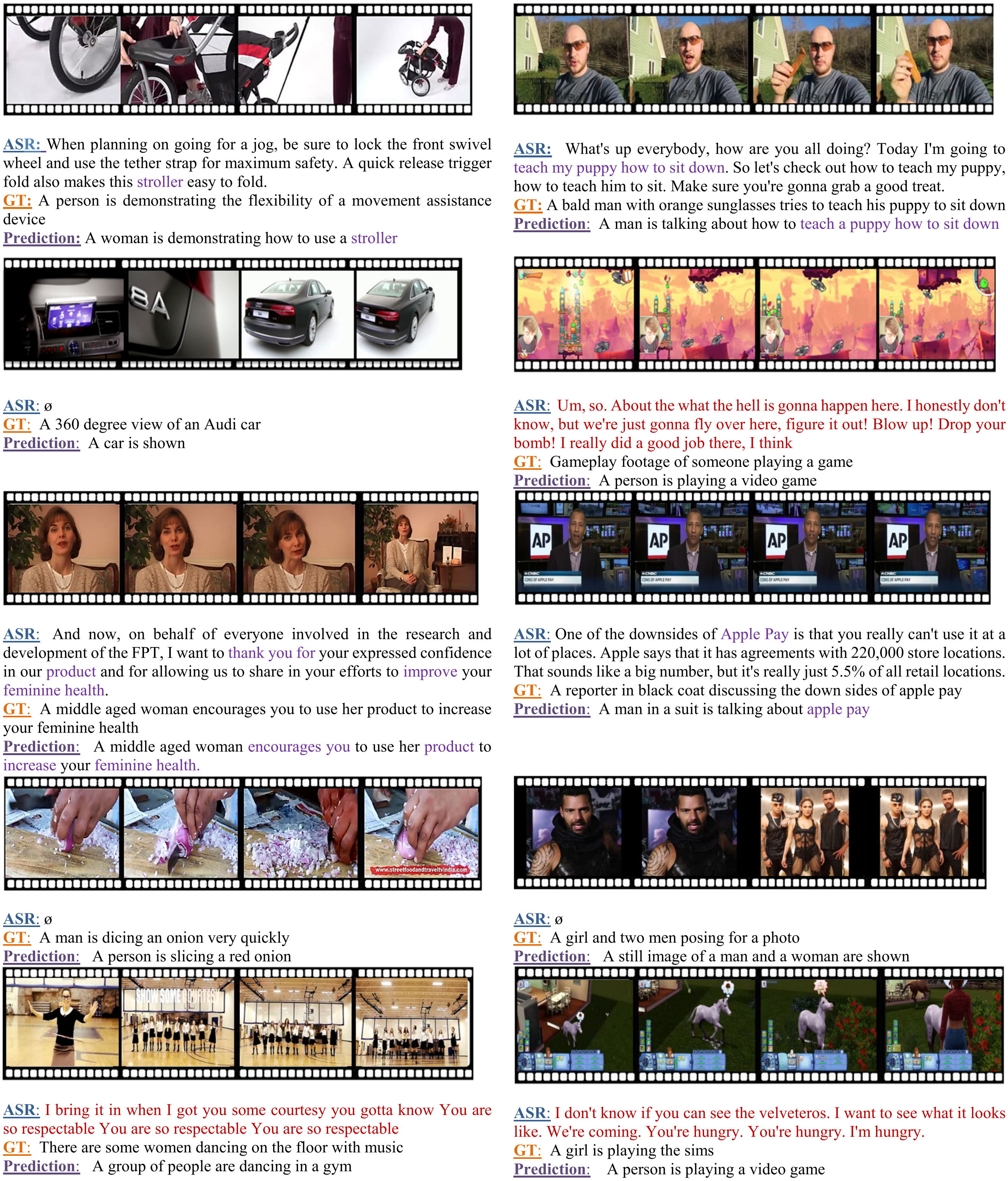


FIGURE 9. **Qualitative results obtained on videos from the MSRVTT dataset.**

TABLE 4 presents the comparison of CapVT with the retained methods on the MSRVTT dataset. CapVT outperforms previous methods by a significant margin of 1.28%, 1.52%, and 1.21% in terms of BLEU4, METEOR, and ROUGE, respectively. Notably, even our model without encoder-decoder initialization (*MAM* + *Cont*) achieves

TABLE 3. Statistics of video captioning models. PT stands for pre-training. *x* stands for unknown.

Method	Size	PT data scale	Hardware (GPUs/TPUs)	Training time
m-PLUG2 [47]	600M	766M	16 NVIDIA A 100 GPUs	-
GIT [8]	681M	800M	*x* NVIDIA A100	-
GIT2 [8]	5.1B	10.5B	*x* NVIDIA A100	-
CLIP-DCD [55]	425M	400M	-	-
VAST [53]	1.3B	324M	64 Tesla V100	-
VideoCoca [54]	2.1B	144M	128 CloudTPU v4	6 hours
UniVL[6]	198M	136M	8 NVIDIA Tesla V100 GPUs	14 days
OA-BTG [21]	-	No PT	-	-
VideoAsMT [12]	286M	136M	-	-
SwinBert [7]	198M	No PT	Nvidia V100 GPUs	-
OpenBook [3]	-	No PT	-	-
CapVT	198M	No PT	2 NVIDIA GeForce 2080 GPUs	4 days

comparable results, highlighting the effectiveness of modality fusion using modality attention and the importance of caption information in guiding the training. We anticipate that further improvements can be achieved by integrating a vision-language, end-to-end pre-training phase on the whole model. The results obtained demonstrate the pertinence of the CapVT model and its potential for achieving superior performance in video captioning tasks.

E. QUALITATIVE RESULTS

Some examples of results obtained on the MSRVTT corpus are illustrated in Figure 9. The results indicate that the quality of the predicted captions is affected by various factors, including the availability of audio and visual information, the complexity of the content and the accuracy of the ASR. When the transcripts are pertinent (with salient words represented in purple in Figure 9), combining textual and visual modalities leads to precise captioning. In contrast, in the absence of the audio or more generally when the transcript channel is not consistent with the content (transcripts represented in red in Figure 9), the predictions rely only on visual cues and are less informative. In some cases, the model lacks access to external knowledge, as illustrated in example 3, where specific information like the car brand remains elusive from both visual and textual cues. To address this limitation, future research axes could explore the integration of external knowledge sources, such as databases or ontologies, offering a promising prospect for enhancing the model's contextual understanding.

TABLE 4. Comparison with state of the art.

Method	BLEU4	METEOR	ROUGE
OA-BTG [21]	0.4140	0.2820	-
VideoAsMT [12]	0.417	0.285	-
UniVL[6]	0.4179	0.2894	0.6087
SwinBert [7]	0.419	0.299	0.621
OpenBook [3]	0.428	0.293	0.617
MAM+Cont	0.4191	0.2803	0.6087
CapVT	**0.4408**	**0.3082**	**0.6291**

Another limitation arises from our offline transcript extraction, illustrated in case 9, where the model lacks awareness that the transcript represents music lyrics. This highlights a challenge where textual data can inadvertently introduce noise into the model's predictions. To mitigate this, integrating audio features emerges as a valuable strategy, not only for recognizing music lyrics but also for scenarios necessitating the recognition of environmental sounds or the capture of emotional context in videos.

In general, the accuracy of the predicted captions is largely influenced by the type and quality of the input data. In all cases, incorporating multimodal approaches can enhance the precision of the predictions.

VI. CONCLUSION

In this work, we have introduced CapVT, a novel architecture that efficiently exploits and combines rich information from both visual and transcript modalities for multimodal video captioning. The proposed modality-attention module and contrastive learning technique make it possible to enhance the representation of inter-modal relationships, leading to a new state-of-the-art performance on the MSRVTT dataset with respect to various evaluation metrics. The proposed model achieves a BLEU4 score of 0.4408, a METEOR score of 0.3082, and a ROUGE score of 0.6291 representing an improvement of 1.28%, 1.52%, and 1.21% respectively with respect to the state of the art. Our comprehensive study of each training strategy demonstrates the effectiveness of the CapVT model and its potential for achieving superior performances for the video captioning task. Furthermore, the study of the effect of the input modalities involved highlights the effectiveness of our training strategies in improving the model's ability to generate accurate captions that rely on both text and visual information.

We have also shown that the gain in performance strongly depends on the nature of the data in different categories. This indicates a need for further research to develop more effective training methods that can take into account in a fine-grained manner the data characteristics of various categories. Future work could explore pre-training on larger datasets to further improve the performance of our approach. Large-scale pre-training allows the model to learn and capture the intricate correlations and interactions between different modalities. It facilitates the comprehension of complex multimodal patterns that may not be discernible in smaller, more constrained datasets. Another potential avenue can be the exploration of knowledge-augmented models. External knowledge sources can enhance the contextual understanding of video content, improve caption accuracy, and ensure domain relevance. They offer potential solutions to handle ambiguous or limited sensory cues, adapt to evolving content, and reduce biases. While knowledge-enhanced NLP models are widely studied, the exploration of knowledge-enhanced vision and multi-modal models is a relatively uncharted territory, presenting an exciting opportunity for further research.

REFERENCES

[1] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[2] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele, "Movie description," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 94–120, 2017.

[3] Z. Zhang, Z. Qi, C. Yuan, Y. Shan, B. Li, Y. Deng, and W. Hu, "Open-book video captioning with retrieve-copy-generate network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9837–9846.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.

[5] P. H. Seo, A. Nagrani, A. Arnab, and C. Schmid, "End-to-end generative pretraining for multimodal video captioning," 2022, *arXiv:2201.08264*.

[6] H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li, J. Li, T. Bharti, and M. Zhou, "UniVL: A unified video and language pre-training model for multimodal understanding and generation," 2020, *arXiv:2002.06353*.

[7] K. Lin, L. Li, C.-C. Lin, F. Ahmed, Z. Gan, Z. Liu, Y. Lu, and L. Wang, "SwinBERT: End-to-end transformers with sparse attention for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17949–17958.

[8] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang, "GIT: A generative image-to-text transformer for vision and language," 2022, *arXiv:2205.14100*.

[9] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020.

[10] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2630–2640.

[11] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7464–7473.

[12] B. Korbar, F. Petroni, R. Girdhar, and L. Torresani, "Video understanding as machine translation," 2020, *arXiv:2006.07203*.

[13] G. Huang, B. Pang, Z. Zhu, C. Rivera, and R. Soricut, "Multimodal pretraining for dense video captioning," 2020, *arXiv:2011.11760*.

[14] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi, "MERLOT: Multimodal neural script knowledge models," 2021, *arXiv:2106.02636*.

[15] H. Mcgurk and J. Macdonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, Dec. 1976.

[16] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5288–5296.

[17] P. Das, C. Xu, R. F. Doell, and J. J. Corso, "A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2634–2641.

[18] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, L. Schmidt, J. Shangguan, J. M. Siskind, J. Waggoner, S. Wang, J. Wei, Y. Yin, and Z. Zhang, "Video in sentences out," 2014, *arXiv:1204.2742*.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[20] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele, "Translating video content to natural language descriptions," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 433–440.

[21] J. Zhang and Y. Peng, "Object-aware aggregation with bidirectional temporal graph for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8327–8336.

[22] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence—Video to text," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4534–4542.

[23] X. Zhang, C. Liu, and F. Chang, "Guidance module network for video captioning," in *Proc. 40th Chin. Control Conf. (CCC)*, Jul. 2021, pp. 7955–7959.

[24] C. Szegedy, S. Iofe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.

[25] H. Xu, B. Li, V. Ramanishka, L. Sigal, and K. Saenko, "Joint event detection and description in continuous video streams," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2019, pp. 396–405.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[27] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.

[28] M. Hemalatha and C. C. Sekhar, "Domain-specific semantics guided approach to video captioning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1576–1585.

[29] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3D residual networks for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 3154–3160.

[30] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4507–4515.

[31] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.

[32] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4534–4542.

[33] Z. Guo, L. Gao, J. Song, X. Xu, J. Shao, and H. T. Shen, "Attention-based LSTM with semantic consistency for videos captioning," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 357–361.

[34] Z. Zhang, Y. Shi, C. Yuan, B. Li, P. Wang, W. Hu, and Z.-J. Zha, "Object relational graph with teacher-recommended learning for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13278–13288.

[35] J. Hessel, B. Pang, Z. Zhu, and R. Soricut, "A case study on combining ASR and visual features for generating instructional video captions," in *Proc. 23rd Conf. Comput. Natural Lang. Learn. (CoNLL)*, 2019, pp. 419–429.

[36] B. Shi, L. Ji, Y. Liang, N. Duan, P. Chen, Z. Niu, and M. Zhou, ''Dense procedure captioning in narrated instructional videos,'' in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6382–6391.

[37] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, ''How2: A large-scale dataset for multimodal language understanding,'' 2018, *arXiv:1811.00347*.

[38] W. Guo, J. Wang, and S. Wang, ''Deep multimodal representation learning: A survey,'' *IEEE Access*, vol. 7, pp. 63373–63394, 2019.

[39] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, ''Multimodal machine learning: A survey and taxonomy,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[40] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, ''VideoCLIP: Contrastive pre-training for zero-shot video-text understanding,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 6787–6800.

[41] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, and J. Gao, ''Vision-language pre-training: Basics, recent advances, and future trends,'' *Found. Trends Comput. Graph. Vis.*, vol. 14, no. 3–4, pp. 163–352, 2022.

[42] G. Li, N. Duan, Y. Fang, M. Gong, and D. Jiang, ''Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training,'' in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 11336–11344.

[43] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, ''VL-BERT: Pre-training of generic visual-linguistic representations,'' 2019, *arXiv:1908.08530*.

[44] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, ''Unified vision-language pre-training for image captioning and VQA,'' in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 13041–13049.

[45] H. Xu, Q. Ye, M. Yan, Y. Shi, J. Ye, Y. Xu, C. Li, B. Bi, Q. Qian, W. Wang, G. Xu, J. Zhang, S. Huang, F. Huang, and J. Zhou, ''MPLUG-2: A modularized multi-modal foundation model across text, image and video,'' 2023, *arXiv:2302.00402*.

[46] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, ''EVA: Exploring the limits of masked visual representation learning at scale,'' 2022, *arXiv:2211.07636*.

[47] J. Lu, D. Batra, D. Parikh, and S. Lee, ''ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,'' 2019, *arXiv:1908.02265*.

[48] H. Tan and M. Bansal, ''LXMERT: Learning cross-modality encoder representations from transformers,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 5100–5111.

[49] D. Hendrycks and K. Gimpel, ''Gaussian error linear units (GELUs),'' 2016, *arXiv:1606.08415*.

[50] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, ''Robust speech recognition via large-scale weak supervision,'' 2022, *arXiv:2212.04356*.

[51] Y. Wu et al., ''Google's neural machine translation system: Bridging the gap between human and machine translation,'' 2016, *arXiv:1609.08144*.

[52] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' 2018, *arXiv:1810.04805*.

[53] S. Chen, H. Li, Q. Wang, Z. Zhao, M. Sun, X. Zhu, and J. Liu, ''VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset,'' 2023, *arXiv:2305.18500*.

[54] S. Yan, T. Zhu, Z. Wang, Y. Cao, M. Zhang, S. Ghosh, Y. Wu, and J. Yu, ''VideoCoCa: Video-text modeling with zero-shot transfer from contrastive captioners,'' 2022, *arXiv:2212.04979*.

[55] B. Yang, T. Zhang, and Y. Zou, ''CLIP meets video captioning: Concept-aware representation learning does matter,'' in *Pattern Recognition and Computer Vision*. Cham, Switzerland: Springer, 2022, pp. 368–381.

[56] X. Wang, G. Chen, G. Qian, P. Gao, X.-Y. Wei, Y. Wang, Y. Tian, and W. Gao, ''Large-scale multi-modal pre-trained models: A comprehensive survey,'' *Mach. Intell. Res.*, vol. 20, no. 4, pp. 447–482, Aug. 2023.

[57] M. C. Schiappa, Y. S. Rawat, and M. Shah, ''Self-supervised learning for videos: A survey,'' *ACM Comput. Surv.*, vol. 55, no. 13s, pp. 1–37, Dec. 2023.

[58] T. Mikolov, K. Chen, G. Corrado, and J. Dean, ''Efficient estimation of word representations in vector space,'' 2013, *arXiv:1301.3781*.

[59] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, ''BLEU: A method for automatic evaluation of machine translation,'' in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2001, pp. 311–318.

[60] A. Lavie and A. Agarwal, ''Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments,'' in *Proc. 2nd Workshop Stat. Mach. Transl. (StatMT)*, 2007, pp. 65–72.

[61] C.-Y. Lin and F. J. Och, ''Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics,'' in *Proc. 42nd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2004, pp. 605–612.

[62] Q. V. Le, N. Jaitly, and G. E. Hinton, ''A simple way to initialize recurrent networks of rectified linear units,'' 2015, *arXiv:1504.00941*.

[63] A. Ratnaparkhi, ''A linear observed time statistical parser based on maximum entropy models,'' in *Proc. 2nd Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 1997, pp. 1–10.

KAOUTHER OUENNICHE received the engineering degree from Ecole Polytechnique de Tunisie, in 2019. She is currently pursuing the Ph.D. degree with Institut Polytechnique de Paris, Télécom SudParis. She then, joined the ARTEMIS Department, Télécom SudParis, in 2020. Her research interests include computer vision, natural language processing, multimodal learning, and content-based video indexing.

RUXANDRA TAPU (Member, IEEE) received the B.S. degree (Hons.) in electronics, telecommunications, and information technology, and the Ph.D. degree in electronics and telecommunication from the University Politehnica of Bucharest, Romania, in 2008 and 2012, respectively, and the Ph.D. degree (Hons.) in informatics from University Paris VI-Pierre et Marie Curie Paris, France. Since 2012, she has been a Senior Researcher with the ARTEMIS Department, Télécom SudParis, France. Her research interests include content-based video indexing and retrieval, pattern recognition, and machine learning techniques.

TITUS ZAHARIA (Member, IEEE) received the engineering degree in electronics and telecommunications and the M.Sc. degree from the Politehnica University of Bucharest, Bucharest, Romania, in 1995 and 1996, respectively, and the Ph.D. degree in mathematics and computer science from University Paris V–Rene Descartes, Paris, France. He joined the ARTEMIS Department, Télécom SudParis, as an Associate Professor, in 2002, where he become a Full Professor, in 2011. His research interests include visual content representation methods, with 2D/3D compression, reconstruction, recognition, computer vision, and indexing applications.