Emotional Video Captioning With Vision-Based Emotion Interpretation Network

Peipei Song, Dan Guo, *Senior Member, IEEE*, Xun Yang, Shengeng Tang, and Meng Wang, *Fellow, IEEE*

***Abstract*— Effectively summarizing and re-expressing video content by natural languages in a more human-like fashion is one of the key topics in the field of multimedia content understanding. Despite good progress made in recent years, existing efforts usually overlooked the emotions in user-generated videos, thus making the generated sentence a bit boring and soulless. To fill the research gap, this paper presents a novel emotional video captioning framework in which we design a Vision-based Emotion Interpretation Network to effectively capture the emotions conveyed in videos and describe the visual content in both factual and emotional languages. Specifically, we first model the emotion distribution over an open psychological vocabulary to predict the emotional state of videos. Then, guided by the discovered emotional state, we incorporate visual context, textual context, and visual-textual relevance into an aggregated multimodal contextual vector to enhance video captioning. Furthermore, we optimize the network in a new emotion-fact coordinated way that involves two losses—*Emotional Indication Loss* and *Factual Contrastive Loss*, which penalize the error of emotion prediction and visual-textual factual relevance, respectively. In other words, we innovatively introduce *emotional representation learning* into an end-to-end video captioning network. Extensive experiments on public benchmark datasets, EmVidCap and EmVidCap-S, demonstrate that our method can significantly outperform the state-of-the-art methods by a large margin. Quantitative ablation studies and qualitative analyses clearly show that our method is able to effectively capture the emotions in videos and thus generate emotional language sentences to interpret the video content.**

***Index Terms*— Emotional video captioning, emotion analysis, emotion-fact coordinated optimization.**

I. Introduction

RAPID development of deep neural networks has made remarkable progress in objective and factual vision understanding, such as image classification [1], object detection [2], action recognition [3], and video captioning [4]. Among these visual tasks, video captioning is a fundamental but more challenging task due to the inherent semantic gap between vision content and natural language, thus attracting increasing research attention in recent years from both computer vision and natural language processing communities. Despite great success achieved by advanced video captioning models, most existing efforts usually overlooked the emotions conveyed in user-generated videos, which can only translate the video content into factual yet boring language sentences [4], [5]. Note that, more and more young people nowadays express emotions by sharing daily life images or short videos on social network platforms, such as WeChat, Twitter, and Instagram. Emotion is an essential factor to describe visual content more accurately and attractively in language [6], [7], which can be effectively captured from human facial expression, action, and pose, *etc*. For example, a scene of *two men are drinking* can express different emotions depending on the men's emotional state (happy or distressed). As shown in Fig. 1 (a∼c), incorporating vision with emotion analysis can enrich the attractiveness and correctness of video descriptions. How to effectively integrate emotion analysis into video captioning to generate more emotional language sentences is critical in the field of video understanding but, as far as we know, has not been well studied.

Preliminary works in image captioning have attempted to subjectively describe the image content in a fixed language style, such as positive or negative [10], humorous or romantic [7], [11], *etc*. Besides, some efforts [6], [12] tried to engage image captioning with emotional traits, such as optimistic, anxious, dramatic, *etc*. However, these advances in image captioning are still limited to a small number of pre-defined emotions, which are not suitable to deal with complex and diverse video content. In this work, we aim to tackle the

Manuscript received 22 June 2022; revised 19 September 2023; accepted 19 January 2024. Date of publication 1 February 2024; date of current version 6 February 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB4500600; in part by the National Natural Science Foundation of China under Grant 62020106007, Grant U22A2094, Grant 62272435, Grant 62272144, Grant 72188101, and Grant U20A20183; and in part by the Major Project of Anhui Province under Grant 202203a05020011. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sen-Ching Samson Cheung. *(Corresponding authors: Dan Guo; Xun Yang; Meng Wang.)*

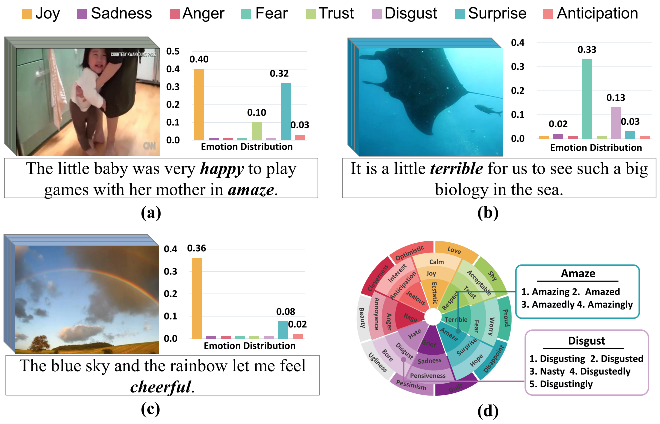
Peipei Song and Xun Yang are with the Department of Electronic Engineering and Information Science, School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: beta.songpp@gmail.com; xyang21@ustc.edu.cn).

Dan Guo is with the Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education, School of Computer Science and Information Engineering, Hefei University of Technology (HFUT), Hefei 230601, China, also with the Hefei Comprehensive National Science Center, Institute of Artificial Intelligence, Hefei 230088, China, and also with Anhui Zhonghuitong Technology Company Ltd., Hefei 230094, China (e-mail: guodan@hfut.edu.cn).

Shengeng Tang is with the Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education, School of Computer Science and Information Engineering, Hefei University of Technology (HFUT), Hefei 230601, China (e-mail: tangsg@hfut.edu.cn).

Meng Wang is with the Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education, School of Computer Science and Information Engineering, Hefei University of Technology (HFUT), Hefei 230601, China, and also with the Hefei Comprehensive National Science Center, China Institute of Artificial Intelligence, Hefei 230088, China (e-mail: eric.mengwang@gmail.com).

Digital Object Identifier 10.1109/TIP.2024.3359045



The little baby was very ***happy*** to play games with her mother in ***amaze***.

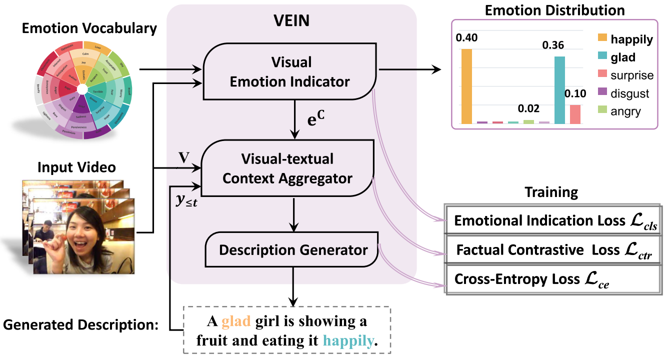
It is a little ***terrible*** for us to see such a big biology in the sea.

The blue sky and the rainbow let me feel ***cheerful***.

Amaze
1. Amazing 2. Amazed
3. Amazedly 4. Amazingly

Disgust
1. Disgusting 2. Disgusted
3. Nasty 4. Disgustedly
5. Disgustingly

Fig. 1. (a∼c): Visual emotion learning of video. (d): *Psychology theory—-Plutchik's Wheel of Emotions* [8], [9]. From this, the emotion vocabulary used in this work includes 34 categories covering totally 179 nouns, adjectives, and adverbs (affective words). In this study, we introduce emotion learning into the captioning model, enabling more human-like descriptions.



A glad girl is showing a fruit and eating it happily.

Fig. 2. A novel Vision-based Emotion Interpretation Network (VEIN) is proposed. VEIN comprises a visual emotion indicator, a visual-textual context aggregator, and a description generator. The emotion indicator first performs emotion distribution learning over a large vocabulary and encodes the top-K words into a compound emotion vector $\mathbf{e}^{\mathbf{C}}$. Then, guided by $\mathbf{e}^{\mathbf{C}}$, the visual-textual context aggregator explores visual, textual, and visual-textual relevant contexts for captioning. For the task, VEIN is optimized with a basic cross-entropy loss $\mathcal{L}_{ce}$, a new emotional indication loss $\mathcal{L}_{cls}$, and a new factual contrastive loss $\mathcal{L}_{ctr}$.

challenging task of emotional video captioning and expect to translate video content into more natural and open emotional descriptions. Wang et al. [9] first contributed a video captioning dataset with emotion expression, named EmVidCap, for this new task. Then they built two captioning modules, *factual* part and *emotional* part, respectively, using the same CNN+LSTM captioning architecture [13] to explore both the factual and emotional information in videos, whose predictions from the two parts are fused as the final output. Song et al. [14] proposed a unified contextual attention network for emotional video captioning, which applied visual and textual attention to capture critical contexts for captioning. Despite their simplicity, the approaches in [9] and [14] do not explicitly model the emotional state conveyed in videos, which easily generate inaccurate emotional sentences due to the significant imbalance between emotion words and common words.

To fill the research gap, we propose to model the visual emotion explicitly by learning an emotion distribution over a large emotion vocabulary before caption decoding in this work. Intuitively, guided by the affective clue, more emotion-specific words can be effectively predicted by the captioning model. In particular, we develop a novel *Vision-based Emotion Interpretation Network* (VEIN) for emotional video captioning. As shown in Fig. 2, VEIN consists of a visual emotion indicator, a visual-textual context aggregator, and a textual description generator. It refers to three main stages: *emotion perceiving*, *video understanding*, and *video describing*, respectively.

For the **emotion perceiving** stage, psychologists have made major successes in building psychological systems with both basic and complex emotion states. For example, the famous *Plutchik's Wheel* [8] is widely applied in the community of affective computing [15], [16]. Inspired by the psychological study, we exploit the vocabulary of *Plutchik's Wheel* as shown in Fig. 1 (d) and perform *emotion distribution learning* over this large and open psychology vocabulary as shown in Fig. 2. To be specific, in our work, the *top-K* emotions with large probabilities are selected as the highly responsive emotions, *e.g.*, *happily* and *glad* in Fig. 2, while we discard the irrelevant emotions with low probabilities, *e.g.*, *disgust* and *angry* in Fig. 2. Then, the word embeddings of *top-K* emotions are aggregated based on their intensity weights into an emotion representation vector $\mathbf{e}^{\mathbf{C}}$. For the **video understanding** stage, guided by the emotion vector $\mathbf{e}^{\mathbf{C}}$, we develop an effective visual-textual context aggregator to exploit the multimodal context. It not only takes into consideration the visual context and textual context (*i.e.*, previously generated words), but also explores the visual-textual contextual relevance to capture the semantic alignment between the video and previously generated words along the timeline. By this way, our proposed VEIN can progressively and effectively understand the video semantics and perceive the context. Finally, for the **video describing** stage, we feed all these contexts into a language decoder to predict the next word until the whole sentence is generated.

In view of the significance of fact and emotion, we consider both of them in a unified captioning optimization framework. We enable an end-to-end optimization, including (1) a widely-used cross-entropy loss $\mathcal{L}_{ce}$ for generating natural and objective captions, (2) an emotional indication loss $\mathcal{L}_{cls}$ that enforces the consistency between the predicted emotion and ground-truth emotion words in the generated caption, and (3) a factual contrastive loss $\mathcal{L}_{ctr}$ that enhances reliable facts by correlating the visual-textual contextual relevance.

Our main contributions are summarized as follows:

- Inspired by the psychological study, we present a simple but effective solution to model the visual emotion explicitly by learning an emotion distribution over a prioritized large vocabulary to instruct the caption generation. Specifically, we learn an emotion representation vector from the given video based on the *top-K* emotions with high confidence.
- We develop an effective multimodal context learning module to enhance emotional video captioning. Under the guidance of the emotion cues, three types of contexts (*i.e.* visual context, textual context, and visual-textual

contextual relevance) are effectively incorporated into the caption generator.
- Extensive experiments on public benchmark datasets with emotion expressions, *i.e.*, EmVidCap and EmVidCap-S, clearly demonstrate the effectiveness of our proposed method using various evaluation metrics. Quantitative and qualitative analyses show that our method can effectively capture the emotions in videos and thus generate emotional sentences to interpret the video content.

The rest of the paper is organized as follows. We overview the related work in Section II and elaborate the proposed VEIN method in Section III. Extensive experiments including quantitative comparison with state-of-the-art methods, ablation study, and visualization analysis are presented in Section IV, followed by a brief conclusion of this work in Section V.

II. Related Work

In this section, we briefly review existing methods referring to traditional video captioning, emotional video captioning, and visual emotion analysis.

A. Traditional Video Captioning

Traditional video captioning aims to generate natural and objective textual sentences for describing videos. Existing efforts can be divided into two stages.

Early works developed the template-based approaches. Researchers used some object or activity classifiers to detect a set of visual concepts and filled these concepts into the pre-defined language template of caption sentence [17], [18]. For example, Krishnamoorthy et al. [19] developed an approach to select the best subject-verb-object triplet as the video caption. However, as a result, the diversity and flexibility of sentences are limited to these predefined templates.

Subsequently, in nowadays stage, sequential learning models are more popular for video captioning. Sequential learning models based on encoder-decoder structure have achieved outstanding success in the field of neural machine translation (NMT) [20]. Inspired by this, researchers introduced various sequential learning based models to address video captioning [13], [21]. Venugopalan et al. [22] used a pre-trained CNN to extract the visual features, then pooled and fed the features to an LSTM decoder for caption generation. However, the pooling strategy ignored the temporal structure in the video. Yao et al. [21] proposed a temporal attention mechanism to summarize the visual feature sequence. Besides, some works proposed to apply spatial attention to focus on different visual regions during captioning. Zhao et al. [23] designed an object-aware tube feature representation by attending on salient objects. Chen and Jiang [24] performed a novel spatial attention on stacked optical flow images with a customized CNN. Later methods progressed by introducing semantic attributes [25], [26] and joint modeling of visual content with compositional text [27], [28]. Other efforts exploited multi-modal information to improve the video captioning performance, such as introducing the object [5], motion [29] and audio [30] features. There are also some recent works focusing on network architecture design, such as CNN [31], Transformer [32], and memory network [33].

B. Emotional Video Captioning

Emotional video captioning is a new emerging task that is still in its infancy. There are merely two work lines referring to the semantic enhancement of caption with emotional factors. The first research line is *stylized captioning* [10], [12], [34], which aims to generate captions in a specified language style, such as romance, pride, and shame. The second line is more generally *emotional video captioning* [9], [14]. Our work belongs to the latter.

For *stylized captioning*, Mathews et al. [10] first proposed a switching RNN model to embed positive or negative sentiment into the generated captions. To be specific, they used two parallel RNNs [35] equipped with a gating mechanism to switch the two RNNs for generating caption: one RNN was trained on a large factual dataset and the other was trained on a small emotional dataset. Chen et al. [7] proposed a style-factual LSTM to incorporate two groups of dynamic attention parameters, which adaptively adjusted the attention weights between the fact and style-related parts in an LSTM. To generalize the capability of captioning models, several methods leveraged unpaired stylized corpus. For example, Gan et al. [11] proposed a semi-supervised framework to leverage both standard vision-caption factual pairs and unpaired stylized language corpus (*e.g.* humorous and romantic sentences) for training and fine-tuning model. Chen et al. [36] designed a domain layer normalization (DLN) mechanism to disentangle the language style from the factual or stylized sentences, which refers to four language styles of fairy tale, romance, humor, and country song lyrics. The aforementioned works handled a sentiment style at once. A few recent works explored multiple but limited and fixed number of styles simultaneously in a single model [37], [38], [39]. In this work, we do not make effort to imitate a specified language style but aim to generate more general emotional descriptions by perceiving the emotion state adaptively in the video.

For *emotional captioning*, Wang et al. [9] proposed the first methodological solution. They trained two vanilla S2VTs [13] on factual and emotional captioning datasets separately. The output probabilities of the two S2VTs are summed to a final probability for caption generation. Song et al. [14] focused on extracting rich context from video and text to improve the quality of emotional captions. They proposed a contextual attention network that introduced a visual attention module and a textual attention module into the LSTM decoder. However, these methods did not do anything explicitly on emotion learning.

In this article, we make efforts to mine the affective clues by: 1) the well-designed video emotion distribution representation covering the open psychological emotion vocabulary, and 2) extracting emotion-guided contexts from visual, textual, and joint visual-textual perspectives.

C. Visual Emotion Analysis

With the popularity of images or short videos on social networks, the research of visual emotion analysis has attracted more and more attention [40], [41]. Most of the current work focuses on addressing it for the visual emotion

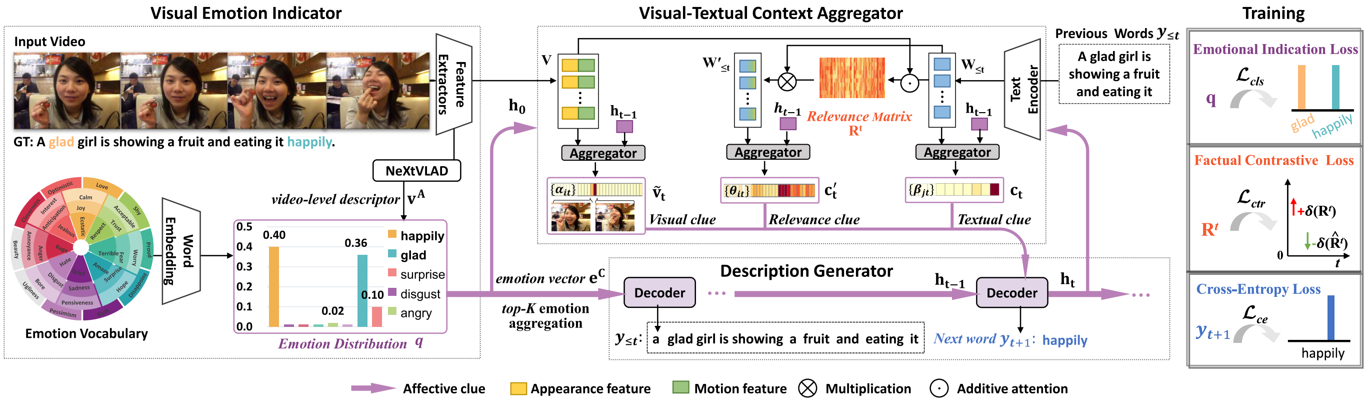


Fig. 3. An overview of the proposed VEIN. First, we learn an emotion distribution $\mathbf{q}$ and choose the *top-K* intensive emotions to integrate a emotion vector $\mathbf{e}^{\mathbf{C}}$. Then, guided by $\mathbf{e}^{\mathbf{C}}$, we aggregate the contexts from video $\mathbf{V}$, partially generated words $y_{\leqslant t}$ and their visual-textual correlation matrix $\mathbf{R}^t$ at each timestamp t. These contexts are fed together into an LSTM-based language decoder (description generator) to produce the video description. Moreover, we propose two new objectives for this task, *i.e.*, imposing on the model with two constraints of emotion distribution $\mathbf{q}$ and visual-textual relevance $\mathbf{R}^t$.

classification task [15], [42]. To bridge the affective gap, researchers primarily disentangled discriminative features that can better distinguish the difference among different emotions. Rao et al. [41] disentangled the emotional clue from image semantics, aesthetics, and low-level features simultaneously to predict the dominant emotion for each image. Yang et al. [42] proposed a stimuli-aware model to recognize emotion from specific stimuli clues, such as color, object, and face. Kosti et al. [43] explored emotion recognition by combining both the facial expression of the person and the global scene in the whole image. In addition, under the consideration of ambiguity and subjectivity of human emotions, noteworthy efforts are devoted to the task of visual emotion distribution learning instead of a single dominant emotion prediction [44], [45]. Yang et al. [45] proposed a well-grounded circular-structured representation to utilize the prior knowledge of pure emotion theory for visual emotion learning.

However, all the above-mentioned works focus on image-based emotion analysis. Nowadays, few emotional methods are developed on videos. For user-generated videos, Zhao et al. [15] utilized spatiotemporal attention to weight emotional-rich image regions and video segments for emotion recognition. Yang et al. [46] considered human portraits to perform human-centered GIF emotion recognition. Mittal et al. [47] developed a time-series perception model that explores the audience's emotions responsive to various movie scenarios. In addition, various downstream tasks are inspired by the visual emotion analysis, such as emotional image retrieval [48], emotional video recommendation [49] and dialogue tasks [50]. Motivated by these works, we propose to address emotional video captioning by injecting emotion learning into the captioning model.

III. Proposed Approach

Our goal is to generate a sentence to describe the given video with emotional expression. As shown in Fig. 3, our proposed VEIN consists of a *visual emotion indicator* (detailed in Sec. III-A), a *visual-textual context aggregator* (detailed in Sec. III-B), and a *description decoder* (detailed in Sec. III-C). The main research questions are two: 1) how to accurately capture the emotions conveyed in the videos (**R1**) and 2) how to effectively model the multimodal context, *i.e.*, visual context, textual context, and visual-textual contextual relevance, for better comprehension of the video (**R2**).

A. Visual Emotion Indicator

Inspired by the psychological study [8], we propose to learn the emotional representation of the video that encodes the human feelings carried in the video. The basic idea is that we model the emotion distribution of the video over a carefully constructed emotion vocabulary and then select the top-K highly responsive emotions to compose an emotion vector $\mathbf{e}^{\mathbf{C}}$. As shown in Fig. 3, we select the *happily* and *glad* as the relevant emotions and discard the *angry* and *disgust* emotions with low confidence. We describe the details as follows.

We first represent the video as a sequence of frame-level feature vectors $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^{N} \in \mathbb{R}^{N \times d_v}$ using pre-trained network models, such as ResNet [51], ResNext [52], or CLIP [53], where N is the number of the sampled frames and d_v denotes the dimension of feature vectors. The second step is to aggregate frame-level features $\{\mathbf{v}_i\}_{i=1}^{N}$ into a compact video-level representation vector $\mathbf{v}^{\mathbf{A}}$ for emotion analysis. For simplicity, we use the aggregation scheme in NeXtVLAD [54]. Formally, each frame-level vector $\mathbf{v}_i$ in $\mathbf{V}$ is first expanded as $\mathbf{v}'_i \in \mathbb{R}^{\lambda d_v}$ via a fully-connected layer, where λ is a width multiplier. Next, we split the expanded vector $\mathbf{v}'_i$ into G groups of lower-dimensional feature vectors $\mathbf{v}''_{ig} \in \mathbb{R}^{\lambda d_v / G}$. Then, each $\mathbf{v}''_{ig}$ is represented as a mixture of residuals from the anchor point $\mathbf{c}_m$ of cluster m. With G groups and M clusters, a compact video descriptor [54] is achieved as follows:

$$\mathbf{v}^{\mathbf{A}} = \sum_{i=1}^{N} \sum_{g=1}^{G} \omega_{igm} (\mathbf{v}''_{ig} - \mathbf{c}_m), \tag{1}$$

where ω_{igm} denotes the weight of i-th video frame vector assigned to the m-th cluster in the g-th group. We set $\lambda = 2$, $G = 8$, and $\lambda M = 4$ in this work, and then $\mathbf{v}^{\mathbf{A}} \in \mathbb{R}^{d_v}$.

After we obtain the aggregated global video representation $\mathbf{v}^{\mathbf{A}}$, the next step is to reveal the emotion tendency of the

TABLE I

EXAMPLE OF THE EMOTION VOCABULARY. WE DISPLAY A PART OF EMOTION CATEGORIES WITH THE TOP 10 PROPORTIONS APPEARED IN EMVIDCAP AND THE CORRESPONDING EMOTION WORDS

Emotion Category	Proportion	Emotion Words
joy	20.0%	joy, happy, joyful, glad, cheerful, delightful, joyfully, happily, gladly, cheerfully, delightfully
calm	11.2%	calm, peaceful, quiet, calmly, quietly, peacefully
surprise	8.4%	surprise, surprised, astonished, wonderful, surprisedly, astonishedly
beauty	8.1%	beauty, beautiful, cute, pretty, beautifully, prettily, cutely
terrible	6.9%	terrible, horrible, scared, awful, sacredly, terribly, awfully
sadness	5.9%	sadness, sad, sorrowful, sadly, sorrowfully
anger	5.8%	anger, annoyed, annoying, angry, angrily, wrathfully
love	4.7%	love, lovely, adorable, lovably, enchantingly
disgust	4.4%	disgust, disgusting, disgusted, nasty, disgustedly, disgustingly
fear	4.2%	fear, afraid, timid, fearful, frightening, fearfully

given video. Given the video representation $\mathbf{v}^{\mathbf{A}}$, we predict the probability distribution of emotion words $\mathbf{q} = (q_1, \cdots, q_{|Voc|}) \in \mathbb{R}^{|Voc|}$ over an open emotion vocabulary, as follows:

$$q_k = \text{softmax}(\mathbf{u}_q^\top \tanh(\mathbf{U}_q \mathbf{v}^{\mathbf{A}} + \mathbf{H}_q \mathbf{e}_k + \mathbf{b}_q)), \tag{2}$$

where $\mathbf{u}_q \in \mathbb{R}^{d_a}$, $\mathbf{U}_q \in \mathbb{R}^{d_a \times d_v}$, $\mathbf{H}_q \in \mathbb{R}^{d_a \times d_w}$, and $\mathbf{b}_q \in \mathbb{R}^{d_a}$ are learnable parameters, and $\mathbf{e}_k \in \mathbb{R}^{d_w}$ is the GloVe embedding of the k-th emotion word in the emotion vocabulary, as shown in Table I. With the emotion distribution $\mathbf{q}$, we select the *top-K* emotion words with large probability scores $\{q_k\}_{k=1}^K$ to calculate a video emotion vector $\mathbf{e}^{\mathbf{C}}$ as shown in Eq. 3:

$$\mathbf{e}^{\mathbf{C}} = \sum\nolimits_{k=1}^{K} q_k \mathbf{e}_k, \tag{3}$$

where $\mathbf{e}^{\mathbf{C}}$ captures the main emotional clues in the video, thus can guide the visual-textual context modeling. The emotion words with low probability scores are usually discarded. So far, we have answered the research question **R1** in this section. Note that we can also use other video aggregate strategies in our visual emotion indicator module to obtain the compact video descriptor. The NeXtVLAD [54] scheme is just applied due to its simplicity and effectiveness.

B. Visual-Textual Context Aggregator

In this section, we introduce how to answer the research question **R2**. It is critical in our work to effectively exploit all relevant and informative contexts for each step of word decoding. The basic idea is that we expect to feed the emotional clue into the modeling of different contexts (*i.e.*, visual, textual, and joint visual-textual contexts) to reach emotion-aware representation and aggregation of contexts for the task of emotional video captioning.

1) Visual Context: We aim to discover the emotional and informative visual frames in the video for the modeling of visual context. As shown in Fig. 3, the frames showing the woman's smiling face can contribute to the prediction of the emotion *happily*. Based on such observation, we design an attention mechanism to model the emotion-rich visual features. As shown in Fig. 3, at the t-th decoding step, we modulate the weights $\{\alpha_{it}\}$ of the frame-level features $\{\mathbf{v}_i\}$ under the guidance of hidden state $\mathbf{h}_{t-1}$ as follows:

$$\alpha_{it} = \text{softmax}(\mathbf{u}_\alpha^\top \tanh(\mathbf{U}_\alpha \mathbf{h}_{t-1} + \mathbf{H}_\alpha \mathbf{v}_i + \mathbf{b}_\alpha)), \tag{4}$$

where $\mathbf{u}_\alpha \in \mathbb{R}^{d_a}$, $\mathbf{U}_\alpha \in \mathbb{R}^{d_a \times d_h}$, $\mathbf{H}_\alpha \in \mathbb{R}^{d_a \times d_v}$, and $\mathbf{b}_\alpha \in \mathbb{R}^{d_a}$ are all learnable network parameters. Note that the initial hidden state $\mathbf{h}_0$ is activated by the emotion vector $\mathbf{e}^{\mathbf{C}}$ by $\mathbf{h}_0 = \text{FC}(\mathbf{e}^{\mathbf{C}})$, which forces the model to pay more attention on the video frames with emotion expression, thus playing a vital role in our proposed VEIN. The FC$(\cdot)$ denotes a fully-connected layer.

Then, the video context representation $\widetilde{\mathbf{v}}_t \in \mathbb{R}^{d_v}$ at the t-th decoding step can be obtained by aggregating the frame-level feature vectors based on the weights $\{\alpha_{it}\}$ in Eq. 4. We deem $\widetilde{\mathbf{v}}_t$ is an emotionally influenced context unit.

$$\widetilde{\mathbf{v}}_t = \sum\nolimits_{i=1}^{N} \alpha_{it} \mathbf{v}_i. \tag{5}$$

2) Textual Context: Here, we consider the contextual transition of previously generated words $\{y_{\leqslant t}\}$. We extract the textual features of $\{y_{\leqslant t}\}$ using GloVe embedding, and further capture the sequential dependency among words using the well-known self-attention [55], [56], [57] module. Formally, the textual features of previous words are transformed into $\mathbf{W}_{\leqslant t} = \{\mathbf{w}_j\}_{j=1}^t \in \mathbb{R}^{t \times d_w} = \Phi([y_1; \cdots; y_t])$, where [;] is a row-wise stacking operator, and Φ denotes the self-attention layer.

As the same to visual context $\widetilde{\mathbf{v}}_t$, we perform the attention mechanism on $\mathbf{W}_{\leqslant t}$ to discover the emotion-aware textual clue. At the t-th decoding step, we assign the weights $\{\beta_{jt}\}$ to textual features $\{\mathbf{w}_j\}$ under the guidance of hidden state $\mathbf{h}_{t-1}$, and the textual features of previous words are aggregated as a textual context vector $\mathbf{c}_t \in \mathbb{R}^{d_w}$ at the t-th decoding step:

$$\begin{cases} \beta_{jt} = \text{softmax}(\mathbf{u}_\beta^\top \tanh(\mathbf{U}_\beta \mathbf{h}_{t-1} + \mathbf{H}_\beta \mathbf{w}_j + \mathbf{b}_\beta)), \\ \mathbf{c}_t = \sum\nolimits_{j=1}^{t-1} \beta_{jt} \mathbf{w}_j, \end{cases} \tag{6}$$

where $\mathbf{u}_\beta \in \mathbb{R}^{d_a}$, $\mathbf{U}_\beta \in \mathbb{R}^{d_a \times d_h}$, $\mathbf{H}_\beta \in \mathbb{R}^{d_a \times d_w}$, and $\mathbf{b}_\beta \in \mathbb{R}^{d_a}$ are learnable parameters. The initial hidden state $\mathbf{h}_0$ is also activated by the emotion vector $\mathbf{e}^{\mathbf{C}}$ in the same way as Eq. 4.

3) Enhanced Visual-Textual Context: Apart from the above contexts $\widetilde{\mathbf{v}}_t$ and $\mathbf{c}_t$, we also explore the frame-word relevance for visual-textual context modeling at each timestamp. In detail, given the partially generated sentence $\mathbf{W}_{\leqslant t}$ and video $\mathbf{V}$, we calculate the relevance matrix $\mathbf{R}^t = \{r_{ij}^t | i \leqslant N, j \leqslant t\} \in \mathbb{R}^{N \times t}$ at the t-th step that captures the frame-word alignment as follows:

$$r_{ij}^t = \mathbf{u}_r^\top \tanh(\mathbf{U}_r \mathbf{v}_i + \mathbf{H}_r \mathbf{w}_j + \mathbf{b}_r), \tag{7}$$

where $\mathbf{u}_r \in \mathbb{R}^{d_a}$, $\mathbf{U}_r \in \mathbb{R}^{d_a \times d_v}$, $\mathbf{H}_r \in \mathbb{R}^{d_a \times d_w}$, and $\mathbf{b}_r \in \mathbb{R}^{d_a}$ are learnable parameters. Based on the relevance matrix $\mathbf{R}^t \in \mathbb{R}^{N \times t}$, we now can aggregate the word embeddings of previously generated words to obtain a contextual feature matrix $\mathbf{W}'_{\leqslant t} = \{\mathbf{w}'_{it}\}_{i=1}^N \in \mathbb{R}^{N \times d_w}$:

$$\mathbf{W}'_{\leqslant t} = \text{softmax}(\mathbf{R}^t)\mathbf{W}_{\leqslant t}, \tag{8}$$

where the operation softmax$(\cdot)$ denotes a row-wise *softmax* operation. Note that both $\mathbf{W}_{\leqslant t} \in \mathbb{R}^{t \times d_w}$ and $\mathbf{R}^t \in \mathbb{R}^{N \times t}$

are two variable-length matrices along the timeline. We then can aggregate the contextual feature matrix $\mathbf{W}'_{\leqslant t} \in \mathbb{R}^{N \times d_w}$ using the attention mechanism in Eq. 6 to obtain an enhanced visual-textual context state vector $\mathbf{c}'_t \in \mathbb{R}^{d_w}$:

$$\begin{cases} \theta_{it} = \text{softmax}(\mathbf{u}_\theta^\top \tanh(\mathbf{U}_\theta \mathbf{h}_{t-1} + \mathbf{H}_\theta \mathbf{w}'_{it} + \mathbf{b}_\theta)), \\ \mathbf{c}'_t = \sum_{i=1}^{N} \theta_{it} \mathbf{w}'_{it}, \end{cases} \tag{9}$$

where $\mathbf{u}_\theta \in \mathbb{R}^{d_a}$, $\mathbf{U}_\theta \in \mathbb{R}^{d_a \times d_h}$, $\mathbf{H}_\theta \in \mathbb{R}^{d_a \times d_w}$, and $\mathbf{b}_\theta \in \mathbb{R}^{d_a}$ are learnable parameters. The initial hidden state vector is also activated by the emotion vector. Finally, we fuse the textual context $\mathbf{c}_t$ and the enhanced context $\mathbf{c}'_t$ as follow:

$$\widetilde{\mathbf{c}}_t = (\mathbf{c}_t + \mathbf{c}'_t). \tag{10}$$

C. Description Generator

So far, we have obtained the emotion indication vector $\mathbf{e}^{\mathbf{C}}$, visual context vector $\widetilde{\mathbf{v}}_t$, and the fused textual context vector $\widetilde{\mathbf{c}}_t$ at the t-th step. We adopt LSTM as the decoder to generate the words by steps. As mentioned previously, we transform $\mathbf{e}^{\mathbf{C}}$ as the initial hidden state $\mathbf{h}_0 \in \mathbb{R}^{d_h}$ of the decoder. It forces the emotion vector $\mathbf{e}^{\mathbf{C}}$ to guide the description generation. For example, if a video's emotion is identified as *sadness* with a large probability, the model inclines to describe the visual content with a sad tone. At each time step, the previously obtained context state vectors $\{\widetilde{\mathbf{v}}_t, \widetilde{\mathbf{c}}_t\}$ and the previous word y_t are fed into the LSTM unit to predict the next word:

$$\begin{cases} \mathbf{h}_0 = \text{FC}(\mathbf{e}^{\mathbf{C}}), & t = 0; \\ \mathbf{h}_t = \text{LSTM}([\widetilde{\mathbf{v}}_t, \widetilde{\mathbf{c}}_t, y_t], \mathbf{h}_{t-1}), & t \in \{1, \ldots L\} \end{cases} \tag{11}$$

where FC is a fully-connected layer, and L is the total length of the to-be-generated sentence. Thus, the word probability prediction that depends on vision $\mathbf{V}$ and emotion $\mathbf{e}^{\mathbf{C}}$ can be formulated as follows:

$$p(y_{t+1} | \mathbf{V}, \mathbf{e}^{\mathbf{C}}, y_{\leqslant t}) = \text{softmax}(\text{FC}(\mathbf{h}_t)). \tag{12}$$

D. Optimization

In this work, we design the optimization objectives of the emotional video captioning network from general, emotional, and factual aspects. At first, we adopt a *cross-entropy loss* $\mathcal{L}_{ce}$ for general captioning training. Then, to generate both emotional and factual descriptions, we design two specific objectives: 1) an *emotional indication loss* $\mathcal{L}_{cls}$ in the emotion encoding phase, which maximizes the probability of the top-ranked candidate emotion words being the ground-truth emotions, and 2) a *factual contrastive loss* $\mathcal{L}_{ctr}$ in the context aggregation phase, which aims to make the visual-textual relevance matrix more distinctive in a weakly-supervised contrastive learning fashion.

1) Cross-Entropy Loss $\mathcal{L}_{ce}$: As a basic objective, we use it to maximize the log-likelihood of each target word y_{t+1} as follows [58], [59]:

$$\mathcal{L}_{ce} = -\sum_t \log p(y_{t+1} | \mathbf{V}, \mathbf{e}^{\mathbf{C}}, y_{\leqslant t}). \tag{13}$$

2) Emotional Indication Loss $\mathcal{L}_{cls}$: We optimize the emotion distribution $\mathbf{q} \in \mathbb{R}^{|Voc|}$ with the automatically obtained emotion labels $\{e_{gt}\}_{gt=1}^{E}$, where $\{e_{gt}\}$ is the intersection set of ground-truth caption and the emotion vocabulary, and E is the set size. To improve the accuracy of the emotion distribution, we design an emotion indication loss $\mathcal{L}_{cls}$ as follows:

$$\mathcal{L}_{cls} = -\sum_{gt}^{E} \delta(e_{gt}) \log(\mathbf{q})^\top, \tag{14}$$

where $\delta(e_{gt}) \in \mathbb{R}^{|Voc|}$ denotes a multi-hot vector, where the value of 1 denotes the occurrence of ground-truth emotion word and otherwise 0.

3) Factual Contrastive Loss $\mathcal{L}_{ctr}$: To generate semantic-rich descriptions, in this work, we further enforce a novel semantic constraint. Inspired by the idea of contrastive learning [60] that discriminates similar but different semantics through the correlation of positive and negative instances, we propose a contrastive loss $\mathcal{L}_{ctr}$ to impose a contrastive constraint on the relevance matrix $\mathbf{R}^t$ (the frame-word alignment of the video and partially generated sentence). The whole process of $\mathcal{L}_{ctr}$ is implemented in a weakly-supervised learning fashion.

We sample positive and negative video-sentence pairs in each training batch. For a video and its generated words ($\mathbf{V}$, $\mathbf{W}_{\leqslant t}$), the relevance matrix $\mathbf{R}^t$ calculated by Eq. 7 is considered as a positive instance. We also construct some negative relevance matrices $\{\widetilde{\mathbf{R}}^t\}$ based on randomly sampled negative pairs ($\widetilde{\mathbf{V}}$, $\mathbf{W}_{\leqslant t}$) or ($\mathbf{V}$, $\widetilde{\mathbf{W}}_{\leqslant t}$). The contrastive loss $\mathcal{L}_{ctr}$ is formulated as:

$$\mathcal{L}_{ctr} = -\sum_{t=1}^{L} \sum_{j=1}^{t} \sum_{i=1}^{N} [\log \sigma(r_{ij}^t) + \log(1 - \sigma(\tilde{r}_{ij}^t))], \tag{15}$$

where $\sigma(\cdot)$ is sigmoid function and the objective $\mathcal{L}_{ctr}$ encourages high relevance scores in positive instances $\mathbf{R}^t$ and penalizes the high relevance scores in negative instances $\widetilde{\mathbf{R}}^t$ simultaneously.

Finally, the total optimized objective is formulated as:

$$\mathcal{L} = \lambda_{ce} \mathcal{L}_{ce} + \lambda_{cls} \mathcal{L}_{cls} + \lambda_{ctr} \mathcal{L}_{ctr}, \tag{16}$$

where λ_{ce}, λ_{cls} and λ_{ctr} are three hyperparameters to modulate the contribution of the three losses.

IV. Experiment

A. Datasets

EmVidCap [9] is a public emotional video captioning dataset, which includes two sub-datasets: EmVidCap-S and EmVidCap-L. **EmVidCap-S** is a small dataset that contains 374 videos originated from factual dataset MSVD [62]. Each video is labeled with roughly 40 emotional captions. Following [9], the dataset is divided into 240/134 videos and 8,169/4,611 sentences for training/testing, respectively. **EmVidCap-L** contains 1,523 videos from VideoEmotion-8 (an emotion prediction dataset) [63]. Each video is annotated around about 17 emotional captions. The dataset is split into 1,141/382 videos and 19,398/6,527 sentences for training/testing. The full EmVidCap dataset contains 27,567 captions and 1,381 videos for training, and 11,138 captions

TABLE II

MAIN COMPARISON ON EMVIDCAP AND EMVIDCAP-S DATASETS. THE RESULTS AND CORRESPONDING FEATURES ARE LISTED

			Semantic							Emotion		Hybrid	
Dataset	Model	Feature	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	Acc_{sw}	Acc_{c}	BFS	CFS
EmVidCap	FT [9]	R152	67.6	47.2	32.0	21.6	20.4	43.1	29.0	51.2	49.6	37.6	33.3
	CANet [14]	R101+RN	68.1	47.7	32.9	22.5	19.7	43.7	34.5	53.7	52.7	38.8	38.2
	SA* [61]	R101+RN	68.4	48.3	33.3	22.4	19.8	44.1	32.4	48.4	45.5	37.8	35.3
	SGN* [4]	R101+RN	68.7	48.9	34.2	24.0	20.1	44.8	35.5	50.4	48.6	39.1	38.3
	CLIP+SA* [53]	CLIP	70.6	51.4	36.7	25.4	21.0	45.9	38.8	53.4	50.7	41.2	41.5
	VEIN (Ours)	R152	67.3	46.9	32.3	21.9	19.6	43.2	30.3	52.2	50.6	37.9	34.5
		R101+RN	69.7	49.9	35.3	24.5	20.7	45.7	37.1	58.1	57.0	41.4	41.2
		CLIP	**72.1**	**52.8**	**37.9**	**27.1**	**21.6**	**46.8**	**39.4**	**59.0**	**57.6**	**43.6**	**43.1**
EmVidCap-S	FT [9]	R152	77.2	60.3	47.4	36.3	29.0	63.4	62.5	69.4	67.1	52.5	63.7
	CANet [14]	R101+RN	78.5	64.0	52.1	41.8	30.8	65.7	74.4	78.7	76.8	57.9	75.1
	SA* [61]	R101+RN	76.3	62.3	51.3	40.3	30.5	64.3	67.0	63.9	61.9	53.9	66.2
	SGN* [4]	R101+RN	77.5	62.7	51.3	41.1	30.6	63.6	71.0	73.9	73.1	56.4	71.5
	CLIP+SA* [53]	CLIP	80.7	67.9	56.3	45.5	33.0	68.2	72.1	68.8	67.2	59.0	71.3
	VEIN (Ours)	R152	78.6	61.6	48.4	37.0	29.4	63.2	64.7	72.3	69.3	53.8	65.9
		R101+RN	79.6	64.4	52.9	42.7	31.7	66.9	71.1	80.1	79.1	59.0	72.8
		CLIP	**82.0**	**68.4**	**57.1**	**45.9**	**33.0**	**69.0**	**79.6**	**82.7**	**82.1**	**62.4**	**80.2**

* indicates the reconstructed results by us. R, RN, and CLIP features are extracted by ResNet [51], 3D-ResNext-101 [52], and CLIP [53], respectively. All the results are reported with percentage (%) as in [9].

over 516 videos for testing. Following [9], we evaluate the model on the EmVidCap and EmVidCap-S datasets. Compared with EmVidCap-S, EmVidCap contains much longer videos (average 23s vs. 10s per video) and more diverse annotations (average 11 tokens vs. 7 tokens).

Furthermore, we also test our model on two well-known datasets for traditional video captioning, MSVD [62] and MSR-VTT [64]. **MSVD** contains 1,970 videos, in which each video is 10∼25 seconds long and annotated with roughly 40 English sentences. The MSVD is separated into 1,200 training, 100 validation, and 670 testing splits [13], [21]. **MSRVTT** is composed of 10,000 videos. Each video is described with 20 English captions. We use the official splits [64], where 6,513 videos for training, 497 videos for validation, and 2,990 videos for testing.

B. Evaluation Metrics

Common standard metrics are used to evaluate the generated sentences [65], [66], *i.e.*, **BLEU-*n***, **METEOR**, **ROUGE**, and **CIDEr** abbreviated to B-n, M, R, and C, respectively. Besides, following the prior work [9], we introduce two emotion metrics $\mathbf{Acc_{sw}}$ and $\mathbf{Acc_{c}}$ to measure the emotion accuracy at word-level and sentence-level. More importantly, new hybrid metrics **BFS** and **CFS** [9] that combine BLEU and CIDEr with emotion metrics are given in Eq. 17. Since both factual and emotional semantics benefit the descriptions, we pay more attention to discuss BFS and CFS in this work.

$$\begin{cases} BFS = k\cdot\sum_{n=1}^{4}\pi_n\cdot\text{BLEU-}n+(1-k)(\frac{Acc_{sw}+Acc_c}{2}), \\ CFS = k\cdot\text{CIDEr}+(1-k)(\frac{Acc_{sw}+Acc_c}{2}), \end{cases} \tag{17}$$

where we set $k = 0.8$ and $\pi_n = \frac{n}{10}$ as in [9].

TABLE III

PERFORMANCE COMPARISON FOR TRADITIONAL VIDEO CAPTIONING ON MSVD AND MSR-VTT DATASETS

		MSVD					MSR-VTT				
Models	Venue	B-4	M	R	C	Size	B-4	M	R	C	Size
SA [61]	CVPR'18	45.3	31.9	64.2	76.2	72.0M	36.3	25.5	58.3	39.9	130.0M
M3 [67]	CVPR'18	52.8	33.3	-	-	-	38.1	26.6	-	-	-
RecNet [61]	CVPR'18	52.3	34.1	69.8	80.3	122.0M	39.1	26.6	59.3	42.7	180.0M
SibNet [68]	MM'18	54.2	34.8	71.7	88.2	-	40.9	27.5	60.2	47.5	-
PickNet [69]	ECCV'18	52.3	33.3	69.6	76.5	-	41.3	27.7	59.8	44.1	-
MARN [70]	CVPR'19	48.6	35.1	71.9	92.2	39.7M	40.4	28.1	60.7	47.1	-
GRU-EVE† [71]	CVPR'19	47.9	35.0	71.5	78.1	-	38.3	28.4	60.7	48.1	-
MGSA [24]	AAAI'19	53.4	35.0	-	86.7	-	42.4	27.6	-	47.5	-
POS+CG [72]	ICCV'19	52.5	34.1	71.3	88.7	-	42.0	28.2	61.6	48.7	172.7M
POS+VCT [73]	ICCV'19	52.8	36.1	71.8	87.8	-	42.3	29.7	62.8	49.1	-
OA-BTG† [74]	CVPR'19	56.9	36.2	-	90.6	-	41.4	28.2	-	46.9	-
STG-KD† [75]	CVPR'20	52.2	36.9	73.9	93.0	-	40.5	28.3	60.9	47.1	-
ORG-TRL† [5]	CVPR'20	54.3	36.4	73.9	95.2	-	43.6	28.8	62.1	50.9	-
PMI-CAP [76]	ECCV'20	54.6	36.4	-	95.1	-	42.1	28.7	-	49.4	-
SAAT† [77]	CVPR'20	46.5	33.5	69.4	81.0	-	40.5	28.2	60.9	49.1	-
SGN [4]	AAAI'21	52.8	35.5	72.9	94.3	53.5M	40.8	28.3	60.8	49.5	55.6M
Swinbert [78]	CVPR'22	**58.2**	**41.3**	**77.5**	**120.6**	2.0G	41.9	29.9	62.1	53.8	2.7G
LSRT† [79]	TIP'22	55.6	37.1	73.5	98.5	-	42.6	28.3	61.0	49.5	-
HRNAT [80]	TIP'22	55.7	36.8	74.1	98.1	-	42.1	28.0	61.6	48.2	-
VEIN (A)	-	54.6	36.9	73.9	94.9	48.8M	44.0	**30.0**	**62.9**	54.3	50.5M
VEIN (A+M)	-	55.7	37.6	74.4	98.9	54.3M	**44.1**	**30.0**	62.6	**55.3**	56.4M

† denotes the methods equipped with an extra object detector. A and M denote appearance and motion features, respectively. The column "Size" lists the model size of each method. The **best** and the second-best methods are highlighted.

C. Implementation Details

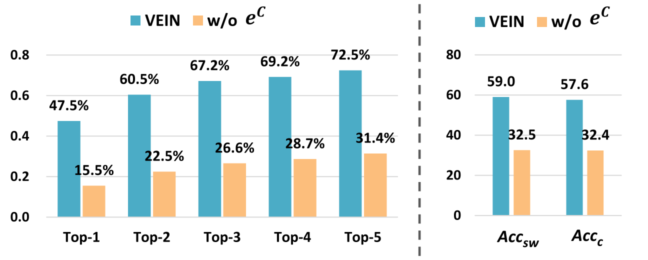
For each video, we sample 30 frames uniformly ($N = 30$) and set the feature dimension to 300 ($d_v = 300$). For the NeXtVLAD setting, we use eight groups and four clusters ($G = 8$ and $K = 4$). About processing the sentence annotations, we tokenize, lowercase and truncate them to 15 words. The word embedding is initialized using GloVe with $d_w = 300$ [81], [82] and the hidden size of LSTM is set to $d_h = 512$. We built two kinds of vocabulary: 1) one is the common vocabulary that consists of words from training sets of EmVidCap (EmVidCap-S) and MSVD as well as special tokens <PAD>, <SOS>, <EOS>, and <UNK>; 2) the other is the emotion vocabulary that is originated from [8] and [9].

TABLE IV
PERFORMANCE COMPARISON FOR STYLIZED IMAGE CAPTIONING ON SENTICAP DATASET [10]

Sentiment	Methods	B-1	B-2	B-3	B-4	M	R	C
Positive	SentiCap [10]	49.1	29.1	17.5	10.8	16.8	36.5	54.4
	StyNet [11]	45.3	-	12.1	-	12.1	-	36.3
	SF-LSTM [7]	50.5	30.8	19.1	12.1	16.6	38.0	60.0
	You et al. [37]	51.2	31.4	19.4	12.3	17.2	38.6	61.1
	MSCap [38]	46.9	-	16.2	-	16.8	-	55.3
	MemCap [39]	50.8	-	17.1	-	16.6	-	54.4
	ERG(Up-Down) [83]	52.4	-	**24.4**	-	18.1	-	77.7
	ERG(VinVL) [83]	53.4	-	23.4	-	18.0	-	75.9
	VEIN	**55.6**	**35.5**	23.2	**15.2**	**20.1**	**42.4**	**79.4**
Negative	SentiCap [10]	50.0	31.2	20.3	13.1	16.8	37.9	61.8
	StyNet [11]	43.7	-	10.6	-	10.9	-	36.6
	SF-LSTM [7]	50.3	31.0	20.1	13.3	16.2	38.0	59.7
	You et al. [37]	52.2	33.6	22.2	14.8	17.1	39.8	70.1
	MSCap [38]	45.5	-	15.4	-	16.2	-	51.6
	MemCap [39]	48.7	-	19.6	-	15.8	-	60.6
	ERG(Up-Down) [83]	52.6	-	21.6	-	18.0	-	68.3
	ERG(VinVL) [83]	53.1	-	21.2	-	18.6	-	70.0
	VEIN	**55.4**	**36.4**	**24.7**	**16.8**	**19.8**	**42.4**	**82.8**

TABLE V
ABLATION STUDIES OF *top-K* EMOTIONS ON EMVIDCAP

top-K	B-1	B-2	B-3	B-4	M	R	C	Acc$_{sw}$	Acc$_c$	BFS	CFS
0	68.6	47.9	32.8	22.3	20.0	44.6	35.0	54.9	53.5	39.0	38.8
1	68.7	48.8	33.9	23.2	20.1	44.5	36.6	57.4	56.0	40.2	40.7
5	**69.7**	**49.9**	**35.3**	**24.5**	**20.7**	**45.7**	**37.1**	58.1	57.0	**41.4**	**41.2**
20	69.6	**49.9**	34.9	24.1	20.3	45.4	**37.1**	56.6	55.2	40.8	40.9
All	68.6	49.0	34.7	24.2	20.6	**45.7**	36.4	**59.5**	**58.1**	41.2	40.9



(a) Inclusion of emotion words (b) Emotion accuracy

Fig. 4. Comparison with the "w/o $\mathbf{e^C}$" that shields the information from the predicted emotions, regarding (a) the frequency of the *top-K* predicted emotion words that appear in the generated caption and (b) emotion metrics on the EmVidCap dataset.

Table I shows an example of the emotion vocabulary. There are totally 14,038 (9,641) words in the common vocabulary for EmVidCap (EmVidCap-S) and $|Voc| = 179$ emotions in the emotion vocabulary.

We use PyTorch on NVIDIA GeForce RTX 2080 Ti GPU for experiments. The model is trained by Adamax optimizer with the learning rate 7×10^{-4} and the batch size is set to 200. We set hyperparameters $\lambda_{ce} = 1$, $\lambda_{ctr} = 1$ and $\lambda_{cls} = 0.2$. Following [9], the model is initialized on the factual dataset MSVD with $\mathcal{L}_{ce}$. In the test stage, the beam-search with a beam size of 5 is used for caption generation. Following existing works [9], [14], we report the best results from a single experiment for a fair comparison.

D. Comparison With State-of-the-Art Methods

1) Comparison on Emotional Video Captioning: There are few works to explore the emotional video captioning task by now. **FT** [9] and **CANet** [14] are the two existing methods for this task. For comparison, we introduce two existing methods in the field of general video captioning, *i.e.*, **SA** [61] and **SGN** [4]. SA is a classical video model and SGN is a brilliant video model achieving SOTA performance. We also discuss the representative pre-trained vision-transformer **CLIP** [53]. We apply CLIP to the model SA, resulting in a competitive method denoted as **CLIP+SA**.

Table II shows the experimental results on EmVidCap and EmVidCap-S datasets. The proposed VEIN achieves new state-of-the-art performances on both datasets. It outperforms the others by a large margin, especially on the emotion metrics. For example, compared to FT, the VEIN improves the Acc$_{sw}$/Acc$_c$ scores from 51.2/49.6 to 59.0/57.6 on EmVidCap, and from 69.4/67.1 to 82.7/82.1 on EmVidCap-S. With the same visual features R101+RN (ResNet-101 [51] and 3D-ResNet-101 [52]), VEIN achieves considerable improvement of 6.7% and 7.9% on BFS and CFS compared to CANet on EmVidCap, respectively. Although the combination CLIP+SA gains remarkable improvements with the advanced CLIP features, VEIN still performs the best with fair settings. On the EmVidCap-S dataset, the CIDEr of CLIP+SA is 72.1, while the VEIN reports much higher results, 79.6. The remarkable improvements demonstrate the effectiveness and advantage of the emotion learning module.

2) Comparison on Factual Video Captioning: To validate the generalization ability, we experiment on factual benchmark datasets MSVD [62] and MSR-VTT [64]. For a fair comparison, we set the emotion vector to zeros, which indicates no emotion learning. As shown in Table III, the VEIN consistently achieves comparable performance to the existing works. It is worth noting that on the MSVD dataset, our method completely surpasses all the compared methods except Swinbert [78], its advantage results from optimizing large-scale parameters. Swinbert [78] trains a video backbone (VidSwin) and a multimodal transformer in an end-to-end manner. As shown in Table III, the size of our model is much smaller than Swinbert [78], 2.7G *vs* 56.4M on MSR-VTT, the difference is nearly 50 times. We have obvious advantages in flexibility and lightweight. Nonetheless, our model outperforms Swinbert [78] on all metrics on the MSR-VTT dataset. VEIN exhibits its effectiveness under both emotional and factual settings of video captioning.

3) Comparison on Stylized Image Captioning: We further evaluate the model on an alternative emotion-related task, *i.e.*, stylized image captioning. Its goal is to generate captions with a desired sentiment (e.g., positive or negative). We extend the method on the popular SentiCap [10] dataset and compare existing methods. As shown in Table IV, our method outperforms existing methods on both positive and negative subsets, e.g., increasing the CIDEr reported by the SOTA method ERG(VinVL) [6] by 3.5 and 12.8, respectively. Our emotion-guided decoding framework effectively provides emotion-related contexts for emotional captioning.

TABLE VI

ABLATION STUDIES OF VISUAL, TEXTUAL, AND THE ENHANCED VISUAL-TEXTUAL CONTEXTS ON EMVIDCAP

Context	B-1	B-2	B-3	B-4	M	R	C	Acc_{sw}	Acc_{c}	BFS	CFS
w/o V-$Clue$	68.6	48.3	33.6	22.9	20.2	45.1	35.8	57.7	55.0	39.9	39.9
w/o T-$Clue$	68.8	48.5	33.5	22.6	19.6	44.4	34.9	53.1	51.2	39.0	38.4
w/o VT-$Clue$	69.3	48.4	33.4	22.7	20.0	44.8	34.2	55.0	53.1	39.4	38.1
w/o $T\&VT$-$Clue$	68.2	48.6	34.0	23.2	19.8	44.7	35.2	51.6	49.2	38.9	38.2
VEIN	**69.7**	**49.9**	**35.3**	**24.5**	**20.7**	**45.7**	**37.1**	**58.1**	**57.0**	**41.4**	**41.2**

"w/o" denotes "without". V-$Clue$, T-$Clue$, and VT-$Clue$ denote visual, textual and the enhanced visual-textual contexts.

TABLE VII

ABLATION STUDIES OF DIFFERENT LOSSES ON EMVIDCAP

$\mathcal{L}_{ce}$	$\mathcal{L}_{ctr}$	$\mathcal{L}_{cls}$	B-1	B-2	B-3	B-4	M	R	C	Acc_{sw}	Acc_{c}	BFS	CFS
✓			67.9	48.0	33.0	22.0	19.0	43.9	33.2	50.6	48.6	38.0	36.5
✓	✓		68.3	48.1	33.1	22.5	19.7	44.6	35.4	52.3	51.2	38.7	38.7
✓		✓	**69.7**	49.4	34.7	24.0	19.9	44.8	37.0	52.7	51.7	40.0	40.0
✓	✓	✓	**69.7**	**49.9**	**35.3**	**24.5**	**20.7**	**45.7**	**37.1**	**58.1**	**57.0**	**41.4**	**41.2**

TABLE VIII

ABLATION STUDIES OF ALTERNATIVE MODULES ON EMVIDCAP

Setting	B-1	B-2	B-3	B-4	M	R	C	Acc_{sw}	Acc_{c}	BFS	CFS
Mean Pool	69.3	49.1	34.7	24.3	20.3	45.2	36.8	56.4	54.1	40.6	40.5
Attention	69.2	48.7	33.6	23.0	19.8	44.8	35.3	54.2	52.3	39.4	38.9
LSTM	**69.8**	49.6	34.7	23.9	**20.8**	45.3	36.4	**58.8**	56.2	41.0	40.7
BERT	**69.8**	49.3	34.1	23.0	19.9	45.0	34.8	54.8	53.1	39.8	38.7
Cross-$\mathbf{R}^{t}$	69.7	49.0	33.7	22.7	20.4	45.3	36.8	57.9	56.0	40.2	40.8
VEIN (NeXtVLAD)	69.7	**49.9**	**35.3**	**24.5**	20.7	**45.7**	**37.1**	58.1	**57.0**	**41.4**	**41.2**

TABLE IX

COMPARISON OF DIFFERENT DECODING STRATEGIES ON EMVIDCAP

Method	B-1	B-2	B-3	B-4	M	R	C	Acc_{sw}	Acc_{c}	BFS	CFS
BS	72.1	52.8	37.9	27.1	21.6	46.8	39.4	59.0	57.6	43.6	43.1
CBS	72.0	52.7	37.9	27.0	21.6	46.8	39.2	59.6	58.1	43.7	43.1

E. Ablation Study

1) Model Test of VEIN: To analyze the proposed method deeply, we conduct various ablation studies on the EmVidCap dataset, referring to testing *top-K* emotions (Table V), different contexts (Table VI) and loss functions (Table VII).

a) Effect of top-K emotions: As shown in Table V, we test the effect of *top-K* emotions with different choices of $K \in \{0, 1, 5, 20, \text{All}\}$. When $K = 0$, we remove $\mathbf{e}^{\mathbf{C}}$ from the model and initialize $\mathbf{h}_0$ with zeros. We can observe from Table V that $K = 0$ performs the worst. The scores of $\text{Acc}_{sw}/\text{Acc}_{c}$ drop from 58.1/57.90 to 54.9/53.5. It indicates that introducing the emotion vector $\mathbf{e}^{\mathbf{C}}$ into the caption model can effectively promote the emotional style of the description sentences. We also observe that the model reaches the highest scores at $K = 5$ with BFS of 41.4 and CFS of 41.2, respectively. When $K = 1$, the small emotion coverage may not be enough to guide diverse attempts of captioning process; when $K = 20$ or All, the emotion vector would be less informative due to the noisy emotional cue.

Besides, Figure 4 shows the frequency of the *top-K* predicted emotion words that appear in the generated caption. The comparison model "w/o e^{C}" shields the information from the predicted emotions. It can be found that the model with emotion guidance effectively promotes both emotion inclusion and emotion accuracy.

b) Effect of context modeling: Here, we test the effect of various contexts—visual ($\tilde{\mathbf{v}}_t$), textual ($\mathbf{c}_t$), and the enhanced visual-textual ($\mathbf{c}'_t$) contexts. From Table VI, "w/o V-$Clue$", "w/o T-$Clue$", "w/o VT-$Clue$", and "w/o $T\&VT$-$Clue$" denote removing $\tilde{\mathbf{v}}_t$, $\mathbf{c}_t$, $\mathbf{c}'_t$, and $(\mathbf{c}_t + \mathbf{c}'_t)$ from the proposed VEIN, respectively. In "w/o V-$Clue$", we use the averagely pooled visual feature $\bar{\mathbf{v}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{v}_i$ to replace $\tilde{\mathbf{v}}_t$. $\bar{\mathbf{v}}$ performs much worse than $\tilde{\mathbf{v}}_t$. Observing Acc_{sw} and Acc_{c}, the textual context T-$Clue$ brings higher improvement of emotion accuracy than V-$Clue$ or VT-$Clue$. The removal of $T\&VT$-$Clue$ (totally removing the textual influence) leads to the worst performance. To summarize, as shown in Table VI, removing either context degrades the performance on all the metrics.

c) Effect of different losses: In our wrok, $\mathcal{L}_{ce}$ is the basic cross-entropy loss. We test the model by adding $\mathcal{L}_{ctr}$ and $\mathcal{L}_{cls}$ step by step. As shown in Table VII, combining either $\mathcal{L}_{cls}$ or $\mathcal{L}_{ctr}$ with $\mathcal{L}_{ce}$, the performance rises with respect to all the metrics. Among them, the effect of introducing $\mathcal{L}_{cls}$ is more obvious. $\mathcal{L}_{cls}$ is crucial to ensure the reliability of emotion guidance during captioning. The best result is achieved with the combination of all three losses. It validates that the emotion-fact coordinated optimization significantly boosts the experimental performances.

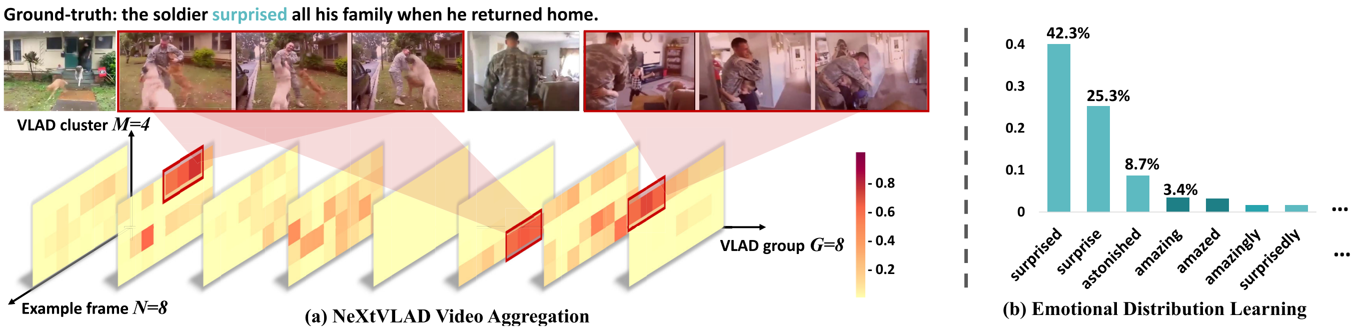
2) Discussion on Alternative Module Implementations: Here, we explore some alternative implementations of our model. All the ablation studies are conducted on EmVidCap dataset and results are shown in Table VIII. Significantly, the results of these alternative implementations shown in Table VIII are still superior to existing approaches [4], [9].

a) Visual feature aggregation of video: In the emotion indicator stage, we use NeXtVLAD [54] to tackle the feature aggregation of video. There are some alternative implementations, such as mean pooling [27], attention [61], and LSTM [13]. As shown in Table VIII, NeXtVLAD performs the best. It can effectively aggregate the feature sequence from a global view of video by multiple times.

b) GloVe vs. bert textual features: In our original setting, we use the GloVe embedding. As shown in Table VIII, ours (VEIN with GloVe) performs better than that with BERT. The reason is that considering the training cost, the BERT is a large pre-trained model and we freeze the pre-trained BERT parameters in experiments. In contrast, when using GloVe, we jointly train it with the VEIN architecture.

c) Visual-textual correlation: As shown in Table VIII, "Cross-$\mathbf{R}^{t}$" denotes the model that adopts a cross-attention mechanism [84] instead of additive attention to calculate $\mathbf{R}^{t}$. Its performance is comparable but slightly inferior to the VEIN with additive attention.

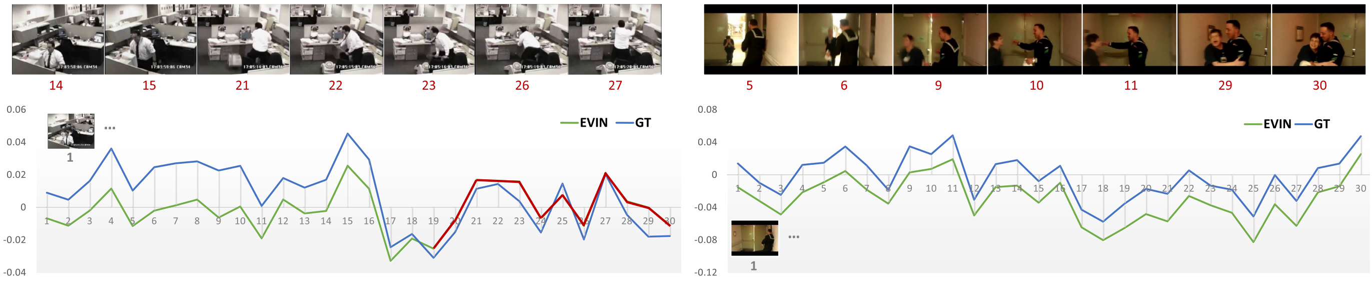
d) BS vs CBS decoding strategy: In the inference stage, we use the beam search (BS) [61] with a beam size of 5 for caption generation. Here, we adopt the constrained beam search (CBS) [85] instead to encourage the inclusion of predicted emotion in the generated caption. We consider the top-1 predicted emotion word as a constraint, which is additionally added to the beam at each decoding time step.



(a) NeXtVLAD Video Aggregation

(b) Emotional Distribution Learning

Fig. 5. Illustrations of aggregation weights $\{\omega_{igm}\}$ in NeXtVLAD (a) and emotion distribution prediction (b). The proposed VEIN attends discriminative frames with obvious emotion intention and filters out irrelevant backgrounds.



(a) GT: a man is working on computer and gets ***angry*** for the noise.
VEIN: a man in the office was ***angry*** with his colleague in the office.

(b) GT: a woman was **surprised** when she opened the door because she saw her husband.
VEIN: the woman was very ***surprised*** when she saw her husband.

Fig. 6. Embedding distance-based evaluation on video and its caption sentences.

As shown in Table IX, the CBS decoding strategy improves Acc_{sw} and Acc_{c} by 0.6 and 0.5, respectively. However, the semantic metrics drop slightly. This may be due to the forced production of emotion words during decoding, which will affect the fluency of sentences.

F. Qualitative Analysis

1) Visualization of Video Aggregation: In Section III-A, we use NeXtVLAD [54] to obtain a video-level descriptor for emotion distribution learning. To display the interpretability of feature aggregation, we show a video instance in Fig. 5, which expresses emotion *surprise* in the video except for the first and fifth background frames. There are three highly attentive regions, such as $\{g = 2, m = 4\}$, $\{g = 6, m = 1\}$ and $\{g = 8, m = 4\}$ in the NeXtVLAD attention map. The two former attentive regions pay more attention to the 2∼4-th frames, while the last region attends to the 6∼8-th frames. We also provide the emotion distribution in Fig. 5 (b), the target emotion *surprised* displays the highest score of 42.3%.

2) Visualization of Video-Caption Distance: Shi et al. pointed out that embedding distance-based evaluation between cross-modal data is effective and can be considered as a supplement to human judgment [86]. The evaluation of video $(\{\mathbf{v}_i\}, i \in \{1, \cdots, N\})$ and caption $\bar{\mathbf{y}}$ is formulated as follows:

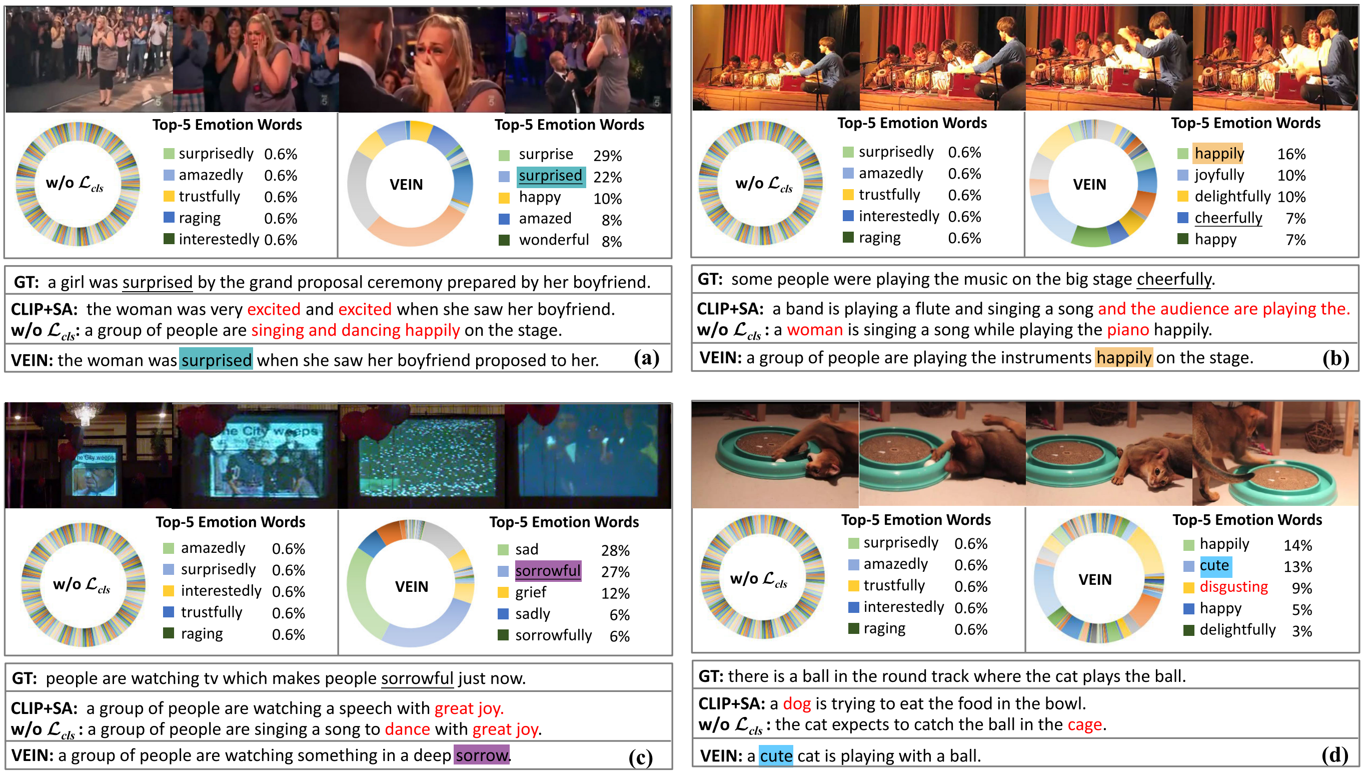
$$Sim(\mathbf{v}_i, \bar{\mathbf{y}}) = \frac{\mathbf{v}_i \cdot \bar{\mathbf{y}}}{||\mathbf{v}_i||_2 \times ||\bar{\mathbf{y}}||_2} \tag{18}$$

where $\mathbf{v}_i$ denotes the i-th visual feature of video and $\bar{\mathbf{y}}$ is the mean pooling of all the word embeddings in the caption.

Figure 6 illustrates two examples to show the video-caption distance based on embedding distance evaluation. Obviously, the VEIN is consistent with the ground-truth (GT). Both the similarity scores of VEIN and GT are relatively high at emotion-specific frames and low at emotionless ones. For example, at the 19∼30-th frames in Fig. 6 (a), the score of VEIN coincides with the GT. The VEIN can interpret visual emotions well.

3) Visualization of Emotion Indication: We visualize the predicted emotion distribution over the large emotion vocabulary and list the Top-5 ones in Fig. 7. "w/o $\mathcal{L}_{cls}$" is incapable of emotion recognition, namely in the case that all the emotions are set with the same intensity to the video; in other words, there is no emotion preference learning before the captioning process. By comparison, VEIN responses to the relevant emotions with strong intensities. For example in Fig. 7, VEIN predicts *surprise*/*surprised* (29%/22%) for videos (a) and *happily*/*joyfully* (16%/10%) for video (b), respectively. Experiment facts demonstrate that the emotion clue is intuitively helpful in generating emotional descriptions.

4) Visualization of Visual-Textual Relevance: Taking video (a) in Fig. 7 as an example, we display the relevance matrix $\mathbf{R}^t$ along the timeline t and the gradually generated sentence in Fig. 8. $\mathbf{R}^t$ is a changeable variable. We have two observations. 1) At each step, the relevance matrix shows that the previously generated words, especially emotion words, are useful to predict the next word. 2) At the 5-th step in this example, the relevance between the video and emotion word *surprised* attends the frames with more emotion clues than other ones (such as background frames). These observations show effective emotional and factual semantics propagation in term of visual-textual correlation.



(a)

GT: a girl was surprised by the grand proposal ceremony prepared by her boyfriend.

CLIP+SA: the woman was very excited and excited when she saw her boyfriend.
w/o $\mathcal{L}_{cls}$: a group of people are singing and dancing happily on the stage.

VEIN: the woman was surprised when she saw her boyfriend proposed to her.

(b)

GT: some people were playing the music on the big stage cheerfully.

CLIP+SA: a band is playing a flute and singing a song and the audience are playing the.
w/o $\mathcal{L}_{cls}$: a woman is singing a song while playing the piano happily.

VEIN: a group of people are playing the instruments happily on the stage.

(c)

GT: people are watching tv which makes people sorrowful just now.

CLIP+SA: a group of people are watching a speech with great joy.
w/o $\mathcal{L}_{cls}$: a group of people are singing a song to dance with great joy.

VEIN: a group of people are watching something in a deep sorrow.

(d)

GT: there is a ball in the round track where the cat plays the ball.

CLIP+SA: a dog is trying to eat the food in the bowl.
w/o $\mathcal{L}_{cls}$: the cat expects to catch the ball in the cage.

VEIN: a cute cat is playing with a ball.

Fig. 7. Visualization results. Underline indicates the emotion words in ground-truth. Red fonts indicate the error generations. VEIN accurately identifies the emotion and generates promising emotional descriptions.

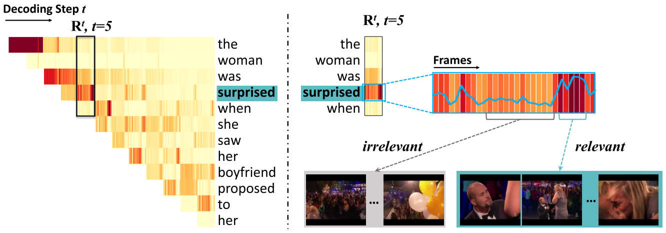
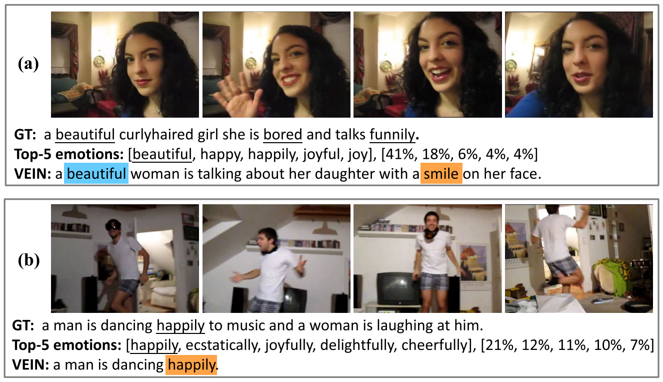


Fig. 8. An example of visual-textual relevance matrix $\mathbf{R}^t$ along timeline. Orange color marks relevance magnitude.



(a)

GT: a beautiful curlyhaired girl she is bored and talks funnily.
Top-5 emotions: [beautiful, happy, happily, joyful, joy], [41%, 18%, 6%, 4%, 4%]
VEIN: a beautiful woman is talking about her daughter with a smile on her face.

(b)

GT: a man is dancing happily to music and a woman is laughing at him.
Top-5 emotions: [happily, ecstatically, joyfully, delightfully, cheerfully], [21%, 12%, 11%, 10%, 7%]
VEIN: a man is dancing happily.

Fig. 9. Two failure cases. *bored* or *disgusted* are annotated from the sound source of the video rather than visual content.

5) Visualization of Generated Captions: By observing Fig. 7 again, the bottom row visualizes the caption sentences generated by CLIP+SA, "w/o $\mathcal{L}_{cls}$", VEIN, and the ground-truth (GT). The VEIN performs better in understanding the video within both factual and emotional modes than the other approaches. Others fail by generating some emotional errors or irrelevant visual descriptions. In example (c), CLIP+SA and "w/o $\mathcal{L}_{cls}$" predict a wrong emotion *great joy* with the crowd fact *dance*; our VEIN describes the accurate emotion *surprised* of the woman who "saw her boyfriend proposed to her". Our superiority may be attributed to the novel design of VEIN by exploiting the visual emotion distribution (affective clue) in the captioning model. Example (d) shows a confusing video, at first, the model predicts two different emotions *happily* and *cute* with similar probabilities of 14% and 13%. However, why does the VEIN finally output *cute* (the 2nd-rank emotion) rather than *happily* (the 1st-rank emotion)? The factual contrastive loss $\mathcal{L}_{ctr}$ and CE loss $\mathcal{L}_{ce}$ restrict the model to generate factual description. Note that both *cute* and *happily* do not appear in the GT label and there is an indeed *cute* cat.

Furthermore, Fig. 9 shows two failure cases. Their ground-truth emotion labels are annotated from the audio source rather than video, such as *talks funnily* in Fig. 9 (a) and *a woman is laughing at him* in Fig. 9 (b). Our method merely handles the vision without audio. If only considering the visual appearance, the prediction of the VEIN seems to be reasonable. The joint audio-visual emotion will be our future research direction.

V. Conclusion

In this paper, we propose a novel Vision-based Emotion Interpretation Network (VEIN) for emotional video description. It considers both emotion and fact two aspects. We perform an emotion distribution learning over a large emotion vocabulary to capture the emotion cue in the video. Guided by the emotion cue, we explore different types of contexts (*i.e*. visual, textual, and enhanced visual-textual contexts) to boost video understanding and multimodal context modeling. Moreover, two new losses for this task—emotional indication loss and factual contrastive loss, are introduced to enhance the optimization of our method. Experimental and visualization results have clearly demonstrated the superiority of our proposed method.

References

[1] C.-F. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 357–366.

[2] P. Sun et al., "Sparse R-CNN: End-to-end object detection with learnable proposals," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14454–14463.

[3] C.-F.-R. Chen et al., "Deep analysis of CNN-based spatio-temporal representations for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6165–6175.

[4] H. Ryu, S. Kang, H. Kang, and C. D. Yoo, "Semantic grouping network for video captioning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 3, pp. 2514–2522.

[5] Z. Zhang et al., "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.

[6] K. Uehara, Y. Mori, Y. Mukuta, and T. Harada, "ViNTER: Image narrative generation with emotion-arc-aware transformer," 2022, *arXiv:2202.07305*.

[7] T. Chen et al., "'Factual' or 'Emotional': Stylized image captioning with adaptive learning and attention," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 519–535.

[8] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of Emotion*. Amsterdam, The Netherlands: Elsevier, 1980, pp. 3–33.

[9] H. Wang, P. Tang, Q. Li, and M. Cheng, "Emotion expression with fact transfer for video description," *IEEE Trans. Multimedia*, vol. 24, pp. 715–727, 2022.

[10] A. P. Mathews, L. Xie, and X. He, "SentiCap: Generating image descriptions with sentiments," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 3574–3580.

[11] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, "StyleNet: Generating attractive visual captions with styles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3137–3146.

[12] K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston, "Engaging image captioning via personality," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 12516–12526.

[13] 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.

[14] P. Song, D. Guo, J. Cheng, and M. Wang, "Contextual attention network for emotional video captioning," *IEEE Trans. Multimedia*, vol. 25, pp. 1858–1867, 2022.

[15] S. Zhao et al., "An end-to-end visual-audio attention network for emotion recognition in user-generated videos," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 303–311.

[16] J. Wei, X. Yang, and Y. Dong, "User-generated video emotion recognition based on key frames," *Multimedia Tools Appl.*, vol. 80, no. 9, pp. 14343–14361, Apr. 2021.

[17] S. Guadarrama et al., "YouTube2Text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2712–2719.

[18] 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.

[19] N. Krishnamoorthy, G. Malkarnenkar, R. Mooney, K. Saenko, and S. Guadarrama, "Generating natural-language video descriptions using text-mined knowledge," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2013, pp. 541–547.

[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[21] L. Yao et al., "Describing videos by exploiting temporal structure," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4507–4515.

[22] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," in *Proc. Conf. 9th Amer. Chapter Assoc. Comput. Linguistics*, 2015, pp. 1494–1504.

[23] B. Zhao et al., "Video captioning with tube features," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 1177–1183.

[24] S. Chen and Y.-G. Jiang, "Motion guided spatial attention for video captioning," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 8191–8198.

[25] Z. Gan et al., "Semantic compositional networks for visual captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5630–5639.

[26] Y. Pan, T. Yao, H. Li, and T. Mei, "Video captioning with transferred semantic attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6504–6512.

[27] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4594–4602.

[28] X. Long, C. Gan, and G. De Melo, "Video captioning with multi-faceted attention," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 173–184, Mar. 2018.

[29] S. Chen and Y. Jiang, "Motion guided region message passing for video captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 1543–1552.

[30] M. Monfort et al., "Spoken moments: Learning joint audio-visual representations from video descriptions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 14871–14881.

[31] H. Xiao, J. Xu, and J. Shi, "Exploring diverse and fine-grained caption for video by incorporating convolutional architecture into LSTM-based model," *Pattern Recognit. Lett.*, vol. 129, pp. 173–180, Jan. 2020.

[32] M. Tang, Z. Wang, Z. Liu, F. Rao, D. Li, and X. Li, "CLIP4Caption: CLIP for video caption," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4858–4862.

[33] M. Qi, J. Qin, D. Huang, Z. Shen, Y. Yang, and J. Luo, "Latent memory-augmented graph transformer for visual storytelling," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4892–4901.

[34] A. Mathews, L. Xie, and X. He, "SemStyle: Learning to generate stylised image captions using unaligned text," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8591–8600.

[35] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137.

[36] C.-K. Chen, Z. Pan, M.-Y. Liu, and M. Sun, "Unsupervised stylish image description generation via domain layer norm," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 8151–8158.

[37] Q. You, H. Jin, and J. Luo, "Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions," 2018, *arXiv:1801.10121*.

[38] L. Guo, J. Liu, P. Yao, J. Li, and H. Lu, "MSCap: Multi-style image captioning with unpaired stylized text," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4204–4213.

[39] W. Zhao, X. Wu, and X. Zhang, "MemCap: Memorizing style knowledge for image captioning," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12984–12992.

[40] W. Zhang, X. He, and W. Lu, "Exploring discriminative representations for image emotion recognition with CNNs," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 515–523, Feb. 2020.

[41] T. Rao, X. Li, and M. Xu, "Learning multi-level deep representations for image emotion classification," *Neural Process. Lett.*, vol. 51, no. 3, pp. 2043–2061, Jun. 2020.

[42] J. Yang, J. Li, X. Wang, Y. Ding, and X. Gao, "Stimuli-aware visual emotion analysis," *IEEE Trans. Image Process.*, vol. 30, pp. 7432–7445, 2021.

[43] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Context based emotion recognition using EMOTIC dataset," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 11, pp. 2755–2766, Nov. 2020.

[44] T. He and X. Jin, "Image emotion distribution learning with graph convolutional networks," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2019, pp. 382–390.

[45] J. Yang, J. Li, L. Li, X. Wang, and X. Gao, “A circular-structured representation for visual emotion distribution learning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4237–4246.

[46] Z. Yang, Y. Zhang, and J. Luo, “Human-centered emotion recognition in animated GIFs,” in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 1090–1095.

[47] T. Mittal, P. Mathur, A. Bera, and D. Manocha, “Affect2MM: Affective analysis of multimedia content using emotion causality,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 5661–5671.

[48] Z. Wei et al., “Learning visual emotion representations from web data,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 13106–13115.

[49] N. Vedula et al., “Multimodal content analysis for effective advertisements on Youtube,” in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 1123–1128.

[50] B. Huber, D. McDuff, C. Brockett, M. Galley, and B. Dolan, “Emotional dialogue generation using image-grounded language models,” in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–12.

[51] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[52] K. Hara, H. Kataoka, and Y. Satoh, “Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet?” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.

[53] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[54] R. Lin, J. Xiao, and J. Fan, “NeXtVLAD: An efficient neural network to aggregate frame-level features for large-scale video classification,” in *Proc. Eur. Conf. Comput. Vis. Workshops*, vol. 11132, 2018, pp. 206–218.

[55] A. Vaswani et al., “Attention is all you need,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[56] K. Li, J. Li, D. Guo, X. Yang, and M. Wang, “Transformer-based visual grounding with cross-modality interaction,” *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 6, pp. 1–19, Nov. 2023.

[57] J. Dong et al., “Partially relevant video retrieval,” in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 246–257.

[58] Y. Li, X. Yang, X. Shang, and T.-S. Chua, “Interventional video relation detection,” in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4091–4099.

[59] Q. Zheng et al., “Progressive localization networks for language-based moment localization,” *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 19, no. 2, pp. 1–21, 2023.

[60] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[61] B. Wang, L. Ma, W. Zhang, and W. Liu, “Reconstruction network for video captioning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7622–7631.

[62] D. Chen and W. B. Dolan, “Collecting highly parallel data for paraphrase evaluation,” in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2011, pp. 190–200.

[63] Y.-G. Jiang, B. Xu, and X. Xue, “Predicting emotions in user-generated videos,” in *Proc. AAAI Conf. Artif. Intell.*, 2014, pp. 73–79.

[64] 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.

[65] D. Guo, Y. Wang, P. Song, and M. Wang, “Recurrent relational memory network for unsupervised image captioning,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2020, pp. 920–926.

[66] S. Tang, R. Hong, D. Guo, and M. Wang, “Gloss semantic-enhanced network with online back-translation for sign language production,” in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 5630–5638.

[67] J. Wang, W. Wang, Y. Huang, L. Wang, and T. Tan, “M3: Multimodal memory modelling for video captioning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7512–7520.

[68] S. Liu, Z. Ren, and J. Yuan, “SibNet: Sibling convolutional encoder for video captioning,” in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1425–1434.

[69] Y. Chen, S. Wang, W. Zhang, and Q. Huang, “Less is more: Picking informative frames for video captioning,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 367–384.

[70] W. Pei, J. Zhang, X. Wang, L. Ke, X. Shen, and Y.-W. Tai, “Memory-attended recurrent network for video captioning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8347–8356.

[71] N. Aafaq, N. Akhtar, W. Liu, S. Z. Gilani, and A. Mian, “Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12487–12496.

[72] B. Wang, L. Ma, W. Zhang, W. Jiang, J. Wang, and W. Liu, “Controllable video captioning with POS sequence guidance based on gated fusion network,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2641–2650.

[73] J. Hou, X. Wu, W. Zhao, J. Luo, and Y. Jia, “Joint syntax representation learning and visual cue translation for video captioning,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8917–8926.

[74] 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.

[75] B. Pan et al., “Spatio-temporal graph for video captioning with knowledge distillation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10870–10879.

[76] S. Chen, W. Jiang, W. Liu, and Y.-G. Jiang, “Learning modality interaction for temporal sentence localization and event captioning in videos,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 333–351.

[77] Q. Zheng, C. Wang, and D. Tao, “Syntax-aware action targeting for video captioning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13096–13105.

[78] K. Lin et al., “SwinBERT: End-to-end transformers with sparse attention for video captioning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17928–17937.

[79] L. Li, X. Gao, J. Deng, Y. Tu, Z.-J. Zha, and Q. Huang, “Long short-term relation transformer with global gating for video captioning,” *IEEE Trans. Image Process.*, vol. 31, pp. 2726–2738, 2022.

[80] L. Gao, Y. Lei, P. Zeng, J. Song, M. Wang, and H. T. Shen, “Hierarchical representation network with auxiliary tasks for video captioning and video question answering,” *IEEE Trans. Image Process.*, vol. 31, pp. 202–215, 2022.

[81] K. Li, D. Guo, and M. Wang, “Proposal-free video grounding with contextual pyramid network,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 3, pp. 1902–1910.

[82] D. Guo, H. Wang, and M. Wang, “Context-aware graph inference with knowledge distillation for visual dialog,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6056–6073, Oct. 2022.

[83] G. Li, Y. Zhai, Z. Lin, and Y. Zhang, “Similar scenes arouse similar emotions: Parallel data augmentation for stylized image captioning,” in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 5363–5372.

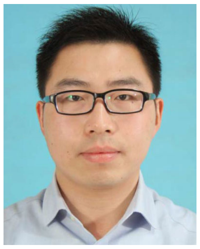
[84] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, “Visual semantic reasoning for image-text matching,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4654–4662.

[85] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “Guided open vocabulary image captioning with constrained beam search,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017.

[86] Y. Shi et al., “EMScore: Evaluating video captioning via coarse-grained and fine-grained embedding matching,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 17929–17938.

Peipei Song received the B.E. degree in electronic information engineering and the Ph.D. degree in signal and information processing from the Hefei University of Technology, China, in 2017 and 2023, respectively. She is currently a Postdoctoral Researcher with the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC). Her research interests include computer vision, natural language processing, and video understanding.

Dan Guo (Senior Member, IEEE) received the B.E. degree in computer science and technology from Yangtze University, China, in 2004, and the Ph.D. degree in system analysis and integration from the Huazhong University of Science and Technology, China, in 2010. She is currently a Professor with the School of Computer Science and Information Engineering, Hefei University of Technology, China. Her research interests include computer vision, machine learning, and intelligent multimedia content analysis. She regularly serves as a PC Member and for top-tier conferences and prestigious journals in multimedia and artificial intelligence, like ACM Multimedia, IJCAI, AAAI, CVPR, and ECCV. She also serves as a SPC Member for IJCAI 2021.



Xun Yang received the Ph.D. degree from the Hefei University of Technology, Hefei, China, in 2017. He is currently a Professor with the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC). From 2015 to 2017, he visited the University of Technology Sydney (UTS), Australia, as a Joint Ph.D. Student. He was a Research Fellow with the NExT++ Research Center, National University of Singapore (NUS), from 2018 to 2021. His current research interests include information retrieval, cross-media analysis and reasoning, and computer vision. He serves as an Associate Editor for IEEE Transactions on Big Data and *Multimedia Systems* journal.

Shengeng Tang received the B.E. degree in computer science and technology from Hunan Normal University, China, in 2017, and the Ph.D. degree in computer application technology from the Hefei University of Technology, China, in 2022. He is currently a Lecturer with the School of Computer Science and Information Engineering, Hefei University of Technology. His research interests include multimedia content analysis and computer vision.

Meng Wang (Fellow, IEEE) received the B.E. and Ph.D. degrees in the special class for the gifted young from the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively. He is currently a Professor with the Hefei University of Technology, China. He has authored over 200 book chapters, journals, and conference papers in his research areas. His current research interests include multimedia content analysis, computer vision, and pattern recognition. He was a recipient of the ACM SIGMM Rising Star Award in 2014. He is an Associate Editor of IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Circuits and Systems for Video Technology, and IEEE Transactions on Neural Networks and Learning Systems.