Emotion-oriented Cross-modal Prompting and Alignment for Human-centric Emotional Video Captioning

Yu Wang, Yuanyuan Liu*, *Member, IEEE*, Shunping Zhou, Yuxuan Huang, Chang Tang, *Senior Member, IEEE,* Wujie Zhou, *Senior Member, IEEE*, Zhe Chen, *Member, IEEE*

***Abstract*—Human-centric Emotional Video Captioning (H-EVC) aims to generate fine-grained, emotion-related sentences for human-based videos, enhancing the understanding of human emotions and facilitating human-computer emotional interaction. However, existing video captioning methods primarily focus on overall event content, often overlooking sufficient subtle emotional clues and interactions in videos. As a result, the generated captions frequently lack emotional information. To address this, we propose a novel Emotion-oriented Cross-modal Prompting and Alignment (ECPA) approach for large foundation models to enhance H-EVC accuracy by effectively modeling fine-grained visual-textual emotion clues and interactions. Using large foundation models, our ECPA introduces two learnable prompting strategies: visual emotion prompting (VEP) and textual emotion prompting (TEP), as well as an emotion-oriented cross-modal alignment (ECA) module. In VEP, we develop two-level learnable visual prompts, *i.e.*, emotion recognition (ER)-level and action unit (AU)-level prompting, to assist pre-trained vision-language foundation models to attend to both coarse and fine emotion-related visual information in videos. In TEP, we correspondingly devise two-level learnable textual prompts, *i.e.*, sentence-level emotional tokens, and word-level masked tokens, for obtaining both whole and local textual prompt representations related to emotions. To further facilitate the interaction and alignment of visual-textual emotion prompt representations, our ECA introduces another two levels of emotion-oriented prompt alignment learning mechanisms: the ER-sentence level and the AU-word level alignment losses. Both enhance the model's ability to capture and integrate both global and local cross-modal emotion semantics, thereby enabling the generation of fine-grained emotional linguistic descriptions in video captioning. Extensive experiments not only demonstrate that our ECPA outperforms existing state-of-the-art approaches on various H-EVC datasets (a relative improvement of 8.48% on MAFW and 10.58% on EmVidCap) by a large margin, but also support zero-shot tasks on two video captioning datasets (MSVD and MSRVTT), underscoring its applicability and generalization capabilities. The project page (including the dataset and demo) can be found in https://github.com/virtuesvvy/ECPA_Emotional_Video_Captioning.**

***Index Terms*—Emotional video captioning, learnable textual prompting, learnable visual prompting, cross-model prompt alignment, emotion-oriented prompt.**

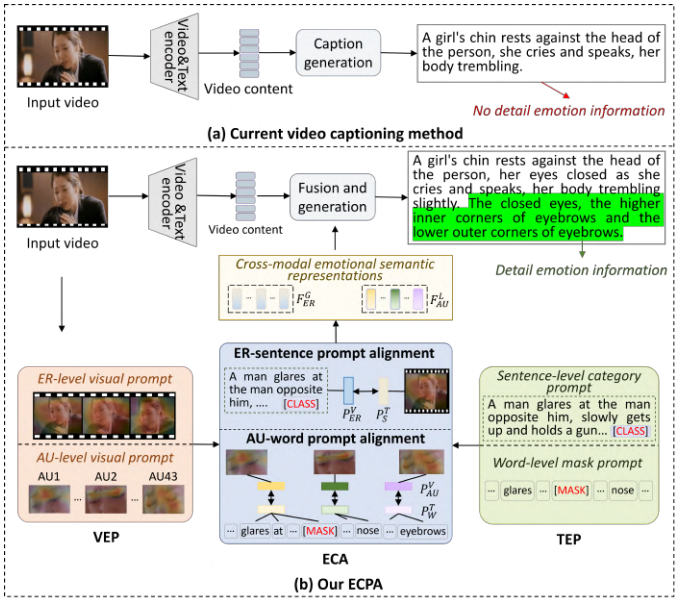

Fig. 1. Comparison of the proposed ECPA with the current video captioning frameworks. (a) The current video captioning method only generates event content descriptions without emotional details. (b) Our ECPA effectively generates human-centric emotion-linguistic sentences containing not only non-emotional event content but also fine-grained emotional information by introducing two learnable emotion prompting modules (VEP and TEP) and an emotion-oriented cross-modal alignment module (ECA). Note: [CLASS] represents the emotional classification of the entire sentence.

This work was supported by the National Natural Science Foundation of China grant (62076227), Natural Science Foundation of Hubei Province grant (2023AFB572) and Hubei Key Laboratory of Intelligent Geo-Information Processing (KLIGIP-2022-B10).

∗Corresponding author: Yuanyuan Liu.

Yu Wang is with the School of Computer Science and the School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China, (e-mail: vvy190701@cug.edu.cn).

Shunping Zhou, Yuxuan Huang, Chang Tang, and Yuanyuan Liu are with the School of Computer Science, China University of Geosciences, Wuhan 430074, China, (e-mail: zhoushunping, cosinehuang, tangchang, liuyy@cug.edu.cn).

Wujie Zhou is with the School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou, 310018, China, (e-mail: wujiezhou@163.com).

Zhe Chen is with Cisco-La Trobe Centre for AI and IoT, the School of Computing, Engineering and Mathematical Sciences, La Trobe University, Australia, (e-mail: zhe.chen@latrobe.edu.au).

I. Introduction

Human-centric Emotional Video Captioning (H-EVC) is a challenging sub-direction of video captioning, which aims to generate emotion-related linguistic sentence descriptions for human-based videos, incorporating both video event content and emotional semantic information [1]. H-EVC is a typical cross-modal generation task involving text and video modalities. It has been applied in various multimedia domains, such as human-computer interaction, emotion understanding,

and multimedia event analysis [2]–[4]. Existing technologies in H-EVC primarily use video captioning approaches. However, these methods easily generate linguistic sentences focused on overall event content rather than emotion-related nuanced semantic information, thus compromising the generation quality of detailed emotion-related captions [5], [6].

Current video captioning methods mainly fall into two categories: visual object extraction-based methods and visual-textual relation learning-based methods. The first category of methods emphasizes extracting different visual object features from videos to enhance the quality of generated captions [7], [8]. Zhang *et al.* [8] propose object-aware aggregation with bidirectional temporal graphs, which captures detailed temporal dynamics for salient objects to automatically generate natural language descriptions of video content. The other category mainly explores the overall relationship between video and text by introducing video-text alignment algorithms [9]. For example, Ye *et al.* [10] bridge video representations and linguistic semantics at three levels, entity, predicate, and sentence, to generate accurate and meaningful captions. Despite advances in video captioning, most existing methods primarily focus on the content of the video event, resulting in captioning sentences that lack emotional information [11]. As indicated in Fig. 1(a), the current video captioning method fails to generate emotion-related descriptions for the human-based video.

Recently, to improve the capability of modeling emotional information in videos, some approaches have introduced the emotion classifier to identify video emotion categories for emotional video captioning. For instance, Wang *et al*. [5] first employ two separate seq2seq modules for video factual content and emotion captioning, respectively. Song *et al*. [12] further propose a unified Contextual Attention Network (CANet) to learn factual content and emotion categories in videos. EPAN [6] identifies emotion categories to guide video caption generation. Despite the progress made in H-EVC, current approaches mainly introduce global video emotion categories without considering fine-grained cross-modal emotion clues and semantic interaction, resulting in sub-optimal performance. Furthermore, recent developments in large foundation models have been developed to enhance the task, such as GPT4 [13]. However, these models still struggle to capture valid sentiment information due to lacking pre-training on large-scale emotional data.

To address the issue mentioned above, inspired by the prompt learning (PL) that can help to efficiently capture downstream task-detailed knowledge in large foundation models [14], we propose a novel **E**motion-oriented **C**ross-modal **P**rompting and **A**lignment (ECPA) approach to augment the H-EVC performance by effectively modeling cross-modal fine-grained emotion semantic clues and interaction. Fig. 1 illustrates the motivation and innovation results of our proposed ECPA. Unlike existing methods, ECPA generates comprehensive linguistic sentences integrating event content with fine-grained emotional semantics through two learnable prompting schemes: visual emotion prompting (VEP) and textual emotion prompting (TEP), along with an emotion-oriented cross-modal alignment (ECA) module. More specifically, as illustrated in Fig. 2, we first use the pre-trained large foundation models as the visual encoder and text encoder to extract visual and textual embedding features. Then, in our VEP, we devise two-level learnable visual prompting schemes to obtain emotion-related visual information. Meanwhile, for our TEP, we also introduce two-level learnable semantic prompt representations to help better depict textual emotion semantics. Furthermore, the ECA module enhances fine-grained emotional interaction between visual and textual emotion prompts by introducing another two-level cross-modal emotion alignment losses, ultimately improving the generation of video descriptions with fine-grained emotional semantics.

Overall, our ECPA offers an effective and generalized approach to augment current video captioning technologies for robust H-EVC performance. It adeptly models and fuses video content with human emotion semantic cues for H-EVC. The major contributions of this study are summarized as follows:

(1) We propose a novel ECPA approach to enhance the generation quality of fine-grained emotional linguistic descriptions for human-centric videos. To the best of our knowledge, we are the first to explore the H-EVC task using cross-modal prompting and alignment.

(2) We devise VEP with two-level learnable visual prompts related to human emotions, *i.e.*, ER- and AU-level, to enhance the pre-trained visual foundation models in extracting both global and local visual emotion features. Meanwhile, we introduce TEP with sentence- and word-level learnable tokens for two-level textual prompting representations, thus depicting human emotion semantic cues from text inputs.

(3) We propose a two-level emotion-oriented alignment learning mechanism, ECA, to facilitate the cross-modal prompt interaction at the ER-sentence and AU-word level, respectively, by introducing a cross-model ER-aware alignment loss and a cross-model AU-aware alignment loss. This further enhances ECPA in capturing and integrating both global and local emotional semantics for fine-gained emotional captioning.

(4) Experimentally, inspired by previous video captioning tasks, we define experimental settings of H-EVC, and conduct extensive experiments on two challenging H-EVC datasets, namely EmVidCap and MAFW. Our approach achieves new state-of-the-art results, with average relative improvements of 8.43%, 5.01%, 3.33%, and 17.02% on MAFW and 12.82%, 20.27%, 4.23%, and 5.01% on EmVidCap, across four various evaluation metrics (BLEU-4, METEOR, ROUGE-L, and CIDER). In addition, we evaluate our ECPA for two zero-shot tasks on MSVD and MSRVTT video captioning datasets, demonstrating the generalization of our approach.

II. Related Work

In this section, we briefly review existing works: video captioning, emotional video captioning, and prompt learning in vision-language models.

A. Video Captioning

Video captioning is a challenging task involving generating open-form video descriptions. Initial approaches leverage

the encoder-decoder framework, using pre-trained CNNs for feature extraction and RNNs for caption generation [8], [15]. For comprehensive feature encoding, Yao *et al.* [15] introduce a temporal attention mechanism to better summarize visual sequences. Subsequent research incorporates spatial attention to identify crucial visual regions for encoding fine-grained visual features [8]. Video Swin Transformer enhances attention computation in both spatial and temporal domains [16]. As the field has evolved, there's an increasing emphasis on cross-modal information, bridging vision and language in video captioning [17]. For example, Aafaq *et al.* [18] exploit the fusion of semantic and visual content, enabling it to generate semantically meaningful event proposals. Seo *et al.* [19] jointly train a multimodal video encoder and a sentence decoder for multimodal video captioning. The latest SwinBERT [20] is positioned as the first end-to-end model for comprehensive video-text interaction. Yet, there's a noticeable gap as none of these models highlight the emotion-oriented salients.

B. Emotional Video Captioning

In the quest to understand emotional expressions in videos, many studies have focused on predicting emotional contexts in user-generated content. Facial expressions have emerged as a vital element for emotion recognition [21]. Yet, generating emotional expressions for video descriptions remains a less-explored area. Wang *et al*. [5] pioneer this new task and release the EVC dataset, EmovidCap. Their Fact Transfer (FT) framework comprises two distinct modules for facts and emotions, respectively. This misses nuanced links between emotions and events, resulting in misleading outcomes. Addressing this gap, recent advancements in multimodal data fusion have been pivotal. For instance, Wang *et al*. [22] apply a multimodal fusion model to generate abstractive emotion causes in conversations based on text, audio, and vision modalities. In addition, Song *et al*. [12] present a unified multi-model that harnesses contextual data from both video and text to create emotional video captions. Considering the importance of emotional understanding in EVC, emotion is first perceived and then used to guide caption generation in the emotion-prior awareness network (EPAN) [6]. Liu *et al*. [23] create a fine-grained H-EVC dataset encompassing detailed descriptions of emotion-related contents, including face Action Units (AUs), setting the stage for more refined tasks. However, current researchers still have some drawbacks: (1) they often provide simple summaries of the emotional atmosphere captured in a video, making it difficult to offer detailed emotion descriptions, such as facial Action Units (AUs); (2) the lack of visual and textual emotion interactive fusion leads to their inability to address the emotional gap between the video and text.

C. Prompt Learning in Vision-Language Models

In recent years, integrating prompts from the NLP domain into Computer Vision (CV) has gained traction, particularly for cross-modal fusion techniques that aim to integrate information from multiple modalities, such as vision and language [24]. These methods ensure that interactive fusion between visual cues and linguistic descriptions is well-coordinated and mutually informative. A notable example is the Context Optimization (CoOp) [7], which leverages continuous learnable textual prompts in bridging NLP and CV techniques. Li *et al.* [25] further align cascaded text semantic prompts with fine-grained visual regions. Additionally, visual-prompt tuning (VPT) takes a deeper exploration into visual prompting [26]. To address the lack of cohesive interaction between vision and language prompts, MaPLe [10] introduces multimodal prompts to enhance the consistency of visual and textual representations. Furthermore, Xu *et al.* [27] propose a method that translates images into captions to serve as context prompts (COP) and introduces hybrid emotion prompts (HEP) from the interaction between visual and textual information.

While progress has been made, a prominent limitation is that the existing emotional prompts are derived from the overall visual and textual emotional information, resulting in sub-optimal H-EVC. Motivated by this observation, our study explores the emotion-aware interaction between vision and textual prompts to formulate cross-modal emotion prompts, for fine-grained emotional captioning.

III. Proposed Method

A. Problem Definition

We first define the problem of the H-EVC task. In H-EVC, the training set with N samples is defined as $D_{train} = (V_i, T_i)_{i=1}^{N}$, where i indexes the samples, V_i represents the i-th video, $T_i = \{w_j\}_{j=0}^{J}$ is the related textual captions of V_i, w_j is the j-th word in the sequence T_i, and J is the number of word Tokens. Therefore, the test set only comprises videos representing $D_{test} = \{V_l\}$, where l indexes over the testing samples. The aim of H-EVC is to generate fine-grained, emotion-related linguistic descriptions for the video V_l, including both event content and human-centric emotional semantic information. As shown in Fig 2(a), current methods for H-EVC mainly consist of two stages, *i.e.*, task-related feature embedding, and video captioning generation [20].

1) Task-related Feature Embedding: This process is crucial for transforming raw data into meaningful features that are directly relevant to the specific tasks our model addresses.

Visual encoder. Given the video input V_i, a visual feature encoder $VE()$ extracts task-related visual embedding features F_V, formulated as $F_V = VE(V_i, \theta_V) \in \mathbb{R}^{M \times d_v}$, where M is the number of tokens and d_v is the feature dimension per token. θ_V are the trainable parameters in the encoder. Notably, various large visual foundation models, such as VidSwin [16] or CLIP-based visual encoder [28], can serve as $VE()$.

Textual encoder. For text sequence input T_i, a text encoder $TE()$ also encodes T_i into task-related visual embedding features, as $F_T = TE(T_i, \theta_T) \in \mathbb{R}^{J \times d_t}$, where J is the number of word tokens and d_t is the dimension of each token. The parameters θ_T are learnable during training. The textual encoder $VE()$ can utilize various large language models, such as BERT [29], CLIP [28], T5 [30] or GPT2 [31].

To balance computational complexity, we use VidSwin [16] for visual feature embedding and BERT [29] for textual feature embedding. We also evaluate the performance of the video

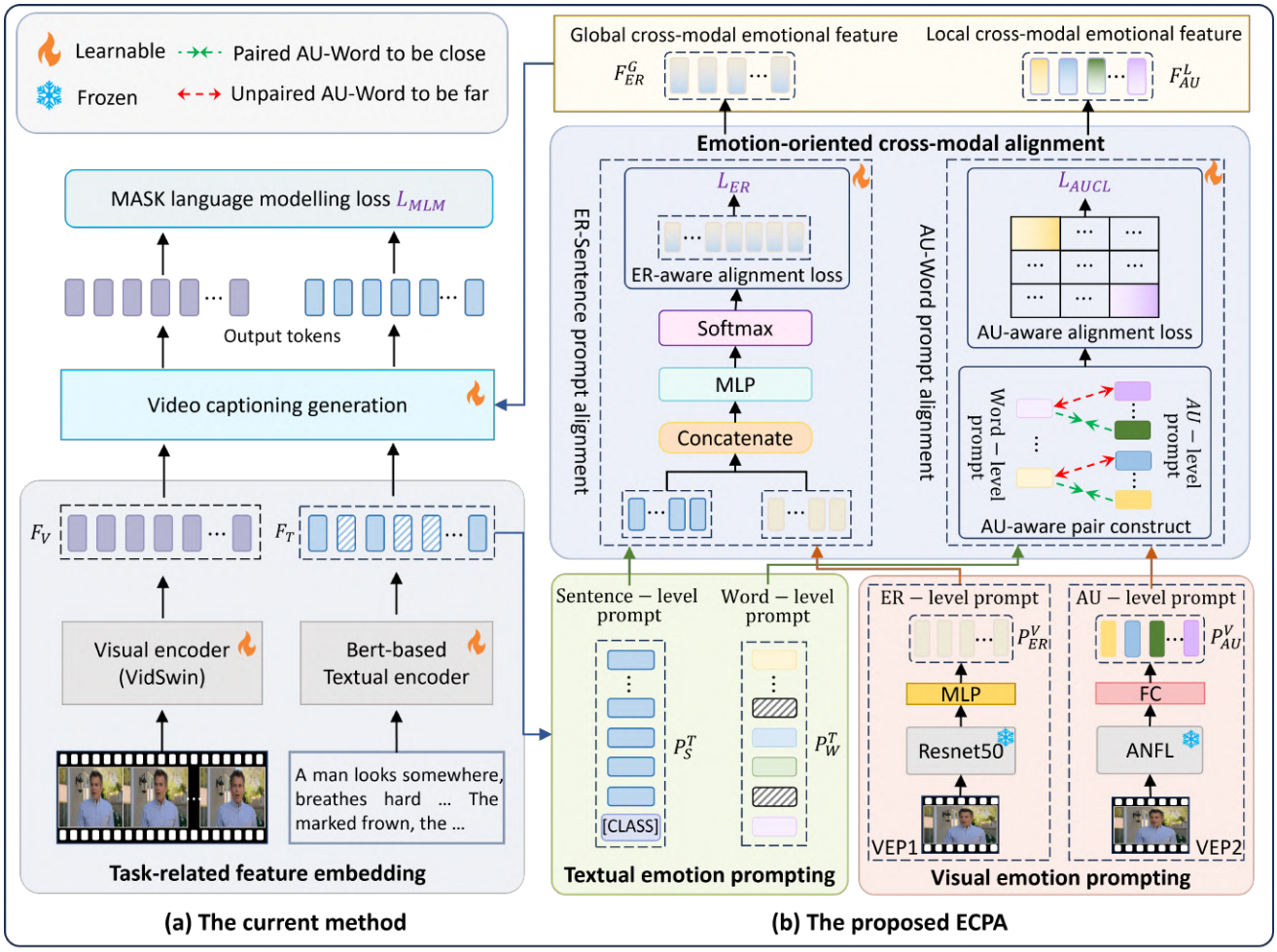

Fig. 2. The training pipeline of our ECPA for H-EVC. With visual and textual feature embedding, ECPA first combines visual emotion prompting (VEP) and textual emotion prompting (TEP) to enhance the pre-trained visual and textual foundation models in extracting both global and local emotion-related prompt representations. Then, an emotion-oriented cross-modal alignment (ECA) module is proposed to facilitate the interaction and alignment of the learned cross-modal prompts, thus improving H-EVC performance for generating both video content and fine-grained human-centric emotion linguistic sentences. Note: different color rectangles inside prompts represent different local tokens, *e.g.*, different AU vectors in VEP or different word tokens in TEP, while different textures of rectangles indicate the fused global and local cross-modal emotional feature in ECA, respectively.

captioning framework with various foundation models, as shown in Table VII of Section IV-E7.

2) Video Captioning Generation: Using the extracted F_V and F_T, a fusion module, typically a Transformer decoder [20], [32] with an MLM loss [33], is employed to integrate these cross-modal features for linguistic sentence generation in seq2seq generation manner. As a result, the caption generation can be written as,

$$T_l = \mathcal{T}rans(F_V, F_T), \tag{1}$$

where T_l is the generated linguistic descriptions.

Although existing models work well in some special scenarios, their overall performance is still limited due to the lack of effective mechanisms for modeling fine-grained emotional semantics and interaction between videos and textual captions. To address this issue, we propose a novel emotion-oriented cross-modal prompting and alignment approach, termed ECPA. This method aims to augment the H-EVC performance by effectively modeling cross-modal fine-grained emotional semantic clues and their interactions within foundation models through a hierarchical structure.

B. Overview of ECPA

Fig. 2 illustrates the overview of how our proposed ECPA enhances the current video captioning framework for the H-EVC task. Specifically, ECPA consists of visual emotion prompting (VEP), textual emotion prompting (TEP), and an emotion-oriented cross-modal alignment module (ECA). First, with the video V_i, its textual caption T_i, and the textual embedding features F_T as inputs, we introduce the VEP to provide both global and local emotion-related visual prompt representations to enhance the visual embedding features. Meanwhile, we devise the TEP, which introduces two-level learnable semantic prompt representations, namely sentence-level and word-level tokens, designed to better depict human emotion semantics for textual embedding features. Formally, we mainly employ the Eq. (2) to obtain cross-modal emotion

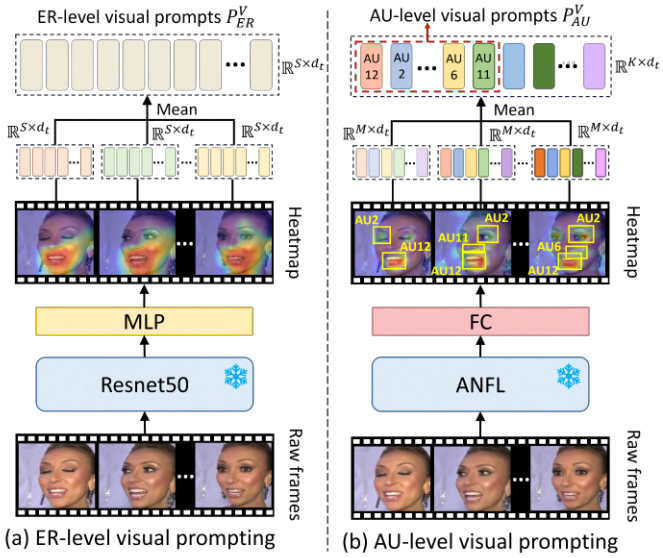

Fig. 3. Detailed pipeline of VEP. (a) ER-level visual prompting, (b) AU-level visual prompting. Note: heatmaps display the captured ER- and AU-level emotion-sensitive information, which effectively complements global and local visual emotion cues to enhance visual embedding features.

prompt representations enhanced by the VEP and TEP:

$$P_{ER}^V, P_{AU}^V = VEP(V_i), \quad P_S^T, P_W^T = TEP(T_i, F_T), \quad (2)$$

where P_{ER}^V, P_{AU}^V, P_S^T, and P_W^T are visual and textual emotion prompting representations at two-level prompting learning, respectively.

Subsequently, the ECA, incorporating a global ER-aware alignment loss and a local AU-aware alignment loss, is proposed to improve visual and textual prompt interaction and alignment at various levels, namely ER-sentence and AU-word levels, thus obtaining effective cross-modal emotional semantic representations. Formally, we formulate the ECA as,

$$F_{ER}^G, F_{AU}^L = ECA(P_{ER}^V, P_{AU}^V, P_S^T, P_W^T), \quad (3)$$

where F_{ER}^G and F_{AU}^L represent global and local cross-modal emotional semantic representations, respectively. Finally, we mainly recast the Eq.1 by incorporating these global and local cross-modal emotional semantic representations, effectively enhancing the generation of fine-grained emotional linguistic sentences for H-EVC.

In the following sections, we will provide detailed explanations of the VEP, TEP, and ECA, respectively.

C. Visual Emotion Prompting

To mine fine-grained emotion-related information from the visual embedding features, we devise a two-level learnable visual emotion prompting (VEP) mechanism comprising ER- and AU-level visual prompting, as illustrated in Fig. 3. Rather than hand-crafted visual prompts [34], [35], such as fixed-size or random image perturbation patches, our designed learnable prompts allow the visual encoder to attend to emotion-sensitive areas, thus better adapting to the H-EVC task.

1) ER-level Visual Prompting: The ER-level visual prompting is designed to acquire global visual emotion knowledge from videos, thereby enhancing the visual embedding features for emotion-related representations. Specifically, with a video input V_i, we first use a pre-trained Resnet50 [36]-based ER classifier, followed by a Multi-Layer Perception (represented as $MLP()$) as the first-level $VEP_1()$ to obtain the global learnable ER-level visual prompt P_{ER}^V, Formally, the ER-level visual prompting is given by:

$$P_{ER}^V = VEP_1(V_i) = MLP(Resnet50(V_i)). \quad (4)$$

2) AU-level Visual Prompting: Studies indicate that facial action units (AUs) effectively encode subtle emotion-related patterns for emotion recognition [37]. To capture subtle and finer visual emotion changes in videos, we devise a learnable AU-level visual prompting module, enhancing the visual embedding features with fine-grained emotional representations. Formally, for video input V_i, we introduce a pre-trained AUs Relationship-aware Node Feature Learning network (ANFL) [38], with a fully-connected (FC) layer as the second-level $VEP_2()$, to model local visual emotion changes, forming fine-grained visual prompt representations for human emotions. This procedure is described as:

$$P_{AU}^V = VEP_2(V_i) = FC(ANFL(V_i)), \quad (5)$$

where $P_{AU}^V = \{a_i\}_{i=1}^K \in \mathbb{R}^{K\times 768}$ represents the extracted AU-level visual prompts. a_i is the i-th AU feature vector extracted by the ANFL and K is the number of selected AUs in the video. To focus on the most reliable AUs, we select TOP-K AUs with the highest confidence scores. In the experimental part, we also discuss the effects of various K (see Fig. 6) and set the K to 3 for the best performance.

It is worth noting that the parameters of both Resnet50 and ANFL are frozen; we only adjust the parameters of our learnable prompts P_{ER}^V and P_{AU}^V. With proper training, both ER- and AU-level visual prompts can be effectively learned, capturing global and local emotional information, respectively. These two-level prompts provide emotion-sensitive cues that complement the visual embedding features.

D. Textual Emotion Prompting

1) Sentence-level Category Prompting: For the sentence-level category prompting, we introduce specific emotional tokens denoted as [CLASS-k] to depict the k-th emotion category. The [CLASS-k] is an emotional phrase, such as "happiness" or "sadness". Usually, manual-designed textual prompts use a fixed template like "a picture of X", where "X" represents a category label. However, this approach is inadequate for describing videos with human emotions. As a result, we instead introduce learnable emotion-specific tokens as the first-level $TEP_1()$ to instantiate P_S^T.

Formally, using the textual embedding features F_T of the whole sentence, we combine the special emotional token [CLASS-k] with F_T, forming learnable textual prompt representations. We define the sentence-level prompting as:

$$P_S^T = TEP_1(F_T) = \{F_T \oplus [\text{CLASS-k}]\} \quad (6)$$

where $\oplus$ attaches together F_T, and [CLASS-k] prompts that it is an emotion classification task.

2) Word-level Mask Prompting: To learn finer word-level emotion prompting representations, TEP further introduces a Masked Language Modeling (MLM) [33]-based local language prompting. Formally, with the text embedding $F_T = \{t_j\}_{j=1}^{N}$, we randomly mask some tokens in F_T, formulated as,

$$P_W^T = TEP_2(F_T) = \{t_1, t_2, [\text{MASK}], t_j, \ldots, [\text{MASK}], \ldots, t_N\}, \tag{7}$$

where [MASK] represents the tokens set to zero, and t_j is the j-th word vector in the embedding feature F_T.

E. Emotion-oriented Cross-modal Alignment

To facilitate effective emotion interaction of visual and textual prompt representations for the H-EVC task, a cross-modal alignment mechanism is crucial. Existing methods that directly splice or fuse these modalities often fail to achieve fine-grained emotion interaction due to insufficient emotion alignment. To this end, we propose a novel Emotion-Oriented Cross-Modal Alignment (ECA) module, a two-level alignment learning mechanism that enhances emotional-semantic interactions between visuals and text both globally and locally by exploiting the intrinsic connections between emotional expressions, as shown in Fig.2(b). The first level alignment, ER-sentence prompt alignment, primarily focuses on the global emotion interaction between the visual and textual prompt representations by introducing a global ER-aware alignment loss. The second level, AU-word prompt alignment, attends to more fine-grained emotional interactions between these representations by introducing a local AU-aware alignment loss. The effective surgery further enhances the ECPA's ability to capture both global and local cross-modal emotional semantic interaction for fine-gained emotion captioning.

1) ER-Sentence Prompt Alignment: Instead of directly combining visual with textual prompting branches, the ER-sentence prompt alignment achieves the global emotion alignment and interaction between them. To this end, we introduce a global ER-aware alignment loss L_{ER} to align the global emotion semantic between the ER-level visual prompt P_{ER}^V and the sentence-level textual prompt P_S^T, thus obtaining the enhanced global cross-modal emotional semantic representations F_{ER}^G. Formally, we first concatenate the P_{ER}^V and P_S^T, represented as, $F_{ER}^G = Concat(P_{ER}^V, P_S^T)$. We then employ a cross-entropy loss CE as the L_{ER} to facilitate effective gradient propagation and mutual synergy between the visual and textual prompt representations. This is be written as,

$$L_{ER} = -\frac{1}{N_b}\sum_{i=1}^{N_b} CE(\text{softmax}(F_{ER}^G), y_i) \tag{8}$$

where N_b denotes the number of samples in the mini-batch and softmax(F_{ER}^G) indicates the predicted probability assigned to the true emotion class y_i for the i-th example by the softmax function. Through minimizing the L_{ER}, we can align mutual visual and textual prompts to achieve the global cross-modal emotion representations F_{ER}^G in Eq. 3.

2) AU-Word Prompt Alignment: Furthermore, to achieve fine-grained cross-modal emotion interaction between AU-level visual and word-level textual prompts, we further propose an AU-word prompt alignment mechanism to uncover local, finer cross-modal emotion representations. To this end, we introduce an AU-aware alignment loss, which helps align the subtle relationships between AU cues in videos and semantic cues in sentences. The detailed pipeline of AU-word prompt alignment is illustrated in Fig. 4.

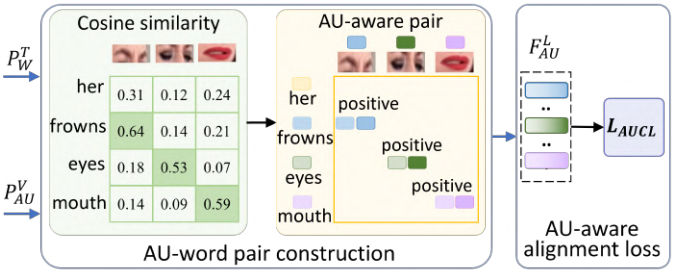

Fig. 4. Pipeline of the AU-word prompt alignment.

Formally, given the obtained AU-level visual prompts $P_{AU}^V = \{a_i\}_{i=1}^{K}$ and word-level MASK textual prompts $P_W^T = \{t_1, t_2, [\text{MASK}], t_j, \ldots, [\text{MASK}], \ldots, t_N\}$, we first use the local cosine similarity to calculate the similarity between per AU and word,

$$cos(a_i, t_j) = \frac{a_i \cdot t_j}{\|a_i\|_2 \cdot \|t_j\|_2}, \tag{9}$$

where $\|\|_2$ represent the L2-norm of vectors. Then, the AU-aware alignment loss is introduced to obtain the finer alignment of each AU prompt a_i with the corresponding word prompt t_j, by maximizing the similarity $cos(a_i, t_j)$. To achieve this, the learning pipeline of the AU-word prompt alignment mechanism contains two main steps below.

AU-word pair construction. To facilitate AU-word prompt alignment, we first construct positive/negative AU-word alignment pairs. Unlike current contrastive learning methods that rely on data augmentation for positive/negative pair construction [39], our approach introduces a novel positive/negative AU-word alignment pair construction method. This new method enhances the alignment between AUs and corresponding textual prompts, thereby improving the model's understanding and representation of facial expressions in a cross-modal context.

With the cross-modal prompts (P_{AU}^V, P_{AU}^W), we introduce the Hungarian [40] method to define the AU-word similarity between each AU and word in the cross-modal prompts. This aims to maximize the alignment of fine-grained emotional semantics and visual changes. We calculate the AU-word similarity as follows:

$$D(P_{AU}^V, P_{AU}^W) = \text{argmax}\left(\sum_{k=1}^{K} \cos(a_i, t_j)\right), \tag{10}$$

where i and j index the sample over P_{AU}^V and P_{AU}^W, respectively.

Using the obtained $D(P_{AU}^V, P_{AU}^W)$, we introduce a threshold-based procedure to construct negative/positive AU-word alignment pairs for achieving AU-word prompt alignment. The process can be formulated as,

$$Y = \begin{cases} 1 & \text{if } D(P_{AU}^V, P_{AU}^W) > T_C, \\ 0 & \text{otherwise.} \end{cases}, \tag{11}$$

where T_C is the pre-defined threshold and Y is the label of the AU-word pair. If $D(P_{AU}^V, P_{AU}^W) > T_C$, we consider the pair as the positive AU-word pair, assigning $Y = 1$; otherwise, we categorize it as the negative AU-word pair with $Y = 0$. In our study, we experimentally set the threshold T_C to 0.3. More discussions on the threshold setting can be seen in Fig. 7 of the experimental part.

AU-aware alignment loss. With the constructed positive/negative AU-word pairs, we devise an AU-aware alignment loss as,

$$\begin{aligned} L_{AUCL}(Y, P_{AU}^V, P_{AU}^W) = (1-Y)\big(D(P_{AU}^V, P_{AU}^W)\big)^2 \\ + Y\big(\max(0, m - D(P_{AU}^V, P_{AU}^W)\big)^2, \end{aligned} \tag{12}$$

where $m = 0.5$ is a margin that sets the maximum distance between the positive and negative pairs in the feature space [41]. Through maximizing the cosine similarity between the positive pair while minimizing the negative pair, the optimization of L_{AUCL} helps to facilitate the fine-grained alignment between per AU and word prompt vector, thus obtaining the locally cross-modal emotion representations F_{AU}^L in Eq. 3.

In general, introducing the ECA module enhances both global and local interaction of visual-textual cross-modal prompting, enhancing both global and local cross-modal emotion representations. This improvement facilitates robust H-EVC performance.

F. Emotion-oriented Content Generation and Overall Objective

1) Emotion-oriented Content Generation: For H-EVC performance, we refer to video captioning and employ a widely-used seq2seq mechanism [20] to generate linguistic descriptions for input videos. We use a cross-modal Transformer encoder, denoted as $Trans(\cdot)$, to ultimately fuse the obtained global and local cross-modal emotion representations, F_{ER}^G and F_{AU}^L, with the original embedding vectors, F_V and F_T. In this study, the Transformer consists of 12-layer encoders, each with 12 self-attention heads. This process aims to generate a comprehensive emotion-oriented content representation F_E for description generation, as delineated in the following equation:

$$F_E = \mathcal{T}rans\Big(k/q/v = Concat(F_V, F_T, F_{ER}^G, F_{AU}^L)\Big). \tag{13}$$

To enable the $\mathcal{T}rans(\cdot)$ to learn and generate the emotion-oriented content semantic descriptions from the masked words, we introduce an MLM loss function [33]. Formally, the MLM loss is defined as follows,

$$L_{MLM} = -\frac{1}{|M|}\sum_{r\in|M|} log(\frac{exp(p_{r,y}))}{\sum_{j=1}^{G} exp(p_{r,q})}), \tag{14}$$

where $|M|$ denotes the number of masked words in a sentence, r indexes each individual masked token in the sequence, and G represents the vocabulary size in the training set. $exp()$ denotes the exponential function operation. $p_{r,y}$ is the predicted score for the r-th correct label of the masked word, and $p_{r,q}$ is the predicted score for the q-th vocabulary item.

2) Overall Learning Objectives: In summary, our ECPA includes three learning objectives, namely, global ER-aware alignment loss L_{ER}, local AU-aware alignment loss L_{AUCL}, and MLM generation loss L_{MLM}. Mathematically, the total loss function can be defined as:

$$L = \lambda_{ER}L_{ER} + \lambda_{AUCL}L_{AUCL} + \lambda_{MLM}L_{MLM}, \tag{15}$$

where λ_{ER}, λ_{AUCL}, λ_{MLM} are dynamic weights for each loss component x. Inspired by [42], Dynamic weights average (DWA) facilitates interaction between different loss functions by dynamically adjusting the weights based on the relative loss reduction rates. Specifically, for each loss component x at the training step t, we calculate the relative loss reduction rate $r_x(t) = L_x(t)/L_x(t-1)$ and update the weights using:

$$r_x(t) = \frac{exp(\frac{r_x(t)}{\tau})}{\sum_{x=1}^{3} exp(\frac{r_x(t)}{\tau})} \tag{16}$$

where τ is a temperature parameter, and the total number of loss components is 3. This approach ensures balanced training and effective integration of various objectives by prioritizing slower-learning tasks with higher weights.

G. Inference

In the inference phase, given the learnable global and local visual prompt representations, F_{ER}^G and F_{AU}^L, and the visual embedding features F_V as inputs, we first employ the cross-modal Transformer encoder to obtain the emotion-oriented content representation F_E. Then, using the F_E, we employ a seq2seq generator [42] with an auto-regressive decoder to sequentially generate continuous word tokens, ultimately obtaining fine-grained, high-quality emotion-related linguistic sentences.

IV. Experiments and Analysis

A. Dataset

To demonstrate the effectiveness of our method, we used two popular H-EVC datasets, MAFW [23] and EmVidCap [5]. And performed the zero-shot evaluation on two widely used video captioning datasets, MSVD [43] and MSRVTT [44].

1) MAFW: MAFW [11] is the first multi-label emotion database containing fine-grained emotion descriptive texts and 11 single emotion categories. These captions detail information about the environment, body movements, AUs, and other emotional elements. The MAFW dataset consists of 6,423 training videos and 1,611 validation videos.

2) EmVidCap: EmVidCap [5] is an emotion video captioning dataset containing 1,897 videos and 41,009 captions with one or more sentiment words. For the H-EVC task. We selected 526 clear-faced videos from EmVidCap for our EVC task, 362 for training, and 164 for testing.

TABLE I
THE IMPORTANT ARCHITECTURE DETAILS AND PARAMETERS OF CORE COMPONENTS IN THE H-EVC FRAMEWORK

Parameters	Feature encoders					Emotion prompting		Multimodal
	VidSwin	CLIP	BERT	T5	GPT2	VEP	TEP	Transformer
Number of layers	12	-	12	24	12	-	-	6
Attention heads	8	-	12	16	12	-	-	8
Hidden size	768	512(T), 1024(V)	768	1024	768	-	-	512
Output dimension	768	512(T), 1024(V)	768	1024	768	768	768	512

3) MSVD: MSVD [43] is a video captioning dataset without emotional captions. It has 1,970 videos with around 40 captions each, totaling approximately 80,000 video-description pairs. MSVD is divided into a training set of 1,200 videos and a test set of 670 videos.

4) MSRVTT: MSRVTT [44] is an open-domain video captioning dataset. Each video clip has 20 ground-truth captions. We use the standard captioning split [20], which has 6.5K training videos and 2.9K testing videos.

B. Evaluation Protocols

Following existing methods [12], we used four standard semantic metrics to evaluate our method, namely BLEU-4 (B-4) [45], METEOR (M) [46], ROUGE-L (R-L) [47] and CIDEr (C) [48]. Briefly, BLEU-4 evaluates 4-gram overlap, METEOR checks word and synonym matches, ROUGE-L focuses on the longest common subsequence, and CIDEr assesses vocabulary relevance. All metrics match the code released on the Microsoft COCO evaluation server [49]. Following the convention [6], we also consider the emotion evaluation with emotion classification accuracy.

C. Implementation Details

We use the Pytorch [50] and DeepSpeed [51] libraries to implement our model. The experiments are conducted on a computer with an AMD Ryzen 9 5950X processor at 3.40 GHz, 64 GB of RAM, and an NVIDIA GeForce GTX 3090 Ti. We use the AdamW optimizer [50] with a learning rate warm-up strategy during the first 10% of the training phase, followed by linear decay. Our cross-modal Transformer has 12 layers with an implicit layer size of 512. We resize the short edge of all video frames to 224. During training, frames are randomly cropped to 224 × 224. In inference, they are center-cropped to the same size. As shown in Table I, we provide the important architecture details and parameters of the feature encoders, VEP, TEP and the multimodal Transformer used in our H-EVC framework, including the number of layers, attention heads, hidden size and output dimension of each module.

D. Overall performance on H-EVC

1) Results on the MAFW dataset: To comprehensively evaluate the performance of the proposed ECPA for H-EVC, we compared ECPA with several state-of-the-art video caption description techniques on the MAFW dataset, including Pos+VCT [52], RecNet$_{local}$ [53], POS+CG [54], SAAT [55], TDPC [56], SMAN [57], SWINBERT [20], Knight [58] and VLTinT [32]. As shown in Table II, compared to the method with the second best overall results, *i.e.* SWINBERT, our method demonstrates the best performance in terms of relative improvement of 8.43%, 5.01%, 3.33%, and 17.02%, in four evaluation metrics, namely BLEU-4, METEOR, ROUGE-L and CIDER, respectively. The generated texts effectively capture key relevant semantic information, *e.g.*, AU phrases (higher CIDER score), while not merely using the same words (smaller rise in ROUGE-L).

TABLE II
COMPARED WITH OTHER METHODS ON THE MAFW DATASET. THE BEST RESULTS ARE IN BOLD, AND THE SECOND-BEST ARE UNDERLINED

Method	Semantic metrics (%)			
	B-4	M	R-L	C
Pos+VCT [52]	6	14.3	25.06	19.85
RecNet$_{local}$ [53]	6.87	11.5	31.93	25.63
POS+CG [54]	4.85	13.46	23.27	15.41
SAAT [55]	4.32	<u>15.73</u>	20.23	3.31
TDPC [56]	9.09	15.49	24.73	23.4
SMAN [57]	9.2	14.3	32.7	20.3
SWINBERT [20]	<u>10.32</u>	15.57	38.11	<u>27.61</u>
Knight [58]	2.56	9.06	17.73	12.57
VLTinT [32]	9.81	15.38	<u>38.25</u>	25.75
ECPA(Our)	**11.35**	**16.46**	**39.81**	**34.38**

2) Results on the EmVidCap Dataset: The comparison results presented in Table III show that our method is still superior to other algorithms on the EmVidCap dataset. We evaluated our method by comparing its performance with the reproduced results of RecNet$_{local}$ [53], POS+CG [54], SAAT [55], TDPC [56], SMAN [57], SWINBERT [20] and Knight [58]. As indicated in Table III, our approach outperforms all other methods and has better accuracies in all four evaluation metrics. Our method outperforms SWINBERT by 12.82%, 20.27%, 4.23%, and 5.01% in BLEU-4, METEOR, ROUGE-L, and CIDER, respectively. These outcomes highlight the ability to produce accurate and relevant text, *i.e.*, emotional words (see BLEU-4 and METEOR results), while also ensuring some level of fluency and semantic relevance that ROUGE-L and CIDEr measure.

TABLE III
COMPARED WITH OTHER METHODS ON THE EMVIDCAP DATASET. THE BEST RESULTS ARE IN BOLD, AND THE SECOND-BEST ARE UNDERLINED

Method	Semantic metrics (%)			
	B-4	M	R-L	C
RecNet$_{local}$ [53]	13.82	15.49	34.54	17.37
POS+CG [54]	7.33	12.03	29.40	4.40
SAAT [55]	7.21	10.21	27.37	7.47
TDPC [56]	5.93	8.88	26.91	2.83
SMAN [57]	15.31	15.60	34.60	14.19
SWINBERT [20]	<u>19.26</u>	17.96	<u>36.97</u>	<u>19.36</u>
Knight [58]	14.69	<u>18.25</u>	32.54	16.11
ECPA(Our)	**21.73**	**21.60**	**38.53**	**20.33**

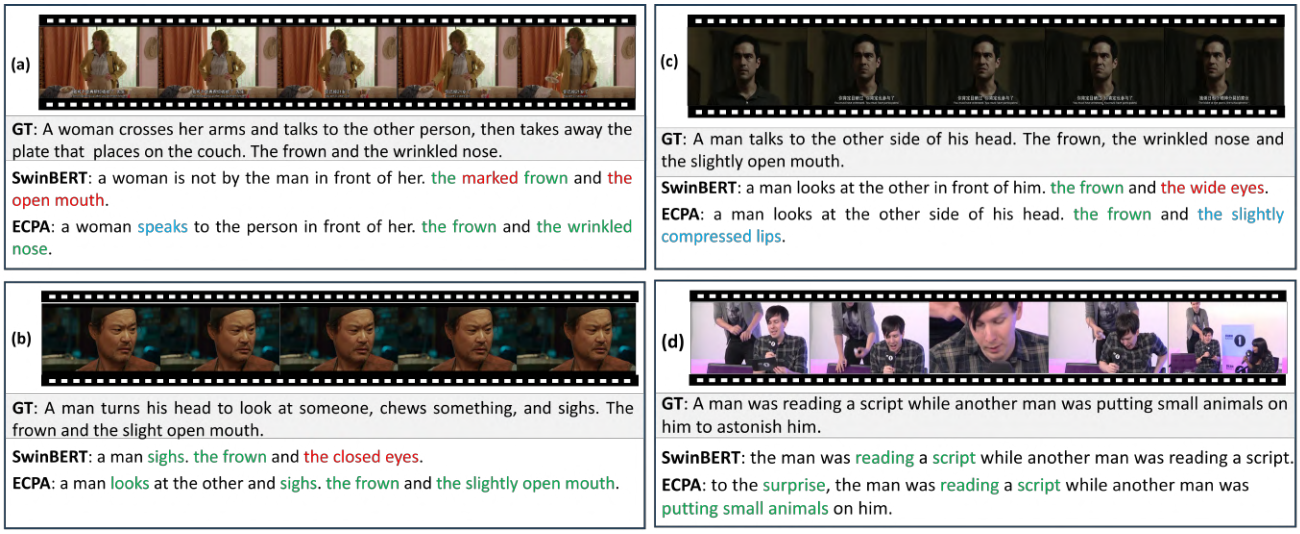

Fig. 5. Captions generated by SWINBERT [20] and ECPA on the (a-c) MAFW and (d) EmVidCap datasets, respectively. Green fonts indicate correctly generated AU-related words, red indicates clearly incorrectly generated words, and blue fonts represent words with semantics close to the ground truth.

3) Examples of H-EVC Performance: Fig. 5 displays examples of emotional captions generated by SWINBERT [20] and ECPA on the MAFW and EmVidCap datasets, respectively. ECPA produced more precise and emotion-consistent captions than SWINBERT, especially in capturing detailed AU-related descriptions. For example, in Fig. 5(a), SWINBERT [14] failed to recognize the AU "*wrinkled nose*", while our ECPA correctly identified it. In Fig. 5(b), SWINBERT [20] generated wrong AU-related words, "*closed eyes*". Fig. 5(c) illustrates an interesting case where generated caption "*slightly compressed lips*" is marked as incorrect compared to the ground truth label "*slightly open mouth*". However, both phrases suggest that the mouth isn't fully open. Recognizing this, we consider "*slightly compressed lips*" as conveying a meaning similar to the original. In Fig. 5(d), SWINBERT [20] missed the emotion word "*astonish*", while our ECPA generates it correctly. Compared to SWINBERT, our ECPA also successfully identified the key actions, *e.g.*, "*reading*" and "*putting*", as well as the small object "*small animal*", as highlighted in green. The cases demonstrate that our ECPA can focus more on key actions and objects to better understand the emotional semantics in the videos, leading to more accurate emotion-related video descriptions across various scenes.

E. Ablation Study

1) Effects of Different Components: To evaluate the effectiveness of different components in ECPA, we performed an ablation study on the MAFW dataset. Table IV presents the ablation results for progressively adding VEP, TEP, and ECA to the baseline Transformer-based video captioning framework. The baseline with only visual and textual features F_V and F_T, obtained results of 10.54%, 15.86%, 38.35%, and 28.26% on the BLEU-4, METEOR, ROUGE-L and CIDER metrics, respectively. Adding VEP led to relative improvements of 21.31%, 18.62%, 0.63% and 23.54%, respectively. Obviously, the best results were obtained when all prompts were incorporated, with an accuracy of 11.35%, 16.46%, 39.81%, and 34.38%, respectively. In addition, Table IV also reports the training time complexities of H-EVC by gradually adding VEP, TEP, and ECA. Referencing Lin *et al.* [20], the training complexity refers to the time for one iteration during training. The results indicate that our ECPA can achieve state-of-the-art performance without introducing much computational cost.

2) Effects of AU-level Visual Prompt Amount: We also discussed the impact of the number of AU-level visual emotion prompts in $VEP_2()$. Fig. 6 presents the performance of evaluation protocols versus different numbers of prompts on the MAFW and ECPA datasets, respectively. Evidently, using 3 AU-level visual emotion prompts resulted in the best performance on both datasets.

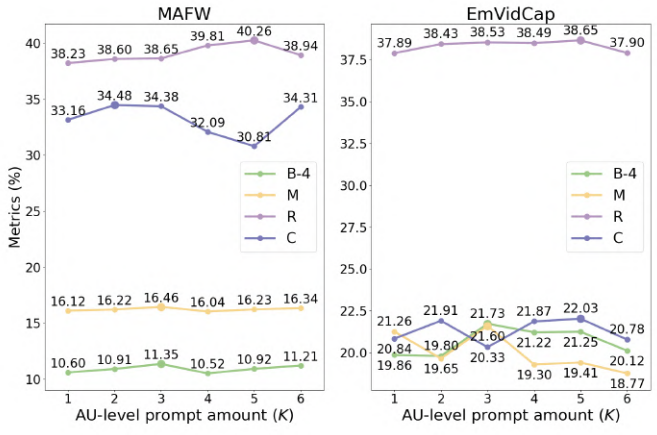

Fig. 6. Effects of the number of AU-level visual emotion prompts.

3) Effects of Different Alignment Levels in ECA: To evaluate the contribution of different alignment levels in ECA, we performed an ablation study on our two-level cross-modal alignment, namely the ER-Sentence prompt alignment and

TABLE IV
IMPACTS OF THREE IMPORTANT COMPONENTS (VEP, TEP, AND ECA) WITH THE COMPUTATIONAL COST ON THE MAFW DATASET. THE BEST RESULTS ARE IN BOLD, AND THE SECOND-BEST ARE UNDERLINED

Baseline	VEP	TEP	ECA	Semantic metrics (%)				Computational cost (s / iter)
				B-4	M	R-L	C	
√				10.32	15.57	38.11	27.61	0.800
√	√			10.54	15.86	38.35	<u>28.26</u>	0.817
√	√	√		<u>10.89</u>	<u>16.00</u>	<u>38.83</u>	28.09	0.849
√	√	√	√	**11.35**	**16.46**	**39.81**	**34.38**	0.952

AU-Word prompt alignment. The results of different cross-modal prompt alignments are provided in Table V. Obviously, the proposed multi-level emotion-oriented cross-modal prompt alignment helps the modal achieve the best performance. Firstly, the ER-Sentence alignment enhanced global emotional semantic relevance, resulting in relative increases of 0.63% and 2.35% in ROUGE-L and CIDEr metrics, respectively. This improvement may stem from the alignment's ability to use a broader and more appropriate vocabulary, capturing the video's emotional context more accurately. Secondly, the AU-Word alignment significantly boosted local linguistic precision, leading to relative increases of 2.13% and 1.86% in BLEU-4 and METEOR scores, respectively. This finer alignment improved the BLEU-4 score by ensuring that caption words accurately reflect the depicted actions, and for METEOR, it also enhanced both syntactic and semantic alignment with the visual data.

TABLE V
EFFECTS OF DIFFERENT CROSS-MODAL ALIGNMENT SCHEMES. THE BEST RESULTS ARE IN BOLD, AND THE SECOND-BEST ARE UNDERLINED

Baseline	ER-Sentence alignment	AU-Word alignment	Semantic metrics (%)			
			B-4	M	R-L	C
√			10.32	15.57	38.11	27.61
	√	√	10.09	14.76	36.73	27.08
√	√		10.54	15.86	38.35	<u>28.26</u>
√		√	<u>10.89</u>	<u>16.00</u>	<u>38.83</u>	28.09
√	√	√	**11.35**	**16.46**	**39.81**	**34.38**

4) Effects of Different Alignment Learning Schemes in ECA: Table VI compares different learning schemes used in ER-Sentence prompt alignment and AU-Word prompt alignment, respectively. In ER-Sentence prompt alignment, our ER-aware alignment loss outperformed others, demonstrating superior results. The similarity-based loss underperformed, likely because it prioritizes aligning general content over specific emotional states, which are better captured by our ER-aware alignment loss. In AU-Word prompt alignment, normal contrast loss did not yield the anticipated results, possibly due to the misclassification of data from the same emotional category as negative samples within a mini-batch. Compared to the contrast loss and cross-entropy loss, our proposed AU-aware alignment loss obtained better semantic performance, with relative improvements of 5.06%, 1.22%, and 18.07% on the BLEU-4, METEOR, ROUGE-L and CIDER metrics, respectively.

5) Effects of Various T_C in AU-Word Emotion Alignment: To discuss the effect of various thresholds T_C in positive/negative construction in AU-word emotion alignment, Fig. 7 presents the performance of H-EVC with varying T_C on the MAFW dataset. The results indicate that the best performance was obtained when T_C was set to 0.3, under all four evaluation protocols. In contrast, raising T_C to 0.4 leads to a drop in performance, likely due to the exclusion of slightly misaligned pairs that retain important contextual or emotional information.

TABLE VI
EFFECTS OF DIFFERENT ALIGNMENT LEARNING SCHEMES IN ECA. THE BEST RESULTS ARE IN BOLD AND THE SECOND-BEST ARE UNDERLINED

Comparison loss strategy	Semantic metrics (%)			
	B-4	M	R-L	C
ER-Sentence prompt alignment				
Global similarity loss	9.98	15.19	37.65	28.56
Similarity-based cross-entropy loss	<u>10.75</u>	<u>16.17</u>	<u>38.26</u>	<u>32.14</u>
ER-aware alignment loss	**11.35**	**16.46**	**39.81**	**34.38**
AU-Word prompt alignment				
Cross-entropy loss	10.67	15.62	<u>38.47</u>	29.06
Normal contrast loss	<u>10.85</u>	<u>15.99</u>	38.34	<u>33.73</u>
AU-aware alignment loss	**11.35**	**16.46**	**39.81**	**34.38**

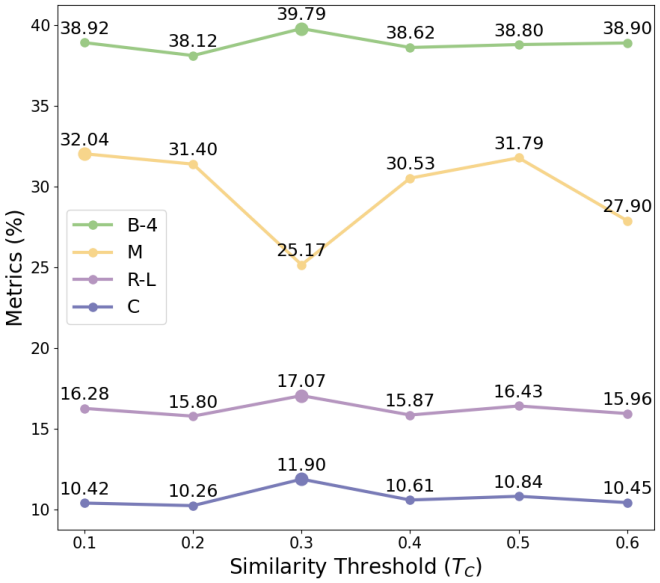

Fig. 7. Performance across positive/negative pair construction thresholds (T_C) on four evaluation protocols.

6) Effects of Various m in AU-Word Emotion Alignment: To find the best m in the AU-word emotion alignment, we conducted experiments over a range of m values. Fig. 8 presents the performance of H-EVC with varying m on the MAFW dataset. The results indicate that the best performance was obtained when m was set to 0.5, consistent with experience settings [59].

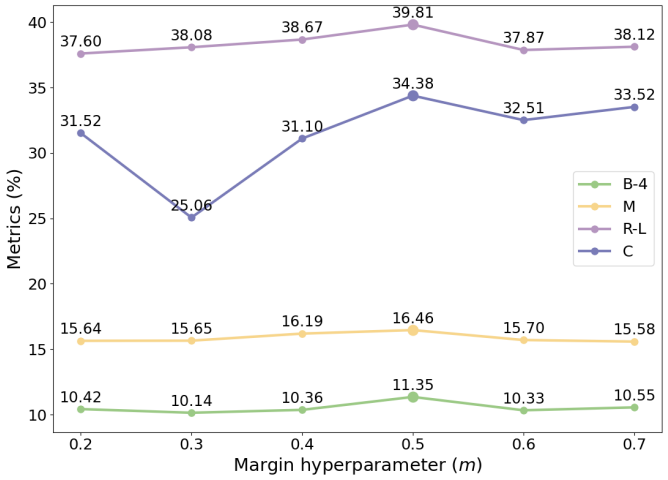

Fig. 8. Performance across margin (m) on four evaluation protocols.

7) Effects of ECPA with Different Frameworks: In addition, we discussed our ECPA approach within various video captioning frameworks to assess the generalization of our method. Four frameworks were tested on the MAFW dataset, which employed various large visual-language foundation models, *e.g.*, VidSwin [16], BERT [29], CLIP [28], T5 [30] and GPT2 [31] as the video or text encoders, respectively. Table VII indicates that the combination of VidSwin and BERT in our ECPA performed best, and integrating ECPA enhanced the performance across all frameworks. This can be attributed to our emotion prompts using VEP and TEP, as well as our multi-level cross-modal interactive alignment module ECA. For further evaluation, we froze the encoders on the two best frameworks. ECPA also outperformed the SOTA CLIP-based approach on B-4 and R-L, indicating its superior linguistic accuracy and alignment with reference texts.

TABLE VII
PERFORMANCES W/O ECPA UNDER DIFFERENT CAPTIONING FRAMEWORKS WITH VARIOUS VISION-LANGUAGE MODELS. THE BEST RESULTS ARE IN BOLD, AND THE SECOND-BEST ARE UNDERLINED

Video encoder	Text encoder	ECPA	Semantic metrics (%)			
			B-4	M	R-L	C
Finetuning the encoders						
VidSwin	BERT		10.32	15.57	38.11	27.61
VidSwin	BERT	√	**11.35**	**16.46**	**39.81**	**34.38**
CLIP	CLIP		9.83	15.32	37.64	28.02
CLIP	CLIP	√	<u>11.13</u>	<u>16.31</u>	<u>39.55</u>	<u>31.25</u>
VidSwin	T5		10.06	15.21	38.54	25.92
VidSwin	T5	√	10.14	15.37	38.75	29.38
VidSwin	GPT2		10.20	15.92	38.94	28.83
VidSwin	GPT2	√	10.21	15.93	38.13	29.77
Freezing the encoders						
VidSwin	BERT	√	**10.31**	15.51	**39.48**	29.39
CLIP	CLIP	√	10.28	**15.55**	39.29	**29.72**

F. Visualization and Qualitative Analysis

1) Convergence Performance: In Fig. 9, we compared the loss convergence curves of different visual-language encoders of VidSwin [16], CLIP [28], GPT2 [31], T5 [30] and BERT [29] in the video captioning framework, with our ECPA approach, on the MAFW dataset. The comparative results demonstrate that our approach helps achieve a better convergence and generalization capability.

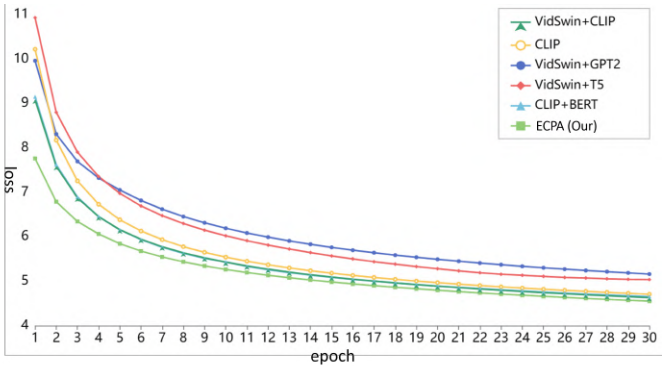

Fig. 9. Comparative analysis of loss convergence performance with various methods on MAFW. Obviously, our ECPA helps achieve a faster convergence during training.

2) Visualization on Attention Weights w/o Our ECPA: To better understand the contribution of our emotion-oriented cross-modal prompting alignment, we visualized attention weights of the textual language feature F_T and visual feature F_V, w/o our ECPA. Fig. 10 presents the visualization of attention weights on a randomly selected video from the MAFW dataset. Higher attention weights are visible in green boxes with our cross-modal prompting and alignment, as shown in Fig. 10(a). It can be concluded that our ECPA effectively enhances the interaction and fusion between vision and language related to emotions.

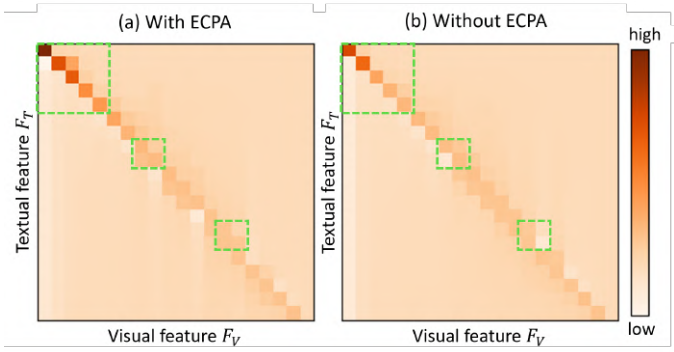

Fig. 10. Attention weights visualization between textual and vision modalities w/o our ECPA.

3) Visualization on Visual Prompts in VEP: To demonstrate the criticality of VEP more intuitively, we visualized the heatmaps for ER- and AU-level visual emotion prompts in Fig. 11(b, c), respectively, on the MAFW dataset. Compared to the video features F_V shown in Fig. 11(a), ER-level visual prompts P_{ER}^V in Fig. 11(b) better captured important visual facial expression information. Furthermore, AU-level visual prompts P_{AU}^V in Fig. 11(c) focused on more fine-grained local AU information, which can reveal subtle local facial changes for understanding the nuanced emotions.

4) Visualization on Cross-modal Alignment by ECA: In Fig. 12(a, b), we visualized the results of the two-layer cross-

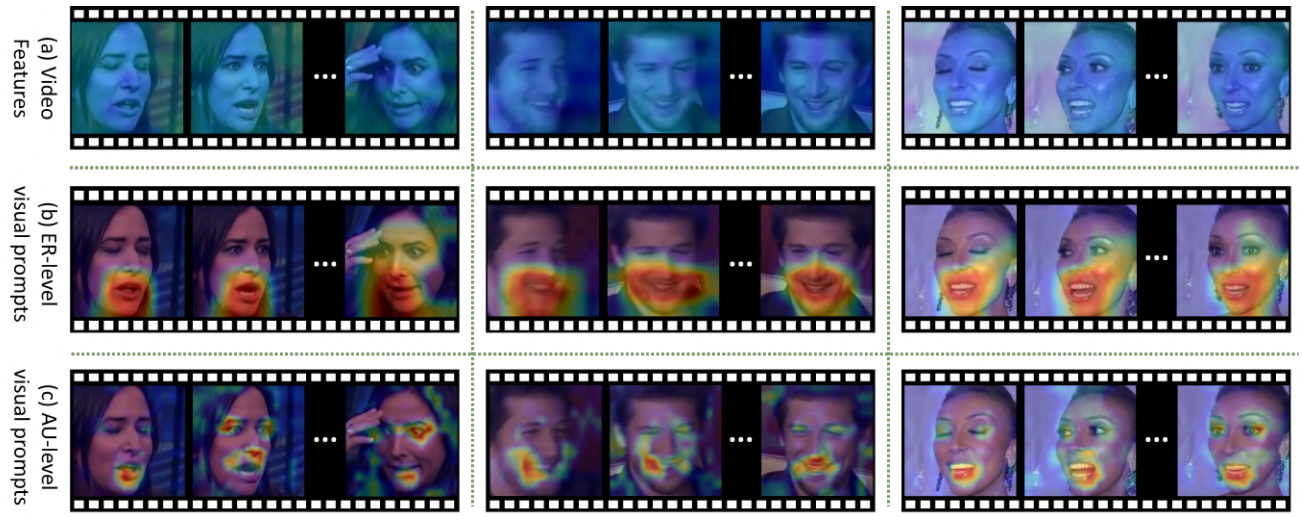

Fig. 11. Visualizing visual emotion prompts at different levels of VEP. (a) Video features, (b) ER-level visual prompts, and (c) AU-level visual prompts.

Fig. 12. Visualizing cross-modal emotion alignment at different levels. (a) Global cross-modal emotion alignment, (b) local cross-modal emotion alignment.

modal alignment mechanism in ECA, *i.e.*, global and local emotion alignment. As shown in Fig. 12(a), global cross-modal alignment effectively captured the overall emotional state of facial expressions, while Fig. 12(b) demonstrates that local cross-modal alignment focuses more on subtle local emotional information, such as the AU states of the mouth or eyes.

5) Examples on Generated Descriptions with Different Prompt Levels: Fig. 13 presents the generated linguistic descriptions for progressively adding ER- and AU-level prompts to the baseline on the MAFW dataset. In Fig. 13(a), the baseline generated incorrect emotion information as "*smiles*" and "*raised lip corners*", and the integration of ER-level prompt yielded emotion-aware descriptions as "*slightly open mouth*" and wrong "*wide eyes*". Further introducing AU-level refined the generated descriptions, correctly capturing AU captions as "*frown*" and "*wrinkled nose*". In Fig. 13(b), the baseline misidentified "*smiles*" and "*raised lip corners*". Our method achieved more refined emotion description as "*wide eyes*" and "*slightly open mouth*". In more challenging scenes like Figures 13(c, d), significant lighting changes can hinder emotion captioning despite using our two-layer prompts. For instance, in Fig. 13(c), while "*frown*" and "*slightly open mouth*" were recognized, ER- and AU- prompting failed to accurately describe eye movements. In Fig. 13(d), ECPA produced an incorrect description of "*compressed lips*". In summary, our approach effectively captures emotional content and changes in videos and generates human emotion-related event and action descriptions in videos by integrating ER and AU-level prompting. However, it is still difficult to generate accurate H-EVC when faced with extremely challenging environments such as dim lighting.

G. Generalization Analysis

1) Evaluation on the Cross-dataset H-EVC task: To verify the generalizability of ECPA across diverse datasets, cross-database validation was performed using two challenging EVC datasets, MAFW and EmVidCap. This validation involved a two-fold approach. First, ECPA was trained on MAFW and tested on ECPA. Then, EmVidCap was used for training and MAFW for testing. Table VIII presents the comparison results of ECPA and other state-of-the-art methods, including Transformer [60] and SWINBERT [20]. ECPA yields better results for reusing across datasets in terms of BLEU-4, METEOR, and

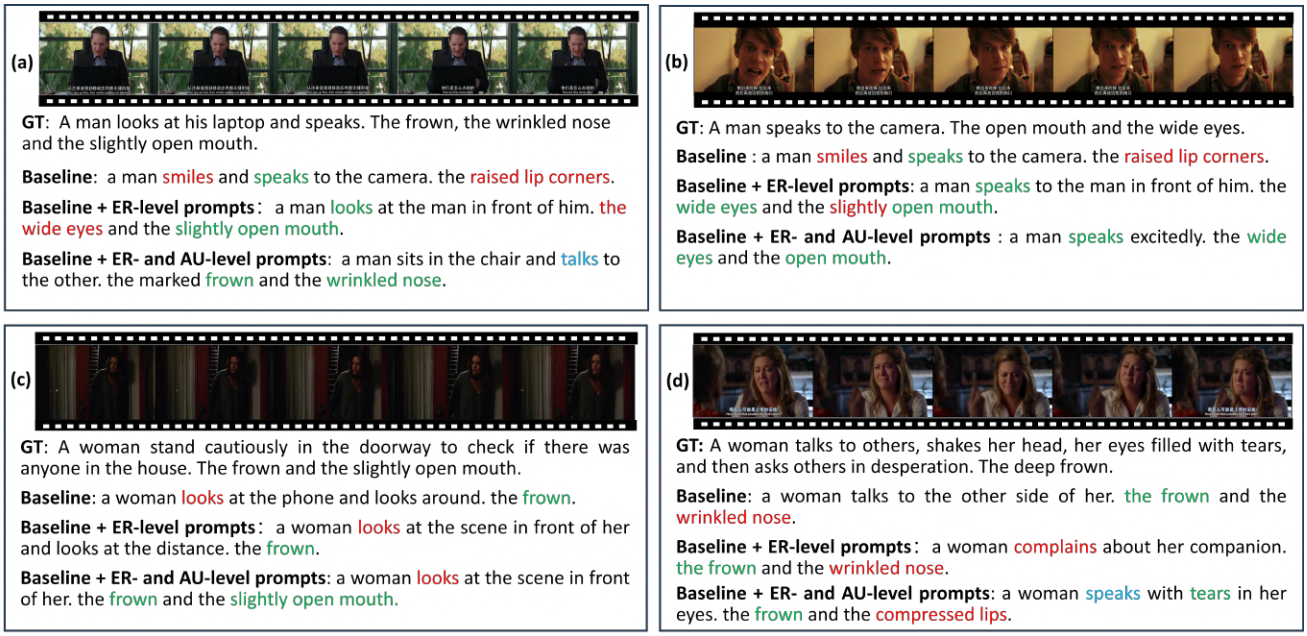

Fig. 13. Captions generated by baseline [60] and with different prompt levels, respectively. Green fonts indicate correctly generated fine-gained emotion words, red indicates error words, and blue fonts represent words with similar semantics to the ground truth.

CIDEr, despite variations in settings (*e.g.*, scene, pose, event, etc.). Despite having lower sentence structure similarity as measured by the ROUGE-L metric compared to SWINBERT, we focus on precise emotional expression and representation.

TABLE VIII
CROSS-DATASET VALIDATION ON MAFW → EMVIDCAP AND EMVIDCAP → MAFW, RESPECTIVELY

Methods	MAFW→EmVidCap				EmVidCap→MAFW			
	B4	M	R-L	C	B4	M	R-L	C
Transformer [60]	1.7	11.8	27.5	20.0	0.3	9.9	20.0	6.5
SWINBERT [20]	2.0	11.3	26.8	20.3	0.1	9.9	**20.6**	6.7
ECPA	**2.2**	**11.9**	**27.8**	**21.0**	**0.4**	**10.6**	15.5	**9.5**

2) Zero-shot Evaluation on the Video Captioning Task: In the zero-shot experiments, we trained our ECPA on MAFW, then performed the zero-shot evaluation on two general video captioning datasets, namely MSVD [43] and MSRVTT [44]. Table IX details the comparative zero-shot performance of ECPA against Transformer [60] and SWINBERT [20]. The results presented in Table IX demonstrate that ECPA outperforms prior methods in all evaluation metrics on the MSVD dataset and achieves the optimal performance on all metrics except for ROUGE-L on MSRVTT. Despite the relatively low results on the cross-task evaluation, they still demonstrate that ECPA generalizes to other task datasets more easily than other methods.

V. CONCLUSION

In this work, we first propose ECPA for fine-grained human-centric emotion captioning, which innovatively integrates visual emotion prompting (VEP) and textual emotion prompting (TEP), as well as emotion-oriented cross-modal alignment

TABLE IX
COMPARISON OF OUR METHOD AND OTHER STATE-OF-THE-ART METHODS FOR THE ZERO-SHOT VIDEO CAPTIONING TASK ON MSVD AND MSRVTT, RESPECTIVELY

Methods	MAFW→MSVD				MAFW→MSRVTT			
	B4	M	R-L	C	B4	M	R-L	C
Transformer [60]	1.5	14.9	31.5	0.7	1.7	12.7	27.6	0.9
SWINBERT [20]	1.5	14.2	31.8	0.8	1.2	15.1	**28.4**	0.7
ECPA	**1.7**	**15.1**	**32.5**	**1.0**	**1.8**	**13.9**	27.2	**1.2**

(ECA). VEP helps extract fine-grained visual emotion prompts at both ER and AU levels, while TEP depicts human emotion semantic cues from text inputs at the sentence- and word-level, respectively. Moreover, ECA further performs ER-sentence and AU-word cross-modal emotion alignment by introducing two novel global and local emotion-aware alignment losses. Evaluations conducted on widely used H-EVC datasets affirm the superiority of our model against state-of-the-art methods. Significantly, our approach not only enhances the accuracy of video content generation but also enriches the understanding of human-centric emotion-related captions. Despite our significant progress in the H-EVC task, we found that the problem of illusions in large models persists when larger-scale foundation models are introduced as the backbone. This leads to the generation of erroneous emotional events and content. In the future, we will delve deeper into this issue by introducing the chaining-based prompting learning mechanism to improve the accuracy and reliability of these models.

REFERENCES

[1] M. A. Butt, A. Qayyum, H. Ali, A. Al-Fuqaha, and J. Qadir, "Towards secure private and trustworthy human-centric embedded machine learning: An emotion-aware facial recognition case study," *Computers & Security*, vol. 125, p. 103058, 2023.

[2] P. Song, D. Guo, X. Yang, S. Tang, and M. Wang, "Emotional video captioning with vision-based emotion interpretation network," *IEEE Transactions on Image Processing*, 2024.

[3] X. Fang, D. Liu, P. Zhou, and Y. Hu, "Multi-modal cross-domain alignment network for video moment retrieval," *IEEE Transactions on Multimedia*, vol. 25, pp. 7517–7532, 2022.

[4] Z. Shao, J. Han, K. Debattista, and Y. Pang, "Textual context-aware dense captioning with diverse words," *IEEE Transactions on Multimedia*, vol. 25, pp. 8753–8766, 2023.

[5] H. Wang, P. Tang, Q. Li, and M. Cheng, "Emotion expression with fact transfer for video description," *IEEE Transactions on Multimedia*, vol. 24, pp. 715–727, 2021.

[6] P. Song, D. Guo, X. Yang, S. Tang, E. Yang, and M. Wang, "Emotion-prior awareness network for emotional video captioning," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 589–600.

[7] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 487–12 496.

[8] J. Zhang and Y. Peng, "Object-aware aggregation with bidirectional temporal graph for video captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8327–8336.

[9] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9837–9846.

[10] H. Ye, G. Li, Y. Qi, S. Wang, Q. Huang, and M.-H. Yang, "Hierarchical modular network for video captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17 939–17 948.

[11] Q. Li, T. Li, H. Wang, and C. W. Chen, "Taking an emotional look at video paragraph captioning," *arXiv preprint arXiv:2203.06356*, 2022.

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

[13] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[14] J. Guo, J. Li, D. Li, A. M. H. Tiong, B. Li, D. Tao, and S. Hoi, "From images to textual prompts: Zero-shot visual question answering with frozen large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 867–10 877.

[15] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4507–4515.

[16] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.

[17] H. Wang, G. Lin, S. C. Hoi, and C. Miao, "Cross-modal graph with meta concepts for video captioning," *IEEE Transactions on Image Processing*, vol. 31, pp. 5150–5162, 2022.

[18] N. Aafaq, A. Mian, W. Liu, N. Akhtar, and M. Shah, "Cross-domain modality fusion for dense video captioning," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 5, pp. 763–777, 2022.

[19] P. H. Seo, A. Nagrani, A. Arnab, and C. Schmid, "End-to-end generative pretraining for multimodal video captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 959–17 968.

[20] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 949–17 958.

[21] M. Wu, W. Su, L. Chen, Z. Liu, W. Cao, and K. Hirota, "Weight-adapted convolution neural network for facial expression recognition in human–robot interaction," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 3, pp. 1473–1484, 2019.

[22] F. Wang, H. Ma, X. Shen, J. Yu, and R. Xia, "Observe before generate: Emotion-cause aware video caption for multimodal emotion cause generation in conversations," in *ACM Multimedia 2024*, 2024.

[23] Y. Liu, W. Dai, C. Feng, W. Wang, G. Yin, J. Zeng, and S. Shan, "Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 24–32.

[24] Z. Xiao, Y. Chen, J. Yao, L. Zhang, Z. Liu, Z. Wu, X. Yu, Y. Pan, L. Zhao, C. Ma *et al.*, "Instruction-vit: Multi-modal prompts for instruction learning in vision transformer," *Information Fusion*, vol. 104, p. 102204, 2024.

[25] J. Li, L. Zhang, K. Zhang, B. Hu, H. Xie, and Z. Mao, "Cascade semantic prompt alignment network for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[26] D. Zhang, X.-J. Wu, T. Xu, and J. Kittler, "Two-stage supervised discrete hashing for cross-modal retrieval," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 11, pp. 7014–7026, 2022.

[27] Q. Xu, Y. Wei, S. Yuan, J. Wu, L. Wang, and C. Wu, "Learning emotional prompt features with multiple views for visual emotion analysis," *Information Fusion*, vol. 108, p. 102366, 2024.

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

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

[30] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[31] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[32] K. Yamazaki, K. Vo, Q. S. Truong, B. Raj, and N. Le, "Vltint: visual-linguistic transformer-in-transformer for coherent video paragraph captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 3081–3090.

[33] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, "Masked language model scoring," *arXiv preprint arXiv:1910.14659*, 2019.

[34] C. Deng, Q. Chen, P. Qin, D. Chen, and Q. Wu, "Prompt switch: Efficient clip adaptation for text-video retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 648–15 658.

[35] Q. He, "Prompting multi-modal image segmentation with semantic grouping," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, 2024, pp. 2094–2102.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] Y. Zhang, W. Dong, B.-G. Hu, and Q. Ji, "Classifier learning with prior probabilities for facial action unit recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5108–5116.

[38] C. Luo, S. Song, W. Xie, L. Shen, and H. Gunes, "Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition," *arXiv preprint arXiv:2205.01782*, 2022.

[39] J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. Chilimbi, and J. Huang, "Vision-language pre-training with triple contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 671–15 680.

[40] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[41] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[42] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1871–1880.

[43] D. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 2011, pp. 190–200.

[44] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5288–5296.

[45] D. Callan and J. Foster, "How interesting and coherent are the stories generated by a large-scale neural language model? comparing human and automatic evaluations of machine-generated text," *Expert Systems*, vol. 40, no. 6, p. e13292, 2023.

[46] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings*

of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.
[47] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.
[48] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.
[49] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.
[50] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.
[51] J. Rasley, S. Rajbhandari, O. Ruwase, and Y. He, "Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 3505–3506.
[52] J. Hou, X. Wu, W. Zhao, J. Luo, and Y. Jia, "Joint syntax representation learning and visual cue translation for video captioning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8918–8927.
[53] B. Wang, L. Ma, W. Zhang, and W. Liu, "Reconstruction network for video captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7622–7631.
[54] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2641–2650.
[55] Q. Zheng, C. Wang, and D. Tao, "Syntax-aware action targeting for video captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 096–13 105.
[56] Y. Song, S. Chen, and Q. Jin, "Towards diverse paragraph captioning for untrimmed videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 245–11 254.
[57] Y. Zheng, Y. Zhang, R. Feng, T. Zhang, and W. Fan, "Stacked multimodal attention network for context-aware video captioning," *IEEE transactions on circuits and systems for video technology*, vol. 32, no. 1, pp. 31–42, 2021.
[58] J. Wang, M. Yan, Y. Zhang, and J. Sang, "From association to generation: Text-only captioning by unsupervised cross-modal mapping," *arXiv preprint arXiv:2304.13273*, 2023.
[59] H. Choi, A. Som, and P. Turaga, "Amc-loss: Angular margin contrastive loss for improved explainability in image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.
[60] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

Yu Wang received the B.S. and M.S. degrees in Surveying and Mapping Engineering from Kunming University of Science and Technology, Kunming, China, in 2016 and 2019, respectively. She is currently a Ph.D. student at China University of Geosciences, Wuhan, China.

Her research interests include computer vision and artificial intelligence in multimodal interaction, affective computing and geospatial artificial intelligence.

Yuanyuan Liu (Member, IEEE) received the Ph.D. degree from Central China Normal University, Wuhan, in 2015. She is currently an Associate Professor at the School of Computer Science, China University of Geosciences (Wuhan). She has published more than 40 peer-reviewed papers, including those in highly regarded journals and conferences such as CVPR, ACM MM, PR, INS, IEEE TGRS, FGR and ICIP, etc. Her research interests include image processing, computer vision, pattern recognition, affective computing, and multimodal interaction, etc.

Shunping Zhou received the B.S. and Ph.D. degrees from the China University of Geosciences, Wuhan, China. He is currently a professor at the School of Computer Science, China University of Geosciences (Wuhan).

His research interests include spatial database technology, geospatial artificial intelligence and computer vision.

Yuxuan Huang received the BS degree in Computer Science and Technology from Southwest Minzu University in 2022. She is currently working toward the Master degree at the China University of Geosciences in Wuhan, China. She has published a paper on ACM MM2024.

Her research interests include computer vision and affective computing.

Chang Tang (Senior Member, IEEE) received the PhD degree from Tianjin University, Tianjin, China, in 2016. He joined the AMRL Lab of the University of Wollongong between September 2014 and September 2015. He is currently an Associate Professor at the School of Computer Science, China University of Geosciences, Wuhan, China. He has published more than 50 peer-reviewed papers, including those in highly regarded journals and conferences such as IEEE TPAMI, TKDE, TMM, THMS, SPL, AAAI, IJCAI, ICCV, CVPR, ACMM, ICME, etc. His current research interests include machine learning and computer vision.

Wujie Zhou (Senior Member, IEEE) is currently an Associate Professor at the School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou, China. He is also a Postdoc Fellow at the Institute of Information and Communication Engineering, Zhejiang University, Hangzhou, China. He has published over 70 peer-reviewed papers, including those in highly regarded journals and conferences such as AAAI, TCI, TIM, MIS, TCDS, TETCI, TIV and PR, etc. His research interests include multimedia signal processing and communication.

Zhe Chen (Member, IEEE) received his PhD from the University of Sydney in 2019. He is now a lecturer at La Trobe University and is affiliated with the Cisco-La Trobe Centre for Artificial Intelligence and Internet of Things. He is a highly-cited researcher with regular publications in top conferences and journals such as CVPR, ECCV, ICCV, IJCV, and TIP. He has also participated in and won championship international computer vision competitions like ImageNet 2017 Detection from Video. His research focuses on visual understanding and deep learning applications on healthcare, robotics, space exploration, and so on.