Emotion-Prior Awareness Network for Emotional Video Captioning

Peipei Song
beta.songpp@gmail.com
University of Science and Technology of China

Dan Guo[*]
guodan@hfut.edu.cn
Hefei University of Technology
Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

Xun Yang[*]
xyang21@ustc.edu.cn
University of Science and Technology of China

Shengeng Tang
tangsg@hfut.edu.cn
Hefei University of Technology

Erkun Yang
yangerkun@xidian.edu.cn
Xidian University

Meng Wang[*]
eric.mengwang@gmail.com
Hefei University of Technology

ABSTRACT

Emotional video captioning (EVC) is an emerging task to describe the factual content with the inherent emotion expressed in a video. It is crucial for the EVC task to effectively perceive subtle and ambiguous visual emotion cues in the stage of caption generation. However, existing captioning methods usually overlooked the learning of emotions in user-generated videos, thus making the generated sentence a bit boring and soulless.

To address this issue, this paper proposes a new emotional captioning perspective in a human-like perception-priority manner, *i.e.*, first perceiving the inherent emotion and then leveraging the perceived emotion cue to support caption generation. Specifically, we devise an Emotion-Prior Awareness Network (EPAN). It mainly benefits from a novel tree-structured emotion learning module involving both catalog-level psychological categories and lexical-level usual words to achieve the goal of explicit and fine-grained emotion perception. Besides, we develop a novel subordinate emotion masking mechanism between the catalog level and lexical level that facilitates coarse-to-fine emotion learning. Afterward, with the emotion prior, we can effectively decode the emotional caption by exploiting the complementation of visual, textual, and emotional semantics. In addition, we also introduce three simple yet effective optimization objectives, which can significantly boost the emotion learning from the perspectives of emotional captioning, hierarchical emotion classification, and emotional contrastive learning. Sufficient experimental results on three benchmark datasets clearly demonstrate the advantages of our proposed EPAN over existing SOTA methods in both semantic and emotional metrics. The extensive ablation study and visualization analysis further reveal the good interpretability of our emotional video captioning method. Code will be made available at https://github.com/songpipi/EPAN.

CCS CONCEPTS

• **Computing methodologies → Scene understanding**.

KEYWORDS

Video captioning, emotion learning, video understanding

ACM Reference Format:
Peipei Song, Dan Guo, Xun Yang, Shengeng Tang, Erkun Yang, and Meng Wang. 2023. Emotion-Prior Awareness Network for Emotional Video Captioning. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3581783.3611726

1 INTRODUCTION

The wide popularity of social networks enables people to habitually express their opinions with images or videos, which raises an urgent demand for visual emotion analysis. Among the large amount of online multimedia data, short videos with dynamic and vibrant storylines often elicit a wide range of emotional reactions from viewers [61, 66]. In recent years, emotion-aware video understanding tasks have drawn increasing attention, including emotional music-video retrieval [38], advertisement recommendation [17], and emotional video captioning (EVC) [43]. The EVC task is an emerging hot topic in multimedia and computer vision communities [36, 43]. As an intersection of vision, emotion, and language studies, EVC requires not only comprehending the video content but also perceiving the intrinsic visual sentiment as well as incorporating the sentimental semantics for caption generation.

To generate high-quality emotional captions, one primary issue is how to accurately capture the intrinsic emotions expressed in videos. For emotion recognition, early research mainly focused on emotion recognition in texts, which contain emotion-specific words, *e.g.*, emotion-label words "happy" and "sad" and emotion-laden words "successful" and "failed". These words straightforwardly elucidate or describe one's sentimental state. Thus, text emotion recognition approaches are mostly based on the word-level analysis of texts to detect the explicit expressions of sentiment

[*]Corresponding authors. This work was supported by the National Key R&D Program of China (2022YFB4500600), the National Natural Science Foundation of China (72188101, 62020106007, 62272144, U20A20183, 62272435, and U22A2094), and the Major Project of Anhui Province (202203a05020011).

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611726

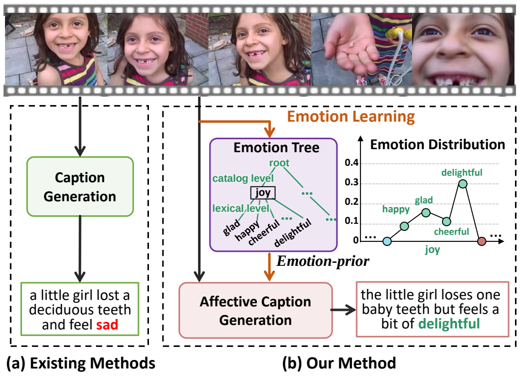


Figure 1: Existing EVC pipelines (a) and ours (b). Different from existing methods [36, 43], we solve this task in a human-like perception-priority manner, first performing tree-structured emotion learning, and then generating captions guided by the emotion priors.

[22, 52, 69]. Different from the emotion-specific words in texts, the visual emotion cues in images or videos are implicit due to the characteristics of ambiguity, subtlety, and heterogeneity [66, 67]. To capture visual emotions, existing visual emotion analysis works have made efforts to enrich affective representations [28, 50, 51, 66], such as by leveraging various visual stimuli (*e.g.*, color and object) or modeling the polarity, type, and intensity of emotion categories [50, 51]. The progress in visual emotion analysis has inspired the study of emotional captioning. Some image-based works adopted a two-stage offline training scheme [1, 23, 29]. They first applied emotion analysis techniques to classify images into several emotion categories, such as positive/negative [23], Ekman's 6 emotions [29], and Mikel's 8 emotions [1]. Then, according to the emotion categories, they learned to generate emotional descriptions. Despite the promising progress, it is still difficult for captioner to generate emotional words as rich, vivid, and natural as real-life language by just considering the coarse-grained emotion category prompt.

For the challenging EVC task, existing attempts [36, 43] focused on implicit emotion learning in a data-driven manner. They directly optimized the semantic mapping between videos and emotional captions as did traditional captioning methods. However, merely modeling semantic relationships is insufficient to fill the affective gap between videos and emotions. As shown in Figure 1 (a), the actual emotion of the "delightful" girl is misinterpreted as "sad" by the factual semantic of "lost teeth". Besides, these methods lack an explicit quantitative analysis of emotion and may be easily affected by the problem of adequate common words and low-frequency emotion words in the training corpus.

To fill the research gap, this work enhances EVC with explicit emotion learning in an end-to-end manner. As shown in Figure 1 (b), we introduce emotion priors through a tree-structured emotion learning scheme involving both psychological categories (catalog level) and usual emotion words (lexical level). Our motivation is two-sided. As well known, in the field of emotion analysis, the psychological categories are more professional than daily words, whose theories are built on the all-around measurement of type, polarity, and intensity to cover human emotions [33]. As well, our goal is to generate caption sentences with emotion words as naturally as possible. Therefore, we leverage the strengths of both psychological emotion categories (from psychological theory) and natural emotion words (from caption annotations) to achieve coarse-to-fine emotion understanding. In our emotion solution, we adopt the psychology emotion catalog and usual emotion words in [43].

Based on the above considerations, we introduce a novel Emotion-Prior Awareness Network (EPAN) for EVC. We propose a human-like perception-priority solution, which first perceives the emotion and then uses it to guide caption generation. As shown in Figure 2, EPAN includes two parts: *Tree-structured Emotion Learning* and *Affective Caption Generation*. For emotion learning, we first perform a quantitative emotion analysis over the psychological categories and obtain the responsive categories. Then, the daily emotion words inherited from the responsive categories are selected for further emotion learning at the lexical level. This operation effectively narrows the range of relevant emotional words. To this end, we design a subordinate emotion mask to filter out irrelevant emotion words along the emotion tree (*i.e.*, from psychological categories to emotion words). The mask operation is embedded into the end-to-end captioning framework to realize cross-validation between the catalog-level and the lexical-level emotion learning. Finally, we effectively exploit the complementation of visual, textual, and emotional semantics for more accurate emotional caption generation. Moreover, to ensure high-quality caption generation, we introduce three simple yet effective emotion-aware learning objectives for training: *1) hierarchical emotional classification loss* for boosting tree-based emotion accuracy, *2) emotional contrastive loss* for differentiating positive and negative emotion-related video-caption pairs, and *3) basic emotional cross-entropy loss* with an emotion penalization term.

The main contributions are summarized as follows:

- We present a new emotion-prior awareness perspective for EVC, which first perceives emotion and then generates captions in a human-like emotion-priority perception manner, rather than simply addressing video-to-caption semantic mapping.
- We devise a tree-structured emotion learning scheme on psycho-emotional categories and the subordinated usual emotion words, where a subordinate mask mechanism is applied between the catalog level and lexical level to filter irrelevant emotion words.
- Three simple yet effective emotion-aware learning objectives involving emotional captioning, emotion recognition, and contrastive emotional learning are specifically devised for this task, which can effectively optimize our proposed EPAN.
- Our method achieves new state-of-the-art performance on three EVC datasets (*i.e.*, EVC-MSVD, EVC-VE, and EVC-Combined), *e.g.*, improving the latest records by +7.0% and +12.3% *w.r.t.* CIDEr and emotion accuracy, respectively, on EVC-Combined. Additional experiments on the SentiCap dataset (image) also demonstrate the generalizability and effectiveness of our EPAN.

2 RELATED WORK

Visual Emotion Analysis. In recent years, various approaches have attempted to reveal the emotion aroused by images or videos [14, 15, 45, 67]. *Image-based emotion analysis* can be divided into several aspects: 1) categorical emotion states (CES) [35, 49, 62] and dimensional emotion space (DES) [18, 26] from the psychological model view, 2) personalized emotion prediction [2, 64] and dominant emotion recognition [48, 49] from individual or public subject view, 3) emotion distribution learning for compound emotion analysis [13, 47, 50], and 4) others, such as that domain adaptation [63, 65] and few/zero-shot learning [44, 59] have been studied for label absence challenges in visual emotion analysis. By comparison, *video-based emotion analysis* has developed slowly. Most works studied crucial spatiotemporal variations of video, such as emotion-related visual regions or segments [28, 31, 57]. And some others strived to acquire emotion-rich clues, *e.g.*, audio features [66], facial features [25], and visual aesthetics [28], *etc.*

Emotional Captioning. The early works are equipped with some predefined emotions, such as positive/negative polarity [19, 68] and psychological categories (*i.e.*, *anger*, *disgust*, *fear*, *happiness*, *sadness*, or *surprise*) [39]. Under the precondition, the model output emotionally customized captions, while the real emotion evoked by visual content was negligible. To break this, some works proposed to generate sentimental captions conditioned on visual emotions [1, 23, 29]. They first used a well-trained visual emotion analyzer to obtain emotion categories and then generate captions. However, this two-stage pipeline deals with emotion understanding only at the category level, which is inadequate to generate rich, vivid, and natural emotional descriptions. Recently, EVC task has attracted interest in the community, which requires describing videos with fine-grained emotion categories and diverse emotion words. The two most related works for this task are [36, 43]. Wang *et al.* [43] proposed a fact decoder and an emotion decoder to jointly generate captions. Song *et al.* [36] utilized an attention mechanism over video and text both to enhance affective semantics. However, both have the drawback as shown in Figure 1 — they lack explicit quantitative analysis of emotion. This leads to their inability to address the affective gap between video and emotions, and are susceptible to the problem of adequate common words and low-frequent emotion words in the training corpus. In this work, we investigate EVC with explicit emotion learning in an end-to-end manner.

3 METHOD

This section describes our proposed Emotion-Prior Awareness Network (EPAN) for the EVC task. As shown in Figure 2, it mainly consists of two parts: 1) *Tree-structured Emotion Learning* that captures the emotion cues in the given video, as described in Sec. 3.1, and 2) *Affective Caption Generation* that leverages the identified affective prior cues to guide the emotional video captioning, as described in Sec. 3.2. We describe how to optimize our method by introducing three emotion-aware learning objectives, which can effectively ensure the emotional and semantic correctness of generated captions, as described in Sec. 3.3.

3.1 Tree-structured Emotion Learning

EVC is a challenging task that requires understanding both factual and emotional semantics from vision simultaneously. Generally, humans perform well on the EVC task due to their strong emotional perceptivity. They usually first observe a video and *perceive* dominant emotions, and then *describe* the visual content driven by the aroused emotions. Inspired by this, we address emotion recognition first, while it is still hard to discover the ambiguous and subtle emotional cues in the video. The main technical contribution of this work is that we propose to enhance emotional video captioning via a tree-structured emotion learning strategy at both the catalog level (psychological categories) and the lexical level (natural words). Our intuition is that psychological categories have the theoretical expertise for quantitative emotion analysis, and natural emotion words are significantly helpful to generate human-like fluent caption sentences. *The research question is how to effectively leverage catalog-lexical emotion structure for learning emotion in videos.*

Emotion Tree: As shown in the example in Figure 1, we manually build an emotion tree with M_c categories and M_w words to encode catalog-lexical emotion word structure, inspired by [43]. On the emotion tree, each natural emotion word is assigned to a psychological category, *e.g.*, “happy and cheerful” is subordinate to the category “joy”. A detailed description about the emotion tree can be found in the Appendix. To achieve the emotion delivery from the catalog level to the lexical level along the tree, we design a subordinate **emotion mask**, which can effectively filter out irrelevant emotion words under the guidance of the selected categories. The emotion mask is embedded in the end-to-end captioning framework, which will be described later. This achieves an online cross-validation between the catalog-level and lexical-level emotion learning based on our constructed emotion tree. In the part of tree-structured emotion learning, both the video and emotion tree are exploited as the input. We will describe how to prepare the video features, and how to encode the catalog-level emotion cues and also the lexicon-level emotion cues in the rest of this section.

3.1.1 Video Feature Preparation.

First, we prepare the original video features. The pre-trained 2D CNN [12, 34] and 3D CNN [9, 11] networks are applied to extract the appearance feature $\mathbf{F}_a$ and motion feature $\mathbf{F}_m$ separately. They are concatenated into $\mathbf{F}=[\phi_a(\mathbf{F}_a);\phi_m(\mathbf{F}_m)]$, where ϕ_a and ϕ_m are two linear layers. To explore the relationship among successive visual features in the video, we apply a basic transformer block [40] to further encode $\mathbf{F}$ and obtain a new video feature $\mathbf{V}\in\mathbb{R}^{N\times d_v}$, where N is the number of video frames/clips. The raw video features $\mathbf{V}$ capture the factual semantics of objects, scenes, *etc.*, but are insufficient to discover the emotional cues in videos. In the following, we endeavor to obtain emotion-discriminative video representations $\mathcal{E}$ using emotional catalog encoder and emotional lexicon encoder. The two encoders cooperate to perform tree-structured emotional learning.

3.1.2 Emotional Catalog Encoder.

In this section, we use psychological terminologies to determine the scope of emotion categories for the video. As shown in Figure 3, the emotional catalog encoder consists of two types of input data, namely the video features $\mathbf{V}$ and the psychological emotion catalog $\mathbf{E}_c=\{\mathbf{c}_m\}_{m=1}^{M_c}$, where $\mathbf{c}_m$ denotes the m-th category such as “joy”. We extract the word embedding of $\mathbf{E}_c$ by GloVe [32]. Queried by video $\mathbf{V}$, the evoked emotion $\mathbf{E}'_c\in\mathbb{R}^{N\times d_e}$ are encoded as below:

$$\mathbf{E}'_c = \text{TransEncoder}_c(\mathbf{V},\mathbf{E}_c)|_{\mathcal{Q}:\ \mathbf{V},\{\mathcal{K},\mathcal{V}\}:\ \mathbf{E}_c}. \tag{1}$$

Then, we utilize the new emotion embedding $\mathbf{E}'_c$ to predict an emotion distribution $\mathbf{P}_c^E$ over the emotion category set. The obtained $\mathbf{P}_c^E$ will contribute to emotional captioning in two ways. 1) We use the $\mathbf{P}_c^E$ to determine subordinate natural emotion words in the subsequent lexicon, such as emotion words “happy, glad, and cheerful,

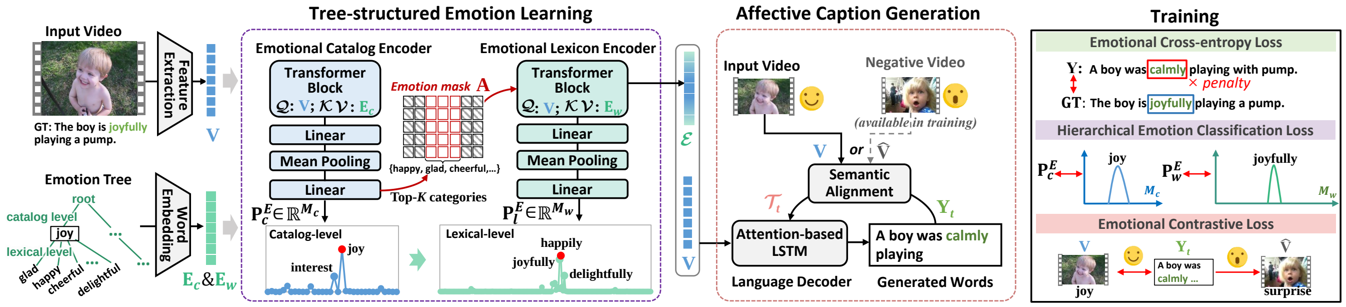
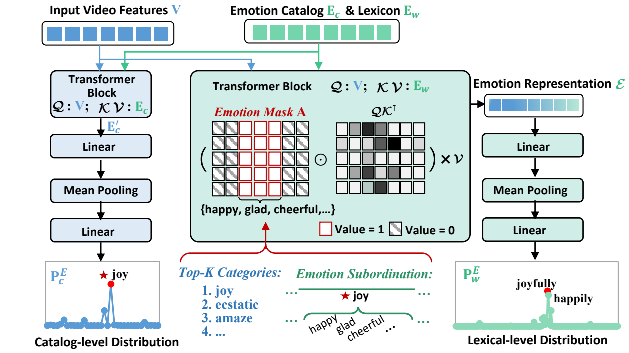


Figure 2: Overview of our EPAN. It mainly consists of two parts: *Tree-structured Emotion Learning* and *Affective Caption Generation*. We design an emotional catalog encoder to recognize psychological categories and an emotional lexicon encoder to select natural emotion words. We retain top-K categories at the catalog level and design a subordinate mask to prevent irrelevant emotion words at the lexical level, and consequently obtain an emotion representation $\mathcal{E}$. We then exploit $\{\mathrm{V}, \mathcal{E}, \mathcal{T}_t\}$ to generate a caption step by step, where $\mathcal{T}_t$ is the textual feature of partially generated sentence Y_t at the t-th step. Finally, three emotion-aware learning objectives are devised to ensure the emotional and semantic correctness of generated sentences.



(a) Emotional Catalog Encoder (b) Emotional Lexicon Encoder

Figure 3: Illustration of the subordinate emotion mask mechanism. Firstly, we select top-K emotion categories from the catalog-level prediction. Based on the tree-structured subordination, the mask of irrelevant natural emotion words is set to zero. In other words, the mask variable A is used to shield irrelevant emotion words in emotion learning.

etc.” under the category “joy”. 2) Based on $\mathbf{P}_c^E$, we design a hierarchical emotion classification loss $\mathcal{L}_{cls}^E$ detailed in Sec. 3.3. Formally, $\mathbf{P}_c^E$ is obtained by mean pooling and linear mapping operations:

$$\mathbf{P}_c^E = \phi_c^2\left(\mathrm{MeanPool}(\phi_c^1(\mathbf{E}_c'))\right) \in \mathbb{R}^{M_c}, \tag{2}$$

where ϕ_c^1 and ϕ_c^2 are two linear layers.

3.1.3 **Emotional Lexicon Encoder.** In this part, we pay more attention to natural emotion words at the lexical level. We utilize the $\mathbf{P}_c^E$ as a categorical guidance to select relevant emotion words from an emotion lexicon $\mathbf{E}_w$. This means that under the professional judgment of psychological classification, irrelevant emotion words do not participate in lexical-level emotion encoding. Here is $\mathbf{E}_w = \{\mathbf{e}_m\}_{m=1}^{M_w}$, where $\mathbf{e}_m$ denotes the m-th natural emotion word, such as “happy”. To achieve the goal of effective emotion learning, we consider only the top-K emotion categories in $\mathbf{P}_c^E$, and construct a *subordinate emotion mask* $\mathbf{A} = [\mathbf{a}_m] \in \mathbb{R}^{N \times M_w}$ as formulated in Eq. (3). If an emotion word $\mathbf{e}_m$ belongs to the top-K emotion categories (symbolized as ∝), the mask value in $\mathbf{a}_m$ is set to 1, otherwise 0.

$$\mathbf{A} = [\mathbf{a}_m] \in \mathbb{R}^{N \times M_w}, \quad m \in [1, M_w],$$
$$s.t.\ \mathbf{a}_m = \begin{cases} \mathbf{1}_N, & \mathbf{e}_m \propto \text{top-}K(\mathbf{P}_c^E), \\ \mathbf{0}_N, & otherwise. \end{cases} \tag{3}$$

where $\mathbf{1}_N$ and $\mathbf{0}_N$ denote all-one and all-zero vectors of size N, respectively. Then, we impose the mask $\mathbf{A}$ onto video $\mathbf{V}$ and emotion lexicon $\mathbf{E}_w$ through a transformer block, and obtain the emotion representation $\mathcal{E} \in \mathbb{R}^{N \times d_e}$ as follows:

$$\mathcal{E} = \mathrm{TransEncoder}_w(\mathbf{V}, \mathbf{E}_w, \mathrm{mask}=\mathbf{A})|_{\mathcal{Q}:\ \mathbf{V}, \{\mathcal{K}, \mathcal{V}\}:\ \mathbf{E}_w}. \tag{4}$$

Based on $\mathcal{E}$, we obtain the lexical-level emotion distribution $\mathbf{P}_w^E$ as below. Similar to $\mathbf{P}_c^E$, the $\mathbf{P}_w^E$ is also used in the design of loss $\mathcal{L}_{cls}^E$ (in Sec. 3.3) for hierarchical emotion learning optimization.

$$\mathbf{P}_w^E = \phi_w^2\left(\mathrm{MeanPool}(\phi_w^1(\mathcal{E}))\right) \in \mathbb{R}^{M_w}, \tag{5}$$

where ϕ_w^1 and ϕ_w^2 are two linear layers.

3.2 Affective Caption Generation

Apart from the emotional cues, previously generated words are also help predict the next word. We enforce visual-textual semantic alignment for caption generation step by step. At the t-th decoding step, we obtain the visually correlated text feature $\mathcal{T}_t$ as follows.

- We encode the partially generated caption and obtain the textual features $\mathbf{W}^{(t)} \in \mathbb{R}^{t \times d_w}$ through GloVe and transformer block.
- Given the video features $\mathbf{V} \in \mathbb{R}^{N \times d_v}$ and the caption features $\mathbf{W}^{(t)}$, we calculate the relevance between visual feature $\mathbf{v}_i$ and textual feature $\mathbf{w}_j^{(t)}$ in each pair.
- The relevance scores are normalized to align all the generated words onto each visual feature and obtain visually correlated text feature $\mathcal{T}_t \in \mathbb{R}^{N \times d_w}$.

$$\mathcal{T}_t = \left\{\sum_{j=1}^{t} \mathrm{softmax}(r_{i,j}^{(t)})\mathbf{w}_j^{(t)}\right\}_{i=1}^{N},$$
$$s.t.\ r_{i,j}^{(t)} = \mathbf{u}_r^\top \tanh(\mathbf{U}_r \mathbf{v}_i + \mathbf{H}_r \mathbf{w}_j^{(t)} + \mathbf{b}_r), \tag{6}$$

where $\mathbf{u}_r$, $\mathbf{U}_r$, $\mathbf{H}_r$, and $\mathbf{b}_r$ are learnable parameters.

With sufficient semantics $\mathbf{V}, \mathcal{E}$, and $\mathcal{T}_t$, we use a standard attention-based LSTM to generate the caption sentence step by step. At the decoding step t, we combine all the sufficient semantics as $\mathbf{C}_t = [\mathbf{V}; \mathcal{E}; \mathcal{T}_t] \in \mathbb{R}^{N \times (d_v + d_e + d_w)}$ and summarize $\mathbf{C}_t$ into a vector $\mathbf{z}_t$ for visual, textual, and emotional context aggregation.

$$\begin{aligned} &\mathbf{z}_t = \sum\nolimits_{i=1}^{N} g_{i,t} \mathbf{c}_{i,t} \in \mathbb{R}^{(d_v+d_e+d_w)}, \\ &s.t.\ g_{i,t} = \text{softmax}\left(\mathbf{u}_g^\top \tanh(\mathbf{U}_g \mathbf{h}_{t-1} + \mathbf{H}_g \mathbf{c}_{i,t} + \mathbf{b}_g)\right), \end{aligned} \tag{7}$$

where $\mathbf{c}_{i,t}$ denotes the i-th feature in $\mathbf{C}_t$; $\mathbf{u}_g$, $\mathbf{U}_g$, $\mathbf{H}_g$, and $\mathbf{b}_g$ are learnable parameters. By using LSTM, we obtain the probability distribution of the next word $\mathbf{y}_{t+1}$ as follows:

$$\begin{cases} \mathbf{h}_{t+1} = \text{LSTM}\left([\mathbf{z}_t; \mathbf{y}_t], \mathbf{h}_t\right); \\ \mathbf{P}(\mathbf{y}_{t+1}) = \mathbf{U}_p \mathbf{h}_{t+1} + \mathbf{b}_p, \end{cases} \tag{8}$$

where $\mathbf{U}_p$ and $\mathbf{b}_p$ are learnable parameters.

3.3 Training

To ensure both the emotional and semantic correctness of the generated captions, we devise three emotion-aware learning objectives from the perspectives of emotional captioning, hierarchical emotion classification, and emotional contrastive learning, respectively.

3.3.1 **Emotional Cross-entropy Loss.** Cross-entropy loss [24, 46] is a fundamental learning objective for video captioning. Existing EVC works [36, 43] directly adopt the traditional cross-entropy loss [58], having weakness in generating low-frequency but meaningful emotion words in the corpus. In this work, we propose a customized cross-entropy loss $\mathcal{L}_{ce}^{E}$ considering the prominence of emotional semantics. We introduce a penalty term on the emotion lexicon and the $\mathcal{L}_{ce}^{E}$ is defined as follows.

$$\begin{aligned} &\mathcal{L}_{ce}^{E} = -\left(1 + \alpha f(\mathbf{y}_t)\right) \sum_t log\mathbf{P}(\mathbf{y}_t), \\ &s.t.\ f(\mathbf{y}_t) = \begin{cases} 1, & if\ \mathbf{y}_t \in \mathbf{E}_w, \\ 0, & otherwise. \end{cases} \end{aligned} \tag{9}$$

where α is a penalty coefficient that controls the cross-entropy benefit on emotion words while $\mathbf{y}_t \in \mathbf{E}_w$.

3.3.2 **Hierarchical Emotion Classification Loss.** As stated in Sec. 3.1, tree-structured emotion learning is a crucial technical contribution in our work. Correspondingly, we propose a hierarchical optimization on emotion classification. Based on the emotional catalog and lexicon encoders, we obtain the emotion distributions $\mathbf{P}_c^E$ at catalog level (Eq. 2) and $\mathbf{P}_w^E$ at lexical level (Eq. 5), respectively. Thus, we build a hierarchical emotional classification loss $\mathcal{L}_{cls}^{E}$. The $\mathcal{L}_{cls}^{E}$ requires the model to predict the probabilities of the correct psychological category and correct emotion word as high as possible. The whole process is formulated as follows:

$$\mathcal{L}_{cls}^{E} = -\sum_{x \in \hat{\mathbf{Y}} \cap \mathbf{E}_c} log\mathbf{P}_c^E(x) - \sum_{x \in \hat{\mathbf{Y}} \cap \mathbf{E}_w} log\mathbf{P}_w^E(x), \tag{10}$$

where x denotes the emotion word or category contained in the ground-truth caption $\hat{\mathbf{Y}}$. $\mathbf{P}_c^E(x)$ and $\mathbf{P}_w^E(x)$ represent the predicted probability of token x in $\mathbf{P}_c^E$ and $\mathbf{P}_w^E$, respectively.

3.3.3 **Emotional Contrastive Loss.** In this part, we focus on semantic emotion verification. To be specific, we select emotionally positive and negative video-caption pairs for contrastive learning [53, 55]. For a generated caption, we take the input video as the positive sample and choose a video *with a different emotion category* in the training batch as the negative sample. We maximize the semantic relevance of the positive video-caption pair while minimizing that of the negative pair. For instance, a caption with "joy" emotion is close to its paired video with the same emotion, but far away from a video with "amaze" emotion. In this way, the captioning model is encouraged to better capture the underlying emotional relationship between the video and the caption.

As the decoding process is implemented step by step, we devise a contrastive learning objective along the timeline. Given video features $\mathbf{V}$ and its generated caption features $\mathbf{W}^{(t)}$ at the t-th decoding step, we calculate the positive relevance score $r_{i,j}^{(t)}$ between textual caption feature and visual feature of the given video in Eq. (6). In the same way, we obtain the negative relevance score $\hat{r}_{i,j}^{(t)}$ between textual caption feature and visual feature of a negative video. The emotional contrastive loss is formulated as follows:

$$\mathcal{L}_{ctr}^{E} = -log[\sigma(\sum_t \sum_{i,j} r_{i,j}^{(t)})] - log[1 - \sigma(\sum_t \sum_{i,j} \hat{r}_{i,j}^{(t)})], \tag{11}$$

where $\sigma(\cdot)$ is a sigmoid function.

At last, our model can be trained end-to-end by minimizing the weighted sum of all losses:

$$\mathcal{L} = \lambda_{ce} \mathcal{L}_{ce}^{E} + \lambda_{cls} \mathcal{L}_{cls}^{E} + \lambda_{ctr} \mathcal{L}_{ctr}^{E}. \tag{12}$$

where λ_{ce}, λ_{cls}, and λ_{ctr} are three hyper-parameters.

4 EXPERIMENT

In this section, we quantitatively and qualitatively conduct extensive experiments to answer the following research questions: **R1:** Can the EPAN effectively improve the performance of existing methods to reach SOTA level? (*w.r.t.* Table 1) **R2:** Can tree-structured emotion learning effectively extract emotional cues from videos and improve the caption quality? (*w.r.t.* Table 2, Figure 4, Figure 5, and Figure 6) **R3:** Will the three emotion-aware learning objectives benefit the model? (*w.r.t.* Table 2 and Figure 8) **R4:** Can EPAN generate semantically and emotionally correct video captions? (*w.r.t.* Figure 7) **R5:** How is the generalization performance of EPAN when extended to images? (*w.r.t.* Table 4)

4.1 Experimental Setup

Dataset. We experiment on three available emotional video-caption benchmarks, *i.e.*, EVC-MSVD [3], EVC-VE [16], and EVC-Combined [43]. **EVC-MSVD** contains 374 videos from the typical caption dataset MSVD [3]. Each video has about 40 emotional captions. The EVC-MSVD is divided into 240/134 videos for training/testing, respectively. **EVC-VE** contains 1,523 videos from the emotion classification dataset VideoEmotion-8 [16]. Each video has around 17 emotional captions. The EVC-VE is divided into 1141/382 videos for training/testing, respectively. Compared with EVC-MSVD, EVC-VE contains longer videos (average 27s vs. 10s per video) and annotations (average 13 vs. 7 tokens per caption). **EVC-Combined** dataset is the combination of EVC-MSVD and EVC-VE. It contains 1,381 videos for training and 516 videos for testing. Since the EVC-Combined dataset covers more scenarios, we pay more attention to the evaluation on this dataset. Besides, **SentiCap** [27] is a public emotional image captioning dataset without emotion polarity label.

Table 1: Performance comparison of emotion, semantic, and hybrid evaluations on EVC-MSVD, EVC-VE, and EVC-Combined.

Features	Methods	Emotion		Semantic							Hybrid	
		Acc_{sw}	Acc_c	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	BFS	CFS
				EVC-MSVD								
R152	FT [43]	69.4	67.1	77.2	60.3	47.4	36.3	29.0	63.4	62.5	52.5	63.7
	CANet [36]	69.6	66.4	78.0	62.1	50.0	38.6	29.8	63.5	63.3	54.1	64.2
	EPAN	**79.0**	**76.1**	**80.4**	**64.7**	**53.0**	**43.1**	**32.0**	**66.4**	**69.8**	**58.8**	**71.4**
R101+RN	CANet [36]	78.7	76.8	78.5	64.0	52.1	41.8	30.8	65.7	74.4	57.9	75.1
	EPAN	**84.4**	**82.8**	**79.8**	**65.8**	**53.5**	**41.9**	**33.0**	**66.8**	**75.7**	**59.9**	**77.3**
CLIP	SA-LSTM† [41]	68.8	67.2	80.7	67.9	56.3	45.5	33.0	68.2	72.1	59.0	71.3
	EPAN	**84.1**	**82.8**	**82.5**	**69.6**	**57.8**	**46.2**	**34.4**	**69.8**	**80.6**	**63.1**	**81.1**
				EVC-VE								
R152	CANet† [36]	42.5	40.2	66.0	44.6	29.1	18.8	17.7	37.7	23.9	33.7	27.4
	EPAN	**47.5**	**45.8**	**68.4**	**46.8**	**31.2**	**20.8**	**18.7**	**38.9**	**26.4**	**36.5**	**30.4**
R101+RN	CANet† [36]	41.9	39.7	66.9	44.8	29.3	19.3	18.2	37.9	23.3	33.9	26.8
	EPAN	**49.3**	**47.9**	**67.7**	**45.9**	**30.9**	**21.1**	**18.5**	**38.3**	**24.8**	**36.6**	**29.5**
CLIP	SA-LSTM† [41]	48.6	47.1	71.0	51.1	34.5	22.5	19.6	40.7	30.2	38.9	33.7
	EPAN	**63.8**	**62.3**	**73.6**	**54.0**	**38.3**	**27.0**	**21.2**	**42.3**	**34.7**	**45.0**	**40.4**
				EVC-Combined								
R152	FT [43]	51.2	49.6	67.6	47.2	32.0	**21.6**	**20.4**	**43.1**	29.0	37.6	33.3
	EPAN	**53.3**	**51.5**	**68.0**	**47.4**	**32.1**	21.5	19.7	42.8	**31.3**	**38.1**	**35.5**
R101+RN	CANet [36]	53.7	52.7	68.1	47.7	32.9	22.5	19.7	43.7	34.5	38.8	38.2
	EPAN	**60.3**	**59.7**	**70.7**	**50.7**	**35.4**	**24.5**	**20.8**	**46.0**	**36.9**	**42.1**	**41.6**
CLIP	SA-LSTM† [41]	53.4	50.7	70.6	51.4	36.7	25.4	21.0	45.9	38.8	41.2	41.5
	EPAN	**69.3**	**67.2**	**74.4**	**55.6**	**39.9**	**28.0**	**23.0**	**47.1**	**43.0**	**47.0**	**48.0**

† reconstructed results. R, RN and CLIP indicate using ResNet [12], 3D-ResNext-101 [11] and CLIP [34] for visual feature extraction, respectively.

We annotate it manually and perform experiments on it to discuss the versatility of the model.

Evaluation Metrics. To measure the factual semantics of generated captions, we use the widely-used metrics [8, 37, 42] of **BLEU**, **METEOR**, **ROUGE**, and **CIDEr**, abbreviated to B, M, R, and C, respectively. Following the convention [36, 43], we also consider the emotion evaluation with the metrics of emotion word accuracy $\mathbf{Acc}_{sw}$ and emotion sentence accuracy $\mathbf{Acc}_c$. Moreover, there are two hybrid metrics **BFS** and **CFS** [43] that combine the emotion evaluation with BLEU and CIDEr metrics, respectively.

Implementation Details. Each video is uniformly sampled into N = 30 keyframes and segments [54, 56, 70]. For fair comparison [36, 43], we extract appearance features using ResNet-101 [21] or ResNet-152 [4, 12] and motion features using 3D-ResNext-101 [5, 11]. Besides, we test the model with the modern CLIP features [34]. About text captions, each sentence is tokenized and truncated into 15 words. We convert all the words into lowercase and remove punctuations. As for emotion learning, there are M_c = 34 emotion categories and M_w = 179 emotion words as in [43]. Both the text caption and emotion word embeddings are initialized by using GloVe [7, 20, 32]. In our work, the feature dimension is set to $d_v = d_w = d_e$ = 300. Finally, for caption generation, we build a vocabulary that contains all words in the training corpus and emotion lexicon. The vocabulary sizes for the EVC-MSVD, EVC-VE, and EVC-Combined datasets are 9,637, 13,980, and 14,034, respectively.

For model implementation, we take the Transformer [10, 40, 60] as our encoder backbone. We set the layer number and the multi-head number to 1 and 1 for both video and text feature preparation, and set them to 4 and 4 for hierarchical emotion encoding (including both catalog and lexicon encoders). The hidden size of the LSTM-based caption decoder is set to d_h = 512. Empirically, we set λ_{ce}, λ_{cls}, and λ_{ctr} to 1, 0.2 and 0.2, respectively. The penalty coefficient is set to α = 0.1. We set the optimal $topK$ = 1 in the following experiments (please see the detailed ablation of different $topK$ in Appendix Table A1). Following the setup in [43], the decoder is initialized on MSVD [3]. For training, we adopt the Adam optimizer [6, 71] with a learning rate of 7e-4, and the batch size is set to 128. During testing, we use beam search with size 5 to generate captions.

4.2 Main Comparison

R1: Comparison with State-of-the-Art Methods. In Table 1, we present the comparison of EPAN and existing methods on the three EVC datasets. We have the following observations:

- Our method achieves the best performances on emotion metrics Acc_{sw} and Acc_c across the three datasets. For instance, on the EVC-MSVD dataset, the proposed EPAN brings remarkable $\text{Acc}_{sw}/\text{Acc}_c$ improvements, *i.e.*, +13.8%/+13.4% improvements over FT [43] and +13.5%/+14.6% improvements over CANet [36].
- Our model also significantly improves the semantic metrics of SOTA methods. On the three datasets, EPAN outperforms FT [43] and CANet [36] by a large margin on almost all metrics. One exception is that on the EVC-Combined dataset, EPAN performs better than FT [43] on BLEU-1~3 and CIDEr while slightly worse on BLEU-4, METEOR, and ROUGE. It may be because our explicit emotional learning on the emotion tree enables diverse descriptions that are not limited to ground truth.
- Furthermore, we investigate the potential of EPAN by using the modern CLIP features [34]. As shown in Table 1, EPAN sets a new record for state-of-the-art performances, reporting 81.1,

Table 2: Ablation studies of the tree-structured emotion learning (TEL) and losses on the EVC-Combined dataset.

Method	Acc_{sw}	Acc_c	B-1	B-2	B-3	B-4	M	R	C	BFS	CFS
w/o TEL	57.2	54.8	71.8	53.0	37.8	26.0	22.0	46.8	40.6	42.8	43.7
w/ catalog	59.9	57.5	73.0	54.6	39.2	28.0	22.6	47.0	41.5	44.7	44.9
w/ lexicon	61.4	59.5	73.8	54.9	39.6	27.9	22.4	47.0	42.6	45.2	46.2
w/ $\mathcal{L}_{ce}$	61.8	60.5	73.1	54.3	39.0	27.6	22.5	46.7	42.5	45.0	46.2
w/o $\mathcal{L}^{E}_{cls}$	62.2	60.2	73.5	54.2	38.8	27.6	22.6	47.0	41.3	44.9	45.3
w/o $\mathcal{L}^{E}_{ctr}$	63.5	61.2	72.6	53.3	38.2	27.2	22.6	46.7	41.7	44.7	45.8
w/ $\mathcal{L}_{ctr}$	68.3	66.8	73.7	54.9	39.5	27.8	22.6	**47.3**	41.7	46.6	46.8
EPAN	**69.3**	**67.2**	**74.4**	**55.6**	**39.9**	**28.0**	**23.0**	47.1	**43.0**	**47.0**	**48.0**

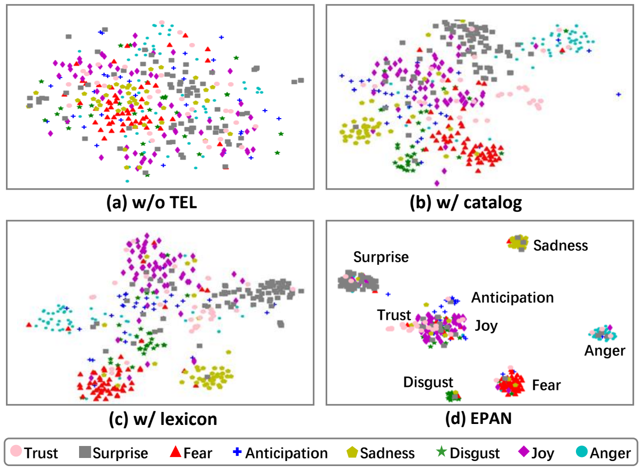


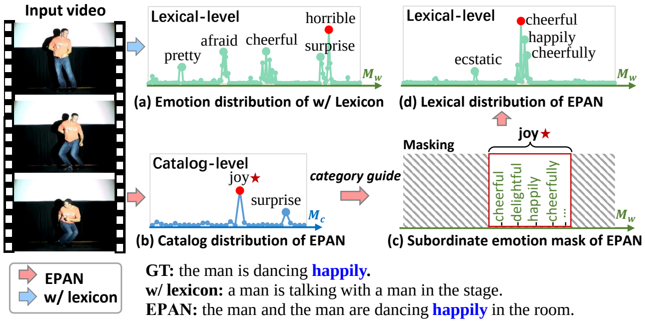
Figure 4: The t-SNE embedding of emotion representation $\mathcal{E}$ on the EVC-VE test set [43], in which all videos are collected from the emotion classification dataset VideoEmotion-8 [16].

40.4, and 48.0 on hybrid metric CFS on the three datasets, respectively. We reconstruct a classical framework SA-LSTM [41] with CLIP features as a comparison baseline for fair experiments. By comparison, despite using the same CLIP features, the CFS of SA-LSTM [41] just achieves 71.3, 33.7, and 41.5 on the three datasets, respectively.

4.3 Ablation Studies

R2: Tree-structured Emotion Learning. As shown in Table 2 (top block), we investigate the effect of the tree-structured emotion learning module by comparing the models of "w/o TEL" (*i.e.*, removing the tree-structured emotion learning module), "w/ catalog" (*i.e.*, only using catalog-level emotion learning module), "w/ lexicon" (*i.e.*, only using lexical-level emotion learning module), and full EPAN. The experimental results show that the emotional hints are indeed favorable for emotional captioning. The baseline "w/o TEL" reports the worst performance. Compared to "w/o TEL", either "w/ catalog" or "w/ lexicon" facilitate the performance improvement obviously. The full EPAN achieves the best results on both semantic and emotional metrics, such as that the Acc_{sw} and CIDEr values reach 69.3 and 43.0, respectively.

We further qualitatively discuss the effect of tree-structured emotion learning from three aspects as below. ***1) t-SNE embedding of emotion representation.*** Our approach learns an emotion-discriminative video representation $\mathcal{E}$ as emotion priors to guide the caption generation. In Figure 4, we display the t-SNE embedding of $\mathcal{E}$ on EVC-VE test set. For "w/o TEL" in Figure 4 (a), the videos with various emotion categories are mixed up and can not distinguish each other. By introducing either catalog-level or lexical-level emotion learning into the model, the videos with the same emotion category get together as shown in Figures 4 (b) and (c). The most obvious emotion clustering happens on the full EPAN in Figure 4 (d). ***2) Role of subordinate emotion mask.*** The subordinate emotion mask is to shield irrelevant emotion words along the emotion tree. As shown in Figure 5 (a), the model without emotion mask (*i.e.*, w/ lexicon) performs the confusing emotion perception. As shown in Figure 5 (b~d), EPAN focuses on the closely relevant natural words guided by the catalog-level "joy", and finally generates the correct emotion word. ***3) Catalog- & lexical-level emotion category accuracy.*** As shown in Figure 6, with the catalog-level emotion learning, the emotion category accuracy achieves 76.9%, where the 76.9% effectively guides the subsequent lexical-level learning, and the emotion category accuracy up to 99.6%. Even 23.1% videos are misclassified at the catalog level, of which 10.9% samples are corrected during lexical-level classification. In the final incorrect emotion categories, we find that 62.5% of them have more than 50% semantic similarity (by using the similarity measurement in [30]) with the ground-truth labels, such as "ecstatic" and "joy".



GT: the man is dancing **happily**.
w/ lexicon: a man is talking with a man in the stage.
EPAN: the man and the man are dancing **happily** in the room.

Figure 5: Example of emotional mask mechanism. Guided by the catalog-level "joy", EPAN focuses on the closely relevant natural words and predicts the correct emotion word.

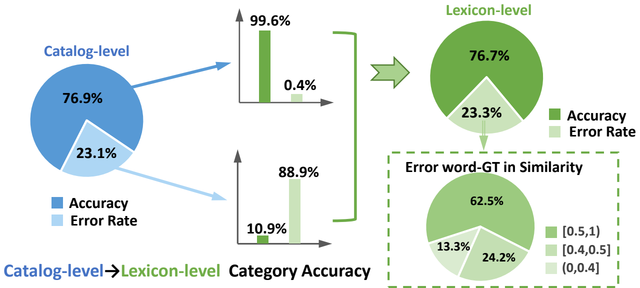
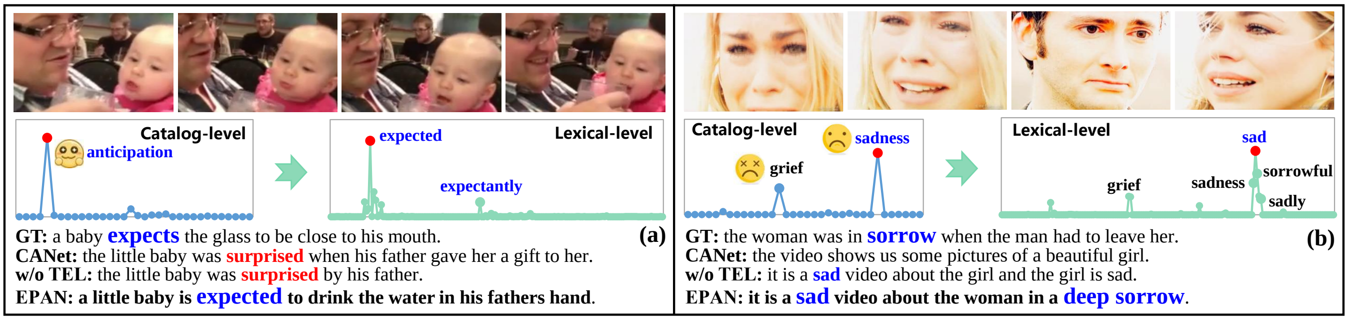


Figure 6: Emotion category accuracy at catalog level and lexical level. The histograms show the accuracy rate from the catalog level to the lexical level.

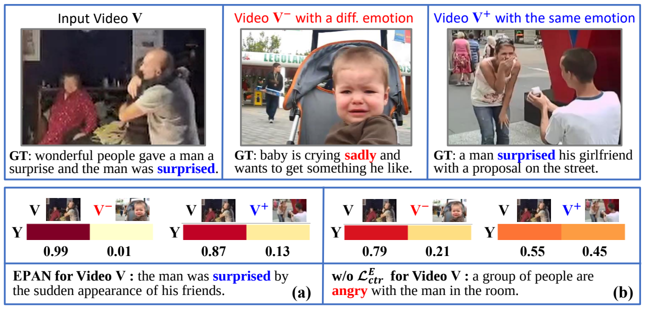
R3: Emotion-aware Optimization Objectives. From Table 2 (bottom block), we observe that the absence of each loss leads to performance degradation. **1)** By replacing $\mathcal{L}^{E}_{ce}$ with the traditional cross-entropy loss [58] (denoted as w/ $\mathcal{L}_{ce}$), the Acc_{sw} drops from 69.3 to 61.8. The penalization term in $\mathcal{L}^{E}_{ce}$ ensures sufficient training for those few but meaningful emotion words in the corpus. **2)** If remove $\mathcal{L}^{E}_{cls}$, the emotion distributions P^{E}_{c} and P^{E}_{w} are not constrained, thereby resulting in a large performance drop, *e.g.*, the Acc_c drops



GT: a baby **expects** the glass to be close to his mouth.
CANet: the little baby was **surprised** when his father gave her a gift to her.
w/o TEL: the little baby was **surprised** by his father.
EPAN: a little baby is expected to drink the water in his fathers hand.

GT: the woman was in **sorrow** when the man had to leave her.
CANet: the video shows us some pictures of a beautiful girl.
w/o TEL: it is a **sad** video about the girl and the girl is sad.
EPAN: it is a sad video about the woman in a deep sorrow.

Figure 7: Affective captioning results from CANet [36], w/o TEL (*i.e.*, removing tree-structured emotion learning module), and EPAN models. Correct emotion words are marked in blue font, while the red font denotes the wrong emotion words.



GT: wonderful people gave a man a surprise and the man was **surprised**.

GT: baby is crying **sadly** and wants to get something he like.

GT: a man **surprised** his girlfriend with a proposal on the street.

EPAN for Video V : the man was **surprised** by the sudden appearance of his friends.

w/o $\mathcal{L}_{ctr}^{E}$ for Video V : a group of people are **angry** with the man in the room.

Figure 8: Visualization of the video-caption relevance computed by EPAN and w/o $\mathcal{L}_{ctr}^{E}$. Y denotes the generated caption for input video V by EPAN or w/o $\mathcal{L}_{ctr}^{E}$.

from 67.2 to 60.2. **3)** As for the contrastive loss $\mathcal{L}_{ctr}^{E}$, it enforces the semantic relevance of emotionally positive video-caption pairs to be much greater than negative ones. The model without $\mathcal{L}_{ctr}^{E}$ shows obvious performance drop, *e.g.*, reducing Acc$_{sw}$/BLEU-1 by 8.4%/2.4%, respectively. **4)** Besides, we study the contrastive strategy of $\mathcal{L}_{ctr}^{E}$. Our emotional contrastive strategy performs much better than using the random contrastive strategy (denoted as w/ $\mathcal{L}_{ctr}$), *e.g.*, lifting the CIDEr from 41.7 to 43.0 and the emotion accuracy Acc$_{sw}$/Acc$_c$ from 68.3/66.8 to 69.3/67.2, respectively.

In Figure 8, we show three videos $\{\mathbf{V}, \mathbf{V}^+, \mathbf{V}^-\}$ sampled by our emotional contrastive strategy, where $\mathbf{V}$ and $\mathbf{V}^+$ are annotated with the same emotion "surprise" and $\mathbf{V}^-$ is annotated with a different emotion "sadness". For video $\mathbf{V}$, we obtain the generated caption $\mathbf{Y}$ by using EPAN or w/o $\mathcal{L}_{ctr}^{E}$. Next, we calculate the relevance scores of Y with $\mathbf{V}$, $\mathbf{V}^-$, and $\mathbf{V}^+$ by using Eq. (6), respectively. We normalize the sum of relevance scores for each two comparable video-caption pairs. As shown in Figure 8, there are two clear observations: 1) the contrastive effect of $\{\mathbf{V},\mathbf{V}^-\}$ (0.99/0.01) is much more obvious than $\{\mathbf{V},\mathbf{V}^+\}$ (0.87/0.13) by using EPAN; 2) without $\mathcal{L}_{ctr}^{E}$, the contrastive effect of w/o $\mathcal{L}_{ctr}^{E}$ becomes less than the EPAN, either in the case $\{\mathbf{V},\mathbf{V}^-\}$ (0.79/0.21) or $\{\mathbf{V},\mathbf{V}^-\}$ (0.55/0.45).

R4: Emotional Video Caption Results. To qualitatively illustrate the advantages of our method, Figure 7 visualizes some caption results from CANet [36], "w/o TEL", and EPAN. In Figure 7 (a), both CANet and "w/o TEL" predict incorrect emotion words, while our EPAN predicts the correct ones (*i.e.*, "surprised vs. expected"). In Figure 7 (b), although all the approaches perform well, our model gives richer and more accurate emotions words "sad" and "sorrow". Anyway, the catalog-level and lexical-level emotion distributions learned by our method are consistent with the captions for each sample. We provide more visualization samples in the appendix.

Table 3: Performance comparison for emotional image captioning on SentiCap dataset.

Method	Acc$_{sw}$	Acc$_c$	B-1	B-2	B-3	B-4	M	R	C	BFS	CFS
SA-LSTM† [41]	83.4	83.4	43.3	24.7	14.0	7.9	13.9	34.2	34.8	30.0	44.5
CANet† [36]	84.5	84.5	45.5	26.1	15.2	9.1	14.8	35.3	42.1	31.3	50.6
w/o TEL	84.5	84.3	44.9	25.2	14.2	8.0	15.0	34.2	42.7	30.5	51.0
w/ catalog	85.4	85.2	46.0	26.7	15.6	8.9	15.1	35.2	44.1	31.6	52.3
w/ lexicon	84.9	84.6	45.5	26.2	14.9	8.4	15.0	35.1	44.5	31.0	52.5
EPAN	**86.1**	**86.1**	**48.6**	**28.6**	**16.7**	**9.5**	**15.8**	**37.0**	**47.4**	**32.7**	**55.2**

In our experiment, the images in SentiCap are labeled with emotion polarity.

4.4 Emotional Image Captioning

R5: Generalization Performance on Images. To further evaluate the versatility of the proposed EPAN framework, we extend the EPAN to the emotional image captioning task by using the polarity-labeled SentiCap dataset. SentiCap refers to image captioning with positive/negative emotion polarity. Table 3 reports the experimental results. Compared to CANet [36] and SA-LSTM [41], our method obtains the best performance in both semantic and emotion metrics, demonstrating great versatility. From the results of the ablation models, we observe the same conclusion as in Table 2: acquiring emotion priors via catalog-level and lexical-level emotional learning effectively improves the caption performance, especially the sentiment consistency (*w.r.t.* Acc$_{sw}$ and Acc$_c$). This further indicates the effectiveness of our method.

5 CONCLUSION

This paper presents a novel Emotion-Prior Awareness Network (EPAN) for emotional video captioning, in which the video emotions are first recognized and then used to guide the caption generation. The EPAN leverages the merits of both psychological categories and daily natural words for emotion understanding. Specifically, it performs a tree-structured emotion learning, and introduces a subordinate emotion mask into an end-to-end captioning framework. In addition, three emotion-aware losses are specifically designed for this task. Extensive experiments demonstrate the effectiveness of our method against the state-of-the-art methods on three benchmark datasets, for both emotional and semantic evaluations.

REFERENCES

[1] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. 2021. Artemis: Affective language for visual art. In *CVPR*. 11569–11579.

[2] Pablo Barros, German Parisi, and Stefan Wermter. 2019. A Personalized Affective Memory Model for Improving Emotion Recognition. In *ICML*, Vol. 97. 485–494.

[3] David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *ACL*. 190–200.

[4] Jianfeng Dong, Xianke Chen, Minsong Zhang, Xun Yang, Shujie Chen, Xirong Li, and Xun Wang. 2022. Partially Relevant Video Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*. 246–257.

[5] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. 2021. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 8 (2021), 4065–4080.

[6] Dan Guo, Kun Li, Zheng-Jun Zha, and Meng Wang. 2019. Dadnet: Dilated-attention-deformable convnet for crowd counting. In *Proceedings of the 27th ACM international conference on multimedia*. 1823–1832.

[7] Dan Guo, Hui Wang, and Meng Wang. 2021. Context-aware graph inference with knowledge distillation for visual dialog. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 10 (2021), 6056–6073.

[8] Dan Guo, Wengang Zhou, Houqiang Li, and Meng Wang. 2018. Hierarchical LSTM for sign language translation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[9] Yanbin Hao, Hao Zhang, Chong-Wah Ngo, and Xiangnan He. 2022. Group contextualization for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 928–938.

[10] Yanbin Hao, Hao Zhang, Chong-Wah Ngo, Qiang Liu, and Xiaojun Hu. 2020. Compact bilinear augmented query structured attention for sport highlights classification. In *Proceedings of the 28th ACM international conference on multimedia*. 628–636.

[11] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?. In *CVPR*. 6546–6555.

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

[13] Tao He and Xiaoming Jin. 2019. Image emotion distribution learning with graph convolutional networks. In *ICMR*. 382–390.

[14] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM CsUR* 51, 6 (2019), 1–36.

[15] Vanita Jain, Fadi Al-Turjman, Gopal Chaudhary, Devang Nayar, Varun Gupta, and Aayush Kumar. 2022. Video captioning: a review of theory, techniques and practices. *Springer MTA* 81, 25 (2022), 35619–35653.

[16] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue. 2014. Predicting emotions in user-generated videos. In *AAAI*. 73–79.

[17] Gihwi Kim, Ilyoung Choi, Qinglong Li, and Jaekyeong Kim. 2021. A CNN-based advertisement recommendation through real-time user face recognition. *Applied Sciences* 11, 20 (2021), 9705.

[18] Dimitrios Kollias and Stefanos Zafeiriou. 2020. Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset. *IEEE Transactions on Affective Computing* 12, 3 (2020), 595–606.

[19] Guodun Li, Yuchen Zhai, Zehao Lin, and Yin Zhang. 2021. Similar scenes arouse similar emotions: Parallel data augmentation for stylized image captioning. In *ACM MM*. 5363–5372.

[20] Kun Li, Dan Guo, and Meng Wang. 2021. Proposal-free video grounding with contextual pyramid network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1902–1910.

[21] Kun Li, Jiaxiu Li, Dan Guo, Xun Yang, and Meng Wang. 2023. Transformer-based Visual Grounding with Cross-modality Interaction. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 6 (2023), 1–19.

[22] Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *AAAI*.

[23] Tong Li, Yunhui Hu, and Xinxiao Wu. 2021. Image Captioning with Inherent Sentiment. In *ICME*. IEEE, 1–6.

[24] Yicong Li, Xun Yang, Xindi Shang, and Tat-Seng Chua. 2021. Interventional video relation detection. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4091–4099.

[25] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *CVPR*. 2554–2562.

[26] Ziyu Ma, Fuyan Ma, Bin Sun, and Shutao Li. 2021. Hybrid mutimodal fusion for dimensional emotion recognition. In *MuSe*. 29–36.

[27] Alexander Mathews, Lexing Xie, and Xuming He. 2016. Senticap: Generating image descriptions with sentiments. In *AAAI*. 3574–3580.

[28] Trisha Mittal, Puneet Mathur, Aniket Bera, and Dinesh Manocha. 2021. Affect2MM: Affective Analysis of Multimedia Content Using Emotion Causality. In *CVPR*. 5661–5671.

[29] Omid Mohamad Nezami, Mark Dras, Peter Anderson, and Len Hamey. 2018. Face-cap: Image captioning using facial expression analysis. In *ECML*. 226–240.

[30] Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *ICLR*.

[31] Michal Muszynski, Leimin Tian, Catherine Lai, Johanna D Moore, Theodoros Kostoulas, Patrizia Lombardo, Thierry Pun, and Guillaume Chanel. 2021. Recognizing induced emotions of movie audiences from multimodal information. *IEEE Transactions on Affective Computing* 12, 1 (2021), 36–52.

[32] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.

[33] Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*. Elsevier, 3–33.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, Vol. 139. 8748–8763.

[35] Tianrong Rao, Xiaoxu Li, and Min Xu. 2020. Learning multi-level deep representations for image emotion classification. *Neural Processing Letters* 51, 3 (2020), 2043–2061.

[36] Peipei Song, Dan Guo, Jun Cheng, and Meng Wang. 2022. Contextual Attention Network for Emotional Video Captioning. *IEEE Transactions on Multimedia* (2022), 1–11. https://doi.org/10.1109/TMM.2022.3183402

[37] Shengeng Tang, Richang Hong, Dan Guo, and Meng Wang. 2022. Gloss semantic-enhanced network with online back-translation for sign language production. In *Proceedings of the 30th ACM International Conference on Multimedia*. 5630–5638.

[38] Yu-Chih Tsai, Tse-Yu Pan, Ting-Yang Kao, Yi-Hsuan Yang, and Min-Chun Hu. 2022. EMVGAN: Emotion-Aware Music-Video Common Representation Learning via Generative Adversarial Networks. In *MMArt*. 13–18.

[39] Kohei Uehara, Yusuke Mori, Yusuke Mukuta, and Tatsuya Harada. 2022. ViNTER: Image Narrative Generation with Emotion-Arc-Aware Transformer. *arXiv preprint arXiv:2202.07305* (2022).

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

[41] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. 2018. Reconstruction network for video captioning. In *CVPR*. 7622–7631.

[42] Bo Wang, Zhao Zhang, Jicong Fan, Mingbo Zhao, Choujun Zhan, and Mingliang Xu. 2022. FineFormer: Fine-Grained Adaptive Object Transformer for Image Captioning. In *2022 IEEE International Conference on Data Mining (ICDM)*. IEEE, 508–517.

[43] Hanli Wang, Pengjie Tang, Qinyu Li, and Meng Cheng. 2022. Emotion Expression with Fact Transfer for Video Description. *IEEE Transactions on Multimedia* 24 (2022), 715–727.

[44] Lin Wang, Xiangmin Xu, Fang Liu, Xiaofen Xing, Bolun Cai, and Weirui Lu. 2019. Robust emotion navigation: Few-shot visual sentiment analysis by auxiliary noisy data. In *ACIIW*. 121–127.

[45] Shangfei Wang and Qiang Ji. 2015. Video affective content analysis: a survey of state-of-the-art methods. *IEEE TAC* 6, 4 (2015), 410–430.

[46] Junbin Xiao, Xindi Shang, Xun Yang, Sheng Tang, and Tat-Seng Chua. 2020. Visual relation grounding in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 447–464.

[47] Haitao Xiong, Hongfu Liu, Bineng Zhong, and Yun Fu. 2019. Structured and sparse annotations for image emotion distribution learning. In *AAAI*. 363–370.

[48] Liwen Xu, Zhengtao Wang, Bin Wu, and Simon Lui. 2022. MDAN: Multi-level Dependent Attention Network for Visual Emotion Analysis. In *CVPR*. 9479–9488.

[49] Jingyuan Yang, Xinbo Gao, Leida Li, Xiumei Wang, and Jinshan Ding. 2021. SOLVER: Scene-Object Interrelated Visual Emotion Reasoning Network. *IEEE Transactions on Image Processing* 30 (2021), 8686–8701.

[50] Jingyuan Yang, Jie Li, Leida Li, Xiumei Wang, and Xinbo Gao. 2021. A Circular-Structured Representation for Visual Emotion Distribution Learning. In *CVPR*. 4237–4246.

[51] Jingyuan Yang, Jie Li, Xiumei Wang, Yuxuan Ding, and Xinbo Gao. 2021. Stimuli-aware visual emotion analysis. *IEEE Transactions on Image Processing* 30 (2021), 7432–7445.

[52] Lin Yang, Yi Shen, Yue Mao, and Longjun Cai. 2022. Hybrid curriculum learning for emotion recognition in conversation. In *AAAI*. 11595–11603.

[53] Xun Yang, Jianfeng Dong, Yixin Cao, Xun Wang, Meng Wang, and Tat-Seng Chua. 2020. Tree-augmented cross-modal encoding for complex-query video retrieval. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 1339–1348.

[54] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1–10.

[55] Xun Yang, Xueliang Liu, Meng Jian, Xinjian Gao, and Meng Wang. 2020. Weakly-supervised video object grounding by exploring spatio-temporal contexts. In *Proceedings of the 28th ACM international conference on multimedia*. 1939–1947.

[56] Xun Yang, Shanshan Wang, Jian Dong, Jianfeng Dong, Meng Wang, and Tat-Seng Chua. 2022. Video moment retrieval with cross-modal neural architecture search. *IEEE Transactions on Image Processing* 31 (2022), 1204–1216.

[57] Zhengyuan Yang, Yixuan Zhang, and Jiebo Luo. 2019. Human-centered emotion recognition in animated gifs. In *ICME*. 1090–1095.

[58] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *ICCV*. 4507–4515.

[59] Chi Zhan, Dongyu She, Sicheng Zhao, Ming-Ming Cheng, and Jufeng Yang. 2019. Zero-shot emotion recognition via affective structural embedding. In *ICCV*. 1151–1160.

[60] Hao Zhang, Yanbin Hao, and Chong-Wah Ngo. 2021. Token shift transformer for video classification. In *Proceedings of the 29th ACM International Conference on Multimedia*. 917–925.

[61] Haimin Zhang and Min Xu. 2021. Recognition of Emotions in User-generated Videos through Frame-level Adaptation and Emotion Intensity Learning. *IEEE Transactions on Multimedia* (2021).

[62] Wei Zhang, Xuanyu He, and Weizhi Lu. 2019. Exploring discriminative representations for image emotion recognition with CNNs. *IEEE Transactions on Multimedia* 22, 2 (2019), 515–523.

[63] Sicheng Zhao, Xuanbai Chen, Xiangyu Yue, Chuang Lin, Pengfei Xu, Ravi Krishna, Jufeng Yang, Guiguang Ding, Alberto L Sangiovanni-Vincentelli, and Kurt Keutzer. 2021. Emotional semantics-preserved and feature-aligned cyclegan for visual emotion adaptation. *IEEE Transactions on Cybernetics* (2021).

[64] Sicheng Zhao, Amir Gholaminejad, Guiguang Ding, Yue Gao, Jungong Han, and Kurt Keutzer. 2019. Personalized emotion recognition by personality-aware high-order learning of physiological signals. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, 1s (2019), 1–18.

[65] Sicheng Zhao, Chuang Lin, Pengfei Xu, Sendong Zhao, Yuchen Guo, Ravi Krishna, Guiguang Ding, and Kurt Keutzer. 2019. Cycleemotiongan: Emotional semantic consistency preserved cyclegan for adapting image emotions. In *AAAI*, Vol. 33. 2620–2627.

[66] Sicheng Zhao, Yunsheng Ma, Yang Gu, Jufeng Yang, Tengfei Xing, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. 2020. An End-to-End visual-audio attention network for emotion recognition in user-generated videos. In *AAAI*. 303–311.

[67] Sicheng Zhao, Xingxu Yao, Jufeng Yang, Guoli Jia, Guiguang Ding, Tat-Seng Chua, Bjoern W Schuller, and Kurt Keutzer. 2021. Affective image content analysis: Two decades review and new perspectives. *IEEE TPAMI* 44, 10 (2021), 6729–6751.

[68] Wentian Zhao, Xinxiao Wu, and Xiaoxun Zhang. 2020. Memcap: Memorizing style knowledge for image captioning. In *AAAI*. 12984–12992.

[69] Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022. Cauain: Causal aware interaction network for emotion recognition in conversations. In *IJCAI*. 4524–4530.

[70] Qi Zheng, Jianfeng Dong, Xiaoye Qu, Xun Yang, Yabing Wang, Pan Zhou, Baolong Liu, and Xun Wang. 2023. Progressive localization networks for language-based moment localization. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 2 (2023), 1–21.

[71] Jinxing Zhou, Dan Guo, and Meng Wang. 2022. Contrastive positive sample propagation along the audio-visual event line. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

A APPENDIX

A.1 Additional Experimental Results

A.1.1 Effect of Different top-K. In Table 4, we test the effect of different top-K. “All” denotes the setup of considering all M_c emotion categories. We observe that the best performance is achieved at K = 1, where Acc$_{sw}$ and Acc$_c$ reach 69.3 and 67.2, respectively. In this case, K = 1 is strict and exclusive, and the emotion representation $\mathcal{E}$ is not expected to be disturbed by the other emotions. From the results in Table 4, the cases of K >1 bring some irrelevant emotional noises. This seems that a dominant emotion benefits describing the video more accurately.

A.1.2 Reliability of Two-level Emotion Learning. **Emotion Recognition Accuracy.** With the catalog- and lexical-level emotion distributions predicted by “w/ catalog”, “w/ lexicon”, and the full EPAN, we calculate the classification accuracy at M_c = 34 categories, respectively. As shown in Table 5, compared with “w/ catalog”, the emotion category accuracy of EPAN is higher. This indicates that lexical-level learning under category guidance can inversely promote category-level learning. Besides, the Top-1 category accuracy of the $\mathbf{P}_w^E$ predicted by EPAN is higher than that by “w/ lexicon”, while Top-3 and Top-5 emotion accuracy is lower. A possible reason is that we adopt K = 1 category filtration in hierarchical emotion learning, making the model focus more on accurate predictions of Top-1 emotion.

In addition, we test the classification accuracy over M_c = 34 categories and M_w = 179 emotion words, respectively. Under a fair experimental setting, we reproduce a video emotion recognition method VAANet [66] for comparison. As shown in Table 6, our EPAN consistently outperforms VAANet [66] at both emotion levels. The results further demonstrate the reliability of our tree-structured emotion learning.

Challenging Emotion Recognition Examples. Figure 9 displays four challenging examples with compound emotion responses. In these complex cases, emotion perception is easily misled by confusing vision (*e.g.*, Figures 9 (a)~(b)), and the emotions are sometimes contradictive (*e.g.*, Figure 9 (c)) or indistinguishable (*e.g.*, Figure 9 (d)). We visualize the emotion distributions predicted by EPAN and by “w/ catalog” and “w/ lexicon” that perform only one-level emotion learning.

Taking Figure 9 (a) as an example, the fact that a man quickly drinks a bottle is amazing, but the vision of drinking a bottle is strange and disgusting. Both “w/ catalog” and “w/ lexical” predict the wrong emotion “disgusting”. And our EPAN accurately recognizes the proper emotion of “surprise” and “amazing” at catalog-level and lexical-level emotion learning. To summarize, for caption generation, we better determine the dominant emotions to describe the videos. All the samples in Figure 9 show the robustness and interpretability of EPAN to achieve this goal. The effectiveness of EPAN is attributed to the subordination relation between catalog-level and lexical-level, which strongly constrains the emotional semantics consistency.

A.2 Model Complexity

In this work, all experiments are conducted using an NVIDIA GeForce RTX 2080 Ti GPU. We used PyTorch framework for development. As shown in Table 7, the model sizes of CANet [36] and our EPAN are 54.86MB and 67.11MB, the training speeds are 8.9ms and 9.4ms per video, and the inference speeds are 5.7ms and 6.0ms per video, respectively. Both models converge within 30 epochs. With comparable computational complexity, our EPAN achieves better performance, *e.g.*, CFS increases by 10% compared with CANet [36] on the EVC-VE dataset.

Table 4: Ablation studies of top-K emotion categories on the EVC-Combined test dataset.

top-K	Acc$_{sw}$	Acc$_c$	B-1	B-2	B-3	B-4	M	R	C	BFS	CFS
1	**69.3**	**67.2**	**74.4**	**55.6**	**39.9**	**28.0**	**23.0**	47.1	43.0	**47.0**	**48.0**
5	64.9	63.6	73.0	53.6	38.3	26.9	22.8	47.1	40.2	45.1	45.0
10	64.4	63.2	73.7	54.4	39.1	27.7	22.3	47.0	42.2	45.6	46.5
20	63.6	61.4	74.0	54.7	39.0	27.1	22.2	**47.2**	41.5	45.2	45.7
All	62.6	60.4	73.7	54.6	39.2	27.7	22.3	**47.2**	**43.4**	45.2	47.0

Table 5: Emotion recognition accuracy of catalog- and lexical-level distributions on the EVC-Combined set.

Method	Distribution	Top-1 Acc	Top-3 Acc	Top-5 Acc
w/ catalog	$\mathbf{P}_c^E$	75.2%	89.0%	94.2%
EPAN		**76.9%**	**90.5%**	**94.8%**
w/ lexicon	$\mathbf{P}_w^E$	75.2%	**85.1%**	**89.1%**
EPAN		**76.7%**	81.2%	83.5%

Table 6: Comparison with video emotion recognition method on the EVC-Combined set.

Emotion Level	Method	Top-1 Acc	Top-3 Acc	Top-5 Acc
Catalog-level	VAANet* [66]	53.9%	86.0%	90.1%
	EPAN	**76.9%**	**90.5%**	**94.8%**
Lexical-level	VAANet* [66]	44.6%	64.9%	78.5%
	EPAN	**66.1%**	**77.5%**	**80.8%**

* For a fair comparison, we reproduce the VAANet [66] by removing the audio branch and adopting the same feature.

Table 7: Analysis of speed and model complexity on the EVC-VE dataset.

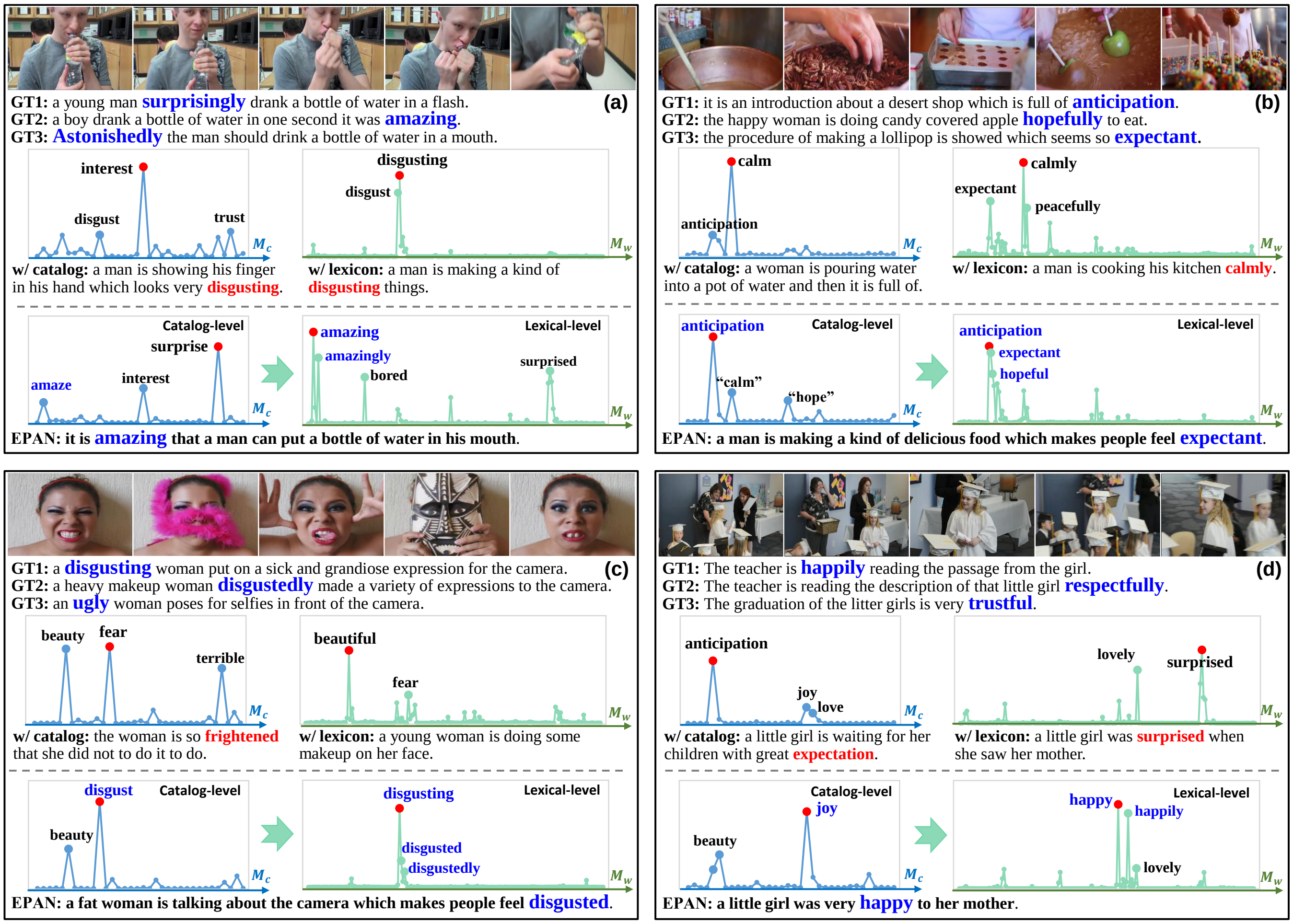
Method	CFS	Speed		Size
		Training	Testing	
CANet* [36]	26.8	8.9ms/video	5.7ms/video	54.86MB
EPAN	29.5	9.4ms/video	6.0ms/video	67.11MB

* indicates our re-implementation.

A.3 Emotion Tree

Figure 10 elaborates on the emotion tree used in this work. Concretely, the catalog level contains 34 emotion categories and the lexical level contains 179 natural emotion words as in [43].

Actually, 179 natural emotion words have been already annotated with each corresponding emotion category label by Wang *et al.*



(a)

GT1: a young man **surprisingly** drank a bottle of water in a flash.
GT2: a boy drank a bottle of water in one second it was **amazing**.
GT3: **Astonishedly** the man should drink a bottle of water in a mouth.

w/ catalog: a man is showing his finger in his hand which looks very **disgusting**.

w/ lexicon: a man is making a kind of **disgusting** things.

EPAN: it is amazing that a man can put a bottle of water in his mouth.

(b)

GT1: it is an introduction about a desert shop which is full of **anticipation**.
GT2: the happy woman is doing candy covered apple **hopefully** to eat.
GT3: the procedure of making a lollipop is showed which seems so **expectant**.

w/ catalog: a woman is pouring water into a pot of water and then it is full of.

w/ lexicon: a man is cooking his kitchen **calmly**.

EPAN: a man is making a kind of delicious food which makes people feel expectant.

(c)

GT1: a **disgusting** woman put on a sick and grandiose expression for the camera.
GT2: a heavy makeup woman **disgustedly** made a variety of expressions to the camera.
GT3: an **ugly** woman poses for selfies in front of the camera.

w/ catalog: the woman is so **frightened** that she did not to do it to do.

w/ lexicon: a young woman is doing some makeup on her face.

EPAN: a fat woman is talking about the camera which makes people feel disgusted.

(d)

GT1: The teacher is **happily** reading the passage from the girl.
GT2: The teacher is reading the description of that little girl **respectfully**.
GT3: The graduation of the litter girls is very **trustful**.

w/ catalog: a little girl is waiting for her children with great **expectation**.

w/ lexicon: a little girl was **surprised** when she saw her mother.

EPAN: a little girl was very happy to her mother.

Figure 9: Visualization of some challenging examples. The EPAN obviously performs better than the ablation variants of "w/ catalog" and "w/ lexicon". This verifies the effectiveness of our tree-structured emotion learning module.

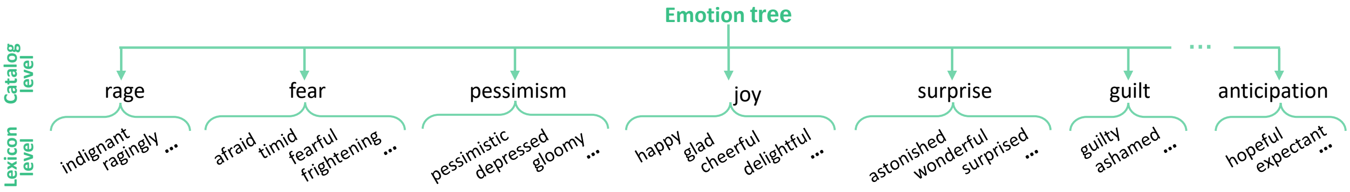


Figure 10: A brief overview of tree-structured emotion architecture.

[43]. Following this protocol, we build the subordinate relationship between each root node and its children nodes (*e.g.*, the relationship between the root node "joy" and its children nodes "happy, glad, cheerful, delightful"), and apply it to the captioning framework. We use this emotion tree for hierarchical emotion learning and emotion-aware optimization in our work. To summarize, our work is the first attempt to leverage explicit and fine-grained emotion modeling for the emotional video captioning task.