CAUSAL-EVC: BREAKING EMOTIONAL SPURIOUS CAUSALITY VIA SPATIOTEMPORAL GROUNDING AND COUNTERFACTUAL INTERVENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Emotional Video Captioning aims to generate factually accurate and emotionally empathetic descriptions. While recent methods have recognized the importance of visual causes to guide emotion perception and caption generation, they fundamentally rely on simple attention matching, which inevitably suffers from causal redundancy and spurious correlations in co-occurrence bias (*e.g.*, misclassifying "sadness" as "joy" on a sunny beach), leading to severe shortcut learning from confusing backgrounds. Furthermore, existing evaluations fail to verify whether models have genuinely mastered causal reasoning or merely exploited background confounders. To address these limitations, we first construct EVC-CauseGround, a comprehensive benchmark with dense spatio-temporal causal annotations. Crucially, it introduces a carefully selected Causal-Faithfulness Subset to explicitly quantify genuine emotion-cause attribution. Second, we propose Causal-EVC, an emotion-grounding captioning framework, which introduces a Motion-guided Causal Spatiotemporal Localization module to precisely decouple causal triggers from background confounders. Besides, we introduce an Interpretable Sparse Emotion Routing module. By synthesizing counterfactual representations and formulating a novel counterfactual contrastive objective, we enforce the model to anchor its emotion predictions strictly on authentic causal triggers instead of confusing background. Extensive experiments show that Causal-EVC not only achieves the best performance on semantic metrics but also exhibits significant advantages in the causal-faithfulness subset, which demonstrates that our model could mine emotional cues from genuine visual causes and mitigate co-occurrence bias for interpretable multimodal emotion understanding.[1]

1 INTRODUCTION

Emotional Video Captioning (EVC) aims to generate factually accurate and emotionally empathetic descriptions for videos (El Assadi et al., 2026; Deng et al., 2026; Zhang et al., 2026). Unlike standard video captioning (Ye et al., 2025a; Li et al., 2026b; Xu et al., 2026), EVC represents a fundamentally more complex endeavor. It requires models to not only cross-modally align factual visual content with natural language, but also accurately perceive implicit affective cues and seamlessly synthesize them into human-like narratives.

Despite recent progress, mainstream EVC methods (Wang et al., 2021; Song et al., 2022; Xu et al., 2024) predominantly map global video representations to predefined emotion spaces via attention patterns. However, this paradigm overlooks the intrinsic nature of visual emotions, which are typically triggered by fine-grained motivational causes rather than global backgrounds. While pioneering efforts like MM-ECPE (Ye et al., 2025b) attempted to explicitly extract emotion-cause pairs, we argue that current cause attribution methods still suffer from two fundamental limitations: **1) Causal Redundancy under Consistent Backgrounds:** As shown in Fig. 1(a), when the contextual cofounder semantically aligns with the true emotion, the background inherently leaks the emotion and generates the causal redundancy. Because MM-ECPE merely maximizes positive mutual information, models effortlessly exploit these environmental confounders to achieve high prediction

[1]Code will be released in the final version of the paper.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

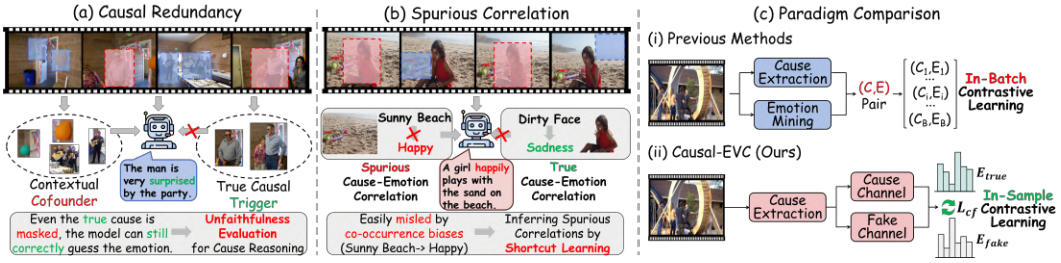

Figure 1: Motivation. Traditional methods suffer from (a) Causal Redundancy and (b) Spurious Correlations, where models exploit shortcut learning and emotional hallucinations due to background confounders. (c) Paradigm Comparison between Causal-EVC and previous methods demonstrate the superiority for genuine emotion-cause reasoning of Causal-EVC.

accuracy while ignoring the authentic causal triggers, failing to penalize this shortcut learning and losing the genuine emotional reasoning capabilities. **2) Spurious Correlations under Inconsistent Backgrounds:** Conversely, when the background conflicts with the true foreground trigger (*e.g.*, a child crying on a sunny beach), traditional patterns are heavily hacked by dataset-level co-occurrence bias (*e.g.*, "Sunny Beach → Joy"). Lacking causal intervention mechanisms, the model is easily misled by the dominant confounder and produces severe emotional hallucinations. Meanwhile, existing benchmarks only evaluate the quality of generated captions and the accuracy of the predicted emotions, which completely fails to quantify the true causal reasoning ability and prevent making lucky statistical guesses based on environmental leakage.

To fundamentally resolve these dilemmas, we tackle the EVC challenge from both benchmark and model perspectives: 1) We construct **EVC-CauseGround**, the first benchmark dedicated to fine-grained emotion causality. We achieve this via two major enhancements: We first introduce dense spatio-temporal causal annotations, defining precise temporal spans and bounding-box tubes for visual causes. Crucially, we collect a pure Causal-Faithfulness Subset, where background leakage is strictly eliminated. Based on it, we design a novel metric ST-IoU to rigorously quantify genuine emotion attribution. 2) We propose **Causal-EVC**, a counterfactual causal grounding framework. Technically, we design a Motion-Guided Causal Spatiotemporal Localization module to guide spatial and temporal attention, generating an accurate causal mask. Guided by it, we physically sever the causal path by synthesizing counterfactual background representations and propose an Interpretable Sparse Emotion Routing module to deduce class-specific emotion probabilities. By enforcing a Counterfactual Contrastive Objective, we explicitly penalize background exploitation, ensuring emotions are reasoned strictly from foreground triggers. As shown in Fig. 2, our model achieves consistent improvements in both caption generation and causal reasoning. In short, our main contributions are summarized as follows:

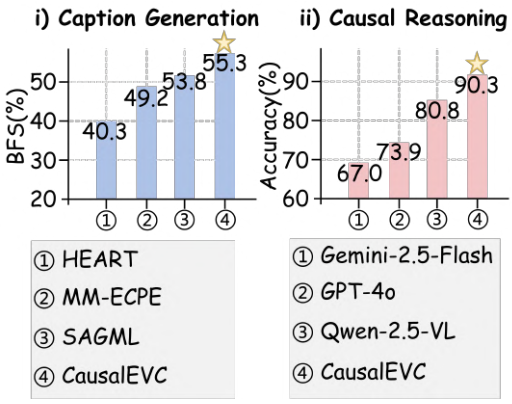

Figure 2: Performance Comparison on the caption generation and causal reasoning.

• **Benchmark**: We construct EVC-CauseGround, a pioneering benchmark with dense spatiotemporal annotations to evaluate the causal attribution capabilities. Besides, we design a convincing metric ST-IoU to explicitly quantify the genuine emotion-cause reasoning ability.

• **Framework**: We propose Causal-EVC, a novel framework with motion-guided cause localization and counterfactual emotion deduction. Besides, we design several objective functions to constrain cause extraction and emotion mining, which breaks spurious correlations and compels the model to reason emotions from true triggers.

• **Performance**: Extensive experiments demonstrate the superiority of Causal-EVC especially in cause grounding, *i.e.,* +17.3%/11.8% improvements on mIoU and emotion accuracy than zero-shot

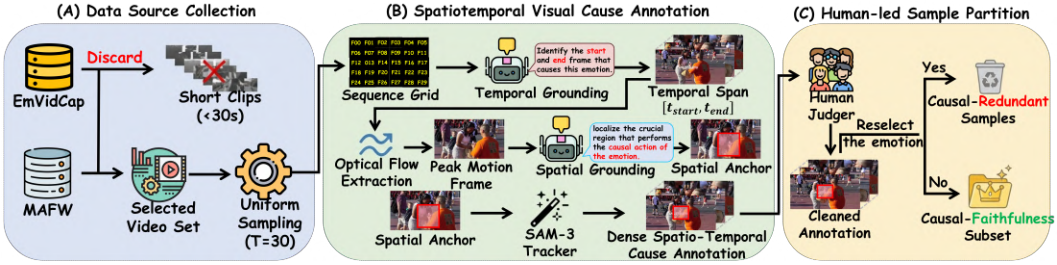

Figure 3: The construction pipeline of the EVC-CauseGround benchmark, which is consist of (A) Data Source Collection, (B) Spatiotemporal Visual Cause Annotation and (C) Human-led Sample Check and Partition, respectively.

Qwen-2.5-VL-7B, which validates that our approach learns genuine emotional causal reasoning capability for better emotional descriptions.

2 EVC-CauseGround: A Visual-Cause Grounding Emotional Descriptive Benchmark

In this paper, we construct EVC-CauseGround, a novel benchmark that extends conventional emotion descriptions with a fine-grained emotional cause annotations. Unlike traditional datasets that provide only holistic labels or coarse utterance-level annotations (Wang et al., 2022; Ma et al., 2024; Zhang et al., 2020), **EVC-CauseGround is the first benchmark to unify fine-grained spatio-temporal causal tubes and a rigorously curated counterfactual subset**[2], establishing a comprehensive benchmark for authentic emotional cause reasoning.

Data Source Collection. We meticulously curate diverse, in-the-wild videos from EmVidCap(Wang et al., 2021) and MAFW(Liu et al., 2022) datasets. Specifically, we discard videos shorter than 5 seconds, as too-short durations make fine-grained causal localization trivial. We finally collect 1,523 videos from EmVidCap and 477 videos from MAFW (averaging $\sim$100s), covering unconstrained scenarios from daily interactions to natural disasters. They are split into 1,500 training and 500 evaluation samples. All selected videos are natively equipped with ground-truth emotion labels and captions, serving as reliable anchors for our subsequent causal annotation.

Spatiotemporal Visual Cause Annotation. To avoid cost-prohibitive manual frame-by-frame drawing, we propose an efficient MLLM-led, tracker-assisted pipeline. We first uniformly sample $T = 30$ frames from each video and concatenate them into a 6×5 grid image with sequence identifiers (e.g., $F00$ to $F29$). We prompt an advanced MLLM (e.g., GPT-5.6 (Singh et al., 2025)) with this grid and the ground-truth emotion to deduce the causal temporal window $[t_{start}, t_{end}]$. To prevent lazy behaviors of MLLMs that include all frames, we constraint that the causal span must not exceed 30% of the total frames. For spatial annotation, we first compute inter-frame optical flow within $[t_{start}, t_{end}]$ to identify the *peak motion frame*, which is the climax of the causal action. By feeding only this high-resolution frame into the MLLM, we obtain a highly confident static bounding box, devoid of multi-frame interference. Utilizing the generated anchor box as an initial prompt, we leverage the zero-shot tracker SAM-3 (Carion et al., 2026) to bidirectionally track the causal object across the temporal span, yielding continuous and dense spatio-temporal causal tubes.

Human-led Sample Check and Partition. While the automated pipeline accelerates annotation, human verification is essential to eliminate hallucinations. Firstly, human annotators are required to check for redundant or missed frames at the boundaries of the temporal spans and spatial regions. Besides, since MLLMs fail to address the *Contextual Leakage* issue, we synthesize a *counterfactual video* for each sample by blacking out the verified causal tube. Then, annotators, blinded to the original labels, are asked to guess the emotion based solely on the masked background. If annotators can still correctly predict the emotion, the sample is flagged as Causal-Redundant (i.e., background confounders are too dominant). Conversely, the video is flagged as Causal-Faithfulness. Finally,

[2]Details and the Pipeline figure of EVC-CauseGround are shown in the Appendix.

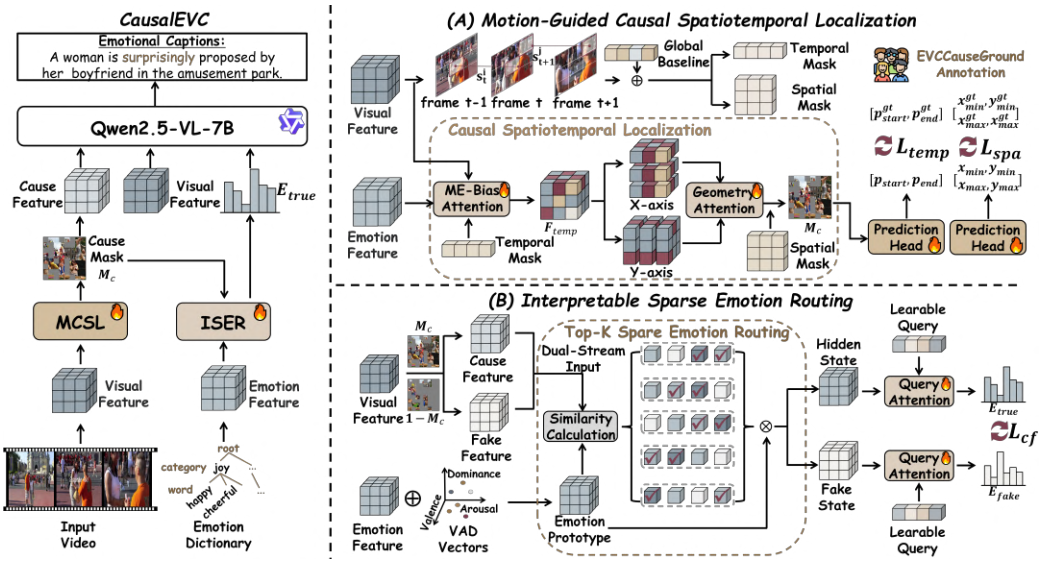

Figure 4: Overview of Causal-EVC. Given the video and emotion dictionary, we perform (A) Motion-Guided Causal Spatiotemporal Localization to extract visual cause masks, and (B) Interpretable Sparse Emotion Routing to mine visual emotion cues for cause-grounding captioning.

we collect a **Causal-Faithfulness Subset** containing 176 videos, which empowers us to evaluate whether a model performs genuine causal reasoning or merely exploits contextual cofounders.

Evaluation Metrics. To effectively evaluate the performance of cause grounding, we design a cascaded metric Spatio-Temporal Grounding Accuracy (ST-IoU). Specifically, we first compute the 1D temporal IoU between the predicted $[t^{pred}_{start}, t^{pred}_{end}]$ and the ground-truth span. If (tIoU $= 0$), the overall score is 0. If $tIoU > 0$, we calculate the 2D spatial IoU ($sIoU$) for each selected frame $t_i \in T_{interest}$. The final ST-IoU is formulated as:

$$\text{ST-IoU} = \text{tIoU} \times \left(\frac{1}{|T_{\text{intersect}}|} \sum_{t_i \in T_{\text{intersect}}} \text{sIoU}_i \right), \tag{1}$$

we report the mIoU, $R@0.3$, and $R@0.5$ to robustly measure causal localization performance.

3 Causal-EVC: A Cause-Grounding Video Captioning Framework

In this section, we describe the Causal-EVC framework for the cause-grounding video captioning task shown in Fig. 4. Given the input video and emotion category, we first encode them to obtain the semantic features. Following the setting of previous works (Song et al., 2023; Ye et al., 2024), we firstly down-sample the video to obtain a frame sequence $\{f_1, f_2 \ldots, f_T\}$, where T denotes the number of frames. Subsequently, we leverage the pre-trained vision encoder of Qwen-2.5-VL (Bai et al., 2025) to extract visual features:

$$\mathcal{V} = \text{VisEncoder}([\text{f}_1, \text{f}_2, \ldots, \text{f}_\text{T}]) \in \mathbb{R}^{\text{T}\times\text{P}\times\text{D}}, \tag{2}$$

where VideoEncoder denotes the visual feature projector and D is the feature dimension. P is the number of patches since each frame is divided into P patches. For the emotion category, we leverage a pre-defined psychology emotional vocabulary dictionary $\Omega = \{w_i\}_{i=1}^{N_w}$, where w_i denotes the i-th emotion word, such as "happy", and N_w denotes the number of emotion words. Then, we leverage the pre-trained Transformer-based emotion embedding, *i.e.,* GloVe (Pennington et al., 2014), to encode these emotion words to $\mathcal{E} \in \mathbb{R}^{N_w \times D}$. The whole process can be formalized as:

$$\mathcal{E} = \text{EmoEncoder}(\Omega) \in \mathbb{R}^{N_w \times D}. \tag{3}$$

These visual and emotional features contain rich semantic information and help to mine accurate emotional causes from video contents.

3.1 MOTION-GUIDED CAUSAL SPATIOTEMPORAL LOCALIZATION

Motion Prior Generation. In video streams, genuine emotional triggers are rarely isolated static objects or a specific token. Instead, they are inherently embedded in the dynamic action states and state transitions. Therefore, we propose to leverage action semantics to serve as a kinematic prior, explicitly collaborating causal localization across both temporal and spatial dimensions. To capture motion prior, we propose a pixel-level motion correlation calculation module, which quantifies motion information through calculating the similarity of all pixel positions between adjacent frames. For the j-th patch of the t-th frame v_t^j, the motion information is quantified as $s_t^j = 1 - \frac{v_{t-1}^j \cdot v_t^j}{||v_{t-1}^j|| ||v_t^j||}$, where $s_t^j \in [0, 2]$ denotes the degree of motion change. The lower the similarity of corresponding patches between adjacent frames, the greater the degree of motion change. Meanwhile, in order to eliminate biases caused by some unexpected disturbances (*i.e.,*camera shake or perspective change), we calculate the global motion baseline through the average of all pixels. The motion information comprehensively considers the global motion baseline and its own motion mask matrix. Finally, we leverage the motion information to generate motion priors:

$$\overline{s} = \frac{1}{T \times P} \sum_{i,j} s_t^j, m_t^j = \mathrm{ReLU}((1 - \lambda_m) s_t^j + \lambda_m \overline{s}), \tag{4}$$

where λ_m is a hyper-parameter to control the weight of the global motion baseline. By aggregating all m_t^j, we obtain an accurate motion prior matrix $M \in \mathbb{R}^{T \times P}$. We decouple the motion prior into temporal and spatial dimensions to achieve coordinated guidance:

$$M_{temp} = \text{Softmax}(\{\sum_{j=1}^{P} m_t^j\}_{t=1}^{T}) \in \mathbb{R}^T, (M_{spa})_t^j = \frac{m_t^j}{\max_{k \in \{1..P\}} m_t^k + \epsilon}, \tag{5}$$

M_{temp} indicates the peaks of the action and M_{spa} represents moving regions in each frame. These constitute a powerful prior that guides two mutually synergistic perspectives on causal exploration.

Temporal Causal Aggregation. To achieve temporal causal aggregation, we introduce a motion-energy bias into the cross-attention mechanism to remove the impact of irrelevant temporal frames:

$$Q_\tau = W_q \mathcal{E}, \quad K_\tau = W_k \mathcal{V}, \quad V_\tau = W_v \mathcal{V}, \tag{6}$$

$$A_\tau = \text{Softmax}\left(\frac{Q_\tau K_\tau^T}{\sqrt{d}} + \alpha \cdot \text{Diag}(\log M_{temp})\right), \mathcal{F}_{temp} = A_\tau V_\tau \in \mathbb{R}^{T \times P \times D}, \tag{7}$$

where α is the scaling parameter and Diag transforms the 1-D curve into a diagonal matrix. The bias term significantly amplifies the weight of frames representing the peak of the movement, thereby focusing more on the emotional climax associated with intense action.

Spatial Causal Localization. Subsequently, we further consider key regions of the emotional causes. We calculate cross-attention along different geometric axes to capture entity boundaries:

$$Q_s^g = W_{s,q}^g \mathcal{E}, \quad K_s^g = W_{s,k}^g \mathcal{F}_{temp}, \quad V_s^g = W_{s,v}^g \mathcal{F}_{temp}, \tag{8}$$

$$A_s^g = \text{Softmax}(Q_s^g K_s^{gT}), \mathcal{O}_s = A_t^y (A_t^x V_s^g)^T \in \mathbb{R}^{T \times P \times D}, \tag{9}$$

where $g \in \{x, y\}$ represents the horizontal and vertical axes. $\mathcal{O}_s$ highlights all salient emotion-related regions in the scene. Finally, we utilize M_{spa} to modulate $\mathcal{O}_s$ and obtain causal features and their corresponding masks:

$$\mathcal{F}_c = \mathcal{O}_s \odot M_{spa} \in \mathbb{R}^{T \times P \times D}, M_c = \sigma(\text{Conv3D}(\mathcal{F}_c)) \in [0, 1]^{T \times P \times 1}, \tag{10}$$

where σ is the Sigmoid activation function. To precisely ground the spatiotemporal labels, we design two MLP heads and objective functions to achieve powerful supervision. Instead of predicting frame-wise and pixel-wise probabilities, we regress the normalized temporal indexes and spatial coordinates. First, we aggregate $\mathcal{F}_c$ to obtain a spatial visual vector $f_s \in \mathbb{R}^{T \times D}$ and a compact visual vector $f_g \in \mathbb{R}^D$:

$$f_s = \text{AvgPool}_P(\mathcal{F}_c), f_g = \text{MaxPool}_T(\text{Conv1D}(f_s)), \tag{11}$$

Then, temporal and spatial heads are used to predict the temporal indexes and spatial coordinates:

$$([\hat{t}_{start}, \hat{t}_{end}], [\hat{x}_{min}^{(t)}, \hat{y}_{min}^{(t)}, \hat{x}_{max}^{(t)}, \hat{y}_{max}^{(t)}]) = (\sigma(\text{MLP}_{temp}(f_g)), \sigma(\text{MLP}_{spa}(f_{s,t}))), \tag{12}$$

where σ is the Sigmoid activation function and MLP is the multi-layer projection.

Causal Grounding Objective. During training, given the ground-truth temporal span $[t^{gt}_{start}, t^{gt}_{end}]$, we apply the SmoothL1 loss (Huber, 1992) to robustly supervise the boundary regression:

$$\mathcal{L}_{temp} = \text{SmoothL1}(\hat{p}_{start}, p^{gt}_{start}) + \text{SmoothL1}(\hat{p}_{end}, p^{gt}_{end}), \tag{13}$$

For the supervision of spatial regions, denoting the predicted and ground-truth bounding box as B^{pred}_t and B^{gt}_t, we employ a combination of SmoothL1 loss and Generalized IoU loss (Hamid et al., 2019), which effectively penalizes scale mismatches and prevents gradient vanishing when there is no overlap between boxes:

$$\mathcal{L}_{GIoU}(B^{pred}_t, B^{gt}_t) = 1 - \left(\text{IoU}(B^{pred}_t, B^{gt}_t) - \frac{\text{Area}(C_t) - \text{Area}(B^{pred}_t \cup B^{gt}_t)}{\text{Area}(C_t)} \right), \tag{14}$$

$$\mathcal{L}_{spa} = \frac{1}{\sum_{t=1}^{T} \mathbb{I}^{gt}_t} \sum_{t=1}^{T} \mathbb{I}^{gt}_t \cdot \left[\lambda_{L1} \mathcal{L}_{L1}(B^{pred}_t, B^{gt}_t) + \lambda_{giou} \mathcal{L}_{GIoU}(B^{pred}_t, B^{gt}_t) \right], \tag{15}$$

where C_t is the minimum enclosing box containing both B^{pred}_t and B^{gt}_t, λ_{L1} and λ_{giou} are balancing weights. $\mathbb{I}^{gt} \in \{0, 1\}^T$ is a temporal indicator mask to ensure only calculating the spatial loss on frames where the genuine causal action occurs.

3.2 Interpretable Sparse Emotion Routing

Emotion Prototype Augmentation. Unlike objective elements, the semantic information of emotion words is characterized by complexity and subjectivity. Relying on a unified text encoder for emotional mining is insufficient. Therefore, we introduce the VAD lexicon (Mohammad, 2018) $R = \{(v_i, a_i, d_i)\}_{i=1}^{N_w} \in \mathbb{R}^{N_w \times 3}$, where each entry is normalized to $[-1, 1]$ (and out-of-lexicon words are set to $\mathbf{0}$), to obtain an interpretable emotion-augmented representation. Given the emotion vocabulary $\mathcal{E}$ and the VAD lexicon, we stack these triplets to obtain the emotion prototype features:

$$\hat{\mathcal{E}} = \text{LayerNorm}(\mathcal{E} \oplus W_e R) \in \mathbb{R}^{N_w \times D}, \tag{16}$$

The $\hat{\mathcal{E}}$ contains both textual semantic and psychological attributes, enhancing the interpretability and facilitating the extraction of emotional semantics from visual information.

Top-K Sparse Emotion Routing. Affective semantics are characterized by significant vagueness and ambiguity. Traditional methods are highly susceptible to interference from irrelevant emotional information. Therefore, we design a sparse emotion routing mechanism. First, instead of using redundant visual features, we employ the cause mask M_c to filter and obtain visual causal features $\mathcal{C} = V \odot M_c$, which are then used to query emotion prototypes. Subsequently, to prevent the traditional attention mechanism from capturing ambiguous emotional semantics, we enforce matching only the top-k emotion categories. Specifically, we first compute the similarity matrix A^e between $\mathcal{C}$ and $\hat{\mathcal{E}}$. Subsequently, for each visual token, we retain only the top-K values and set the rest to $-\infty$, thereby obtaining the sparse attention matrix $\hat{A}_e$ to route the emotion extraction:

$$\hat{A}^e_{i,j} = \begin{cases} \frac{\exp(A^e_{i,j})}{\sum_{m \in \text{Top-K}(A^e_i)} \exp(A_{i,m})}, & \text{if } j \in \text{Top-K}(A^e_i) \\ 0, & \text{otherwise} \end{cases}, \mathcal{H}_{true} = \hat{A}^e(\hat{\mathcal{E}} W_h) \in \mathbb{R}^{T \times D}, \tag{17}$$

$\mathcal{H}_{true}$ denotes the visual hidden states related to the emotions. Finally, we leverage a set of learnable category query vectors $\mathbf{q}_{emo} \in \mathbb{R}^{N_w \times D}$ to mine emotions:

$$\hat{\mathcal{E}}_{true} = \text{MHCA}(\text{Query} = \mathbf{q}_{emo}, \text{Key} = \mathcal{H}_{true}, \text{Value} = \mathcal{H}_{true}) \in \mathbb{R}^{N_w \times D}, \tag{18}$$

besides, to achieve the accurate supervision for emotion mining, we introduce a counterfactual causal alignment module. Specifically, we replace the visual causal features $\mathcal{C}$ in the aforementioned process with counterfactual visual features $\hat{\mathcal{C}} = V \odot (1 - M_c)$ to obtain counterfactual emotion features $\hat{\mathcal{E}}_{fake}$. Then, we calculate the cosine similarity between the ground-truth and the predicted emotion features to obtain the emotion logits:

$$z_{true} = \text{Cos}(\mathcal{E}_{gt}, \hat{\mathcal{E}}_{true}) \in \mathbb{R}^{N_w}, z_{fake} = \text{Cos}(\mathcal{E}_{gt}, \hat{\mathcal{E}}_{fake}) \in \mathbb{R}^{N_w}, \tag{19}$$

Table 1: Performance for emotional video captioning task on two benchmarks. The best results are highlighted in bold.

Methods	Emotion		Semantic					Hybrid	
	$\text{Acc}_{sw}\uparrow$	$\text{Acc}_{c}\uparrow$	BLEU-1↑	BLEU-4↑	METEOR↑	ROUGE↑	CIDEr↑	BFS↑	CFS↑
EVC-MSVD									
SA-LSTM (2018)	68.8	67.2	80.7	45.5	33.0	68.2	72.1	59.0	71.3
FT (2021)	69.4	67.1	77.2	36.3	29.0	63.4	62.5	52.5	63.7
SGN (2021)	73.9	73.1	77.5	41.1	30.6	63.6	71.0	56.4	71.5
CANet (2022)	78.7	76.8	78.5	41.8	30.8	65.7	74.4	57.9	75.1
VEIN (2024)	82.7	82.1	82.0	45.9	33.0	69.0	79.6	62.4	80.2
EPAN (2023)	84.1	82.8	82.5	46.2	34.4	69.8	80.6	63.1	81.1
DCGN (2024)	86.5	85.7	84.5	48.7	35.7	71.0	85.2	65.7	86.6
GLM-EER (2026)	88.2	86.6	92.2	68.3	43.8	80.1	153.9	78.8	140.6
MM-ECPE (2025b)	90.4	89.1	96.9	**71.4**	45.5	83.1	168.3	81.8	152.6
SAGML (2026)	93.5	92.3	93.6	53.8	40.8	77.2	96.1	71.5	95.5
Causal-EVC(Ours)	**94.2**	**92.9**	**97.4**	71.2	**46.2**	**84.6**	**169.8**	**82.7**	**154.6**
EVC-VE									
SA-LSTM (2018)	48.6	47.1	71.0	22.5	19.6	40.7	30.2	38.9	33.7
CANet (2022)	41.9	39.7	66.9	19.3	18.2	37.9	23.3	33.9	26.8
VEIN (2024)	57.4	56.8	71.6	26.3	20.9	41.7	33.4	43.0	39.2
EPAN (2023)	63.8	62.3	73.6	27.0	21.2	42.3	34.7	45.0	40.4
DCGN (2024)	71.0	69.4	74.5	28.1	23.4	47.7	41.5	47.3	46.9
HEART (2026)	55.8	54.2	71.5	22.7	19.7	41.2	28.8	40.3	34.0
MM-ECPE (2025b)	73.4	72.3	76.8	28.9	24.7	49.5	65.2	49.2	66.7
SAGML (2026)	80.2	79.6	82.7	31.8	26.9	52.7	71.8	53.8	73.4
Causal-EVC(Ours)	**82.8**	**81.5**	**85.1**	**32.4**	**27.7**	**54.6**	**74.0**	**55.3**	**75.6**

where $\mathcal{E}_{gt}$ denotes the prototype embedding of the ground-truth emotion labels. Then, we employ an NCE loss (Fu et al., 2024; Radford et al., 2021; Liu et al., 2026; Li et al., 2026a) on (z_{true}, z_{fake}) pair to guide the model to learn accurate emotion features from genuine causal features, rather than relying on shortcut learning from spurious visual backgrounds.

$$\mathcal{L}_{cf} = -\log \frac{\exp(z^{(y)}_{true}/\tau_{cf})}{\exp(z^{(y)}_{true}/\tau_{cf}) + \exp(z^{(y)}_{fake}/\tau_{cf})}, \tag{20}$$

where τ_{cf} is the temperature parameter. After obtaining sufficient semantic information including video, cause, and emotion features, we concatenate them to generate a unified multi-modal semantic representation $M = [\mathcal{V}; \mathcal{C}; \mathcal{E}_{true}] \in \mathbb{R}^{(T\times 2P+N_w)\times D}$. Finally, we send them into the decoder with Qwen-2.5-VL model (Bai et al., 2025) to generate the emotional captions.

4 Results and Discussion

4.1 Main Results

As shown in Table 1, Causal-EVC consistently achieves the best performance across almost all metrics, *i.e.,* improving +4.2% with Acc_{sw} than MM-ECPE and +15.7% with BFS than SAGML on EVC-MSVD. More notably, on the larger EVC-VE benchmark, Causal-EVC also outperforms SAGML, *i.e.,* +2.4%/+3.0% improvements with Acc_c/BLEU-4. These prove that Causal-EVC effectively achieves better emotional description generation. Besides, Table 2 reports the cause-grounding performance on EVC-CauseGround. Causal-EVC demonstrates substantial dominance across all metrics. Compared to specialized visual grounding models, our model leads +67.2%/13.7% than VTimeLLM (Huang et al., 2024)/VideoChat-T (Zeng et al., 2025) on $R@0.3$. While these models excel at objective localization, they are prone to hallucinations when it comes to emotional visual attribution due to their lack of specific emotion understanding components. Besides, compared to general MLLMs, our model leads +52.1%/33.0% than Gemini-2.5-Flash (Comanici et al., 2025)/GPT-4o (Hurst et al., 2024) on mIoU. Despite the powerful comprehension capabilities of these models, the dual requirements of emotional understanding and visual grounding exacerbate the frequency of hallucinations, which demonstrates that Causal-EVC improves the authentic and robust emotion-cause reasoning and alleviates the co-occurrence bias.

Table 2: Results on EVC-CauseGround.

Methods	$mIoU$	$R@0.3$	$R@0.5$	Acc_{sw}
MM-ECPE (2025b)	18.6	28.7	12.6	53.1
Momentor (2024)	21.5	35.8	20.4	58.5
VTimeLLM (2024)	23.7	42.1	22.0	61.4
VideoChat-T (2025)	37.6	61.9	40.8	78.4
Qwen2.5-VL-7B (2025)	39.8	62.5	41.3	80.8
Gemini-2.5-Flash (2025)	30.7	50.9	27.1	67.0
GPT-4o (2024)	35.1	55.5	33.2	73.9
Ours	**46.7**	**70.4**	**49.2**	**90.3**

Table 3: The ablation study for dense MCSL.

Components			Acc_{sw}	BLEU-4	CIDEr	CFS
Motion	Temporal	Spatial				
×	×	×	92.9	67.8	154.4	142.0
	✓	×	89.4	68.6	156.9	143.2
	×	✓	91.2	69.7	158.5	144.9
	✓	✓	90.8	70.3	160.2	146.2
✓	×	×	90.5	67.2	152.8	140.2
	✓	×	93.4	70.1	160.6	147.0
	×	✓	93.7	70.6	163.9	149.7
	✓	✓	**94.2**	**71.2**	**169.8**	**154.6**

Table 4: The ablation study for proposed losses.

Components			Acc_{sw}	BLEU-4	CIDEr	CFS
$\mathcal{L}_{cf}$	$\mathcal{L}_{temp}$	$\mathcal{L}_{spa}$				
×	×	×	87.3	66.7	153.2	139.9
	✓	×	87.8	68.5	157.4	143.4
	×	✓	87.4	68.9	158.3	144.0
	✓	✓	88.4	70.1	164.0	148.8
✓	×	×	90.8	67.2	155.9	142.8
	✓	×	92.1	69.0	159.3	145.8
	×	✓	91.6	69.4	160.2	146.4
	✓	✓	**94.2**	**71.2**	**169.8**	**154.6**

4.2 Ablation Studies

Table 5: The ablation study for proposed modules.

Components		Acc_{sw}	BLEU-4	CIDEr	CFS
MCSL	ISER				
×	×	82.1	65.3	150.9	137.0
✓	×	85.7	70.2	165.4	149.4
×	✓	92.9	67.8	154.4	142.0
✓	✓	**94.2**	**71.2**	**169.8**	**154.6**

Discussion on proposed modules. As shown in Table 5, we make an ablation study on proposed modules. We observe that each module could improve its performance independently. Specifically, MCSL effectively localizes the cause cues, which helps to generate more accurate captions for crucial emotion-related events, resulting in obvious improvements on semantic metrics, *i.e.,* +7.5%/+9.1% improvements on BLEU-4/CFS. Besides, ISER effectively utilizes causal masks and counterfactual interventions to achieve accurate emotion cue extraction, resulting in obvious improvements in emotion accuracy, *i.e.,* +13.2% improvements on Acc_{sw}. Finally, we combine two modules to achieve the best performance, *i.e.,* +14.7%/+12.8% improvements on Acc_{sw}/CFS.

Discussion on MCSL. We further discuss the dense components in MCSL module in Table 3. We first observe that solely relying on motion dynamics to generate the causal mask leads to a performance degradation. We analyze that pixel-wise differences lack semantic discriminability and are easily corrupted by environmental disturbances, such as swaying leaves. However, when motion information is used as a kinematic prior to guide both spatial and temporal localization, the model achieves a substantial performance surge. This result strongly validates that raw motion dynamics are inherently suited as guiding priors rather than standalone causes, while the genuine causal triggers fundamentally reside in the structural semantic entities executing these dynamic actions.

Table 6: Results of different Top-K routing.

Top-K	Acc_{sw}	BLEU-4	CIDEr	CFS
1	49.6	51.5	115.9	102.5
5	88.6	67.4	149.8	137.4
10	94.2	71.2	169.8	154.6
20	92.9	70.1	161.8	147.8
50	92.2	68.9	156.3	143.3

Discussion on Top-K Routing. As shown in Table 6, we discuss the performance of different Top-K routing. First, we observe that an excessively tight constraint (i.e., $K = 1$) leads to a performance collapse. This degradation occurs because forcing a single-prototype selection severely undermines the model's fault tolerance and fails to accommodate the emotional diversity nature. Conversely, scaling K to an overly large value (e.g., $K = 50$) causes trivial performance. An unconstrained routing introduces redundant and irrelevant emotion prototypes, which compromises the sparsity bottleneck and dilutes the authentic causal emotion features with background semantic noise. Overall, we set $K = 10$ as the default configuration, which strikes the optimal trade-off between affective expressiveness and noise suppression.

Discussion on objective functions. Besides, we conduct an ablation study to discuss the effectiveness of three specialized objective functions $\mathcal{L}_{temp}$, $\mathcal{L}_{spa}$, and $\mathcal{L}_{cf}$. As shown in Table 4, applying $\mathcal{L}_{cf}$ brings significant improvements on emotion accuracy. It isolates spurious background information by counterfactual contrastive learning, ensuring that emotion predictions stem from genuine causal triggers. Besides, applying $\mathcal{L}_{temp}$ and $\mathcal{L}_{spa}$ brings remarkable improvements on semantic metrics. They leverage ground-truth spatiotemporal tube annotations to supervise cause grounding, ensuring the accuracy of key cause-related words in emotional captions. Finally, we employ these objective functions in combination to achieve end-to-end training, ensuring the accuracy in both emotional description and causal localization.

4.3 Qualitative Results



Figure 5: Comparison of caption generation between MM-ECPE and our Causal-EVC.

Case Study of Causal Localization. We first make a comparison of caption generation between MM-ECPE and Causal-EVC shown in Fig. 5. MM-ECPE generates incorrect emotions 'annoyed' and 'expected', due to co-occurrence bias for 'throwing stones' and insufficient localization for 'sicken bread'. Besides, we display the effect of cause localization in Fig. 6. We first observe from the Temporal Energy Curve that the baseline is easily misled by the early preparation of the ice cream, peaking prematurely. Instead, our curve surges precisely when the girl experiences a sudden sensory discomfort. Consequently, our Causal-EVC predicts a tighter temporal span that aligns with GT labels than baselines. More crucially, from the spatial heatmaps, we observe that our model successfully eliminates the spurious correlation. Tricked by the prior bias "ice-cream→happy", the baseline focuses predominantly on the ice cream region, mistakenly inferring a "Happy" emotion while entirely overlooking the child's adverse reaction. In contrast, Causal-EVC penalizes the confounder and concentrates on the disgust expressions of the girl, inferring a "Disgust" emotion successfully.

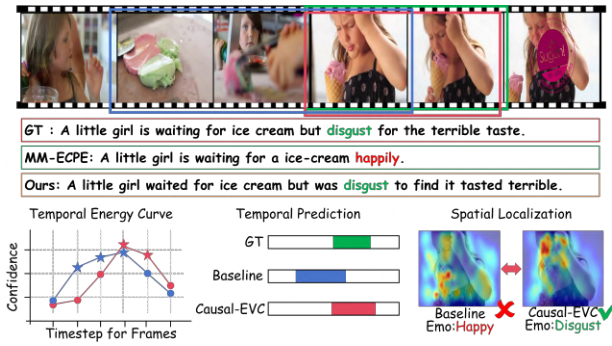

Figure 6: Case study for cause localization.

t-SNE Visualization of Cause Representation. We visualize the t-SNE distributions of $\mathcal{C}$ across emotion spaces alongside $\hat{C}$. As shown in Fig. 7, we first observe that compared MM-ECPE, our Causal-EVC achieves better semantic clustering. This stems from the MCSL module, which synergistically leverages motion priors to steer cause localization under strict coordinate supervision. More crucially, In MM-ECPE, $\hat{C}$ are systematically entangled within the respective emotion clusters. This confirms our findings that conventional in-batch contrastive learning on $(\mathcal{C}, \mathcal{E})$ suffers from shortcut learning, mistakenly encoding background contexts as emotional evidence. Instead, under our Causal-EVC, $\hat{C}$ are uniformly distributed across the entire latent space. This validates that our $\mathcal{L}_{cf}$ on $(C, \hat{C})$ pairs successfully severs the spurious learning from environmental contexts, strictly compelling to deduce affective semantics exclusively from authentic cause triggers.

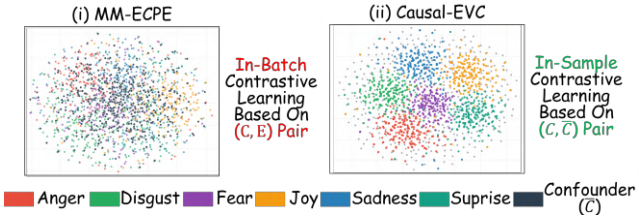

Figure 7: t-SNE visualization of cause representation.

5 Conclusion

In this paper, we propose **Causal-EVC**, a cause-grounding captioning framework. To address causal redundancy and spurious correlations, we establish **EVC-CauseGround**, a pioneering benchmark featuring dense spatio-temporal causal tubes, a curated Causal-Faithfulness Subset, and a convinced metric. Technically, Causal-EVC introduces a Motion-guided Causal Localization module with threshold-free heads to decouple foreground causal triggers into a 3D mask. Furthermore, we design an Interpretable Sparse Emotion Routing module to deduce class-specific emotion probabilities via sparse prototype routing, enforcing a Counterfactual Contrastive Objective on synthesized background representations to strictly penalize shortcut learning. Extensive experiments demonstrate that Causal-EVC achieves the best performance in both text generation and causal faithfulness.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the language of the manuscript, writing and debugging portions of the experiment code, and some steps in constructing the dataset. We have not used generative AI tools for the motivation or the production of the reported results and figures. These are carried out directly by the authors. All AI-assisted code and data was tested and verified by running the experiments reported in this paper, and the resulting claims were checked against the run outputs by the authors. We take responsibility for the final content of this work, including the data, text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Jiaming An, Zixiang Ding, Ke Li, and Rui Xia. Global-view and speaker-aware emotion cause extraction in conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3814–3823, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Satanjeev Banerjee and Alon 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*, pp. 65–72, 2005.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International Conference on Learning Representations*, volume 2026, pp. 138846–138923, 2026.

David Chen and William 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*, pp. 190–200, 2011.

Xinhao Chen, Chong Yang, Changzhi Sun, Man Lan, and Aimin Zhou. From coarse to fine: A distillation method for fine-grained emotion-causal span pair extraction in conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17790–17798, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Yicheng Deng, Hideaki Hayashi, and Hajime Nagahara. From pixels to semantics: Unified facial action representation learning for micro-expression analysis. In *International Conference on Learning Representations*, volume 2026, pp. 105640–105659, 2026.

Adnan El Assadi, Isaac Chung, Roman Solomatin, Niklas Muennighoff, and Kenneth Enevoldsen. Hume: Measuring the human-model performance gap in text embedding tasks. In *International Conference on Learning Representations*, volume 2026, pp. 71873–71904, 2026.

Zheren Fu, Lei Zhang, Hou Xia, and Zhendong Mao. Linguistic-aware patch slimming framework for fine-grained cross-modal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26307–26316, 2024.

Junyu Gao, Mengyuan Chen, and Changsheng Xu. Vectorized evidential learning for weakly-supervised temporal action localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15949 – 15963, 2023.

Rezatofighi Hamid, Tsoi Nathan, Gwak JunYoung, Sadeghian Amir, Reid Ian, and Savarese Silvio. Generalized intersection over union: a metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 658–666, 2019.

Tingting Han, Yuxuan Gong, Sicheng Zhao, Min Tan, Zhou Yu, and Hongxun Yao. Heart: Emotionally grounded video captioning via hierarchical emotion-aligned representation. *IEEE Transactions on Affective Computing*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL `https://arxiv.org/abs/2106.09685`.

Guimin Hu, Yi Zhao, and Guangming Lu. Emotion prediction oriented method with multiple supervisions for emotion-cause pair extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1141–1152, 2023.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14271–14280. IEEE, 2024.

Xiaoyu Huang, Weidong Chen, Bo Hu, and Zhendong Mao. Graph mixture of experts and memory-augmented routers for multivariate time series anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17476–17484, 2025.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue. Predicting emotions in user-generated videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28, 2014.

Yuda Jin, Weidong Chen, Yuanhe Tian, Yan Song, and Chenggang Yan. Improving radiology report generation with multi-grained abnormality prediction. *Neurocomputing*, 600:128122, 2024a.

Yuda Jin, Weidong Chen, Yuanhe Tian, Yan Song, Chenggang Yan, and Zhendong Mao. Improving radiology report generation with d 2-net: When diffusion meets discriminator. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2215–2219. IEEE, 2024b.

Xincheng Ju, Dong Zhang, Junhui Li, Shoushan Li, and Guodong Zhou. Enhanced generative framework with llms for multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Multimedia*, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Bobo Li, Hao Fei, Fei Li, Tat-seng Chua, and Donghong Ji. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

Hengyou Li, Xinyan Liu, Guorong Li, Shuhui Wang, Laiyun Qing, and Qingming Huang. Boost tracking by natural language with prompt-guided grounding. *IEEE Transactions on Intelligent Transportation Systems*, 26(1):1088–1100, 2025. doi: 10.1109/TITS.2024.3492263.

Mingyuan Li, Tong Jia, Hao Wang, Bowen Ma, Shiyi Guo, Da Cai, and Dongyue Chen. Cspcl: Category semantic prior contrastive learning for deformable detr-based prohibited item detectors. *Advances in Neural Information Processing Systems*, 38:141729–141750, 2026a.

Shihao Li, Yuanxing Zhang, Jiangtao Wu, Zhide Lei, Yiwen He, Runzhe Wen, Chenxi Liao, Chengkang Jiang, An Ping, Shuo Gao, et al. If-vidcap: Can video caption models follow instructions? In *International Conference on Learning Representations*, volume 2026, pp. 23157–23201, 2026b.

Wei Li, Yang Li, Vlad Pandelea, Mengshi Ge, Luyao Zhu, and Erik Cambria. Ecpec: Emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(3):1754–1765, 2022.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Zefeng Lin, Weidong Chen, Yan Song, and Yongdong Zhang. Prompting few-shot multi-hop question generation via comprehending type-aware semantics. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3730–3740, 2024.

Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. Bootstrapping large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18635–18643, 2024.

Xiaohao Liu, Xiaobo Xia, See-Kiong Ng, and Tat-Seng Chua. Continual multimodal contrastive learning. *Advances in Neural Information Processing Systems*, 38:48414–48455, 2026.

Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang 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*, pp. 24–32, 2022.

Chong Ma, Shengbo Chen, Pengjie Tang, Hong Rao, and Hanli Wang. Glm-eer: Global-local memory and emotion evaluation refinement for emotional video description. *Expert Systems with Applications*, pp. 132042, 2026.

Heqing Ma, Jianfei Yu, Fanfan Wang, Hanyu Cao, and Rui Xia. From extraction to generation: multimodal emotion-cause pair generation in conversations. *IEEE Transactions on Affective Computing*, 2024.

Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 174–184, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: Advancing video large language model with fine-grained temporal reasoning, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Hobin Ryu, Sunghun Kang, Haeyong Kang, and Chang D Yoo. Semantic grouping network for video captioning. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2514–2522, 2021.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Peipei Song, Dan Guo, Jun Cheng, and Meng Wang. Contextual attention network for emotional video captioning. *IEEE Transactions on Multimedia*, 2022.

Peipei Song, Dan Guo, Xun Yang, Shengeng Tang, Erkun Yang, and Meng Wang. Emotion-prior awareness network for emotional video captioning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 589–600, 2023.

Peipei Song, Dan Guo, Xun Yang, Shengeng Tang, and Meng Wang. Emotional video captioning with vision-based emotion interpretation network. *IEEE Transactions on Image Processing*, 2024.

Peipei Song, Long Zhang, Long Lan, Weidong Chen, Dan Guo, Xun Yang, and Meng Wang. Towards efficient partially relevant video retrieval with active moment discovering. *IEEE Transactions on Multimedia*, 2025.

Shengeng Tang, Dan Guo, Richang Hong, and Meng Wang. Graph-based multimodal sequential embedding for sign language translation. *IEEE Transactions on Multimedia*, 24:4433–4445, 2021.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7622–7631, 2018.

Botao Wang, Keke Tang, and Peican Zhu. Enhancing emotion-cause pair extraction in conversations via center event detection and reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10773–10783, 2024a.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. Multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(3):1832–1844, 2022.

Fanfan Wang, Heqing Ma, Xiangqing Shen, Jianfei Yu, and Rui Xia. Observe before generate: Emotion-cause aware video caption for multimodal emotion cause generation in conversations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5820–5828, 2024b.

Hanli Wang, Pengjie Tang, Qinyu Li, and Meng Cheng. Emotion expression with fact transfer for video description. *IEEE Transactions on Multimedia*, 24:715–727, 2021.

Junbo Wang, Liangyu Fu, Yuke Li, Xuecheng Wu, and Zhiyong Wang. Adaptive emotional video captioning via affective heterogeneous graph reasoning and multi-task joint learning. *arXiv preprint arXiv:2607.29045*, 2026.

Yujing Wang, Ruotong Fang, Xing Huang, Zhiyuan Han, Xiaoqing Lin, Yuhao Shan, and Tong Chen. Emotion-qwen-vl: A fully fine-tuned multimodal large language model for micro-expression visual question answering. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13972–13978, 2025.

Rui Xia and Zixiang Ding. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1003–1012, 2019.

Ning Xu, Yifei Gao, Ting-Ting Zhang, Hongshuo Tian, and An-An Liu. Cross-modal coherence-enhanced feedback prompting for news captioning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9369–9377, 2024.

Yifan Xu, Xinhao Li, Yichun Yang, Desen Meng, Rui Huang, and Limin Wang. Carebench: A fine-grained benchmark for video captioning and retrieval. In *International Conference on Learning Representations*, volume 2026, pp. 51124–51150, 2026.

Cheng Ye, Weidong Chen, Jingyu Li, Lei Zhang, and Zhendong Mao. Dual-path collaborative generation network for emotional video captioning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 496–505, 2024.

Cheng Ye, Weidong Chen, Bo Hu, Lei Zhang, Yongdong Zhang, and Zhendong Mao. Improving video summarization by exploring the coherence between corresponding captions. *IEEE Transactions on Image Processing*, 2025a.

Cheng Ye, Weidong Chen, Peipei Song, Xinyan Liu, Lei Zhang, and Zhendong Mao. Multi-round mutual emotion-cause pair extraction for emotion-attributed video captioning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025b.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

Keunwoo Yu, Zheyuan Zhang, Fengyuan Hu, Shane Storks, and Joyce Chai. Eliciting in-context learning in vision-language models for videos through curated data distributional properties. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20416–20431, 2024.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. Timesuite: Improving mllms for long video understanding via grounded tuning. In *International Conference on Learning Representations*, volume 2025, pp. 38057–38081, 2025.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, YiFan Zhang, Renrui Zhang, et al. Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 48767–48807, 2026.

Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10665–10674. IEEE, 2020.

Peican Zhu, Botao Wang, Keke Tang, Haifeng Zhang, Xiaodong Cui, and Zhen Wang. A knowledge-guided graph attention network for emotion-cause pair extraction. *Knowledge-Based Systems*, 286:111342, 2024.

A Related Work

A.1 Emotion-cause Pair Extraction

Emotion cause analysis has received increasing attention due to its potential to promote emotional exploration. Emotion-cause pair extraction(ECPE) is one of the most representative tasks (Tang et al., 2021; Li et al., 2022; Wang et al., 2024a;a; An et al., 2023; Song et al., 2025; Lin et al., 2024; Li et al., 2025). Xia and Ding (Xia & Ding, 2019) firstly propose this task to extract emotions and their causes in documents. To this day, recent research has focused more on analyzing emotion-cause pairs in diverse textual scenarios such as articles, stories, microblogs, or conversations. Zhu *et al.* (Zhu et al., 2024) propose a novel knowledge-guided graph attention network, which designs two guiding principles: inter-clause dependency graph and inter-pair possibility graph, to aggregate features between clause pairs for emotion-clause and cause-clause pair extraction. Chen *et al.* (Chen et al., 2024) propose to utilize the causal discourse knowledge in a knowledge distillation way, which designs a teacher model to learn to predict causal connective words between utterances and then guide a student model in identifying both the fine-grained emotion labels and causal spans. Hu *et al.* (Hu et al., 2023) analyze that emotion extraction is more crucial to the ECPE task than cause extraction and propose an approach oriented toward emotion prediction, which aims to fully exploit the potential of emotion prediction to enhance ECPE. Furthermore, they design a synchronization mechanism to share their optimizations in the training process.

Besides, some researchers have extended the ECPE task to the multimodal field (Li et al., 2024; Jin et al., 2024b; Lin et al., 2024; Jin et al., 2024b; Liu et al., 2024; Wang et al., 2022; Ma et al., 2024; Ju et al., 2025; Huang et al., 2025; Jin et al., 2024b;a; Gao et al., 2023). Wang *et al.* (Wang et al., 2024b) design a two-stage framework, which first generates emotion-cause aware video captions and then facilitates the generation of emotion causes. With the help of corresponding captions, the model improves the accuracy of pair extractions. However, all existing multimodal ECPE still perform on the text modality while treating image and audio modalities as auxiliary information. In contrast, our work focuses on the cross-modal scenario (visual causes + textual emotions).

A.2 Emotional Video Captioning

Emotional video captioning is proposed to solve the problem of boring and soulless captions generated by traditional video captioning model, which first mines emotional clues hidden in videos and collaborates visual contents to generate vivid captions. Wang *et al.* (Wang et al., 2021) firstly propose this task and release a new dataset. They design two independent prediction networks for factual and emotional captioning separately and finally generate emotional descriptions by weighted average scores. Song *et al.* (Song et al., 2022) propose a contextual attention network to recognize and describe the fact and emotion in the video by semantic-rich context learning. Song *et al.* (Song et al., 2023) propose a novel tree-structured emotion learning module to achieve explicit emotion perception. Song *et al.* (Song et al., 2024) incorporate visual context, textual context, and visual-textual relevance into an aggregated multimodal contextual vector to enhance video captioning. Ye *et al.* (Ye et al., 2024) propose a dual-path collaborative generation network to dynamically perceive emotions at different time steps and design an emotion-adaptive decoder to solve the problem of overemphasis on the role of emotional guidance. Wang *et al.* (Wang et al., 2025) introduce two learnable prompting strategies: visual emotion and textual emotion prompting, to learn emotional cue representations and further design two levels of objective functions: the ER-sentence level and the AU-word level alignment losses, to facilitate the interaction and alignment.

All aforementioned works directly perceive the emotional cues from visual contents and ignore that emotional cues have intrinsic motivational causes reflected in the video content. To the best of our knowledge, we are the first work to notice the visual causes of emotions, improving the accuracy and interpretability of emotion perception and effectively alleviating the mutual interference of multiple factors in emotional caption generation.

B THEORETICAL ANALYSIS: FROM THE INFORMATION THEORY PERSPECTIVE

To provide a rigorous theoretical foundation for our Counterfactual Contrastive Objective ($\mathcal{L}_{cf}$), we analyze the fundamental flaws of existing positive-alignment methods (e.g., standard InfoNCE) from an information-theoretic perspective and demonstrate why our interventional approach strictly guarantees causal faithfulness.

The Flaw of Standard InfoNCE. Let V, Y and C denote the random variables for the video, emotion, and true foreground cause, respectively. Traditional EVC models optimize the representation by applying InfoNCE loss between the video representation and the emotion embedding. Mathematically, minimizing the standard InfoNCE loss is equivalent to maximizing the Mutual Information (MI) between the video and the emotion:

$$\min \mathcal{L}_{InfoNCE} \iff \max I(V;Y) \tag{21}$$

However, according to our structural causal model, the video is composed of true causes C and the background confounders U. By the mutual information chain rule, $I(V;Y)$ can be decomposed as:

$$I(V;Y) = I(C,U;Y) = I(C;Y) + I(U;Y|C) \tag{22}$$

In the datasets, the confounder U (e.g., a "party" scene) frequently co-occurs with the emotion Y ("happy") independently of true causes C (e.g., "opening gifts"). Thus, the conditional mutual information $I(U;Y|C) > 0$. When a model blindly maximizes $I(V;Y)$, it inevitably maximizes the spurious correlation $I(U;Y|C)$. This explains why existing models suffer from Contextual Leakage. They learn to predict emotions by maximizing the information from the easily accessible background confounders U rather than highly dynamic and harder-to-capture true causes C.

The Superiority of Counterfactual Intervention. To achieve authentic causal reasoning, our goal is to exclusively maximize $I(C;Y)$ while strictly minimizing the spurious leakage $I(U;Y)$. Our framework accomplishes this through the explicitly generated causal mask $\mathcal{M}_c$. We decouple the video into the foreground cause representation $\mathcal{C} = \mathcal{V} \odot \mathcal{M}_c$ and the counterfactual background confounder $U = \mathcal{V} \odot (1 - \mathcal{M}_c)$.

Our proposed Counterfactual Contrastive Objective ($\mathcal{L}_{cf}$) is formulated as:

$$\mathcal{L}_{cf} = -\mathbb{E}\left[\log \frac{\exp(\text{sim}(\mathcal{C}, \mathcal{E})/\tau)}{\exp(\text{sim}(\mathcal{C}, \mathcal{E})/\tau) + \exp(\text{sim}(U, \mathcal{E})/\tau)}\right] \tag{23}$$

According to the variational bounds of mutual information (e.g., InfoNCE as an MI estimator), optimizing $\mathcal{L}_{cf}$ intrinsically acts on the ratio of conditional densities. By treating the true cause $\mathcal{C}$ as the positive sample and the decoupled confounder U as the strict hard negative sample within the same video context, our objective function tightly bounds the difference in mutual information:

$$\min \mathcal{L}_{cf} \implies \max \Big(I(\mathcal{C};\mathcal{E}) - I(U;\mathcal{E}) \Big) \tag{24}$$

This formulation yields a profound theoretical guarantee: our model is explicitly penalized if the background confounder U contains any predictive information about the emotion $\mathcal{E}$. To minimize the loss, the optimization process forces the causal mask $\mathcal{M}_c$ to absorb all emotion-relevant semantics into $\mathcal{C}$, driving the spurious mutual information $I(U;\mathcal{E})$ towards zero.

Therefore, compared to traditional positive alignment that inadvertently absorbs environmental biases, our counterfactual intervention rigorously severs the backdoor path $V \to U \to Y$. It ensures that the emotion prediction is authentically anchored to the interventional distribution $P(Y|do(C))$, thereby achieving superior causal faithfulness.

C MORE DETAILS OF EVC-CAUSEGROUND

C.1 COMPARISON WITH EXISTING BENCHMARK

In this paper, we construct **EVC-CauseGround**, a novel benchmark that extends conventional emotion descriptions with a comprehensive suite of fine-grained cause annotations. Table A.1 presents a

Table A.1: **Comparison of EVC-CauseGround with existing datasets across different domains.** "V", "A", and "T" stand for video, audio, and text modalities, respectively. Our dataset is the first to simultaneously provide fine-grained spatio-temporal causal grounding and a rigorously diagnosed counterfactual subset for real-world emotional videos.

Dataset	Modality	Emotion Labels	Emotion Captions	Temporal Grounding	Spatial (Dense Tubes)	Counterfactual Subset
Emotion Classification & Captioning						
V-Emo (2014)	V, A	✓	×	×	×	×
EmVidCap (2021)	V, T	✓	✓	×	×	×
Emotion-Cause Pair Extraction (ECPE)						
ECF (2022)	V, A, T	✓	×	✓ (Utterance-level)	×	×
MECESD (2024)	V, A, T	✓	✓	✓ (Utterance-level)	×	×
Spatio-Temporal Grounding & Causal Reasoning						
VidSTG (2020)	V, T	×	✓	✓	✓	×
CLEVRER (2019)	V, T	×	✓	✓	✓	✓(Synthetic)
EVC-CauseGround (Ours)	V, T	✓	✓	**✓(Frame-level)**	✓	**✓(Real-world)**

systematic comparison between existing emotion/causal reasoning datasets and our proposed EVC-CauseGround. Traditional emotion classification and captioning datasets exclusively provide holistic emotion labels or descriptions, fundamentally lacking the annotation of causal triggers, which makes it impossible to assess the true causal reasoning ability. Conversely, while some Emotion-Cause Pair Extraction datasets attempt to address causality, they not only omit comprehensive emotion captions required for generation evaluation, but their causal annotations are also restricted to coarse-grained utterance levels, entirely neglecting fine-grained visual cause localization. Therefore, **EVC-CauseGround stands as the first benchmark to unify emotion labels, rich emotional captions, dense spatio-temporal visual cause annotations, and a meticulously curated counterfactual subset**. It establishes a solid foundation for comprehensively evaluating the dual capabilities of EVC models: high-quality caption generation and authentic emotional cause reasoning.

C.2 Details of Data Sources

The foundation of a robust causal reasoning benchmark relies heavily on the quality and diversity of its video sources. To effectively evaluate the ability to decouple genuine visual emotion triggers from spurious correlations, the selected videos must exhibit highly diverse, open-world scenes (in-the-wild) and sufficient temporal durations. This structural complexity is crucial for providing abundant environmental contexts, acting as potential environmental confounders, against which the model must learn to pinpoint the true causal regions. Driven by these principles, we meticulously curate raw videos from two original datasets.

First, we source a substantial portion of our data from the widely-used EmVidCap (Wang et al., 2021) benchmark. The original **EmVidCap** consists of two subsets: EmVidCap-S and EmVidCap-L. However, we explicitly discard the videos from EmVidCap-S. The videos in EmVidCap-S are exceedingly short (typically under 5 seconds) and feature monotonic scenes. In such cases, the causal action spans nearly the entire video duration, rendering fine-grained spatiotemporal cause annotation virtually meaningless. Therefore, we exclusively select videos from **EmVidCap-L**. EmVidCap-L comprises videos primarily collected from user-generated content (UGC) platforms, offering a much broader spectrum of daily events and dynamic human interactions. It contains 1523 videos with an average duration of approximately 100 seconds, providing an ideal temporal canvas for fine-grained causal grounding. Second, to further scale up the benchmark and amplify the diversity, we incorporate a carefully selected subset of videos from **MAFW** (Liu et al., 2022). MAFW is a large-scale, in-the-wild emotion dataset sourced from various internet platforms (e.g., YouTube and movies) and features an extensive array of unconstrained environments, ranging from extreme sports and natural disasters to bustling streets and crowded parties. The original MAFW dataset contains 10,045 video clips. We filter and select 477 videos from MAFW that exhibit prominent foreground dynamic actions accompanied by complex background scenes. By amalgamating these two sources, we have successfully collected a total of 2k high-quality emotional videos, which are split into 1500 videos for training and 500 videos for evaluation. Crucially, all selected videos are natively equipped with emotion labels and descriptive emotional captions provided by their original datasets. This inherent semantic richness serves as a reliable ground-truth anchor for our subsequent pipeline to deduce and annotate the precise spatiotemporal causes without suffering from emotion hallucination.

C.3 SPATIOTEMPORAL VISUAL CAUSE ANNOTATION

Unlike previous datasets that rely on exhaustive, cost-prohibitive manual drawing, we propose a novel MLLM-led, Tracker-assisted Coarse-to-Fine Pipeline to efficiently generate high-fidelity spatio-temporal causal tubes. This pipeline explicitly circumvents the inherent temporal hallucinations and spatial coordinate drifting typical of current MLLMs when processing long-form videos. The annotation process involves three meticulously designed stages:

Temporal Span Grounding via Sequence Grid Prompting. Current Video-LLMs struggle with precise temporal localization due to context-length constraints and memory bottlenecks when processing native video streams. To bypass this, we map the temporal dimension into a unified spatial layout. Specifically, we uniformly sample $T = 30$ frames from each video, implicitly aligning with the input space of most mainstream EVC models. We then concatenate these frames into a highly structured 6×5 grid image, with each frame explicitly watermarked with a sequence identifier (e.g., $F00$ to $F29$). We prompt an advanced MLLM (e.g., GPT-5.5 (Singh et al., 2025)) with this grid, alongside the ground-truth emotion label and caption, to reversely deduce the precise temporal window $[t_{start}, t_{end}]$ that directly triggers the emotion. Crucially, to prevent the MLLM from exhibiting "lazy behaviors" (i.e., trivially selecting the entire video sequence as the cause), we enforce a strict length-constraining prompt. We mandate that the predicted causal span must not exceed 30% of the total frames (e.g., max 9 frames). This forces the model to pinpoint the absolute climax or peak moment of the causal action rather than the redundant pre- or post-action context.

Spatial Grounding via Peak Motion Anchoring. Directly prompting an MLLM to generate bounding boxes across multiple consecutive frames frequently leads to severe spatial hallucinations and bounding box jittering. To ensure pixel-perfect spatial accuracy, we decouple temporal and spatial grounding. Within the localized temporal span $[t_{start}, t_{end}]$, we employ the optical flow mechanism to calculate inter-frame pixel displacement. This allows us to automatically identify the peak motion frame, the exact moment where the dynamic causal action is most pronounced. We then feed only this single, high-resolution peak frame into the MLLM, prompting it to generate a highly confident, static bounding box surrounding the causal agent or action. By reducing the cognitive load to a single image, the MLLM achieves remarkable spatial precision devoid of multi-frame interference.

Dense Causal Tube Generation via Zero-shot Tracking. A single bounding box is insufficient for dynamic causal reasoning. To extrapolate the spatial annotation across the entire temporal event, we leverage the state-of-the-art zero-shot visual tracker, SAM-3 (Carion et al., 2026). Utilizing the MLLM-generated bounding box on the peak frame as an initial prompt, SAM-3 bidirectionally tracks the causal object forward to $t_{end}$and backward to t_{start}. This seamless integration of MLLM reasoning and traditional visual tracking yields continuous, smooth, and dense spatio-temporal causal tubes, achieving human-level annotation quality at a fraction of the cost.

D EXPERIMENTAL SETTINGS

Datasets. We experiment on two public EVC benchmarks, *e.g.*, EVC-MSVD (Chen & Dolan, 2011) and EVC-VE (Jiang et al., 2014). **EVC-MSVD** contains 240/134 videos and 8,169/4,611 sentences for training/testing by additionally embedding emotion words into the caption sentences of the traditional video captioning dataset MSVD (Chen & Dolan, 2011). **EVC-VE** is built based on a traditional emotional video prediction dataset VideoEmotion-8 (Jiang et al., 2014) and is divided into 1,141/382 videos and 19,398/6,527 sentences for training/testing, respectively.

Evaluation Metrics. To evaluate the accuracy of emotion in generated sentences effectively, following previous work (Wang et al., 2021), we consider the emotion word accuracy $\mathrm{Acc_{sw}}$ (Wang et al., 2021) and emotion sentence accuracy $\mathrm{Acc_c}$ (Wang et al., 2021). Additionally, to measure the semantic matching degree between the generated sentence and the ground-truth label, we use the standard metrics, *e.g.*, **BLEU** (Papineni et al., 2002), **METEOR** (Banerjee & Lavie, 2005), **ROUGE** (Lin, 2004), and **CIDEr** (Vedantam et al., 2015). Moreover, there are two hybrid metrics **BFS** (Wang et al., 2021) and **CFS** (Wang et al., 2021) that combine the emotion evaluation with BLEU and CIDEr metrics, respectively.

Implementation Details. For each video, we sample $T = 30$ frames and resize them to 224×224 with central cropping for pixel values. Following (Wang et al., 2021) and (Song et al., 2023), we

build an overall vocabulary of size 32,128 that contains all words in the corpus and set the number of emotion words to $N_w = 179$. The embedding dimensions are constructed to $d_E = 300$ and $D = 1408$. In training, we adopt Qwen-2.5-VL-7B (Raffel et al., 2020) as the language decoder and integrate a LoRA adapter (Hu et al., 2021) with $r = 16$ and $\alpha = 32$ to align the visual and emotional latent space. We adopt the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 7e-4, and the batch size is set to 4. We set the global motion parameter λ_m, the penalty parameter θ, the objective function weights λ_e, λ_{cls}, λ_{temp}, λ_{spa}, and λ_{cf} to 0.2, 0.1, 0.8, 0.3, 0.2, 0.2, and 0.1. The maximum length of captions is set to 15 and the size of beam search is set to 4. All experiments are implemented on 4 NVIDIA A800 GPUs.

E BASELINES

We consider several state-of-the-art methods to make comparisons and divide them into two categories: (i) Traditional Captioning Methods and (ii) Emotional Captioning Methods.

(i) Traditional Captioning Methods

1) **SA-LSTM** (CVPR18') (Wang et al., 2018) proposes a reconstruction network to leverage both the forward (video to sentence) and backward (sentence to video) flows. The forward flow produces the sentence description based on the encoded video semantic features and the backward flow reproduces the video features based on the hidden state sequence generated by the decoder.

2) **CANet** (TMM22') (Song et al., 2022) proposes a contextual attention network to recognize and describe the fact and emotion by semantic-rich context learning, which first extracts visual and textual features from both video and previously generated words and then applies the attention mechanism to capture informative contexts for captioning.

3) **VideoBLIP** (EMNLP24') (Yu et al., 2024) proposes a training paradigm that induces in-context learning over video and text by capturing key properties of pre-training data found by prior work to be crucial for improving the ability of LLMs to generate video descriptions.

(ii) Emotional Captioning Methods

1) **FT** (TMM21') (Wang et al., 2021) introduces a visual emotion analyzer to detect implicit emotional cues and a dual-stream network that integrates emotional and factual semantics for caption generation by a weighted sum operation.

2) **EPAN** (MM23') (Song et al., 2023) introduces a tree-structured emotion repository to enable hierarchical emotion perception and an emotion-prior awareness network to achieve the explicit and fine-grained emotion perception by an emotion masking mechanism and then decode the emotional caption by exploiting the multimodal semantic cues.

3) **VEIN** (TIP24') (Song et al., 2024) designs a vision-based emotion interpretation network, which first models the emotion distribution over an open psychological vocabulary and then incorporates visual context, textual context, and visual-textual relevance into an aggregated multimodal contextual vector to enhance video captioning.

4) **DCGN** (MM24') (Ye et al., 2024) propose a framework to leverage fact-emotion collaborative learning to address the challenge of dynamic emotion changes during caption generation.

5) **MM-ECPE** (MM25') (Ye et al., 2025b) focus on the importance of emotional causes in emotional exploration and propose a multi-round mutual learning network to jointly extract emotion-cause pairs for LLM-based caption generation.

6) **HEART** (TAFFC26') (Han et al., 2026) introduces a hierarchical semantic extraction module to decompose visual content into entity-, action-, and event-level representations, and a temporal pyramid module to capture temporal dependencies for temporally coherent captioning.

7) **GLM-EER** (ESWA26') (Ma et al., 2026) designs a global-local memory module to learn abstract emotional representation across time dimensions and a Emotion Evaluation Refinement module to enhance emotion learning by dynamically adjusting cross-entropy loss weights.

8) **SAGML** (Arxiv26') (Wang et al., 2026) proposes an adaptive EVC framework via affective heterogeneous graph and multi-task language modeling, which constructs a soft affective heterogeneous graph to allow visually supported lexical emotions to remain recoverable.

F HUMAN EVALUATION

Due to the highly subjective nature of the EVC task, it is a crucial criterion for evaluation whether the generated descriptions conform to human preferences. Therefore, we add human evaluation to fully measure the quality of emotional descriptions. Inspired by the work (Wang et al., 2021), we designs four metrics, including (1) emotion accuracy: assess the accuracy of emotional expression, (2) relevance: assess whether the generated description is relevant to the video content, (3) coherence: assess the coherence and readability of the description, and (4) usability: how useful would the description be for a person (especially a blind person) to understand what is happening in the video. The score of each metric ranges from 1 to 10. Subsequently, we invite 20 participants with excellent English skills to score a subset of 50 video-caption pairs on these four metrics and calculate the average score on each metric. We make a comparison with two types of models on human evaluation: the state-of-the-art models on EVC and several large vision-language models (LVLMs).

Comparison with SOTA methods. As shown in Table A.2, compared with the SOTA methods for EVC, our model is significantly ahead in all four metrics. Our proposed MCSL captures fine-grained cause localization, thereby generating more accurate and diverse vocabulary. Furthermore, our proposed ISER utilizes VAD constraints to enhance the interpretability of emotions, making the generated descriptions more easily empathized with and understood by humans. These show that our model does not simply imitate the reference sentences, but learns the representation of emotional and factual semantics.

Comparison with LVLMs. As shown in Table A.2, despite the outstanding performance of LVLMs, our model still outperforms them in human evaluations. Due to their powerful ability in objective understanding and generation, LVLMs could produce a wide variety of descriptions. However, limited by insufficient emotion understanding and the influence of visual hallucinations, they sometimes generate factual and emotional contents irrelevant to the video, leading to lower human ratings. Conversely, our model filters out emotion-irrelevant visual information and visual-irrelevant emotional information through fine-grained visual and emotional refinement, thereby generating emotional descriptions that are more closely aligned with the video content.

Table A.2: The comparison with SOTA methods on human evaluations.

Metric	Accuracy	Relevance	Coherence	Usability
VEIN (2024)	5.47	5.18	6.27	5.04
EPAN (2023)	6.24	5.97	7.13	5.76
DCGN (2024)	6.88	6.43	7.68	6.32
VideoBLIP (2024)	5.91	5.54	7.20	4.56
MM-ECPE (2025b)	6.40	6.75	8.04	7.36
GPT-4o (2024)	6.98	6.04	7.24	6.62
VideoLLaMA3 (2025)	7.22	6.20	7.72	6.84
CausalEVC(Ours)	**7.52**	**6.85**	**8.16**	**7.58**