Omni-SILA: Towards Omni-scene Driven Visual Sentiment Identifying, Locating and Attributing in Videos

Jiamin Luo
20204027003@stu.suda.edu.cn
School of Computer Science and Technology, Soochow University
Suzhou, China

Jingjing Wang*
djingwang@suda.edu.cn
School of Computer Science and Technology, Soochow University
Suzhou, China

Junxiao Ma
20235227109@stu.suda.edu.cn
School of Computer Science and Technology, Soochow University
Suzhou, China

Yujie Jin
yjjin0727@stu.suda.edu.cn
School of Computer Science and Technology, Soochow University
Suzhou, China

Shoushan Li
lishoushan@suda.edu.cn
School of Computer Science and Technology, Soochow University
Suzhou, China

Guodong Zhou
gdzhou@suda.edu.cn
School of Computer Science and Technology, Soochow University
Suzhou, China

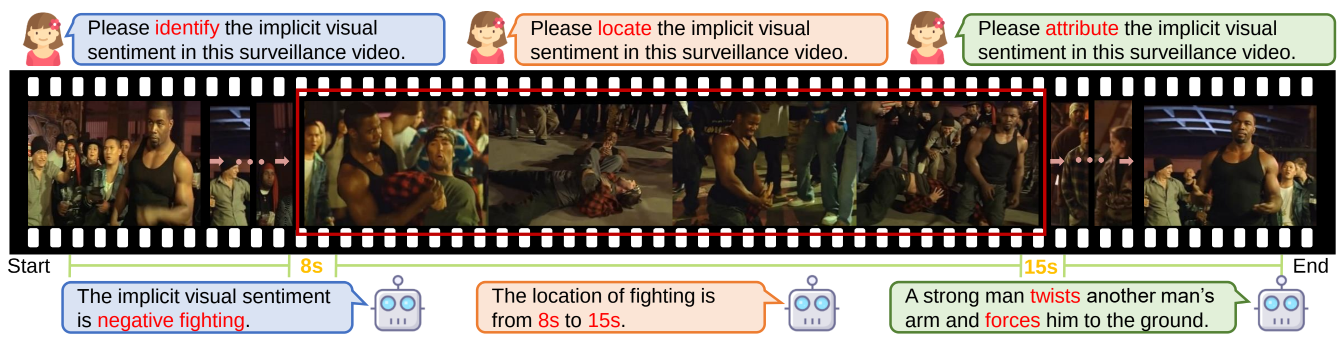

Figure 1: A sample from our constructed implicit Omni-SILA dataset to illustrate the Omni-SILA task, where the proposed ICM approach is required to identify, locate and attribute the negative implicit visual sentiment *fighting* in this surveillance video.

Abstract

Prior studies on Visual Sentiment Understanding (VSU) primarily rely on the explicit scene information (e.g., facial expression) to judge visual sentiments, which largely ignore implicit scene information (e.g., human action, objection relation and visual background), while such information is critical for precisely discovering visual sentiments. Motivated by this, this paper proposes a new **Omni**-scene driven visual **S**entiment **I**dentifying, **L**ocating and **A**ttributing in videos (Omni-SILA) task, aiming to interactively and precisely identify, locate and attribute visual sentiments through both explicit and implicit scene information. Furthermore, this paper believes that this Omni-SILA task faces two key challenges: modeling scene and highlighting implicit scene beyond explicit. To this end, this paper proposes an **I**mplicit-enhanced **C**ausal **M**oE (ICM) approach for addressing the Omni-SILA task. Specifically, a **S**cene-**B**alanced **M**oE (SBM) and an **I**mplicit-**E**nhanced **C**ausal (IEC) blocks are tailored to model scene information and highlight the implicit scene information beyond explicit, respectively. Extensive experimental results on our constructed explicit and implicit Omni-SILA datasets demonstrate the great advantage of the proposed ICM approach over advanced Video-LLMs.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**.

Keywords

Omni-Scene Information, Implicit-enhanced Causal MoE Framework, Visual Sentiment Identifying, Locating and Attributing

ACM Reference Format:
Jiamin Luo, Jingjing Wang, Junxiao Ma, Yujie Jin, Shoushan Li, and Guodong Zhou. 2025. Omni-SILA: Towards Omni-scene Driven Visual Sentiment Identifying, Locating and Attributing in Videos. In *Proceedings of the ACM Web Conference 2025 (WWW '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3696410.3714642

1 Introduction

Visual Sentiment Understanding (VSU) [55, 69] focuses on leveraging explicit scene information (e.g., facial expression) to understand the sentiments of images or videos. However, there exist many

*Corresponding author: Jingjing Wang

WWW '25, April 28-May 2, 2025, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1274-6/25/04
https://doi.org/10.1145/3696410.3714642

surveillance videos in the real world, where implicit scene information (e.g., human action, object relation and visual background) can more truly reflect visual sentiments compared with explicit scene information. In light of this, this paper defines the need to rely on implicit scene information to precisely identify visual sentiments as implicit visual sentiments, such as robbery, shooting and other negative implicit visual sentiments under surveillance videos. More importantly, current VSU studies mainly focus on identifying the visual sentiments, yet they ignore exploring when and why these sentiments occur. Nevertheless, this information is critical for sentiment applications, such as effectively filtering negative or abnormal contents in the video to safeguard the mental health of children and adolescents [11, 40, 47].

Building on these considerations, this paper proposes a new **Omni**-scene driven visual **S**entiment **I**dentifying, **L**ocating and **A**ttributing in videos (Omni-SILA) task, which leverages Video-centred Large Language Models (Video-LLMs) for interactive visual sentiment identification, location and attribution. This task aims to identify what is the visual sentiment, locate when it occurs and attribute why this sentiment through both explicit and implicit scene information. Specifically, the Omni-SILA task identifies, locates and attributes the visual sentiment segments through interactions with LLM. As shown in Figure 1, a strong man is fighting another man during the timestamps from 8s to 15s, where the LLM is asked to identify, locate and attribute this *fighting* implicit visual sentiment. In this paper, we explore two major challenges when leveraging Video-LLMs to comprehend omni-scene (i.e., both explicit and implicit scene) information for addressing the Omni-SILA task.

On one hand, how to model explicit and implicit scene information is challenging. Existing Video-LLMs primarily devote to modeling general visual information for various video understanding tasks. Factually, while LLMs encode vast amounts of world knowledge, they lack the capacity to perceive scenes [29, 35]. Compared to general visual information, explicit and implicit scene information is crucial in the Omni-SILA task. Taking Figure 1 as an example, the negative implicit visual sentiment *fighting* in the video is clearly conveyed through the action *twists an arm* and *forces to ground*. However, due to the heterogeneity of these omni-scene information (i.e., various model structures and encoders), a single, fixed-capacity transformer-based model fails to capitalize on this inherent redundancy, making Video-LLMs difficult to capture important scene information. Recently, MoE has shown scalability in multi-modal heterogeneous representation fusion tasks [43]. Inspired by this, we take advantage of the MoE architecture to model explicit and implicit scene information in videos, thereby evoking the omni-scene perceptive ability of Video-LLMs.

On the other hand, how to highlight the implicit scene information beyond the explicit is challenging. Since explicit scene information (e.g., facial expression) has richer sentiment semantics than implicit scene information (e.g., subtle actions), it is easier for models to model explicit scene information, resulting in modeling bias for explicit and implicit scene information. However, compared with explicit scene information, implicit scene information has more reliable sentiment discriminability and often reflects real visual sentiments as reported by Lian et al. [30]. For the example in Figure 1, the strong man is *laughing* while *twists another man's arm* and *forces him to the ground*, where the facial expression *laughing* contradicts the negative *fighting* visual sentiment conveyed by the actions of *twists the arm* and *forces to ground*. Recently, causal intervention [37] has shown capability in mitigating biases among different information [58]. Inspired by this, we take advantage of causal intervention to highlight the implicit scene information beyond the explicit, thereby mitigating the modeling bias to improve a comprehensive understanding of visual sentiments.

To tackle the above challenges, this paper proposes an **I**mplicit-enhanced **C**ausal **M**oE (ICM) approach, aiming to identify, locate and attribute visual sentiments in videos. Specifically, a **S**cene-**B**alanced **M**oE (SBM) module is designed to model both explicit and implicit scene information. Furthermore, an **I**mplicit-**E**nhanced **C**ausal (IEC) module is tailored to highlight the implicit scene information beyond the explicit. Moreover, this paper constructs two explicit and implicit Omni-SILA datasets to evaluate the effectiveness of our ICM approach. Comprehensive experiments demonstrate that ICM outperforms several advanced Video-LLMs across multiple evaluation metrics. This justifies the importance of omni-scene information for identifying, locating and attributing visual sentiment, and the effectiveness of ICM for capturing such information.

2 Related Work

• **Visual Sentiment Understanding.** Previous studies on Visual Sentiment Understanding (VSU) utilize multiple affective information to predict the overall sentiment of images [55, 72] or videos [61, 64]. For image, traditional studies extract sentiment features to analyze sentiments [54, 56, 71], while recent studies fine-tune LLMs via instructions to predict sentiments [52]. For videos, traditional studies require pre-processing video features and predict sentiments by elaborating fusion strategies [16, 45, 46, 63] or learning representations [18, 19, 57, 62]. To achieve end-to-end goal, some studies [2, 50, 69] input the entire videos, and explore the location of segments that convey different sentiments or anomalies. Recently, a few studies gradually explore the causes of anomalies [10] and sentiments [30] via Video-LLMs. However, these efforts have not addressed visual sentiment identification, location and attribution of videos at the same time. Different from all the above studies, this paper proposes a new Omni-SILA task to interactively answer what, when and why are the visual sentiment through omni-scene information, aiming to precisely identify and locate, as well as reasonably attribute visual sentiments in videos.

• **Video-centred Large Language Models.** Recently, large language models (LLMs) [36], such as LLaMA [44] and Vicuna [6], have shown remarkable abilities. Given the multimodal nature of the world, some studies [3, 25, 32, 75] have explored using LLMs to enhance visual understanding. Building on these, Video-LLMs have extended into the more sophisticated video area. According to the role of LLMs, Video-LLMs can be broadly categorized into three types. **(1)** LLMs as text decoders means LLMs decode embeddings from the video encoder into text outputs, including Video-ChatGPT [34], Video-LLaMA [66], Valley [33], Otter [24], mPLUG-Owl [60], Video-LLaVA [31], Chat-UniVi [20], VideoChat [26] and MiniGPT4-Video [1]. **(2)** LLMs as regressors means LLMs can predict continuous values, including TimeChat [39], GroundingGPT [28], HawkEye [48] and Holmes-VAD [67]. **(3)** LLMs as hidden layers

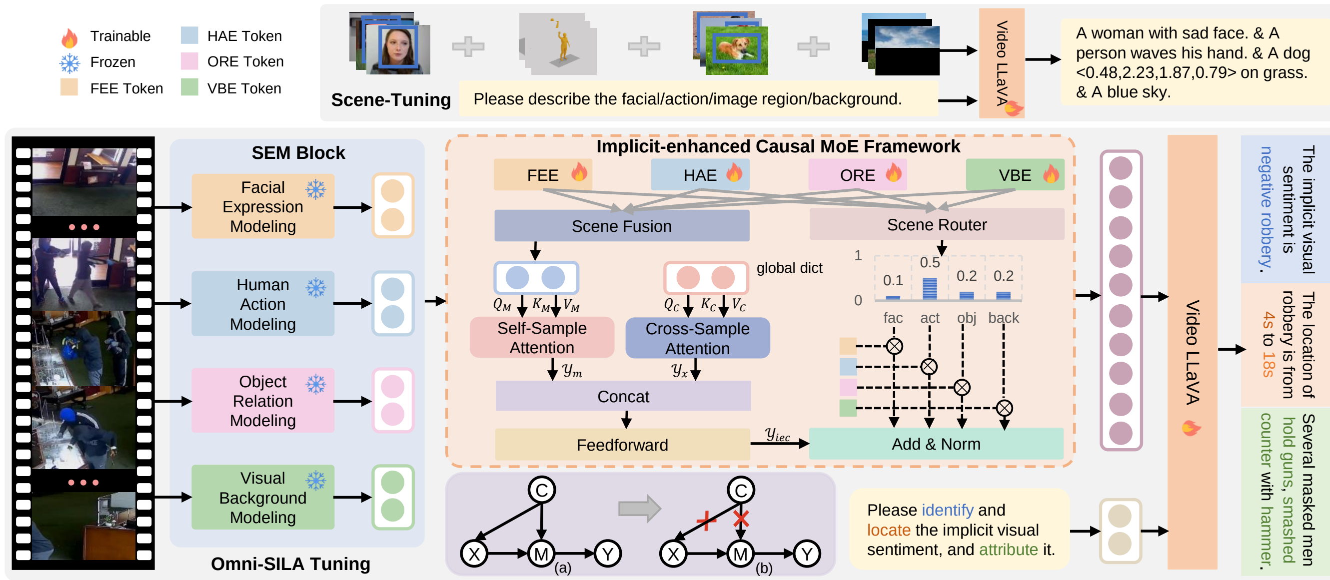

Figure 2: The overall architecture of our ICM approach, consisting of a Scene-Enriched Modeling (SEM) block and an Implicit-enhanced Causal MoE framework, which comprises a Scene-Balanced MoE (SBM) block (right, see Section 3.2) and an Implicit-Enhanced Causal (IEC) block (left, see Section 3.3), where (a) and (b) are causal graphs for IEC block. FEE, HAE, ORE and VBE represent Facial Expression Expert, Human Action Expert, Object Relation Expert and Visual Background Expert.

means LLMs connect to a designed task-specific head to perform tasks, including OneLLM [15], VITRON [12] and GPT4Video [49]. Although the aforementioned Video-LLMs studies make significant progress in video understanding, they remain limitations in their ability to perceive omni-scene information and are unable to analyze harmful video content. Therefore, this paper proposes the ICM approach, aiming to evoke the omni-scene perception capabilities of Video-LLMs and highlight the implicit scenes beyond the explicit.

3 Approach

In this section, we formulate our Omni-SILA task as follows. Given a video v consisting of N segments, each segment n is labeled with a time t, visual sentiment s and cause c. The goal of Omni-SILA is to interactively identify, locate and attribute the visual sentiment within v. Thus, the model generates a set of segments $\{(t_1, s_1, c_1), ..., (t_i, s_i, c_i), ..., (t_n, s_n, c_n)\}$, where t_i, s_i and c_i denote the time, visual sentiment and cause for each video segment v_i. In this paper, we propose an **I**mplicit-enhanced **C**ausal **M**oE (ICM) approach to address the Omni-SILA task, which involves two challenges: modeling scene and highlighting implicit scene beyond explicit. To address these challenges, we design **S**cene-**B**alanced **M**oE (SBM) and **I**mplicit-**E**nhanced **C**ausal (IEC) blocks. Particularly, we choose the open-sourced Video-LLaVA [31] as the backbone, which achieves state-of-the-art performance on most video understanding benchmarks. The overall framework is shown in Figure 2.

3.1 Scene-Enriched Modeling Block

Given a set of video segments $v = [v_1, ..., v_i, ..., v_n]$, we leverage four blocks to model omni-scene information as shown in Figure 2. **Facial Expression Modeling** is used to model explicit facial expression by MTCNN [68], which is a widely-used network to detect faces and key point locations in each video segments v_i, and them employ CNN to obtain the facial expression representation x_{f}. **Human Action Modeling** is used to model implicit human action by HigherHRNet [5], which is a well-studied network to detect the location of action key points or parts (e.g., elbow, wrist, etc) in each video segment v_i, and encode action heatmaps to obtain the human action representation x_{a}. **Object Relation Modeling** is used to model implicit object relations from each video segment v_i by RelTR [7], which is a well-studied model to generate the relations between subjects and objects, and extract the visual feature context and entity representations to obtain object relations representation x_{o}. **Visual Background Modeling** is used to model implicit visual backgrounds from each video segment v_i by ViT [8] and SAM-V1 [22], which are two advanced visual encoding and segmenting tools to segment the visual backgrounds with pure black to fill out the masked parts, and transform them into ViT to obtain the final visual background representation x_{b}.

3.2 Scene-Balanced MoE Block

In this study, we design a **S**cene-**B**alanced **M**oE (SBM) block to model scene information. Specifically, we address two crucial questions: (1) how to model different types of scene information; (2) how to balance the contributions of different scene information for the Omni-SILA task. We will provide comprehensive answers to these two questions in the subsequent section.

Scene Experts are introduced to answer question (1), which model both explicit and implicit scene information inspired by Han et al. [15], consisting of Facial Expression Expert (FEE), Human Action Expert (HAE), Object Relation Expert (ORE) and Visual Background Expert (VBE) four scene experts. Each scene expert is a stack of transformer layers, aiming to dynamically learn different scene information. As shown in Figure 2, our four scene experts operate externally to the LLM, enabling effective alignment of various scene information. Formally, for the representations $x_i, i \in \{\mathrm{f}, \mathrm{a}, \mathrm{o}, \mathrm{b}\}$ of the four scene modeling blocks, the output representation h_i of each expert Expert_i can be denotes as: $h_i = \mathrm{Expert}_i(x_i)$, where Expert_i represents the general term of FEE (f), HAE (a), ORE (o) and VBE (b) four scene experts.

Balanced MoE is leveraged to answer question (2), which balances different scene information contributions, managed by a dynamic scene router R as shown in Figure 2. Balanced MoE is structured as a straightforward MLP that processes input features h of four scene experts and computes routing weights for each expert, effectively functioning as a soft router [38]. Formally, the output y_{moe} of the balanced MoE can be denoted as follows:

$$y_{\mathrm{moe}} = \mathrm{LayerNorm}(\sum\nolimits_{j=1}^{L} g_j(h) \times E_j(h)) \tag{1}$$

where $g_j(h)$ and $E_j(h)$ denote the corresponding weight and the output of the j-th scene expert, and L is the number of scene experts.

To obtain $g(h)$, we computer the gating probability P of each scene expert for input h, formulated as: $g(h) = \mathbf{P} = \mathrm{softmax}(\mathbf{W} \cdot \mathbf{h})$, where $\mathbf{W} \in \mathbb{R}^{L \times d}$ is a learnable parameter for scene router R, and d is the hidden dimension of each expert. $\mathbf{P}$ is a vector size L and $\mathbf{P_j}$ denotes the probability of the j-th scene expert E_j to process h.

Furthermore, to optimize the scene router R, we design a router loss with balancing constraints $\mathcal{L}_{\mathrm{rb}}$, encouraging R to dynamically adjust the contributions of all scene experts, formulated as:

$$\mathcal{L}_{\mathrm{rb}} = -\alpha \cdot \sum\nolimits_{j=1}^{L} \mathbf{P}_j * \log(\mathbf{P}_j) + \beta \cdot L \cdot \sum\nolimits_{j=1}^{L} \mathbf{G}_j * \mathbf{H}_j \tag{2}$$

The first term with the hyper-parameter α measures the contribution of various scene information, encouraging the scene router R to assign a different weight to each scene expert within the constraints of $\mathbf{P}$, thereby preventing R from uniformly assigning the same weight and leading to wrong visual sentiment judgment. We expect the routing mechanism to select the scene experts that are more important for the Omni-SILA task, thus we minimize the entropy of the gating probability distribution $\mathbf{P}$ to ensure that each input feature h_i could be assigned the appropriate weight coefficient. The second term with the hyper-parameter β balances scene experts of different sizes (since the output dimension d of four scene modeling blocks are different), forcing the model not to pay too much attention to scene experts with high dimensions, while ignoring scene experts with low dimensions during the learning process. $\mathbf{G}_j = \frac{1}{L}\sum_{j=1}^{L} 1\{e_j \in E_j\} \times d$ represents the average dimension of the hidden state of the scene expert e_j on the entire input h, which imports the influence of scene expert sizes d when the model focuses more on large scene experts, the loss rises, which direct the model to more economically utilize smaller scene experts. $\mathbf{H}_j = \frac{1}{L}\sum_{j=1}^{L} \mathbf{P_j}$ represents the gating probability assigned to e_j.

3.3 Implicit-Enhanced Causal Block

In this study, we take advantage of the causal intervention technique [37] and design an **I**mplicit-**E**nhanced **C**ausal (IEC) block to highlight implicit scene beyond explicit. Specifically, there are also two crucial questions to be answered: (1) how to highlight implicit scene information through the front-door adjustment strategy [37]; (2) how to implement this front-door adjustment strategy in the Omni-SILA task. Next, we will answer the two questions.

Causal Intervention Graph is introduced to answer question (1), which formulates the causal relations among the scene information X, the fusion scene features M, visual sentiment outputs Y, and confounding factors C as shown in Figure 2 (a). In this graph, $\mathrm{X} \rightarrow \mathrm{M} \rightarrow \mathrm{Y}$ represents the desired causal effect from the scene information X to visual sentiment outputs Y, with the fusion scene features M serving as a mediator. $\mathrm{X} \leftarrow \mathrm{C} \rightarrow \mathrm{Y}$ represents the causal effect of the invisible confounding factors C on both scene information X and visual sentiment outputs Y.

To highlight implicit scene information, we consider mitigating the modeling bias between X and C present in the path $\mathrm{M} \rightarrow \mathrm{Y}$, thus we leverage *do*-operator [37] to block the back-door path $\mathrm{M} \leftarrow \mathrm{X} \leftarrow \mathrm{C} \rightarrow \mathrm{Y}$ through conditioning on X as shown in Figure 2 (b). Then, we utilize the front-door adjustment strategy to analyze the causal effect of $\mathrm{X} \rightarrow \mathrm{Y}$, denoted as: $P(\mathrm{Y} = y|do(\mathrm{X} = x)) = \sum_m P(m|x) \sum_x P(x)[P(y|x, m)]$.

Deconfounded Causal Attention is leveraged to answer question (2), which implements the front-door adjustment strategy through the utilization of attention mechanisms. Given the expensive computational cost of network forward propagation, we use the Normalized Weighted Geometric Mean (NWGM) [41, 53] approximation. Therefore, we sample X, M and compute $P(\mathrm{Y}|do(\mathrm{X}))$ through feeding them into the network, and then leverage NWGM approximation to achieve the goal of deconfounding scene biases, represented as: $P(\mathrm{Y}|do(\mathrm{X})) \approx \mathrm{softmax}[f(y_x, y_m)]$, where $f(.)$ followed by a softmax layer is a network, which is used to parameterize the predictive distribution $P(y|x, m)$. In addition, $y_m = \sum_m P(\mathrm{M} = m|p(\mathrm{X}))\boldsymbol{m}$ and $y_x = \sum_x P(\mathrm{X} = x|q(\mathrm{X}))\boldsymbol{x}$ estimate the self-sampling and cross-sampling, where the variables m, x correspond to the embedding vectors of $\boldsymbol{m}, \boldsymbol{x}$. $p(.)$ and $q(.)$ are query embedding functions parameterized as networks. Therefore, we utilize the attention mechanism to estimate the self-sampling y_m and cross-sampling y_x as shown in Figure 2:

$$y_m = \mathbf{V}_{\boldsymbol{M}} \cdot \mathrm{softmax}(\mathbf{Q}_{\boldsymbol{M}}^{\top}\mathbf{K}_{\boldsymbol{M}}) \tag{3}$$

where Eq.(3) denotes self-sampling attention to compute intrinsic effect of fusion scene features M and confounding factors C.

$$y_x = \mathbf{V}_{\boldsymbol{C}} \cdot \mathrm{softmax}(\mathbf{Q}_{\boldsymbol{C}}^{\top}\mathbf{K}_{\boldsymbol{C}}) \tag{4}$$

where Eq.(4) represents the cross-sampling attention to compute the mutual effect between the fusion scene features M and confounding factors C. In the implementation of two equations, $\mathbf{Q}_{\boldsymbol{M}}$ and $\mathbf{Q}_{\boldsymbol{C}}$ are derived from $p(\mathrm{X})$ and $q(\mathrm{X})$. $\mathbf{K}_{\boldsymbol{M}}$ and $\mathbf{V}_{\boldsymbol{M}}$ are obtained from the current input sample, while $\mathbf{K}_{\boldsymbol{C}}$ and $\mathbf{V}_{\boldsymbol{C}}$ come from other samples in the training set, serving as the global dictionary compressed from the whole training dataset. Specifically, we initialize this dictionary by using K-means clustering [17] on all the embeddings of samples in the training set. To obtain the final output y_{iec} of the IEC block, we employ an FFN to integrate the self-sampling estimation y_m ans cross-sampling estimation y_x, formulated as: $y_{iec} = \mathrm{FFN}(y_m + y_x)$.

3.4 Two-Stage Training Optimization

Due to the lack of scene perception abilities in Video-LLaVA, we design a two-stage training process, where scene-tuning stage is pre-tuned to perceive omni-scene information, while Omni-SILA

Table 1: Comparison of several Video-LLMs and our ICM approach on Explicit and Implicit Omni-SILA dataset for identifying and locating sentiments. The ↓ beside FNRs indicates the lower the metric, the better the performance. Bold and underlined indicate the highest and second-highest performance, respectively (the same below).

Approach	Explicit Omni-SILA Dataset							Implicit Omni-SILA Dataset						
	Acc	F2	FNRs↓	mAP@IoU 0.1	mAP@IoU 0.2	mAP@IoU 0.3	Avg	Acc	F2	FNRs↓	mAP@IoU 0.1	mAP@IoU 0.2	mAP@IoU 0.3	Avg
mPLUG-Owl	60.33	59.57	71.37	30.30	12.20	3.36	15.29	28.88	30.06	73.98	31.42	13.21	4.46	16.36
PandaGPT	64.22	64.12	49.12	28.28	17.17	7.98	17.81	32.48	33.87	49.62	29.36	18.28	8.87	18.83
Valley	65.75	65.01	56.07	31.35	15.15	6.76	17.75	34.66	35.94	53.49	32.24	16.26	7.66	18.75
VideoChat	66.57	65.80	44.50	30.93	20.62	8.25	22.63	35.12	36.44	50.79	31.96	21.73	9.26	20.98
Video-ChatGPT	67.88	66.84	61.26	25.56	18.89	10.00	18.15	37.82	39.31	61.47	26.65	19.91	11.03	19.19
ChatUniVi	67.23	66.57	61.81	18.82	10.61	9.05	12.83	37.95	38.88	62.52	19.89	11.62	10.02	13.84
Video-LLaVA	68.19	67.08	44.32	31.41	15.78	8.82	18.67	40.02	41.88	50.34	32.41	16.79	9.92	19.71
ICM	**71.41**	**70.21**	**33.38**	**31.91**	**23.39**	**18.75**	**25.21**	**47.39**	**48.36**	**32.76**	**34.79**	**26.14**	**19.08**	**27.88**
w/o SBM	69.32	68.36	37.92	30.33	22.23	15.59	22.72	43.18	44.52	40.11	32.44	23.49	15.65	23.68
w/o IEC	69.71	68.82	35.85	31.27	23.23	16.53	23.68	44.12	45.08	38.62	33.18	24.20	16.23	24.54
w/o scene-tuning	67.87	66.32	43.59	26.80	18.51	12.05	19.12	39.64	40.76	49.74	27.88	19.65	13.14	20.23

tuning stage is trained to address the Omni-SILA task better via the perception abilities of scene information, detailed as follows.

For **Scene-Tuning** stage, we utilize four manually annotated instruction datasets (detailed in Section 4.1) to pre-tune Video-LLaVA, aiming to evoke the scene perception abilities of Video-LLaVA, where the model is asked to "*Please describe the facial/action/image region/background*". For **Omni-SILA Tuning** stage, we meticulously construct an Omni-SILA dataset (detailed in Section 4.1) to make our ICM approach better tackling the Omni-SILA task through instruction tuning, where the ICM approach is asked through the instruction "*Please identify and locate the implicit visual sentiment, and attribute it*" as shown in Figure 2. Note that the instruction will be text tokenized inside Video-LLaVA to obtain the textual token y_t, which is added with the normalized combination of the SBM block output y_{moe} and IEC block output y_{iec}. Thus, the input of the LLM inside Video-LLaVA will be "$\text{Norm}(y_{\text{moe}} + y_{\text{iec}}) + y_t + y_v$", where y_v denotes the visual features encoded by intrinsic visual encoder LanguageBind [74] of Video-LLaVA. Moreover, the whole loss of our ICM approach can be represented as $\mathcal{L} = \mathcal{L}_{\text{lm}} + \mathcal{L}_{\text{rb}}$, where $\mathcal{L}_{\text{lm}}$ is the original language modeling loss of the LLM.

4 Experimental Settings

4.1 Datasets Construction

To assess the effectiveness of our ICM approach for the Omni-SILA task, we construct instruction datasets for two stages.

For **Scene-Tuning** stage, we choose CMU-MOSEI [64], HumanML3D [14], RefCOCO [21] and Place365 [73] four datasets and manually construct instructions to improve Video-LLaVA's ability in understanding facial expression, human action, object relations and visual backgrounds four scenes. For instance in Figure 2, with the instruction "*Please describe the facial/action/image region/background*", the responses are "*A woman with sad face./A person waves his hand./A dog <0.48,2.23,1.87,0.79> on the grass./A blue sky.*". Particularly, we use SAM-V1 [22] to segment objects and capture visual backgrounds in Place365. Since CMU-MOSEI and HumanML3D contain over 20K videos, we sample frames at an appropriate rate to obtain 200K frames. To ensure scene data balance, we randomly select 200K images from RefCOCO and Place365.

For **Omni-SILA Tuning** stage, we construct an Omni-SILA tuning dataset consisting of 202K video clips, and we sample 8 frames for each video clip, resulting in 1.62M frames. This dataset consists of an explicit Omni-SILA dataset (training: 52K videos, test: 25K videos) and an implicit Omni-SILA dataset (training: 102K videos, test: 23K videos). The explicit Omni-SILA dataset is based on public TSL-300 [69], which contains explicit *positive*, *negative* and *neutral* three visual sentiment types. Due to its lack of sentiment attributions, we leverage GPT-4V [59] twice to generate and summarize the visual sentiment attribution of each frame from four scene aspects, and manually check and adjust inappropriate attributions. Implicit Omni-SILA dataset is based on public CUVA [10], which contains implicit *Fighting* (1), *Animals Hurting People* (2), *Water Incidents* (3), *Vandalism* (4), *Traffic Accidents* (5), *Robbery* (6), *Theft* (7), *Traffic Violations* (8), *Fire* (9), *Pedestrian Incidents* (10), *Forbidden to Burn* (11), *Normal* twelve visual sentiment types. We manually construct instructions for each video clip. Specifically, with the beginning of instruction "*You will be presented with a video. After watching the video*", we ask the model to identify "*please identify the explicit/implicit visual sentiments in the video*", locate "*please locate the timestamp when ...*", and attribute "*please attribute ... considering facial expression, human action, object relations and visual backgrounds*" visual sentiments. The corresponding responses are "*The explicit/implicit visual sentiment is ...*", "*The location of ... is from 4s to 18s.*", "*The attribution is several ...*". Particularly, due to the goal of identifying, locating and attributing visual sentiments, we evaluate our ICM approach by inferring three tasks with the same instruction on both explicit and implicit Omni-SILA datasets.

4.2 Baselines

Due to the requirement of interaction and pre-processing videos, traditional VSU approaches are not directly suitable for our Omni-SILA task. Therefore, we choose several advanced Video-LLMs as baselines. **mPLUG-Owl** [60] equips LLMs with multimodal abilities via modular learning. **PandaGPT** [42] shows impressive and emergent cross-modal capabilities across six modalities: image/video, text, audio, depth, thermal and inertial measurement units. **Valley** [33] introduces a simple projection to unify video, image and language modalities with LLMs. **VideoChat** [26] designs a VideoChat-Text

Table 2: Comparison of several Video-LLMs and our ICM approach on Explicit and Implicit Omni-SILA datasets for visual sentiments attributing, where GPT-based and Human indicate two methods to evaluate the Atr-R metric.

Approach	Explicit Omni-SILA Dataset					Implicit Omni-SILA Dataset				
	Sem-R	Sem-C	Sen-A	Atr-R		Sem-R	Sem-C	Sen-A	Atr-R	
				GPT-based	Human				GPT-based	Human
mPLUG-Owl	0.216	42.06	53.23	6.57	2.92	0.506	59.65	69.68	5.51	2.09
PandaGPT	0.231	43.12	56.72	6.73	3.08	0.516	60.47	72.65	5.68	2.68
Valley	0.252	45.41	57.93	6.94	3.36	0.535	63.36	70.72	6.07	2.46
VideoChat	0.243	45.06	58.16	7.06	3.51	0.527	63.79	71.65	6.29	2.95
Video-ChatGPT	0.272	48.55	60.54	7.39	3.89	0.558	65.53	75.37	6.75	3.42
ChatUniVi	0.254	46.69	59.33	7.24	3.72	0.532	63.57	74.54	6.87	3.15
Video-LLaVA	0.266	47.27	61.40	7.95	4.05	0.547	64.45	74.12	7.04	3.38
ICM	**0.290**	**54.79**	**65.38**	**9.02**	**4.95**	**0.599**	**73.57**	**81.94**	**8.89**	**4.74**
w/o SBM	0.275	49.76	63.44	8.36	4.21	0.561	68.07	77.22	7.51	3.77
w/o IEC	0.280	50.22	63.62	8.52	4.26	0.567	69.65	78.66	7.87	3.95
w/o scene tuning	0.252	45.92	60.53	7.76	3.96	0.539	63.79	73.63	6.86	3.19

module to convert video streams into text and a VideoChat-Embed module to encode videos into embeddings. **Video-ChatGPT** [34] combines the capabilities of LLMs with a pre-trained visual encoder optimized for spatio-temporal video representations. **ChatUniVi** [20] uses dynamic visual tokens to uniformly represent images and videos, and leverages multi-scale representations to capture both high-level semantic concepts and low-level visual details. **Video-LLaVA** [31] aligns the representation of images and videos to a unified visual feature space, and uses a shared projection layer to map these unified visual representations to the LLMs.

4.3 Implementation Details

Since the above models target different tasks and employ different experimental settings, for a fair and thorough comparison, we re-implement these models and leverage their released codes to obtain experimental results on our Omni-SILA datasets. In our experiments, all the Video-LLMs size is 7B. The hyper-parameters of these baselines remain the same setting reported by their public papers. The others are tuned according to the best performance. For ICM approach, during the training period, we use AdamW as the optimizer, with an initial learning rate 2e-5 and a warmup ratio 0.03. We fine-tune Video-LLaVA (7B) using LoRA for both scene-tuning stage and Omni-SILA stage, and we set the dimension, scaling factor, dropout rate of the LoRA matrix to be 16, 64 and 0.05, while keeping other parameters at their default values. The parameters of MTCNN, HigherHRNet, RelTR and ViT are frozen during training stages. The number of experts in Causal MoE is 4, and the layers of each expert are set to be 8. The hyper-parameters α and β of $\mathcal{L}_{rb}$ are set to be 1e-4 and 1e-2. ICM approach is trained for three epochs with a batch size of 8. All training runs on 1 NVIDIA A100 GPU with 40GB GPU memory. It takes around 18h for scene-tuning stage, 62h for training Omni-SILA stage and 16h for inference.

4.4 Evaluation Metrics

To comprehensively evaluate the performance of various models on the Omni-SILA task, we use commonly used metrics and design additional task-specific ones. We categorized these evaluation metrics into three tasks, as described below.

• **Visual Sentiment Identifying (VSI)**. We leverage Accuracy (Acc) to evaluate the performance of VSI following Yang et al. [57]. Besides, we prioritize Recall over Precision and report F2-score [69].

• **Visual Sentiment Locating (VSL)**. Following prior studies [23, 65], we use mAP@IoU metric to evaluate VSL performance. This metric is calculated as the mean Average Precision (mAP) under different intersections over union (IoU) thresholds (0.1, 0.2 and 0.3). More importantly, we emphasize false-negative rates (FNRs) [27, 70], denoted as: $\frac{number\ of\ false-negative\ frames}{number\ of\ positive\ frames}$, which refer to the rates of "*misclassifying a positive/normal frame as negative*". FNRs indicate that it is preferable to classify all timestamps as negative than to miss any timestamp associated with negative sentiments, as this could lead to serious criminal events.

• **Visual Sentiment Attributing (VSA)**. We design four specific metrics to comprehensively evaluate the accuracy and rationality of generated sentiment attributions. Specifically, semantic relevance (Sem-R) leverages the Rouge score to measure the relevance between generated attribution and true cause. Semantic consistency (Sem-C) leverages cosine similarity to assess the consistency of generated attribution and true cause. Sentiment accuracy (Sen-A) calculates the accuracy between generated attribution and true sentiment label. Attribution rationality (Atr-R) employs both automatic and human evaluations to assess the rationality of generated attributions. For automatic evaluation, we use ChatGPT [36] to score based on two criteria: sentiment overlap and sentiment clue overlap, on a scale of 1 to 10. For human evaluation, three annotators are recruited to rate the rationality of the generated attributions on a scale from 1 to 6, where 1 denotes "*completely wrong*" and 6 denotes "*completely correct*". After obtaining individual scores, we average all the scores to report as the final results. Moreover, t-test [4, 13] is used to evaluate the significance of the performance.

5 Results and Discussion

5.1 Experimental Results

Table 1 and Table 2 show the performance comparison of different approaches on our Omni-SILA task (including VSI, VSL and VSA). From this table, we can see that: **(1)** For VSI, our ICM approach outperforms all the Video-LLMs on the implicit Omni-SILA dataset, and achieves comparable results on the explicit Omni-SILA dataset. For instance, compared to the best-performing Video-LLaVA, ICM achieves average improvements by 6.93% (p-value<0.01) on implicit Omni-SILA dataset and 3.18% (p-value<0.05) on explicit Omni-SILA dataset. This indicates that identifying implicit visual sentiments

Table 3: The effectiveness study of various scenes in Omni-SILA, where ✓ means that we capture the current scene. All the experiments are conducted on Explicit and Implicit Omni-SILA datasets and evaluate VSI, VSL and VSA three tasks. Fac, Act, Obj and Back are short for facial expression, human action, object relation and visual background scene, respectively.

Fac	Act	Obj	Back	Explicit Omni-SILA Dataset						Implicit Omni-SILA Dataset					
				Acc	F2	FNRs↓	mAP@tIoU	Sen-A	Atr-R	Acc	F2	FNRs↓	mAP@tIoU	Sen-A	Atr-R
	✓	✓	✓	68.81	67.79	36.65	23.25	63.29	12.38	46.18	45.93	34.52	25.77	79.26	12.70
✓		✓	✓	68.66	67.87	37.52	22.64	63.48	12.16	43.89	44.97	37.89	24.62	77.78	11.36
✓	✓		✓	70.03	69.06	34.97	24.07	64.15	13.01	45.34	46.22	34.54	26.31	79.41	12.39
✓	✓	✓		69.16	68.32	36.07	23.01	63.62	12.66	44.57	45.49	35.86	25.39	78.36	11.78
✓	✓	✓	✓	71.41	70.21	33.38	25.51	65.38	13.97	47.39	48.36	32.76	27.88	81.94	13.63

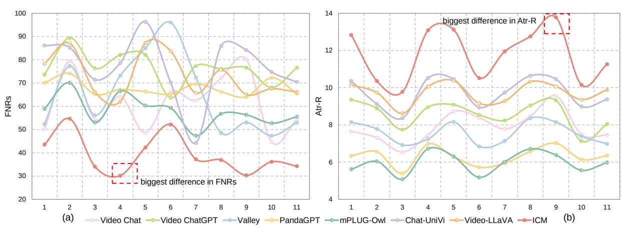

Figure 3: Two line charts to compare several well-performing Video-LLMs with our ICM approach on 11 implicit visual sentiments of FNRs (a) and Atr-R (b) two metrics, and the red boxes indicate the categories *Vandalism* of FNRs and *Fire* of Atr-R where the performance difference is biggest.

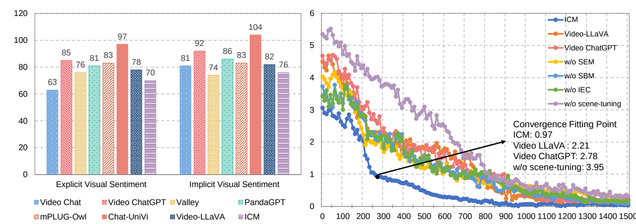

Figure 4: Two statistical charts to illustrate the efficiency of our ICM approach. The histogram (a) compares the inference time of ICM with baselines, while the line chart (b) shows the convergence of training losses of ICM, two well-performing Video-LLMs and the variants of ICM across training steps.

is more challenging than the explicit, and justifies the effectiveness of ICM in identifying what is visual sentiment. **(2)** For VSL, similar to the results on VSI task, our ICM approach outperforms all the baselines on the implicit Omni-SILA dataset while achieves comparable results on the explicit. For instance, compared to the best-performing results underlined, ICM achieves the average improvements by 5.44% (p-value<0.01) on the implicit and 3.65% (p-value<0.05) on the explicit. Particularly, our ICM approach surpasses all the Video-LLMs on FNRs by 17.58% (p-value<0.01) on the implicit and 10.94% (p-value<0.01) on explicit compared with the best results underlined. This again justifies the challenge in locating the implicit visual sentiments, and demonstrates the effectiveness of ICM in locating when the visual sentiment occurs. **(3)** For VSA, our ICM approach outperforms all the Video-LLMs on both implicit and explicit Omni-SILA datasets. Specifically, compared to the best-performing approach on all VSA metrics, ICM achieves total improvements of 5.9%, 14.28%, 10.55% and 5.18 on Sem-R, Sem-C, Sen-A and Atr-R in two datasets. Statistical significance tests show that these improvements are significant (p-value<0.01). This demonstrates that ICM can better attribute why are both explicit and implicit visual sentiments compared to advanced Video-LLMs, and further justifies the importance of omni-scene information.

5.2 Contributions of Key Components

To further study contributions of key components in ICM, we conduct a series of ablations studies, the results of which are detailed in Table 1 and Table 2. From these tables, we can see that: **(1) w/o IEC** block shows inferior performance compared to ICM, with an average decrease of VSI, VSL and VSA tasks by 4.82%(p-value<0.05), 6.6%(p-value<0.01) and 10.37%(p-value<0.01). This indicates the existence of bias between explicit and implicit information, and further justifies the effectiveness of IEC block to highlight the implicit scene information beyond the explicit. **(2) w/o SBM** block shows inferior performance compared to ICM, with an average decrease of VSI, VSL and VSA tasks by 5.99%(p-value<0.01), 9.29%(p-value<0.01) and 13.12%(p-value<0.01). This indicates the effectiveness of SBM block in modeling and balancing the explicit and implicit scene information via the MoE architecture, encouraging us to model heterogeneous information via MoE. **(3) w/o scene tuning** exhibits obvious inferior performance compared to ICM, with an average decreases of VSI, VSL and VSA tasks by 11.39%(p-value<0.01), 20.47%(p-value<0.01) and 23.72%(p-value<0.01). This confirms that the backbone lacks the ability to understand omni-scene information. This further demonstrates the necessity and effectiveness of pre-tuning, and encourages us to introduce more high-quality datasets to improve the scene understanding ability of Video-LLMs.

5.3 Effectiveness Study of Scene Information

To delve deeper into the impact of various scene information, we conduct a series of ablations studies, the results of which are detailed in Table 3. From this table, we can see that: **(1) w/o Facial Expression Modeling** shows more obvious inferior performance on the explicit than the implicit Omni-SILA dataset, with the total decrease of VSI, VSL and VSA by 5.02%(p-value<0.01), 5.53%(p-value<0.01) and 3.68%(p-value<0.05) on the explicit; 3.64%(p-value<0.05), 3.87%(p-value<0.05) and 3.61%(p-value<0.05) on the implicit. This is reasonable that most of videos in the implicit dataset may have no visible faces. **(2) w/o Human Action Modeling** exhibits obvious inferior performance on both explicit and implicit Omni-SILA datasets, with the total decrease of VSI, VSL and VSA by 5.99%(p-value<0.01), 7.7%(p-value<0.01) and 5.07%(p-value<0.01). This indicates that human action information is important to recognize explicit and implicit visual sentiments. For example, we can precisely identify, locate and attribute the *theft* via *an obvious human action of stealing*.

(3) w/o Object Relation Modeling shows slight inferior performance on both explicit and implicit Omni-SILA datasets. This is reasonable that the visual sentiment of visual object relations is very subtle, unless a specific scene like *<a man, holding, guns>* can be identified as *robbery*. **(4) w/o Visual Background Modeling** exhibits obvious inferior performance on both explicit and implicit Omni-SILA datasets, with the total decrease of VSI, VSL and VSA by 4.92%(p-value<0.05), 5.39%(p-value<0.01) and 4.25%(p-value<0.05). This indicates that video background is also important to recognize explicit and implicit visual sentiments. For example, we can precisely identify, locate and attribute the *fire* via the *forest on fire with flames raging to the sky* background.

5.4 Applicative Study of ICM Approach

To study the applicability of ICM, we compare the FNRs and Atr-R of ICM with other Video-LLMs. From Table 1, we can see that ICM performs the best on the metric of FNRs. For example, ICM outperforms the best-performing Video-LLaVA by 10.94% (p-value<0.01) and 17.58% (p-value<0.01) on the explicit and implicit Omni-SILA dataset respectively. From Table 2, ICM achieves state-of-the-art performance on both implicit and explicit Omni-SILA datasets. These results indicate that ICM is effective in reducing the rates of FNRs and providing reasonable attributions, which is important in application. Furthermore, recognizing that the implicit Omni-SILA dataset comprises 11 distinct real-world crimes, we perform an analysis of each negative implicit visual sentiment on the performance of FNRs (Figure 3 (a)) and Atr-R (Figure 3 (b)). From two figures, we can see that ICM surpasses all other Video-LLMs across 11 crimes. Particularly, ICM performs best on *Vandalism* (4) in FNRs and *Fire* (9) in Atr-R. This indicates that ICM is effective in reducing FNRs and improving the interpretability of negative implicit visual sentiments, encouraging us to consider the omni-scene information, such as *human action* in *Vandalism* and *visual background* in *Fire*, for precisely identifying, locating and attributing visual sentiments.

5.5 Efficiency Analysis of ICM Approach

To study the efficacy of ICM, we compare the inference time of ICM with other Video-LLMs (Figure 4 (a)), and analyze the convergence of training loss for Video-ChatGPT, Video-LLaVA, ICM and its variants over different training steps (Figure 4 (b)). As shown in Figure 4 (a), we can see that ICM achieves little difference in inference time compared with other Video-LLMs. This is reasonable because MoE can improve the efficiency of inference [9, 51], encouraging us to model omni-scene information via MoE architecture. From Figure 4 (b), we can see that: **(1)** ICM shows fast convergence compared to Video-LLMs. At the convergence fitting point, the loss of ICM is 0.97, while Video-LLaVA is 2.21. This indicates that ICM has more high efficiency than other Video-LLMs, which further shows the potential of ICM for quicker training times and less source use, thereby improving its applicative use in real-world applications. **(2)** ICM shows fast convergence compared to its variants, indicating that the integration of MoE architecture and causal intervention can accelerate the convergence process. **(3)** ICM shows fast convergence compared to without scene-tuning, where the loss is 3.95 at the convergence fitting point. This again justifies the importance of scene understanding before Omni-SILA tuning.

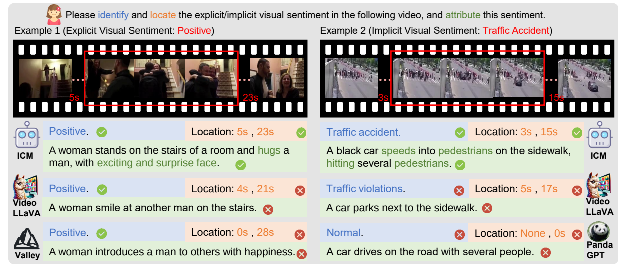

Figure 5: Two samples to compare ICM with other baselines.

5.6 Qualitative Analysis

As illustrated in Figure 5, we provide a qualitative analysis to intuitively compare the performance of ICM with other Video-LLMs on the Omni-SILA task. Specifically, we randomly select two samples from each of explicit and implicit Omni-SILA datasets, asking these approaches to "*Identify and locate the visual sentiment in the following video, and attribute this sentiment*". Due to the space limit, we choose the top-3 well-performing approaches. From this figure, we can see that: **(1)** Identifying and locating implicit visual sentiment is more challenging than the explicit. For instance, Video-LLaVA can roughly locate and precisely identify *positive* sentiment in Example 1, but has difficulties in locating *traffic accident* in Example 2. However, ICM can precisely identify and locate the *traffic accident*. **(2)** Attributing both explicit and implicit visual sentiments is challenging. All the advanced Video-LLMs are difficult to attribute visual sentiments, even some approaches are nonsense. While ICM can provide reasonable attributions due to the capture of omni-scene information, such as *surprise* face, *hugging* action in Example 1, and *speeds, hitting* action, *pedestrians* visual background in Example 2. This again justifies the importance of omni-scene information, and effectiveness of ICM for capturing such information.

6 Conclusion

In this paper, we address a new Omni-SILA task, aiming to identify, locate and attribute the visual sentiment in videos, and we propose an ICM approach to address this task via omni-scene information. The core components of ICM involve SBM and IEC blocks to effectively model scene information and highlight the implicit scene information beyond the explicit. Experimental results on our constructed Omni-SILA datasets demonstrate the superior performance of ICM over several advanced Video-LLMs. In our future work, we would like to train a Video-LLM supporting more signals like audio from scratch to further boost the performance of sentiment identifying, locating and attributing in videos. In addition, we would like to leverage some light-weighting technologies (e.g., LLM distillation and compression) to further improve ICM's efficiency.

Acknowledgments

We thank our anonymous reviewers for their helpful comments. This work was supported by three NSFC grants, i.e., No.62006166, No.62376178 and No.62076175. This work was also supported by a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). This work was also partially supported by Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

[1] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. 2024. MiniGPT4-Video: Advancing Multimodal LLMs for Video Understanding with Interleaved Visual-Textual Tokens. *CoRR* (2024).

[2] Houlun Chen, Xin Wang, Xiaohan Lan, Hong Chen, Xuguang Duan, Jia Jia, and Wenwu Zhu. 2023. Curriculum-Listener: Consistency- and Complementarity-Aware Audio-Enhanced Temporal Sentence Grounding. In *Proceedings of ACM MM 2023*. 3117–3128.

[3] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. *CoRR* (2023).

[4] Xiao Chen, Changlong Sun, Jingjing Wang, Shoushan Li, Luo Si, Min Zhang, and Guodong Zhou. 2020. Aspect Sentiment Classification with Document-level Sentiment Preference Modeling. In *Proceedings of ACL 2020*. Association for Computational Linguistics, 3667–3677.

[5] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. 2020. HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. In *Proceedings of CVPR 2020*. 5385–5394.

[6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

[7] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. 2023. RelTR: Relation Transformer for Scene Graph Generation. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 9 (2023), 11169–11183.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of ICLR 2021*.

[9] Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Jun Zhao, Wei Shen, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. LoRAMoE: Revolutionizing Mixture of Experts for Maintaining World Knowledge in Language Model Alignment. *CoRR* (2023).

[10] Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, Dajiu Huang, Jing Feng, Linli Chen, Can Zhang, Xuhuan Li, Hao Zhang, Jianhang Chen, Qimei Cui, and Xiaofeng Tao. 2024. Uncovering what, why and How: A Comprehensive Benchmark for Causation Understanding of Video Anomaly. In *Proceedings of CVPR 2024*. IEEE, 18793–18803.

[11] Florian Eyben, Felix Weninger, Nicolas Lehment, Björn Schuller, and Gerhard Rigoll. 2013. Affective video retrieval: Violence detection in Hollywood movies by large-scale segmental feature extraction. *PloS one* 8, 12 (2013), e78506.

[12] Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024. VITRON: A Unified Pixel-level Vision LLM for Understanding, Generating, Segmenting, Editing.

[13] Xiaoya Gao, Jingjing Wang, Shoushan Li, Min Zhang, and Guodong Zhou. 2022. Cognition-driven multimodal personality classification. *Sci. China Inf. Sci.* 65, 10 (2022).

[14] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating Diverse and Natural 3D Human Motions from Text. In *Proceedings of CVPR 2022*. 5142–5151.

[15] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2024. OneLLM: One Framework to Align All Modalities with Language. In *Proceedings of CVPR 2024*. 26574–26585.

[16] Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In *Proceedings of EMNLP 2021*. 9180–9192.

[17] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* 28, 1 (1979), 100–108.

[18] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *Proceedings of ACM MM 2020*. 1122–1131.

[19] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. In *Proceedings of EMNLP 2022*. 7837–7851.

[20] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding. In *Proceedings of CVPR 2024*. 13700–13710.

[21] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of EMNLP 2014*. 787–798.

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. 2023. Segment Anything. In *Proceedings of ICCV 2023*. 3992–4003.

[23] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. 2021. Weakly-supervised Temporal Action Localization by Uncertainty Modeling. In *Proceedings of AAAI 2021*. 1854–1862.

[24] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023. Otter: A Multi-Modal Model with In-Context Instruction Tuning. *CoRR* (2023).

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of ICML 2023*. 19730–19742.

[26] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. VideoChat: Chat-Centric Video Understanding. *CoRR* (2023).

[27] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. 2014. Anomaly Detection and Localization in Crowded Scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 1 (2014), 18–32.

[28] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, Zhida Huang, and Tao Wang. 2024. GroundingGPT: Language Enhanced Multi-modal Grounding Model. In *Proceedings of ACL 2024*. 6657–6678.

[29] Zeju Li, Chao Zhang, Xiaoyan Wang, Ruilong Ren, Yifan Xu, Ruifei Ma, Xiangde Liu, and Rong Wei. 2024. 3DMIT: 3D Multi-Modal Instruction Tuning for Scene Understanding. In *ICME 2024-Workshops*. 1–5.

[30] Zheng Lian, Haiyang Sun, Licai Sun, Jiangyan Yi, Bin Liu, and Jianhua Tao. 2024. AffectGPT: Dataset and Framework for Explainable Multimodal Emotion Recognition. *CoRR* (2024).

[31] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. *CoRR* (2023).

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Proceedings of NeurIPS 2023*.

[33] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. 2023. Valley: Video Assistant with Large Language model Enhanced abilitY. *CoRR* (2023).

[34] Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *Proceedings of ACL 2024*. 12585–12602.

[35] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional Chain-of-Thought Prompting for Large Multimodal Models. In *Proceedings of CVPR 2024*. 14420–14431.

[36] OpenAI. 2023. GPT-4 Technical Report. *CoRR* (2023).

[37] Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

[38] Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. 2024. From Sparse to Soft Mixtures of Experts. In *Proceedings of ICLR 2024*.

[39] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2023. TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding. *CoRR* (2023).

[40] Christian Schulze, Dominik Henter, Damian Borth, and Andreas Dengel. 2014. Automatic Detection of CSA Media by Multi-modal Feature Fusion for Law Enforcement Support. In *Proceedings of ICMR 2014*. 353.

[41] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1 (2014), 1929–1958.

[42] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. PandaGPT: One Model To Instruction-Follow Them All. *CoRR* (2023).

[43] Peter T. Szymanski and Michael D. Lemmon. 1993. Adaptive mixtures of local experts are source coding solutions. In *Proceedings of ICNN 1993*. 1391–1396.

[44] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR* (2023).

[45] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of ACL 2019*. 6558–6569.

[46] Yao-Hung Hubert Tsai, Martin Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Multimodal Routing: Improving Local and Global Interpretability of Multimodal Language Analysis. In *Proceedings of EMNLP 2020*. 1823–1833.

[47] Paulo Vitorino, Sandra Avila, Mauricio Perez, and Anderson Rocha. 2018. Leveraging deep neural networks to fight child pornography in the age of social media. *J. Vis. Commun. Image Represent.* 50 (2018), 303–313.

[48] Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. 2024. HawkEye: Training Video-Text LLMs for Grounding Text in Videos. *CoRR* (2024).

[49] Zhanyu Wang, Longyue Wang, Zhen Zhao, Minghao Wu, Chenyang Lyu, Huayang Li, Deng Cai, Luping Zhou, Shuming Shi, and Zhaopeng Tu. 2023.

GPT4Video: A Unified Multimodal Large Language Model for Instruction-Followed Understanding and Safety-Aware Generation. *CoRR* (2023).

[50] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. 2024. VadCLIP: Adapting Vision-Language Models for Weakly Supervised Video Anomaly Detection. In *Proceedings of AAAI 2024*. 6074–6082.

[51] Xun Wu, Shaohan Huang, and Furu Wei. 2024. Mixture of LoRA Experts. In *Proceedings of ICLR 2024*.

[52] Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, and Wen-Huang Cheng. 2024. EmoVIT: Revolutionizing Emotion Insights with Visual Instruction Tuning. In *Proceedings of CVPR 2024*. 26586–26595.

[53] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of ICML 2015*. 2048–2057.

[54] Liwen Xu, Zhengtao Wang, Bin Wu, and Simon Lui. 2022. MDAN: Multi-level Dependent Attention Network for Visual Emotion Analysis. In *Proceedings of CVPR 2022*. 9469–9478.

[55] Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. 2023. EmoSet: A Large-scale Visual Emotion Dataset with Rich Attributes. In *Proceedings of ICCV 2023*. 20326–20337.

[56] Jingyuan Yang, Jie Li, Xiumei Wang, Yuxuan Ding, and Xinbo Gao. 2021. Stimuli-Aware Visual Emotion Analysis. *IEEE Trans. Image Process.* 30 (2021), 7432–7445.

[57] Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis. In *Proceedings of ACL 2023*. 7617–7630.

[58] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. 2021. Causal Attention for Vision-Language Tasks. In *Proceedings of CVPR 2021*. 9847–9857.

[59] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). *CoRR* (2023).

[60] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. *CoRR* (2023).

[61] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality. In *Proceedings of ACL 2020*. 3718–3727.

[62] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis. In *Proceedings of AAAI 2021*. 10790–10797.

[63] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of EMNLP 2017*. 1103–1114.

[64] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of ACL 2018*. 2236–2246.

[65] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. 2021. CoLA: Weakly-Supervised Temporal Action Localization With Snippet Contrastive Learning. In *Proceedings of CVPR 2021*. 16010–16019.

[66] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In *Proceedings of EMNLP 2023*. 543–553.

[67] Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Chuchu Han, Xiaonan Huang, Changxin Gao, Yuehuan Wang, and Nong Sang. 2024. Holmes-VAD: Towards Unbiased and Explainable Video Anomaly Detection via Multi-modal LLM. *CoRR* (2024).

[68] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Process. Lett.* 23, 10 (2016), 1499–1503.

[69] Zhicheng Zhang and Jufeng Yang. 2022. Temporal Sentiment Localization: Listen and Look in Untrimmed Videos. In *Proceedings of ACM MM 2022*. 199–208.

[70] Jianing Zhao, Jingjing Wang, Yujie Jin, Jiamin Luo, and Guodong Zhou. 2024. Hawkeye: Discovering and Grounding Implicit Anomalous Sentiment in Recon-videos via Scene-enhanced Video Large Language Model. In *Proceedings of ACM MM 2024*. ACM, 592–601.

[71] Sicheng Zhao, Zizhou Jia, Hui Chen, Leida Li, Guiguang Ding, and Kurt Keutzer. 2019. PDANet: Polarity-consistent Deep Attention Network for Fine-grained Visual Emotion Regression. In *Proceedings of ACM MM 2019*. 192–201.

[72] Sicheng Zhao, Xingxu Yao, Jufeng Yang, Guoli Jia, Guiguang Ding, Tat-Seng Chua, Björn W. Schuller, and Kurt Keutzer. 2022. Affective Image Content Analysis: Two Decades Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 10 (2022), 6729–6751.

[73] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 Million Image Database for Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 6 (2018), 1452–1464.

[74] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Caiwan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2024. LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment. In *Proceedings of ICLR 2024*.

[75] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *Proceedings of ICLR 2024*.