

Figure 5: Equipment for Collecting Subjective Responses of Video-SME dataset. During data acquisition, participants wear an EEG Device, facing a Video Display, with an Eye-Tracking Device below to monitor gaze. Video durations and subjective responses are recorded on an Integration Display for analysis.

A Analysis of EEG Raw Signal

Electroencephalography (EEG) stands as a pivotal method for recording the electrical activity of the brain. This is achieved through the placement of electrodes across the scalp, shown in Figure 5, to detect electrical signals from neurons. These signals are instrumental in delineating the brain’s activity patterns across various cognitive states, providing deep insights into how the brain orchestrates complex psychological emotion and cognitive processes.

EEG signals are characterized by multiple frequency bands, among which Alpha and Beta waves are paramount, each corresponding to distinct functional states of the brain [36, 38, 65]. We categorize different EEG bands according to the following frequency definitions:


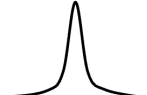
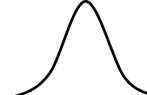
1. Alpha Waves (8–13 Hz) are emblematic of the brain’s state of relaxation and idleness. They are further categorized into three sub-bands based on their frequency range. 1) Alpha1 (8–8.9 Hz): This band is predominant when an individual is in a relaxed state with closed eyes, marking the onset of relaxation. 2) Alpha2 (9–10.9 Hz): These waves are more pronounced when the individual is relaxed yet maintains a level of alertness. 3) Alpha3 (11–12.9 Hz): This band appears as the brain relaxes further while remaining somewhat awake. The gradation within Alpha waves from Alpha1 to Alpha3 illustrates the brain’s transition from alertness to relaxation, elucidating their significance in cognitive tasks and adaptive processes.

2. Beta waves (13–30 Hz) are integral to the brain’s alertness, focused attention, and cognitive processing. 1) Beta1 (13–18 Hz): Associated with mild cognitive activities and focused attention, like reading or simple thought processes. 2) Beta2 (18–22 Hz): These waves intensify during complex cognitive tasks such as problem-solving and decision-making. 3) Beta3 (22–30 Hz): This range signifies highly focused attention and rapid cognitive processing, indicative of active information processing.

In this research, we investigate the analysis of EEG signals and the gathering of eye movement metrics to reveal the authentic subjective feedback from various demographic groups in response to advertisements. The Video-SME dataset, derived from raw signals and specific frequency bands, aids in improving the understanding of videos, especially those related to advertising.

B Data distribution of Subjectivity Task in Video-SME

Table 6: Categories & Distribution of Subjectivity Task

Task	Engagement	Emotion	EMR
Cls-1	non-cognitive [0, 1) prop. 55.0%	negative ($-\infty, -6$) prop. 29.7%	not attended [0, 0.45) prop. 29.3%
Cls-2	cognitive [1, $+\infty$) prop. 45.0%	neutral ($-6, 6$) prop. 39.0%	partially attended [0.45, 0.6) prop. 41.6%
Cls-3	–	positive ($6, +\infty$) prop. 31.2%	fully attended [0.6, $+\infty$) prop. 29.0%
Distri.			

We initiate our analysis by performing a comprehensive histogram distribution assessment of Engagement, Emotion, and EMR indicators, anchored to Audience Profiles benchmarks. As delineated in Table 6, the concluding row graphically encapsulates the distribution of each evaluated SRI. The table’s initial three rows delineate each category’s designation, value range, and the proportion of data attributed to the respective category. For instance, the Engagement indicator is bifurcated into two categories: the first, termed “non-cognitive,” spans a value range from 0 to 1 (non-inclusive), with 55.0% of observations classified under this category.

Leveraging the statistical insights derived, we proceed to discretize the extant SRI values into distinct categories, thereby facilitating the structuring of the Subjectivity Task associated with Video-SME.

C Method Algorithm and Training Hyperparameters

We introduce the Hypergraph Multi-modal Large Language Model (HMLLM), a novel approach designed to integrate and process multi-modal data, taking into full account both subjective and objective elements to comprehend advertising videos. Our training procedure is detailed in Algorithm 1. The method is grounded in the utilization of hypergraphs and large language model to effectively handle complex relationships within and across modalities.

We divide the training process into the following stages and adapt appropriate training hyperparameters, as shown in Table 7.

Algorithm 1: HMLLM: Hypergraph Multi-modal Large Language Model

Input : Video Key Frames $F = \{f_0, f_1, \dots, f_N\}$,
 Textual Prompts T ,
 Ground Truth Y_{gt} ,
 Warm Up Epoch E_0 ,
 Fine-tune Epoch E_1 ,

Initialization: $X_0 \leftarrow \text{Pre_process}(F_v)$
 $F_v \leftarrow \text{Visual_Encoder}(X_0)$
 $\text{SALM} \leftarrow \text{Initialize_SALM}(F_v, T)$
 $\text{HL-Gate} \leftarrow \text{OFF}$
 $Q \leftarrow \text{Initialize_Query}()$

// Stage I: SALM Warm Up

- 1 **for** $i \leftarrow 1$ to E_0 **do**
- 2 $K, V \leftarrow \text{QFormer}(F_v)$
- 3 $F_p \leftarrow \text{SALM_Projector}(F_v, K, V)$
- 4 $\text{SALM} \leftarrow \text{SALM_Train}(\text{SALM}, F_p, T)$
- 5 $\tilde{Y}_{qa} \leftarrow \text{SALM}(F_p, T)$
- 6 $\mathcal{L}_{ITG} \leftarrow \text{ITG_Loss}(\tilde{Y}_{qa}, Y_{gt})$
- 7 $\text{SALM_Optimizer}(\mathcal{L}_{ITG})$

// Stage II: SAL-HL Fine-tuning

- 8 $\text{HL-Gate} \leftarrow \text{SRI_Contained}(T)$ // Set HL-Gate ON.
- 9 **for** $i \leftarrow 1$ to E_1 **do**
- 10 $F_{pv} \leftarrow \text{Feature_Mixer}(F_p, F_v)$
- 11 $F_{\text{frame_level}} \leftarrow \text{Pool}(F_{pv})$
- 12 $\mathcal{G} \leftarrow \text{Construct_Hypergraph}(F_{\text{frame_level}}, \mathcal{R})$
- 13 $\tilde{Y}_{sri} \leftarrow \text{HGNN}(\mathcal{G}, F_{\text{frame_level}})$
- 14 $\mathcal{L}_{CE} \leftarrow \text{Cross_Entropy}(\tilde{Y}_{sri}, Y_{sri})$
 // \mathcal{L}_{ITG} is obtained same with the stage I
- 15 $\mathcal{L} \leftarrow \text{Combined_Loss}(\mathcal{L}_{ITG}, \mathcal{L}_{CE}, \lambda)$
- 16 $\text{Joint_Optimizer}(\text{SALM}, \text{HGNN}, \mathcal{L})$

C.1 Initialization for Model Parameters

The process begins with the extraction of key frames from the input video, which are then pre-processed to standardize the input format. These initial visual features are encoded using a visual encoder, producing a set of feature vectors. Simultaneously, textual prompts are prepared for processing. A query set is initialized, marking the starting point for our model’s learning process.

C.2 Stage I: SALM Warm Up

Stage I is dedicated to the training of the SRI-Aware Language Model (SALM), with the goal of enhancing its capabilities in language generation and reasoning inference. In this stage, visual features are converted into key-value pairs using a query-former mechanism, essential for the attention processes. These features are then fed through the SALM projector, which enriches the model’s understanding by integrating textual prompts.

The training of SALM spans 10 epochs, as detailed in Table 7, with a particular focus on minimizing the Image-Text Grounding (ITG) loss. This step is crucial for ensuring that the model’s outputs are in alignment with the ground truth, thereby optimizing performance.

C.3 Stage II: SAL-HL Fine-tuning

Stage II shifts the focus to fine-tuning the SRI-Aware Language Hypergraph Learning (SAL_HL) component, with the objective of enhancing the model’s capability to mimic the subjective perceptual capacities of the brain. After completing the initial warm-up phase, the model enters the fine-tuning stage, marked by the activation of the Hypergraph Learning (HL) gate. This process enriches the model’s multi-modal context by combining these features into a pooled frame-level representation. This representation then forms the foundation for constructing a hypergraph that captures the intricate interconnections among data points.

Following the construction of the hypergraph, a Hypergraph Neural Network (HGNN) is employed to process the hypergraph structure, allowing the model to leverage the complex connections present within the data. The output generated by the HGNN undergoes fine-tuning, utilizing a Cross-Entropy loss in conjunction with the Image-Text Grounding (ITG) loss that was emphasized during the warm-up phase. This amalgamation of loss functions serves as a directive for the optimization process, targeting both the SRI-Aware Language Model (SALM) and the HGNN components. This strategic approach ensures a unified and coherent learning experience throughout the two distinct stages of the model’s training, fostering a comprehensive understanding and adaptation to the intricacies of the data.

Following [42], during Stage 2, we incorporate Low-Rank Adaptation (LoRA) [30] modules into the SALM with a configuration of rank 16, an alpha value of 32, and a dropout rate of 0.1. Within the HGNN, it is imperative to adjust the input based on pooled frame-level representation, setting the Number of Vertices and Hyperedges to 8×98 and the Channel of Vertex Representation to 1024. For the hypergraph’s internal configuration, we adhere to the commonly used settings as illustrated in the Table 7, ensuring a balance between training effectiveness and model size.

C.4 Summary

HMLLM stands as a comprehensive framework that leverages the strengths of hypergraph structures and multi-modal data integration. Through its two-stage training process, it achieves a deep understanding of the relationships within and across modalities, paving the way for advanced applications in multi-modal data processing and generation.

D Computational Complexity and Resource Utilization of HMLLM

We have recorded the training time on eight A100 GPUs, each with 40GB of memory, and the inference time on a single A100 GPU with the same specifications. Our proposed HMLLM is generally on par with other models supporting video multi-modalities in terms of parameter count, training time, and inference time.

E Ablation Study on Effect of λ .

In the course of training HMLLM, a series of ablation studies were carried out on the λ in Equation 7, the results of which are detailed in Table 9. The integration of the SAL-HL Module significantly bolstered the model’s proficiency in capturing subjective metrics, culminating in optimal performance at a λ value of 0.1. Beyond this

Table 7: Training Hyperparameters for different stages.

config	Stage1	Stage2
	SALM Warm Up	SAL-HL Fine-tune
input frame	8	
input resolution	224	
max text length	512	
optimizer	AdamW	
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$	
weight decay	0.02	
learning rate schedule	cosine decay	
learning rate	1e-4	2e-5
batch size	128	64
warmup epochs	0.5	1
total epochs	10	20
λ of \mathcal{L}	0	0.1
augmentation	flip, MultiScaleCrop [0.5, 1]	
Vertices number of Hypergraph	-	8*98
Hyperedges number of Hypergraph	-	8*98
Channel of Vertex representation	-	1024
K of hypergraph construction	-	3, 4, 5
Average out-degree of vertices	-	12.5
In-degree of hyperedges	-	3, 4, 5

Table 8: Comparative Analysis of Model Parameters

Model	Video-Chat2	HMLLM (Ours)
Total Parameters (Billions)	7.2	7.2
Training Time (H/Epoch)	13.5	14.3
Inference Time (s/Video)	6.2	6.2
ACC on Task2	49.27	50.52

Table 9: Results of λ on Procotol2 of the Subjectivity Task.

λ	Engagement		Emotion		EMR	
	ACC	F1	ACC	F1	ACC	F1
0.0	60.06	60.02	39.66	40.24	44.06	45.51
0.05	62.69	63.66	42.62	42.46	52.22	54.78
0.1	64.43	64.65	43.20	48.84	51.96	56.24
0.2	62.38	63.28	38.06	39.79	50.57	50.56
0.5	60.86	60.83	40.17	42.34	50.89	52.47

threshold, any further increase in λ resulted in a slight decrease in performance, likely due to an overemphasis on SAL-HL features at the expense of the SALM’s inferential capabilities. Despite this, HMLLM consistently surpasses the baseline model ($\lambda = 0.0$) in terms of inference strength, demonstrating the beneficial impact of the hypergraph integration on the model’s overall performance.

F Zero-shot Prompt for Subjectivity Task

For the Subjectivity Task, we conducted Zero-shot inference on commercial and open-source models. During this process, we tested various Prompts to enable reasoning by the language models. The Prompts we selected are as follows.

As an AI model simulating EEG indicator analysis, your task is to systematically evaluate a user’s feelings after viewing video or frames content using the provided EEG indicators: Cognitive Engagement (CE), Emotional Recognition (ER), and Eye Movement Ratio (EMR).

Effective analysis requires detailed video content information themes, narrative structure, visual and auditory elements and viewer attributes, including age, gender, preferences, and experiences. You will integrate the EEG indicators’ definitions and probabilities to deduce the viewer’s cognitive engagement, emotional response, and attention level.

This involves assessing how video elements may attract or repel the viewer, grounded in psychological principles and media consumption research. But you don’t need to output the reasoning process, only the final result.

1. Cognitive Engagement (CE) Definitions:

- A. Non-Cognition:(55% probability)The viewer shows minimal interest or understanding, resulting in low EEG activity.
- B. Cognition:(45% probability)The viewer understands and relates to the video, evidenced by increased EEG activity.

2. Emotional Recognition (ER) Definitions:

- A. Negative:(30% probability)Dislikes certain video elements (e.g., conflict, unappealing objects).
- B. Neutral:(40% probability)Feels indifferent towards the video content (e.g., mundane tasks).
- C. Positive:(30% probability)Experiences enjoyment or excitement (e.g., appealing scenes).

3. Eye Movement Ratio (EMR) Definitions:

- A. Not Attended:(30% probability)Viewing ratio ≤ 0.45 , possibly due to unattractive visuals or cognitive dissonance.
- B. Partially Attended:(40% probability)Viewing ratio between 0.45 and 0.6, suggesting some attractive elements.
- C. Fully Attended:(30% probability)Viewing ratio > 0.6 , indicating high appeal and mood enhancement.

Please ensure your analysis follows this format with no additional output:

CE: B;

ER: B;

EMR: C

Finally, analyze the specified question without extraneous , focusing on the indicators’ specific index based on probabilities.

{[The Question in Test Set]}

It is worth noting that to prevent the model from overfitting to specific choices, we randomized the options and tested them three times, taking the average result as the conclusive outcome.

G Visualization and Qualitative Analysis

In the realm of advertising, the use of metaphors, scenic portrayals and related content is prevalent. Our Video-SME dataset is meticulously crafted to support both subjective and objective analyses, thereby offering a comprehensive understanding of video advertising content. It uniquely bridges the gap between these analyses, with objective comprehension bolstering subjective interpretation. This fusion enables the exploration of qualitative aspects such as Engagement, Emotion and Eye Movement Ratio (EMR) across various demographics.

As shown in Figure 6, Part A showcases the Video-SME’s ground truth, distinguished by its detailed annotations and extensive response lengths. Meanwhile, Part B delineates a comparative analysis



Figure 6: More qualitative analysis of Video-SME. Green signifies accurate descriptions, while red denotes incorrect or hallucinatory responses.

among the outputs generated by the Gemini-Pro-vision, Video-LLaVA, and VideoChat2 models against our HMLLM.

For instance, an energy drink advertisement as shown in the bottom of Figure 6, HMLLM uniquely captures both the overt (a motorcycle rider and a camel) and the covert (the product's essence

of vitality and adventure) elements of the advertisement. This comprehensive analysis extends to the advertisement's main audience, design principles, visual narratives, and product attributes, showcasing our model's superior capability in extracting and interpreting complex thematic elements.