SCG-HF: Semantic Consistency Grouping for Hierarchical Fusion in Video Emotion Recognition

Yujing Rao*
School of Computer Science
Wuhan University
Wuhan, Hubei
raoyujing@whu.edu.cn

Muci Li*
School of Computer Science
Wuhan University
Wuhan, Hubei
2023302121229@whu.edu.cn

He Li †
School of Computer Science
Wuhan University
Wuhan, Hubei
lihe404@whu.edu.cn

Mang Ye †
School of Computer Science
Wuhan University
Wuhan, Hubei
yemang@whu.edu.cn

Abstract

Video Emotion Recognition (VER) aims to identify and understand the emotional states of characters by analyzing visual and auditory information in videos. However, conventional rigid frame sampling strategies for long videos tend to fragment continuous emotional expressions. Moreover, existing fusion methods often fail to properly model temporal dependencies and may suffer from future information leakage. To address these limitations, we propose ***SCG-HF***, a novel framework that preserves semantic consistency through adaptive grouping and hierarchical fusion. Specifically, we introduce ***Semantic Consistency Grouping (SCG)***, which dynamically segments videos into semantically coherent units based on feature similarity, thereby maintaining the integrity of emotional patterns. Furthermore, we design a three-level ***Hierarchical Fusion (HF)*** architecture to capture emotional dynamics at different temporal granularities: (i) frame-level refinement enhances subtle local emotional cues; (ii) segment-level history-guided fusion models temporal evolution across segments while strictly preventing future information leakage via attention; and (iii) sample-level contrastive alignment synchronizes audio and visual representations in a shared latent space. Extensive experiments on the VideoEmotion-8 and Ekman-6 benchmarks demonstrate that SCG-HF achieves state-of-the-art performance.

1 Introduction

Video emotion understanding refers to the process of automatically recognizing and interpreting the emotions or emotional states expressed in a video by analyzing its visual and auditory information [1, 2]. With the rapid growth of short video platforms, video has become a primary medium for humans to express emotions and opinions. Emotion analysis plays a significant role in various fields such as video content recommendation, social interaction, and advertising [3, 4]. Accurate emotion understanding not only enhances user experience but also drives advancements in human-computer interaction, emotional computing, and other related technologies.

*These authors contributed equally to this work.
†Corresponding authors: He Li and Mang Ye

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

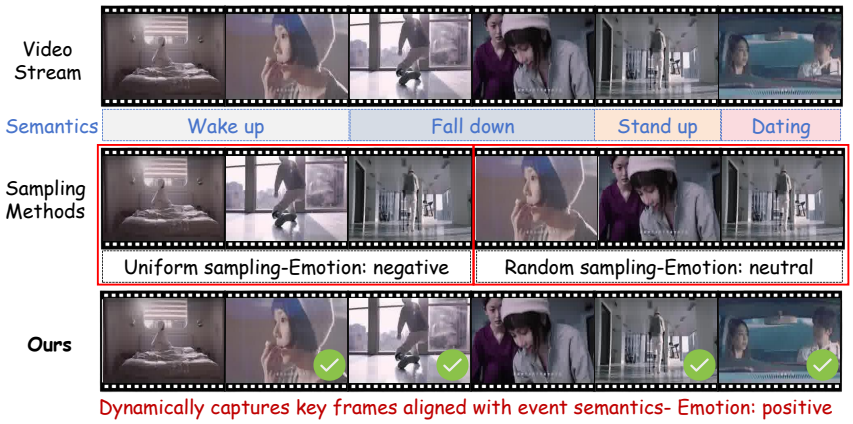

Figure 1: Illustration of different sampling strategies using a rehabilitation video. Uniform and random sampling may omit events needed to interpret the overall storyline. SCG organizes the video into semantic-consistency-based segments while retaining their temporal order, providing structured temporal units for subsequent emotion modeling.

Due to limitations in computational resources and annotation costs, most studies adopt video-level labeling methods instead of frame-level labeling [5–7]. By analyzing the overall video through limited audio-visual samples, these approaches reduce workload but sacrifice semantic integrity and temporal consistency, bringing new challenges to VER. However, these methods often focus on misleading segments. Fig. 1 illustrates this limitation with a rehabilitation video. The storyline involves a brief fall but primarily emphasizes resilience and social life. Uniform and random sampling strategies may select the negative falling frames or redundant non-emotion segments while ignoring true emotional peaks like standing up. Consequently, the illustrated samples may suggest a negative or neutral interpretation.

Related studies have explored weakly supervised temporal localization. For example, Xu *et al.* [8] proposed multimodal temporal forgery localization using only video-level annotations. Such forgery localization addresses a different objective from semantic grouping for video emotion recognition. Regardless of sampling strategy, the prerequisite for selecting keyframes is to segment the video, which is often based on uniform segmentation. Nevertheless, these approaches depend heavily on complex priors, limiting scalability. Unlike single-shot speech videos, user-generated videos typically contain numerous scene transitions. These transitions are not meaningless; they may indicate changes in semantics or different visual perspectives within the same scene. For example, in some videos, environmental information may appear only briefly. Environmental cues often determine the overall emotional tone of a video, and such seemingly secondary information serves as an important complement to human emotional expression [9, 10]. We use the term visual semantics to refer to high-level visual content within a single video, such as relative stability in scene structure, object composition, and contextual layout. This use of visual semantics relates to discussions of cinematic representation [11] and contextual information in image-based emotion recognition [12], without equating visual transitions with emotional changes. Therefore, a key challenge is ***designing a sampling strategy that preserves semantically coherent emotional content.***

Moreover, videos convey richer information than images by integrating both visual and auditory modalities [13]. Earlier multimodal approaches explored concatenation, addition, and tensor-based fusion [14, 15], while recent studies employ cross-modal attention to adaptively balance different modalities [16, 17]. However, visual and audio signals often differ in temporal granularity and intensity, leading to significant modal heterogeneity [18] and potential modal conflicts [19].

Therefore, effective modeling must address modal conflicts while capturing the dynamic evolution of emotions over time. A key challenge is ***how to achieve stable fusion that overcomes modal differences while preserving temporal emotional dynamics.***

To address these challenges, we introduce ***SCG-HF***, a framework that maintains semantic consistency through an adaptive grouping strategy for dynamic sampling and a hierarchical feature aggregation architecture. Specifically, the model extracts fine details at the frame level, captures emotional

changes across segments at the segment level, and aligns audio and visual features at the sample level to ensure a robust representation.

The main contributions are summarized as follows:

- We propose SCG, an adaptive segmentation strategy that uses visual semantic transitions to organize temporally ordered video segments for subsequent emotion modeling.
- We develop a three-level hierarchical mechanism that integrates *frame-level Local Attention Feature Refinement*, *segment-level History-Guided Fusion*, and *sample-level contrastive alignment* to capture emotional evolution across granularities while bridging the gap between audio and visual modalities.
- We show through extensive experiments that SCG-HF achieves state-of-the-art performance on several video emotion datasets.

2 Related Work

2.1 General Video Sampling and Modeling

Since analyzing every frame in a long video sequence is computationally prohibitive, selecting frames containing highly discriminative signals is crucial. Early research predominantly utilized random sampling to simulate temporal diversity. However, subsequent studies observed that task-relevant information typically concentrates in specific temporal regions and prompted a shift toward keyframe selection. Given that salient content is sparse in videos, video-level annotations often overlook fine-grained temporal granularity. To address this issue, Yang *et al.* [20] investigated the use of annotated timestamps to explicitly isolate important video segments. This shows that irrelevant frames shift decision boundaries and reduce model confidence.

Beyond video sampling, effective video modeling is critical for understanding. Masked autoencoders have emerged as a robust framework for learning representations from sparse data, specifically addressing the high spatiotemporal redundancy that characterizes video. For visual modeling, Wang *et al.* [21] introduced VideoMAE V2, which uses dual masking to scale self-supervised video representation learning. However, these pixel-reconstruction methods often prioritize high-frequency intensity over local structural details. To address this, Wei *et al.* [22] proposed reconstructing HOG features to encourage the model to focus on local structural information. In the audio domain, MAE has also proven effective. For instance, Huang *et al.* [23] encoded masked audio spectrogram patches, feeding only visible tokens into the encoder to achieve superior performance. Similarly, Gong *et al.* [24] adapted masked patch modeling to learn robust spectral representations from unlabeled data in AudioSet and LibriSpeech, significantly boosting recognition accuracy.

2.2 Specialized Video Emotion Recognition

VER aims to identify human emotional states from video data. Current research mainly focuses on three key areas: basic feature extraction, contextual integration, and temporal modeling.

Early studies relied heavily on handcrafted descriptors like LBP-TOP [25] and STLMBP [26]. However, deep learning has shifted the paradigm toward multimodal feature learning, demonstrating superior performance. For instance, Zhao *et al.* [27] introduced the dual-branch VAANet framework to process visual and audio modalities separately, allowing the model to extract more distinctive modality-specific features.

Beyond feature extraction, integrating contextual information greatly enhances recognition accuracy. Building on the VAANet backbone, Zhang *et al.* [28] proposed visual erasing and fusion modules to capture broader contextual cues extending beyond single keyframes. Furthermore, temporal reasoning is vital since emotions are dynamically evolving rather than static. To address this, Guo *et al.* [29] proposed StimuVAR, which leverages Multimodal Large Language Models (MLLMs) to reason about emotional stimuli. Similarly, Zhang *et al.* [30] improved performance by modeling the temporal distribution of emotion instead of static labels.

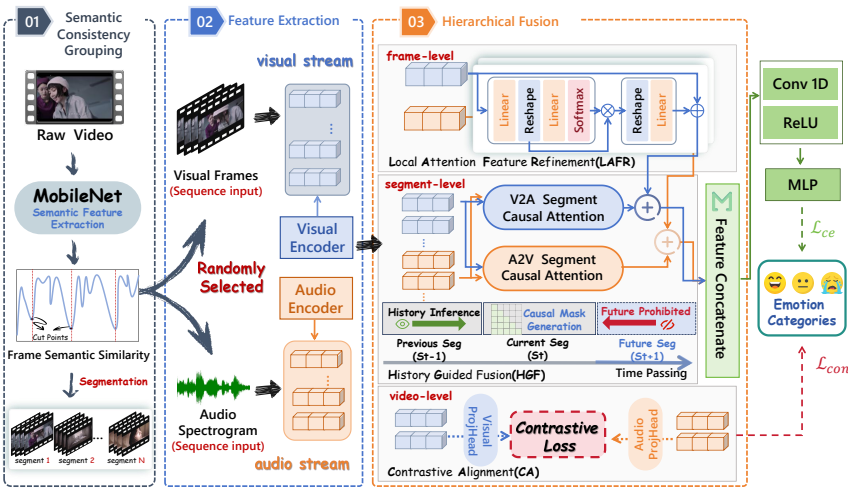

Figure 2: Overview of SCG-HF. The framework features SCG-based adaptive segmentation followed by hierarchical fusion at three levels: (i) frame-level local refinement, (ii) segment-level temporal modeling, and (iii) sample-level contrastive alignment.

3 Method

3.1 Overview

As illustrated in Fig. 2, the proposed framework models multimodal emotional dynamics in a hierarchical and temporally coherent manner. It first applies the **Semantic Consistency Grouping (SCG)** strategy to adaptively divide an input video into semantically homogeneous segments, providing structured units for subsequent analysis.

Based on these segments, the framework performs hierarchical fusion at three levels. At the **frame level**, **Local Attention Feature Refinement (LAFR)** is employed to enhance fine-grained emotional cues within each segment. At the **segment level**, the **History-Guided Fusion (HGF)** module captures temporal dependencies and emotional transitions across consecutive segments. At the **sample level**, contrastive alignment encourages agreement between audio and visual representations.

Finally, the hierarchically fused multimodal features are aggregated through weighted temporal pooling and fed into a classification head for emotion prediction. Through adaptive segmentation and multi-level fusion, SCG-HF effectively integrates local emotional cues, segment-wise temporal dynamics, and global cross-modal alignment, providing a unified framework for robust VER.

3.2 Semantic Consistency Grouping

Existing methods rely on fixed-length segmentation followed by random frame sampling [27, 28], which implicitly assumes that emotional content is uniformly distributed over time [31]. However, in practical scenarios, emotional expressions are often organized around semantic events or scene structures, resulting in strong intra-segment consistency and abrupt inter-segment transitions. Consequently, fixed segmentation may mix heterogeneous emotional states within the same segment or fail to capture short yet emotionally salient moments.

Motivated by these observations, we propose **Semantic Consistency Grouping (SCG)**, an adaptive segmentation strategy for emotion modeling. The detailed procedure of SCG is summarized in Algorithm 1. The core intuition of SCG is that frames sharing similar high-level visual semantics are likely to belong to the same emotional context, while significant drops in semantic similarity indicate potential scene changes or emotional shifts.

Specifically, given an input video, we uniformly sample frames at a fixed temporal interval and extract deep visual features using a pre-trained MobileNetV3-Small network [32]. Let $\mathcal{F} \in \mathbb{R}^{N \times D}$ denote the resulting feature matrix, where N is the number of sampled frames and D is the feature dimension. The feature vector of the i-th frame is represented as $\mathbf{f}_i$. To quantify semantic continuity

Algorithm 1 Semantic Consistency Grouping (SCG)

Require: Video V, target segments K, stride Δ, threshold $\tau = 0.05$
Ensure: Segments $\mathcal{G} = \{G_1, \ldots, G_K\}$

$\mathbf{f}_i \leftarrow$ MOBILENETV3-SMALL(SAMPLEFRAMES(V, Δ)); $s_i \leftarrow \cos(\mathbf{f}_i, \mathbf{f}_{i+1})$, $i{=}1, \ldots, N{-}1$
$\mathcal{M} \leftarrow \{i : 1{<}i{<}N{-}1,\ s_i{<}\min(s_{i-1}, s_{i+1})\}$; $\mathcal{M}^+ \leftarrow \{i \in \mathcal{M} : \delta_i{>}\tau\}$, $\delta_i{=}\min(s_{i-1}{-}s_i, s_{i+1}{-}s_i)$
$\mathcal{C} \leftarrow \mathcal{M}^+$ if $|\mathcal{M}^+| \geq K{-}1$ else $\mathcal{M}$; sort $\mathcal{C}$ in ascending s_i; $m \leftarrow \max(1, \lfloor (N{-}1)/K \rfloor)$; $\mathcal{B} \leftarrow \varnothing$
for $i \in \mathcal{C}$ **do**
  if $\forall b \in \mathcal{B} : |i{-}b| \geq m$ **then**
    $\mathcal{B} \leftarrow \mathcal{B} \cup \{i\}$
  end if
  if $|\mathcal{B}| = K - 1$ **then**
    break
  end if
end for{Greedy with fixed minimum interval}
$e \leftarrow K{-}1{-}|\mathcal{B}|$; $\mathcal{B} \leftarrow$ SORTUNIQUE$\big(\mathcal{B} \cup \{k \cdot \max(1, \lfloor \frac{N-1}{e+1} \rfloor)\}_{k=1}^{e}\big)$ {Uniform padding; may still leave $|\mathcal{B}|{<}K{-}1$}
if $|\mathcal{B}|{=}K{-}1$ **then**
  $t_j \leftarrow b_j/N \cdot T_V$ with $(b_0, \ldots, b_K) = (0, \mathcal{B}[0{:}K{-}1], N)$
  $\mathcal{G} \leftarrow$ FFMPEGSPLIT$(V, \{t_j\})$
else
  $\mathcal{G} \leftarrow \varnothing$
end if
if $|\mathcal{G}| \neq K$ **then**
  return UNIFORMSPLIT(V, K) {Unified fallback}
end if
return $\mathcal{G}$

between adjacent frames, we compute their cosine similarity as:

$$s_i = \frac{\mathbf{f}_i \cdot \mathbf{f}_{i+1}}{\|\mathbf{f}_i\| \cdot \|\mathbf{f}_{i+1}\|}, \quad i = 1, \ldots, N-1 \tag{1}$$

This produces a similarity sequence $\mathcal{S} = \{s_i\}$ that characterizes the temporal smoothness of visual semantics. High similarity values correspond to stable semantic content, whereas low similarity values indicate abrupt changes in scene structure or object composition, which are often correlated with emotional transitions.

In the similarity sequence $\mathcal{S}$, local minima are regarded as candidate semantic boundaries:

$$\mathcal{M} = \{i \mid s_i < s_{i-1} \text{ and } s_i < s_{i+1},\ 1 < i < N-1\} \tag{2}$$

Each candidate boundary reflects a point where semantic similarity drops relative to its temporal neighborhood. To further measure the strength of such transitions, we define a significance score

$$\delta_i = \min(s_{i-1} - s_i, s_{i+1} - s_i), \tag{3}$$

which captures the prominence of the similarity decrease and helps distinguish true semantic changes from minor fluctuations caused by noise or motion variation.

Based on δ_i, we separate strong semantic transitions from minor fluctuations: candidates with $\delta_i > \tau$ (with $\tau = 0.05$ in our experiments) form the high-significance set $\mathcal{M}^+$. If $|\mathcal{M}^+| \geq K - 1$, we set the candidate pool $\mathcal{C} = \mathcal{M}^+$; otherwise, we include every local minimum and set $\mathcal{C} = \mathcal{M}$.

To partition the video into K segments, we select $K - 1$ boundaries from $\mathcal{C}$ by a greedy procedure with a minimum-interval constraint. Within the pre-filtered pool $\mathcal{C}$, candidates are prioritized by their adjacent-frame similarity in *ascending* order: a low s_i directly indicates a large dissimilarity between frames i and $i + 1$, and therefore a more likely boundary position. Concretely, a candidate i is accepted as a boundary iff it satisfies

$$\forall b \in \mathcal{B} :\ |i - b| \ \geq\ m, \qquad m = \max\big(1, \lfloor (N-1)/K \rfloor\big), \tag{4}$$

where the constraint is vacuously satisfied when $\mathcal{B} = \varnothing$. The procedure terminates as soon as $|\mathcal{B}| = K - 1$ or the candidate pool is exhausted. During greedy selection, the minimum interval m

separates accepted internal boundaries; this constraint does not apply to video endpoints or uniformly added fallback boundaries.

When the candidate pool is sparse—e.g., in videos with limited visual variation—the greedy procedure may terminate with fewer than $K-1$ boundaries. In this case, we first supplement $\mathcal{B}$ with uniformly spaced frame indices $\{k \cdot \text{step}\}_{k=1}^{\text{extra}}$, where $\text{extra} = (K-1) - |\mathcal{B}|$ and $\text{step} = \max(1, \lfloor (N-1)/(\text{extra}+1) \rfloor)$, and discard any duplicates. If after this supplementation the boundary count is still smaller than $K-1$, or if the downstream FFmpeg-based splitting produces fewer than K valid segments (e.g., due to keyframe alignment errors), the entire video is re-segmented uniformly into K equal-length segments. This two-level fallback guarantees that the output always contains exactly K segments, while still preserving semantic transitions whenever they can be reliably detected.

By grouping frames according to semantic consistency rather than uniform temporal intervals, SCG generates semantically coherent segments that better align with underlying emotional contexts. These segments provide reliable inputs for subsequent hierarchical fusion modules, facilitating more effective modeling of both fine-grained emotional cues and long-term emotional evolution.

3.3 Fine-grained Hierarchical Fusion

We organize the input into a set of semantically consistent groups $\mathcal{G}$ containing K segments based on the detected boundaries. This segmentation induces a structured temporal representation in which each segment corresponds to a coherent semantic and emotional context. A parallel extraction strategy is adopted to capture modality-specific representations while preserving the distinct statistical properties of audio and visual signals.

For each segment, we sample a clip of $L = 16$ consecutive frames, with a randomly selected starting position during training and a centered clip during validation and testing. A 3D ResNet-101 encodes the visual clip into a 2048-dimensional feature vector, while a 2D ResNet-50 processes the log-Mel spectrogram to produce a 768-dimensional audio representation. We define the resulting feature sequence for a modality m as

$$\mathcal{F}^m = [\mathbf{f}_1^m, \mathbf{f}_2^m, \ldots, \mathbf{f}_K^m] \in \mathbb{R}^{K \times d_m}. \tag{5}$$

This independent extraction introduces an inductive bias that allows each modality to learn discriminative representations according to its own temporal and spectral characteristics. Non-local blocks [33] are further integrated to capture long-range dependencies and improve robustness to local variations.

To integrate multimodal information across different temporal resolutions, we design a fine-grained hierarchical fusion module composed of two complementary branches.

Frame-level Local Attention Feature Refinement. Conventional temporal aggregation tends to smooth out short-term fluctuations and may discard transient emotional cues. LAFR refines representations at the frame level before temporal abstraction, encouraging the model to preserve locally salient emotional patterns. Specifically, attention maps are learned to adaptively reweight audio and visual features, and residual connections are employed to maintain the original information flow while amplifying emotionally informative components. This design can be viewed as adaptive feature reweighting that enhances discriminability without introducing heavy temporal modeling.

Segment-level History-Guided Fusion. To capture emotional dynamics over time, HGF performs cross-modal attention at the segment level with an explicit temporal ordering constraint. The attention mechanism is restricted such that each segment can only attend to its past and current segments, preventing attention to future segments within HGF. This constraint aligns the learned representations with the natural temporal process of emotion perception, where the current emotional state should be inferred solely from historical context. As a result, the model learns temporally consistent cross-modal dependencies across segments. The outputs of LAFR and HGF are combined via element-wise addition to form the fused multimodal representation, integrating both fine-grained local cues and segment-level temporal reasoning. This restriction applies to HGF attention; SCG segmentation is computed from the complete input video.

Video-level Contrastive Alignment. While frame- and segment-level fusion capture local and temporal interactions, they do not explicitly enforce global semantic consistency between modalities. To this end, we introduce a video-level contrastive alignment objective that regularizes the learned representations in a shared latent space. Temporal features are first aggregated by average pooling and then projected into a common embedding space using modality-specific projection heads. The

cross-modal contrastive loss $\mathcal{L}_{con}$ treats visual and audio embeddings from the same video as positive pairs and embeddings from different videos as negative pairs. This objective can be interpreted as maximizing agreement between modalities at the sample level, encouraging modality-invariant emotion representations and reducing cross-modal representation drift across the dataset:

$$\mathcal{L}_{con} = -\frac{1}{2N}\sum_{i=1}^{N}\Big[\log\frac{\exp(\text{sim}(z_i^v, z_i^a)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(z_i^v, z_j^a)/\tau)} + \log\frac{\exp(\text{sim}(z_i^a, z_i^v)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(z_i^a, z_j^v)/\tau)}\Big], \quad (6)$$

Here, N denotes the mini-batch size, $\text{sim}(\cdot)$ denotes cosine similarity, and $\tau = 0.1$ is the contrastive temperature, distinct from the boundary-selection threshold used in SCG.

3.4 Emotion Classification and Loss Function

Emotional expressions in videos are often sparse and temporally localized, while long neutral intervals may dominate the sequence. Uniform temporal aggregation therefore introduces noise by treating all time steps equally and diluting short but informative emotional cues. To address this issue, we introduce a Temporal Aggregation Classification module that learns to emphasize emotionally salient moments during feature aggregation.

Specifically, we employ a temporal attention mechanism to assign adaptive importance weights to fused features based on their discriminative relevance. This design introduces an inductive bias that favors sparse yet informative temporal evidence, enabling the model to focus on critical emotional segments rather than background content. The weighted aggregation produces a compact video-level representation for emotion prediction. Given the SCG segments, we jointly train the recognition network using a joint optimization objective that combines emotion classification and cross-modal alignment. The total loss is defined as

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda\mathcal{L}_{con}, \quad (7)$$

where $\mathcal{L}_{ce}$ denotes the cross-entropy loss for emotion classification and $\mathcal{L}_{con}$ represents the contrastive alignment loss. The hyperparameter λ balances discriminative supervision and representation regularization. In our implementation, λ is set to 0.2.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluated our framework on three datasets:

- *VideoEmotion-8* [34]: 1,101 social media videos obtained from YouTube and Flickr in 8 categories (avg. 107 s), challenging for its emotional changes over time. Each video is assigned one of eight emotion categories following Plutchik's Wheel of Emotions theory [35].
- *Ekman-6* [36]: 1,637 videos (avg. 112 s) in 6 basic categories according to Ekman's emotion theory [37], a standard benchmark for long video analysis.
- *eMotions* [38]: 5,000 short-form videos sampled from the full dataset, collected from Douyin, Kuaishou and TikTok with six categories: excitation, fear, neutral, relaxation, sadness, and tension.

Evaluation Metrics. To comprehensively assess the performance of our framework, we employ two standard metrics: accuracy (Acc) and macro-averaged F1 (F1). Accuracy indicates the ratio of correctly predicted observations to the total observations. Macro-F1 is the unweighted mean of class-wise F1 scores, giving equal weight to each emotion category.

Implementation Details. For scene segmentation, we employ a lightweight MobileNetV3-Small network [32] to extract frame-level visual features for efficiency and robustness. All frames are resized to 256×256, center-cropped to 224×224, and normalized using ImageNet statistics. Cosine similarity between adjacent frames is computed to detect semantic boundaries, yielding 15 segmentation points and partitioning each video into 16 segments.

For backbone preprocessing, visual and audio streams are sampled independently. From each segment, 16 consecutive visual frames and 100 audio feature frames are extracted. Visual inputs are resized to 112×112. Following [27, 28], we adopt modality-specific encoders: a 3D ResNet pretrained

on Kinetics-400 for visual representation learning and a 2D ResNet pretrained on ImageNet for log-Mel-based audio feature extraction. Non-local blocks are used to enhance temporal modeling.

During training, data augmentation is applied to visual inputs and dropout is set to 0.35 to alleviate overfitting. The model is optimized using stochastic gradient descent with a batch size of 32 and a learning rate of 2×10^{-4}. We follow the standard dataset splits, using a 1:1 train-test ratio for Ekman-6 and a 2:1 ratio for VideoEmotion-8. For eMotions, we use a 5,000-video subset as a standalone benchmark and construct a fixed, stratified, non-overlapping split of 3,200/800/1,000 videos for training, validation, and testing. All experiments are implemented in PyTorch and conducted on two NVIDIA RTX 4090 GPUs.

4.2 Comparison with State-of-the-Art Methods

We compare SCG-HF with representative state-of-the-art methods on the VideoEmotion-8 and Ekman-6 datasets. The baselines are divided into two groups. The first group consists of general video representation learning methods such as AudioMAE [23], TCL [39], SSAST [24], MaskFeat [22], MILAN [40], MAEst [41], MGMAE [42], and VMAE [21]. The second group includes specialized VER approaches such as StimuVAR [29], MoLAN [43], CTEN [28], BGA [44], VAANet [27], and MART [30], which are specifically designed to model affective cues and temporal dynamics for emotion understanding. Together, these methods represent both generic multimodal pre-training techniques and task-oriented emotion modeling approaches, enabling a comprehensive comparison.

Table 1: Comparison with state-of-the-art methods on VideoEmotion-8 and Ekman-6 datasets.

Method	VideoEmotion-8		Ekman-6		Method	VideoEmotion-8		Ekman-6	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)		Acc (%)	F1 (%)	Acc (%)	F1 (%)
General Video Understanding					***Specialized Video Emotion Recognition***				
AudioMAE	41.39	40.37	45.24	44.38	StimuVAR	40.50	–	46.20	–
TCL	44.44	41.93	46.85	46.35	MoLAN*	48.01	45.00	49.48	48.93
SSAST	44.44	43.84	50.06	49.97	CTEN*	48.30	46.10	51.43	51.39
MaskFeat	46.39	44.76	<u>52.41</u>	<u>52.40</u>	BGA	49.20	–	51.22	–
MILAN	47.22	45.72	50.80	50.40	VAANet*	<u>50.14</u>	46.43	48.85	48.33
MAEst	48.89	<u>47.53</u>	49.44	48.66	MART*	49.43	46.94	51.56	51.33
MGMAE	49.28	45.88	51.98	52.36	**Ours**	**51.56**	**48.37**	**55.21**	**54.60**
VMAE	49.44	46.62	50.93	50.07	$\Delta SOTA$	↑1.42	↑0.84	↑2.80	↑2.20

Table 1 presents the quantitative comparison with state-of-the-art methods on the VideoEmotion-8 and Ekman-6 datasets. Methods marked with an asterisk (*) indicate results reproduced by us using the official implementations and recommended settings, while unmarked baselines use published results, including the general-method comparison in MART [30].

Overall, SCG-HF achieves the highest accuracy and F1 among the methods compared on both datasets. On VideoEmotion-8, SCG-HF achieves 51.56% accuracy and 48.37% F1, exceeding the best competing accuracy achieved by VAANet (50.14%) by 1.42 percentage points and the best competing F1 achieved by MAEst (47.53%) by 0.84 percentage points. On Ekman-6, SCG-HF achieves 55.21% accuracy and 54.60% F1, exceeding the corresponding best competing results by 2.80 percentage points and 2.20 percentage points, respectively. These results demonstrate that aligning temporal segmentation with semantic and emotional transitions leads to more discriminative representations. By integrating fine-grained local cues with segment-level temporal modeling, our framework effectively captures both short-term emotional signals and long-term emotional evolution.

In an additional three-seed evaluation, SCG-HF achieves the highest mean accuracy and macro-F1 among four reproduced baselines on VideoEmotion-8, Ekman-6, and eMotions. Full results are reported in Appendix A (Table 7).

4.3 Ablation Studies

Validation of SCG Segmentation Quality on VEATIC. SCG does not assume that semantic boundaries are emotion boundaries. Semantic consistency is used only as a weak temporal prior for emotion modeling. To fill this gap, we conduct quantitative analysis on VEATIC [45], a dataset equipped with dense frame-level valence-arousal affective annotations. We compare uniform segmentation, Random-20 random segmentation and our SCG from three perspectives: boundary emotion distance (± 1 s window), adjacent-segment emotion distance, and within-segment emotion consistency. As

reported in Table 2, SCG obtains superior performance on all three metrics, supporting the use of semantic consistency as a temporal prior for affective modeling.

Table 2: Quantitative comparison of segmentation strategies on VEATIC.

Metric	Uniform	Random-20	SCG
Boundary emotion distance	0.022	0.022	**0.023**
Adjacent-segment emotion distance	0.084	0.083	**0.090**
Within-segment emotion consistency	0.979	0.980	**0.982**

Component Effectiveness Analysis. To evaluate the contribution of each module, we perform ablation studies by separately adding individual components to the backbone CTEN model. The quantitative results are summarized in Table 3. Overall, each component consistently improves performance on both datasets, indicating that all proposed modules contribute positively to VER.

Table 3: Impact of individual components across VideoEmotion-8 and Ekman-6.

Method	VideoEmotion-8		Ekman-6		Method	VideoEmotion-8		Ekman-6	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)		Acc (%)	F1 (%)	Acc (%)	F1 (%)
BASE	48.30	46.10	51.43	51.39	**Ours (Full)**	**51.56**	**48.37**	**55.21**	**54.60**
w/ SCG	50.00$_{\uparrow 1.70}$	47.26$_{\uparrow 1.16}$	53.75$_{\uparrow 2.32}$	53.47$_{\uparrow 2.08}$	w/o SCG	51.14$_{\downarrow 0.42}$	46.51$_{\downarrow 1.86}$	52.50$_{\downarrow 2.71}$	52.56$_{\downarrow 2.04}$
w/ CA	49.69$_{\uparrow 1.39}$	46.21$_{\uparrow 0.11}$	54.56$_{\uparrow 3.13}$	54.29$_{\uparrow 2.90}$	w/o CA	47.50$_{\downarrow 4.06}$	44.72$_{\downarrow 3.65}$	51.30$_{\downarrow 3.91}$	51.39$_{\downarrow 3.21}$
w/ LAFR	49.15$_{\uparrow 0.85}$	47.57$_{\uparrow 1.47}$	53.26$_{\uparrow 1.83}$	53.10$_{\uparrow 1.71}$	w/o LAFR	49.06$_{\downarrow 2.50}$	43.54$_{\downarrow 4.83}$	53.91$_{\downarrow 1.30}$	53.85$_{\downarrow 0.75}$
w/ HGF	48.86$_{\uparrow 0.56}$	46.71$_{\uparrow 0.61}$	52.47$_{\uparrow 1.04}$	52.60$_{\uparrow 1.21}$	w/o HGF	48.75$_{\downarrow 2.81}$	43.61$_{\downarrow 4.76}$	54.04$_{\downarrow 1.17}$	53.41$_{\downarrow 1.19}$

Effectiveness of SCG. SCG consistently improves performance on both datasets, with Acc gains of up to 1.70 percentage points on VideoEmotion-8 and 2.32 percentage points on Ekman-6. This shows that adaptive segmentation aligned with semantic transitions provides more reliable temporal units for emotion modeling.

Effectiveness of LAFR (frame level). LAFR yields notable improvements, particularly in F1 (up to +1.47 percentage points on VideoEmotion-8), indicating that refining frame-level representations helps preserve subtle and transient emotional cues that would otherwise be diluted by temporal aggregation.

Effectiveness of HGF (segment level). HGF introduces performance gains, validating the importance of modeling temporal dependencies across segments. These results suggest that incorporating historical context leads to more temporally consistent emotional representations.

Effectiveness of CA (sample level). CA provides the largest improvement on Ekman-6, highlighting the benefit of enforcing cross-modal consistency. It confirms that aligning audio and visual representations in a shared embedding space is crucial for emotion prediction.

These results indicate that each component contributes at a distinct granularity, namely the frame, segment, or sample level, jointly yielding gains within a unified framework.

Analysis of Sampling Strategies. We compare SCG with uniform sampling across several representative architectures. As shown in Table 4, SCG improves accuracy for all three backbones and improves F1 for VAANet and CTEN. For MART, accuracy increases from 49.43% to 50.31%, while F1 decreases from 46.94% to 44.96%.

Table 4: Comparison of different sampling strategies used by different methods on VideoEmotion-8.

Sampling Method	VAANet		CTEN		MART	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
Uniform	<u>50.14</u>	46.43	<u>48.30</u>	46.10	<u>49.43</u>	46.94
TDC[46]	44.65	42.64	44.10	41.35	45.14	43.28
StimuVAR[29]	42.61	39.40	43.18	38.18	40.62	38.22
SCG	**50.29**	47.44	**50.00**	47.26	**50.31**	44.96

Under uniform sampling, our framework achieves 51.14% accuracy on VideoEmotion-8, outperforming the competing baselines even when they adopt SCG. By further integrating SCG, our model attains 51.56% accuracy and 48.37% F1, achieving the best overall performance. These results indicate that preserving semantic coherence during sampling is critical for modeling emotional dynamics. Moreover, the combination of SCG and hierarchical fusion enables more stable temporal representations, effectively maintaining emotional consistency throughout video analysis.

Effect of Feature Extractor. We evaluate the robustness of different backbones within the SCG module in Table 5. Lightweight backbones, such as MobileNetV3-Small, achieve performance on par with deeper models like ResNet-101. The result implies that the power of SCG primarily stems from modeling semantic transitions rather than from the depth of the underlying feature extractor.

Moreover, this robustness across backbones confirms that SCG provides an efficient mechanism for semantic grouping, making it suitable for practical applications under limited computational budgets.

Table 5: Impact of different backbones for the SCG on VideoEmotion-8.

Method	Param	Acc (%)	F1 (%)	FPS	Method	Param	Acc (%)	F1 (%)	FPS
MobileNetV3-Small	2.9M	**51.56**	**48.37**	5712.77	ResNet-18	11.7M	49.06	45.36	1596.78
MobileNetV3-Large	5.4M	<u>50.31</u>	<u>46.02</u>	1925.96	ResNet-50	25.6M	<u>50.31</u>	45.58	437.27
EfficientNet-B3	12.2M	49.69	45.73	492.00	ResNet-101	44.6M	50.00	45.55	282.76

Effect of Segment Number. We analyze model sensitivity to the number of segments K. Table 6 shows performance peaks at 16 segments among the tested settings.

Table 6: Sensitivity analysis of segment number.

Seg. Number	4	8	16	24	32
Acc / F1 (%)	47.50 / 43.72	49.06 / 45.44	**51.56 / 48.37**	49.69 / <u>46.02</u>	<u>50.00</u> / 43.33

When the segment number is too small, each segment tends to contain multiple heterogeneous scenes, which weakens the semantic coherence within a group and hinders the extraction of focused emotional cues. In contrast, using an excessively large number of segments leads to fragmented semantic units and introduces intra-segment noise, making the learned representations less stable and discriminative.

These observations suggest that the segment number controls a trade-off between semantic richness and feature purity. Setting $K = 16$ provides an effective balance between preserving sufficient temporal structure and maintaining coherent emotional contexts, yielding the optimal performance on VideoEmotion-8. This result also indicates that SCG benefits from a moderate level of temporal granularity rather than overly coarse or overly fine segmentation.

Sensitivity Analysis of Hyperparameters. To evaluate the robustness of SCG-HF with respect to hyperparameter choices, we conduct a sensitivity analysis on two key parameters: the temperature τ in Eq. (6) and the loss weight λ in Eq. (7). The performance is shown in Fig. 3. As shown in the figure, the model achieves its best performance when τ is around 0.1 on the VideoEmotion-8 dataset. A similar trend is observed for the loss weight λ, which reaches its optimum near 0.2. The responses vary by dataset and metric; the Ekman-6 curves do not indicate a shared optimum for both parameters.

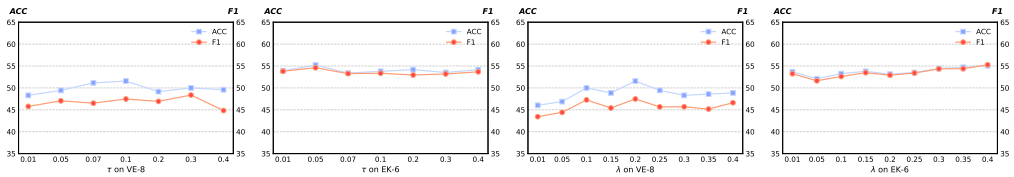

Figure 3: Sensitivity to τ (left two panels) and λ (right two panels) on VideoEmotion-8 and Ekman-6.

5 Conclusion

We propose SCG-HF, a framework that ensures semantic consistency via SCG-based grouping and a three-tier hierarchy: frame-level refinement, segment-level temporal modeling, and sample-level contrastive alignment. Experiments on VideoEmotion-8, Ekman-6 and eMotions demonstrate its effectiveness and generalization. The proposed method may support downstream applications such as content recommendation and accessibility. However, its use in affective surveillance or profiling could raise privacy concerns, and emotion predictions may be unreliable in sensitive contexts.

Indirect validation of semantic-emotion alignment. The main classification benchmarks provide only video-level labels. The VEATIC analysis offers complementary evidence from continuous affective annotations, but does not establish a one-to-one correspondence between semantic boundaries and ground-truth emotional transitions.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Key Research and Development Program of China (Grant No. 2024YFC3308403), in part by the National Natural Science Foundation of China (Grant Nos. T254100759, 62506269). This work was also supported by the WHU-Kingsoft Joint Lab. The supercomputing system at the Supercomputing Center of Wuhan University supported the numerical calculations in this paper. The authors declare no competing interests.

References

[1] Shangfei Wang and Qiang Ji. Video affective content analysis: A survey of state-of-the-art methods. *IEEE Transactions on Affective Computing*, 6(4):410–430, 2015. doi: 10.1109/TAFFC.2015.2432791.

[2] Yiyang Fang, Jian Liang, Wenke Huang, He Li, Kehua Su, and Mang Ye. Catch your emotion: Sharpening emotion perception in multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 16029–16039, 2025.

[3] C. Vinola and K. Vimaladevi. A survey on human emotion recognition approaches, databases and applications. *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, 14(2): 24–44, 2015. doi: 10.5565/rev/elcvia.795.

[4] Anvita Saxena, Ashish Khanna, and Deepak Gupta. Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems*, 2(1):53–79, 2020. doi: 10.33969/AIS.2020.21005.

[5] Yung-Hsuan Lai, Janek Ebbers, Yu-Chiang Frank Wang, François Germain, Michael Jeffrey Jones, and Moitreya Chatterjee. UWAV: Uncertainty-weighted weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13561–13570, 2025.

[6] Suhang Cai, Xiaohao Peng, Chong Wang, Xiaojie Cai, and Jiangbo Qian. GV-VAD: Exploring video generation for weakly-supervised video anomaly detection, 2025. arXiv preprint arXiv:2508.00312.

[7] Wenbo Xu, Junyan Wu, Wei Lu, Xiangyang Luo, and Qian Wang. A multimodal deviation perceiving framework for weakly-supervised temporal forgery localization, 2025. arXiv preprint arXiv:2507.16596.

[8] Wenbo Xu, Wei Lu, and Xiangyang Luo. Weakly supervised multimodal temporal forgery localization via multitask learning, 2025. arXiv preprint arXiv:2508.02179.

[9] Meriem Zerkouk, Miloud Mihoubi, and Belkacem Chikhaoui. Contextual attention-based multimodal fusion of LLM and CNN for sentiment analysis, 2025. arXiv preprint arXiv:2508.13196.

[10] Xianbing Zhao, Shengzun Yang, Buzhou Tang, and Ronghuan Jiang. Towards multimodal sentiment analysis via contrastive cross-modal retrieval augmentation and hierachical prompts, 2025. arXiv preprint arXiv:2508.07666.

[11] André Bazin. *What Is Cinema? Volume I*. University of California Press, Berkeley, CA, 2004.

[12] Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1960–1968, 2017. doi: 10.1109/CVPR.2017.212.

[13] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 169–176, 2011. doi: 10.1145/2070481.2070509.

[14] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, 2017. doi: 10.18653/v1/D17-1115.

[15] Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019. doi: 10.1109/access.2019.2916887.

[16] Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 25:4909–4921, 2023. doi: 10.1109/TMM.2022.3183830.

[17] Meng Li, Zhenfang Zhu, Kefeng Li, and Hongli Pei. Diversity and balance: Multimodal sentiment analysis using multimodal-prefixed and cross-modal attention. *IEEE Transactions on Affective Computing*, 16(1):250–263, 2025. doi: 10.1109/TAFFC.2024.3430045.

[18] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

[19] Peng Yan, Weina Chu, and Hongyuan Wu. A multimodal deep learning model for optimizing music emotion recognition through temporal and semantic feature integration. *Journal of Circuits, Systems and Computers*, 35(13):2650058, 2026. doi: 10.1142/S0218126626500581.

[20] Shuonan Yang, Tailin Chen, Rahul Singh, Jiangbei Yue, Jianbo Jiao, and Zeyu Fu. Revealing temporal label noise in multimodal hateful video classification, 2025. arXiv preprint arXiv:2508.04900.

[21] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023.

[22] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022.

[23] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *Advances in Neural Information Processing Systems*, volume 35, pages 28708–28720, 2022. doi: 10.52202/068431-2081.

[24] Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. SSAST: Self-supervised audio spectrogram transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36 (10):10699–10709, 2022. doi: 10.1609/aaai.v36i10.21315.

[25] Guoying Zhao and Matti Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007. doi: 10.1109/tpami.2007.1110.

[26] Xiaohua Huang, Qiuhai He, Xiaopeng Hong, Guoying Zhao, and Matti Pietikäinen. Improved spatiotemporal local monogenic binary pattern for emotion recognition in the wild. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 514–520, 2014. doi: 10.1145/2663204.2666278.

[27] Sicheng Zhao, Yunsheng Ma, Yang Gu, Jufeng Yang, Tengfei Xing, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. An end-to-end visual-audio attention network for emotion recognition in user-generated videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(1): 303–311, 2020. doi: 10.1609/aaai.v34i01.5364.

[28] Zhicheng Zhang, Lijuan Wang, and Jufeng Yang. Weakly supervised video emotion detection and prediction via cross-modal temporal erasing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18888–18897, 2023. doi: 10.1109/CVPR52729.2023.01811.

[29] Yuxiang Guo, Faizan Siddiqui, Yang Zhao, Rama Chellappa, and Shao-Yuan Lo. StimuVAR: Spatiotemporal stimuli-aware video affective reasoning with multimodal large language models. *International Journal of Computer Vision*, 133(9):6456–6472, 2025. doi: 10.1007/s11263-025-02495-3.

[30] Zhicheng Zhang, Pancheng Zhao, Eunil Park, and Jufeng Yang. MART: Masked affective representation learning via masked temporal distribution distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12830–12840, 2024.

[31] Kaicheng Fu, Changde Du, Jie Peng, Kunpeng Wang, Shuangchen Zhao, Xiaoyu Chen, and Huiguang He. EmoGrowth: Incremental multi-label emotion decoding with augmented emotional relation graph. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 17746–17763, 2025.

[32] Andrew Howard, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, Yukun Zhu, Ruoming Pang, Hartwig Adam, and Quoc Le. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019. doi: 10.1109/ICCV.2019.00140.

[33] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[34] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue. Predicting emotions in user-generated videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1):73–79, 2014. doi: 10.1609/aaai.v28i1.8724.

[35] Robert Plutchik. A general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3–33. Academic Press, 1980. ISBN 978-0-12-558701-3. doi: 10.1016/B978-0-12-558701-3.50007-7.

[36] Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li, and Leonid Sigal. Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. *IEEE Transactions on Affective Computing*, 9(2):255–270, 2018. doi: 10.1109/TAFFC.2016.2622690.

[37] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200, 1992. doi: 10.1080/02699939208411068.

[38] Xuecheng Wu, Heli Sun, Junxiao Xue, Jiayu Nie, Xiangyan Kong, Ruofan Zhai, Danlei Huang, and Liang He. Towards emotion analysis in short-form videos: A large-scale dataset and baseline. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, pages 1497–1506, 2025. doi: 10.1145/3731715.3733453.

[39] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022.

[40] Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and Sun-Yuan Kung. MILAN: Masked image pretraining on language assisted representation, 2022. arXiv preprint arXiv:2208.06049.

[41] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. In *Advances in Neural Information Processing Systems*, volume 35, pages 35946–35958, 2022. doi: 10.52202/068431-2605.

[42] Bingkun Huang, Zhiyu Zhao, Guozhen Zhang, Yu Qiao, and Limin Wang. MGMAE: Motion guided masking for video masked autoencoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13493–13504, 2023.

[43] Xingle Xu, Yongkang Liu, Dexian Cai, Shi Feng, Xiaocui Yang, Daling Wang, and Yifei Zhang. MoLAN: A unified modality-aware noise dynamic editing framework for multimodal sentiment analysis, 2025. arXiv preprint arXiv:2508.09145.

[44] Kaicheng Fu, Changde Du, Shengpei Wang, and Huiguang He. Improved video emotion recognition with alignment of CNN and human brain representations. *IEEE Transactions on Affective Computing*, 15(3):1026–1040, 2024. doi: 10.1109/TAFFC.2023.3316173.

[45] Zhihang Ren, Jefferson Ortega, Yifan Wang, Zhimin Chen, Yunhui Guo, Stella X. Yu, and David Whitney. VEATIC: Video-based emotion and affect tracking in context dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4455–4465, 2024. doi: 10.1109/WACV57701.2024.00441.

[46] Haoran Hao, Jiaming Han, Yiyuan Zhang, and Xiangyu Yue. Multimodal long video modeling based on temporal dynamic context, 2025. arXiv preprint arXiv:2504.10443.

A Robustness Across Random Seeds

To assess robustness, we conduct three runs using random seeds 1024, 2048, and 3072 for SCG-HF and four reproduced baselines (VAANet, CTEN, MoLAN, and MART) on VideoEmotion-8, Ekman-6, and eMotions in Table 7. All runs used fixed dataset splits, and the results are reported as the arithmetic mean ± sample standard deviation across all three runs. The remaining baseline results were taken from their original publications and were not included in the repeated-run analysis.

Table 7: Robustness Across Three Random Seeds.

Method	VideoEmotion-8		Ekman-6		eMotions	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
VAANet	48.42 ± 1.98	44.90 ± 1.87	51.34 ± 1.11	49.73 ± 2.90	55.43 ± 0.84	47.11 ± 2.27
CTEN	44.26 ± 3.68	42.30 ± 3.71	46.60 ± 1.49	44.86 ± 1.11	48.50 ± 2.08	39.80 ± 3.30
MoLAN	45.70 ± 2.47	41.98 ± 2.70	48.50 ± 0.70	47.14 ± 1.89	51.53 ± 3.05	37.78 ± 6.60
MART	47.51 ± 1.76	44.55 ± 3.33	50.48 ± 0.96	49.06 ± 1.84	55.43 ± 1.01	45.47 ± 3.72
SCG-HF	**50.04 ± 1.50**	**47.04 ± 1.38**	**54.49 ± 0.65**	**53.67 ± 0.83**	**56.83 ± 1.87**	**53.30 ± 2.99**

B Semantic vs. emotional consistency

High-level visual semantic boundaries are more likely to coincide with emotional shifts than arbitrary fixed-length boundaries, making SCG a better structural prior for temporal emotion modeling.

Since VideoEmotion-8 provides only video-level labels, segment-level emotion shifts cannot be verified against ground truth. Using a pre-trained model, we observed changes in predicted emotion labels between adjacent segments at rates of 43.77% for uniform segmentation and 55.34% for SCG-based segmentation. For a video with ground-truth label 1, uniform segmentation yields the label sequence 5, 5, 4, 4, 5, 5, 6, 4, 5, 1, 6, 4, 5, 6, 1, 1 (change rate: 11/15), whereas SCG yields 5, 6, 4, 1, 5, 4, 4, 1, 6, 5, 4, 4, 5, 3, 5, 1 (change rate: 13/15).

Table 8: Comparison of uniform and SCG segmentation: predicted emotions and segment lengths (ground-truth label: Ant.)

Seg.	Uniform frame range	Uni. pred.	SCG frame range	SCG pred.	SCG length
1	1–397	Sad.	1–10	Sad.	10
2	398–794	Sad.	11–720	Sur.	710
3	795–1191	Joy	721–1220	Joy	500
4	1192–1588	Joy	1221–1585	Ant.	365
5	1589–1985	Sad.	1586–1690	Sad.	105
6	1986–2382	Sad.	1691–2225	Joy	535
7	2383–2779	Sur.	2226–2780	Joy	555
8	2780–3176	Joy	2781–3170	Ant.	390
9	3177–3573	Sad.	3171–3370	Sur.	200
10	3574–3970	Ant.	3371–3815	Sad.	445
11	3971–4367	Sur.	3816–4244	Joy	429
12	4368–4764	Joy	4245–4754	Joy	510
13	4765–5161	Sad.	4755–5174	Sad.	420
14	5162–5558	Sur.	5175–5609	Fea.	435
15	5559–5955	Ant.	5610–6159	Sad.	550
16	5956–6352	Ant.	6160–6358	Ant.	199

As shown in Table 8, we list the predicted emotion for each segment under both uniform and SCG segmentation, along with the number of frames per segment (uniform segmentation yields roughly 397 frames per segment). Both strategies use the same sampling budget; the listed frame ranges and segment lengths describe temporal coverage, not the number of sampled encoder inputs, and the ranges need not coincide. These results show a higher rate of changes in predicted segment labels under SCG. This is a descriptive observation rather than direct evidence of ground-truth emotional boundaries.

Regarding cross-domain stability, SCG is built on cosine similarity between features from a MobileNet backbone, without requiring dataset-specific boundary annotations. Our two primary benchmarks already differ substantially: VideoEmotion-8 consists of YouTube and Flickr videos across 8 emotion categories, while Ekman-6 covers social media videos across 6 categories. SCG yields consistent improvements on both, supporting its applicability across the evaluated datasets.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's contributions: SCG adaptive segmentation, three-level hierarchical fusion (LAFR, HGF, CA), and benchmark comparisons on VideoEmotion-8 and Ekman-6, with an additional three-seed evaluation including eMotions in Appendix A. These claims are supported by the method description in Sec. 3 and experimental results in Sec. 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the limitations in Sec. 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present theoretical results; it proposes an empirical method and architecture.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Sec. 4.1 provides dataset details, data splits, hyperparameters (learning rate, batch size, dropout, $\lambda = 0.2$), optimizer (SGD), feature extractors, and preprocessing steps sufficient for a trained researcher to reproduce the main results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We use three publicly available benchmarks, VideoEmotion-8, Ekman-6 and eMotions, all of which are widely adopted in the video emotion recognition community and accessible through their original releases (cited in Section 4.1). Detailed implementation settings, including backbone architectures, optimizer configurations, learning rate, batch size, data splits, and all key hyperparameters, are reported in Section 4.1 to facilitate reproduction. The full source code will be made publicly available as soon as possible.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sec. 4.1 specifies all training and test details: frame sampling (16 frames per segment), input sizes (224×224 for SCG, 112×112 for backbones), backbone architectures (MobileNetV3-Small, 3D ResNet-101, 2D ResNet-50), augmentation, dropout=0.35, SGD with a learning rate of 2×10^{-4}, batch size 32, and train/test splits (1:1 for Ekman-6, 2:1 for VideoEmotion-8). In addition, we evaluate SCG-HF on the 5,000-video eMotions [38] subset, treating it as a standalone dataset. We construct a fixed, stratified, non-overlapping split of 3,200/800/1,000 videos for training, validation, and testing. The test set is used only for final evaluation, and SCG-HF and four reproduced baselines are evaluated with three random seeds in Appendix A. Table 7 reports the mean $\pm$ sample standard deviation of this evaluation. SCG-HF has the highest mean accuracy and macro-F1 among these methods; these summaries do not establish a win for every individual seed.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 7 reports the arithmetic mean and sample standard deviation across three random seeds with fixed dataset splits. The standard deviations describe variability across runs; no formal statistical significance test is claimed.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sec. 4.1 states: "All experiments are implemented in PyTorch and conducted on two NVIDIA RTX 4090 GPUs." This provides the type and count of compute workers.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research does not involve human subjects, sensitive data, or ethical violations.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Both positive applications (e.g., content recommendation, accessibility) and potential negative impacts (e.g., affective surveillance, profiling, unreliable predictions in sensitive contexts) are discussed at the end of Sec. 5.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models (e.g., generative models) or scraped datasets; thus no safeguards for misuse are needed.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use existing public datasets (VideoEmotion-8, Ekman-6 and eMotions) and pretrained models (MobileNetV3-Small, 3D ResNet, etc.), and properly credit their original creators by citing the corresponding papers and official sources. We follow the publicly stated licenses and terms of use for these assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce new datasets, code packages, or model checkpoints; therefore no asset documentation is required.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human subjects; all data come from existing public benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects research is conducted; thus IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method does not use large language models as an important component; LLMs appear only in a baseline (StimuVAR) for comparison.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.