OSVAR: Oddly Satisfying Video Affective Reasoning with Psychophysics-Driven multi-modal learning

Anonymous ACL submission

Abstract

Oddly Satisfying Videos (OSVs) elicit psychological comfort through precise audio-visual stimuli. However, existing MLLMs predominantly focus on high-level semantic recognition, often overlooking fine-grained sensory dynamics and underlying affective mechanisms. To bridge this gap, we present **OSVAR**, a psychophysics-driven multimodal framework designed for Oddly Satisfying Video Affective Reasoning. The proposed OSVAR injects domain-specific sensory priors into multimodal models through three distinct mechanisms: (1) **Visual Haptics**: which models motion predictability via optical flow intensity to capture the “visual order” inherent in satisfying content; (2) **Acoustic Purity**: which aligns features with ASMR triggers via constraints on dynamic range, non-speech probability, and timbre consistency; and (3) **Synesthesia**: which enforces cross-modal congruence via a fine-grained synchronization loss. Extensive experimental results on the constructed dataset demonstrate that OSVAR significantly outperforms state-of-the-art baselines in multiple affective reasoning tasks, offering a novel direction for sensory-aware multimodal understanding.

1 Introduction

The genre of "Oddly Satisfying Videos" (OSVs) has evolved from a niche subculture of the 2010s (Faramarzi, 2018) into a global phenomenon. Characterized by repetitive, rhythmic, and precision-executed actions like kinetic sand slicing or paint mixing, these videos are consumed not for instructional value, but for intrinsic sensory gratification (Watson, 2021). While seemingly recreational, the appeal of OSVs is rooted in robust neuroscientific principles. Neurologically, the Mirror Neuron System (MNS) facilitates embodied simulation, creating visual haptics, where viewers vicariously experience texture and resistance solely through visual stimuli (Fadiga et al., 1995; Maeda et al., 2002; Patuzzo et al., 2003). Psychologically, the hyper-predictability of these videos aligns with the brain’s regulation of uncertainty, modulating neural reward systems (Vanhersecke, 2021). Physiologically, this sensory integration often culminates in Autonomous Sensory Meridian Response (ASMR), a phenomenon driven by *synesthesia*, defined as the strict cross-modal congruence between visual action and auditory feedback (Sumpf et al., 2015).

Despite this profound psychophysical basis, current Multimodal Large Language Models (MLLMs) faces a critical bottleneck of "semantic-sensory gap" (Guo et al., 2025). Existing MLLMs (Tong et al., 2022; Li et al., 2024b; Jin et al., 2024; Ye et al., 2024) prioritize high-level semantic recognition (e.g., identifying ”a person cutting soap”) while neglecting the fine-grained physical dynamics (e.g., flow smoothness and textural brittleness) that drive human affective responses. Furthermore, the oddly satisfying experience is intrinsically multimodal on the temporal isomorphism between sight and sound, and conventional audio-visual frameworks (Sardari et al., 2024; Mo and Morgado, 2023; Vilaca et al., 2025) typically treat modalities as loosely coupled streams, failing to verify *sensory congruence*. Without modeling this micro-level synchronization, models cannot distinguish between a genuinely satisfying video and one with asynchronous or disjointed audio.

Furthermore, most public datasets (Yang et al., 2003; Saikh et al., 2022; Rawal et al., 2024; Fu et al., 2025; Swetha et al., 2025) focus on factual captioning or standard QA, limiting models to generating superficial descriptions (e.g., "This is satisfying") without providing grounded reasoning based on psychophysical priors. This semantic-sensory gap prevents MLLMs from achieving human-like empathy in understanding sensory comfort. Therefore, constructing a high-quality, inter-

pretable dataset is pivotal to bridge the affective gap between low-level sensory cues and high-level psychological perception. Such detailed annotations are essential for empowering models to understand *why* specific visual-auditory patterns effectively alleviate stress and induce satisfaction.

To bridge this semantic-sensory gap, we propose **OSVAR** (**O**ddly **S**atisfying **V**ideo **A**ffective **R**easoning), a novel multimodal framework driven by psychophysical priors. Unlike generic MLLMs, OSVAR explicitly models the predictability and congruence inherent in oddly satisfying content. Specifically, we first propose a motion predictability module utilizing optical flow intensity to capture the "visual order" (e.g., smoothness and rhythm) of physical dynamics. For the auditory modality, we impose strict auditory texture constraints, including low dynamic range, non-speech probability, and timbre consistency, to align the encoder with the acoustic purity of ASMR triggers. Crucially, to computationally model synesthesia, we design a fine-grained audio-visual synchronization loss, forcing the model to learn the precise temporal alignment between visual action and auditory content.

Furthermore, addressing the scarcity of reasoning data, we construct an automated Multi-Expert Data Pipeline. Leveraging specialized vision and audio expert to excavate fine-grained semantic details, we curate a large-scale dataset comprising 12,000 oddly satisfying videos, richly annotated with psychophysical stress-relieving attributes. Extensive experiments, encompassing both classification benchmarks and interpretable evaluations, validate the effectiveness of our proposed method.

The main contributions of this paper are summarized as follows:

- We propose **OSVAR**, the first MLLM framework to ground affective reasoning in psychophysical priors, specifically designed for affective reasoning in Oddly Satisfying Videos.
- We construct a comprehensive OSV Dataset (**OSVD**) with 10 fine-grained categories via an automated expert-model pipeline, addressing the data scarcity in affective reasoning.
- Extensive experiments demonstrate the effectiveness of OSVAR, which achieves state-of-the-art performance in binary satisfaction judgment, fine-grained category classification, and the rationality of affective explanations.

2 Related work

2.1 MLLMs and Audio-Visual Synesthetic Learning

Recent advancements in MLLMs, such as Qwen-VL (Xu et al., 2025), LLaVA-Next-Video (Li et al., 2024a), and GPT-4V (Yang et al., 2023), have demonstrated their remarkable capabilities in general video understanding. Although these models excel at semantic analysis, a significant semantic-sensory gap remains: while existing models can describe *what* is happening (e.g., "cutting soap"), they struggle to perceive *how* it feels (e.g., the tactile smoothness or auditory crispness).

In the realm of audio-visual synesthetic learning, traditional approaches like ImageBind (Girdhar et al., 2023) and AudioCLIP (Guzhov et al., 2022) have successfully aligned modalities into a shared embedding space. Yet, these methods typically focus on coarse-grained semantic correspondence (e.g., matching a dog image to a bark). They often overlook the fine-grained temporal synchronization and textural congruence required for affective reasoning. Recent method like Synchformer (Iashin et al., 2024) have begun to focus audio-visual synchronization, trying to segment videos in pieces to align video and audio precisely. In oddly satisfying videos, the sensation of relief relies on Synesthesia—the precise, microsecond-level alignment between visual kinematics and auditory feedback, which is crucial for oddly satisfying feature extracting.

2.2 Video Affective Reasoning

Video affective reasoning methods can be categorized into two classes: analyzing the emotions of characters in a video and predicting viewers' affective response to a video. The former relies on facial expression recognition, which have been widely studied (Srivastava et al., 2023; Lee et al., 2019; Kosti et al., 2019; Yang et al., 2024). We target the latter, which is more challenging, since it requires not only video content understanding but also commonsense knowledge of human reaction.

A recent work named StimuVAR (Guo et al., 2025) have begun to explore user-side induced emotions, they use event-driven frame sampling and emotion-triggered tube selection strategies to capture the key frame and tube for affective stimuli. However, StimuVAR suffers from limited generalization, particularly for OSVs, where satisfaction emerges not from abrupt stimuli, but from conti-

nuity, rhythm, and predictability. Existing frameworks lack the capability to model this "process-oriented" affect, pushing us to design a model that can interpret the soothing dynamics of continuous visual-audio flows rather than just detecting sudden visual stimulus.

2.3 Scientific Analysis of Oddly Satisfying Videos

The psychological appeal of oddly satisfying videos (OSVs) is not arbitrary. it is deeply rooted in specific neuroscientific mechanisms. We analyze three core frameworks that explain this phenomenon.

Mirror neurons and visual haptics. Research indicates that observing physical actions triggers the Mirror Neuron System (MNS), creating a "visual haptic" simulation where viewers mentally feel the texture and resistance of materials (Fadiga et al., 1995; Maeda et al., 2002; Patuzzo et al., 2003). This justifies our incorporation of optical flow features, enabling the model to perceive the physical dynamics essential for this embodied simulation.

Predictive coding and dopamine loops. According to Predictive Coding Theory (Friston, 2012), the brain seeks to minimize sensory uncertainty. OSVs provide hyper-predictable visual patterns (e.g., perfect slicing). The successful prediction of these smooth motions triggers reward responses (dopamine release). We map this predictability prior to our motion level prediction module, training the model to recognize ordered motion.

ASMR and synesthesia. ASMR relies on specific acoustic properties—typically low dynamic range and non-speech textures (Yusaira and Bennett, 2021). Furthermore, the immersion depends on Synesthesia, in other words cross-modal congruence (Poerio et al., 2018). We operationalize these biological findings into our loss landscape: auditory texture is modeled via explicit audio constraints, while synesthesia is enforced through a fine-grained InfoNCE loss, ensuring the model learns the precise harmony between video and audio.

3 Methodology

3.1 Overview

We present **OSVAR**, a unified multimodal framework designed to decode the psychophysical mechanisms underlying sensory satisfaction. As illustrated in Figure 1, we introduce a specialized input formulation and a psychophysics-driven training paradigm to transcend generic semantic understanding and capture the subtle sensory dynamics of satisfaction.

Formally, the framework operates on a multimodal input triplet denoted as $I = \{V_{rgb}, V_{flow}, A_{raw}\}$. Here, V_{rgb} represents the sequence of raw RGB frames, which provides essential appearance information such as object identity and material texture. V_{flow} denotes the optical flow sequence extracted from adjacent frames to explicitly capture the fluidity and rhythm of physical dynamics. A_{raw} denotes the raw audio waveform in form of Mel-spectrogram to encode the fine-grained auditory textures essential for ASMR-like triggers.

The core training objective is to inject psychophysical priors—specifically *visual haptics*, *acoustic purity*, and *synesthesia*—into the latent embedding space via three auxiliary modules: the *visual perception enhancement module*, the *audio conditional constraint module*, and the *audio-visual synesthesia module*. To translate these sensory features into linguistic reasoning, we employ a structured fine-tuning strategy. Specifically, we design the prompt to enforce a hierarchical reasoning chain, requiring the model to sequentially output a binary satisfaction judgment, a fine-grained category classification (e.g., *kinetic sand*, *precision slicing*), and finally, a grounded affective explanation. This structured output format ensures that the generated text is not merely descriptive but is logically grounded in the specific sensory cues extracted by our auxiliary modules.

3.2 Visual Perception Enhancement Module: Modeling Motion Predictability

The visual allure of oddly satisfying videos stems not merely from static objects, but from the visual order of their movement, such as the laminar flow of fluids or the rhythmic slicing of kinetic sand. Standard RGB encoders, however, often struggle to decouple these fine-grained physical dynamics from background semantics, leading to a deficiency in visual haptics. To bridge this gap, we introduce a dual-stream mechanism centered on optical flow to model motion predictability.

We employ a visual encoder to extract feature representations for both RGB frames (F_{rgb}) and optical flow (F_{flow}). To fuse these modalities ef-

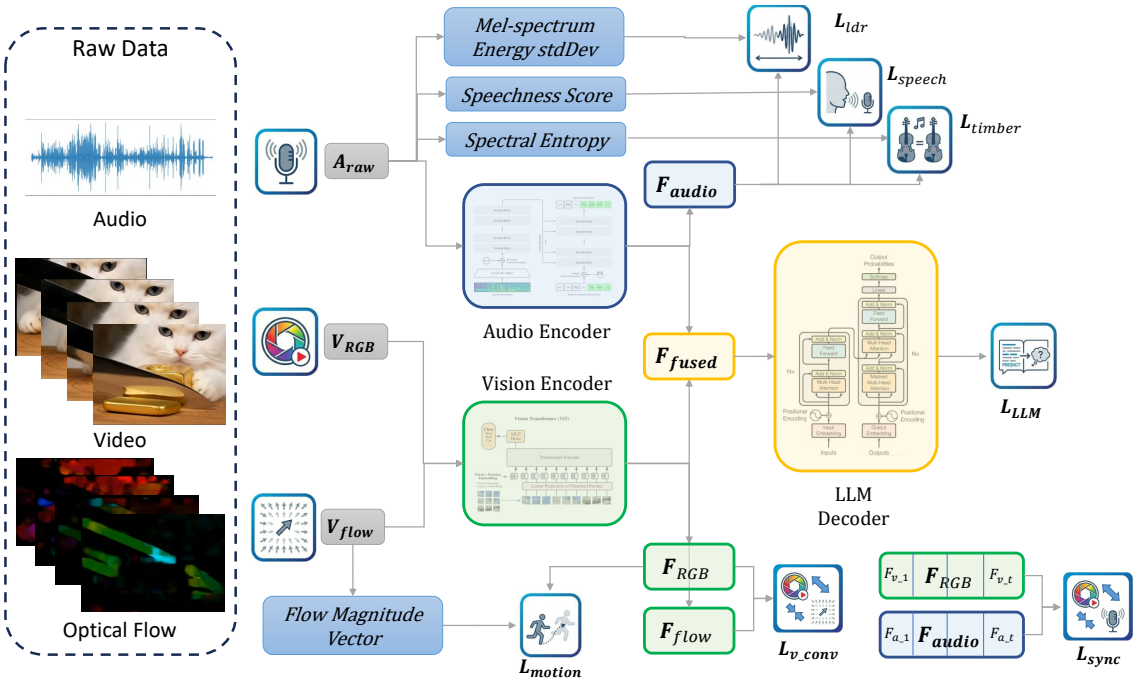

Figure 1: **The architecture of OSVAR.** The **Visual Perception Enhancement Module** leverages optical flow and motion prediction to explicitly model the fluidity and visual haptics of physical dynamics, while the **Audio Conditional Constraint Module** enforces acoustic purity through low dynamic range, non-speech, and timbre regularization to encode sensory texture. Crucially, the **Audio-Visual Synesthesia Module** captures the micro-level temporal congruence between visual action and auditory feedback. Finally, we employ prompt fine-tuning to translate these psychophysical priors into linguistic reasoning.

fectively, we first implement a contrastive alignment strategy. By applying a contrastive loss, we enforce the semantic alignment between the static appearance of an object (e.g., a knife) and its corresponding kinematic tendencies (e.g., a smooth downward motion). The loss is formulated as:

$$\mathcal{L}_{conv} = -\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(f_{rgb}^i, f_{flow}^i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(f_{rgb}^i, f_{flow}^j)/\tau)}, \tag{1}$$

where N denotes the batch size and τ is the temperature parameter.

Furthermore, to explicitly encode motion predictability, we introduce a motion prediction head. This MLP-based module takes the RGB features F_{rgb} as input and is trained to regress the average optical flow magnitude vector of the scene. By minimizing the regression loss

$$\mathcal{L}_{motion} = ||\text{MLP}(F_{rgb}) - \text{Magnitude}(V_{flow})||_2^2, \tag{2}$$

we compel the visual encoder to infer the underlying physical laws of motion, such as smoothness and velocity, directly from visual cues, thereby simulating the brain's predictive coding mechanism.

3.3 Audio Conditional Constraint Module: Encoding Auditory Texture

In the auditory domain, the sensation of relief is characterized by acoustic purity. Unlike generic audio recognition which focuses on semantic events (e.g., "a dog barking"), affective reasoning requires discerning the texture and stability of sound. To align the audio encoder with these specific physical characteristics, we impose a set of three distinct constraints, formally defined as follows:

1. Low Dynamic Range Constraint. Psychologically, satisfying sounds are typically consistent, devoid of abrupt, jarring noises (e.g., screams or explosions) that induce a startle response. We quantify this attribute using spectrum energy dynamic range, calculated as the standard deviation of the frame-level Mel-spectrogram energy, denoted as σ_E. To enforce the encoder to capture this temporal stability, we employ a predictor head $\mathcal{H}_{ldr}$ to regress this statistic:

$$\mathcal{L}_{ldr} = ||\mathcal{H}_{ldr}(F_{audio}) - \sigma_E||_2^2, \tag{3}$$

where F_{audio} represents the global audio feature embedding.

2. Non-speech probability Constraint. Deep im-

mersion relies on physical sounds (e.g., crunching, flowing). Clear human speech often activates language processing regions (e.g., Broca's area (Flinker et al., 2015)), breaking the sensory immersion. To penalize semantic interference, we leverage a teacher-student distillation approach. We utilize the pre-trained MLLM(in zero-shot mode) to generate a speech probability score $S_{teacher} \in [0, 1]$ for each clip. A lightweight predictor $\mathcal{H}_{speech}$ is then optimized to approximate this teacher distribution, compelling the encoder to prioritize physical textures over verbal semantics:

$$\mathcal{L}_{speech} = ||\mathcal{H}_{speech}(F_{audio}) - S_{teacher}||_2^2. \quad (4)$$

3. Timbre Consistency Constraint. High-quality ASMR triggers, such as white noise or rain sounds, exhibit a stable spectral structure over time. To capture this, we utilize spectral entropy as a proxy for timbre consistency. We compute the normalized spectral entropy $H = -\sum p_i \log p_i$ on the frequency spectrum. The model is trained to infer this structural complexity via the regression loss:

$$\mathcal{L}_{timbre} = ||\mathcal{H}_{timbre}(F_{audio}) - H||_2^2. \quad (5)$$

By minimizing the summation of these constraints, i.e., $\mathcal{L}_{audio} = \mathcal{L}_{ldr} + \mathcal{L}_{speech} + \mathcal{L}_{timbre}$, OSVAR explicitly learns to encode the acoustic properties essential for psychological relief.

3.4 Audio-Visual Synesthesia Module

The most critical component of our framework addresses synesthesia—the psychological phenomenon where satisfaction arises from the precise temporal congruence between visual action and auditory feedback. Conventional video-text alignment methods typically aggregate features globally, thereby ignoring the micro-level synchronization required for satisfying sensation.

To capture this fine-grained interplay, we employ an InfoNCE-based synesthetic alignment mechanism. We segment the input video and audio streams into T synchronized segments. For each temporal segment t, the visual feature v_t and the corresponding audio feature a_t constitute a positive pair, while mismatched pairs serve as negative pairs. The synchronization loss is defined as:

$$\mathcal{L}_{sync} = -\sum_{t=1}^{T} \log \frac{\exp(\text{sim}(v_t, a_t)/\tau)}{\sum_{j \neq t} \exp(\text{sim}(v_t, a_j)/\tau)}. \quad (6)$$

Minimizing $\mathcal{L}_{sync}$ forces the model to discriminate strictly synchronized signals from asynchronous ones. Ideally, this models the sensory congruence, ensuring that a specific auditory texture (e.g., a "crunch") is validated only if it perfectly aligns temporally with the corresponding visual event (e.g., a "fracture"), thus mimicking the human neural response to synesthetic stimuli.

3.5 Optimization Objective

Consequently, the overall training of OSVAR is formulated as a multi-modal learning problem. The total objective function is a weighted sum of the primary fine-tuning task and the proposed sensory-aware auxiliary tasks:

$$\mathcal{L}_{total} = \mathcal{L}_{LLM} + \lambda_1 \mathcal{L}_{motion} + \lambda_2 \mathcal{L}_{conv} + \lambda_3(\mathcal{L}_{ldr} + \mathcal{L}_{speech} + \mathcal{L}_{timbre}) + \lambda_4 \mathcal{L}_{sync} \quad (7)$$

where $\mathcal{L}_{LLM}$ represents the standard auto-regressive cross-entropy loss of baseline language models, and $\lambda_{1..4}$ are hyperparameters balancing the contribution of each psychophysical constraint. This composite objective ensures that OSVAR simultaneously acquires high-level semantic reasoning capabilities and fine-grained, affect-driven sensory perception.

4 Experiments

4.1 Experimental Settings

Dataset. To facilitate rigorous evaluation, we construct the **Oddly Satisfying Video Dataset (OSVD)**, a large-scale multimodal corpus generated via an automated curation pipeline. The dataset comprises 12,000 video-text pairs, meticulously annotated with a hierarchical label structure: (1) A *binary label* indicating satisfaction status; (2) A *fine-grained label* covering 11 specific categories (e.g., *Kinetic Sand*, *Soap Cutting*, *Unknown*); and (3) An *affective explanation* that provides detailed reasoning for the satisfaction mechanism. To construct these explanations, we employ a multi-stage pipeline: rich visual and acoustic semantic details are first independently extracted by the expert models Molmo2-8B (Deitke et al., 2024) and Music-Flamingo (Ghosh et al., 2025), respectively. Subsequently, Gemini (Gemini Team and Google, 2024) serves as a reasoning core to synthesize these multi-modal insights, generating a holistic interpretation of the video's stress-relieving effects. We randomly

partition the dataset into training, validation, and testing sets with a ratio of 4:1:1.

Baselines. We benchmark OSVAR against a comprehensive suite of state-of-the-art models, categorized into three groups:

- **Open-source MLLMs:** We include *Video-LLaMA 2* (Zhang et al., 2023), *LLaVA-NeXT-Video* (Li et al., 2024a), and the recent *Molmo2-8B* (Deitke et al., 2024) to evaluate general video understanding capabilities.
- **Foundation Base:** We utilize the *Qwen2.5-Omni* (Jin Xu, 2025) model (in both zero-shot and few-shot settings) as a baseline for our backbone architecture.
- **Proprietary SOTA:** We also compare against closed-source frontiers, *Doubao-seed1.6* (Volcano Engine, 2024) to assess the gap with commercial-grade systems.

Evaluation Metrics. Our evaluation protocol is three-fold: (1) **Classification Metrics:** We report accuracy for both binary (satisfying/not) and fine-grained (11-class) tasks. (2) **Generation Metrics:** We calculate the *semantic similarity* (using embeddings) between the generated explanations and the expert-annotated ground truth. (3) **LLM-as-a-Judge:** Following recent trends (Li et al., 2025), we employ Gemini-3 as an impartial judge to score generated explanations (0-5 scale) across three dimensions: *rationality* (logical coherence), *sensory detail* (richness of texture descriptions), and *AV correlation* (audio-visual alignment).

4.2 Main Results

Table 1 presents a comprehensive quantitative evaluation across binary satisfaction judgment, fine-grained classification, and semantic explanation quality.

Initially, a distinct performance disparity is observed between standard open-source MLLMs and our proposed framework. As shown in the first block, general-purpose models such as Video-LLaMA 2 and Molmo2-8B exhibit limited capability in fine-grained reasoning, achieving only 22.7% and 28.2% accuracy on the 11-class task, respectively. This suboptimal performance empirically corroborates the existence of the *semantic-sensory gap*: without explicit guidance, generic video encoders struggle to perceive the micro-level physical dynamics (e.g., flow viscosity) and audio-visual synchronization essential for distinguishing specific satisfaction categories. In stark contrast, OSVAR (Full Model) significantly elevates this baseline, demonstrating that psychophysical priors are indispensable for decoding affective content.

More strikingly, OSVAR challenges the prevailing assumption that model scale is the sole determinant of performance. In the 11-class fine-grained classification, our model achieves a state-of-the-art accuracy of 66.7%, surpassing not only the strongest open-source baselines but also the proprietary giant, doubao-seed1.6 by 3.9%. This result is particularly profound as it suggests that for specialized psychophysical tasks, a compact model equipped with domain-specific sensory mechanisms can outperform significantly larger foundation models that rely on generic pre-training.

While the proprietary doubao-seed1.6 maintains a slight edge in semantic similarity (0.76), OSVAR demonstrates remarkable competitiveness. It successfully outperforms the commercial Doubao-seed1.6 model in binary classification by 1.9%. This indicates that while foundation models excel at general semantics, OSVAR provides a highly efficient and potent alternative for domain-specific affective reasoning, achieving above-SOTA performance with a fraction of the computational resources.

4.3 Comparison with Specialized VAR Methods

To further delineate the efficacy of our motion-centric paradigm, we benchmark OSVAR against StimuVAR (Guo et al., 2025), a specialized Video Affective Reasoning (VAR) framework designed for arousal detection via key-frame extraction. As detailed in Table 2, the results yield two pivotal insights regarding the nature of satisfaction reasoning.

First, OSVAR (Full Model) significantly outperforms the standard StimuVAR baseline, achieving a substantial gain of 23.8% in fine-grained accuracy. This performance disparity can be attributed to the fundamental difference in affective triggers: while StimuVAR relies on detecting abrupt frames or editing points to gauge arousal, the sensation of satisfaction in OSVs is intrinsically tied to *continuous temporal processes*. StimuVAR's discrete sampling strategy inadvertently discards these critical motion dynamics, whereas OSVAR's flow-based approach effectively preserves the visual haptics essential for fine-grained classification.

Table 1: **Main Comparison with State-of-the-Art Methods.** We report the binary accuracy, fine-grained (11-Class) accuracy, and semantic similarity (Sem. Sim) with ground truth explanations. OSVAR significantly outperforms open-source baselines and achieves competitive performance against proprietary giants in fine-grained reasoning.

Category	Model	Binary Acc (%)	11-Class Acc (%)	Sem. Sim (0-1)
Open-source	Video-LLaMA 2 (Zhang et al., 2023)	56.2	22.7	0.38
	LLaVA-NeXT-Video (Li et al., 2024a)	58.9	21.3	0.41
	Molmo2-8B (Deitke et al., 2024)	65.3	28.2	0.45
Base Model	Qwen2.5-Omni 7B (Jin Xu, 2025) (Zero-shot)	70.3	34.4	0.49
	Qwen2.5-Omni 7B(Few-shot)	74.1	38.2	0.52
Ours	**OSVAR (Auto-reg)**	83.3	45.8	0.61
Close-source	doubao-seed1.6 (Volcano Engine, 2024)	86.7	62.8	**0.76**
Ours	**OSVAR (Full Model)**	**88.6**	**66.7**	0.73

Table 2: **Comparison with Specialized VAR Methods.** Our approach outperforms StimuVAR for satisfaction classification.

Method Configuration	Binary Acc	11-Class Acc
StimuVAR (Guo et al., 2025)	64.8	32.5
StimuVAR (Auto-reg)	82.4	44.9
StimuVAR + our Audio Loss	84.6	48.1
OSVAR (Full Model)	**88.6**	**66.7**

Table 3: **Ablation Study.** Each component contributes to the performance, with audio-visual synthetic module the greatest.

Config	Components			Binary Acc	11-Class Acc
	Video	Audio	Sync		
Base (auto-reg)	×	×	×	83.3	45.8
+ Video	✓	×	×	85.7	55.4
+ Audio	×	✓	×	83.4	46.2
+ AV-Sync	×	×	✓	86.3	57.2
Full	✓	✓	✓	**88.6**	**66.7**

Second, combing StimuVAR with our audio supplementary module can improve semantic understanding of the model, but there still exist a 4% gap between this combing methods and our OSVAR.

4.4 Ablation Study

To disentangle the contribution of each psychophysical constraint to the final performance, we conduct a systematic ablation study. As detailed in Table 3, the results validate the indispensability of each proposed module.

The introduction of the visual perception enhancement Module yields a marked improvement of +9.6% in fine-grained accuracy. This substantial gain underscores the hypothesis that satisfaction categories are defined not merely by static objects (e.g., sand), but by their kinematic dynamics. By explicitly modeling motion predictability, the model effectively captures the distinct visual haptics required for fine-grained classification.

Notably, the audio-visual synesthesia module emerges as the most dominant individual component, boosting the 11-class accuracy by +11.4%. This result empirically confirms that the core of the oddly satisfying experience lies in synesthesia. Without this synchronization constraint, even a powerful MLLM fails to verify whether a sound genuinely originates from the visual interaction, leading to misclassifications in complex scenarios.

An interesting observation is that the audio constraint alone provides only marginal gains. This suggests that while acoustic purity is a necessary condition for satisfaction, it is not sufficient for discrimination without visual context. However, when integrated into the full model, these components demonstrate a powerful synergistic effect, elevating the performance to 66.7%. This peak performance indicates that OSVAR successfully learns a holistic representation where visual haptics, acoustic purity, and synesthesia mutually reinforce one another to decode the complex affect of satisfaction.

4.5 Evaluation of Explanation Quality

We also employ "LLM-as-a-Judge" to evaluate the reasoning quality. As shown in Table 4, the baseline Qwen2.5-Omni typically generates generic descriptions, resulting in low scores for Sensory Detail (2.8/5). Conversely, OSVAR achieves a high Sensory Detail score of 4.5/5, indicating its ability to articulate specific textural properties. Moreover, the AV Correlation score improves drastically to 4.2, validating that our synesthesia module successfully teaches the model to ground its reasoning in the synchronization of sight and sound.

Table 4: **LLM-as-a-Judge Evaluation.** Scores (0-5) by Gemini-3 on three psychophysical dimensions.

Model	Rationality	Sensory	AV-Corr	Avg
Video-LLaMA 2	2.5	1.9	1.5	1.97
Qwen2.5-Omni	3.5	2.8	2.8	3.03
OSVAR (Ours)	**4.6**	**4.5**	**4.2**	**4.43**

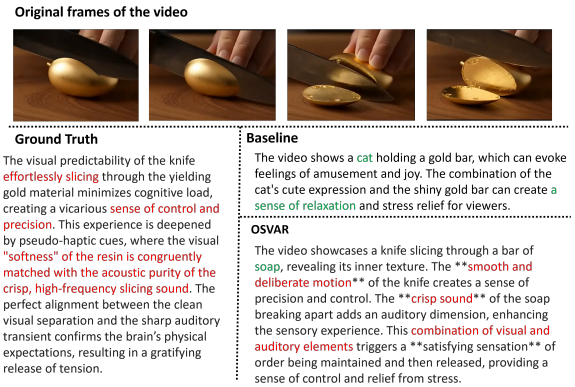

Figure 2: A case study to demonstrate the supercity of our OSVAR framework in affective reasoning.

4.6 Case Study

Figure 2 presents a representative sample (slicing gold egg) to illustrate the qualitative disparity between models. The ground truth explicitly attributes the sensation of relief to "pseudo-haptic cues", linking the "visual predictability" and "acoustic purity" to a psychological state of vicarious control. In stark contrast, the baseline model suffers from severe object hallucination and affective misalignment. It misidentifies the gold soap as a cat holding a gold bar and defaults to a generic emotional guess (amusement, cute). This failure highlights the limitation of standard encoders: they prioritize high-level semantic objects (nouns) while overlooking the fine-grained physical dynamics (verbs and adjectives) essential for satisfaction reasoning.

Conversely, OSVAR demonstrates a human-like understanding of the decompression mechanism. First, it accurately grounds the visual content ("knife slicing through soap"), proving the effectiveness of our visual perception enhancement in capturing distinct material textures. Second, it explicitly articulates the "smooth and deliberate motion" (visual flow) and the "crisp sound" (auditory texture), aligning perfectly with our audio conditional constraints. Crucially, OSVAR bridges the sensory cues to the psychological outcome, explaining how the combination triggers a "sense of order being maintained." This confirms that OSVAR does not merely caption the scene but decodes the underlying logic of satisfaction.

5 Conclusion

In this paper, we introduce **OSVAR**, the first multimodal framework dedicated to decoding the psychophysical mechanisms behind oddly satisfying videos. By bridging the gap between low-level sensory signals and high-level affective reasoning, we propose a novel training paradigm that injects priors of visual haptics, acoustic purity, and synesthesia into Large Language Models. Extensive experiments on our newly curated OSVD dataset demonstrate that OSVAR significantly outperforms state-of-the-art MLLMs in fine-grained satisfaction classification and explanation generation. Beyond technical metrics, this work offers a promising direction for digital psychotherapy, illustrating how AI can act not merely as a passive observer, but as an empathetic partner capable of understanding human sensory relief.

6 Limitations

Despite the promising results, our current framework entails two primary limitations. First, the dependence on explicit optical flow extraction introduces significant computational overhead, potentially hindering real-time deployment on edge devices. Future work could explore implicit motion modeling or lightweight flow distillation techniques. Second, our model currently predicts the *consensus* of satisfaction, overlooking the inherent subjectivity of affective perception (e.g., *Misophonia*, where specific sounds trigger negative reactions in certain individuals). We plan to extend OSVAR towards personalized affective reasoning, incorporating user-specific feedback loops to tailor the sensory analysis to individual psychological profiles.

7 Ethical Considerations

7.1 Data Usage

This study utilizes publicly accessible videos from YouTube, adhering strictly to their respective terms of service. We only collect data that is openly available to all users, avoiding any private or sensitive information. To protect user privacy, all personal identifiers have been removed or anonymized prior to analysis. We recognize the ethical responsibility

to handle user-generated data respectfully and ensure that our analyses do not infringe on individual rights or lead to unintended harm.

7.2 Code and Transparency

We are committed to transparency and reproducibility in our research. To this end, we plan to release the code, fine-tuned models, and the OSVD dataset (subject to platform policies and privacy considerations) to enable verification and further research by the community. Detailed documentation and usage guidelines will accompany the release to promote responsible use. We acknowledge the importance of providing clear explanations of model capabilities and limitations to prevent misuse or overinterpretation of our findings.

References

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Weihs, Alasdair Tran, Sujay Kumar, Waimant Tan, Jenia Jitsev, Luka Salvador, Peter Jansen, Kiana Ehsani, Ehsan Sayyad, Daniel McDuff, and 5 others. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

L. Fadiga, L. Fogassi, G. Pavesi, and G. Rizzolatti. 1995. Motor facilitation during action observation: A magnetic stimulation study. *Journal of Neurophysiology*, 73(6):2608–2611.

Sabrina Faramarzi. 2018. The odd psychology behind oddly satisfying slime videos. *Wired. https://www.wired.co.uk/article/oddly-satisfying-videos-explained-psychology-youtube [access 15 VI 2019]*.

A. Flinker, A. Korzeniewska, A. Y. Shestyuk, P. J. Franaszczuk, N. F. Sztaheler, R. T. Knight, and N. E. Crone. 2015. Redefining the role of Broca's area in speech. *Proceedings of the National Academy of Sciences*, 112(9):2871–2875.

Karl Friston. 2012. Predictive coding, precision and synchrony. *Cognitive neuroscience*, 3(3-4):238–239.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Gemini Team and Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sanggil Lee, Zhifeng Kong, Joao Felipe Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, and 1 others. 2025. Music flamingo: Scaling music understanding in audio language models. *arXiv preprint arXiv:2511.10289*.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190.

Yuxiang Guo, Faizan Siddiqui, Yang Zhao, Rama Chellappa, and Shao-Yuan Lo. 2025. Stimuvar: Spatiotemporal stimuli-aware video affective reasoning with multimodal large language models. *International Journal of Computer Vision*, pages 1–17.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. 2024. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5325–5329. IEEE.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710.

Jinzheng He Hangrui Hu Ting He Shuai Bai Keqin Chen Jialin Wang Yang Fan Kai Dang Bin Zhang Xiong Wang Yunfei Chu Junyang Lin Jin Xu, Zhifang Guo. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. 2019. Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2755–2766.

Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. 2019. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10143–10152.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024a. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Fumiko Maeda, Galit Kleiner-Fisman, and Alvaro Pascual-Leone. 2002. Motor facilitation while observing hand actions: specificity of the effect and role of observer's orientation. *Journal of neurophysiology*, 87(3):1329–1335.

Shentong Mo and Pedro Morgado. 2023. A unified audio-visual learning framework for localization, separation, and recognition. In *International Conference on Machine Learning*, pages 25006–25017. PMLR.

Simone Patuzzo, Antonio Fiaschi, and Paolo Manganotti. 2003. Modulation of motor cortex excitability in the left hemisphere during action observation: a single-and paired-pulse transcranial magnetic stimulation study of self-and non-self-action observation. *Neuropsychologia*, 41(9):1272–1278.

Giulia Lara Poerio, Emma Blakey, Thomas J Hostler, and Theresa Veltri. 2018. More than a feeling: Autonomous sensory meridian response (asmr) is characterized by reliable changes in affect and physiology. *PloS one*, 13(6):e0196645.

Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. 2024. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.

Faegheh Sardari, Armin Mustafa, Philip JB Jackson, and Adrian Hilton. 2024. Coleaf: A contrastive-collaborative learning framework for weakly supervised audio-visual video parsing. In *European Conference on Computer Vision*, pages 1–17. Springer.

Dhruv Srivastava, Aditya Kumar Singh, and Makarand Tapaswi. 2023. How you feelin'? learning emotions and mental states in movie scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2517–2528.

Maria Sumpf, Sebastian Jentschke, and Stefan Koelsch. 2015. Effects of aesthetic chills on a cardiac signature of emotionality. *PloS one*, 10(6):e0130117.

Sirnam Swetha, Hilde Kuehne, and Mubarak Shah. 2025. Timelogic: A temporal logic benchmark for video qa. *arXiv preprint arXiv:2501.07214*.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pretraining. *Advances in neural information processing systems*, 35:10078–10093.

Zoé Vanhersecke. 2021. 'oddly satisfying'? or WEIRD satisfaction... LSE Psychological & Behavioural Science Blog. Accessed: 2026-01-06.

Luis Vilaca, Yi Yu, and Paula Viana. 2025. A survey of recent advances and challenges in deep audio-visual correlation learning. *ACM Computing Surveys*, 57(12):1–46.

Volcano Engine. 2024. Doubao-seed-1.6. https://www.volcengine.com/product/doubao. Large Language Model, accessed via Volcano Ark Platform.

Meg Watson. 2021. Hypnotic loops and self-soothing sounds: the rise of #oddlysatisfying and visual ASMR. The Guardian. Retrieved 9 April 2022.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Dingkang Yang, Kun Yang, Mingcheng Li, Shunli Wang, Shuaibing Wang, and Lihua Zhang. 2024. Robust emotion recognition in context debiasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12447–12457.

Hui Yang, Lekha Chaisorn, Yunlong Zhao, Shi-Yong Neo, and Tat-Seng Chua. 2003. Videoqa: question answering on news video. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 632–641.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 13040–13051.

Fathima Yusaira and Cathlyn Niranjana Bennett. 2021. Influence of autonomous sensory meridian response on relaxation states: An experimental study. *NeuroRegulation*, 8(4):184–184.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.