Why Do Vision Language Models Struggle To Recognize Human Emotions?

Madhav Agarwal, Sotirios A. Tsaftaris, Laura Sevilla-Lara, Steven McDonagh

The University of Edinburgh, UK
{madhav.agarwal, s.tsaftaris, l.sevilla, s.mcdonagh}@ed.ac.uk

Abstract. Understanding emotions is a fundamental ability for intelligent systems to be able to interact with humans. Vision-language models (VLMs) have made tremendous progress in the last few years for many visual tasks, potentially offering a promising solution for understanding emotions. However, it is surprising that even the most sophisticated contemporary VLMs struggle to recognize human emotions or to outperform even specialized vision-only classifiers. In this paper we ask the question 'Why do VLMs struggle to recognize human emotions?', and observe that the inherently continuous and dynamic task of facial expression recognition (DFER) exposes two critical VLM vulnerabilities. First, emotion datasets are naturally long-tailed, and the web-scale data used to pre-train VLMs exacerbates this head-class bias, causing them to systematically collapse rare, under-represented emotions into common categories. We propose alternative sampling strategies that prevent favoring common concepts. Second, temporal information is critical for understanding emotions. However, VLMs are unable to represent temporal information over dense frame sequences, as they are limited by context size and the number of tokens that can fit in memory, which poses a clear challenge for emotion recognition. We demonstrate that the sparse temporal sampling strategy used in VLMs is inherently misaligned with the fleeting nature of micro-expressions (0.25–0.5 seconds), which are often the most critical affective signal. As a diagnostic probe, we propose a multi-stage context enrichment strategy that utilizes the information from 'in-between' frames by first converting them into natural language summaries. This enriched textual context is provided as input to the VLM alongside sparse keyframes, preventing attentional dilution from excessive visual data while preserving the emotional trajectory.

Keywords: Temporal Understanding · Vision-language models · Emotion Recognition

1 Introduction

Comprehending human emotion is central to natural social interaction and a prerequisite for trustworthy, human-centred AI. Emotion recognition infers a

Project Page: https://madhav1ag.github.io/vlm-temporal-emotion-gap

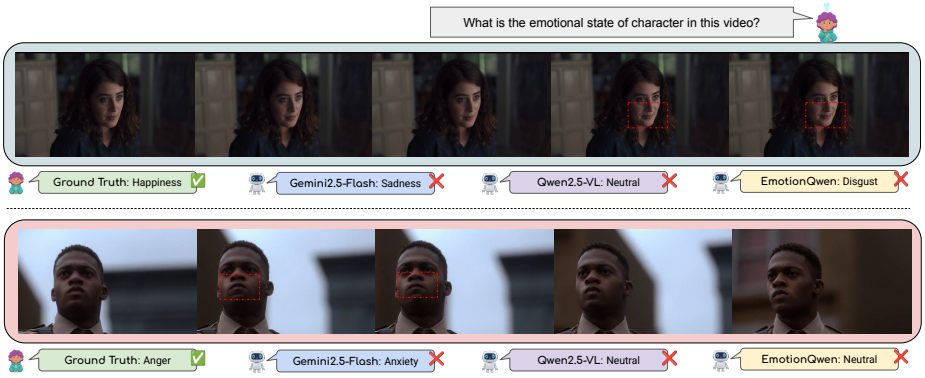

Fig. 1: Qualitative Failure Analysis of SOTA VLMs on Video Emotion Recognition. VLMs like Gemini2.5-Flash, Qwen2.5-VL and EmotionQwen misclassify emotions by failing to capture subtle, temporal cues. Examples include overlooking a brief smile (top row, 'Happiness' misclassified as 'Sadness/Neutral/Disgust') or a tensed facial expression (bottom row, 'Anger' misclassified as 'Neutral/Anxiety')

person's affective state from behavioral signals (e.g., facial expressions, voice, gesture), requiring the integration of subtle, dynamic patterns that unfold over time. Humans excel at this by picking up on fleeting micro-expressions (0.25–0.5 seconds) [18, 23, 45], modulate interpretations by context, and cope with a naturally long-tailed distribution in which common states (*e.g.*, neutral) dominate while rare but consequential emotions occur infrequently.

Equipping digital systems with comparable sensitivity would bring tangible benefits: from more supportive mental-health screening and safer conversational agents to richer assistive technologies for education and care. Realising these gains requires models that respect the temporal nature of affect and perform robustly across both common and rare emotional states, aligning machine perception more closely with the way humans read and respond to one another.

Contemporary Vision–Language Models (VLMs) such as the Gemini-series [14,59] and Qwen-series [4,5,7,62] can demonstrate strong performance when recognizing static objects and in grounding long textual descriptions. However, they do not yet exhibit robust emotion recognition abilities. Inferring human affective state from video demands temporal modeling to track how emotions evolve [29] and the capturing of short-lived micro-expressions. In practice, contemporary VLMs often conflate closely related states (*e.g.*, sadness versus disappointment). We hypothesize this gap arises because the task of dynamic emotion recognition inherently relies on overcoming two long-standing challenges that contemporary VLMs are uniquely vulnerable to: firstly, the inherently long-tailed distribution of affective categories, which induces head-class bias, whereby VLMs over-predict common emotions and under-recognize rare but consequential concepts. Secondly, inadequate temporal representation under limited context windows (token budgets), which promotes a 'bag-of-frames' sampling strategy. This is fundamentally misaligned with the fleeting nature of micro-expressions and disrupts the temporal dynamics essential for accurate emotion classification.

We investigate the root causes of these VLM performance deficiencies. We ask three questions: 1) How large is the performance gap between VLMs and vision-only classifiers in recognizing human emotions from video? 2) Why does this gap arise? and 3) How can it be mitigated?

We adopt video-based emotion classification as a probe task; it couples fine-grained spatial sensitivity to subtle facial muscle movements with temporal fidelity and short-lived micro-expressions, making it a stringent benchmark for high-fidelity temporal modeling. This setting also reflects the intrinsically long-tailed distribution of affective categories. We quantify failures, linking them to two root causes: (i) head-class bias from long-tailed data and (ii) weak temporal modeling under fixed token budgets that promote sparse, order-agnostic frame sampling.

On our first issue, we highlight that VLMs trained on internet-scale data induce natural biases and inherit skewed concept and class-frequency distributions [46, 49]. Since underlying training data specifications are often proprietary, we cannot directly quantify emotion class prevalence. However, we note that vision-language datasets are predominantly constructed by scraping web images and their associated HTML alt-texts. Because VLMs align visual features with these text embeddings during contrastive pre-training, the lexical frequency of an emotion concept in human-generated text directly dictates its representation density in the multimodal latent space. Building on this premise, we utilize historical lexical frequencies of emotion terms in Google Books Ngrams [38, 47] as a representative proxy for this pre-training distribution bias. This provides a high-level correlative indicator of conceptual prevalence in human-generated data, where we observe a pronounced long-tail. Class sparsity is inversely associated with per-class classification accuracy. Rarer emotions incur larger error rates in contemporary VLMs (see Sec. 3.3). The direct, empirical validation of this hypothesis, via controlled dataset balancing, is presented in Sec. 4.1.

Our second key issue stems from observations that VLM abilities to retain and prioritise relevant information degrade as context length increases [41, 67]. This limitation is exacerbated in subtle video understanding tasks such as emotion recognition, where fine-grained cues are diluted by redundant frames under a fixed token budget. To assess temporal understanding we vary frame-sampling strategies and the effective frame rate, and probe order sensitivity by shuffling frames. We observe a quasi-bell-shaped relationship between performance and the number of frames available during inference: accuracy initially improves with denser evidence but declines once the token budget is saturated. Even at its peak, we observe that VLM performance remains below vision-only classifier baselines. Moreover, we observe that accuracy is largely invariant to frame ordering, indicating weak temporal modelling and a reliance on order-agnostic aggregation of per-frame appearance.

To mitigate these two fundamental failures, we propose corresponding 'plug-and-play' solutions. To address the long-tail bias, we demonstrate the efficacy of a decoupled training strategy using a balanced dataset (Sec. 4.1). More centrally, to solve the temporal bottleneck, we introduce a multi-stage context enrichment

strategy. This inference-time pipeline addresses the core conflict between temporal density, needed for micro-expressions, and token budget limitations. Instead of merely selecting from a sparse set of frames which risks information loss from induced gaps, our approach enriches the context. We firstly use a VLM to perform modality translation on the 'in-between' frames, converting subtle motion cues into high-level natural language summaries. This enriched textual context, which VLMs excel at processing, is then fed to the VLM alongside the original sparse keyframes to perform final classification, allowing the model to capture the fleeting dynamics it would otherwise have missed. Crucially, our objective is not to engineer a task-specific model to marginally surpass current Dynamic Facial Expression Recognition (DFER) state-of-the-art benchmarks. Specialized models utilizing 3D-CNNs or spatial-temporal transformers currently dominate these metrics, but they lack the open-ended reasoning, zero-shot generalization, and conversational capabilities of VLMs. Our primary contribution is therefore not to chase these benchmarks but to provide a new understanding of VLM behavior for affective computing. We systematically diagnose why generalized, massive-scale VLMs fail at affective tasks, and demonstrate that unlocking their temporal bottlenecks is a critical stepping stone toward achieving human-level proficiency in affective scene understanding.

Our main contributions can be summarised as:

- We provide empirical evidence that long-tailed distributions, characteristic of internet-scale data, correlate with per-class accuracy, explaining underperformance in the emotion recognition task.
- We analyse temporal limitations of contemporary VLMs, using video-based emotion classification as a probe to quantify performance and dissect root failure causes.
- As a secondary, diagnostic intervention, we propose a plug-and-play strategy that strengthens temporal modeling while keeping track of micro-expressions, within fixed token budget, yielding consistent improvements in performance on emotion classification tasks.

2 Related Work

2.1 Vision-Language Models for Videos

The rise of Large Language Models (LLMs) [4,8,9,60] has driven the use of vision encoders with language models to handle multi-modal tasks. Vision-Language Models [1, 7, 14, 59] have shown promising results in tasks like image captioning and visual question answering. These approaches are further extended to incorporate videos directly [2, 44, 68, 69], by employing cross-modal attention and aligning video features with the LLM latent space. Although these models demonstrate promising early results on coarse-grained video understanding tasks [27, 64], they exhibit limited capacity to attend to specific video segments and to perform temporal reasoning [3, 17, 41, 61].

In particular, recent benchmarks highlight that VLMs often exploit spatial and textual biases rather than temporal reasoning [15], revealing a critical dis-

parity between their observed performance and their underlying reasoning capabilities. The performance deficit in temporal reasoning stems from several foundational failures. Overly long context windows may introduce large amounts of irrelevant context, which can mislead the model and diminish its ability to focus on task-relevant information [25, 31, 41, 67]. However, the challenge extends beyond the *size* of the context window to the more fundamental limitation that current models cannot *utilize* it effectively [3]. Most notably, with large context windows, models succumb to irrelevant distractors, highlighting a deeper attention failure rooted in the 'bag-of-words' tendency in foundational spatial reasoning [50]. The problem is compounded by a positional encoding breakdown [21, 54], which arises due to naïve handling of continuous video frames and discrete language tokens in a similar manner. In contrast to previous work, we focus on highlighting how these fundamental issues are a direct cause of failure in complex temporal tasks like emotion understanding, along with an inference strategy with enriched context.

A complementary line of work improves affective recognition by pairing facial-language models with structured or specialist facial cues. For DFER, DK-CLIP [35] and PE-CLIP [53] inject AU-grounded textual prompts into CLIP, Fine-CLIPER [11] fuses landmarks, segmentation masks, and MLLM-generated descriptions, and AU-DFER [39] adds AU–expression priors to conventional networks. For static FER, ExpLLM [34] derives a chain-of-thought from detected Action Units to guide an MLLM. Dedicated facial-video MLLMs such as FaVChat [73] further report strong DFER results, surpassing a 72B generalist VLM on DFEW by a wide margin. Rather than contradicting our analysis, these works reinforce it: each succeeds by supplying generalist VLMs with the additional structure or specialisation our diagnosis shows they lack.

2.2 Long-Tail Bias

Long-tailed data distributions are known to degrade deep learning performance [20, 71], with bias embedding deeply in the feature representations, not just the final predictions [40]. Therefore, contemporary solutions aim to mitigate the final classification output and also seek to correct the underlying distortion in the feature space. Long-tail distribution bias is not limited to classification; it also prevails in regression [75], semantic segmentation, and instance segmentation [65, 70]. A standard approach for handling long-tail imbalance is through class rebalancing [19, 42, 66], which alters the training distribution by over-sampling tail classes [10] or under-sampling head classes. While simple, these methods risk overfitting on the tail or discarding valuable information from the head. Information augmentation techniques are also used to enrich the sparse tail class by creating synthetic samples using generative models like GANs [22, 33, 51] or LLMs [13, 63]. Such approaches typically incur substantial computational overhead and are currently infeasible for video-based classification tasks using VLMs, which are already pre-trained on massive web-scale data. Alternative prior work decouples training into a two-stage approach [32, 48, 57] in which a generalised feature extractor is trained firstly on a naturally imbalanced dataset, followed by

classifier finetuning using a class-balanced sampler. We adopt pretraining plus class-balanced finetuning to probe the long-tail hypothesis with minimal confounds, deferring more advanced mitigation (*e.g.* logit adjustment, reweighting, tail augmentation) to avoid added complexity and compute while keeping the focus on diagnosis.

3 Systematic Diagnosis of VLM Deficiencies

We designed experiments to validate our hypothesis: *class imbalance in training data and architectural limitations of VLMs lead to poor performance on temporal understanding tasks, like emotion classification.* We perform multi-class emotion classification using both VLMs and vision-only classifiers towards thorough experimental evaluation; experimental details follow.

3.1 Dataset and Evaluation Metrics

To evaluate dynamic emotion recognition, we utilize two in-the-wild datasets: MAFW [43] (11-class emotions) and DFEW [30] (7-class emotions). Standard DFER evaluation conventionally relies on two primary metrics: Weighted Average Recall (WAR), representing overall accuracy by accounting for class proportions, and Unweighted Average Recall (UAR), which averages recall across all classes regardless of frequency. UAR is uniquely critical as it treats every emotion category equally, preventing the majority classes from masking minority class failures. Standard WAR metrics often mask failures on the minority tail in imbalanced datasets like DFEW. We therefore utilise balanced testing splits, sampling 495 videos (45 videos per class) for MAFW and 700 videos (100 videos per class) for DFEW, to ensure a rigorous evaluation free from majority-class bias and use the remaining videos for training. Due to this uniformity, UAR and WAR are mathematically identical in our evaluations.

3.2 Models

We used representative examples from four model categories for our experimental work: closed-source general-purpose VLMs (Gemini2.5-Flash [14]), open-source general-purpose VLMs (Qwen2.5-VL [7], Qwen2.5-Omni [68], Qwen3-VL [6], Video-LLaVA [36], LLaVA-NeXT-Video [72], InternVL-3.0 [74]), open-source task-specific VLMs (EmotionQwen [28]), and open-source task-specific vision-only classifiers (MAE-DFER [55] and HiCMAE [56]). The task-specific models are models designed specifically for emotion understanding. Where possible, open-source VLM parameter counts were kept within the 7–8B range for fair comparison.

3.3 Long-tail distribution effects

Real-world data distributions with respect to model performance: Empirical category frequencies in many naturalistic and web-scale datasets are

heavy-tailed and well approximated by Pareto and Zipf-like distributions [52,58]: a small number of object categories are exceptionally common, while the vast majority are rare.

Early in the deep-learning era, computer vision models were often evaluated on artificially balanced datasets like ImageNet [16] and MS-COCO [37], whereas modern VLMs are trained on larger, web-scraped data. Consequently, the underlying data distribution is often opaque, particularly the class-conditional sampling frequencies and long-tail coverage relative to any chosen label set. A class distribution is considered long-tailed when a small minority of head classes account for a substantial share of all training examples, while the majority of tail classes each have few examples and their frequencies decline in a heavy-tailed (often power-law) manner. In practice, the analysis of common large-scale datasets, used for fine-tuning VLMs, exhibit long-tail patterns [49].

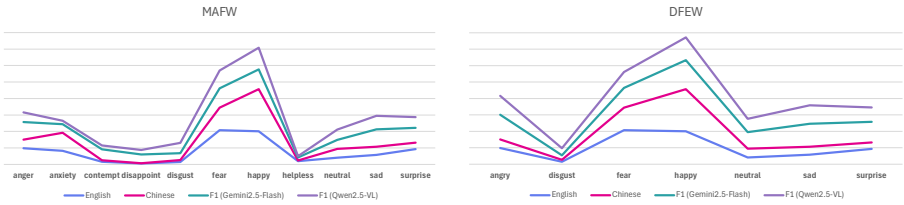

Fig. 2: Correlation between Lexical Frequency and VLM Accuracy. We plot per-class F1 scores for VLMs against the lexical frequency of emotion terms in the Google Books Ngram corpus. The positive correlation suggests a head-class bias: rarer emotions (e.g., 'contempt', 'helplessness') have lower lexical frequency and correspond to weaker VLM performance.

As an initial probing analysis, we estimate the data distribution for eleven classes from the MAFW [43] dataset. For VLMs trained on image-text pairs, we note that text frequency dictates supervision. With proprietary pre-training data unavailable, we utilize historical lexical frequencies in Google Books Ngrams [38, 47] as a representative proxy for this 'linguistic bottleneck'. As established in our results, the strong correlation between Ngram frequency and VLM error rates provides suggestive empirical support for this proxy. We selected the per-class frequency from the July 2024 corpora for each emotion label in English and Chinese languages. We compare Ngram frequency with the per-class F1 score of Gemini2.5-Flash [14] and Qwen2.5-VL [7] inference results, using a uniformly balanced test set from MAFW [43], in a zero-shot setting (Figure 2). We calculate the Pearson correlation coefficient for each pair to estimate the relationship between model predictive accuracy and the occurrence of each class in Google Books Ngram data. Unlike prior work establishing long-tail inheritance for general objects via synonym counts in pretraining captions [49], we show this bias in emotion recognition couples to historical lexical prevalence, with cross-lingual consistency suggesting a deeper cultural and linguistic skew beyond raw pretraining statistics. The Pearson correlation coefficient is calculated using:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{1}$$

	English	Chinese	Gemini2.5-Flash	Qwen2.5-VL
English	1.000 (—)	0.8825 (0.0003)	0.7927 (0.0036)	0.8041 (0.0029)
Chinese	0.8825 (0.0003)	1.000 (—)	0.6343 (0.0361)	0.7547 (0.0073)
Gemini2.5-Flash	0.7927 (0.0036)	0.6343 (0.0361)	1.000 (—)	0.8176 (0.0021)
Qwen2.5-VL	0.8041 (0.0029)	0.7547 (0.0073)	0.8176 (0.0021)	1.000 (—)

Table 1: Correlation between per-class occurrence in Google Books Ngram data (English, Chinese) and F1 accuracy of Gemini2.5-Flash and Qwen2.5-VL on MAFW. We report correlation coefficients and (p-values).

where x and y are the two signals and $\bar{x}$ and $\bar{y}$ are their respective means, and n is the number of data points (here emotion classes). Statistical significance is assessed using a two-tailed t-test, with p-values across model-language pairs (Gemini2.5-Flash and Qwen2.5-VL; English and Chinese) indicating that all correlations are statistically significant ($p < 0.05$; see Table 1). VLMs are trained predominantly on English web data, so English Ngrams are expected to be the closer proxy; Chinese affective lemmas are also more polysemous, inflating raw frequency relative to affective usage.The cross-lingual consistency of the correlation, rather than the absolute frequencies, is what strengthens the long-tail interpretation.

We further examine whether similar correlations emerge in specialist (vision-only) emotion classifiers, thus controlling for VLM-specific pretraining and context windows. Essentially our objective is to test whether the observed accuracy–frequency coupling reflects the long-tailed training distribution itself. We selected two state-of-the-art models MAE-DFER [55] and HiCMAE [56], which are based on the MAE [24] architecture and pre-trained on Voxceleb2 data [12]. Voxceleb2 contains no emotion classification labels yet we treat this as a large audio-visual resource, capable of producing a strong, general purpose feature extractor. We fine-tuned the models on the training split of the MAFW dataset (Sec. 3.1), which exhibits class imbalance. Some of the classes are high-frequency: anger, happiness, neutral, sadness, surprise, while others are nuanced or long-tail: contempt, disappointment, helplessness. We evaluate the performance of the fine-tuned models on the evenly sampled testing split. We observe a similar performance pattern in specialized vision-only video classification models. For instance, the F1 score for happiness was 0.70 in MAE-DFER, while low-frequency classes like contempt and disappointment has 0.18 and 0.22, respectively. Similarly, HiCMAE has an F1 score of 0.69 for happiness, but 0.04 and 0.16 for helplessness and disappointment, respectively. These findings are consistent with our conjecture that web-scale pretraining data, ingested by VLMs, are long-tailed over emotion categories, contributing to weaker performance on under-represented emotions.

Validating the performance on a class-balanced dataset: As Gemini2.5-Flash cannot be fine-tuned (code unavailable), we report its per-class behaviour

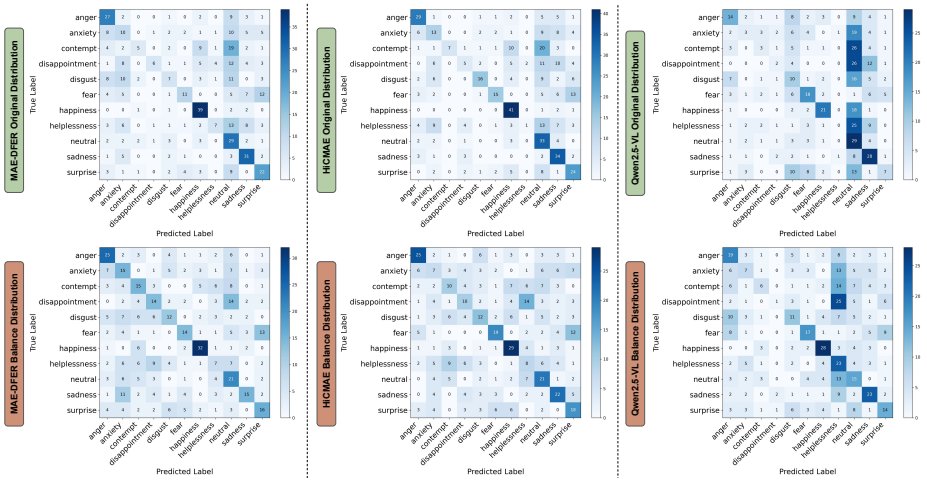

Fig. 3: Confusion matrices for the classification task on MAFW dataset. Original data distribution (top row); fine-tuning with balanced dataset distribution (bottom row).

as a measurement of bias rather than a mitigation result. This is relevant for downstream users of closed models, who cannot retrain but can use such estimates to calibrate trust. The model achieved F1 scores of 0.60 for happiness, 0.09 for helplessness, and 0.26 for disappointment. These results are comparable to those obtained with Qwen2.5-VL, which recorded F1 scores of 0.54 for happiness, 0.0 for helplessness, and 0.07 for disappointment. This indicates that current VLMs have not effectively learned to recognize nuanced emotional expressions; instead, they tend to rely on memorization of the most frequent emotion labels.

When presented with videos depicting rare emotions such as disappointment, the investigated models often resort to predicting the high-frequency or common emotion categories (Figure. 3, *e.g.* top-row, last-column). The confusion matrices reveal strong diagonals for high-frequency classes (*e.g.*, neutral), while predictions for long-tail emotions (*e.g.*, contempt, disappointment) are scattered and inconsistent. This pattern refutes the assumption that class confusion may stem solely from semantic similarity (e.g., Sadness vs. Disappointment). Crucially, our error analysis reveals *asymmetric* confusion: rare classes systematically collapse into common classes (Disappointment → Sadness), but the reverse almost never holds. Pure semantic confusion is bidirectional, whereas this pronounced asymmetry confirms that models default to high-frequency priors rather than failing due to semantic ambiguity. We formalise this with a Directional Asymmetry Index in the supplementary material. Overall, the results highlight a key limitation of current VLMs: training on long-tailed datasets leads to the systematic collapse of low-frequency emotions toward high-frequency categories, thereby diminishing sensitivity to subtle emotional variations.

3.4 Temporal Understanding

Do VLMs leverage frame ordering information? Emotion recognition relies on temporal cues that involve micro- and macro-expressions [18,23,45], evolving over time. This suggests that temporal information likely plays a key role. To examine how crucial temporal information is to task success, we modify the input video frame ordering. For a given input video V with $[I_1, I_2, \ldots, I_L]$ frames, we randomly shuffle the frame order. We evaluate the performance of both VLMs and vision-only classifier models with two sets of input videos containing identical visual (*i.e.* spatial and appearance) information: one with the original chronological temporal ordering and one with the temporal frames shuffled (FS). The specialized temporal models (MAE-DFER, HiCMAE) show a significant (**15–16%**) performance drop in macro-averaged F1 score. In contrast, VLM model performance remains largely indifferent to frame shuffling (see Table 2). Across varying model scales and specific visual integration techniques, the vulnerability remains strikingly consistent: VLMs performance remains largely indifferent to random temporal shuffling, confirming an order-agnostic 'bag-of-frames' processing strategy in all instances.

Our experiment highlights a critical limitation in the temporal awareness and reasoning of current VLMs. This insensitivity provides strong evidence of an over-reliance on frame-level spatial features, indicating a failure to reason across temporal information. We demonstrate that models are not merely failing to reason, but are architecturally defaulting to a 'bag-of-frames' processing strategy. Furthermore, this suggests that the order-aware processing imposed by causal attention masking, which is intended to implicitly guide inter-visual token interactions and guarantee architectural temporal ordering, is also failing. Our frame-shuffling probe provides a real-world corollary that complements SpookyBench [61] by evaluating naturalistic, high-resolution video rather than noise-based patterns, holding spatial content constant while removing only temporal order. While both works find an over-reliance on spatial cues, we identify a distinct failure mode: redundancy-driven attentional dilution. Unlike the failure on noise sequences, we show that accuracy drops as visual token density increases and saturates the model budget, revealing an architectural inability to isolate fleeting signals from clear but redundant frames.

Context Windows and Irrelevant Distractors: Contemporary VLMs process visual frames by dividing them into tokens. These visual tokens are treated analogously to text tokens within the model architecture. Although this design choice appears straightforward and easily adaptable, it overlooks the fundamental distinctions between the modalities. Text tokens are discrete semantic units and categorical: equality is all-or-nothing. Thus 'cat' and 'dog' are distinct labels; in a standard one-hot representation, their vectors are orthogonal. Video frames, however, are dense, continuous, and highly correlated; two adjacent frames in a video are often highly similar. This signifies that a large portion of video information is typically redundant and, for a specific targeted query, often even irrelevant. This problem is compounded by a mechanistic imbalance: video token

Model	MAFW			DFEW		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
MAE-DFER	0.4394	0.3919	0.3602	0.6883	0.6086	0.5645
MAE-DFER FS	0.4959	0.3354	0.3041	0.5006	0.5229	0.4802
HiCMAE	0.5040	0.4404	0.3993	0.6935	0.6214	0.5725
HiCMAE FS	0.4979	0.3778	0.3345	0.4807	0.5329	0.4797
Qwen2.5-VL	0.2849	0.2727	0.2449	0.5053	0.4614	0.4552
Qwen2.5-VL FS	0.3527	0.2727	0.2506	0.5015	0.4643	0.4534
Qwen2.5-Omni	0.4517	0.3253	0.306	0.5557	0.4700	0.4296
Qwen2.5-Omni FS	0.4502	0.3172	0.2972	0.5276	0.4614	0.4226
Qwen3-VL	0.4180	0.3313	0.2738	0.6380	0.5843	0.5511
Qwen3-VL FS	0.3198	0.3273	0.2615	0.6468	0.5871	0.5538
EmotionQwen	0.3478	0.2925	0.2581	0.5521	0.5329	0.5010
EmotionQwen FS	0.3185	0.3057	0.2517	0.5383	0.4968	0.4895
Video-LLaVA	0.1126	0.1630	0.0870	0.1490	0.2800	0.1654
Video-LLaVA FS	0.0811	0.1569	0.0814	0.1514	0.2570	0.1531
LLaVA-NeXT-V	0.1853	0.2040	0.1438	0.5351	0.3474	0.2969
LLaVA-NeXT-V FS	0.1705	0.1980	0.1382	0.4366	0.3314	0.2712
InternVL-3.0	0.3441	0.2828	0.2445	0.5561	0.5357	0.5044
InternVL-3.0 FS	0.3370	0.2667	0.2326	0.5669	0.5214	0.4985
Gemini2.5-Flash	0.4112	0.3869	0.3758	0.6412	0.6331	0.6008
Gemini2.5-Flash FS	0.4286	0.3817	0.3626	0.6246	0.6120	0.5778

Table 2: Model performance on two datasets (MAFW and DFEW) for normal videos vs Frame Shuffled (FS) videos using vision-only classifiers (top section) and VLM-based architectures (bottom section).

embeddings possess a significantly larger vector norm than text token embeddings [50]. Given that attention mechanisms already struggle to model extended temporal context, enlarging the context with distracting and irrelevant content can further exacerbate this limitation. This phenomenon is a video-centric manifestation of the 'lost-in-the-middle' problem [41], where model attention is confused by noise, and performance degrades as sequence length increases.

To evaluate the performance of the model with a varying amount of visual information, we perform an experiment that controls the number of video input frames. We perform this experiment on Qwen2.5-VL [7] and EmotionQwen [28], as their open-source nature allows for fine-grained control over frame-sampling inputs. Closed-source models like Gemini2.5-Flash [14] do not provide this level of API control, and are thus omitted from this specific test. We started by sampling one frame every second from video input and instruct the Qwen2.5-VL [7] model to infer the emotional state from the given sampled frames only. We then increase the number of frames per second to 5, 10, 15, 20, and 25. We observe that, initially, model performance improves when it enjoys access to more frames *i.e.*, more visual information is available to infer the emotional state. However, as we increase the number of frames, model performance starts to degrade, and even eventually drops below the performance obtained when using one frame per second. We repeat the experiment with EmotionQwen [28] and note similar performance trends. (Additional results in supplementary material)

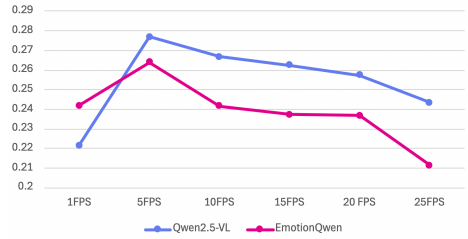

Fig. 4: VLM Performance vs. Input Frame Rate. We plot the macro-averaged F1-score of Qwen2.5-VL and EmotionQwen as a function of frames-per-second (FPS) on MAFW dataset. The quasi-bell-shaped curve shows performance initially improving (1–5 FPS) but then degrading significantly ($>$ 5 FPS), demonstrating attentional dilution from redundant visual tokens.

The performance curves shown in Figure 4 provide empirical evidence in support of our hypothesis that attention mechanisms in VLMs degrade with excessive information and fail to retain long-term context. The initial performance gain is expected, and the subsequent degradation at high frame rates is a notable finding that highlights underlying architectural limitations. This behavior is particularly damaging for complex temporal tasks like emotion understanding, which likely require fine-grained temporal sensitivity to reliably classify micro-expressions. The quasi-bell-shaped curve (Figure 4) provides compelling empirical evidence that VLM attention is not just limited, but is actively diluted by high-density, redundant visual input. This confirms that redundancy and distractors, and a limited ability to exploit long context, constitute a principal bottleneck for temporal reasoning in contemporary VLMs.

4 Enhancing Emotions via Distribution and Dynamics

4.1 Mitigating Long-Tail Bias

Furthering our high-level correlative analysis (Section 3.3), we next consider direct empirical validation of the long-tail hypothesis. To evidence that data imbalance is a primary cause of failure, we employed a 'decoupled' training strategy. We take advantage of the strong feature extraction capabilities learned by models trained on large-scale data. We sampled ∼1500 videos from the MAFW dataset, such that each class has an equal number of samples. We fine-tune both MAE-DFER and HiCMAE on the new uniformly sampled balanced distribution. Despite a six-fold reduction in available fine-tuning data, the distribution of predicted labels at inference is appreciably more uniform across classes, indicating reduced head-class bias. We repeat a similar experiment with Qwen2.5-VL, using LoRA-based finetuning [26]. As shown in Figure 3, this balanced finetuning validates our long-tail hypothesis. It improves tail-class performance and yields a more uniform prediction distribution, confirming the issue is data bias rather than an inherent incapacity for these emotions. In summary, our decoupled pretrain–then–balanced finetuning is simple yet effective: it preserves web-scale representations while correcting head-class bias without introducing

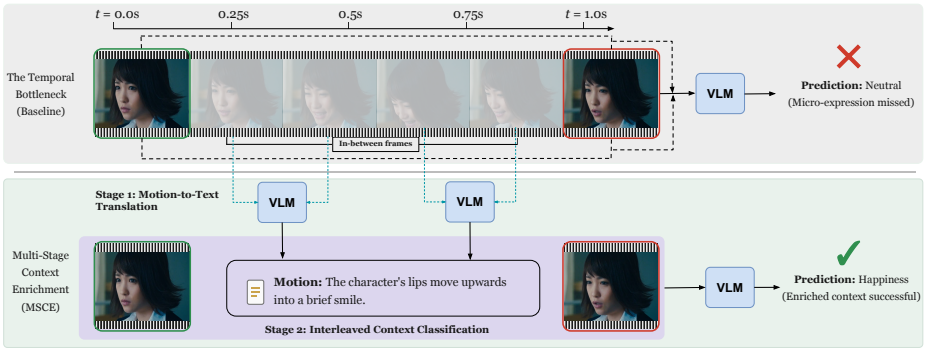

Fig. 5: Multi-Stage Context Enrichment (MSCE) is a two-stage inference pipeline. Top (baseline): a VLM given only sparse keyframes misses the fleeting micro-expression in the in-between frames and predicts *Neutral*. Bottom (MSCE): in Stage 1 (Motion-to-Text Translation), the VLM converts the in-between frames into a compact motion description of the micro-expression; in Stage 2 (Interleaved Context Classification), this textual summary is interleaved with the sparse keyframes for a final prediction, recovering the correct label (*Happiness*).

additional modeling or data-generation confounds. This lightweight step yields consistent gains across all models, providing a pragmatic and clean mitigation when the pretraining distribution is opaque and retraining is infeasible.

4.2 Multi-Stage Context Enrichment for Temporal Reasoning

Our analysis in Section 3.4 provides empirical evidence that VLMs struggle with temporal reasoning due to two competing problems: 1) sparse sampling (*e.g.*, 1 FPS) misses fleeting micro-expressions, and 2) dense sampling (*e.g.* >10 FPS) overwhelms context windows with redundant distractions. This results in a critical problem: diagnostically salient affective cues frequently reside in frames that sparse sampling necessarily excludes under context-window constraints. Unlike frame selection methods such as TCoT [3], which discard content in unsampled temporal gaps and risk losing information, we propose a Multi-Stage Context Enrichment (MSCE) strategy. Rather than discarding these gaps, MSCE translates their inherent motion cues into compact natural-language summaries, a modality that vision-language models (VLMs) process highly effectively. By converting high-volume, redundant visual tokens into low-volume, information-dense text tokens, we augment the model's effective context while respecting the token budget. This strategy is illustrated in Figure 5 and is executed in two stages.

MSCE Stage 1: Motion-to-Text Translation: For video V, we sample n frames $K = \{k_1, k_2, \ldots, k_n\}$, corresponding to our baseline sparse sampling strategy (e.g., 1 FPS dictates that k_1 is at 0.0 seconds, k_2 is at 1.0 seconds, etc.). We refer to this as a set of keyframes and this sparse sampling creates $n-1$ temporal gaps, containing unobserved intermediate frames. For each temporal gap j (containing frames between k_i and k_{i+1}), we sample a small, denser

set of intermediate frames, $G_j = \{g_{j_1}, g_{j_2}, \ldots, g_{j_m}\}$. These frames are sampled uniformly, to maximize input information. The set G_j is intentionally kept small ($m = 4$ in our experiments) to avoid overwhelming VLMs and minimize attention dilution. The set G_j are then processed through the VLM to provide a textual summary of the facial temporal movements, and to capture micro-expressions. This granular text, containing motion information, serves as a surrogate for the temporal information lost under sparse frame sampling, and provides high-level context, absent from the visual inputs. At the end of Stage 1, we have a list of $n-1$ textual summaries, $T = \{t_1, t_2, ..., t_{n-1}\}$, where each t_i describes motion that the sparse keyframes K have missed.

MSCE Stage 2: Interleaved Context Classification: The second stage synthesizes sparsely sampled frames K with the generated textual motion summaries to provide a holistic emotion classification. We create a single text prompt that inter-leaves the keyframes from K with the text summaries from T in chronological order. The input to the final VLM is structured as:
$\{k_1, t_1, k_2, t_2, \ldots, k_{n-1}, t_{n-1}, k_n\}$. The VLM is then prompted to make a final classification based on this complete, enriched context using prompt:

```
Analyze the following sequence of sparse keyframes and the
detailed motion descriptions for the gaps between them.
FrameID 1:{k_1}, Motion:{t_1},···, FrameID n:{k_n}.
Question: {question}
Emotion Labels choices: {List of Emotion Labels}
```

	MAFW			DFEW		
Model	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Qwen2.5-VL	0.2849	0.2727	0.2449	0.5053	0.4614	0.4552
Qwen2.5-VL +MSCE	0.3293	0.3091	0.2731	0.5204	0.4900	0.4820
EmotionQwen	0.3478	0.2925	0.2581	0.5521	0.5329	0.5010
EmotionQwen +MSCE	0.3617	0.3061	0.2683	0.5635	0.5429	0.5147
LLaVA-NeXT-V	0.1853	0.2040	0.1438	0.5351	0.3474	0.2969
LLaVA-NeXT-V +MSCE	0.2294	0.2323	0.1715	0.5167	0.3714	0.3171

Table 3: Our Multi-Stage Context Enrichment (MSCE) strategy results in consistent F1 score improvement compared with baseline sparse sampling.

This interleaved architecture directly remedies the core failures identified in Section 3.4. We further contrast MSCE with the frame-selection method TCoT [3] under a matched Qwen2.5-VL setup. Because micro-expressions last only 0.25–0.5 s, a coarse macro-level selection search such as TCoT [3] discards the unselected gaps that carry vital affective signal, leaving its F1 at or below the sparse baseline (0.241 vs. 0.245 on MAFW, 0.450 vs. 0.455 on DFEW), whereas MSCE improves over both (0.273 and 0.482). Our strategy yields a consistent performance improvement (Table 3), further demonstrating that reasoning over temporally intermediate content is critical. While these gains are moderate, they are diagnostically conclusive for our temporal hypothesis. They empirically prove

that critical affective cues reside in the temporal gaps discarded by sparse sampling, and demonstrate that these cues can be recovered and utilized by VLMs via modality translation. The textual summary t_i acts as a 'temporal bridge', providing explicit, high-level semantic context that spans the visual gap between keyframe k_i and k_{i+1}. This allows the model to reason over the evolution of an expression, *e.g.* a micro-expression fleetingly appearing and disappearing, and consequently, it moves beyond the order-agnostic aggregation of per-frame spatial features. We provide initial evidence that modality conversion can preserve essential temporal information while avoiding visual-token overload.

5 Limitations and Discussion

While MSCE provides measurable gains, its implementation highlights three structural frontiers in long-form video understanding. First, MSCE functions as a strategic intervention rather than a final resolution. While it improves access to temporal cues, we acknowledge that text summarization may introduce generation noise. However, this dependency strategically shifts the burden from the model's weak long-term temporal retention to its strong short-term descriptive capabilities. By interleaving text summaries with original keyframes, the text serves as a supporting signal rather than a sole dependency, with empirical results confirming that the recovered temporal information outweighs any noise introduced by the process. Future work may pursue video-native architectures to permanently resolve the underlying deficit in long-horizon temporal modeling. Second, current VLM attention degrades with very long visual contexts, which constrains sensitivity to micro-expressions; potential remedies include hierarchical or sparse temporal attention and improved long-sequence positional encodings. Third, our long-tail analysis relies on Google Books Ngram frequencies as a correlative, non-causal, proxy for training priors. Given the opacity of web-scale corpora, additional frequency proxies can strengthen the analysis.

6 Conclusion

We investigate the critical failure of modern VLMs to comprehend human emotion and empirically demonstrate that this problem stems from two fundamental flaws. First, we provide quantitative evidence that VLM failures are data-driven. By correlating model accuracy with lexical frequency, we show that VLMs inherit a long-tail bias from their pretraining data, causing them to conflate rare emotions with their high-frequency counterparts. Second, we demonstrate that this limitation is inherently architectural. Empirically, VLMs struggle with temporal modeling, lacking sensitivity to frame order. Moreover, performance deteriorates as visual input density increases, indicating limited capacity to exploit the dense context required to detect micro-expressions. Our experiments suggest that these shortcomings, while potentially addressable, are deeply rooted in current data and architectural practices. Consequently, advancing toward robust affective computing will require fundamental shifts, including a thorough reconsideration of VLM training regimes and temporal modeling architectures.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
3. Arnab, A., Iscen, A., Caron, M., Fathi, A., Schmid, C.: Temporal chain of thought: Long-video understanding by thinking in frames. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
4. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
5. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
6. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
7. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
8. Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., et al.: Deepseek llm: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954 (2024)
9. Cai, Z., Cao, M., Chen, H., Chen, K., Chen, K., Chen, X., Chen, X., Chen, Z., Chen, Z., Chu, P., et al.: Internlm2 technical report. arXiv preprint arXiv:2403.17297 (2024)
10. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
11. Chen, H., Huang, H., Dong, J., Zheng, M., Shao, D.: Finecliper: Multi-modal fine-grained clip for dynamic facial expression recognition with adapters. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 2301–2310 (2024)
12. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. In: INTERSPEECH (2018)
13. Cloutier, N.A., Japkowicz, N.: Fine-tuned generative llm oversampling can improve performance over traditional techniques on multiclass imbalanced text classification. In: 2023 IEEE International conference on big data (BigData). pp. 5181–5186. IEEE (2023)
14. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
15. Cores, D., Dorkenwald, M., Mucientes, M., Snoek, C.G.M., Asano, Y.M.: Lost in time: A new temporal benchmark for videollms. In: 36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025. BMVA (2025)

16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. IEEE (2009)
17. Ding, X., Wang, L.: Do language models understand time? In: The First International Workshop on Transformative Insights in Multifaceted Evaluation at The Web Conference 2025 (2025)
18. Ekman, P., Friesen, W.V.: Nonverbal leakage and clues to deception. Psychiatry **32**(1), 88–106 (1969)
19. Estabrooks, A., Jo, T., Japkowicz, N.: A multiple resampling method for learning from imbalanced data sets. Computational intelligence **20**(1), 18–36 (2004)
20. Fang, C., Zhang, D., Zheng, W., Li, X., Yang, L., Cheng, L., Han, J.: Revisiting long-tailed image classification: Survey and benchmarks with new evaluation metrics. arXiv preprint arXiv:2302.01507 (2023)
21. Ge, J., Chen, Z., Lin, J., Zhu, J., Liu, X., Dai, J., Zhu, X.: V2pe: Improving multimodal long-context capability of vision-language models with variable visual position encoding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21070–21084 (2025)
22. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
23. Haggard, E.A., Isaacs, K.S.: Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In: Methods of research in psychotherapy, pp. 154–165. Springer (1966)
24. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
25. Hsieh, C.P., Sun, S., Kriman, S., Acharya, S., Rekesh, D., Jia, F., Ginsburg, B.: RULER: What's the real context size of your long-context language models? In: First Conference on Language Modeling (2024)
26. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
27. Huang, B., Wang, X., Chen, H., Song, Z., Zhu, W.: Vtimellm: Empower llm to grasp video moments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14271–14280 (2024)
28. Huang, D., Li, Q., Yan, C., Cheng, Z., Huang, Y., Li, X., Li, B., Wang, X., Lian, Z., Peng, X.: Emotion-qwen: Training hybrid experts for unified emotion and general vision-language understanding. arXiv preprint arXiv:2505.06685 (2025)
29. Jack, R.E., Garrod, O.G., Schyns, P.G.: Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. Current biology **24**(2), 187–192 (2014)
30. Jiang, X., Zong, Y., Zheng, W., Tang, C., Xia, W., Lu, C., Liu, J.: Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2881–2889 (2020)
31. Kahatapitiya, K., Ranasinghe, K., Park, J., Ryoo, M.S.: Language repository for long video understanding. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 5627–5646. Association for Computational Linguistics, Vienna, Austria (Jul 2025)

32. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: Eighth International Conference on Learning Representations (ICLR) (2020)
33. Khorram, S., Jiang, M., Shahbazi, M., Danesh, M.H., Fuxin, L.: Taming the tail in class-conditional gans: Knowledge sharing via unconditional training at lower resolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7580–7590 (2024)
34. Lan, X., Xue, J., Qi, J., Jiang, D., Lu, K., Chua, T.S.: Expllm: Towards chain of thought for facial expression recognition. IEEE Transactions on Multimedia **27**, 3069–3081 (2025)
35. Li, L., Zheng, Y., Liu, S., Xu, X., Li, T.: Domain knowledge enhanced vision-language pretrained model for dynamic facial expression recognition. In: Proceedings of the 32nd ACM International Conference on Multimedia. p. 5673–5682. MM '24, Association for Computing Machinery, New York, NY, USA (2024)
36. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)
37. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
38. Lin, Y., Michel, J.B., Aiden, E.L., Orwant, J., Brockman, W., Petrov, S.: Syntactic annotations for the google books ngram corpus. In: Proceedings of the ACL 2012 System Demonstrations. p. 169–174. ACL '12, Association for Computational Linguistics, USA (2012)
39. Liu, F., Gu, L., Shi, C., Fu, X.: Action unit enhance dynamic facial expression recognition. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 5597–5606 (2025)
40. Liu, J., Sun, Y., Han, C., Dou, Z., Li, W.: Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2970–2979 (2020)
41. Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P.: Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics **12**, 157–173 (2024)
42. Liu, X.Y., Wu, J., Zhou, Z.H.: Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **39**(2), 539–550 (2009)
43. Liu, Y., Dai, W., Feng, C., Wang, W., Yin, G., Zeng, J., Shan, S.: Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In: Proceedings of the 30th ACM international conference on multimedia. pp. 24–32 (2022)
44. Maaz, M., Rasheed, H., Khan, S., Khan, F.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12585–12602 (2024)
45. Matsumoto, D., Hwang, H.S.: Evidence for training the ability to read microexpressions of emotion. Motivation and emotion **35**(2), 181–191 (2011)
46. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) **54**(6), 1–35 (2021)

47. Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Team, G.B., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., et al.: Quantitative analysis of culture using millions of digitized books. science **331**(6014), 176–182 (2011)
48. Nam, G., Jang, S., Lee, J.: Decoupled training for long-tailed classification with stochastic representations. In: The Eleventh International Conference on Learning Representations (2023)
49. Parashar, S., Lin, Z., Liu, T., Dong, X., Li, Y., Ramanan, D., Caverlee, J., Kong, S.: The neglected tails in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12988–12997 (2024)
50. Qi, J., Liu, J., Tang, H., Zhu, Z.: Beyond semantics: Rediscovering spatial awareness in vision-language models. arXiv preprint arXiv:2503.17349 (2025)
51. Rangwani, H., Mopuri, K.R., Babu, R.V.: Class balancing gan with a classifier in the loop. In: Uncertainty in Artificial Intelligence. pp. 1618–1627. PMLR (2021)
52. Reed, W.J.: The pareto, zipf and other power laws. Economics letters **74**(1), 15–19 (2001)
53. Saadi, I., Hadid, A., Cunningham, D.W., Taleb-Ahmed, A., El Hillali, Y.: Pe-clip: A parameter-efficient fine-tuning of vision language models for dynamic facial expression recognition. ACM Transactions on Multimedia Computing, Communications and Applications (2025)
54. Shi, Y., Long, Q., Wu, Y., Wang, W.: Causality matters: How temporal information emerges in video language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 9006–9014 (2026)
55. Sun, L., Lian, Z., Liu, B., Tao, J.: Mae-dfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In: Proceedings of the 31st ACM International Conference on Multimedia. p. 6110–6121 (2023)
56. Sun, L., Lian, Z., Liu, B., Tao, J.: Hicmae: Hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition. Information Fusion **108**, 102382 (2024)
57. Sun, S., Lu, H., Li, J., Xie, Y., Li, T., Yang, X., Zhang, L., Yan, J.: Rethinking classifier re-training in long-tailed recognition: Label over-smooth can balance. In: The Thirteenth International Conference on Learning Representations (2025)
58. Tang, K., Huang, J., Zhang, H.: Long-tailed classification by keeping the good and removing the bad momentum causal effect. Advances in neural information processing systems **33**, 1513–1524 (2020)
59. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
60. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
61. Upadhyay, U., Ranjan, M., Shen, Z., Elhoseiny, M.: Time blindness: Why video-language models can't see what humans can. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2026)
62. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
63. Wang, P., Zhao, Z., Wen, H., Wang, F., Wang, B., Zhang, Q., Wang, Y.: Llm-autoda: Large language model-driven automatic data augmentation for long-tailed problems. Advances in Neural Information Processing Systems **37**, 64915–64941 (2024)

64. Wang, T., Zhang, J., Zheng, F., Jiang, W., Cheng, R., Luo, P.: Learning grounded vision-language representation for versatile understanding in untrimmed videos. arXiv preprint arXiv:2303.06378 (2023)
65. Wang, Y., Fei, J., Wang, H., Li, W., Bao, T., Wu, L., Zhao, R., Shen, Y.: Balancing logit variation for long-tailed semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19561–19573 (2023)
66. Wongvorachan, T., He, S., Bulut, O.: A comparison of undersampling, oversampling, and smote methods for dealing with imbalanced classification in educational data mining. Information **14**(1), 54 (2023)
67. Wu, T.H., Biamby, G., Quenum, J., Gupta, R., Gonzalez, J.E., Darrell, T., Chan, D.: Visual haystacks: A vision-centric needle-in-a-haystack benchmark. In: The Thirteenth International Conference on Learning Representations (2025)
68. Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., et al.: Qwen2. 5-omni technical report. arXiv preprint arXiv:2503.20215 (2025)
69. Zhang, H., Li, X., Bing, L.: Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In: Feng, Y., Lefever, E. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 543–553. Association for Computational Linguistics, Singapore (Dec 2023)
70. Zhang, S., Li, Z., Yan, S., He, X., Sun, J.: Distribution alignment: A unified framework for long-tail visual recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2361–2370 (2021)
71. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(9), 10795–10816 (2023)
72. Zhang, Y., Li, B., Liu, h., Lee, Y.j., Gui, L., Fu, D., Feng, J., Liu, Z., Li, C.: Llava-next: A strong zero-shot video understanding model (April 2024), https://llava-vl.github.io/blog/2024-04-30-llava-next-video/
73. Zhao, F., Tan, S., Qiu, X., Xun, L., Jiang, W., Zheng, J., Fan, H., Gao, J., Yan, D., Li, M.: Favchat: Hierarchical prompt-query guided facial video understanding with data-efficient grpo (2026)
74. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)
75. Zhu, K., Fu, M., Shao, J., Liu, T., Wu, J.: Rectify the regression bias in long-tailed object detection. In: European Conference on Computer Vision. pp. 198–214. Springer (2024)