REWEIGHTING FRAMEWISE ATTENTION IN VIDEO TRANSFORMERS FOR FACIAL EXPRESSION UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding facial expressions in unconstrained videos requires modeling subtle and localized facial dynamics. Despite substantial advances enabled by large-scale self-supervised pretraining, recent Vision Transformer (ViT)-based video models often remain biased toward dominant global motions and coarse temporal dynamics, limiting their sensitivity to fine-grained facial variations. To address this limitation, we propose MiRA (Marginal-induced Attention Redistribution), a plug-in framework for ViT backbones that enhances spatio-temporal selectivity for fine-grained facial cues without introducing additional trainable parameters. In its principled *exact mode*, it derives frame-level confidence and intra-frame concentration from post-softmax attention, capturing each frame's aggregate attention mass and its spatial concentration, respectively. Together, these quantities form a framewise importance score, which is then transformed into a prior for reallocating attention mass across frames based on global salience and within-frame localization. To further improve efficiency, we introduce *flashLite mode*, a lightweight pre-softmax approximation that derives framewise logit modulation from key-based proxy statistics and incorporates it into FlashAttention kernels while preserving the effectiveness of the exact formulation. Extensive experiments demonstrate that MiRA achieves state-of-the-art performance among video-only methods while remaining competitive with multimodal approaches.

1 INTRODUCTION

Learning effective representations of facial dynamics in videos requires capturing subtle spatio-temporal expression cues, such as transient wrinkles, frowns, and fine-grained muscle movements. These nuanced cues are often entangled with dominant global motions (*e.g.*, head movements) and extrinsic variations (*e.g.*, camera motion), making expression-relevant facial changes difficult to discern. Under such conditions, video-based facial expression understanding calls for selectively allocating attention across frames to emphasize informative expression dynamics while suppressing redundant or motion-dominated frames.

Recent advances in general video understanding have demonstrated impressive performance on video benchmarks through large-scale self-supervised pretraining and downstream finetuning. Various paradigms have been explored for video representation learning, including contrastive approaches (Han et al., 2020; Qian et al., 2021; Dave et al., 2022), self-distillation frameworks (Alwassel et al., 2020; Bardes et al., 2024), and, more recently, masked video modeling based on Masked Autoencoders (MAEs) (Wei et al., 2022; Wang et al., 2022a; Feichtenhofer et al., 2022; Tong et al., 2022; Wang et al., 2023b; Bardes et al., 2024; Bandara et al., 2023; Huang et al., 2023; Hwang et al., 2024; Pei et al., 2024; Gupta et al., 2023; Gundavarapu et al., 2024). However, these models implicitly prioritize dominant scene-level dynamics, often focusing on coarse object motions. When applied to facial videos, subtle intra-facial variations are often treated as redundant when co-occurring with more pronounced global movements. To address this limitation, we investigate how generic ViT-based video models can be adapted to better capture facial dynamics.

Despite differences in formulation, most modern video models are built upon ViTs (Vaswani et al., 2017; Dosovitskiy et al., 2021; Arnab et al., 2021), which have become the canonical backbone for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

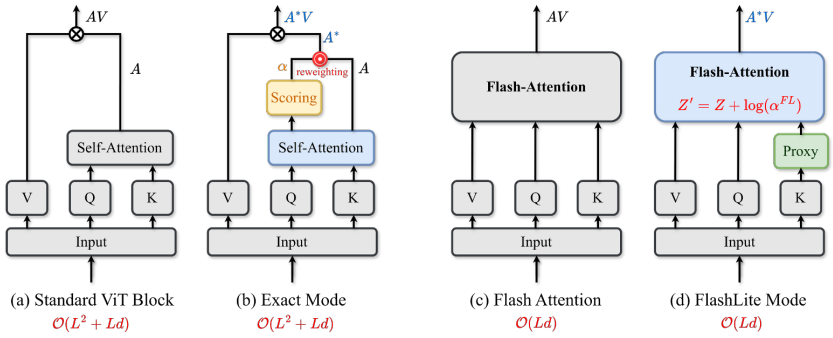


Figure 1: **Frame-marginal attention reweighting modules.** (a) Standard attention materializes the full post-softmax attention map A before computing AV. (b) *Exact mode* derives frame-level confidence and intra-frame concentration from A to guide the redistribution of attention mass across frames, requiring additional full-map passes. (c) FlashAttention computes attention in tiles without materializing the full map, reducing both activation memory and memory traffic. (d) *FlashLite mode* derives framewise logit modulation from key-based proxy statistics before attention computation and incorporates it into the fused computation in (c). (*Notation:* L: number of tokens; d: feature dimension; $Z = QK^\top/\sqrt{d_K}$: attention logits; $\mathcal{O}(\cdot)$: activation-memory complexity.)

video representation learning. Motivated by their broad applicability, we redesign the self-attention mechanism in ViTs to shift attention from dominant global motions toward fine-grained facial contexts, enabling more effective modeling of subtle facial dynamics while retaining the strengths of the original ViT backbone.

We propose MiRA, a plug-in frame-marginal attention redistribution module for ViT backbones, and first formulate its principled *exact mode*. Without relying on explicit face cropping or tight facial alignment, the module redistributes attention in facial videos according to frame-level confidence and intra-frame concentration statistics, promoting sharper and more localized attention patterns while reducing the influence of dominant global motion. Given the post-softmax self-attention maps of a ViT block, the proposed method derives two complementary frame-level statistics: a *confidence score*, measuring the aggregate attention mass assigned to each frame across the sequence, and an *intra-frame concentration score*, quantifying spatial concentration of the attention distribution via inverse entropy. These statistics are combined into a framewise importance score and transformed into a prior that guides attention redistribution toward salient and spatially localized facial dynamics.

The exact mode, while effective, incurs additional memory I/O overhead since deriving its post-softmax statistics and applying redistribution require repeated memory access to the full attention map (see fig. 1). This constraint becomes particularly restrictive in large-scale vision transformer training, where fused kernels such as FlashAttention (Dao et al., 2022; Dao, 2024) are often a practical necessity for memory- and runtime-efficient computation, owing to their tiled, on-the-fly execution that avoids full attention-map materialization. To bypass this dependency, we introduce *flashLite mode*, a pre-softmax approximation that derives framewise logit modulation from key-based proxy statistics and incorporates it into the fused attention computation. Empirically, flashLite preserves the effectiveness of the exact mode while inheriting FlashAttention's efficiency benefits.

Our contributions are summarized as follows:

- We propose MiRA, a plug-in frame-marginal attention redistribution module for ViT backbones that enhances sensitivity to subtle facial dynamics without introducing additional trainable parameters.
- We further introduce flashLite mode, a practical pre-softmax approximation of the redistribution performed by the principled exact mode, closely matching its recognition performance while inheriting FlashAttention's memory and runtime efficiency.
- We demonstrate consistent improvements over strong ViT baselines on large-scale video FER benchmarks without relying on task-specific heuristics (*e.g.*, tight facial alignment), which are often impractical in scalable pretraining and real-world video settings.

2 Related Work

Video Representation Learning. The success of ViTs in image recognition has inspired extensive efforts to extend them to video understanding tasks, where the goal is to model rich spatio-temporal dependencies across frames (Arnab et al., 2021). Early approaches adapt the vanilla ViT by incorporating temporal attention to jointly learn spatial and temporal correlations within video clips. To improve efficiency and scalability, subsequent works introduced architectural refinements, including factorized spatio-temporal attentions (Bertasius et al., 2021), locality bias through windowed attention (Liu et al., 2022c), and multiscale or pooled attention mechanisms that capture motion at different temporal resolutions (Arnab et al., 2021; Li et al., 2022c; Zhang et al., 2021; Li et al., 2022b). In parallel, masked prediction has emerged as a dominant paradigm for self-supervised video representation learning. Inspired by the success of masked language modeling (Devlin et al., 2019) and masked autoencoders in the image domain (He et al., 2022; Zhou et al., 2022; Oquab et al., 2023), these approaches train models to reconstruct missing spatio-temporal information from sparsely observed inputs (Wei et al., 2022; Wang et al., 2022a; Feichtenhofer et al., 2022; Tong et al., 2022). More recent developments enhance temporal reasoning by enforcing consistency across paired frames (Pei et al., 2024; Gupta et al., 2023), introducing adaptive and content-aware masking strategies (Wu et al., 2023; Huang et al., 2023; Bandara et al., 2023; Hwang et al., 2024), and scaling pretraining to longer input sequences to extend temporal coverage (Gundavarapu et al., 2024).

Facial Dynamics Modeling in Video. The main challenge of facial expression modeling in videos lies in capturing fine-grained intra-facial dynamics under diverse and unconstrained conditions. A growing body of work has proposed transformer-based models that directly model spatio-temporal facial dynamics, incorporating mechanisms such as temporal attention, context encoding, or landmark guidance (Ma et al., 2022; Liu et al., 2022b; Wang et al., 2023a; Li et al., 2023; Liu et al., 2025). Building on this trend, masked autoencoding has also been adapted to facial video, showing strong effectiveness for dynamic expression recognition. Recent advances include self-supervised MAE adaptations (Sun et al., 2023; Chen et al., 2024), hierarchical or contrastive extensions (Sun et al., 2024b; Zhang et al., 2024), and multimodal variants that integrate audio-visual information (Sun et al., 2024a; Chumachenko et al., 2024; Wu et al., 2025; Cheng et al., 2024). These efforts underscore a growing shift toward self-supervised and transformer-based frameworks as the dominant direction for facial expression understanding.

Despite significant progress, reliably modeling subtle facial dynamics remains challenging. General video models learn broadly transferable spatio-temporal representations, but their objectives are not specifically designed to emphasize localized facial variations. Facial expression methods, meanwhile, employ task-specific architectures, multimodal extensions of general video backbones, or facial alignment with limited scalability to suppress global motion, yet the underlying representational bias remains insufficiently addressed within the backbone itself. Bridging these lines of work, our method retains the generality and scalability of standard ViT backbones while directly adapting their attention allocation to prioritize subtle intra-facial dynamics over dominant global motion.

3 Method

We introduce MiRA, a frame-marginal attention reweighting mechanism that combines frame-level confidence and intra-frame concentration to identify informative facial moments and enhance spatio-temporal selectivity (fig. 2). We first present *exact mode*, a principled formulation that performs frame-marginal reweighting directly in post-softmax attention space, and then introduce *flashLite mode*, a lightweight approximation that realizes the underlying redistribution principle in a form compatible with FlashAttention kernels (Dao et al., 2022; Dao, 2024). Standard background on video transformers is provided in Appendix A.1.

3.1 Frame-Marginal Attention Reweighting: *Exact Mode*

Frame-Level Scoring. Let $\mathcal{I}_t = \{i_{t,n} = (t-1)N + n\}_{n=1}^{N}$ denote the set of patch-token indices corresponding to frame t in a transformer block, where $t = 1, \ldots, T$. For simplicity, we define each temporal unit as a *frame* t, where each frame corresponds to a temporal snippet at index t determined by the temporal granularity of the tokenizer. Given the head-averaged attention matrix $\bar{A} \in \mathbb{R}^{L \times L}$,

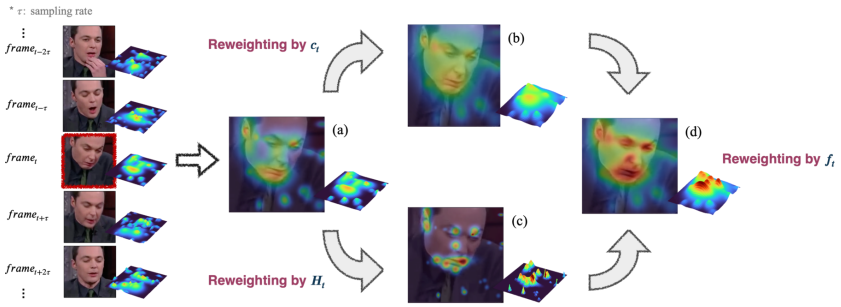


Figure 2: **Illustration of the effects of frame-level confidence and intra-frame concentration.** In-the-wild facial videos exhibit temporal redundancy and dominant global motions (e.g., head movements), which can overshadow subtle facial dynamics. (a) Baseline attention focuses on dominant scene-level variations and coarse global motions. (b) The frame-level confidence c_t emphasizes informative frames while suppressing redundancy. (c) The intra-frame concentration H_t favors frames with spatially concentrated attention responses. (d) Their combination improves spatio-temporal selectivity toward informative facial moments.

where $L = TN$, we define the *frame-level confidence score* as

$$c_t = \frac{1}{L} \sum_{q=1}^{L} \sum_{i_{t,n} \in \mathcal{I}_t} \bar{A}_{q,i_{t,n}}, \tag{1}$$

which measures the marginal attention mass assigned to frame t when treated as keys, indicating how strongly it is attended to across all queries. We further normalize c_t across frames, *i.e.*, $c_t \leftarrow c_t / \sum_{t'} c_{t'}$. To quantify the concentration of attention over tokens within each frame, we first define a normalized intra-frame attention distribution:

$$p_{t,n} = \frac{\sum_{q=1}^{L} \bar{A}_{q,i_{t,n}}}{\sum_{q=1}^{L} \sum_{j \in \mathcal{I}_t} \bar{A}_{q,j}}. \tag{2}$$

Here, $p_{t,n}$ defines the relative attention distribution over tokens within frame t. Based on this distribution, we define the *intra-frame concentration score* via inverse entropy:

$$H_t = \left(-\sum_{n=1}^{N} p_{t,n} \log p_{t,n} + \epsilon \right)^{-1}, \tag{3}$$

which assigns larger values to more spatially concentrated attention patterns, encouraging localized attention over diffuse global patterns. While c_t captures global salience, H_t reflects the spatial selectivity of attention within a frame, jointly guiding the model to identify facial expression cues ranging from global changes to subtle, region-specific dynamics.

To enable attention redistribution using these complementary signals, we define a frame-marginal importance score:

$$f_t = w_{\text{con}}\, c_t + w_{\text{ent}}\, H_t, \tag{4}$$

where w_{con} and w_{ent} are fixed weights that balance the confidence and concentration terms. To use f_t as a prior for attention reweighting, we convert it into a normalized distribution over frames. We first apply min–max normalization across frames to mitigate scale variations, and then transform the normalized scores into a probability distribution using a power-softmax:

$$f'_t = \frac{f_t - \min_{t'} f_{t'}}{\max_{t'} f_{t'} - \min_{t'} f_{t'} + \epsilon}, \quad \pi_t = (1 - \lambda_\pi) \frac{(f'_t)^\beta}{\sum_{t'} (f'_{t'})^\beta} + \frac{\lambda_\pi}{T}. \tag{5}$$

Here, β controls the sharpness of the distribution, while λ_π mixes a uniform component to prevent degenerate solutions. The resulting prior π_t defines a *frame-marginal redistribution* of attention, reallocating the frame-level aggregated attention mass across frames based on the confidence and concentration scores, thereby enhancing spatio-temporal selectivity. This aligns with the challenges in unconstrained facial videos, where attention is often biased toward dominant motions and temporal redundancy, obscuring informative facial moments.

Attention Reweighting. The resulting scaling factors are applied to the original multi-head attention A, while the scoring functions are computed from the head-averaged attention $\bar{A}$. Directly replacing A with π_t may distort learned attention patterns, as π_t encodes only frame-level marginals and does not preserve the original query-conditioned distribution over keys. To address this, we perform a ratio-based alignment that adjusts the frame-level attention distribution toward π_t while preserving the original query–key attention structure. Conceptually, this can be interpreted as a marginal-alignment rescaling, analogous to a single step of Sinkhorn normalization (Cuturi, 2013), a procedure that alternates row- and column-wise normalizations to match prescribed marginals.

We define the scaling factors as

$$\alpha_t = \texttt{clip}\left(\frac{\pi_t}{c_t + \epsilon}, \alpha_{\text{min}}, \alpha_{\text{max}}\right), \tag{6}$$

where $\texttt{clip}(\cdot)$ bounds the scaling to $[\alpha_{\min}, \alpha_{\max}]$. This scaling adjusts the frame-level attention mass by amplifying underrepresented frames ($c_t < \pi_t$) and attenuating overrepresented ones ($c_t > \pi_t$), thereby aligning the attention with the target prior π_t while reinforcing frame-level selectivity and spatial localization. Importantly, the ratio-based formulation preserves the relative query-conditioned attention distribution over key tokens within each frame while adjusting the frame-level attention allocation. Applying α_t rescales all key tokens in frame t, followed by a small uniform mixing and row-wise normalization:

$$\tilde{A}_{q,i_{t,n}} = \alpha_t A_{q,i_{t,n}}, \quad \hat{A}_{q,i} = (1 - \epsilon_{\text{floor}})\tilde{A}_{q,i} + \frac{\epsilon_{\text{floor}}}{L}, \quad A'_{q,i} = \frac{\hat{A}_{q,i}}{\sum_{j=1}^{L} \hat{A}_{q,j}}. \tag{7}$$

Here, $\tilde{A}_{q,i_{t,n}}$ denotes the frame-level rescaled attention, $\hat{A}_{q,i}$ introduces a small uniform component to prevent vanishing or overly peaky values, and $A'_{q,i}$ enforces a valid probability distribution via row-wise normalization. A final interpolation with the original attention preserves the learned query–key structure:

$$A^\star = (1 - \eta)A + \eta A', \tag{8}$$

while preventing excessive deviations from the original attention structure.

3.2 Approximation for FlashAttention: *FlashLite Mode*

While the exact mode of MiRA does not introduce additional trainable parameters, it remains computationally demanding because reweighting is applied after softmax in self-attention. This requires materializing the full post-softmax attention map $A \in \mathbb{R}^{L\times L}$ and retaining it during backpropagation. Computing frame-level statistics, applying attention rescaling, and performing row-wise normalization as in equation 7 introduce additional full-map operations, resulting in extra passes over the $\mathcal{O}(L^2)$ map and increased memory traffic and compute beyond standard attention computation. In contrast, FlashAttention (Dao et al., 2022; Dao, 2024) requires only $\mathcal{O}(L)$ additional memory through streaming attention computation: it tiles Q and K, computes blocks of $QK^\top$, applies blockwise softmax with numerically stable accumulators, and immediately contracts with V to produce $Y = AV$. This streaming design avoids materializing the full attention map, significantly reducing memory I/O. However, our exact formulation is not directly compatible with this design, as it requires access to post-softmax attention values that are not retained in the streaming computation.

To overcome this limitation, we integrate the frame-level scaling factors into the attention computation as a pre-softmax bias. Instead of rescaling the attention after normalization as in equation 7, the scaling factors are incorporated directly into the logits as biases:

$$Z'_{q,i_{t,n}} = Z_{q,i_{t,n}} + \log \alpha_t, \quad A' = \texttt{softmax}(Z'), \tag{9}$$

where $Z = QK^\top/\sqrt{d_m}$ denotes the logits for a single attention head, and d_m is the per-head query/key dimensionality. Since post-softmax scaling is followed by renormalization, multiplying an attention entry by α_t is equivalent to scaling its softmax numerator, *i.e.*, $\alpha_t \exp(z_i) = \exp(z_i + \log \alpha_t)$, and thus to adding $\log \alpha_t$ to the logits before the softmax. Unlike the exact mode, which performs redistribution directly in the attention probability space after softmax normalization, flashLite applies framewise modulation in logit space before softmax. Consequently, the two modes require different parameterizations to achieve stable framewise redistribution.

Key-Based Proxy for α_t. The scaling factors α_t in the exact mode depend on frame-level attention marginals derived from the attention map, which is not materialized in FlashAttention due to its streaming computation. To enable framewise modulation without explicit attention reconstruction, flashLite instead approximates the required frame-level statistics directly from the multi-head key representations K. Since attention logits are computed from query–key similarities, tokens with larger key energy can contribute to larger query–key logit magnitudes, providing a lightweight surrogate for estimating frame-level salience. Let M denote the number of attention heads, and $d_m = d_K/M$ the dimensionality of each head. Given the multi-head key tensor $K \in \mathbb{R}^{M\times L\times d_m}$, the per-token key energy $E \in \mathbb{R}^{T\times N}$ is defined as

$$E_{t,n} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{d_m}\left\|K_{m,i_{t,n}}\right\|_2^2, \tag{10}$$

where $E_{t,n}$ represents the average squared key energy across attention heads for token n in frame t. Using $E_{t,n}$, we compute the corresponding scores as

$$\tilde{c}_t = \sum_{n=1}^{N} E_{t,n}, \quad \tilde{p}_{t,n} = \frac{E_{t,n}}{\sum_{n'=1}^{N} E_{t,n'}}, \quad \tilde{H}_t = \left(-\sum_{n=1}^{N}\tilde{p}_{t,n}\,\log\tilde{p}_{t,n} + \epsilon\right)^{-1}, \tag{11}$$

where $\tilde{c}_t$ and $\tilde{H}_t$ approximate the attention-based confidence equation 1 and inverse entropy scores equation 3, respectively.

The two proxies are combined into the frame-marginal importance score $\tilde{f}_t$ following equation 4. Since the key-energy surrogate is computed prior to softmax normalization, its framewise statistics exhibit weaker inter-frame contrast than post-softmax attention marginals. We therefore standardize $\tilde{f}_t$ across frames and convert it into a frame prior using a temperature-controlled softmax:

$$\tilde{f}'_t = \frac{\tilde{f}_t - \mu_{\tilde{f}}}{\sigma_{\tilde{f}} + \epsilon}, \quad \pi_t^{\mathrm{FL}} = (1-\lambda_\pi)\frac{\exp(\tau\tilde{f}'_t)}{\sum_{t'}\exp(\tau\tilde{f}'_{t'})} + \frac{\lambda_\pi}{T}, \tag{12}$$

where $\mu_{\tilde{f}}$ and $\sigma_{\tilde{f}}$ denote the mean and standard deviation of $\tilde{f}_t$ over frames, while τ controls the sharpness of the resulting frame prior. Unlike the exact mode, which redistributes normalized attention marginals in probability space after softmax, flashLite performs modulation directly in logit space through additive biases. Accordingly, FlashLite employs a temperature-controlled softmax instead of the power transformation in equation 5 to compensate for the weaker inter-frame contrast of the key-based proxy. The resulting prior is converted into a unit-centered multiplicative factor, *i.e.*, $\alpha_t^{\mathrm{FL}} = \texttt{clip}(T\cdot\pi_t^{\mathrm{FL}}, \alpha_{\min}, \alpha_{\max})$. In contrast to the exact formulation, we do not use the ratio $\pi_t^{\mathrm{FL}}/\tilde{c}_t$, since both terms are derived from the same key-energy surrogate. Such a ratio would substantially suppress framewise contrast, weakening the intended corrective effect of the modulation and causing the scaling factors to collapse toward near-uniform values. Finally, these scaling factors are injected into the logits following equation 9, with α_t replaced by α_t^{FL}, enabling framewise redistribution without disrupting the streaming computation pattern of FlashAttention. Further analysis and justification of this approximation are provided in Appendix A.2.

4 Experiments

We adopt VideoMAE (Tong et al., 2022) as the backbone under the standard pretrain–finetune paradigm. During self-supervised pretraining, the model learns spatio-temporal representations via masked reconstruction, and the same ViT encoder is subsequently finetuned for downstream facial expression recognition tasks. The proposed MiRA module is integrated into each self-attention block in a plug-and-play manner without introducing additional learnable parameters. Further details on implementation and model configurations are provided in Appendices A.3 and A.4.

4.1 Datasets

For pretraining, we use VoxCeleb2 (Chung et al., 2018), a large-scale face video corpus with over 1.2M real-world utterances spanning diverse speakers and natural affective variations. For finetuning, we evaluate on large-scale in-the-wild benchmarks: DFEW (Jiang et al., 2020),

Table 1: Effect of reweighting blocks (κ).

Dataset	Metric	$\kappa = 2$	$\kappa = 6$	$\kappa = 10$	$\kappa = 12$
DFEW-f.1	UAR	62.81	62.75	61.65	**63.28**
	WAR	75.14	74.46	74.63	**75.57**
MAFW-f.1	UAR	36.74	36.55	36.62	**38.81**
	WAR	48.59	48.53	47.77	**49.40**

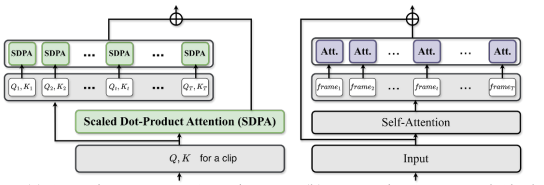
Table 2: Effect of w_{con} and w_{ent}.

		$\kappa = 2$				$\kappa = 12$			
		DFEW-f.1		MAFW-f.1		DFEW-f.1		MAFW-f.1	
w_{con}	w_{ent}	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR
0.3	0.7	62.01	**75.18**	36.05	48.42	62.06	75.05	36.50	**49.51**
0.5	0.5	**62.81**	75.14	36.74	48.59	**63.28**	75.57	**38.81**	49.40
0.7	0.3	61.10	74.97	**36.87**	**48.69**	62.31	**75.78**	37.80	48.20

Table 3: Comparison with baseline variants on DFEW fold-1.

Method	UAR	WAR
Vanilla	59.94	72.90
(a) Reused Intra Att.	59.31	72.36
(b) Framewise Aug. Block	62.68	74.93
(c) Param-Matched (b)	60.83	72.62
Ours finetuned ($\kappa = 2$)	62.81	75.14
Ours finetuned ($\kappa = 12$)	63.28	75.57
Ours pretrained	**63.88**	**76.21**

Figure 3: Illustration of baseline attention variants.



(a) Reused Intra-Frame Attention (b) Framewise Augmented Block

MAFW (Liu et al., 2022a), and FERV39k (Wang et al., 2022b). DFEW and MAFW include around 12K and 10K clips with 7 and 11 emotion categories, respectively, sourced from movies and TV shows. We follow their standard 5-fold cross-validation protocols. FERV39k consists of about 39K video clips with 7 expression categories, drawn from online videos across 22 real-world scenes.

4.2 Ablation Study

We primarily evaluate the exact mode as the canonical form to isolate the intrinsic effect of the proposed formulation, with ViT-B as the vanilla architecture. All ablations are conducted on DFEW and MAFW with a shared VideoMAE-based ViT-B backbone, pretrained on VoxCeleb2 and finetuned on the target datasets. Unless otherwise specified, results are reported on the fold-1 split, as data characteristics across folds are largely similar. Hyperparameters are initialized with default values (w_{con}=0.5, w_{ent}=0.5, β=1.5), tuned via validation, and then fixed for all experiments.

Effect of Reweighting Depth. We study the impact of applying the reweighting module at different transformer depths, as its role may vary between upper-layer refinement and full-backbone adaptation. To isolate the intrinsic effect of the proposed module, we perform ablations by finetuning a fixed VideoMAE backbone pretrained on VoxCeleb2. For a 12-layer ViT-B, we vary the number of topmost reweighting blocks (κ). As shown in table 1, full-depth reweighting (κ=12) achieves the best UAR/WAR on both DFEW and MAFW, while restricting the module to the top layers (κ=2) remains competitive with only a minor drop. In contrast, intermediate settings (κ=6 or 10) result in lower performance. We hypothesize that shallow reweighting primarily refines high-level semantics, whereas full-depth application enables coherent adaptation across the entire feature hierarchy, while intermediate depths may introduce inconsistencies across the attention hierarchy. Based on these observations, κ=12 yields the best performance, whereas κ=2 may offer an efficient trade-off.

Effect of w_{con} and w_{ent}. Under the same experimental setting, we examine the effect of weighting factors in the composite score. The frame-marginal importance score f_t in equation 4 combines frame-level confidence c_t and intra-frame concentration H_t, weighted by w_{con} and w_{ent}, respectively. We vary $[w_{\text{con}}, w_{\text{ent}}]$ to assess sensitivity. As shown in table 2, performance differences across settings are minor, with balanced weighting ($[0.5, 0.5]$) achieving near-optimal results. A sensitivity analysis of the power factor β is provided in Appendix A.5.

Beyond Naïve Framewise Modeling. We compare against three baseline variants that enhance framewise or temporal modeling: (a) *reused intra-frame attention*, adding a branch that confines attention within each frame and combines it with global attention; (b) *framewise augmented block* (+20% params), which augments the transformer with a framewise block operating independently within each frame; and (c) *parameter-matched version of (b)*, where global head dimensions are reduced to match the parameter budget of the vanilla model, as illustrated in fig. 3. To ensure a fair comparison, all baselines are trained consistently from pretraining on VoxCeleb2 through finetuning

Table 4: Performance comparison of Exact and FlashLite on ViT-B over 5 folds.

DFEW	Fold1		Fold2		Fold3		Fold4		Fold5	
	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR
Exact	63.88	76.21	64.32	75.10	65.67	75.64	64.28	74.35	68.28	78.00
FlashLite	63.77	76.51	64.16	74.93	65.06	76.15	64.87	75.63	67.69	78.04

MAFW	Fold1		Fold2		Fold3		Fold4		Fold5	
	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR
Exact	39.67	50.87	42.44	55.63	49.69	62.19	48.36	63.37	46.02	61.39
FlashLite	39.20	51.20	43.65	55.53	48.81	61.54	49.83	63.48	45.91	60.13

Table 5: Comparison with general video models. (**Best**, Second)

Method	DFEW-f.1		MAFW-f.1	
	UAR	WAR	UAR	WAR
VideoMAE (Tong et al., 2022)	59.94	72.90	38.80	48.80
VideoMAEv2 (Wang et al., 2023b)	59.16	73.56	37.40	49.51
AdaMAE (Bandara et al., 2023)	61.73	75.31	37.59	50.27
MGMAE (Huang et al., 2023)	60.95	75.01	37.68	50.27
EVEREST (Hwang et al., 2024)	54.83	66.85	36.26	47.77
ViT-B Exact	**63.88**	76.21	**39.67**	50.87
ViT-B FlashLite	63.77	**76.51**	39.20	**51.20**

on the target dataset. In table 3, "finetuned" denotes applying MiRA only during downstream finetuning, whereas "pretrained" denotes applying it during both pretraining and finetuning. Our models outperform all baselines, suggesting that naive framewise enhancements alone do not fully account for the observed gains. Facial expressions are often temporally sparse and spatially localized, requiring the model to focus on informative moments. By jointly leveraging frame-level confidence and intra-frame concentration, our approach demonstrates improved capacity to model these dynamics.

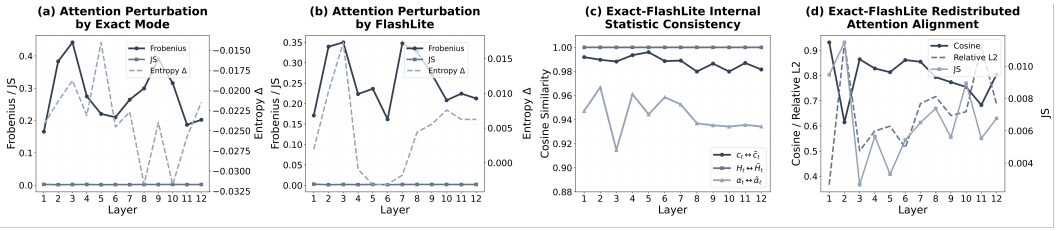


Figure 4: Behavioral analysis of Exact and FlashLite modes. (a,b) Attention perturbation relative to vanilla attention, measured by Frobenius distance on attention maps, together with JS divergence and entropy difference on framewise temporal attention distributions. (c) Correspondence between the exact mode statistics (c_t, H_t, α_t) and their flashLite counterparts $(\tilde{c}_t, \tilde{H}_t, \alpha_t^{\mathrm{FL}})$. (d) Alignment between redistributed attentions, measured using cosine similarity (directional agreement) and relative ℓ_2 distance (relative deviation magnitude).

Exact Mode vs. FlashLite Mode. We conduct layer-wise analyses of attention perturbation, internal statistics, and redistributed attention alignment for both modes. In fig. 4(a,b), each mode is compared against vanilla attention to examine how modulation alters transformer attention across layers. Frobenius distance $\|A_{\mathrm{reweighted}} - A_{\mathrm{vanilla}}\|_F$ measures attention perturbation magnitude, indicating that both exact and flashLite consistently modify transformer attention relative to vanilla attention. JS divergence is computed between temporal attention distributions p_{vanilla} and $p_{\mathrm{reweighted}}$, obtained by marginalizing attention over spatial tokens within each frame. Its consistently low values indicate that the global temporal allocation pattern remains largely preserved after modulation. Likewise, entropy differences $\mathcal{H}(p_{\mathrm{reweighted}}) - \mathcal{H}(p_{\mathrm{vanilla}})$ are computed on the same temporal attention distributions. Their consistently small magnitudes suggest only modest variations in concentration behavior. Overall, these analyses suggest that both exact and flashLite introduce structured redistribution while preserving the original temporal attention organization. Figure 4(c) compares the internal statistics used for modulation, showing that flashLite closely preserves the internal modulation dynamics of the exact formulation. Figure 4(d) directly compares redistributed attentions between the two modes, indicating that redistribution derived from key-space statistics closely approximates that of the exact formulation. Consistent with these analyses, exact and flashLite exhibit nearly identical recognition performance across all folds on DFEW and MAFW (Table 4). Together with this close approximation, flashLite provides a latency reduction of 20% (400.6 ms $\rightarrow$ 318.7 ms), corresponding to a throughput gain of about 26% on a single H100 GPU in practice. Detailed computational analysis is provided in Appendix A.6.

4.3 Main Results

Comparison with General Video Models. All models are pretrained on VoxCeleb2 under identical settings and evaluated on the fold-1 splits of DFEW and MAFW. Building upon VideoMAE, VideoMAEv2 focuses on scaling to larger models by improving efficiency through sparse masking. AdaMAE employs a classification-based sampler to prioritize informative tokens, MGMAE

Table 6: Comparison with state-of-the-art methods on DFEW, MAFW, and FERV39k. Results on DFEW and MAFW are averaged over 5 folds. See Appendix A.7 for an extended comparison and class-wise confusion matrices. (**Best** / Second in video-only methods. Pretrain: **U**nsupervised and **S**upervised. Modality: **V**ideo, **I**mage, **A**udio, and **L**anguage.)

Method	Pretrain	Modality	Face Align.	Params (M)	DFEW		MAFW		FERV39k	
					UAR	WAR	UAR	WAR	UAR	WAR
M3DFEL (Wang et al., 2023a)	S	V			56.10	69.25	–	–	35.94	47.67
Former-DFER (Zhao & Liu, 2021)	S	V		18	53.69	65.70	31.16	43.27	37.20	46.85
NR-DFERNet (Li et al., 2022a)	S	V			54.21	68.19	–	–	33.99	45.97
STT (Ma et al., 2022)	S	V			54.58	66.65	–	–	37.76	48.11
LOGO-Former (Ma et al., 2023)	S	V			54.21	66.98	–	–	38.22	48.13
MoCo (He et al., 2020)	U	V		32	53.47	67.45	–	–	–	–
VideoMAE* (Tong et al., 2022)	U	V	✓	86	63.60	74.60	40.87	53.51	43.33	52.39
Video Swin-B (Liu et al., 2022c)	S	V		88	59.38	71.90	–	–	–	–
MAE-DFER (Sun et al., 2023)	U	V	✓	85	63.41	74.43	41.62	54.31	43.12	52.07
SVFAP (Sun et al., 2024b)	U	V	✓	78	62.83	74.27	41.19	54.28	42.14	52.29
ViT-B Exact	U	V	×	86	65.29	75.86	45.24	58.69	–	–
ViT-B FlashLite	U	V	×	86	65.11	76.25	45.48	58.38	43.81	53.80
ViT-L FlashLite	U	V	×	305	67.75	77.32	46.80	60.04	44.41	54.62
ViT-H FlashLite	U	V	×	633	**68.25**	**78.24**	**48.22**	**61.45**	**45.12**	**55.64**
MMA-DFER (Chumachenko et al., 2024)	U	V+A	✓		66.85	77.43	44.25	58.45	–	–
HiCMAE (Sun et al., 2024a)	U	V+A	✓	81	63.76	75.01	42.65	56.17	–	–
AVF-MAE++-B (Wu et al., 2025)	U+S	V+A	✓	169	63.74	75.42	43.10	57.50	–	–
AVF-MAE++-L (Wu et al., 2025)	U+S	V+A	✓	303	65.14	76.24	45.36	59.13	–	–
AVF-MAE++-H (Wu et al., 2025)	U+S	V+A	✓	521	66.88	77.45	46.05	60.24	–	–
Emotion-LLaMA (Cheng et al., 2024)		V+A+L			64.21	77.06	–	–	–	–
EmoThinker (Xu et al., 2026)		V+A+L			65.23	78.13	–	–	–	–

enforces temporal consistency via motion-guided masking, and EVEREST reduces redundancy by focusing on information-dense frames, all aiming to enhance spatio-temporal token correlation. As shown in table 5, our approach outperforms these baselines, demonstrating the effectiveness of frame-marginal modulation for characterizing fine-grained facial dynamics.

Performance on Facial Expression Recognition. As shown in table 6, MiRA achieves state-of-the-art performance within video-only methods and remains competitive with multimodal and hybrid pretraining approaches. Audio–visual models rely on more extensive audio–visual pretraining, often with supervised objectives. Other multimodal approaches leverage textual supervision or pretrained language knowledge to enrich facial semantics. Despite relying solely on video input, our results suggest that effective modeling of facial dynamics remains critical for representing affective context. This aligns with our modeling strategy, which is designed to reduce the influence of nuisance global motion and temporal redundancy while enhancing sensitivity to localized intra-facial cues. This, in turn, allows our approach to operate without task-specific preprocessing, unlike prior works that rely on heuristics (*e.g.*, tight facial cropping or alignment) to constrain the model to intra-facial regions. This alignment-free design is applicable to large-scale and diverse data for which tight facial alignment is impractical. Finally, by improving computational efficiency, flashLite enables scaling to larger models, allowing us to extend from ViT-B to ViT-L and ViT-H and consistently obtain gains as model capacity increases. We further assess the pretrained representations through frozen-feature probing on micro-expression datasets and qualitative analysis of learned attention allocation, as detailed in Appendices A.8 and A.9, respectively.

5 Conclusion

This work investigates learning facial dynamics from videos in scalable settings, where subtle expression cues are often obscured by temporally redundant content and dominant global motion. We revisit self-attention and introduce a frame-marginal attention redistribution mechanism that promotes selective and localized spatio-temporal attention, together with an efficient approximation compatible with large-scale training. Across both pretraining and downstream evaluations, the proposed approach consistently improves performance while preserving computational efficiency. Future work will explore integrating frame-marginal attention redistribution into emerging video foundation models and multimodal pretraining frameworks.

AI USE STATEMENT

ChatGPT (OpenAI) was used to assist with language editing, literature searches, and clarity checks for the mathematical exposition, experimental presentation, and schematic figures. The core method, implementation, and experiments were developed and conducted by the authors. ChatGPT was used only to review the authors' implementation from a usability perspective and to improve its user-facing documentation. It was not used to collect or generate datasets, execute experiments, or produce experimental measurements. All AI-assisted suggestions were critically reviewed and independently verified by the authors, who take full responsibility for the final content of this paper.

ETHICS STATEMENT

This work studies facial expression recognition using existing public research benchmarks and does not collect or introduce new human-subject data. Observed facial expressions do not necessarily provide a complete or reliable indication of an individual's internal emotional state, and models trained on existing datasets may reflect demographic, cultural, contextual, and annotation biases. Such systems may also raise privacy concerns or be misused in sensitive decision-making contexts. The results should therefore be interpreted within the scope of the evaluated research benchmarks and should not be used as a basis for inferring individuals' mental states in high-stakes applications.

REPRODUCIBILITY STATEMENT

The Exact and FlashLite formulations are described in Section 3 and Algorithm 1. Implementation details, model configurations, training settings, and computational evaluation protocols are provided in Appendix A. The code and pretrained models will be made publicly available after the review process.

REFERENCES

Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Adv. Neural Inform. Process. Syst.*, 33:9758–9770, 2020.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Int. Conf. Comput. Vis.*, 2021.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL `https://arxiv.org/abs/2502.13923`.

Wele Gedara Chaminda Bandara, Naman Patel, Ali Gholami, Mehdi Nikkhah, Motilal Agrawal, and Vishal M Patel. Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 14507–14517, 2023.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024.

Xianye Ben, Yi Ren, Junping Zhang, Su-Jing Wang, Kidiyo Kpalma, Weixiao Meng, and Yong-Jin Liu. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9):5826–5846, 2021.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Int. Conf. Mach. Learn.*, 2021.

Ashutosh Chaubey, Jiacheng Pang, Maksim Siniukov, and Mohammad Soleymani. AVERE: Improving audiovisual emotion reasoning with preference optimization. In *Int. Conf. Learn. Represent.*, 2026.

Yin Chen, Jia Li, Yu Zhang, Zhenzhen Hu, Shiguang Shan, Meng Wang, and Richang Hong. Unilearn: Enhancing dynamic facial expression recognition through unified pre-training and fine-tuning on images and videos. *arXiv e-prints*, pp. arXiv:2409.06154v1, 2024.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander G Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Adv. Neural Inform. Process. Syst.*, 37:110805–110853, 2024.

Kateryna Chumachenko, Alexandros Iosifidis, and Moncef Gabbouj. Mma-dfer: Multimodal adaptation of unimodal models for dynamic facial expression recognition in-the-wild. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4673–4682, 2024.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep Speaker Recognition. In *Interspeech 2018*, pp. 1086–1090, 2018. doi: 10.21437/Interspeech.2018-1929.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Adv. Neural Inform. Process. Syst.*, 26, 2013.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *Int. Conf. Learn. Represent.*, 2024.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Adv. Neural Inform. Process. Syst.*, 35:16344–16359, 2022.

Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, 219:103406, 2022.

Adrian K Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. Samm: A spontaneous micro-facial movement dataset. *IEEE transactions on affective computing*, 9(1): 116–129, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent.*, 2021.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Int. Conf. Comput. Vis.* IEEE, 2021.

Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *Adv. Neural Inform. Process. Syst.*, 35:35946–35958, 2022.

Niki Maria Foteinopoulou and Ioannis Patras. Emoclip: A vision-language method for zero-shot video facial expression recognition. In *2024 IEEE 18th international conference on automatic face and gesture recognition (FG)*, pp. 1–10. IEEE, 2024.

Nitesh B Gundavarapu, Luke Friedman, Raghav Goyal, Chaitra Hegde, Eirikur Agustsson, Sagar Waghmare, Mikhail Sirotenko, Ming-Hsuan Yang, Tobias Weyand, Boqing Gong, et al. Extending video masked autoencoders to 128 frames. *Adv. Neural Inform. Process. Syst.*, 37:121376–121400, 2024.

Agrim Gupta, Jiajun Wu, Jia Deng, and Fei-Fei Li. Siamese masked autoencoders. *Adv. Neural Inform. Process. Syst.*, 36:40676–40693, 2023.

Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Adv. Neural Inform. Process. Syst.*, 33:5679–5690, 2020.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 6546–6555, 2018.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

Bingkun Huang, Zhiyu Zhao, Guozhen Zhang, Yu Qiao, and Limin Wang. Mgmae: Motion guided masking for video masked autoencoding. In *Int. Conf. Comput. Vis.*, 2023.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Sunil Hwang, Jaehong Yoon, Youngwan Lee, and Sung Ju Hwang. Everest: Efficient masked video autoencoder by removing redundant spatiotemporal tokens. In *Int. Conf. Mach. Learn.*, 2024.

Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. 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.

Hanting Li, Mingzhe Sui, Zhaoqing Zhu, and Feng Zhao. Nr-dfernet: Noise-robust network for dynamic facial expression recognition. *arXiv preprint arXiv:2206.04975*, 2022a.

Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Intensity-aware loss for dynamic facial expression recognition in the wild. In *AAAI*, pp. 67–75, 2023.

Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Cliper: A unified vision-language framework for in-the-wild facial expression recognition. In *2024 IEEE international conference on multimedia and expo (ICME)*, pp. 1–6. IEEE, 2024.

Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. UniFormer: Unified transformer for efficient spatial-temporal representation learning. In *Int. Conf. Learn. Represent.*, 2022b.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022c.

Feng Liu, Hanyang Wang, and Siyuan Shen. Robust dynamic facial expression recognition. *IEEE transactions on biometrics, behavior, and identity science*, 7(4):563–572, 2025.

Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. 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, 2022a.

Yuanyuan Liu, Chuanxu Feng, Xiaohui Yuan, Lin Zhou, Wenbin Wang, Jie Qin, and Zhongwen Luo. Clip-aware expressive feature learning for video-based facial expression recognition. *Information Sciences*, 598:182–195, 2022b.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE, 2022c.

Fuyan Ma, Bin Sun, and Shutao Li. Spatio-temporal transformer for dynamic facial expression recognition in the wild. *arXiv preprint arXiv:2205.04749*, 2022.

Fuyan Ma, Bin Sun, and Shutao Li. Logo-former: Local-global spatio-temporal transformer for dynamic facial expression recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Gensheng Pei, Tao Chen, Xiruo Jiang, Huafeng Liu, Zeren Sun, and Yazhou Yao. Videomac: Video masked autoencoders meet convnets. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 22733–22743, 2024.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

Pritam Sarkar, Aaron Posen, and Ali Etemad. Avcaffe: A large scale audio-visual dataset of cognitive load and affect for remote work. In *AAAI*, pp. 76–85, 2023.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Mae-dfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 6110–6121, 2023.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Hicmae: Hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition. *Information Fusion*, 108:102382, 2024a.

Licai Sun, Zheng Lian, Kexin Wang, Yu He, Mingyu Xu, Haiyang Sun, Bin Liu, and Jianhua Tao. Svfap: Self-supervised video facial affect perceiver. *IEEE Transactions on Affective Computing*, 16(1):405–422, 2024b.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Adv. Neural Inform. Process. Syst.*, 30, 2017.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Adv. Neural Inform. Process. Syst.*, 35: 10078–10093, 2022.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Int. Conf. Comput. Vis.*, pp. 4489–4497, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Adv. Neural Inform. Process. Syst.*, 30, 2017.

Hanyang Wang, Bo Li, Shuang Wu, Siyuan Shen, Feng Liu, Shouhong Ding, and Aimin Zhou. Rethinking the learning paradigm for dynamic facial expression recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 17958–17968, 2023a.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 14549–14560, 2023b.

Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022a.

Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022b.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

Qiangqiang Wu, Tianyu Yang, Ziquan Liu, Baoyuan Wu, Ying Shan, and Antoni B Chan. Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 14561–14571, 2023.

Xuecheng Wu, Heli Sun, Yifan Wang, Jiayu Nie, Jie Zhang, Yabing Wang, Junxiao Xue, and Liang He. Avf-mae++: Scaling affective video facial masked autoencoders via efficient audio-visual self-supervised learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 9142–9153, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

Qinfu Xu, Liyuan Pan, Yiwei Wei, Shaozu Yuan, Jiaqi Chen, and Tianyu Liu. Emothinker: Advancing visual-acoustic emotion analysis via structural token selection and chain-of-thought reasoning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 1672–1682, 2026.

Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions. In *Int. Conf. Comput. Vis.*, 2021.

Zhicheng Zhang, Pancheng Zhao, Eunil Park, and Jufeng Yang. Mart: Masked affective representation learning via masked temporal distribution distillation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 12830–12840, 2024.

Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 1553–1561, 2021.

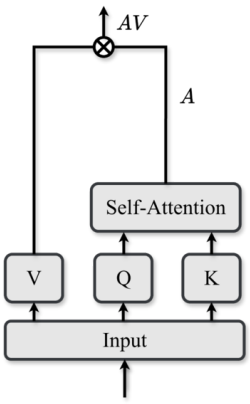
Zengqun Zhao and Ioannis Patras. Prompting visual-language models for dynamic facial expression recognition. *arXiv preprint arXiv:2308.13382*, 2023.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. In *Int. Conf. Learn. Represent.*, 2022.

A Appendix

A.1 Preliminary on Video Transformers

In recent ViT-based video representation learning methods, an input video is divided into non-overlapping tubelet patches (Arnab et al., 2021). Let T, N_h, and N_w denote the numbers of splits along the temporal, height, and width dimensions, respectively. The total number of tubelets is then $L = T \times N$, where $N = N_h \times N_w$. Each tubelet is embedded into a vector $x \in \mathbb{R}^d$ via a patch embedding module, and stacking all embeddings forms $X \in \mathbb{R}^{L\times d}$. The embedding sequence X is then processed by a stack of transformer blocks, as illustrated in fig. 5, to encode spatio-temporal representations. Within each block, consider a single head of multi-head self-attention. Let $d_m = d_K/M$ denote the per-head query/key dimension, where M is the number of attention heads. The queries, keys, and values for this head are defined as $Q = XW_Q \in \mathbb{R}^{L\times d_m}$, $K = XW_K \in \mathbb{R}^{L\times d_m}$, and $V = XW_V \in \mathbb{R}^{L\times d_v}$, where d_v denotes the per-head value dimension. The scaled dot-product attention is computed as $A = \texttt{softmax}\big(QK^\top/\sqrt{d_m}\big) \in \mathbb{R}^{L\times L}$, where $A_{q,k}$ denotes the (q, k)-th element in query–key order. The output of this head is obtained as $Y = AV$. The same self-attention operation is applied across all tubelet tokens, with frame-level attention allocation emerging implicitly from content-dependent query–key affinities and positional encodings. Note that T does not necessarily correspond to the original number of video frames, as it depends on the temporal stride of the tokenizer that constructs each tubelet (*e.g.*, a single tubelet may span multiple consecutive frames) (Arnab et al., 2021; Fan et al., 2021; Liu et al., 2022c).



(a) Standard ViT Block

Figure 5: **Transformer block.**

A.2 Approximation Rationale for FlashLite Mode

In self-attention, query–key logits depend on the magnitudes and directional alignment of the query and key vectors. Although key magnitude alone does not determine the resulting attention mass, the frame-sum energy defined in equation 10–equation 11 provides a query-independent proxy for the potential salience of each frame and thus serves as a practical surrogate for the attention-derived confidence c_t. Similarly, the inverse entropy of the normalized key-energy distribution, $\tilde{H}_t$, serves as a proxy for whether the spatial energy within a frame is diffuse or concentrated, mirroring the role of H_t. Using these proxy statistics, flashLite constructs framewise scaling factors α_t^{FL} without requiring access to the attention map A. For fixed positive scaling factors and in the absence of additional smoothing or interpolation, post-softmax scaling followed by row-wise normalization is algebraically equivalent to adding the corresponding log-biases before the softmax. Therefore, the approximation in flashLite does not arise from moving the modulation from post-softmax attention to pre-softmax logits. Instead, flashLite differs from the exact formulation because its scaling factors are estimated from key-based proxy statistics and constructed using a different prior and scaling parameterization from the attention-derived ratio used in the exact mode. Because the proxy statistics are aggregated across attention heads, the resulting framewise factors are shared across heads, consistent with the head-aggregated design of the exact mode. Finally, applying the modulation in logit space preserves the fused streaming pipeline of FlashAttention and avoids materializing the $L \times L$ attention map, thereby retaining its activation-memory and memory-traffic benefits during both training and inference.

A.3 Implementation Details

Backbone Context. We adopt VideoMAE (Tong et al., 2022) as the baseline under the pretrain–finetune paradigm. It learns video representations via self-supervised pretraining on unlabeled videos and adapts the same ViT backbone to downstream tasks by supervised finetuning. In pretraining, the model masks 90% of spatio-temporal tubelet patches and reconstructs the masked tokens from the visible subset, using a frame-shared mask to preserve spatial alignment and maintain consistent frame-wise token indexing. For classification, we finetune the encoder with a lightweight task head. We integrate MiRA into each self-attention block as a drop-in module that performs at-

tention redistribution and framewise modulation; it adds no learnable parameters and preserves the original transformer pipeline while adaptively reweighting attention.

Stabilizing Confidence Score with EMA. Compared to the inverse entropy H_t, the confidence score c_t tends to exhibit higher variance, as it aggregates raw attention magnitudes across all queries. As a result, its value may fluctuate significantly depending on batch composition. To mitigate this variability during pretraining, we maintain a dataset-level prior by applying an exponential moving average (EMA) (Polyak & Juditsky, 1992; Tarvainen & Valpola, 2017) to the batch-averaged confidence values:

$$\bar{c}_t \leftarrow m\,\bar{c}_t + (1-m)\,c_t^{\text{batch}}, \quad m \in [0,1), \tag{13}$$

where m denotes the momentum coefficient controlling the trade-off between the historical estimate $\bar{c}_t$ and the current batch statistic c_t^{batch}. This EMA-based smoothing provides a more stable estimate of frame-level importance across training iterations, which is particularly beneficial during large-scale pretraining. In contrast, during finetuning, we use instance-level statistics without EMA, allowing the model to adapt more flexibly to sample-specific dynamics and task-dependent variations.

FlashLite Temperature. For flashLite, the temperature parameter τ in equation 12 is fixed to 1.7 across all datasets and model scales. The parameter controls the sharpness of the frame prior distribution derived from the standardized frame importance scores. Larger values produce more selective framewise modulation by emphasizing high-response frames, whereas smaller values yield smoother redistribution across frames. We empirically found $\tau = 1.7$ to provide a stable balance between selective emphasis and optimization stability.

Numerical Safeguards. To ensure numerical stability, normalized probabilities are clamped from below at 10^{-12} when computing entropy, and $\epsilon = 10^{-6}$ is added to the inverse-entropy denominators in equation 3 and equation 11. A smoothing factor $\lambda_\pi = 0.1$ in equation 5 and equation 12 mixes a fraction of the uniform distribution into each prior to avoid frame collapse. Additionally, in the exact mode, a uniform floor mass ($\epsilon_{\text{floor}} = 10^{-3}$) is added before row-wise renormalization to prevent query distributions from collapsing to zero.

Exact vs. FlashLite Modes. We summarize the practical differences between the exact and flashLite modes, as implemented in algorithm 1. Exact mode operates on the explicitly materialized attention map $A \in \mathbb{R}^{B\times M\times L\times L}$, from which the head-averaged attention $\bar{A}$ is used to compute the frame-level confidence c_t and intra-frame concentration H_t. The composite prior π_t is then constructed, and a ratio-based alignment is applied to adjust the frame-level marginal distribution. This reweighting is performed after softmax, followed by uniform smoothing, row-wise renormalization, and residual interpolation with the original attention.

FlashLite mode, in contrast, avoids materializing A by replacing attention-derived statistics with key-based proxy statistics. Specifically, the proxy scores $(\tilde{c}_t, \tilde{H}_t)$ are computed from the key-energy tensor $E \in \mathbb{R}^{B\times T\times N}$, following the same confidence–concentration formulation used in exact mode. The resulting prior π_t^{FL} is converted into frame-level scaling factors, which are injected as additive log-biases into the attention logits. This enables framewise modulation to be incorporated into the pre-softmax computation, thereby preserving compatibility with fused FlashAttention kernels.

The two modes differ in where the modulation is applied within the attention pipeline. Exact mode performs *post-softmax* attention redistribution with explicit row-wise renormalization, whereas FlashLite applies a corresponding *pre-softmax* modulation via logit bias. Despite this difference, both modes share the same confidence–concentration principle while differing in the source of statistics, the prior and scaling parameterizations, and the point at which modulation is applied.

A.4 Model Configurations

We follow a two-stage training pipeline consisting of large-scale self-supervised pretraining on VoxCeleb2 and supervised finetuning for downstream facial emotion recognition. Most hyperparameters are kept consistent with the VideoMAE framework (Tong et al., 2022) to ensure fair comparison with established baselines. A detailed summary of training configurations is provided in table 7.

Algorithm 1 Pseudocode of **Exact** and **FlashLite** modes in a transformer block.

Exact mode

```
X: video tokens [B,L,d], L=T*N
# B: batch, T: frames, N: tokens/frame
# d: embed dim, Hh: heads, dh: head dim

q,k,v = QKV(X)                          # [B,Hh,L,dh]

A_orig = Softmax(q @ k.T/sqrt(dh))      # [B,Hh,L,L]
A_bar = MeanOverHeads(A_orig)           # [B,L,L]

c = FrameConfidence(A_bar)              # [B,T]
p = FrameTokenDistribution(A_bar)       # [B,T,N]
H = IntraFrameConcentration(p)          # [B,T]

c_score = c                             # [B,T]
if pretrain:
    c_score = EMA(c)                    # [B,T]

f = w_con*c_score + w_ent*H             # [B,T]
f = MinMaxNorm(f)                       # [B,T]
pi = PowerSoftmax(f,beta)               # [B,T]
pi = ((1-lambda_pi)*pi + lambda_pi/T)
                                        # [B,T]

alpha = clip(pi/(c_score+eps),alpha_min,alpha_max)
                                        # [B,T]

A_update = ScaleKeysByFrame(A_orig,alpha)
                                        # [B,Hh,L,L]
A_update = ((1-eps_floor)*A_update + eps_floor/L)
                                        # [B,Hh,L,L]
A_update = RowNormalize(A_update)       # [B,Hh,L,L]

A_final = ((1-eta)*A_orig + eta*A_update)
                                        # [B,Hh,L,L]

O = A_final @ v                         # [B,Hh,L,dh]
Y = OutputProjection(O)                 # [B,L,d]
```

FlashLite mode

```
X: video tokens [B,L,d], L=T*N
# B: batch, T: frames, N: tokens/frame
# d: embed dim, Hh: heads, dh: head dim

q,k,v = QKV(X)                          # [B,Hh,L,dh]

E = HeadAveragedKeyEnergy(k)            # [B,T,N]

c_tilde = FrameConfidence(E)            # [B,T]
p_tilde = FrameTokenDistribution(E)     # [B,T,N]
H_tilde = IntraFrameConcentration(p_tilde)
                                        # [B,T]

c_score = c_tilde                       # [B,T]
if pretrain:
    c_score = EMA(c_tilde)              # [B,T]

f_tilde = (w_con*c_score + w_ent*H_tilde)
                                        # [B,T]
f_tilde = Standardize(f_tilde)          # [B,T]
pi_FL = Softmax(tau*f_tilde)            # [B,T]
pi_FL = ((1-lambda_pi)*pi_FL + lambda_pi/T)
                                        # [B,T]

alpha_FL = clip(T*pi_FL,alpha_min,alpha_max)
                                        # [B,T]

O = FlashAttention(q,k,v,frame_bias=log(alpha_FL))
                                        # [B,Hh,L,dh]
Y = OutputProjection(O)                 # [B,L,d]
```

Backbone Architecture. We adopt ViT-B/16, ViT-L/16, and ViT-H/16 as the backbone, following the standard VideoMAE (Tong et al., 2022) configuration. Videos are partitioned into non-overlapping $2 \times 16 \times 16$ spatio-temporal tubelets and linearly projected into a sequence of tokens. The tokens are processed by a ViT encoder with multi-head self-attention, while a lightweight decoder is used only during pretraining for masked reconstruction.

Pretraining. We pretrain all models on VoxCeleb2 using a masked video modeling objective with a tube masking ratio of 0.9. Each input clip consists of 16 RGB frames at 224×224 resolution, sampled every 4 frames. Training is conducted on NVIDIA H100 80GB GPUs, using 32 GPUs for ViT-B and 40 GPUs for both ViT-L and ViT-H. We adopt relatively lightweight data augmentation, consisting of MultiScaleCrop with scale factors $\{1.0, 0.875, 0.75, 0.66\}$ and standard normalization, following the original VideoMAE recipe. This design emphasizes representation learning from masked reconstruction rather than aggressive input perturbation.

Finetuning. For downstream tasks, the encoder is initialized from the 200-epoch self-supervised checkpoint. The pretraining decoder is removed, and a mean-pooling classification head is applied over the token embeddings. Finetuning is typically performed on a single GPU with a batch size of 16 for 100 epochs using AdamW with layer-wise learning rate decay, while minor adjustments to the batch size or number of GPUs are made for larger datasets to accommodate practical runtime constraints. In contrast to pretraining, we employ stronger data augmentation, including RandAugment, color jitter, random erasing, and mixup/cutmix regularization, to improve generalization. For evaluation, we follow the standard multi-view testing protocol used in video classification. Each video is uniformly divided into multiple temporal segments, and multiple spatial crops are extracted per segment. The model processes all segment–crop pairs independently, and the prediction logits are averaged to produce the final video-level output.

Table 7: Training configurations for self-supervised pretraining and supervised finetuning across ViT-B/16, ViT-L/16, and ViT-H/16 backbones.

Configurations	**Pretraining**	**Finetuning**
backbone	ViT-B/16, ViT-L/16, ViT-H/16	ViT-B/16, ViT-L/16, ViT-H/16
dataset	VoxCeleb2	target datasets
input resolution	224×224	224×224
frames per clip	16	16
sampling rate	4	4
tubelet size	$2 \times 16 \times 16$	$2 \times 16 \times 16$
masking ratio	0.9	–
ViT decoder	4-layer / 8-layer / 8-layer	–
pooling	–	mean pooling over tokens
global batch size	128×32 / 64×40 / 64×40	16×1
optimizer	AdamW	AdamW
base learning rate	1.5e−4 / 7e−5 / 8e−5	1e−3
min learning rate	1e−5	1e−6
optimizer momentum	$\beta_1 = 0.9,\ \beta_2 = 0.95$	$\beta_1 = 0.9,\ \beta_2 = 0.999$
weight decay	0.05	0.05
layer-wise decay	–	0.75
lr schedule	cosine decay	cosine decay
warmup epochs	20	5
training epochs	200	100 (best epoch selected)
drop path	–	0.1
label smoothing	–	0.1
augmentation	MultiScaleCrop, normalization	RandAug, color jitter 0.4, random erasing 0.25, mixup 0.8, cutmix 1.0

A.5 Sensitivity to the Power Factor β

In equation 5, β controls the sharpness of the transformed importance score distribution before normalization, thereby determining how strongly the prior emphasizes salient frames. A larger β increases the contrast among frame-level importance scores f_t, whereas a smaller β compresses these differences, producing a more uniform prior π_t. To assess its influence, we first evaluate three representative values (0.7, 1.0, 1.5), as shown in table 8. β=1.5 generally provides the strongest performance across datasets and configurations (κ=2 and κ=12). Guided by this observation, we further examine values in the neighborhood of β=1.5, where performance remains consistently strong with only marginal fluctuations across settings.

Table 8: Effect of the power factor β under different κ settings. (**Best**, <u>Second</u>)

	$\kappa = 2$				$\kappa = 12$			
	DFEW-f.1		**MAFW-f.1**		**DFEW-f.1**		**MAFW-f.1**	
	UAR	WAR	UAR	WAR	UAR	WAR	UAR	WAR
$\beta = 0.7$	55.87	68.56	36.38	48.37	54.03	66.98	37.97	49.29
$\beta = 1.0$	56.62	68.94	36.45	48.26	55.01	67.88	<u>38.53</u>	**49.62**
$\beta = 1.4$	61.66	<u>75.18</u>	35.92	48.42	62.51	75.14	38.42	49.29
$\beta = 1.5$	**62.81**	75.14	**36.74**	**48.59**	**63.28**	**75.57**	**38.81**	49.40
$\beta = 1.6$	<u>61.93</u>	**75.48**	<u>36.68</u>	<u>48.53</u>	<u>62.85</u>	75.01	36.78	<u>49.51</u>
$\beta = 1.7$	61.81	<u>75.18</u>	36.31	48.26	62.66	<u>75.18</u>	38.29	48.75

A.6 Computational Efficiency Analysis

table 9 summarizes the computational characteristics of our variants on a single H100 GPU. We report FLOPs, peak memory, throughput, and the batch sizes used in pretraining and finetuning. FlashLite approximates the functional effect of exact-mode reweighting and integrates it into the

Table 9: Runtime and memory comparison of our ViT-B/16 and ViT-L/16 training variants on a single H100 GPU. FLOPs are reported per 16-frame 224×224 clip under the full-token convention. Encoder+decoder FLOPs apply only to the self-supervised pretraining stage (using a 4-layer / 8-layer transformer decoder for ViT-B / ViT-L, respectively), while finetuning uses the encoder only. Peak memory denotes the maximum allocated GPU memory, and iteration time is measured in milliseconds per training step. Speedup is reported relative to the Exact ViT-B/16 baseline within each stage.

Self-supervised pretraining (per 16-frame 224×224 clip, 1×H100)							
Method	Backbone	Enc FLOPs(G)	Enc+Dec FLOPs(G)	Peak Mem(GB)	Iter Time(ms)	Speedup (×)	Batch
Exact	ViT-B/16	357	394	24.6	400.6	1.00	96
FlashLite	ViT-B/16	357	394	21.6	318.7	1.26	96
FlashLite	ViT-L/16	1189	1226	58.5	558.9	0.72	96
Supervised finetuning (per 16-frame 224×224 clip, 1×H100)							
Method	Backbone	Enc FLOPs(G)	Enc+Dec FLOPs(G)	Peak Mem(GB)	Iter Time(ms)	Speedup (×)	Batch
Exact	ViT-B/16	357	–	71.4	700.8	1.00	16
FlashLite	ViT-B/16	357	–	10.2	169.9	4.12	16
FlashLite	ViT-L/16	1189	–	27.1	414.2	1.69	16

attention computation, enabling compatibility with FlashAttention kernels (Dao et al., 2022; Dao, 2024). As a result, flashLite inherits the memory and throughput efficiency of FlashAttention, with larger gains under full-token finetuning.

Both memory and runtime are primarily governed by attention computation, whose cost scales with the number of active tokens and associated memory accesses. During pretraining, masked autoencoding removes 90% of patches, substantially reducing the number of visible tokens and thus limiting both memory usage and the throughput gap between exact and flashLite. In contrast, during finetuning, all tokens are processed, making attention the dominant factor in both memory and runtime. In this regime, exact mode incurs additional overhead due to explicit reweighting passes over attention maps, whereas flashLite avoids these by integrating reweighting into the attention computation. This leads to substantial improvements in both memory efficiency and training throughput, achieving an 85.7% reduction in peak memory and a **4.1**× speedup for ViT-B/16, while remaining consistently more efficient for ViT-L/16.

A.7 Detailed Comparisons on Facial Expression Recognition

For clarity of evaluation, we first provide a brief description of the metrics (WAR and UAR), which are the standard evaluation metrics adopted by prior state-of-the-art methods in facial emotion recognition. Then, we present confusion matrix analyses to further examine class-wise performance and error patterns.

Main Evaluation Metrics. We employ weighted average recall (WAR) and unweighted average recall (UAR), following standard evaluation protocols in prior state-of-the-art facial emotion recognition studies. Let C denote the number of classes, N_i the number of samples in class i, and R_i the recall of class i, defined as

$$R_i = \frac{TP_i}{TP_i + FN_i},$$

where TP_i and FN_i denote the number of true positives and false negatives for class i, respectively. WAR is defined as the class-frequency-weighted average of recalls:

$$\text{WAR} = \frac{\sum_{i=1}^{C} N_i R_i}{\sum_{i=1}^{C} N_i}.$$

In standard single-label classification, each sample belongs to exactly one class, implying $TP_i + FN_i = N_i$ and thus $R_i = TP_i/N_i$. Substituting this into WAR gives

$$\text{WAR} = \frac{\sum_{i=1}^{C} N_i \cdot \frac{TP_i}{N_i}}{\sum_{i=1}^{C} N_i} = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} N_i}.$$

Since $\sum_i TP_i$ is the total number of correctly classified samples and $\sum_i N_i$ is the total number of samples, WAR is equivalent to the overall accuracy:

$$\text{ACC} = \frac{\sum_i TP_i}{\sum_i N_i} = \text{WAR}.$$

In contrast, UAR computes the average recall across classes:

$$\text{UAR} = \frac{1}{C} \sum_{i=1}^{C} R_i.$$

When the class distribution is balanced (i.e., N_i is identical for all i), the weighting coefficients in WAR become uniform, making WAR equivalent to UAR. However, in imbalanced settings, WAR reflects performance under the empirical data distribution and therefore places greater emphasis on frequent classes. In contrast, UAR corresponds to the average recall under equal weighting across classes, assigning equal importance to all classes regardless of frequency. From this perspective, WAR measures overall classification accuracy, whereas UAR provides a class-balanced evaluation metric. Since facial emotion recognition datasets are typically imbalanced, UAR is often preferred for evaluating balanced recognition performance across emotion categories.

Extended State-of-the-Art Comparison. To complement the compact comparison in the main paper, table 10 provides an extended comparison with additional video-only and multimodal methods on DFEW, MAFW, and FERV39k.

Detailed Results with Confusion Matrices. To provide a finer-grained analysis of the main results, we present confusion matrices for each dataset. These analyses complement the extended comparisons above by revealing class-wise performance and error patterns. As shown in Figs. 9, 10, and 11, the confusion matrices on DFEW, MAFW, and FERV39k provide insights into class-wise recognition performance and the dominant misclassification patterns across emotions.

A.8 Probing Pretrained Representations on Micro-Expression Datasets

To analyze the quality of pretrained representations, we perform frozen-feature k-NN probing under a leave-one-subject-out (LOSO) protocol, where each subject is used as the test set and the remaining subjects form the reference set. Unlike conventional FER, where large intra-class variation necessitates task-specific classification, SAMM (Davison et al., 2016) and MMEW (Ben et al., 2021) are tightly face-aligned micro-expression datasets in which the observable changes across frames are extremely subtle, as illustrated in fig. 6. This setting makes recognition primarily dependent on highly fine-grained temporal dynamics and enables a direct evaluation of the intrinsic quality of pretrained representations. We use a three-class setting (negative, positive, and surprise) for SAMM to mitigate label ambiguity and focus on subtle temporal cues. For MMEW, we use a four-class setting (negative, positive, surprise, and neutral) under the same LOSO evaluation protocol. Features from the frozen backbone are evaluated using cosine-similarity-based k-NN. Table 11 reports Top-1 accuracy averaged over all subjects for different values of k. Our method consistently improves performance across datasets and k values, indicating that the learned representations better capture subtle temporal dynamics.

A.9 Qualitative Analysis of Learned Attention Allocation

Because masked video reconstruction favors cues that most effectively reduce reconstruction error, pretrained attention tends to emphasize dominant appearance variations rather than subtle facial dynamics. To examine how MiRA affects learned attention allocation, figs. 7 and 8 present global and region-conditioned attention maps for representative video samples. We first consider the first

Table 10: Comparison with extended state-of-the-art methods on DFEW, MAFW, and FERV39k. Results on DFEW and MAFW are averaged over 5 folds. (**Best** / Second in video-only methods. Pretrain: **U**nsupervised and **S**upervised. Modality: **V**ideo, **I**mage, **A**udio, and **L**anguage.)

Method	Pretrain	Modality	Face Align.	Params (M)	DFEW		MAFW		FERV39k	
					UAR	WAR	UAR	WAR	UAR	WAR
R(2+1)D-18 (Sarkar et al., 2023)		V		31	42.79	53.22	–	–	–	–
C3D (Tran et al., 2015)		V		79	42.74	53.54	31.17	42.25	–	–
3D ResNet-18 (Hara et al., 2018)		V		33	46.52	58.27	–	–	–	–
IAL (Li et al., 2023)	S	V		19	55.71	69.24	–	–	35.82	48.54
M3DFEL (Wang et al., 2023a)	S	V			56.10	69.25	–	–	35.94	47.67
Former-DFER (Zhao & Liu, 2021)	S	V		18	53.69	65.70	31.16	43.27	37.20	46.85
NR-DFERNet (Li et al., 2022a)	S	V			54.21	68.19	–	–	33.99	45.97
STT (Ma et al., 2022)	S	V			54.58	66.65	–	–	37.76	48.11
LOGO-Former (Ma et al., 2023)	S	V			54.21	66.98	–	–	38.22	48.13
MoCo (He et al., 2020)	U	V		32	53.47	67.45	–	–	–	–
VideoMAE* (Tong et al., 2022)	U	V	✓	86	63.60	74.60	40.87	53.51	43.33	52.39
Video Swin-B (Liu et al., 2022c)	S	V		88	59.38	71.90	–	–	–	–
MAE-DFER (Sun et al., 2023)	U	V	✓	85	63.41	74.43	41.62	54.31	43.12	52.07
SVFAP (Sun et al., 2024b)	U	V	✓	78	62.83	74.27	41.19	54.28	42.14	52.29
ViT-B Exact	U	V	×	86	65.29	75.86	45.24	58.69	–	–
ViT-B FlashLite	U	V	×	86	65.11	76.25	45.48	58.38	43.81	53.80
ViT-L FlashLite	U	V	×	305	67.75	77.32	46.80	60.04	44.41	54.62
ViT-H FlashLite	U	V	×	633	**68.25**	**78.24**	**48.22**	**61.45**	**45.12**	**55.64**
MMA-DFER (Chumachenko et al., 2024)	U	V+A	✓		66.85	77.43	44.25	58.45	–	–
HiCMAE (Sun et al., 2024a)	U	V+A	✓	81	63.76	75.01	42.65	56.17	–	–
AVF-MAE++-B (Wu et al., 2025)	U+S	V+A	✓	169	63.74	75.42	43.10	57.50	–	–
AVF-MAE++-L (Wu et al., 2025)	U+S	V+A	✓	303	65.14	76.24	45.36	59.13	–	–
AVF-MAE++-H (Wu et al., 2025)	U+S	V+A	✓	521	66.88	77.45	46.05	60.24	–	–
CLIPER (Li et al., 2024)	U	V+L	✓		57.56	70.84	–	–	41.23	51.34
DFER-CLIP (Zhao & Patras, 2023)	U	V+L	✓	153	59.61	71.25	39.89	52.55	41.27	51.65
EmoCLIP (Foteinopoulou & Patras, 2024)	U	V+L			58.04	62.12	31.41	36.18	34.24	41.46
GPT-4o (Hurst et al., 2024)		V+A+L			50.57	57.19	38.29	48.82	–	–
Qwen2.5-VL (Bai et al., 2025)		V+A+L			50.04	59.95	–	–	–	–
Qwen2.5-Omni (Xu et al., 2025)		V+A+L			51.48	62.24	–	–	–	–
AVERE (Chaubey et al., 2026)		V+A+L			58.54	64.24	–	–	–	–
Emotion-LLaMA (Cheng et al., 2024)		V+A+L			64.21	77.06	–	–	–	–
EmoThinker (Xu et al., 2026)		V+A+L			65.23	78.13	–	–	–	–

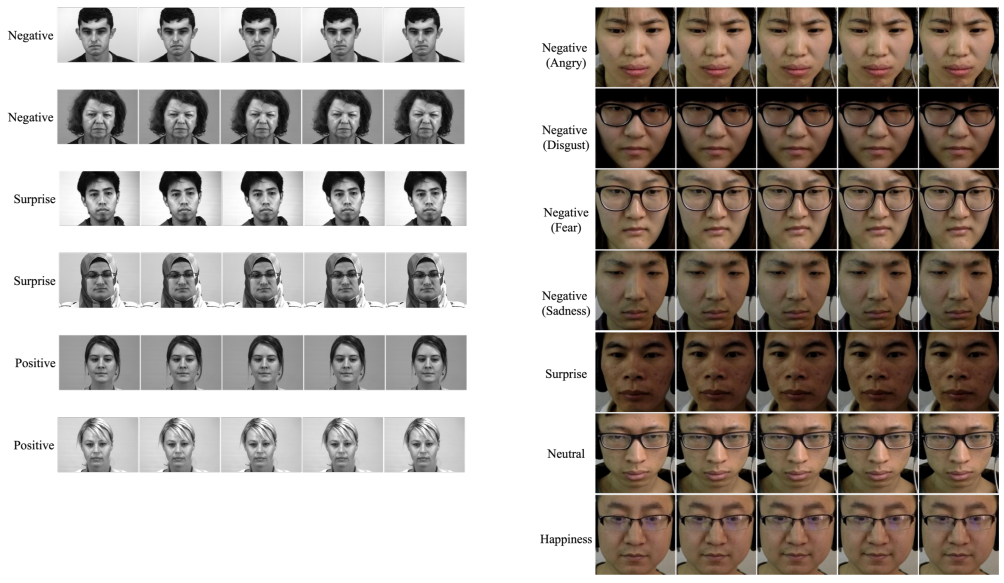


Figure 6: **Representative examples from SAMM (left) and MMEW (right).** Facial regions are tightly aligned across frames, with only subtle temporal variations observable.

Table 11: Evaluation of representation quality via k-NN probing with frozen video features on micro-expression datasets, reported as Top-1 accuracy for different values of k. (**Best** / Second)

Model Family	Method	SAMM			MMEW		
		$k=1$	$k=3$	$k=5$	$k=1$	$k=3$	$k=5$
General Video Models	VideoMAE	0.6876	0.7213	0.7270	0.3124	0.2883	0.3074
	MGMAE	0.7011	0.7350	0.7524	0.2881	0.2903	0.2975
	AdaMAE	0.6276	0.6730	0.7126	0.1674	0.1895	0.2443
FER Models	SVFAP	0.6898	0.6963	0.7373	0.2684	0.2596	0.2710
	MAE-DFER	0.5897	0.6970	0.7568	0.3106	0.2845	**0.3122**
	ViT-B FlashLite	**0.7329**	**0.7369**	**0.7730**	**0.3148**	**0.3037**	0.3118

two examples in fig. 7 in detail. In their left panels, the VideoMAE baseline consistently emphasizes visually salient boundary regions induced by head-pose variation. In the first example, attention concentrates on the left cheek during head turning, whereas in the second example it accumulates around outer facial contours under lateral head motion. In contrast, MiRA focuses on a smaller subset of frames while distributing attention across multiple facial regions, yielding more face-centered attention patterns. The corresponding region-conditioned attention maps further show how a selected facial region interacts with the remaining spatio-temporal locations. VideoMAE attention remains largely confined to the selected frame, with substantial attention extending to surrounding background regions. By contrast, MiRA establishes stronger interactions with facial regions in other frames, enabling complementary facial information to be aggregated across time. Similar behavior can also be observed in the remaining samples, providing further qualitative context for the observations above.

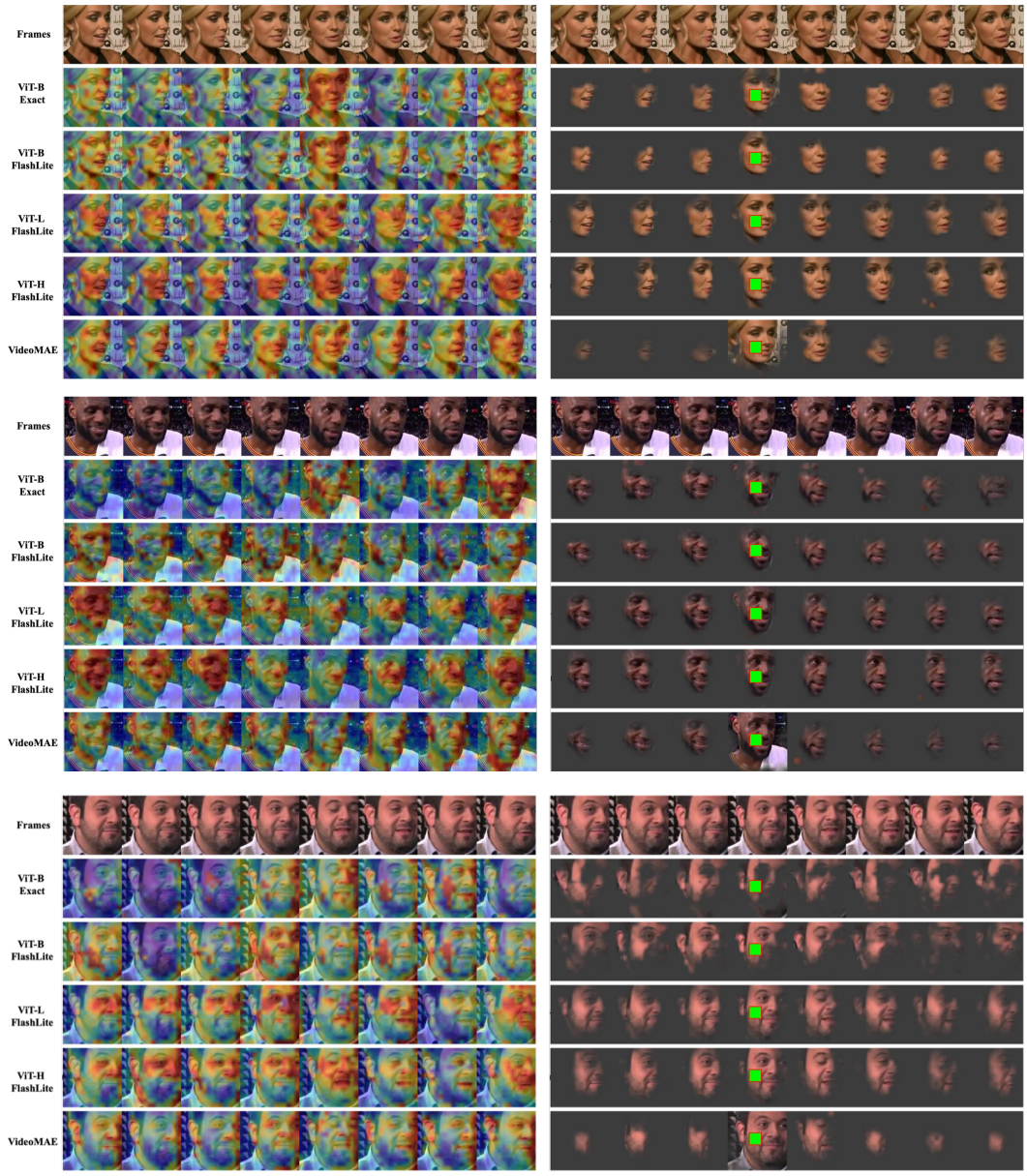


Figure 7: **Global and region-conditioned attention visualizations during masked video reconstruction pretraining.** Each sample group shows a different input video using the same visualization protocol, with model variants arranged by row. Left: global attention visualizations obtained by modulating image visibility according to attention accumulated over the full spatio-temporal sequence. Right: region-conditioned attention heatmaps associated with the selected facial region (green box), illustrating the interaction strength between the highlighted region and the remaining spatio-temporal locations across the video sequence.

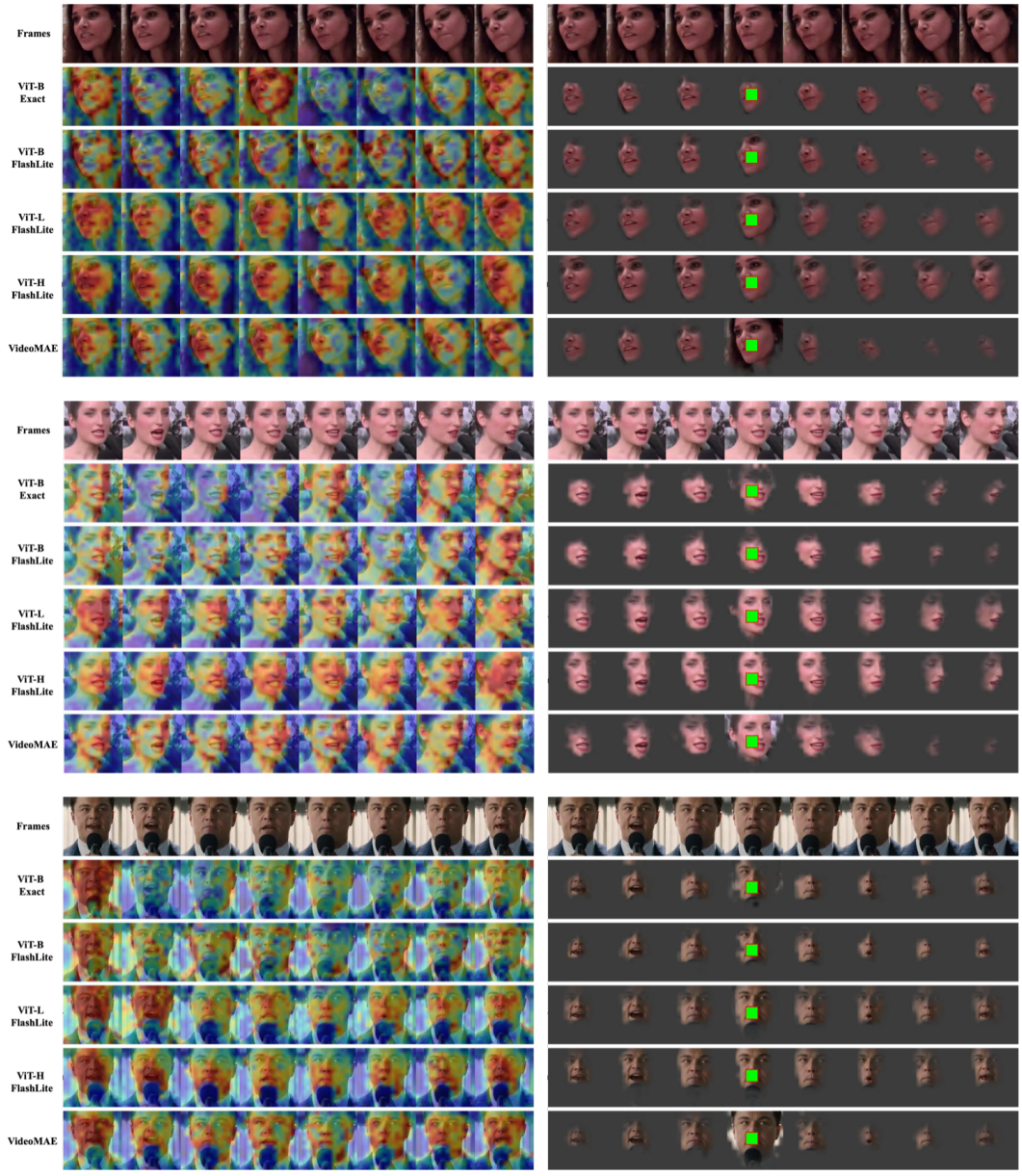


Figure 8: **Global and region-conditioned attention visualizations during masked video reconstruction pretraining (continued from fig. 7).** Each sample group shows a different input video using the same visualization protocol, with model variants arranged by row. Left: global attention visualizations obtained by modulating image visibility according to attention accumulated over the full spatio-temporal sequence. Right: region-conditioned attention heatmaps associated with the selected facial region (green box), illustrating the interaction strength between the highlighted region and the remaining spatio-temporal locations across the video sequence.

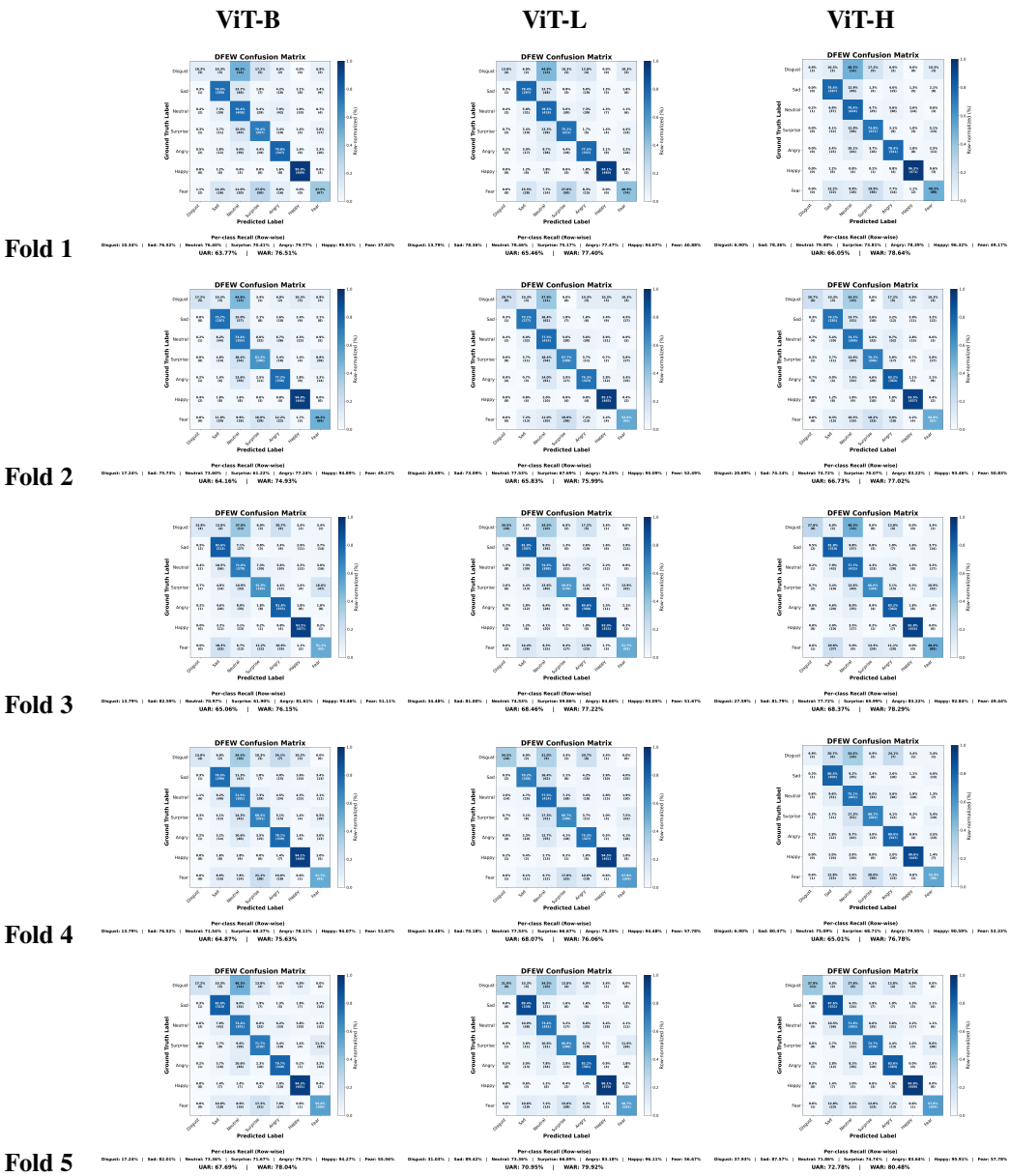


Figure 9: **Confusion matrices on DFEW across all five folds. Best viewed digitally in color with zoom.** Rows correspond to the five cross-validation folds, and columns correspond to ViT-B, ViT-L, and ViT-H with FlashLite.

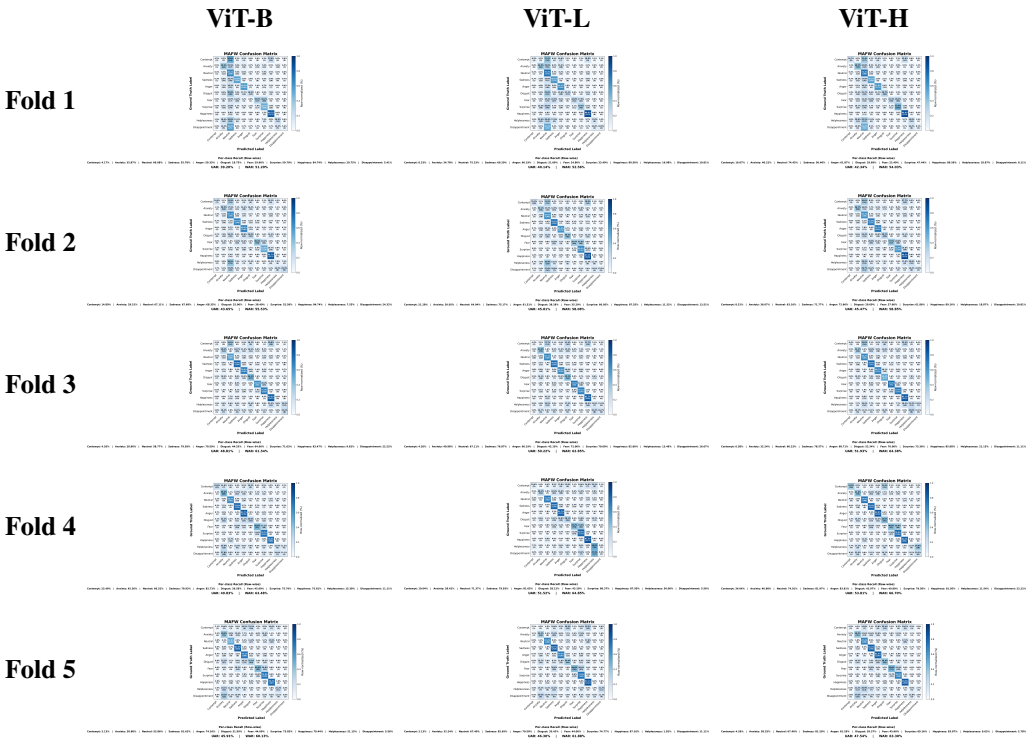


Figure 10: **Confusion matrices on MAFW across all five folds. Best viewed digitally in color with zoom.** Rows correspond to the five cross-validation folds, and columns correspond to ViT-B, ViT-L, and ViT-H with FlashLite.

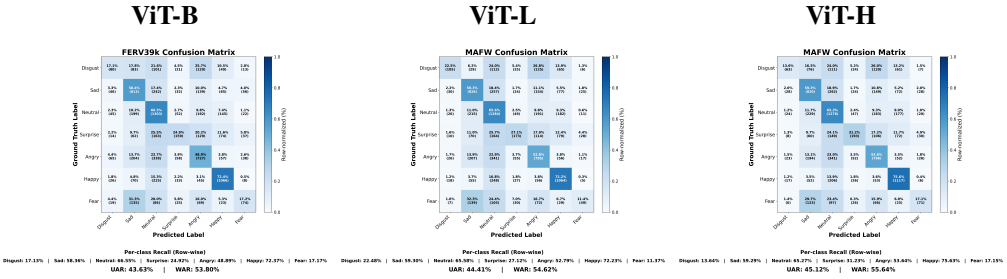


Figure 11: **Confusion matrices on FERV39k. Best viewed digitally in color with zoom.** Columns correspond to ViT-B, ViT-L, and ViT-H with FlashLite.