CLIP-AUTT: Test-Time Personalization with Action Unit Prompting for Fine-Grained Video Emotion Recognition

Muhammad Osama Zeeshan[1], Masoumeh Sharafi[1], Benoît Savary[3], Alessandro Lameiras Koerich[2], Marco Pedersoli[1], and Eric Granger[1]

[1] LIVIA, ILLS, Dept. of Systems Engineering, ETS Montreal, Canada
[2] LIVIA, Dept. of Software and IT Engineering, ETS Montreal, Canada
[3] Dept. of Computer Science, École Polytechnique, Paris, France
{muhammad-osama.zeeshan.1,masoumeh.sharafi.1}@ens.etsmtl.ca
{marco.pedersoli, alessandro.koerich, eric.granger}@etsmtl.ca

Abstract. Personalization in emotion recognition (ER) is essential for an accurate interpretation of subtle and subject-specific expressive patterns. Recent advances in vision–language models (VLMs) such as CLIP demonstrate strong potential for leveraging joint image–text representations in ER. However, CLIP-based methods either depend on CLIP's contrastive pretraining or on LLMs to generate descriptive text prompts, which are noisy, computationally expensive, and fail to capture fine-grained expressions, leading to degraded performance. In this work, we leverage Action Units (AUs) as structured textual prompts within CLIP to model fine-grained facial expressions. AUs encode the subtle muscle activations underlying expressions, providing localized and interpretable semantic cues for more robust ER. We introduce CLIP-AU, a lightweight AU–guided temporal learning method that integrates interpretable AU semantics into CLIP. It learns generic, subject-agnostic representations by aligning AU prompts with facial dynamics, enabling fine-grained ER without CLIP fine-tuning or LLM-generated text supervision. Although CLIP-AU models fine-grained AU semantics, it does not adapt to subject-specific variability in subtle expressions. To address this limitation, we propose CLIP-AUTT, a video-based test-time personalization method that dynamically adapts AU prompts to videos from unseen subjects. By combining entropy-guided temporal window selection with prompt tuning, CLIP-AUTT enables subject-specific adaptation while preserving temporal consistency. Our experiments on three challenging video-based subtle ER datasets — BioVid, StressID, and BAH — indicate that CLIP-AU and CLIP-AUTT outperform state-of-the-art CLIP-based FER and TTA methods. Code: https://github.com/osamazeeshan/CLIP-AUTT.

Keywords: Personalization · Fine-Grained Emotions · AUs · TTA

1 Introduction

Large-scale vision–language models (VLMs) have demonstrated remarkable success across diverse multimodal tasks [2, 16, 28]. Among them, CLIP [28] has

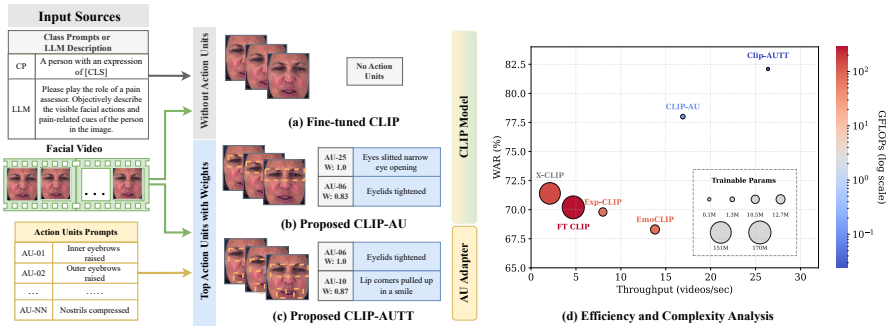

Fig. 1: **(a)** Existing Video ER methods based on CLIP fine-tuning use class prompts or generate LLM prompts that remain fixed at inference time, providing no subject-specific adaptation. **(b)** Our `CLIP-AU` introduces AU-guided pretraining to learn fine-grained expression cues. **(c)** Our `CLIP-AUTT` extends this for test-time personalization, combining entropy-guided temporal window selection with AU prompt tuning to adapt CLIP to each target subject. **(d)** Illustration of throughput (videos/sec), trainable parameters (M), and GFLOPs against `WAR`. Lower GFLOPs indicate better efficiency.

emerged as a powerful VLM that learns a unified representation space between vision and language, enabling flexible cross-modal understanding and adaptation to downstream tasks such as image classification and retrieval [17, 37, 38]. However, despite their versatility, CLIP-based models often struggle with specialized tasks such as recognizing subtle and fine-grained facial expressions. Unlike coarse emotion classification, subtle expressions are low-intensity and arise from localized facial muscle activations, which are difficult to capture using generic semantic prompts commonly used in CLIP-based approaches. Several recent studies have explored the use of CLIP for facial expression recognition (FER) [8, 11, 44, 46], demonstrating the effectiveness of multimodal alignment for macro-level emotion classification. However, these approaches primarily target prototypical basic emotions and often rely on extensive fine-tuning on emotion-labeled datasets, which limits generalization and increases computational cost. Other methods [8, 46] rely on static class-prompts or large language models to generate textual prompts (Fig. 1a), introducing noisy or overly generic supervision that is less suitable for modeling subtle and fine-grained facial expressions.

In this paper, we leverage fine-grained Action Unit (AU) semantics within CLIP to model subtle and subject-specific facial expressions. AUs [25] provide structured and interpretable descriptions of individual facial muscle activations, forming the building blocks of complex expressions [6, 19, 24]. We first introduced `CLIP-AU` that incorporate AU textual descriptions directly as structured text prompts within CLIP's text encoder. Each AU prompt precisely describes a specific facial muscle activation (e.g., AU-06 cheek raiser, AU-25 lips part, AU-01 inner brow raiser), enabling the model to learn fine-grained visual–semantic correspondences at the muscle level as shown in Fig. 1b. Importantly, subtle facial expressions arise from combinations of multiple AUs, resulting in a one-to-many relationship between a single expression and several AU activations.

By computing AU–video similarity scores in CLIP's cross-modal space, `CLIP-AU` captures these compositional patterns without requiring AU ground-truth annotations. Instead of predicting AUs explicitly, we train a lightweight classifier on AU–video similarity features to map learned AU activation patterns directly to subtle expression categories. This design allows `CLIP-AU` to exploit structured AU semantics without AU supervision, avoiding costly CLIP fine-tuning, or LLM-generated prompts while remaining computationally efficient (Fig. 1d).

Personalization adapts models to the diverse expressive patterns of individual users [30, 42] and is essential for modeling subtle facial expressions, particularly in healthcare applications such as pain and stress assessment. Unlike prototypical macro-emotions, subtle expressions are low-intensity, temporally evolving, and highly individualized. The same AU activation (e.g., AU-01 eyebrow raise) may indicate pain for one subject while reflecting a neutral expression for another. This subject-specific variability makes subtle expressions inherently difficult to model using generic representations. Beyond this personalization challenge, additional variability arises from extrinsic factors such as cultural background, or camera conditions [13]. While such factors affect most emotion recognition systems, subtle expressions are particularly vulnerable to misinterpretation due to their weak intensity and individualized nature. To address this subject-specific variability in subtle facial expressions, we introduce `CLIP-AUTT`, a video-based test-time personalization method that dynamically adapts AU prompts at inference to account for individual expressive differences (Fig. 1c). Unlike SoTA test-time adaptation (TTA) methods for CLIP models [1, 14, 21, 32] that focus on static image classification, `CLIP-AUTT` explicitly models temporal dynamics to capture the gradual emergence of subtle and fine-grained facial expressions. Since peak expressive cues often appear only within specific temporal segments, we introduce an entropy-guided temporal window selection strategy to isolate the most confident and expressive window based on AU–video similarity. The selected window is then used for AU prompt tuning via entropy minimization, allowing the AU representations to adapt to how the target subject expresses the subtle emotion. `CLIP-AUTT` operates on subject-specific videos, performing adaptation independently for each subject video and resetting the adapted parameters afterward to maintain stable personalization. The method performs unsupervised subject-level test-time adaptation for temporally consistent, subject-specific modeling of subtle expressions. As shown in Fig. 1(d), under identical NVIDIA A100 48GB settings, `CLIP-AUTT` achieves the best performance with substantially fewer trainable parameters, lower computational cost, and higher video throughput than competing methods.

Main contributions: **(1)** `CLIP-AU`: An AU-guided temporal learning is proposed that integrates AU semantics into the CLIP model, enabling it to recognize fine-grained and subtle facial expressions without requiring AU labels or LLM-generated descriptions. **(2)** `CLIP-AUTT`: A video-based test-time personalization method that dynamically adapts to unseen target subjects by optimizing AU prompts and selecting key expressive windows, enabling subject-specific modeling during inference without additional supervision. **(3)** An extensive set of

experiments was conducted to validate our methods under two settings – (1) `CLIP-AU`: fine-grained video alignment and (2) `CLIP-AUTT`: video-based test-time personalization – on three challenging video ER datasets – BioVid [34], StressID [7], and BAH [12] – demonstrating that our methods outperform state-of-the-art CLIP-based FER and test-time adaptation methods.

2 Related Work

CLIP-based FER Models. CLIP-style models learn a joint image–text embedding via contrastive pretraining on large web corpora; at inference, they score the similarity between an image and text prompts (e.g., "a photo of a [CLS] face") to enable zero-/few-shot classification without task-specific heads [28]. In FER, this typically means crafting emotion prompts (sometimes with prompt ensembles), optionally tuning a small set of parameters (e.g., adapters/LoRA), or enriching the text side with descriptions derived from LLMs, then classifying by the nearest text prototype. Recent works extend CLIP to FER along several axes: adapter-tuned, fine-grained prompting for dynamic FER (FineCLIPER) [8]; zero-shot *video* FER via sample-level captions (EmoCLIP) [11]; multimodal pretraining from verbal and nonverbal cues (EmotionCLIP) [44]; and LLM-augmented text prototypes with learned alignment (Exp-CLIP) [46]. Despite their zero-/few-shot positioning, these approaches still require training data.

Facial Action Units. Facial Action Coding System (FACS) provides a compositional, muscle-level description of facial movements via Action Units (AUs), which many works exploit as structured priors for expression recognition [10]. Recent AU-aware methods leverage AUs as supervision, auxiliary tasks, or intermediate reasoning cues. Li et al. use AU-assisted meta multi-task learning to improve compound expression recognition on RAF-CE [19], while Liu et al. incorporate features from a facial AU detector into a multi-modal FER pipeline for in-the-wild videos [24]. On the AU-detection side, FG-Net and MCM learn generalizable AU representations that support cross-corpus facial behaviour analysis [39, 45]. Beyond conventional FER, Emotion-LLaMA integrates FACS-style AU patterns into a multimodal LLM for emotion recognition and reasoning [9], and Belharbi et al. [6] guide FER with a spatial AU codebook that supervises interpretable attention maps.

Personalization in FER. Standard FER literature largely targets models trained and evaluated under subject-independent protocols without personalization [3, 18, 33]. FER personalization is the subject-specific adaptation of a pre-trained model during training, fine-tuning, or inference to mitigate subject-level distribution shifts arising from anatomical, cultural, personality, and contextual variability in real-world settings. Despite notable gains from supervised personalization methods [5, 29], these approaches depend on per-subject labeled data, which is costly to collect and raises privacy concerns. To address these challenges, several multi-source domain adaptation (MSDA) methods [40–42] have been proposed. These techniques leverage labeled data from multiple sources to learn domain-invariant representations for a target subject by improving ro-

bustness under cross-subject shifts. However, in many real-world applications, the source data cannot be retained or shared due to privacy constraints, which motivates source-free domain adaptation (SFDA). SFDA adapts by using only a pretrained source model and unlabeled target data via self-training, entropy minimization, and prototype refinement [30, 31].

Test-time Domain Adaptation. TTA updates a trained model at test time using unlabeled target test data captured to mitigate distribution shift, offering practical gains when source data are unavailable and full retraining is infeasible [22, 23, 35, 36]. While effective, conventional TTA often drifts under confirmation bias, overfits to short or non-i.i.d. streams, and depends on BatchNorm updates that fail on LayerNorm-only backbones (e.g., ViT). To address these issues, CLIP-based TTA leverages CLIP's strong zero-shot priors and text anchors to enable label-free, structure-aware adaptation through lightweight parameter updates that remain stable (via zero-init and EMA) while preserving the pretrained mapping and reducing test-time shifty [1, 14, 21, 32, 47]. Loss functions aligned with CLIP contrastive pre-training allow to mitigate pseudo-label drift and class collapse [15], prompt-ensemble strategies with periodic weight averaging yield robust updates—including single-image adaptation [27], and backdoor analyses surface security risks for CLIP-style models [20].

3 Proposed Method

State-of-the-art CLIP-based ER approaches [8, 11, 44, 46] associate video frames with expression-related text prompts, typically using generic templates such as *"a person with an expression of [CLS]"* or descriptions generated by large language models. While effective for macro-level emotion classification, such prompts overlook the localized facial muscle activations that characterize subtle and subject-specific expressions. Moreover, these methods often require fine-tuning of CLIP on emotion-labeled datasets, limiting generalization and increasing computational cost. Subtle facial expressions arise from specific combinations of localized facial muscle movements. To address this limitation, we propose `CLIP-AU`, which replaces class-level prompts with textual descriptions of individual AUs to explicitly model fine-grained facial muscle activations. Building upon this, we introduce `CLIP-AUTT`, a test-time adaptation method that enables subject-specific personalization for subtle ER.

Preliminaries. The proposed method builds upon a frozen VLM composed of a **visual encoder** $\mathcal{E}_v$ and a **textual encoder** $\mathcal{E}_t$, which extract visual and textual embeddings, respectively. For the textual modality, we define a set of $N = 46$ AU descriptions $\{a_i\}_{i=1}^{N}$, where each a_i is a short natural-language phrase describing a localized facial muscle movement (e.g., "eyebrow raised", "nose wrinkled", "lip corner pulled"). The encoded AU features are further refined through a lightweight **AU adapter** $\mathcal{A}_t$ that specializes the text embeddings for expression-related semantics. The **temporal module** processes sequential frame embeddings from $\mathcal{E}_v$ and consists of a 1D convolutional encoder denoted as $\mathcal{T}(\cdot)$, a gated linear unit (GLU), and an average pooling operation to capture temporal

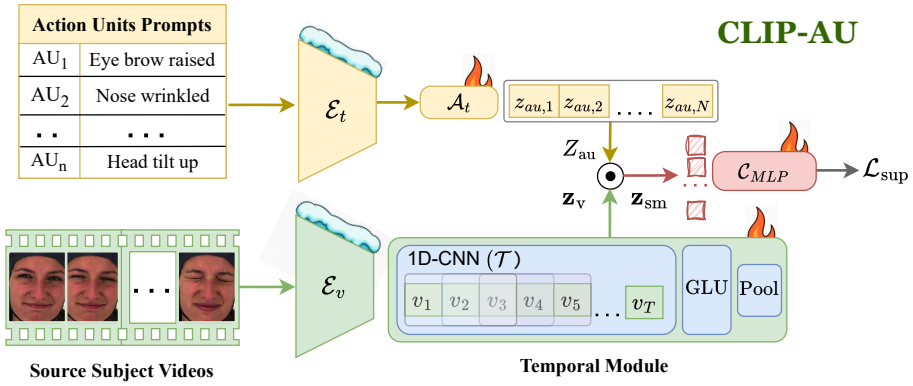

Fig. 2: Overview of the `CLIP-AU`. The model aligns AU text embeddings with temporally encoded video representations to capture fine-grained facial dynamics for ER.

dynamics. A two-layer **emotion classifier** $\mathcal{C}_{\text{MLP}}$ maps the AU–video similarity representations to emotion logits. Throughout this paper, Z_{au} denotes the set of AU embeddings, $\mathbf{z}_v$ represents the temporally aggregated video embedding, $\mathbf{s}_i$ denotes the AU–window similarity vector for the i-th temporal segment, and $p(c|\mathbf{X})$ refers to the predicted probability for expression class c for the video $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times 3}$. All these components are consistently used across both the `CLIP-AU` and `CLIP-AUTT` methods.

3.1 `CLIP-AU`: Fine-Grained Video Alignment

During pretraining, $\mathcal{E}_t$ and $\mathcal{E}_v$ encoders are kept frozen to preserve their rich multimodal representations. A lightweight **AU adapter** ($\mathcal{A}_t$) is introduced after the $\mathcal{E}_t$ to specialize the textual embeddings toward AU-specific semantics that are critical for recognizing subtle facial expressions. While the frozen text encoder retains the general linguistic and visual alignment knowledge of CLIP, the AU adapter learns to extract task-relevant and fine-grained facial cues that are discriminative for emotion recognition. In parallel, a **temporal module** processes frame-level embeddings from the $\mathcal{E}_v$ to capture motion-aware temporal dynamics, enabling the model to encode expressive variations over time. Finally, an **emotion classifier** is trained to predict categorical emotions based on the learned AU features. The overall objective is to align these refined textual and visual representations for robust AU–video correspondence, as illustrated in Fig. 2.

Given a source video $\mathbf{X} = \{I_t\}_{t=1}^{T}$, each frame $I_t \in \mathbb{R}^{H\times W\times 3}$ is first encoded by the frozen image encoder $\mathcal{E}_v(\cdot)$ to obtain frame-level embeddings:

$$\mathbf{v}_t = \mathcal{E}_v(I_t), \quad V = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_T] \in \mathbb{R}^{T\times d}. \tag{1}$$

To capture the temporal information of facial expressions, sequence V is processed using a 1D convolutional temporal encoder ($\mathcal{T}$), followed by GLUs and an average pooling layer:

$$\mathbf{z}_v = \text{Pool}\big(\text{GLU}(\mathcal{T}(V))\big) \in \mathbb{R}^{d}. \tag{2}$$

$\mathcal{T}(\cdot)$ extracts motion-aware representations by modeling local temporal continuity and fine-grained frame dependencies within a video sequence. It efficiently aggregates short-term dynamics, enabling the representation $\mathbf{z}_{\mathrm{v}}$ to capture subtle temporal variations in facial motion, such as transient cues linked to pain, stress, or ambivalence. GLU further enhances this capability by adaptively emphasizing salient temporal changes while suppressing neutral or low-variance frames, resulting in a temporally discriminative and noise-robust visual embedding.

For the textual modality, each AU description a_i is encoded by the frozen text encoder $\mathcal{E}_t$ and refined through the AU adapter $\mathcal{A}_{\mathrm{t}}(\cdot)$:

$$z_{\mathrm{au},i} = \mathcal{A}_{\mathrm{t}}(\mathcal{E}_t(a_i)), \quad Z_{\mathrm{au}} = [z_{\mathrm{au},1}, \ldots, z_{\mathrm{au},N}] \in \mathbb{R}^{N\times d} \tag{3}$$

Then, cosine similarity is computed between the temporal-aware visual embedding z_v and each AU embedding $z_{\mathrm{au},i}$:

$$z_{\mathrm{sm},i} = \frac{z_v^\top z_{\mathrm{au},i}}{\|z_v\|\,\|z_{\mathrm{au},i}\|}, \quad \mathbf{z}_{\mathrm{sm}} = [z_{\mathrm{sm},1}, \ldots, z_{\mathrm{sm},N}] \in \mathbb{R}^N \tag{4}$$

The resulting similarity vector $\mathbf{z}_{\mathrm{sm}}$ represents the activation strengths of all AUs across the video and is passed to an MLP-based emotion classifier, which is trained using a supervised loss:

$$\mathcal{L}_{\mathrm{sup}} = -\frac{1}{M}\sum_{m=1}^{M}\sum_{c=1}^{C} y_{m,c}\log p_{m,c}, \tag{5}$$

where M denotes the total number of training videos, p_c denotes the predicted probability for class c obtained from $\mathcal{C}_{\mathrm{MLP}}(\mathbf{z}_{\mathrm{sm}})$. This formulation enables `CLIP-AU` to align AU-level semantics with temporally salient visual features, facilitating robust recognition of subtle and fine-grained facial expressions.

3.2 `CLIP-AUTT`: Video Test-time Personalization

Although `CLIP-AU` effectively learns fine-grained AU–video alignment, its performance may degrade on unseen subjects due to differences in facial appearance, expression intensity, and motion dynamics. To address these challenges, we introduce `CLIP-AUTT`, a test-time adaptation method that performs unsupervised, subject-specific personalization through **Entropy-Guided Temporal Window Selection** and **AU Prompt Tuning**, as illustrated in Fig. 3.

Entropy-Guided Temporal Window Selection. Facial expressions associated with subtle and fine-grained emotions (e.g., pain or stress) typically evolve gradually over time, reaching a peak intensity before fading. Consequently, not all frames in a video are equally informative; Early and late segments often contain low-intensity or ambiguous facial cues that may not reliably reflect the underlying expression. Since each video corresponds to a single subtle expression, accurately identifying the temporally coherent peak segment is crucial for

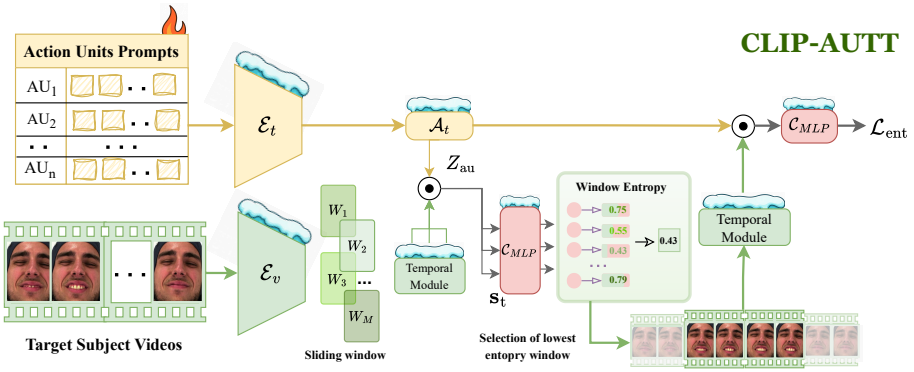

Fig. 3: Overview of the `CLIP-AUTT`. Given a target video, `CLIP-AUTT` applies a sliding window with a temporal module to capture temporally coherent AU patterns and estimates window-level entropy from AU similarity to select the most expressive segment. The selected window is then used for AU prompt tuning, adapting AU embeddings to the target subject that results in a improve subject-specific personalized model.

reliable recognition. This becomes particularly important in `CLIP-AUTT`, as test-time AU prompt tuning relies on selecting a confident and expressive segment to accurately adapt AU semantics to the target subject. Adapting on weak or ambiguous segments may misguide the AU semantics and degrade personalization. To address this, `CLIP-AUTT` employs an entropy-guided temporal window selection strategy grounded in AU–video semantic alignment, focusing adaptation on the most confident and expressive window.

Given a target video $X = \{I_t\}_{t=1}^{T}$, each frame is first encoded using the visual encoder $\mathbf{v}_t = \mathcal{E}_v(I_t)$. To capture the temporal evolution of subtle and fine-grained facial expressions, a sliding window of length L is applied over the frame embeddings to generate $M = T - L + 1$ overlapping windows:

$$W_i = \{\mathbf{v}_t\}_{t=i}^{i+L-1}, \quad i = 1, \ldots, M. \tag{6}$$

where W_i denotes a contiguous set of L frame embeddings representing a short temporal segment of the video. Each window is processed by the temporal module to capture short-term facial dynamics:

$$\mathbf{z}_i = \mathrm{Pool}\big(\mathrm{GLU}(\mathcal{T}(W_i))\big) \in \mathbb{R}^d. \tag{7}$$

Let $\mathbf{z}_i$ denote the embedding of the i-th temporally aggregated window, and $\{\mathbf{z}_{\mathrm{au},k}\}_{k=1}^{N}$ the text embeddings of N AU prompts. We compute the cosine similarity between $\mathbf{z}_i$ and each AU embedding and stack the resulting scores to form an AU similarity vector:

$$s_{i,k} = \frac{\mathbf{z}_i^\top \mathbf{z}_{\mathrm{au},k}}{\|\mathbf{z}_i\| \, \|\mathbf{z}_{\mathrm{au},k}\|}, \quad \mathbf{s}_i = [s_{i,1}, \ldots, s_{i,N}] \in \mathbb{R}^N. \tag{8}$$

here, $s_{i,k}$ quantifies the semantic alignment between the i-th window and the k-th AU, while $\mathbf{s}_i$ provides a structured AU-level representation of the facial

expression in window i. The AU similarity vector is then passed to the pretrained classifier $\mathcal{C}_{\text{MLP}}$ to estimate expression probabilities:

$$\mathbf{p}_i = \text{softmax}\big(\mathcal{C}_{\text{MLP}}(\mathbf{s}_i)\big), \tag{9}$$

where $\mathbf{p}_i \in \mathbb{R}^C$ denotes the predicted probability distribution over C expression categories. To quantify prediction confidence, we compute the Shannon entropy of the window-level distribution:

$$h_i = -\sum_{c=1}^{C} p_i(c) \log p_i(c). \tag{10}$$

A lower entropy indicates higher confidence and stronger semantic consistency between AU activations and the predicted expression. The most expressive and reliable segment is selected by $i^* = \arg\min_i \ h_i$. Only the selected window W_{i^*} is used for subsequent test-time AU prompt adaptation. By restricting personalization to the most confident window, we ensure that AU tuning is guided by temporally coherent and discriminative expression evidence.

AU Prompt Tuning. Recent studies have demonstrated that prompt tuning can effectively adapt large-scale VLMs to new tasks and unseen domains without full fine-tuning. Methods such as PromptAlign [1] and TPT [32] have shown that learning or optimizing textual prompts enables efficient adaptation while preserving the pretrained model's generalization ability. Motivated by these successes, we investigate whether prompt tuning can also facilitate personalized adaptation in video-based facial expression recognition, where inter-subject variability plays a crucial role. In this context, each subject expresses emotions through distinct combinations of facial muscle activations. For instance, one individual may display pain through lip stretching and eye closure, while another may exhibit it via brow raising and cheek tension. To capture these personalized patterns, `CLIP-AUTT` introduces AU prompt tuning, which optimizes Action Unit text embeddings in an unsupervised, subject-specific manner during test time to adaptively align AU semantics with individual expression styles while keeping all other modules frozen.

Using the selected key window from Sec. 3.2, their temporal embeddings are obtained through the temporal module, which consists of an encoder $\mathcal{T}(\cdot)$, a GLU, and an average pooling operation:

$$\mathbf{z}_{\text{v}} = \text{Pool}\big(\text{GLU}(\mathcal{T}(W_{i^*}))\big) \in \mathbb{R}^d, \tag{11}$$

where W_{i^*} denotes the selected temporal window embeddings. The resulting temporal representation $\mathbf{z}_{\text{v}}$ is then compared with the AU embeddings $\{z_{\text{au},i}\}_{i=1}^N$ using cosine similarity:

$$z'_{sm,i} = \frac{z_v^\top z_{\text{au},i}}{\|z_v\| \, \|z_{\text{au},i}\|}, \quad \mathbf{z}'_{\text{sm}} = [s'_1, \ldots, s'_N] \in \mathbb{R}^N. \tag{12}$$

The similarity vector $\mathbf{z}'_{sm}$ is passed through the classifier $\mathcal{C}_{\text{MLP}}$ to produce class probabilities, and the AU prompts are optimized via entropy minimization:

$$\mathcal{L}_{\text{ent}} = -\sum_{c=1}^{C} p(c|\mathbf{X}) \log p(c|\mathbf{X}), \tag{13}$$

where $p(c|\mathbf{X}) = \text{softmax}(\mathcal{C}_{\text{MLP}}(\mathbf{z}'_{\text{sm}}))_c$. This adaptation step tunes the AU prompt embeddings Z_{au} to align with each distinctive subject activation pattern, improving the AU–video alignment and enhancing personalized expression recognition under unseen conditions.

4 Results and Discussion

4.1 Experimental Protocol

Datasets. `CLIP-AU` evaluated on three challenging datasets. **BioVid Heat Pain (Part A)** [34] with 87 subjects, following [40], we use 77 as source and 10 as target subjects. **StressID** [7] dataset contains recordings of different participants captured in a controlled setting while performing 11 tasks designed to trigger the stress response of each individual. We use 44 subjects as the source and 10 subjects as the target. **Behavioural Ambivalence/Hesitancy (BAH)** [12] consists of participants from diverse demographics in uncontrolled, real-world settings. We use 143 as the source and 10 subjects as the target for personalization. See suppl. materials for the complete list of target subjects.
Implementation Details. All models are optimized using the AdamW optimizer with a learning rate of 1×10^{-3} and a weight decay of 1×10^{-4}. For `CLIP-AU`, we use a batch size of 8, processing eight video clips simultaneously. The AU adapter and emotion classifier are implemented as two-layer MLPs. The temporal encoder is a lightweight 1D CNN composed of a single linear layer, followed by a Gated Linear Unit (GLU) and average pooling for temporal aggregation. For `CLIP-AUTT`, the model operates in a video-level test-time setting, where one video is processed at a time for each subject. Thus, the batch size is set to 1.
Baseline Methods. We report results using two standard evaluation metrics. Weighted Average Recall (`WAR`), which measures the overall classification accuracy. `F1`-score, which provides a balanced measure of precision and recall under imbalanced conditions. We include two sets of baseline comparisons in our study. First, for `CLIP-AU`, we compare our approach against existing video-based FER methods. These include the CLIP-ViT-B/32 [28] backbone without any fine-tuning, CLIP-ViT-B/32†, which performs complete end-to-end fine-tuning of the CLIP model, and EmoCLIP (Adapter) [11], where only the adapter layers are trained. We also compare with X-CLIP [26] and Exp-CLIP [46], which train a projection head and leverage LLMs to extract facial expression descriptions for supervision. Second, for `CLIP-AUTT`, we compare against several image-based SoTA TTA methods, including TPT [32], TDA [14], DPE [43], PromptAlign [1], and ReTA [21]. Further, we adapt the action recognition video-based T3AL [22]

Table 1: Comparison with CLIP-based FER and TTA methods. Results are averaged over 10 target subjects per dataset. *CLIP-ViT-B/32*† denotes full CLIP fine-tuning. *ZS: Zero-shot*; *FT: Fine-tuning*; *TTA: Test-time adaptation.*

Settings	Method	BioVid		StressID		BAH	
		WAR ↑	F1 ↑	WAR ↑	F1 ↑	WAR ↑	F1 ↑
ZS	CLIP-ViT-B/32 [28] (ICML'21)	50.0	33.3	60.4	34.8	39.5	28.1
FT	CLIP-ViT-B/32† [28] (ICML'21)	69.7	66.6	**67.0**	44.5	60.4	39.8
	EmoCLIP [11] (FG'24)	67.7	63.4	63.5	35.9	56.2	36.5
	X-CLIP [26] (ECCV'22)	70.9	57.9	62.3	41.3	63.0	39.2
	Exp-CLIP [46] (WACV'25)	70.2	66.7	63.1	44.5	62.2	38.5
	CLIP-AU	**78.0**	**74.8**	66.5	**58.5**	**68.3**	**40.3**
TTA	TPT [32] (NeurIPS'22)	71.1	67.5	70.9	57.9	65.6	39.7
	TDA [14] (CVPR'24)	71.4	68.2	69.7	49.9	65.2	39.9
	DPE [43] (NeurIPS'24)	73.1	69.6	71.3	54.2	66.7	39.4
	PromptAlign [1] (NeurIPS'23)	75.3	71.6	74.6	53.2	67.1	39.7
	ReTA [21] (ACMMM'25)	75.1	71.3	71.8	52.8	67.6	39.8
	T3AL [22] (CVPR'24)	76.1	72.9	75.9	59.4	67.9	40.7
	CLIP-AUTT	**81.5**	**78.0**	**80.8**	**77.9**	**69.8**	**41.1**

Table 2: CLIP-AUTT subject-wise comparison with TTA methods on the BioVid dataset. Best results are in bold.

Method	Woman-27		Man-36		Woman-43		Man-25		Woman-25		Woman-65		Woman-21	
	WAR	F1	WAR	F1	WAR	F1	WAR	F1	WAR	F1	WAR	F1	WAR	F1
TPT [32]	91.9	92.0	53.0	51.5	61.5	49.6	88.9	87.0	76.0	79.5	98.9	99.0	79.0	79.0
TDA [14]	93.0	94.9	52.6	50.3	63.9	48.0	87.5	86.2	76.3	80.2	100.0	100.0	79.5	80.8
DPE [43]	91.7	90.4	55.0	51.5	85.0	73.1	84.7	83.6	80.0	79.8	100.0	100.0	73.0	72.0
PromptAlign [1]	93.8	94.2	65.2	60.1	70.0	61.5	92.0	90.5	80.0	80.2	100.0	100.0	85.0	83.5
ReTA [21]	93.8	94.2	65.2	60.1	69.5	60.8	90.2	88.0	80.0	80.2	100.0	100.0	85.0	83.5
T3AL [22]	94.0	**95.8**	66.5	62.0	70.0	61.5	94.2	93.9	83.9	83.0	100.0	100.0	85.2	83.9
CLIP-AUTT	**95.0**	94.9	**75.0**	**74.9**	**100.0**	**100.0**	**97.5**	**97.5**	**92.5**	**92.4**	**100.0**	**100.0**	**87.5**	**87.3**

method for our classification task. To ensure a fair comparison, we use the CLIP-ViT-B/32† backbone, which is fine-tuned on the source subject of each dataset.
Prompts. For all methods, we use the prompt *"a person with an expression of [CLS]"*, which has been shown to work effectively in prior FER work [46].

4.2 Comparison with State-of-the-Art Methods

Fine-tuning. Table 1 shows that CLIP-AU improves over all video-based CLIP FER baselines. On BioVid it reaches 78.0/74.8, and on BAH it is best across both matrices (68.3/40.3). On StressID, EmoCLIP† is slightly higher in WAR (67.0 vs 66.5), but CLIP-AU achieves a decisively higher F1 (58.5, +14.0%), indicating substantially more balanced per-class predictions under class imbalance. See suppl. materials for t-SNE analysis showing improved class separation.
Test-time adaptation. Against TTA baselines in Table 1, CLIP-AUTT sets a new state of the art. We compare to image-level TTA methods (TPT, TDA, DPE, PromptAlign, ReTA) and the video-level TTA method (T3AL). CLIP-AUTT lifts BioVid to 81.5/78.0, boosts StressID to 80.8/77.9 (large margins in F1),

Table 3: CLIP-AUTT subject-wise comparison with TTA methods on the StressID dataset. Best results are in bold.

Method	Man								Woman					
	Sub-1		Sub-2		Sub-3		Sub-4		Sub-5		Sub-6		Sub-7	
	WAR	F1	WAR	F1	WAR	F1	WAR	F1	WAR	F1	WAR	F1	WAR	F1
TPT [32]	74.7	70.7	90.0	86.0	77.5	78.7	51.6	43.6	63.4	50.0	54.0	40.0	91.9	50.0
TDA [14]	66.9	45.5	75.6	50.1	80.0	80.0	45.2	40.0	75.3	45.6	81.8	42.8	91.9	50.0
DPE [43]	64.0	40.9	73.9	49.0	80.0	80.0	55.0	40.6	73.0	46.0	81.8	42.8	92.0	56.6
PromptAlign [1]	69.9	41.2	78.6	45.6	80.0	80.0	61.6	45.0	81.3	58.0	81.8	42.8	90.9	45.4
ReTA [21]	64.5	43.2	82.0	53.0	80.0	80.0	50.5	41.0	80.0	53.3	81.8	42.8	93.7	59.3
T3AL [22]	71.3	46.3	80.3	48.6	80.0	80.0	62.7	48.6	77.0	68.0	**82.0**	44.0	93.9	66.0
CLIP-AUTT	**90.9**	**90.6**	**90.9**	**90.6**	**100.0**	**100.0**	**63.6**	**63.3**	**81.8**	**80.3**	72.7	**68.6**	**100.0**	**100.0**

and improves BAH to 69.8/41.1. These gains stem from temporally adapting AU prompts at test time and calibrating to each individual, which strengthens per-class balance and robustness under shift.

Personalization. Tables 2 and 3 report subject-wise results on BioVid and StressID. CLIP-AUTT adapts to each person's specific AU patterns, enabling more reliable expression predictions and achieving strong performance across most subject–metric pairs. It achieves perfect scores on easier identities (e.g., BioVid *Woman-43*; StressID *Sub-7*) and delivers large gains on harder ones (e.g., StressID *Sub-4*: 63.6/63.3). Even when a baseline slightly edges WAR for a subject (e.g., *Sub-6*), CLIP-AUTT consistently raises F1, indicating less bias toward frequent classes. Results for the BAH dataset and all *10* target subjects are provided in the suppl. materials.

4.3 Qualitative Visualization

AU Analysis. Figure. 4 qualitatively compares CLIP-AU and CLIP-AUTT on subjects from the BioVid dataset. While CLIP-AU captures general facial muscle activations correlated with emotion classes, CLIP-AUTT dynamically adapts to each subject's expressive behavior, identifying more meaningful AU combinations such as fine-grained eye and mouth movements that reflect subject-specific activation patterns. Since the dataset does not provide ground-truth AU annotations, we additionally estimate AU activations using an external AU detector, Openface [4], to assess the reliability of the predicted AUs. The AU patterns predicted by CLIP-AUTT show stronger agreement with those detected by OpenFace and align better with the underlying expression, resulting in higher classification confidence. These examples demonstrate how CLIP-AU learns a generic understanding of AU semantics across subjects, while CLIP-AUTT further refines these representations through test-time adaptation to achieve personalized AU alignment without requiring AU supervision. More visualization in suppl. material.

Representative Temporal Window Selection. Figure 5 visualizes the most expressive temporal windows, their prediction uncertainty, and the minimum-entropy segment selected for adaptation. The selected window yields the most

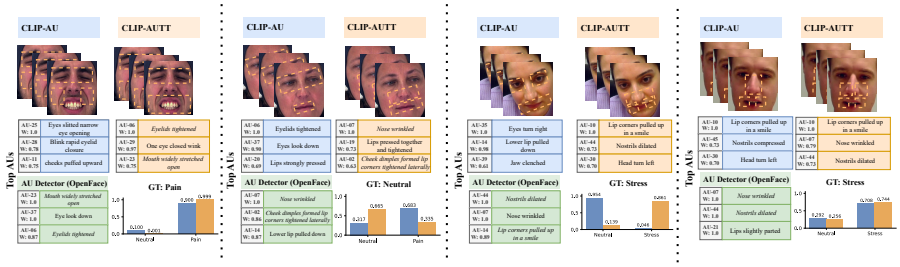

Fig. 4: Qualitative comparison of predicted top-activated AUs and class predictions for `CLIP-AU` and `CLIP-AUTT`, alongside AU activations estimated by an external detector (OpenFace). `CLIP-AUTT` shows stronger agreement with the externally estimated AUs and better alignment with the underlying expression.

concentrated class distribution with the lowest entropy $h_3 = 0.493$ and, therefore, the most confident emotion prediction. Qualitatively, it also contains the clearest and most temporally consistent facial-expression cues, while the remaining windows exhibit weaker or more ambiguous evidence. Using this informative segment provides a more reliable signal for AU prompt adaptation and helps reduce updates based on uncertain frames.

Dataset Variability Analysis. To better understand dataset insights, Figure. 6 visualizes representative samples from BioVid, StressID, and BAH. BioVid and StressID are collected under controlled laboratory conditions with stable lighting and minimal incidental motion. In contrast, BAH is recorded in semi-controlled settings where subjects capture videos using personal devices, resulting in larger variations in illumination, camera positioning, and incidental facial movements (e.g., speaking). These factors can obscure the low-intensity facial cues associated with subtle expressions, making reliable recognition more challenging.

4.4 Ablation Studies

Impact of Temporal Window Selection. Table 4 (left) examines the impact of varying the temporal window length for `CLIP-AUTT` on BioVid. A shorter window length generally provides a better accuracy–efficiency trade-off, with 16 frames yielding the best performance. Importantly, `CLIP-AUTT` remains relatively stable across different window lengths, as the entropy-based selection reliably identifies the most expressive temporal segments. See suppl. material for extended ablation.

Are Action Units Really Helping? To assess whether our AU adapter truly captures subject-specific AU semantics, we replace AU prompts with an ensemble of 46 generic FER class prompts (23 neutral and 23 emotion-specific templates). As shown in Figure 4 (right), training with these generic CPs yields

Win.	WAR	F1
8	78.8	75.4
16	81.5	78.0
32	79.8	76.3
64	80.0	76.6
72	79.0	75.4

Method	BioVid		StressID	
	WAR	F1	WAR	F1
Fine-tuning				
Ens. CPs	69.5	63.9	61.3	49.9
CLIP-AU	**78.0**	**74.8**	**66.5**	**58.5**
TTA				
Ens. CPs	77.5	74.0	44.1	31.0
CLIP-AUTT	**81.5**	**78.0**	**80.8**	**77.9**

Table 4: Left. Temporal-window selection, performance across window lengths (L). **Right.** Prompts-design ablations, AU prompts versus generic class prompts under fine-tuning and TTA.

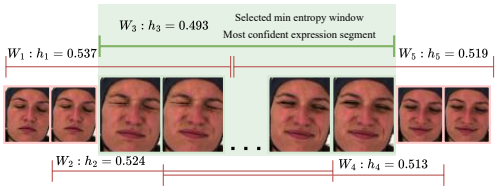

Fig. 5: Selection of the candidate temporal window with the minimum-entropy score. The selected segment contains the most confident and temporally coherent facial-expression cues that are used for personalization.

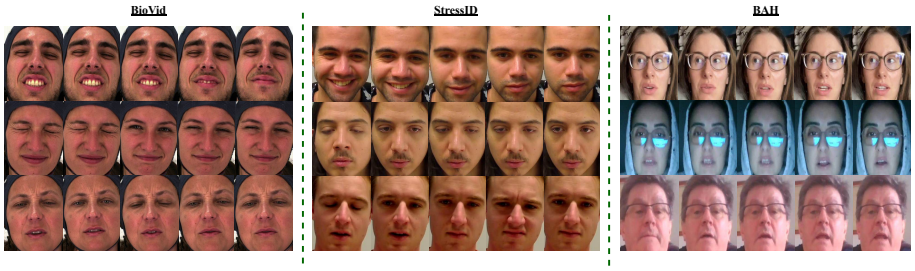

Fig. 6: Samples from BioVid, StressID, and BAH highlight increased subject variability, facial movements (e.g., speaking or head motion), and environmental factors that can obscure subtle expression cues.

moderate performance under pretraining but collapses under TTA, most notably on *StressID*, with nearly a 15% drop. This demonstrates that class-level textual cues lack the granularity needed to model subtle, individualized expressions. In contrast, AU-based personalization consistently improves performance by enabling the adapter to align with subject-specific AU activation patterns and fine-grained facial dynamics. See the suppl. materials for CP templates and $CP + AU$ prompts ablation.

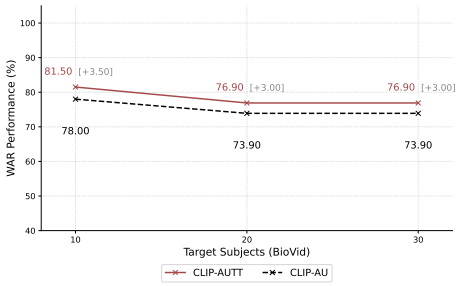

Fig. 7: Robustness analysis with increasing target subjects. Comparing `CLIP-AU` and `CLIP-AUTT`.

Robustness to Subject Variability. To evaluate robustness under increasing subject diversity, we progressively increase the number of target subjects from 10 to 30 while comparing `CLIP-AUTT` with the subject-agnostic baseline `CLIP-AU` (Figure. 7). As the target pool grows, performance naturally reflects the increased subject heterogeneity and the reduced number of source subjects

available for training. Importantly, the performance trend remains smooth and consistent, indicating stable generalization. Across all target sizes, CLIP-AUTT consistently outperforms CLIP-AU by about three WAR points, demonstrating reliable personalization benefits independent of the target pool size.

Individual Component Analysis. As shown in Table 5. We isolate the contributions of minimum-entropy window selection and AU prompt tuning. Starting from the CLIP-AU baseline, window selection alone provides modest gains of 0.7% WAR and 0.4% F1, while AU prompt tuning yields larger improvements of 2.0% WAR and 1.6% F1. Combining both components achieves the best performance, improving the baseline by 3.5% and 3.2% for WAR and F1 respectively. These results show that all the components were significantly informative, while the AU prompt tuning is the main contributor to guiding the adaptation process.

Win. Sel.	AU Tune	WAR	F1
		78.0	74.8
✓		78.7	75.2
	✓	80.0	76.4
✓	✓	**81.5**	**78.0**

Table 5: Individual component analysis on BioVid dataset.

5 Conclusion

Our study highlights a key challenge in facial expression recognition by capturing the individuality and variability of human emotional expression. While existing CLIP-based ER models rely on generic text prompts or LLM-generated descriptions to align visual features with emotion categories, they fail to model the localized AU-level cues, and the diversity of subject-specific facial behaviors. Motivated by this limitation, we introduced two approaches: CLIP-AU, which integrates fine-grained AU-aware representations, and CLIP-AUTT, which adapts AU prompts to each subject at test time. Our approach learns distinctive AU activation patterns for each subject, enabling adaptive modeling of individual expressive traits. For instance, pain may manifest through "eye tightening" and "mouth opening" in some individuals, whereas others exhibit cues such as "cheek raising" or "rapid eyelid closure." By adapting AU semantics to each subject, our method learns these personalized activation patterns without requiring subject-level supervision, enabling robust and fine-grained emotion recognition.

Acknowledgment

This research was partially supported by the Natural Sciences and Engineering Research Council of Canada, Fonds de recherche du Québec – Santé, Canada Foundation for Innovation, and the Digital Research Alliance of Canada.

References

1. Abdul Samadh, J., Gani, M.H., Hussein, N., Khattak, M.U., Naseer, M.M., Shahbaz Khan, F., Khan, S.H.: Align your prompts: Test-time prompting with dis-

tribution alignment for zero-shot generalization. Advances in Neural Information Processing Systems **36**, 80396–80413 (2023) 3, 5, 9, 10, 11, 12
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) 1
3. Aslam, M.H., Zeeshan, M.O., Belharbi, S., Pedersoli, M., Koerich, A.L., Bacon, S., Granger, E.: Distilling privileged multimodal information for expression recognition using optimal transport. In: 18th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–10 (2024) 4
4. Baltrušaitis, T., Robinson, P., Morency, L.P.: Openface: an open source facial behavior analysis toolkit. In: 2016 IEEE winter conference on applications of computer vision (WACV). pp. 1–10. IEEE (2016) 12
5. Barros, P., Parisi, G., Wermter, S.: A personalized affective memory model for improving emotion recognition. In: International Conference on Machine Learning. pp. 485–494. PMLR (2019) 4
6. Belharbi, S., Pedersoli, M., Koerich, A.L., Bacon, S., Granger, E.: Guided interpretable facial expression recognition via spatial action unit cues. In: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–10. IEEE (2024) 2, 4
7. Chaptoukaev, H., Strizhkova, V., Panariello, M., Dalpaos, B., Reka, A., Manera, V., Thümmler, S., Ismailova, E., Todisco, M., Zuluaga, M.A., et al.: Stressid: a multimodal dataset for stress identification. Advances in Neural Information Processing Systems **36**, 29798–29811 (2023) 4, 10
8. 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) 2, 4, 5
9. Cheng, Z., Cheng, Z.Q., He, J.Y., Wang, K., Lin, Y., Lian, Z., Peng, X., Hauptmann, A.: Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. Advances in Neural Information Processing Systems **37**, 110805–110853 (2024) 4
10. Ekman, P., Friesen, W.V.: Facial action coding system. Environmental Psychology & Nonverbal Behavior (1978) 4
11. Foteinopoulou, N.M., Patras, I.: 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) 2, 4, 5, 10, 11
12. González-González, M., Belharbi, S., Zeeshan, M.O., Sharafi, M., Aslam, M.H., Pedersoli, M., Koerich, A.L., Bacon, S.L., Granger, E.: Bah dataset for ambivalence/hesitancy recognition in videos for behavioural change. arXiv preprint arXiv:2505.19328 (2025) 4, 10
13. Jack, R.E., Blais, C., Scheepers, C., Schyns, P.G., Caldara, R.: Cultural confusions show that facial expressions are not universal. Current biology **19**(18), 1543–1548 (2009) 3
14. Karmanov, A., Guan, D., Lu, S., El Saddik, A., Xing, E.: Efficient test-time adaptation of vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14162–14171 (2024) 3, 5, 10, 11, 12

15. Lafon, M., Hakim, G.A.V., Rambour, C., Desrosier, C., Thome, N.: Cliptta: Robust contrastive vision-language test-time adaptation. arXiv preprint arXiv:2507.14312 (2025) 5
16. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023) 1
17. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022) 2
18. Li, S., Deng, W.: Deep emotion transfer network for cross-database facial expression recognition. In: 24th International Conference on Pattern Recognition (ICPR). pp. 3092–3099. IEEE, Piscataway, USA (2018) 4
19. Li, X., Deng, W., Li, S., Li, Y.: Compound expression recognition in-the-wild with au-assisted meta multi-task learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5735–5744 (2023) 2, 4
20. Liang, S., Zhu, M., Liu, A., Wu, B., Cao, X., Chang, E.C.: Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24645–24654 (2024) 5
21. Liang, Y., Chen, H., Xiong, Y., Zhou, Z., Lyu, M., Lin, Z., Niu, S., Zhao, S., Han, J., Ding, G.: Advancing reliable test-time adaptation of vision-language models under visual variations. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 4788–4797 (2025) 3, 5, 10, 11, 12
22. Liberatori, B., Conti, A., Rota, P., Wang, Y., Ricci, E.: Test-time zero-shot temporal action localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18720–18729 (2024) 5, 10, 11, 12
23. Lin, W., Mirza, M.J., Kozinski, M., Possegger, H., Kuehne, H., Bischof, H.: Video test-time adaptation for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22952–22961 (2023) 5
24. Liu, C., Zhang, X., Liu, X., Zhang, T., Meng, L., Liu, Y., Deng, Y., Jiang, W.: Facial expression recognition based on multi-modal features for videos in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5872–5879 (2023) 2, 4
25. Martinez, B., Valstar, M.F., Jiang, B., Pantic, M.: Automatic analysis of facial actions: A survey. IEEE transactions on affective computing **10**(3), 325–347 (2017) 2
26. Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., Ling, H.: Expanding language-image pretrained models for general video recognition. In: European Conference on Computer Vision. pp. 1–18 (2022) 10, 11
27. Osowiechi, D., Noori, M., Vargas Hakim, G., Yazdanpanah, M., Bahri, A., Cheraghalikhani, M., Dastani, S., Beizaee, F., Ayed, I., Desrosiers, C.: Watt: Weight average test time adaptation of clip. Advances in neural information processing systems **37**, 48015–48044 (2024) 5
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021) 1, 4, 10, 11
29. Rescigno, M., Spezialetti, M., Rossi, S.: Personalized models for facial emotion recognition through transfer learning. Multimedia Tools and Applications **79**(47), 35811–35828 (2020) 4

30. Sharafi, M., Belharbi, S., Zeeshan, M.O., Salem, H.B., Etemad, A., Koerich, A.L., Pedersoli, M., Bacon, S., Granger, E.: Personalized feature translation for expression recognition: An efficient source-free domain adaptation method. The Fourteenth International Conference on Learning Representations (2026) 3, 5
31. Sharafi, M., Ollivier, E., Zeeshan, M.O., Belharbi, S., Koerich, A.L., Pedersoli, M., Bacon, S., Granger, E.: Disentangled source-free personalization for facial expression recognition with neutral target data. In: 2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–10. IEEE (2025) 5
32. Shu, M., Nie, W., Huang, D.A., Yu, Z.: Test-time prompt tuning for zero-shot generalization in vision-language models. Advances in Neural Information Processing Systems **35**, 14274–14289 (2022) 3, 5, 9, 10, 11, 12
33. Waligora, P., Aslam, M.H., Zeeshan, M.O., Belharbi, S., Koerich, A.L., Pedersoli, M., Bacon, S., Granger, E.: Joint multimodal transformer for emotion recognition in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4625–4635 (2024) 4
34. Walter, S., Gruss, S., Ehleiter, H., Tan, J., Traue, H.C., Werner, P., Al-Hamadi, A., Crawcour, S., Andrade, A.O., da Silva, G.M.: The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In: 2013 IEEE international conference on cybernetics (CYBCO). pp. 128–131. IEEE (2013) 4, 10
35. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020) 5
36. Wang, Z., Chi, Z., Wu, Y., Gu, L., Liu, Z., Plataniotis, K., Wang, Y.: Distribution alignment for fully test-time adaptation with dynamic online data streams. In: European Conference on Computer Vision. pp. 332–349. Springer (2024) 5
37. Wu, W., Wang, X., Luo, H., Wang, J., Yang, Y., Ouyang, W.: Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6620–6630 (2023) 2
38. Yan, S., Xiong, X., Nagrani, A., Arnab, A., Wang, Z., Ge, W., Ross, D., Schmid, C.: Unloc: A unified framework for video localization tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13623–13633 (2023) 2
39. Yin, Y., Chang, D., Song, G., Sang, S., Zhi, T., Liu, J., Luo, L., Soleymani, M.: Fg-net: Facial action unit detection with generalizable pyramidal features. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6099–6108 (2024) 4
40. Zeeshan, M.O., Aslam, M.H., Belharbi, S., Koerich, A.L., Pedersoli, M., Bacon, S., Granger, E.: Subject-based domain adaptation for facial expression recognition. In: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–10. IEEE (2024) 4, 10
41. Zeeshan, M.O., Gillet, N., Koerich, A.L., Pedersoli, M., Bremond, F., Granger, E.: Musaco: Multimodal subject-specific selection and adaptation for expression recognition with co-training. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3606–3616 (2026) 4
42. Zeeshan, M.O., Pedersoli, M., Koerich, A.L., Granger, E.: Progressive multi-source domain adaptation for personalized facial expression recognition. IEEE Transactions on Affective Computing (2025) 3, 4

43. Zhang, C., Stepputtis, S., Sycara, K., Xie, Y.: Dual prototype evolving for test-time generalization of vision-language models. Advances in Neural Information Processing Systems **37**, 32111–32136 (2024) 10, 11, 12
44. Zhang, S., Pan, Y., Wang, J.Z.: Learning emotion representations from verbal and nonverbal communication. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18993–19004 (2023) 2, 4, 5
45. Zhang, X., Yang, H., Wang, T., Li, X., Yin, L.: Multimodal channel-mixing: Channel and spatial masked autoencoder on facial action unit detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6077–6086 (2024) 4
46. Zhao, Z., Cao, Y., Gong, S., Patras, I.: Enhancing zero-shot facial expression recognition by llm knowledge transfer. in 2025 ieee/cvf winter conference on applications of computer vision (wacv) (2025) 2, 4, 5, 10, 11
47. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 5