ALIGN-TAPER: ALIGNING VLMS WITH TEST-TIME PERSONALIZATION FOR FINE-GRAINED MULTIMODAL EMOTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video-based emotion recognition (ER) is inherently multimodal, as subtle affective states are expressed through subject-specific combinations of facial, vocal, and language cues. Test-time adaptation (TTA) methods have been successfully employed to personalize vision–language models (VLMs) during inference for facial ER. However, state-of-the-art VLMs mainly rely on visual–textual alignment and often overlook complementary information from other modalities. To address this limitation, we propose ALIGN-TAPER-FT, an efficient multimodal approach that augments visual semantics based on action units (AUs) with an acoustic cluster dictionary. Audio is incorporated as an additional modality, providing complementary acoustic cues for fine-grained ER. Specifically, source-domain acoustic representations are clustered into an audio dictionary, and a Similarity-to-Probability Dictionary (`SPD`) module maps cluster similarities directly to class probabilities without requiring a trainable audio classifier. These probabilities are fused with AU-guided visual predictions during source pretraining. A generic source-trained model is often insufficient for unseen target subjects because expressive behavior and modality reliability vary across individuals. To address this, we introduce ALIGN-TAPER-TTA, a test-time personalization method that selects the visual window whose prediction is both confident and consistent with the audio class probabilities. The synchronized audio segment is then employed with the selected visual window to adapt AU prompts and a lightweight subject-specific correction parameter. Experiments on `StressID` and `BAH` datasets show that ALIGN-TAPER achieves state-of-the-art performance on CLIP and TTA methods for ER. Results further show stronger subject-wise personalization across target subjects. Our code is included in the supplementary materials and will be released upon publication.

1 INTRODUCTION

Understanding human emotions in video is a fundamental problem in affective computing, with applications in human–computer interaction, mental health assessment, and behavioral understanding Aslam et al. (2023); Zeeshan et al. (2024); Sharafi et al. (2025); Zeeshan et al. (2025); Praveen et al. (2026). Video-based ER is inherently multimodal Praveen et al. (2022); Zeeshan et al. (2026a); Aslam et al. (2024); Tzirakis et al. (2017), as subtle emotional states are rarely expressed through a single cue. Instead, they typically emerge from complementary facial, vocal, and language patterns that vary across individuals and contexts. Furthermore, recognizing subtle and fine-grained emotions remains challenging because these cues are often low-intensity, ambiguous, and highly subject-dependent. Different individuals may express the same emotional state through different facial and acoustic patterns, making test-time personalization important for maintaining performance on unseen, unlabeled target subjects.

Recent VLMs such as CLIP have shown strong capabilities in learning aligned visual and textual representations, enabling flexible semantic reasoning and transfer to downstream ER tasks Wu et al. (2023); Li et al. (2022); Yan et al. (2023). However, most VLM-based ER methods remain limited to visual–textual alignment and can overlook complementary acoustic evidence. Several studies have incorporated AUs Martinez et al. (2017) into ER models Belharbi et al. (2024); Li et al. (2023);

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.

Liu et al. (2023). AUs provide structured and interpretable descriptions of localized facial muscle activations. More recently, Zeeshan et al. (2026b) introduced AU descriptions as structured textual prompts within CLIP, enabling fine-grained visual–semantic alignment and subject-specific prompt adaptation without requiring AU annotations or descriptions generated by large language models (LLMs). Nevertheless, existing AU-guided and TTA-based VLM methods remain restricted to the visual modality, limiting their ability to exploit acoustic cues when facial expressions are weak, ambiguous, or subject-dependent. Motivated by this, we introduce audio as an additional modality to capture vocal patterns that provide meaningful acoustic cues and complement AU-based visual semantics for fine-grained ER.

In this paper, we introduce a modality-specific acoustic alignment dictionary to combine audio cues with AU-guided visual–textual representations for fine-grained multimodal ER. AUs provide a structured semantic description of facial muscle activation patterns for the visual modality Martinez et al. (2017). Similarly, our goal is to capture recurring acoustic patterns from the audio signal that are informative for understanding a human emotional state and use them as guidance for more reliable fine-grained ER. To this end, we propose ALIGN-TAPER-FT, a multimodal source-pretraining approach that combines the AU-based visual branch with a modality-specific audio dictionary. Instead of training a large audio encoder or introducing an additional audio classifier, source-domain acoustic representations are extracted from a well-defined OpenSMILE Eyben et al. (2010; 2016) feature extractor and clustered to construct an acoustic-cluster dictionary, where each cluster prototype captures a representative audio pattern associated with emotional behavior, such as loudness variation, voice instability, reduced pitch, and unvoiced regions. Given an input audio sample, its similarities to the dictionary clusters are converted into class probabilities through the Similarity-to-Probability Dictionary (`SPD`) module and fused with the AU-guided visual prediction. Multimodal source pretraining captures generic audiovisual patterns for fine-grained ER, but it does not fully address subject-specific variability of unseen target subjects at inference. TTA methods Sharafi et al. (2025; 2026a) have been proposed for personalization of vision encoders and VLMs without access to target labels. Recently, CLIP-AUTT Zeeshan et al. (2026b) performs AU-prompt adaptation for personalized ER but remains restricted to the visual–textual modality and cannot exploit complementary acoustic cues.

To address this limitation, we propose ALIGN-TAPER-TTA, a test-time personalization method that assists AU-prompt adaptation with audio-guided window selection and subject-specific modality weighting. Given an unlabeled target video, our method first evaluates consecutive visual windows using the AU-guided visual encoder and computes the visual uncertainty of each window. The complete audio waveform is processed once through the acoustic-cluster dictionary and `SPD` module to obtain a clip-level audio prediction, which serves as acoustic guidance for all candidate visual windows. The window with the lowest multimodal selection score is chosen as the most informative region. We then extract the corresponding visual window and its synchronized audio segment, Finally, the probabilities from the selected audiovisual region are combined through late fusion. During TTA, only the AU prompts and a lightweight subject-specific fusion parameter are optimized, and both are reset to the source initialization after each target subject. This preserves the source-subject multimodal prior while allowing the model to adjust the visual–audio balance to the expressive behavior of each unseen target subject.

The main contributions of this paper are as follows: **(1)** ALIGN-TAPER-FT: a multimodal source-pretraining method that combines AU-guided visual–textual representations with a modality-specific acoustic alignment dictionary. The dictionary is constructed from source-domain acoustic clusters and integrated through the `SPD` module, enabling efficient audiovisual ER without training a large audio encoder or an additional audio classifier. **(2)** ALIGN-TAPER-TTA: an audiovisual test-time personalization method to adapt to unseen target subjects. It selects an informative synchronized audiovisual region using visual and acoustic guidance, and adapts AU prompts together with a lightweight subject-specific fusion correction initialized from the source-learned modality weight, without requiring target annotations. **(3)** An extensive set of experiments under subject-independent and test-time personalization settings for fine-grained ER. Results on challenging video-based `StressID` and `BAH` datasets demonstrate the effectiveness of the proposed ALIGN-TAPER-FT and ALIGN-TAPER-TTA methods compared with state-of-the-art emotion-recognition and TTA methods.

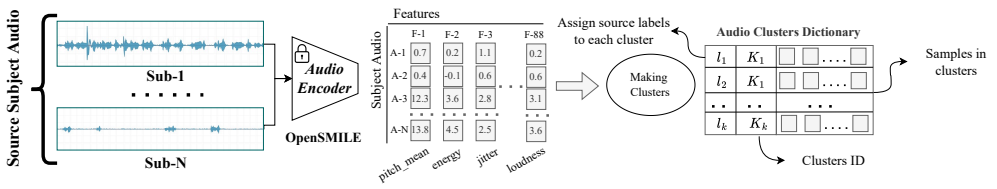


Figure 1: **Modality-Specific Audio Dictionary.** OpenSMILE-extracted acoustic features are clustered based on sample-level feature similarity to form a dictionary. Each cluster is then assigned a label based on the dominant class among its samples.

2 Related Work

CLIP-based FER Models. CLIP learns a shared visual–textual embedding space through contrastive pretraining on image–text pairs Radford et al. (2021). For ER, images are matched with emotion-related prompts, and recent methods improve this alignment using prompt engineering, richer textual descriptions, or lightweight adaptation. FineCLIPER combines adapters with fine-grained prompts for dynamic ER Chen et al. (2024) while EmoCLIP uses sample-specific descriptions for zero-shot video ER Foteinopoulou & Patras (2024). EmotionCLIP learns from verbal and nonverbal emotional signals Zhang et al. (2023); and Exp-CLIP enhances image–text alignment using LLM-generated descriptions Zhao et al. (2025). TTA-CaP introduces cache-based test-time personalization for video FER Sharafi et al. (2026b). Despite CLIP's strong transferability, these approaches generally rely on source-domain training or task-specific adaptation. Recently, CLIP-AUTT Zeeshan et al. (2026b) introduced a CLIP-based method that learns from diverse source individuals and then applies TTA to adapt to target subjects. However, it is limited to visual–textual modalities, which restricts adaptation when recognition cues are available from other modalities. In this paper, we extend this direction by incorporating audio as an additional modality to guide visual–textual representations and capture more diverse audiovisual patterns for fine-grained ER.

Audio–Guided Multimodal ER. Audio provides complementary affective information to visual cues through prosody, voice quality, intensity, and other paralinguistic characteristics. Prior methods Ortega et al. (2019); Praveen et al. (2022); Mittal et al. (2020) integrate acoustic and visual representations through feature- or decision-level fusion, joint learning, and cross-modal interactions. More recently, audio foundation models have provided general-purpose and affect-aware representations Koudounas et al. (2025); Ghosh et al. (2026), but their computational and memory costs can limit TTA. In contrast, OpenSMILE Eyben et al. (2016) provides fixed and interpretable acoustic descriptors, such as pitch, loudness, shimmer, and voice-quality measures. Rather than treating audio as a generic feature vector and combining it with visual embeddings. In this paper, we use these acoustic descriptors from OpenSmile to construct a modality-specific dictionary. The dictionary is designed to capture recurring vocal patterns that provide structured audio guidance. Similar to how CLIP-AUTT Zeeshan et al. (2026b) uses AU prompts to guide the visual modality, our acoustic-cluster dictionary guides the audio modality, enabling more informative multimodal learning for fine-grained ER.

3 Proposed Align-Taper Methodology

The goal of Align-Taper is to perform fine-grained ER while enabling efficient test-time personalization. Instead of fusing raw multimodal embeddings through concatenation or generic alignment, we build a modality-specific audio dictionary. During Align-Taper-FT, the model learns a generic audiovisual representation from diverse source subjects by using dictionary-based audio cues to guide AU-visual predictions. At test time, Align-Taper-TTA adapts the model to each target subject by tuning AU prompts with a lightweight subject-specific correction parameter.

3.1 Modality-Specific Audio Dictionary

AU prompts serve as fine-grained visual–semantic anchors for representing facial expressions. To complement these visual semantics with acoustic information, we construct a compact modality-specific audio dictionary whose features represent recurring acoustic patterns. The dictionary is

created once using the source data and remains fixed during both multimodal source training and test-time personalization, as illustrated in Fig. 1.

Acoustic feature extraction. Let $\mathcal{S}_a = \{(a_i, y_i)\}_{i=1}^{N_s}$ denote the source audio set, where a_i is the waveform associated with the i-th source video, $y_i \in \mathcal{Y}$ is its emotion label, and N_s is the number of source samples. Fixed-dimensional acoustic descriptors are extracted using OpenSMILE Eyben et al. (2016). In particular, its functional configuration summarizes prosodic, spectral, and voice-quality characteristics into an d_a-dimensional representation, where $d_a = 88$. The acoustic representation of a_i is defined as: $\mathbf{x}_i = g_{\mathrm{OS}}(a_i) \in \mathbb{R}^{d_a}$, where $g_{\mathrm{OS}}(\cdot)$ denotes the OpenSMILE audio encoder. Extracted representations are stacked to form

$$\mathbf{X}_{\mathrm{raw}} = \left[\mathbf{x}_1^\top, \ldots, \mathbf{x}_{N_s}^\top\right]^\top \in \mathbb{R}^{N_s \times d_a}. \tag{1}$$

Since acoustic descriptors have heterogeneous numerical ranges, we standardize each OpenSMILE feature using source-domain statistics before clustering. Let $\mathbf{z}_i \in \mathbb{R}^{d_a}$ denote the standardized acoustic representation of the i-th source audio sample.

Acoustic cluster construction. Rather than defining one prototype directly from each expression emotion class, we use HDBSCAN McInnes & Healy (2017) to discover K naturally occurring acoustic clusters from the standardized source representations. Low-density samples identified as noise are excluded from dictionary construction. For each valid cluster k, let $\mathcal{I}_k$ denote the set of source samples assigned to that cluster. A cluster prototype is obtained by averaging the standardized representations in $\mathcal{I}_k$, followed by ℓ_2-normalization for cosine-based dictionary matching:

$$\widetilde{\mathbf{d}}_k = \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \mathbf{z}_i, \qquad \mathbf{d}_k = \frac{\widetilde{\mathbf{d}}_k}{\left\|\widetilde{\mathbf{d}}_k\right\|_2}. \tag{2}$$

Each normalized prototype summarizes a recurring source-domain acoustic pattern associated with a discovered cluster rather than an individual training sample. Afterwards, each cluster is associated with an emotion class using the majority label among the samples assigned to that cluster:

$$\ell_k = \arg\max_{c \in \mathcal{Y}} \sum_{i \in \mathcal{I}_k} \mathbb{1}[y_i = c], \tag{3}$$

where $\mathbb{1}[\cdot]$ is the indicator function. This preserves label-independent discovery of acoustic structures while providing the class association required for dictionary-based emotion prediction. The resulting audio dictionary is defined as

$$\mathcal{D}_a = \{(\mathbf{d}_k, \ell_k)\}_{k=1}^{K}, \qquad \mathbf{D}_a = \left[\mathbf{d}_1^\top, \ldots, \mathbf{d}_K^\top\right]^\top \in \mathbb{R}^{K \times d_a}. \tag{4}$$

Together with the AU-based visual-semantic representations, the audio dictionary forms the multimodal representation used in the subsequent source-training and test-time personalization stages.

3.2 Multimodal Source Pretraining

After constructing the source-domain audio dictionary, we integrate it with the AU-guided visual representation for multimodal source pretraining. As shown in Fig. 2, the proposed model consists of a trainable visual classification head and a fixed dictionary-based audio branch. No separate audio classifier is introduced. Instead, the similarities between an audio sample and the dictionary clusters are converted into class probabilities using the Similarity-to-Probability Dictionary (`SPD`) module, based on the class association of each cluster. The resulting visual and audio predictions are then combined through late fusion.

Let $\mathcal{V}_i^s = \{I_{i,f}\}_{f=1}^{F_i}$ denote the i-th source video with F_i frames. Following Zeeshan et al. (2026b), ALIGN-TAPER-FT uses frozen CLIP visual and text encoders to compute similarities between video frames and AU prompts. The frame-level visual features are temporally aggregated and compared with AU text embeddings to obtain an AU-similarity representation, which is then mapped by a lightweight visual head to emotion-class probabilities, $\mathbf{p}_i^v = h_\theta\left(\mathcal{V}_i^s\right)$, where $\mathbf{p}_i^v \in \mathbb{R}^C$ denotes the visual prediction over C emotion classes. The CLIP encoders remain frozen, while the lightweight temporal visual head is optimized on the source domain.

Similarity-to-Probability Dictionary. For the synchronized source audio a_i, OpenSMILE features are extracted and standardized using the source-domain statistics. The resulting representation $\mathbf{z}_i$ is ℓ_2-normalized and compared with every cluster in the fixed audio dictionary, $\mathbf{s}_i^a = \mathbf{z}_i \mathbf{D}_a^\top \in \mathbb{R}^K$, where $s_{i,k}^a$ denotes the cosine similarity between the input audio representation and the prototype of the k-th acoustic cluster. Let $\ell_k \in \mathcal{Y}$ denote the emotion class associated with cluster k, and let $\mathcal{K}_c = \{k \mid \ell_k = c\}$ denote the set of clusters associated with class c. The `SPD` module averages the similarities of all clusters associated with the same class to obtain the class-level audio score:

$$o_{i,c}^a = \frac{1}{|\mathcal{K}_c|} \sum_{k \in \mathcal{K}_c} s_{i,k}^a. \tag{5}$$

The resulting class scores are converted into audio probabilities as $\mathbf{p}_i^a$. Since the scores are obtained directly from the fixed acoustic-cluster dictionary, the audio branch does not require a trainable audio encoder or classifier.

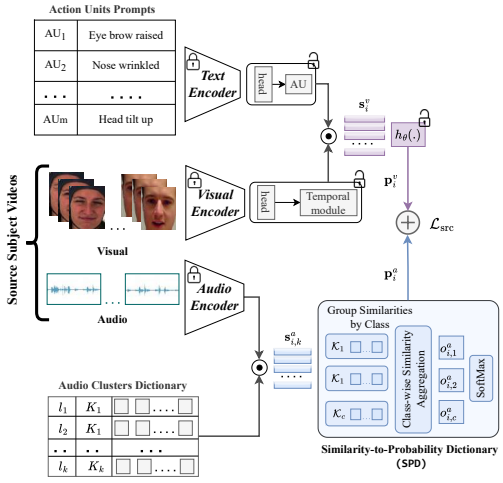


Figure 2: **Overview of ALIGN-TAPER-FT.** The AU-guided visual model uses frozen CLIP and a learnable temporal classifier for emotion prediction. In parallel, audio features are compared with a fixed acoustic-cluster dictionary and converted into class-level probabilities through an `SPD` module. Visual and audio predictions are combined through late fusion and trained with the supervised loss.

The visual and audio class probabilities are combined through weighted late fusion, $\mathbf{p}_i^{av} = (1 - \alpha_s)\mathbf{p}_i^v + \alpha_s \mathbf{p}_i^a$, where $\alpha_s \in [0, 1]$ controls the contribution of the audio modality and $1 - \alpha_s$ controls the visual contribution. We parameterize the source fusion weight as, $\alpha_s = \mathrm{sigmoid}(\beta_s)$, where β_s is a single learnable scalar. This provides a source-derived modality prior that is subsequently used to initialize subject-specific fusion during test-time personalization.

Given source labels y_i, the model is optimized by the negative log-likelihood of the fused prediction:

$$\mathcal{L}_{\mathrm{src}} = -\frac{1}{N_s} \sum_{i=1}^{N_s} \log p_{i,y_i}^{av}. \tag{6}$$

Only the lightweight visual head and scalar fusion parameter are optimized, while the CLIP encoders, AU text representations, OpenSMILE extractor, and audio dictionary remain fixed.

3.3 MULTIMODAL TEST-TIME PERSONALIZATION

Although multimodal source pretraining learns a generic audiovisual model, facial and vocal expressions can vary considerably across individuals. We therefore introduced a test-time personalization method to jointly adapt the AU prompts and the relative contribution of the visual and audio modalities. As shown in Fig. 3 for an unseen target subject, the proposed method first identifies the most informative temporal window using multimodal guidance and extracts the corresponding synchronized visual and audio segments. The selected audiovisual region is then used to adapt the AU prompts and subject-specific fusion weights without requiring target annotations.

Multimodal-guided informative window selection. Let $\mathcal{V}_i^t = \{I_{i,f}\}_{f=0}^{F_i}$ and a_i denote the visual sequence and synchronized audio waveform of the i-th target video. Given a window length L, overlapping visual windows are defined as

$$\mathcal{W}_{i,m} = \{I_{i,f}\}_{f=m}^{m+L-1}, \qquad m = 0, \ldots, T_i - L. \tag{7}$$

Each candidate window is processed by the AU-guided visual branch to obtain the class probability $\mathbf{p}_{i,m}^v$. Its informativeness is measured using normalized prediction entropy:

$$\mathcal{H}_{i,m}^v = -\frac{1}{\log C} \sum_{c=1}^{C} p_{i,m,c}^v \log\left(p_{i,m,c}^v + \epsilon\right), \tag{8}$$

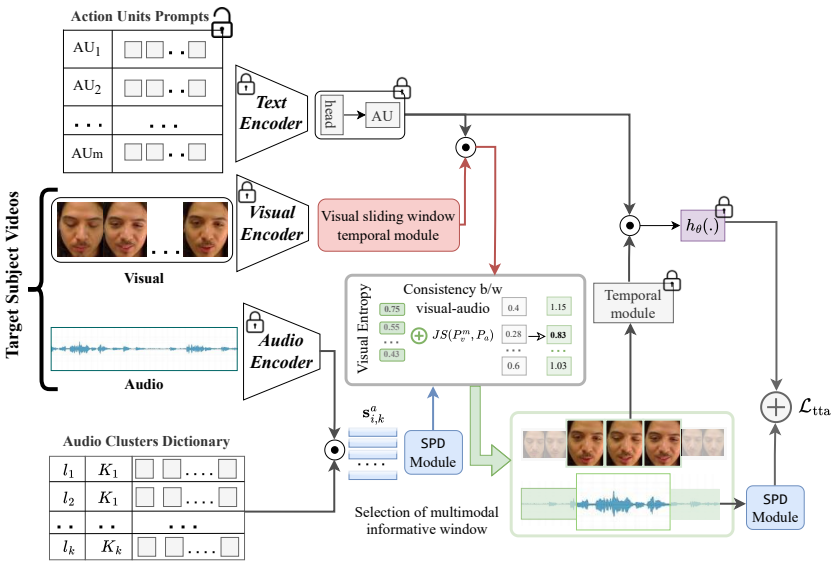


Figure 3: **Overview of ALIGN-TAPER-TTA.** Visual windows are ranked using visual entropy and audio-guided JS divergence. The lowest-score window and its synchronized audio segment are selected for personalized late fusion, where AU prompts and subject-specific modality weights are adapted while the remaining components are frozen.

where C denotes the number of emotion classes. The complete audio waveform is processed only once using OpenSMILE, the fixed acoustic-cluster dictionary, and the `SPD` module, producing a clip-level audio probability $\mathbf{p}_i^a = f_a(a_i; \mathcal{D}_a)$. This probability is shared across all candidate visual windows. To measure the agreement between each visual window and the audio prediction, we compute the normalized Jensen–Shannon divergence:

$$\mathbf{m}_{i,m} = \frac{1}{2}\left(\mathbf{p}_{i,m}^v + \mathbf{p}_i^a\right), \qquad \mathcal{D}_{i,m}^{av} = \frac{\mathrm{KL}\left(\mathbf{p}_{i,m}^v \,\|\, \mathbf{m}_{i,m}\right) + \mathrm{KL}\left(\mathbf{p}_i^a \,\|\, \mathbf{m}_{i,m}\right)}{2\log 2} \tag{9}$$

A low divergence indicates agreement between the visual and audio-acoustic prediction. The final selection score combines the visual score with audiovisual consistency:

$$\mathcal{S}_{i,m} = \mathcal{H}_{i,m}^v + \lambda_{\mathrm{a}} \mathcal{D}_{i,m}^{av}, \qquad m_i^* = \arg\min_m \mathcal{S}_{i,m}, \tag{10}$$

where λ_{a} controls the strength of audio guidance. Thus, the selected window is visually confident while remaining consistent with the overall acoustic prediction.

Synchronized audiovisual region. After selecting the informative visual window, the corresponding audio segment is extracted from the same temporal interval. Let S_i denote the number of samples in the complete waveform. The synchronized audiovisual region is defined as

$$\begin{aligned} s_i^{\text{start}} &= \left\lfloor \frac{m_i^*}{T_i} S_i \right\rfloor, \qquad s_i^{\text{end}} = \left\lfloor \frac{m_i^* + L}{T_i} S_i \right\rfloor, \\ \mathcal{W}_i^* &= \mathcal{W}_{i,m_i^*}, \qquad a_i^* = a_i[s_i^{\text{start}} : s_i^{\text{end}}]. \end{aligned} \tag{11}$$

The selected visual window $\mathcal{W}_i^*$ and audio segment a_i^* define the same temporal region. For window-selection guidance, however, we use the full-video audio prediction, since short audio chunks may lose broader acoustic context such as pauses, loudness variation, and spectral patterns. A synchronized-local-audio ablation is reported in the appendix.

Subject-specific multimodal adaptation. The AU prompts are initialized from the source model and adapted to the target subject. In addition, we personalize the contribution of the audio modality during late fusion. During source pretraining, we learn a scalar fusion parameter β_s, which defines the source-domain audio weight as $\alpha_s = \sigma(\beta_s)$, where $\sigma(\cdot)$ denotes the sigmoid function. For each

target subject u, we introduce a lightweight learnable parameter δ_u, and compute the subject-specific audio weight as, $\alpha_u = \sigma(\beta_s + \delta_u)$. Thus, the source-learned fusion parameter is retained as a prior, while δ_u allows the model to adjust the visual–audio balance for the target subject.

Given the selected visual window and its synchronized audio segment, the visual and audio produce $\mathbf{p}_i^{v,*}$ and $\mathbf{p}_i^{a,*}$, respectively. The final prediction is obtained through subject-specific late fusion: $\mathbf{p}_i^* = (1-\alpha_u)\mathbf{p}_i^{v,*} + \alpha_u \mathbf{p}_i^{a,*}$. Here, α_u controls the contribution of the audio modality, while $1-\alpha_u$ controls the contribution of the visual modality. The target model is optimized by minimizing the prediction entropy over the unlabeled videos:

$$\mathcal{L}_{\text{tta}} = -\frac{1}{N_u} \sum_{i=1}^{N_u} \sum_{c=1}^{C} p_{i,c}^* \log\left(p_{i,c}^* + \epsilon\right), \tag{12}$$

where N_u is the number of test videos available for the target subject, and only the AU prompts and the subject-specific correction δ_u are updated.

4 Results and Discussion

Datasets. ALIGN-TAPER is evaluated on two challenging datasets for video-based fine-grained multimodal ER. `StressID` Chaptoukaev et al. (2023) contains synchronized audiovisual recordings with 44 subjects as the source and 10 subjects as the target. Behavioural Ambivalence/Hesitancy (`BAH`) González-González et al. (2025) contains 143 subjects as the source domain and retains 10 unseen subjects for TTA. More details on the dataset and target subjects are in the appendix.

Implementation details. The visual module follows the protocol of Zeeshan et al. (2026b). For audio, we extract 88-dimensional eGeMAPSv02 descriptors using OpenSMILE Eyben et al. (2016). During ALIGN-TAPER-FT, video clips are processed with a batch size of 8, and all trainable parameters are optimized using AdamW with a learning rate of 1×10^{-3} and weight decay of 1×10^{-4}. For ALIGN-TAPER-TTA, personalization is performed independently for each target subject, using a batch size of 1 to process one video at a time. The model is reset to its source-initialized state after each subject. For ALIGN-TAPER-TTA, we use an episodic subject-level TTA protocol. Personalization is performed independently for each target subject, which may contain one or multiple videos. Videos are processed sequentially with batch size 1, allowing adaptation even when only a single target video is available, and the model is reset to its source-initialized state after each subject. All TTA baselines follow the same data-access protocol, using only unlabeled target videos from the current subject. We report Weighted Average Recall (`WAR`) and `F1`. Following Zeeshan et al. (2026b), we compare with ER methods under fine-tuning and with established baselines under TTA.

4.1 Comparison with State-of-the-Art Methods

Fine-tuning. Table 1 shows that ALIGN-TAPER-FT substantially improves over the CLIP-based ER baselines on `StressID`. It achieves 76.0% `WAR` and 70.4% `F1`-score, outperforming the strongest fine-tuning baseline by 9.5% and 11.9%, respectively. A similar trend is observed on `BAH`, where ALIGN-TAPER-TTA improves `WAR` by 1.7% and `F1` by 19.1%. The substantial gain in `F1` highlights the ability of ALIGN-TAPER-TTA to better handle difficult and imbalanced classes. These gains show that the acoustic-cluster dictionary complements AU-guided visual cues without training an audio encoder or an additional audio classifier.

Test-time adaptation. Among the TTA methods in Table 1, ALIGN-TAPER-TTA achieves the best performance on `StressID`, reaching 81.9% `WAR` and 79.4% `F1`-score. It improves over CLIP-AUTT by 1.1% in `WAR` and 1.5% in `F1`-score. On `BAH`, we observe a similar trend, with ALIGN-TAPER-TTA improving `WAR` by 2.9% and `F1` by 29.1%. The large gain in `F1` indicates that ALIGN-TAPER-TTA better handles difficult and imbalanced classes. This suggests that audio-guided informative-window selection and subject-specific modality weighting complement AU-prompt adaptation by capturing variations in both facial and acoustic expression patterns.

Personalization. Table 2 reports the subject-wise results for `StressID` target subjects. ALIGN-TAPER-TTA achieves the best performance across 6 subjects, including perfect `WAR` and `F1`-scores for *Sub-2* and *Sub-6*. The method also provides substantial improvements for more challenging subjects. For example, on *Sub-3*, ALIGN-TAPER-TTA improves over CLIP-AUTT from $54.6\%/47.6\%$

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

Setting	Method	StressID		BAH	
		WAR ↑	F1 ↑	WAR ↑	F1 ↑
ZS	CLIP-ViT-B/32 Radford et al. (2021) (ICML'21)	60.4	34.8	39.5	28.1
FT	CLIP-ViT-B/32† Radford et al. (2021) (ICML'21)	67.0	44.5	60.4	39.8
	EmoCLIP Foteinopoulou & Patras (2024) (FG'24)	63.5	35.9	56.2	36.5
	X-CLIP Ni et al. (2022) (ECCV'22)	62.3	41.3	63.0	39.2
	Exp-CLIP Zhao et al. (2025) (WACV'25)	63.1	44.5	62.2	38.5
	CLIP-AU Zeeshan et al. (2026b) (ECCV'26)	66.5	58.5	68.3	40.3
	ALIGN-TAPER-FT	**76.0**	**70.4**	**70.0**	**59.4**
TTA	TPT Shu et al. (2022) (NeurIPS'22)	70.9	57.9	65.6	39.7
	TDA Karmanov et al. (2024) (CVPR'24)	69.7	49.9	65.2	39.9
	DPE Zhang et al. (2024) (NeurIPS'24)	71.3	54.2	66.7	39.4
	PromptAlign Abdul Samadh et al. (2023) (NeurIPS'23)	74.6	53.2	67.1	39.7
	ReTA Liang et al. (2025) (ACMMM'25)	71.8	52.8	67.6	39.8
	T3AL Liberatori et al. (2024) (CVPR'24)	75.9	59.4	67.9	40.7
	CLIP-AUTT Zeeshan et al. (2026b) (ECCV'26)	80.8	77.9	69.8	41.1
	ALIGN-TAPER-TTA	**81.9**	**79.4**	**72.7**	**70.2**

Table 2: Subject-wise comparison of ALIGN-TAPER-TTA with existing TTA methods on six target subjects from the StressID dataset. Best results are highlighted in bold.

Method	Men						Women					
	Sub-1		Sub-2		Sub-3		Sub-4		Sub-5		Sub-6	
	WAR	F1	WAR	F1	WAR	F1	WAR	F1	WAR	F1	WAR	F1
TPT Shu et al. (2022)	51.6	43.6	90.0	86.0	52.2	36.4	66.6	41.1	63.4	50.0	91.9	50.0
TDA Karmanov et al. (2024)	45.2	40.0	75.6	50.1	49.6	35.0	41.2	29.6	75.3	45.6	91.9	50.0
DPE Zhang et al. (2024)	55.0	40.6	73.9	49.0	52.6	49.8	50.9	56.0	73.0	46.0	92.0	56.6
PromptAlign Abdul Samadh et al. (2023)	61.6	45.0	78.6	45.6	51.2	48.1	60.4	44.0	81.3	58.0	90.9	45.5
ReTA Liang et al. (2025)	50.5	41.0	82.0	53.0	46.8	36.1	49.3	39.1	80.0	53.3	93.7	59.3
T3AL Liberatori et al. (2024)	62.7	48.6	80.3	48.6	54.0	51.0	68.0	61.8	77.0	68.0	93.9	66.0
CLIP-AUTT Zeeshan et al. (2026b)	63.6	63.3	90.9	90.6	54.6	47.6	**72.7**	61.1	81.8	80.3	**100.0**	**100.0**
ALIGN-TAPER-TTA	**72.7**	**71.8**	**100.0**	**100.0**	**72.7**	**71.8**	**72.7**	**71.8**	**90.9**	**89.5**	**100.0**	**100.0**

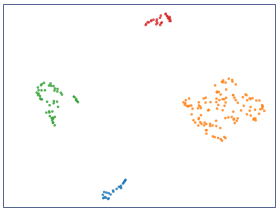
to 72.7%/71.8% in WAR/F1. On *Sub-4*, both methods obtain 72.7% WAR, while ALIGN-TAPER-TTA raises the F1 from 61.1% to 71.8%, indicating more balanced class-level predictions. The consistent subject-wise gains demonstrate the benefit of learning the subject-specific correction parameter while adapting the AU prompts to each target individual. Additional subject-wise results and BAH dataset results are presented in the appendix.

Interpretability Analysis of the Acoustic Dictionary. We analyze the acoustic clusters learned from StressID audio to assess whether the dictionary captures meaningful patterns. As shown in Fig. 4 (Left), the dictionary forms two stress-dominant and two neutral-dominant clusters, capturing diverse acoustic sub-patterns within each class. Since eGeMAPS Eyben et al. (2016) provides interpretable vocal descriptors, the stress clusters are mainly associated with spectral energy, loudness variability, spectral slope, and loudness level, reflecting changes in vocal intensity and spectral structure. In contrast, the neutral clusters are linked to spectral balance, pitch level, unvoiced duration, and reduced pitch, capturing calmer voice and pause-related patterns. This shows that the acoustic dictionary captures diverse class-specific cues, improving robustness to subject-specific vocal patterns.

4.2 ABLATION STUDIES

Audio–Visual Fusion Techniques. To evaluate the contribution of the proposed modality-specific audio dictionary, we compare it with common audiovisual fusion strategies, including concatenation,

(a) Acoustic dictionary interpretability



Class	Top Audio Acoustic Cues
Stress	Spectral energy, loudness variability
Stress	Spectral slope, loudness level
Neutral	Spectral balance, pitch level
Neutral	Unvoiced duration, reduced pitch

(b) Audio–visual training ablation

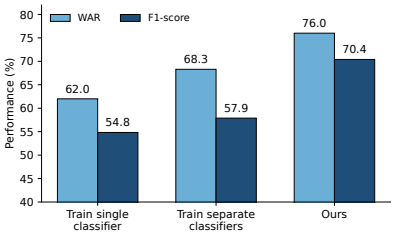


Figure 4: Analysis of the proposed acoustic dictionary and audio–visual training strategy. **(a)** UMAP visualization of learned OpenSMILE acoustic clusters with their dominant class and main acoustic cues. **(b)** Ablation on audio–visual training strategies under the same training setting.

gated fusion, and cross-attention Tzirakis et al. (2017); Praveen et al. (2022); Aslam et al. (2024). As shown in Table 3 (Left), cross-attention and gated fusion provide limited gains. Cross-attention depends on reliable frame–audio correspondence, which is difficult with compact acoustic features, while gated fusion may learn unstable modality weights and over-emphasize noisy audio cues. Concatenation performs better by adding acoustic information with fewer assumptions, but still treats audio as a generic feature stream. In contrast, ALIGN-TAPER-FT improves WAR by 1.6% and F1-score by 1.4% over concatenation, suggesting that the modality-specific audio dictionary better captures the relationship between acoustic patterns and emotion classes, thereby providing more effective guidance to the visual modality for fine-grained ER.

Component Analysis of ALIGN-TAPER-TTA. Table 3 (Right) analyzes the two main components of ALIGN-TAPER-TTA. AU prompt tuning alone improves the model by adapting the visual–textual representation to the target subject, reaching 77.0% `WAR` and 72.03% `F1`. Window selection alone achieves stronger performance, indicating that selecting an informative audiovisual region provides a reliable adaptation signal. Combining AU prompt tuning with window selection gives the best result, improving performance to 81.9% `WAR` and 79.40% `F1`. These results show that prompt adaptation and informative-window selection are complementary for subject-specific personalization.

Training Strategies for Visual–Audio Learning. We evaluate different visual–audio training strategies. As shown in Fig. 4 (Right), directly concatenating AU–visual and audio–dictionary similarities performs poorly because the higher-dimensional visual vector dominates the classifier. Training separate visual and audio classifiers with late fusion improves performance, but the audio classifier remains weak due to the limited dictionary-similarity input. In contrast, ALIGN-TAPER-FT keeps the acoustic dictionary fixed and directly maps audio–cluster similarities to class probabilities through `SPD`, providing more stable and reliable audio guidance.

Table 3: **Left:** Audio–visual fusion techniques v/s ALIGN-TAPER-FT. **Right:** Component analysis of ALIGN-TAPER-TTA.

Fusion Type	WAR	F1
Gated fusion	60.1	51.3
Cross-attention	64.3	51.3
Concatenation	74.4	69.0
ALIGN-TAPER-FT	76.0	70.4

AU	Win.	WAR	F1
✓		77.0	72.0
	✓	80.2	74.1
✓	✓	81.9	79.4

5 CONCLUSION

We propose ALIGN-TAPER, an efficient multimodal test-time personalization approach for video-based ER that combines AU-guided visual semantics with an audio acoustic dictionary. The modality-specific audio dictionary is developed to exploit complementary non-visual cues when facial expressions are subtle or ambiguous, without relying on large trainable audio encoders or generic feature fusion. In ALIGN-TAPER-FT, source-subject audio representations are clustered to construct an acoustic-cluster dictionary, where each cluster captures recurring acoustic patterns relevant to emotional states. Additionally, to address subject-specific variability at inference, ALIGN-TAPER-TTA selects audiovisual regions whose visual predictions are confident and consistent with audio evidence, then adapts AU prompts together with a lightweight subject-specific correction parameter. Experiments on `StressID` and `BAH` show that ALIGN-TAPER achieves state-of-the-art results among CLIP-based ER and TTA methods, with stronger subject-wise personalization.

AI USE STATEMENT

In this work, we used AI tools to assist with manuscript editing, improving readability, and grammar checks. We have not used generative AI tools to generate datasets, fabricate experimental results, or make any sort of final scientific claims. All AI-assisted text, code, figures, and interpretations were reviewed, edited, and verified by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide implementation details in the main paper and supplementary material, including the model algorithm, audio-dictionary construction, HDBSCAN hyperparameters, optimization settings, evaluation metrics, subject-independent source/target splits, and source code. The supplementary material also describes the OpenSMILE eGeMAPSv02 acoustic features, preprocessing steps, and additional subject-wise results. Source code is included in the supplementary materials and will be made publicly available to support reproducibility.

REFERENCES

Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman H Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. *Advances in Neural Information Processing Systems*, 36:80396–80413, 2023.

Muhammad Haseeb Aslam, Muhammad Osama Zeeshan, Marco Pedersoli, Alessandro L Koerich, Simon Bacon, and Eric Granger. Privileged knowledge distillation for dimensional emotion recognition in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3338–3347, 2023.

Muhammad Haseeb Aslam, Muhammad Osama Zeeshan, Soufiane Belharbi, Marco Pedersoli, Alessandro Lameiras Koerich, Simon Bacon, and Eric Granger. 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.

Soufiane Belharbi, Marco Pedersoli, Alessandro Lameiras Koerich, Simon Bacon, and Eric Granger. 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.

Hava Chaptoukaev, Valeriya Strizhkova, Michele Panariello, Bianca Dalpaos, Aglind Reka, Valeria Manera, Susanne Thümmler, Esma Ismailova, Massimiliano Todisco, Maria A Zuluaga, et al. Stressid: a multimodal dataset for stress identification. *Advances in Neural Information Processing Systems*, 36:29798–29811, 2023.

Haodong Chen, Haojian Huang, Junhao Dong, Mingzhe Zheng, and Dian Shao. Finecliper: Multimodal fine-grained clip for dynamic facial expression recognition with adapters. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 2301–2310, 2024.

Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010.

Florian Eyben, Klaus R Scherer, Bjorn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2016.

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.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Advances in Neural Information Processing Systems*, 38:41819–41886, 2026.

Manuela González-González, Soufiane Belharbi, Muhammad Osama Zeeshan, Masoumeh Sharafi, Muhammad Haseeb Aslam, Marco Pedersoli, Alessandro Lameiras Koerich, Simon L Bacon, and Eric Granger. Bah dataset for ambivalence/hesitancy recognition in videos for behavioural change. *arXiv preprint arXiv:2505.19328*, 2025.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. 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.

Alkis Koudounas, Moreno La Quatra, Sabato Marco Siniscalchi, and Elena Baralis. voc2vec: A foundation model for non-verbal vocalization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Ximan Li, Weihong Deng, Shan Li, and Yong Li. 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.

Yiwen Liang, Hui Chen, Yizhe Xiong, Zihan Zhou, Mengyao Lyu, Zijia Lin, Shuaicheng Niu, Sicheng Zhao, Jungong Han, and Guiguang Ding. Advancing reliable test-time adaptation of vision-language models under visual variations. *arXiv preprint arXiv:2507.09500*, 2025.

Benedetta Liberatori, Alessandro Conti, Paolo Rota, Yiming Wang, and Elisa Ricci. Test-time zero-shot temporal action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18720–18729, 2024.

Chuanhe Liu, Xinjie Zhang, Xiaolong Liu, Tenggan Zhang, Liyu Meng, Yuchen Liu, Yuanyuan Deng, and Wenqiang Jiang. 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.

Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing*, 10(3):325–347, 2017.

Leland McInnes and John Healy. Accelerated hierarchical density based clustering. In *2017 IEEE international conference on data mining workshops (ICDMW)*, pp. 33–42. IEEE, 2017.

Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 1359–1367, 2020.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pp. 1–18, 2022.

Juan DS Ortega, Patrick Cardinal, and Alessandro L Koerich. Emotion recognition using fusion of audio and video features. *arXiv preprint arXiv:1906.10623*, 2019.

R Gnana Praveen, Wheidima Carneiro de Melo, Nasib Ullah, Haseeb Aslam, Osama Zeeshan, Théo Denorme, Marco Pedersoli, Alessandro L Koerich, Simon Bacon, Patrick Cardinal, et al. A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. In *Proceedings of the IEEE/CVF CVPR*, pp. 2486–2495, 2022.

R Gnana Praveen, Patrick Cardinal, and Eric Granger. Weakly supervised learning for facial affective behavior analysis: A review. *IEEE Transactions on Affective Computing*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Masoumeh Sharafi, Emma Ollivier, Muhammad Osama Zeeshan, Soufiane Belharbi, Alessandro Lameiras Koerich, Marco Pedersoli, Simon Bacon, and Eric Granger. 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.

Masoumeh Sharafi, Soufiane Belharbi, Muhammad Osama Zeeshan, Houssem Ben Salem, Ali Etemad, Alessandro Lameiras Koerich, Marco Pedersoli, Simon Bacon, and Eric Granger. Personalized feature translation for expression recognition: An efficient source-free domain adaptation method. *The Fourteenth International Conference on Learning Representations*, 2026a.

Masoumeh Sharafi, Muhammad Osama Zeeshan, Soufiane Belharbi, Alessandro Lameiras Koerich, Marco Pedersoli, and Eric Granger. Test-time adaptation via cache personalization for facial expression recognition in videos. *arXiv preprint arXiv:2603.21309*, 2026b.

Manli Shu, Weili Nie, De-An Huang, and Zhiding Yu. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35: 14274–14289, 2022.

Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *Journal of selected topics in signal processing*, 11(8):1301–1309, 2017.

Wenhao Wu, Xiaohan Wang, Haipeng Luo, Jingdong Wang, Yi Yang, and Wanli Ouyang. 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.

Shen Yan, Xuehan Xiong, Arsha Nagrani, Anurag Arnab, Zhonghao Wang, Weina Ge, David Ross, and Cordelia Schmid. Unloc: A unified framework for video localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13623–13633, 2023.

Muhammad Osama Zeeshan, Muhammad Haseeb Aslam, Soufiane Belharbi, Alessandro Lameiras Koerich, Marco Pedersoli, Simon Bacon, and Eric Granger. 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.

Muhammad Osama Zeeshan, Marco Pedersoli, Alessandro Lameiras Koerich, and Eric Granger. Progressive multi-source domain adaptation for personalized facial expression recognition. *IEEE Transactions on Affective Computing*, 2025.

Muhammad Osama Zeeshan, Natacha Gillet, Alessandro Lameiras Koerich, Marco Pedersoli, Francois Bremond, and Eric Granger. Musaco: Multimodal subject-specific selection and adaptation for expression recognition with co-training. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3606–3616. IEEE, 2026a.

Muhammad Osama Zeeshan, Masoumeh Sharafi, Benoît Savary, Alessandro Lameiras Koerich, Marco Pedersoli, and Eric Granger. Clip-autt: Test-time personalization with action unit prompting for fine-grained video emotion recognition. In *European Conference on Computer Vision*, pp. 680–698. Springer, 2026b.

Ce Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Dual prototype evolving for test-time generalization of vision-language models. *Advances in Neural Information Processing Systems*, 37:32111–32136, 2024.

Sitao Zhang, Yimu Pan, and James Z Wang. 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.

Zengqun Zhao, Yu Cao, Shaogang Gong, and Ioannis Patras. Enhancing zero-shot facial expression recognition by llm knowledge transfer. in 2025 ieee/cvf winter conference on applications of computer vision (wacv), 2025.

maketitle

APPENDIX

A. Dataset Details 13
A.1 StressID Dataset 13
A.2 BAH Dataset 13
A.3 Video-preprocessing 13
B. Implementation Detail 14
B.1 Algorithm: Modality-Specific Audio Dictionary 14
B.2 OpenSMILE eGeMAPS Acoustic Parameter Set 16
B.3 HDBSCAN Hyperparameter Analysis for Audio Dictionary 16
C. Additional Results on ALIGN-TAPER-TTA 16
C.1 StressID Personalization Results 16
C.2 BAH Personalization Results 16
C.3 Full-Video Audio Guidance vs. Synchronized Audio 17

A DATASET DETAILS

A.1 STRESSID

Chaptoukaev et al. (2023) is a controlled video-based stress-recognition dataset containing synchronized audio and visual recordings of participants performing 11 tasks designed to elicit subject-specific stress responses. The dataset is suitable for evaluating personalization because stress-related facial and vocal cues can vary substantially across individuals, even under the same task protocol. Following the protocol of Zeeshan et al. (2026b), we use 44 subjects for source training and reserve 10 unseen target subjects for test-time personalization: kycf, uymz, h8s1, ctzy, p9i3, 7h5u, g7r2, b9w0, r3zm, and x1q3.

A.2 BEHAVIOURAL AMBIVALENCE/HESITANCY (BAH)

González-González et al. (2025) is a challenging in-the-wild dataset for fine-grained behavioural emotion recognition. It contains audiovisual recordings of participants from diverse demographic backgrounds collected under uncontrolled, real-world conditions, leading to natural variation in facial expressions, vocal patterns, pose, illumination, and recording quality. This makes BAH particularly suitable for evaluating whether test-time personalization can adapt to unseen subjects under realistic distribution shifts. We use 143 subjects as the source domain and retain 10 unseen subjects for personalization: 82711, 82687, 82585, 82592, 82598, 82632, 82681, 82683, 82708, and 82714.

A.3 VIDEO-PREPROCESSING

Although the input data are videos, our model processes them as sequences of extracted face frames. For each video, frames are first extracted and faces are cropped before being passed to the visual encoder. In StressID, the original videos are recorded at $5fps$, and we sample frames at $1fps$. During ALIGN-TAPER-FT (source pre-training), we use a temporal window of 40 consecutive frames. For ALIGN-TAPER-TTA (test-time personalization), the multimodal-guided informative window selection module selects windows of size 16 frames. In BAH, videos are recorded at 24 fps and are typically longer video. Therefore, we use a larger temporal window of 100 consecutive frames. For both datasets, windows are generated with stride 1. During ALIGN-TAPER-FT, we randomly select consecutive frame window from each video/clip. Since each video is associated with a single emotion label, any sampled window inherits the video-level label. This random window sampling acts

Algorithm 1 Audio Dictionary Construction

Require: Labeled source audio set $\mathcal{S}_a = \{(a_i, y_i)\}_{i=1}^{N_s}$
Require: OpenSMILE extractor $g_{\mathrm{OS}}(\cdot)$; HDBSCAN
Ensure: Audio dictionary $\mathcal{D}_a = \{(\mathbf{d}_k, \ell_k)\}_{k=1}^{K}$ and cluster index sets $\{\mathcal{I}_k\}_{k=1}^{K}$
Initialize $\mathbf{X}_{\mathrm{raw}} \leftarrow [\,]$
for $i = 1$ to N_s **do**
    $\mathbf{x}_i \leftarrow g_{\mathrm{OS}}(a_i)$
    Append $\mathbf{x}_i$ to $\mathbf{X}_{\mathrm{raw}}$
end for
Compute source statistics $\boldsymbol{\mu}_a$ and $\boldsymbol{\sigma}_a$ from $\mathbf{X}_{\mathrm{raw}}$
for $i = 1$ to N_s **do**
    $\mathbf{z}_i \leftarrow (\mathbf{x}_i - \boldsymbol{\mu}_a)/(\boldsymbol{\sigma}_a + \epsilon)$
end for
$\{\mathcal{I}_k\}_{k=1}^{K} \leftarrow \mathrm{HDBSCAN}(\{\mathbf{z}_i\}_{i=1}^{N_s})$
Remove low-density noise samples from the discovered clusters
Initialize $\mathcal{D}_a \leftarrow \emptyset$
for $k = 1$ to K **do**
    Compute cluster prototype: $\widetilde{\mathbf{d}}_k \leftarrow \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \mathbf{z}_i$
    Normalize prototype: $\mathbf{d}_k \leftarrow \widetilde{\mathbf{d}}_k / \|\widetilde{\mathbf{d}}_k\|_2$
    Assign cluster label from source labels: $\ell_k \leftarrow \arg\max_{q \in \mathcal{Y}} \sum_{i \in \mathcal{I}_k} \mathbb{1}[y_i = q]$
    $\mathcal{D}_a \leftarrow \mathcal{D}_a \cup \{(\mathbf{d}_k, \ell_k)\}$
end for
return $\mathcal{D}_a, \{\mathcal{I}_k\}_{k=1}^{K}$

as temporal augmentation, exposing the model to different portions of the same video and reducing overfitting to a fixed temporal segment. Each frame is resized, center-cropped, converted to a tensor, and normalized using the CLIP image normalization statistics. The same preprocessing is applied to source and target frames, with antialiasing used during target resizing.

B Implementation Detail

All video inputs are processed as sequences of extracted frames rather, where each frame is resized to 224×224, center-cropped, and normalized. For the OpenSMILE audio encoder, all waveforms are resampled to $16\,\mathrm{kHz}$. During source pre-training, a random consecutive frame window is selected from each video, and the corresponding synchronized audio segment is extracted using the same temporal interval. This ensures that the visual and audio inputs describe the same portion of the video. Since each video is assigned a single emotion label, the selected window inherits the video-level label. Random window sampling also acts as temporal augmentation by exposing the model to different parts of the same video during training. All experiments are run with a fixed seed of 1999. The same seed is used for Python, NumPy, PyTorch CPU, and CUDA random number generators. We also disable cuDNN benchmarking and enable deterministic cuDNN execution when applicable. All reported results are produced using PyTorch 2.2.0 with CUDA 12.8 on an NVIDIA A100-SXM4-40GB GPU.

B.1 Algorithm: Modality-Specific Audio Dictionary

Algorithm 1 summarizes the construction of the proposed audio dictionary. Given labeled source audio samples, we first extract fixed-dimensional OpenSMILE acoustic features and standardize them using source-domain statistics. HDBSCAN is then applied to the standardized representations to discover valid acoustic clusters, while low-density noise samples are excluded. For each cluster k, the corresponding source-sample index set $\mathcal{I}_k$ is used to compute a normalized cluster prototype $\mathbf{d}_k$. The cluster is then assigned a label ℓ_k based on the dominant source class among the samples in $\mathcal{I}_k$. The final output is the audio dictionary $\mathcal{D}_a$, containing each acoustic prototype and its associated source-derived class label, together with the cluster index sets used for analysis and dictionary-based prediction.

Table 4: Grouped summary of the OpenSMILE eGeMAPS 88-dimensional acoustic parameter set used for the construction of a modality-specific audio dictionary.

#	Feature Group	Description
A. Minimalistic Low-Level Descriptors (18 LLDs)		
1	Pitch / F0	Logarithmic fundamental frequency on a semitone scale, capturing pitch level and pitch variation.
2	Jitter	Deviations in consecutive F0 period lengths, reflecting pitch-period irregularity and voice instability.
3	Formant 1–3 frequency	Center frequencies of the first, second, and third formants, describing vocal-tract resonance patterns.
4	Formant 1 bandwidth	Bandwidth of the first formant, capturing spread around the first vocal-tract resonance.
5	Shimmer	Differences in peak amplitudes of consecutive F0 periods, reflecting amplitude instability.
6	Loudness	Estimate of perceived signal intensity from the auditory spectrum, capturing vocal energy.
7	Harmonics-to-Noise Ratio	Ratio between harmonic and noise-like components, describing voice quality.
8	Alpha ratio	Ratio of energy between low-frequency and high-frequency bands, capturing spectral balance.
9	Hammarberg index	Ratio between dominant low- and high-frequency spectral peaks, reflecting spectral emphasis.
10	Spectral slope	Regression slope of the logarithmic power spectrum in low- and mid-frequency bands.
11	Formant 1–3 relative energy	Energy around the first, second, and third formant regions relative to the F0 spectral peak.
12	Harmonic difference H1–H2	Energy ratio between the first and second harmonics, related to phonation characteristics.
13	Harmonic difference H1–A3	Energy ratio between the first harmonic and the strongest harmonic in the third-formant region.
B. Functionals Applied to the 18 LLDs		
14	Mean and coefficient of variation	Arithmetic mean and coefficient of variation are applied to all 18 LLDs, producing **36 parameters**.
15	Pitch-specific functionals	Additional pitch functionals include percentiles, percentile range, and rising/falling slope statistics, adding **8 parameters** and increasing the total to **44 parameters**.
16	Loudness-specific functionals	The same additional functionals are applied to loudness, adding **8 parameters** and increasing the total to **52 parameters**.
17	Unvoiced spectral descriptors	Mean alpha ratio, Hammarberg index, and spectral slopes are computed over unvoiced segments, adding **4 parameters** and increasing the total to **56 parameters**.
C. Temporal Features		
18	Loudness peaks	Number of loudness peaks per second, adding **1 parameter**.
19	Voiced-region duration	Mean and standard deviation of continuously voiced regions, adding **2 parameters**.
20	Unvoiced-region duration	Mean and standard deviation of unvoiced regions, approximating pauses, adding **2 parameters**.
21	Voiced segments per second	Number of continuous voiced regions per second, adding **1 parameter**. Together, temporal features add **6 parameters**, increasing the total to **62 parameters**.
D. Extended eGeMAPS Descriptors		
22	MFCC 1–4	Low-order Mel-frequency cepstral coefficients, capturing coarse spectral shape and energy distribution.
23	Spectral flux	Difference between spectra of consecutive frames, capturing dynamic spectral changes.
24	Formant 2–3 bandwidth	Bandwidths of the second and third formants, completing the formant bandwidth descriptors.
25	Extended functionals	Mean and coefficient of variation are applied to the 7 extended LLDs, and additional voiced/unvoiced spectral-flux and MFCC descriptors are included, adding **25 parameters**.
26	Equivalent sound level	Overall sound pressure/intensity level of the audio signal, adding **1 parameter**. The extended set adds **26 parameters**, extending the GeMAPSv to **88 parameters**.
Total: 88 acoustic parameters		

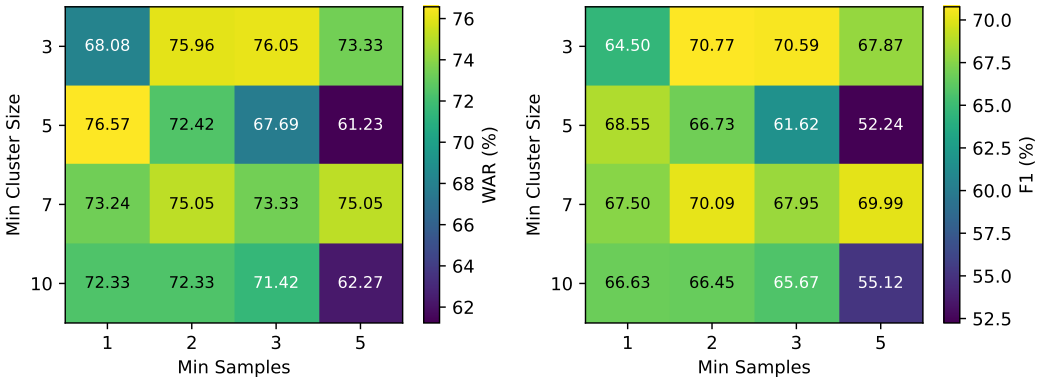


Figure 5: HDBSCAN hyperparameter analysis for audio dictionary construction on `StressID`. Each cell reports the average performance over 10 target subjects.

B.2 OpenSMILE eGeMAPS Acoustic Feature Set

Table 4 summarizes the OpenSMILE eGeMAPS Eyben et al. (2016) acoustic parameter set used in our audio dictionary creation. The set starts from 18 low-level descriptors (LLDs) that capture frequency-related, energy/amplitude, and spectral-balance properties of speech, such as pitch, jitter, shimmer, loudness, HNR, formants, spectral slope, and harmonic differences. OpenSMILE then applies statistical functionals to these descriptors, including mean, coefficient of variation, percentiles, percentile range, and rising/falling slope statistics, increasing the representation to 56 parameters. Six temporal parameters describing loudness peaks, voiced regions, unvoiced regions, and voiced-segment rate further increase the minimalistic set to 62 parameters. Finally, the extended eGeMAPS set adds MFCCs, spectral flux, formant bandwidths, additional voiced/unvoiced functionals, and equivalent sound level, resulting in 88 acoustic parameters. These parameters provide interpretable vocal cues that are relevant for emotion-related acoustic patterns. For example, pitch, loudness, jitter, shimmer, spectral balance, and voiced/unvoiced duration can reflect changes in vocal intensity, voice stability, phonation, and pause structure. In the main paper *Sec.4.2*, the *Interpretability Analysis of the Acoustic Dictionary* shows how these OpenSMILE parameters form meaningful acoustic clusters, where each cluster is associated with a dominant emotion class and characterized by its top acoustic cues. This supports the use of the acoustic-cluster dictionary as an interpretable audio representation for fine-grained ER.

B.3 HDBSCAN Hyperparameter Analysis for Audio Dictionary

To construct the acoustic dictionary, we use HDBSCAN McInnes & Healy (2017) rather than standard DBSCAN. Unlike DBSCAN, which depends on a fixed global neighborhood radius, HDBSCAN builds a hierarchy of density-based clusters and can better handle clusters with different densities. This is useful for OpenSMILE acoustic representations, where stress and neutral vocal patterns may form clusters of different sizes and densities. HDBSCAN also identifies low-density samples as noise, preventing unstable samples from being used as dictionary prototypes. We search over `min_cluster_size` and `min_samples`, which control the minimum cluster size and the conservativeness of noise assignment, respectively. Figure 5 reports average `WAR`/`F1` over the 10 `StressID` target subjects. Overall, the heatmaps show that moderate HDBSCAN settings provide the most reliable acoustic dictionaries. Very small or very conservative configurations lead to less stable performance, while increasing `min_samples` too much often reduces both `WAR` and `F1`. This suggests that assigning too many samples as low-density noise can reduce dictionary coverage and weaken the audio guidance. In contrast, moderate values preserve compact but diverse acoustic clusters, allowing the dictionary to capture recurring vocal patterns that are useful for test-time personalization.

C Additional Results on Align-Taper-TTA

C.1 StressID Personalization Results

Table 5 reports subject-wise personalization results on the 10 unseen `StressID` target subjects. Overall, Align-Taper-TTA achieves the best average performance, reaching 81.9% `WAR` and 79.4% `F1`. Compared with CLIP-AUTT, this improves the average `WAR` by 1.1% and `F1` by 1.5%. The gain is larger compared with general TTA baselines, especially over T3AL, where Align-Taper-TTA improves `WAR` by 6.0% and `F1` by 20.0%. These results show that incorporating acoustic guidance improves personalization beyond visual–textual prompt adaptation alone. The improvements are also visible at the subject level, particularly for Sub-2, Sub-3, Sub-5, Sub-8, and Sub-9, where audio-guided adaptation provides stronger or more balanced predictions.

C.2 BAH Personalization Results

Table 6 shows the subject-wise results on the more challenging `BAH` dataset. Align-Taper-TTA achieves the best average performance with 72.7% `WAR` and 70.2% `F1`, improving over CLIP-AUTT by 2.9% in `WAR` and 29.1% in `F1`. The gain in `F1` is especially important: Align-Taper-TTA outperforms all baselines on every target subject, often by a large margin. This indicates that the proposed audio dictionary is particularly effective for difficult and imbalanced target-subject

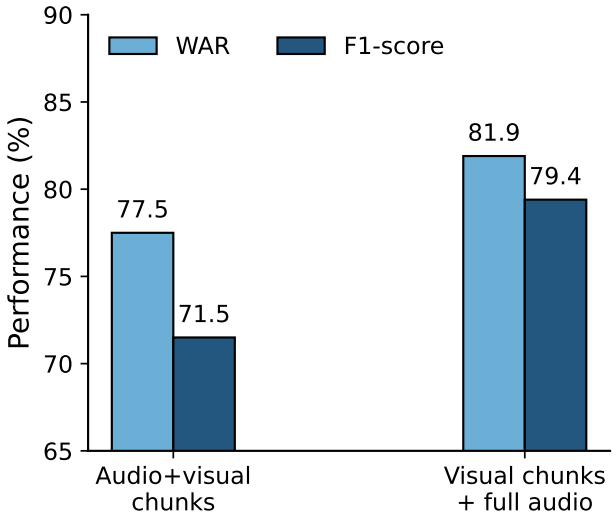


Figure 6: Ablation comparing synchronized local audio chunks with complete-video audio guidance for visual-window selection.

cases, where visual-only personalization may produce biased or unstable predictions. By capturing recurring acoustic patterns and using them as structured guidance, the added audio modality improves subject-wise personalization and leads to more reliable fine-grained ER across unseen individuals.

C.3 Full-Video Audio Guidance vs. Synchronized Audio

We further analyze whether audio should be used as synchronized local chunks or as complete-video context for guiding visual-window selection. As shown in Figure 6 In the synchronized-audio variant, both the visual input and the audio input are cropped to the same temporal window. This setting achieves 77.5% `WAR` and 71.5% `F1`. In contrast, using local visual windows with complete-video audio guidance achieves 81.9% `WAR` and 79.4% `F1`, improving `WAR` by 4.4% and `F1` by 7.9%. These results suggest that audio benefits from longer temporal context. Short audio chunks may miss important vocal cues such as pauses, loudness variation, and spectral patterns, while each visual frame still preserves the full facial configuration. Therefore, full-video audio provides more reliable guidance for selecting informative visual windows.

Table 5: Subject-wise personalization performance on 10 StressID target subjects. Only WAR and F1 are reported; the best value in each column is in **Bold**.

Method	Sub-1	Sub-2	Sub-3	Sub-4	Sub-5	Sub-6	Sub-7	Sub-8	Sub-9	Sub-10	Avg
					WAR (%)						
TPT Shu et al. (2022)	74.7	51.6	90.0	88.0	52.2	77.5	66.6	63.4	54.0	91.9	70.9
TDA Karmanov et al. (2024)	66.9	45.2	75.6	90.1	49.6	80.0	41.2	75.3	81.8	91.9	69.7
DPE Zhang et al. (2024)	64.0	55.0	73.9	90.0	52.6	80.0	50.9	73.0	81.8	92.0	71.3
PromptAlign Abdul Samadh et al. (2023)	69.9	61.6	78.6	**90.9**	51.2	80.0	60.4	81.3	81.8	90.9	74.6
ReTA Liang et al. (2025)	64.5	50.5	82.0	90.0	46.8	80.0	49.3	80.0	81.8	93.7	71.8
T3AL Liberatori et al. (2024)	71.3	62.7	80.3	90.2	54.0	80.0	68.0	77.0	82.0	93.9	75.9
CLIP-AUTT Zeeshan et al. (2026b)	**90.9**	63.6	90.9	81.0	54.6	**100.0**	**72.7**	81.8	72.7	**100.0**	80.8
ALIGN-TAPER-TTA	73.1	**72.7**	**100.0**	73.6	**72.7**	80.6	**72.7**	**90.9**	**82.3**	**100.0**	**81.9**
					F1 (%)						
TPT Shu et al. (2022)	70.7	43.6	86.0	**83.1**	36.4	78.7	41.1	50.0	40.0	50.0	57.9
TDA Karmanov et al. (2024)	45.5	40.0	50.1	81.0	35.0	80.0	29.6	45.6	42.9	50.0	49.9
DPE Zhang et al. (2024)	41.0	40.6	49.0	81.0	49.8	80.0	56.0	46.0	42.9	56.6	54.2
PromptAlign Abdul Samadh et al. (2023)	41.3	45.0	45.6	82.3	48.1	80.0	44.0	58.0	42.9	45.5	53.2
ReTA Liang et al. (2025)	43.2	41.0	53.0	80.9	36.1	80.0	39.1	53.3	42.9	59.3	52.8
T3AL Liberatori et al. (2024)	46.3	48.6	48.6	80.1	51.0	80.0	61.8	68.0	44.0	66.0	59.4
CLIP-AUTT Zeeshan et al. (2026b)	**90.6**	63.3	90.6	77.0	47.6	**100.0**	61.1	80.3	68.6	**100.0**	77.9
ALIGN-TAPER-TTA	71.8	**71.8**	**100.0**	68.6	**71.8**	79.1	**71.8**	**89.5**	**69.4**	**100.0**	**79.4**

Table 6: Subject-wise personalization performance on 10 BAH target subjects. Only WAR and F1 are reported; the best value in each column is in **Bold**.

Method	Sub-1	Sub-2	Sub-3	Sub-4	Sub-5	Sub-6	Sub-7	Sub-8	Sub-9	Sub-10	Avg
					WAR (%)						
TPT Shu et al. (2022)	62.0	53.7	73.9	80.0	81.2	71.0	58.0	55.0	69.0	53.0	65.6
TDA Karmanov et al. (2024)	58.0	53.7	75.0	80.3	79.0	71.6	58.0	55.9	69.0	51.7	65.2
DPE Zhang et al. (2024)	58.6	59.6	79.6	80.5	83.4	67.0	59.0	54.0	71.9	54.0	66.7
PromptAlign Abdul Samadh et al. (2023)	59.1	59.1	72.0	84.5	84.1	72.4	58.5	55.9	71.5	54.9	67.1
ReTA Liang et al. (2025)	61.4	60.0	72.0	84.0	84.1	72.4	58.5	55.9	**72.0**	56.0	67.6
T3AL Liberatori et al. (2024)	61.0	58.9	75.0	85.6	83.6	74.7	60.0	56.0	71.0	53.6	67.9
CLIP-AUTT Zeeshan et al. (2026b)	54.7	61.3	**87.2**	**88.4**	**85.9**	**76.4**	61.6	56.7	**72.0**	54.1	69.8
ALIGN-TAPER-TTA	**69.8**	**69.2**	63.6	85.7	81.2	73.1	**66.7**	**69.4**	70.6	**77.8**	**72.7**
					F1 (%)						
TPT Shu et al. (2022)	40.6	35.7	43.7	45.9	45.6	43.1	35.0	34.7	40.7	32.9	39.7
TDA Karmanov et al. (2024)	40.0	39.7	44.4	46.0	42.0	42.6	35.0	35.5	40.7	33.2	39.9
DPE Zhang et al. (2024)	34.8	37.1	44.2	46.0	45.3	43.0	37.2	32.5	41.0	33.2	39.4
PromptAlign Abdul Samadh et al. (2023)	25.0	37.0	46.9	51.0	46.1	43.5	41.0	32.5	41.0	33.2	39.7
ReTA Liang et al. (2025)	25.0	36.9	41.9	51.0	46.1	51.2	40.0	32.5	41.0	33.2	39.8
T3AL Liberatori et al. (2024)	41.6	37.0	44.7	46.0	45.2	43.0	38.0	36.2	41.0	35.0	40.7
CLIP-AUTT Zeeshan et al. (2026b)	38.8	38.0	46.6	46.9	46.2	43.3	38.1	36.2	41.9	35.1	41.1
ALIGN-TAPER-TTA	**67.2**	**60.9**	**63.6**	**81.5**	**78.2**	**70.9**	**66.3**	**66.3**	**70.2**	**77.5**	**70.2**