TEST-TIME ADAPTATION WITH ONLINE PERSONALIZED ENERGY-BASED CACHE FOR FINE-GRAINED VIDEO EXPRESSION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Facial expression recognition (FER) in videos remains challenging because models must identify subtle temporally evolving affective states that vary significantly across target individuals. Although vision-language models provide transferable visual-semantic representations, models trained on subject-independent source data often degrade under subject-specific distribution shifts at inference time. State-of-the-art test-time adaptation (TTA) methods typically optimize models during inference, increasing computational cost and latency. Cache-based approaches avoid parameter updates but typically require accumulating sufficient target samples to construct reliable class prototypes. This is difficult at the beginning of adaptation and when some classes are rarely observed. To alleviate these limitations, existing methods may store source prototypes, but these are not personalized to the current target subject. This paper introduces Energy-Based Cache Personalization (`EB-CaP`), a subject-based online TTA method for video FER that samples class-specific prototypes personalized to each target video on-the-fly. Unlike existing cache-based methods, `EB-CaP` does not require observing and accumulating large amounts of target data or storing diverse source prototypes across subjects. Instead, it relies on a lightweight energy-based model (EBM) to sample class-wise prototypes from the current unlabeled video and populate a personalized cache online. The energy function relies only on the pretrained CLIP model, where the similarity between the visual embedding of the target video and the class text embeddings guides the energy-based sampling process. In parallel, positive and negative caches store reliable and uncertain target embeddings, respectively. An adaptive entropy gating follows the evolving confidence distribution to control cache updates, while a diversity gate prevents redundant samples from dominating the memory. Predictions are refined by combining cache-derived scores with the current CLIP scores. Experiments on 3 challenging video datasets for video FER – `BioVid`, `StressID`, and `BAH` – indicate that `EB-CaP` can outperform state-of-the-art TTA methods, while maintaining low computational and memory overhead. Our code included in Appendix, and will be made public.

1 INTRODUCTION

Fine-grained FER in videos is challenging as closely related affective states may differ only in subtle facial movements, expression intensity, or temporal evolution. These cues also vary across individuals, causing the same state to appear differently across subjects while different states may exhibit similar facial patterns. Consequently, models trained on fixed source subjects often generalize poorly to unseen individuals. This limitation is particularly important in many applications including human–computer interaction (Pu & Nie, 2023), health monitoring (Gaya-Morey et al., 2025), and the assessment of pain (Sharafi et al., 2025; 2026a;b; Zeeshan et al., 2024; 2025a;b), depression (De Melo et al., 2020), and stress (Calvo & D'Mello, 2010). Reliable fine-grained video FER, therefore, requires representations that can capture subtle expression cues, temporal dynamics, and subject-specific variability.

VLMs, and in particular CLIP models (Radford et al., 2021), provide a promising foundation for fine-grained FER by transferring semantic knowledge learned from large-scale visual–text

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

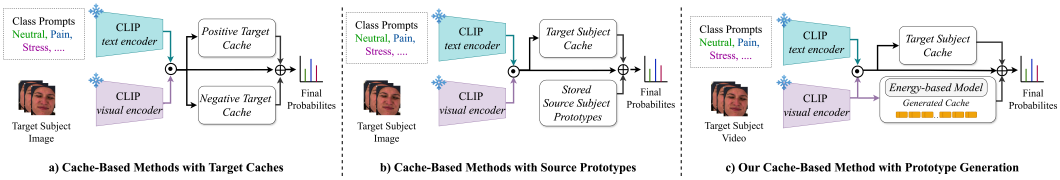

Figure 1: Cache-based TTA methods from target subject data. **(a)** State-of-the-art methods with dynamic positive and negative target caches may accumulate confident but incorrect pseudo-labels (Karmanov et al., 2024). **(b)** A prototype-based method mitigates this issue by constructing prototypes from source domain samples with ground-truth labels and storing them at test time (Sharafi et al., 2026b). **(c)** In contrast, `EB-CaP` samples subject-specific, class-wise prototypes online from each unlabeled target video using an EBM.

data (Yu et al., 2022; Li et al., 2021; 2023). Recent CLIP-based models for FER have further improved recognition through task-specific prompting, temporal modeling, and richer textual supervision (Foteinopoulou & Patras, 2024; Zhao & Patras, 2023; Zhao et al., 2025). However, these approaches mainly learn subject-independent representations and may degrade when the facial behavior, appearance, or recording conditions of an unseen subject differ from those observed during source training (Shu et al., 2023; Tu et al., 2023). This limitation is especially important in fine-grained FER, where small subject-specific variations can alter predictions between closely related affective states. TTA can address this issue by personalizing a model to target subjects based on unlabeled videos captured at inference time. State-of-the-art CLIP-based TTA methods are broadly optimization-based or cache-based. Optimization-based methods update prompts, normalization statistics, or lightweight model components using objectives such as entropy minimization and prediction consistency (Shu et al., 2022; Feng et al., 2023; Abdul Samadh et al., 2023; Yoon et al., 2024; Zhang et al., 2024b); however, repeated online optimization increases inference costs and may cause unstable adaptation when target data are ambiguous or incorrectly predicted.

Cache-based TTA offers a more efficient alternative, where target embeddings and their pseudo-labels are stored and retrieved to refine subsequent predictions (Karmanov et al., 2024; Zhang et al., 2024a; Huang et al., 2025; Nguyen et al., 2025; Liang et al., 2025; Zhu et al., 2025; Zhai et al., 2025; Chen et al., 2025b;a; Guan et al., 2025). Since model parameters remain fixed, these methods are suitable for real-time applications. However, they often require sufficient target samples to form reliable class prototypes, creating a cold-start problem when classes are rare or initially absent. Some methods store source prototypes (Sharafi et al., 2026b), but these require extra storage and are not personalized to the current subject. The diversity of observed samples limits target caches. In video FER, highly correlated temporal windows can cause redundant high-confidence embeddings to dominate the cache, while incorrect pseudo-labels may propagate across ambiguous transitions. An effective cache-based method should therefore construct reliable, diverse, and personalized prototypes without requiring large amounts of accumulated target data.

This paper introduces Energy-Based Cache Personalization (`EB-CaP`), a cache-based TTA method for video FER that operates without model optimization at inference time. It samples online using an energy-based model (EBM) (Grathwohl et al., 2019) to generate class-wise prototypes from each unlabeled target video. Figure 1 compares three cache construction strategies: target-only methods accumulate positive and negative target embeddings but suffer from cache cold start (Figure. 1(a)); prototype-based methods store prototypes derived from labeled source data, providing more reliable class references when target caches contain high-confidence but incorrectly pseudo-labeled samples (Figure. 1 (b)); and `EB-CaP` generates subject-specific, class-wise prototypes online from each unlabeled target video without storing source data or labels (Figure. 1(c)). The sampling process starts from the current video representation and is stochastic, enabling the generation of multiple diverse prototypes for each class. To implement this sampling strategy, `EB-CaP` uses a lightweight class-conditional EBM whose energy is guided by the similarity between the target visual embedding and the CLIP class text embeddings. Generated prototypes are stored in a personalized sampled cache that is refreshed for every incoming video. In parallel, a positive cache stores reliable target embeddings, while a negative cache retains uncertain or class-conflicting embeddings as negative evidence. These target caches are maintained across videos of the same subject and reset for each new subject. Their updates are controlled by an adaptive entropy gate and a diversity gate that limits redundant

samples. During inference, cosine similarities between the incoming video and all three caches are converted into class scores and combined with the base CLIP scores to obtain the final prediction.

The main contributions of this work are summarized as follows. **(1)** `EB-CaP`: an online strategy that rapidly samples diverse class-wise embeddings personalized to each target video. It addresses the cache cold start problem without accumulating target data or retaining source prototypes. A lightweight EBM guided by CLIP visual-text similarity enables this sampling. **(2)** Adaptive entropy and diversity gates for reliable cache construction. The entropy gate updates the caches based on each incoming video, while the diversity gate selects informative and non-redundant embeddings, reducing noisy pseudo-labels and repeated cache entries. **(3)** An extensive set of experiments on three challenging fine-grained video FER benchmarks – `BioVid` (Walter et al., 2013), `StressID` (Chaptoukaev et al., 2023), and `BAH` (González-González et al., 2025) – show that `EB-CaP` outperforms state-of-the-art optimization-based and cache-based TTA methods while maintaining low computational and memory overhead.

2 Related Work

Vision-Language Models for FER. CLIP enables prompt-based recognition by matching visual features with semantic text representations (Radford et al., 2021), making it a promising foundation for FER. However, facial expressions are highly fine-grained and are influenced by subject identity, temporal dynamics, expression intensity, and recording conditions. Recent CLIP-based FER methods improve recognition by incorporating richer textual guidance, temporal modeling, expression descriptions, language priors, and parameter-efficient adaptation (Li et al., 2024; Zhao & Patras, 2023; Foteinopoulou & Patras, 2024; Zhao et al., 2025; Ma et al., 2025). These approaches strengthen the alignment between facial appearance and semantic expression concepts and can improve generalization beyond conventional discriminative FER models. Nevertheless, most existing methods focus on training-time adaptation, prompt design, or subject-independent video representations. As a result, the model remains fixed with respect to the characteristics of an unseen target subject at inference time. This limitation is particularly relevant for fine-grained video FER, where small subject-specific variations in facial appearance and temporal behavior can substantially affect predictions between closely related expression classes, motivating test-time personalization.

Test-Time Adaptation. TTA adapts pretrained models to unlabeled test data under distribution shift, allowing a model to respond to changes in the target distribution without requiring target annotations. Early approaches employ self-supervised objectives or entropy minimization, while later methods improve robustness under noisy, small-batch, and non-stationary streams through selective updates, teacher–student learning, and memory mechanisms (Sun et al., 2020; Zeeshan et al., 2026; Wang et al., 2022; Yuan et al., 2023). Although these strategies can improve target-domain performance, repeated parameter updates may accumulate prediction errors, cause model drift, and increase inference-time computation. For VLMs, TPT and PromptAlign adapt textual prompts or align test-time feature statistics, while other methods employ prompt ensembling, weight averaging, or CLIP-specific adaptation objectives (Shu et al., 2022; Abdul Samadh et al., 2023; Osowiechi et al., 2024; Lafon et al., 2025). Such methods typically require repeated forward and backward passes during inference and can remain sensitive to noisy pseudo-labels, optimization hyperparameters, and limited target observations. These limitations are especially important for online video FER, where adaptation must be performed efficiently as temporally correlated samples arrive sequentially, motivating approaches that personalize predictions without repeatedly updating the recognition model.

Cache-Based TTA. Cache-based TTA adapts predictions by storing and retrieving target representations while keeping the backbone frozen. TDA (Karmanov et al., 2024) uses positive and negative caches, DPE (Zhang et al., 2024a) maintains evolving visual and textual prototypes, while ReTA (Liang et al., 2025), MCP (Chen et al., 2025b), and CRG (Zhai et al., 2025) improve cache reliability through selective updates, complementary memories, or noise reduction. However, these methods still depend on target samples observed during inference, which can cause a cold-start problem and leave rare or unseen classes poorly represented. In video FER, highly correlated temporal windows can also introduce redundant cache entries, while incorrect pseudo-labels may propagate through subsequent updates. Moreover, accumulated caches may not adequately capture the subject-specific expression patterns present in each incoming video. This is limiting when only a few target observations are available, as the cache may provide insufficient evidence for reliable adaptation.

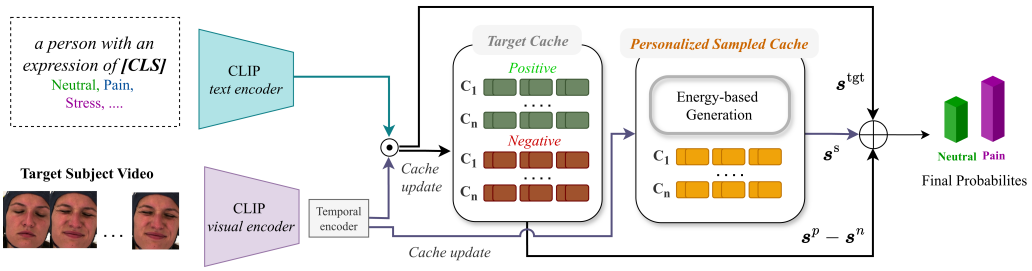

Figure 2: Overview of EB-CaP for online TTA. Frozen CLIP encoders and a temporal encoder produce the current video representation and CLIP class scores. An energy-based model (EBM) samples class-conditioned embeddings initialized from the target video to construct a personalized sampled cache covering all classes. In parallel, adaptive entropy and diversity gates control updates to the positive and negative target caches. Final predictions combine the current CLIP prediction with the caches.

These limitations motivate constructing diverse, class-complete representations directly from each incoming target video without relying on prior target accumulation or stored source prototypes.

3 TTA THROUGH CACHE PERSONALIZATION

The core of EB-CaP is a personalized sampled cache that generates diverse class-wise prototypes from each incoming target video. This cache is rebuilt online for every video, providing immediate subject-adaptive representations without requiring the accumulation of large target data or the retention of source prototypes. EB-CaP further maintains two target caches across videos from the same subject: a positive cache of reliable target embeddings and a negative cache of uncertain embeddings. These target caches are reset when a new subject is encountered. For prediction, the current video embedding is compared with all cached representations, and the resulting similarities are converted into cache scores and combined with the CLIP scores. Adaptive entropy and diversity gates control updates to the target caches.

Figure 2 summarizes the proposed method. Let tgt denote a test target subject and let $\mathcal{X}^{\mathrm{tgt}}$ represent an unlabeled video sequence of length T from the target subject tgt. A frozen CLIP image encoder $E_{\mathrm{I}} : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^d$ maps each frame to an embedding $\mathbf{v}^{\mathrm{tgt}}$. A temporal Transformer encoder $E_{\mathrm{V}} : (\mathbb{R}^d)^L \to \mathbb{R}^d$ models the dependencies among L consecutive frame embeddings and produces the video-window representation $\mathbf{z}^{\mathrm{tgt}}$. The temporal encoder is trained on the source data and remains fixed during test-time adaptation. Let $\mathcal{C}$ denote the set of expression classes, and let p_c be the textual prompt associated with class $c \in \mathcal{C}$. The frozen CLIP text encoder E_{T} produces the normalized class embedding $\mathbf{e}_c$. Given the current video representation $\mathbf{z}^{\mathrm{tgt}}$, the CLIP class score for class c is computed as $s^{\mathrm{tgt}}(c) = \eta \cos(\mathbf{z}^{\mathrm{tgt}}, \mathbf{e}_c)$, where $\eta > 0$ is the fixed CLIP logit scale.

3.1 PERSONALIZED SAMPLED CACHE

For each target video, a personalized sampled cache is constructed by drawing class-conditioned embeddings initialized from its normalized video representation $\mathbf{z}^{\mathrm{tgt}}$. The frozen recognition model, parameterized by θ and trained on the source data, defines a learned class-conditional energy through its class compatibility score (Grathwohl et al., 2019):

$$E_\theta(\mathbf{z}, c) = -\mathbf{z}^\top \mathbf{e}_c, \tag{1}$$

where $\mathbf{e}_c$ is the normalized class embedding produced by the frozen model, and lower energy indicates greater compatibility between representation $\mathbf{z}$ and class c.

For every class $c \in \mathcal{C}$, a separate stochastic gradient Langevin dynamics (SGLD) (Welling & Teh, 2011) chain is initialized from the current target representation: $\mathbf{z}_0 = \mathbf{z}^{\mathrm{tgt}}$. The chain is subsequently refined as:

$$\widetilde{\mathbf{z}}_{k+1} = \widetilde{\mathbf{z}}_k - \frac{\alpha}{2}\nabla_{\widetilde{\mathbf{z}}_k} E_\theta(\widetilde{\mathbf{z}}_k, c) + \sqrt{\alpha}\,\sigma\boldsymbol{\epsilon}_k, \qquad \boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

where α is the SGLD step size, σ controls the noise magnitude, and $\boldsymbol{\epsilon}_k$ is independently sampled at each iteration from the standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The sampled embedding is ℓ_2-normalized after each refinement step to remain in the normalized representation space of the recognition model.

Once the sampled embedding is predicted as the conditioning class c, the chain is stopped to encourage diversity without introducing an additional confidence threshold. Specifically, sampling stops when $\arg\max_{c' \in \mathcal{C}} \widetilde{\mathbf{z}}_k^\top \mathbf{e}_{c'} = c$ or when the maximum number of refinement steps is reached. The final sampled embedding is then ℓ_2-normalized before being stored in the cache. The recognition model remains fixed throughout this process, and gradients are computed only with respect to the sampled embedding.

After sampling, the personalized sampled cache is defined as $\mathcal{K}^s = \{\widetilde{\mathbf{z}}_{c,j} \mid c \in \mathcal{C},\, j = 1, \ldots, m_s\}$, where m_s embeddings are sampled for every class, and the class index c indicates the corresponding conditioning class. Multiple chains can be sampled in parallel, including several samples from the same class. Since only the embedding is updated and the chains are stopped early, sampling requires few iterations and introduces limited computational overhead.

3.2 Target Caches

In addition to $\mathcal{K}^s$, `EB-CaP` maintains a positive target cache $\mathcal{K}^p$ and a negative target cache $\mathcal{K}^n$. Both caches are updated sequentially as the temporal representations of the current video are processed. Their capacities are bounded per class to maintain constant memory usage. Retrieval is completed before the current representation is considered for insertion, ensuring that the current sample does not contribute to its own refined prediction.

Positive cache. The positive cache is defined as $\mathcal{K}^p = \{\mathbf{z}^p_{c,j} \mid c \in \mathcal{C},\, j = 1, \ldots, m_p\}$, where $\mathbf{z}^p_{c,j}$ is the j-th accepted target embedding with pseudo-label c, and m_p is the current number of positive entries for that class. Each class partition has a fixed capacity m_p. The predictive entropy of each entry is retained only as metadata for reliability-based cache replacement.

Negative cache. The negative cache is defined as $\mathcal{K}^n = \{\mathbf{z}^n_{c,j} \mid c \in \mathcal{C},\, j = 1, \ldots, m_n\}$, where $\mathbf{z}^n_{c,j}$ is the j-th uncertain target embedding for which class c is the least likely prediction, and m_n is the current number of negative entries for that class. Each class partition has a fixed capacity m^-, and the stored entropy is used only for cache replacement.

3.3 Online TTA per User Video

For each target video, `EB-CaP` processes the temporal representations sequentially. The current representation $\mathbf{z}^{\mathrm{tgt}}$ produces CLIP logits and retrieves evidence from $\mathcal{K}^s$, $\mathcal{K}^p$, and $\mathcal{K}^n$. The corresponding cache logits are combined with the current CLIP logits to obtain refined class scores. Cache updates are then controlled by an adaptive entropy gate for reliability and a diversity gate for representational coverage.

Adaptive entropy gate. The adaptive entropy gate controls whether the current target representation updates the dynamic caches. Normalized predictive entropy, $\bar{H} = -\sum_{c=1}^{C} p_c \log p_c / \log C$, is computed from the refined class probabilities, such that $\bar{H} \in [0, 1]$, where lower entropy indicates higher confidence. During the first five temporal representations of each video, fixed thresholds of 0.5 and 0.8 are used for the positive and negative caches, respectively. Afterwards, the thresholds are adapted using the running mean μ and standard deviation σ of the predictive entropy values observed within the current video, as $\tau_p = \mu - \sigma$ and $\tau_n = \mu + \sigma$. Samples with entropy below τ_p are assigned their most probable class and considered for the positive cache, while samples with entropy between τ_p and τ_n are assigned their least probable class and considered for the negative cache. Samples above τ_n are rejected. The running entropy statistics are reset at the beginning of each video.

Diversity gate. While entropy measures candidate reliability, it does not indicate whether the candidate adds new information to the cache. Diversity is measured as the mean feature-wise variance of the embeddings stored in the corresponding class partition of $\mathcal{K}^p$ or $\mathcal{K}^n$. For a candidate assigned to class c, this variance is computed before and after temporarily adding the candidate, and the diversity gain is defined as $\Delta D = D_{\mathrm{after}} - D_{\mathrm{before}}$. A positive gain indicates that the candidate increases

the spread of the class representations, whereas a non-positive gain indicates limited or redundant information. The candidate is retained when $\Delta D > 0$.

3.4 Cache Retrieval and Final Prediction

For the current representation $\mathbf{z}^{\mathrm{tgt}}$, cosine similarities are computed with all embeddings stored in the sampled, positive, and negative caches. For each class, the similarity-weighted contributions of all cache entries associated with that class are aggregated to produce one sampled-cache score $\boldsymbol{s}^s$, positive-cache score $\boldsymbol{s}^p$, and negative-cache score $\boldsymbol{s}^n$. Thus, no single prototype is selected; all available prototypes contribute to the corresponding class score. The refined class scores are obtained as

$$\boldsymbol{s}^{\mathrm{fuse}} = \boldsymbol{s}^{\mathrm{tgt}} + \boldsymbol{s}^s + \boldsymbol{s}^p - \boldsymbol{s}^n, \tag{3}$$

The final prediction is obtained by selecting the class with the highest fused score, $\widehat{y} = \arg\max_{c \in \mathcal{C}} s^{\mathrm{fuse}}(c)$.

4 Results and Discussion

4.1 Experimental Methodology

Datasets. Three challenging public video datasets are used to evaluate the proposed approach: `BioVid` (Walter et al., 2013), `StressID` (Chaptoukaev et al., 2023), and `BAH` (González-González et al., 2025). In line with previous studies (Zeeshan et al., 2024; Sharafi et al., 2025; Zeeshan et al., 2025a;b), subject-independent (cross-subject) splits are employed to construct the source and target sets. `BioVid` Heat Pain Database (Part A) consists of controlled heat-pain recordings spanning several intensity levels plus a neutral condition, accompanied by synchronized facial video; the source set comprises 77 subjects, while the remaining 10 form the target set. `StressID` provides facial recordings gathered during stress-inducing tasks along with corresponding self-reported affect labels, split into 44 source subjects and 10 target subjects. `BAH` is a large-scale, multimodal dataset for recognizing ambivalence and hesitancy, captured via webcam and annotated at the frame level; it is divided into 143 source subjects and 10 target subjects. To further assess personalization at a larger scale, we additionally evaluate `EB-CaP` with increasing numbers of randomly selected held-out target subjects on `BioVid`; results are reported in the Appendix. Performance is evaluated using weighted average recall (`WAR`) and macro F1-score (`F1`), which measure overall classification accuracy and the average F1-score across classes, respectively.

Implementation Details. All experiments were conducted on a single NVIDIA A100 GPU with 48 GB of memory, using CLIP ViT-B/32 as the backbone and a batch size of 16. Each target video is processed online using sliding temporal windows of L=16 frames with a stride of 2, without access to future frames. For each target video, the class-conditional sampler generates m_s=3 embeddings per class using at most 20 SGLD iterations, with a step size of 0.01 and a noise scale of 0.1. The positive and negative cache capacities are set to m_p=5 and m_n=4 embeddings per class, respectively. The running entropy statistics are reset at the beginning of each video. During the first five temporal representations, a positive entropy threshold of 0.5 and a negative entropy threshold of 0.8 are used for warm-up. The thresholds are then adapted using the entropy statistics accumulated within the same video, with both adaptation coefficients set to 1. The diversity gate retains a candidate only when it produces a positive variance gain, $\Delta D > 0$, and negative cache scores are fused using unit weights. The same configuration is used for all datasets.

Baseline comparisons cover both CLIP-based FER approaches and TTA methods. For CLIP-based FER, `EB-CaP` is benchmarked against the frozen CLIP-ViT-B/32 backbone in a zero-shot setting, CLIP-ViT-B/32† (Foteinopoulou & Patras, 2024) (CLIP fine-tuned on the source domain), and Emo-CLIP (Foteinopoulou & Patras, 2024), which trains only an adapter. Further baselines include X-CLIP (Ni et al., 2022) and Exp-CLIP (Zhao et al., 2025), which employ trainable projection modules; Exp-CLIP additionally incorporates LLM-generated expression descriptions. For TTA, `EB-CaP` is compared with methods originally developed for image recognition, including TPT (Shu et al., 2022), TDA (Karmanov et al., 2024), DPE (Zhang et al., 2024a), PromptAlign (Abdul Samadh et al., 2023), ReTA (Liang et al., 2025), MCP (Chen et al., 2025b), and CRG (Zhai et al., 2025), together with an adapted version of T3AL (Liberatori et al., 2024), originally proposed for video-based

Table 1: Performance (F1 score and WAR) of zero-shot (ZS), fine-tuning (FT), and TTA on CLIP models for FER. Results are averaged over 10 target subjects for each dataset. CLIP-ViT-B/32† denotes full CLIP fine-tuning.

Protocol	Method	BioVid		StressID		BAH	
		WAR ↑	F1 ↑	WAR ↑	F1 ↑	WAR ↑	F1 ↑
ZS	CLIP-ViT-B/32 (Radford et al., 2021) (ICML'21)	50.0	33.3	60.4	34.8	39.5	28.1
FT	CLIP-ViT-B/32 (Radford et al., 2021) (ICML'21)	69.7	66.6	67.0	44.5	60.4	39.8
	EmoCLIP (Foteinopoulou & Patras, 2024) (FG'24)	67.7	63.4	63.5	35.9	56.2	36.5
	X-CLIP (Ni et al., 2022) (ECCV'22)	70.9	57.9	62.3	41.3	63.0	39.2
	Exp-CLIP (Zhao et al., 2025) (WACV'25)	70.2	66.7	63.1	44.5	62.2	38.5
FT + TTA	TPT (Shu et al., 2022) (NeurIPS'22)	71.1	67.5	70.9	57.9	65.6	39.7
	TDA (Karmanov et al., 2024) (CVPR'24)	71.4	68.2	69.7	49.9	65.2	39.9
	DPE (Zhang et al., 2024a) (NeurIPS'24)	73.1	69.6	71.3	54.2	66.7	39.4
	PromptAlign (Abdul Samadh et al., 2023) (NeurIPS'23)	75.3	71.6	74.6	53.2	67.1	39.7
	ReTA (Liang et al., 2025) (ACMMM'25)	75.1	71.3	71.8	52.8	67.6	39.8
	MCP (Chen et al., 2025b) (ICCV'25)	75.7	73.0	75.5	58.2	67.6	40.1
	T3AL (Liberatori et al., 2024) (CVPR'24)	76.1	72.9	75.9	59.4	67.9	40.7
	CRG (Zhai et al., 2025) (ICLR'25)	76.4	73.8	76.5	60.3	68.1	40.8
	EB-CaP (ours)	**81.0**	**78.2**	**81.2**	**67.0**	**68.9**	**41.0**

Table 2: F1 score per subject on the BioVid dataset for EB-CaP and competing TTA methods. Best results are shown in **bold**.

Method	Sub-1	Sub-2	Sub-3	Sub-4	Sub-5	Sub-6	Sub-7	Sub-8	Sub-9	Sub-10	Avg.
TPT (Shu et al., 2022)	92.0	51.5	40.0	49.6	79.0	87.0	79.5	43.4	99.0	54.0	67.5
TDA (Karmanov et al., 2024)	94.9	50.3	43.5	48.0	80.8	86.2	80.2	43.5	**100.0**	55.0	68.2
DPE (Zhang et al., 2024a)	90.4	51.5	43.5	73.1	72.0	83.6	79.8	48.5	**100.0**	54.4	69.6
PromptAlign (Abdul Samadh et al., 2023)	94.2	60.1	42.5	61.5	83.5	90.5	80.2	43.5	**100.0**	60.0	71.6
ReTA (Liang et al., 2025)	94.2	60.1	43.5	60.8	83.5	88.0	80.2	43.5	**100.0**	60.0	71.3
MCP (Chen et al., 2025b)	94.5	62.0	43.5	64.0	86.0	92.0	82.0	44.0	**100.0**	62.0	73.0
T3AL (Liberatori et al., 2024)	**95.8**	62.0	**43.9**	61.5	83.9	93.9	83.0	43.5	**100.0**	62.0	72.9
CRG (Zhai et al., 2025)	95.0	63.0	43.5	66.0	87.0	93.0	83.0	44.5	**100.0**	63.0	73.8
EB-CaP (ours)	94.4	**66.0**	33.3	**73.0**	**99.5**	**97.1**	**89.0**	**47.0**	**100.0**	**83.3**	**78.2**

Table 3: F1 score per subject on StressID for EB-CaP and other TTA methods.

Method	Sub-1	Sub-2	Sub-3	Sub-4	Sub-5	Sub-6	Sub-7	Sub-8	Sub-9	Sub-10	Avg.
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., 2024a)	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
MCP (Chen et al., 2025b)	55.0	44.5	60.0	78.3	43.6	79.1	52.2	61.2	44.0	65.0	58.2
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
CRG (Zhai et al., 2025)	58.3	46.2	63.5	81.0	47.0	81.0	58.0	65.0	44.0	59.0	60.3
EB-CaP (ours)	**77.0**	**50.0**	**87.0**	82.0	46.1	**84.2**	**62.0**	**68.5**	**46.2**	**67.1**	**67.0**

Table 4: F1 score per subject on BAH for EB-CaP and other TTA methods.

Method	Sub-1	Sub-2	Sub-3	Sub-4	Sub-5	Sub-6	Sub-7	Sub-8	Sub-9	Sub-10	Avg.
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., 2024a)	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
MCP (Chen et al., 2025b)	38.5	37.8	44.5	48.0	45.5	45.0	39.0	33.5	40.7	28.5	40.1
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
CRG (Zhai et al., 2025)	40.5	38.5	44.6	48.5	45.8	46.0	40.0	34.0	40.8	29.3	40.8
EB-CaP (ours)	**41.8**	**39.9**	**44.7**	45.9	**46.2**	**43.6**	**41.0**	32.0	40.4	**35.2**	**41.0**

action recognition. All baselines use the same CLIP backbone, prompts, preprocessing, source fine-tuning, target-subject access, and experimental protocol. Consistent with prior CLIP-based FER research (Zhao et al., 2025), the prompt template "*a person with an expression of [CLS]*" is used across all TTA baselines and EB-CaP. Further details on the baseline implementations and prompt selection are provided in the Appendix.

4.2 Comparison with State-of-the-Art Methods

Table 1 compares zero-shot CLIP, fine-tuned FER models, and TTA approaches on BioVid, StressID, and BAH. Fine-tuning and test-time adaptation consistently improve over zero-shot

Table 5: Complexity of TTA methods. Run time per batch ($B = 16$), memory, and average WAR on BioVid.

Method	Time (ms)	Memory (MB)	WAR (%)
TPT	771.2	2800	71.1
PromptAlign	900.0	3100	75.3
TDA	**97.8**	2124	71.4
DPE	263.9	1700	73.1
ReTA	174.0	2390	75.1
MCP	145.0	1450	75.7
T3AL	520.0	2200	76.1
CRG	205.0	1850	76.4
EB-CaP	220.0	**612.0**	**81.0**

Table 6: Personalized cache construction on BioVid. EB-CaP uses target-initialized energy-based sampling without source information. *Prototype-based caching requires source-derived class prototypes.

Method	WAR (%)
GMM-based	74.0
Classifier-based	77.6
Generator-based	79.1
Prototype* (Sharafi et al., 2026b)	81.5
Energy-based with noise initialization	78.2
Energy-based with target initialization (ours)	**81.0**

inference, highlighting the importance of adapting CLIP representations to subject-specific distribution shifts. Among the methods, EB-CaP achieves the highest WAR and F1 on all three datasets. The gains are pronounced on BioVid and StressID. Compared with CRG, the strongest competing TTA method, EB-CaP improves WAR by 4.6 and 4.7 percentage points and F1 by 4.4 and 6.7 points on BioVid and StressID, respectively. On BAH, the improvement is more modest, suggesting that the ambiguity between expression classes is less easily resolved through test-time personalization. All TTA methods use the CLIP backbone, input data, and evaluation protocol. To isolate our adaptation strategy from temporal modeling and the sampled cache, we augment MCP with these components and report the results in the supplementary material. The improvements over MCP remain below those of EB-CaP, indicating that the gains cannot be attributed solely to these components.

The subject-wise results in Tables 2–4 show that the improvements are not driven by only a few target subjects. On StressID, EB-CaP achieves the highest F1 for eight of the ten subjects and improves the performance. On BioVid, the gains are distributed across several subjects, with large improvements where competing methods remain less reliable, leading to the highest average F1. Sub-9 reaches perfect F1 across all methods, suggesting that neutral and pain expressions are well separated for this subject and easier to recognize independently of the adaptation strategy. In contrast, subjects such as Sub-3 remain challenging across methods, indicating that subject-specific expression separability influences recognition difficulty. Results on BAH are more mixed, with smaller differences among methods and subject-level variability. This reflects the greater difficulty of this dataset and indicates that test-time personalization cannot resolve all ambiguous cases equally well. Overall, the results suggest that combining personalized sampled representations with reliable positive and complementary negative target evidence provides more robust subject adaptation than relying only on model optimization or accumulated target observations. Importantly, these improvements are obtained without updating the recognition model parameters during inference.

Computational Complexity. Table 5 examines the accuracy–efficiency trade-off on BioVid. EB-CaP achieves the highest WAR while requiring only 612 MB of GPU memory, substantially less than all competing TTA methods. Its runtime of 220 ms per batch is higher than lightweight cache-based approaches such as TDA, ReTA, MCP, and CRG, but remains considerably lower than optimization-intensive methods including TPT, PromptAlign, DPE, and T3AL. Thus, the additional class-conditioned sampling introduces moderate runtime overhead while avoiding the large memory cost associated with test-time optimization. Overall, EB-CaP provides a favorable trade-off between recognition performance, memory consumption, and inference cost.

4.3 Ablation Studies

Cache construction. Table 6 compares personalized cache construction strategies. The GMM-, classifier-, and generator-based variants rely on representations learned from labeled source data, while noise-initialized energy-based sampling generates class-conditioned embeddings without using the current target representation. The prototype-based strategy (Sharafi et al., 2026b) is included as a reference but requires source-derived class prototypes at test time, achieving 81.5% WAR. In contrast, the proposed target-initialized energy-based strategy (ours) uses no source samples, labels, or prototypes and initializes each SGLD chain from the current unlabeled target video. Despite

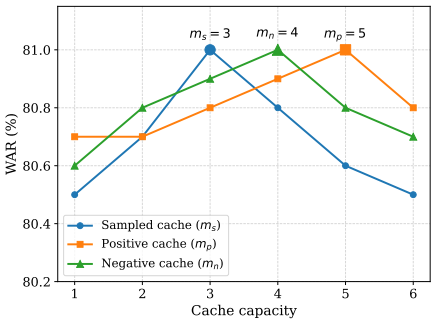

Figure 3: Sensitivity of the sampled, positive, and negative cache capacities on BioVid.

Table 7: Component ablation of EB-CaP on BioVid. Each row removes one component from the full method.

Method	WAR (%)
EB-CaP	**81.0**
w/o Personalized Sampled Cache	78.6
w/o Positive Target Cache	79.3
w/o Negative Target Cache	80.0
w/o Adaptive Entropy Gate	78.0
w/o Diversity Gate	79.1
w/o Target Initialization	78.2

this constraint, EB-CaP achieves a comparable 81.0% WAR, only 0.5 percentage points below the source-prototype variant, showing that effective class-specific caches can be constructed online from unlabeled target data without retaining source information.

Cache capacity. Figure 3 studies the sampled, positive, and negative cache capacities in EB-CaP. Each capacity is varied independently while the remaining two are fixed. Performance remains stable across the evaluated settings, indicating limited sensitivity to moderate changes in cache size. For the personalized sampled cache, performance improves up to m_s=3, after which additional embeddings provide no further benefit. A similar trend is observed for the target caches, with the positive and negative caches reaching their highest performance at m_p=5 and m_n=4, respectively. These results suggest that small caches are sufficient to capture useful subject-specific information, while larger capacities may introduce redundant entries without improving recognition. The resulting configuration, m_s=3, m_p=5, and m_n=4, is therefore used for BioVid.

Component ablation. Table 7 evaluates the contribution of each component of EB-CaP on BioVid. Removing the personalized sampled cache causes a notable drop in performance, from 81.0% to 78.6% WAR, highlighting the importance of generating class-specific representations from the current target video. The positive and negative target caches also provide complementary gains, while the entropy and diversity gates improve cache reliability by controlling which samples are retained. Replacing target initialization further degrades performance, confirming the benefit of conditioning the energy-based sampling process on the current target representation. Overall, the full EB-CaP configuration provides the strongest performance, showing that the proposed components work together to improve test-time personalization.

5 Conclusion

This paper introduced EB-CaP, a personalized TTA method for fine-grained video FER. For each unlabeled target video, EB-CaP constructs a personalized sampled cache by generating class-wise embeddings using SGLD that is initialized from the current video representation. This cache is complemented by positive and negative target caches that retain reliable and uncertain observations. Adaptive entropy and diversity gates control cache updates to reduce noisy pseudo-label accumulation and redundant representations. The resulting cache-derived class scores are fused with the current CLIP scores while the recognition model remains fixed. Experiments on three challenging video datasets including BioVid, StressID, and BAH demonstrate that EB-CaP consistently improves over existing fine-tuning and TTA methods without requiring source data, target labels, or test-time model updates.

Appendix. Includes architectural and implementation details, the complete online TTA algorithm, TTA baseline implementation details, subject-wise results, and extended analyses on energy-based sampling, cold-start behavior, target-subject scaling, cache-diversity criteria, temporally augmented TTA baselines, class-prompt selection, and fusion-weight sensitivity.

AI USE STATEMENT

AI-based tools were used during the preparation of this manuscript to support language refinement, proofreading, and improvements in clarity and presentation. These tools were not used to generate datasets, produce or alter experimental results, or determine the scientific conclusions of the study. Any AI-assisted content, including text, code, figures, and interpretations, was carefully inspected and validated by the authors. The authors remain fully responsible for the accuracy, integrity, and final content of the manuscript.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the proposed method, experimental protocol, dataset splits, implementation settings, and hyperparameters in the main paper and Appendix. Additional ablations and implementation details are also provided in the Appendix. We further include the source code with the submission to enable reproduction of the reported experiments.

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.

Rafael A Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37, 2010.

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.

Si Chen, Yujia Chen, Xiaotian Yin, Xin Liu, Huakai Lai, and Tianzhu Zhang. Paf: Prototype adaptive fusion for test-time adaptation of vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3007–3016, 2025a.

Xinyu Chen, Haotian Zhai, Can Zhang, Xiupeng Shi, and Ruirui Li. Multi-cache enhanced prototype learning for test-time generalization of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2281–2291, 2025b.

Wheidima Carneiro De Melo, Eric Granger, and Abdenour Hadid. A deep multiscale spatiotemporal network for assessing depression from facial dynamics. *IEEE transactions on affective computing*, 13(3):1581–1592, 2020.

Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2704–2714, 2023.

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.

F Xavier Gaya-Morey, Jose M Buades-Rubio, Philippe Palanque, Raquel Lacuesta, and Cristina Manresa-Yee. Deep learning-based facial expression recognition for the elderly: A systematic review. *arXiv preprint arXiv:2502.02618*, 2025.

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.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *ICLR*, 2019.

Zenghao Guan, Zhou Yucan, Wu Liu, and Xiaoyan Gu. Statistics caching test-time adaptation for vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Fanding Huang, Jingyan Jiang, Qinting Jiang, Hebei Li, Faisal Nadeem Khan, and Zhi Wang. Cosmic: Clique-oriented semantic multi-space integration for robust clip test-time adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9772–9781, 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.

Marc Lafon, Gustavo Adolfo Vargas Hakim, Clément Rambour, Christian Desrosier, and Nicolas Thome. Cliptta: Robust contrastive vision-language test-time adaptation. *arXiv preprint arXiv:2507.14312*, 2025.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 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.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.

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. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4788–4797, 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.

Fuyan Ma, Yiran He, Bin Sun, and Shutao Li. Multimodal prompt alignment for facial expression recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Khanh-Binh Nguyen, Phuoc-Nguyen Bui, Hyunseung Choo, and Duc Thanh Nguyen. Adaptive cache enhancement for test-time adaptation of vision-language models. *arXiv preprint arXiv:2508.07570*, 2025.

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.

David Osowiechi, Mehrdad Noori, Gustavo Vargas Hakim, Moslem Yazdanpanah, Ali Bahri, Milad Cheraghalikhani, Sahar Dastani, Farzad Beizaee, Ismail Ayed, and Christian Desrosiers. Watt: Weight average test time adaptation of clip. *Advances in neural information processing systems*, 37:48015–48044, 2024.

Jing Pu and Xinxin Nie. Convolutional channel attentional facial expression recognition network and its application in human–computer interaction. *IEEE Access*, 11:129412–129424, 2023.

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.

M. Sharafi, M. O. Belharbi, S. Zeeshan, H. Ben Salem, A. Etemad, A.L. Koerich, M. Pedersoli, S. Bacon, and E. Granger. Personalized feature translation for expression recognition: An efficient source-free domain adaptation method. In *ICLR*, 2026a.

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)*, 2025.

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.

Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. Clipood: Generalizing clip to out-of-distributions. In *International conference on machine learning*, pp. 31716–31731. PMLR, 2023.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Weijie Tu, Weijian Deng, and Tom Gedeon. A closer look at the robustness of contrastive language-image pre-training (clip). *Advances in Neural Information Processing Systems*, 36:13678–13691, 2023.

Steffen Walter, Sascha Gruss, Hagen Ehleiter, Junwen Tan, Harald C Traue, Philipp Werner, Ayoub Al-Hamadi, Stephen Crawcour, Adriano O Andrade, and Gustavo Moreira da Silva. 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.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688, 2011.

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang D Yoo. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. *arXiv preprint arXiv:2403.14119*, 2024.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15922–15932, 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, 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. *arXiv preprint arXiv:2508.12522*, 2025a.

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*, 2025b.

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. *arXiv preprint arXiv:2603.27999*, 2026.

Haotian Zhai, Xinyu Chen, Can Zhang, Tianming Sha, and Ruirui Li. Mitigating cache noise in test-time adaptation for large vision-language models. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025.

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, 2024a.

Ding-Chu Zhang, Zhi Zhou, and Yu-Feng Li. Robust test-time adaptation for zero-shot prompt tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16714–16722, 2024b.

Zengqun Zhao and Ioannis Patras. Prompting visual-language models for dynamic facial expression recognition. In *BMVC*, 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.

Xingyu Zhu, Shuo Wang, Beier Zhu, Miaoge Li, Yunfan Li, Junfeng Fang, Zhicai Wang, Dongsheng Wang, and Hanwang Zhang. Dynamic multimodal prototype learning in vision-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2501–2511, 2025.

A APPENDIX

Table of Contents

1 Algorithmic Details 14

2 TTA Baseline Implementation Details 14

3 Subject-Wise Results 15

4 Analysis of Energy-Based Sampling 17

5 Cold-Start Analysis 17

6 Scaling with the Number of Target Subjects 17

7 Analysis of Diversity Criteria 17

8 Comparison with Temporally Augmented TTA 18

9 Class Prompt Selection 19

10 Fusion Weight Sensitivity 20

B ALGORITHMIC DETAILS

This section provides additional architectural and algorithmic details of the proposed online test-time adaptation pipeline. The recognition model consists of a CLIP ViT-B/32 image encoder followed by a temporal Transformer that aggregates frame-level visual representations into a video-level embedding. The temporal encoder contains four Transformer layers with eight attention heads, a feed-forward dimension of 2048, dropout of 0.1, and an embedding dimension of 512. A learnable classification token is appended to the sequence of frame embeddings, and its output representation is used as the temporal embedding for classification and test-time adaptation. The temporal encoder is initialized from scratch and jointly trained with the CLIP image encoder on the labeled source subjects for 30 epochs using AdamW with a learning rate of 10^{-4}. Training uses a contrastive objective between the temporal visual representations and their corresponding class-text embeddings.

At test time, the image encoder, temporal encoder, and CLIP text encoder are frozen. Algorithm 1 summarizes the complete online inference procedure of `EB-CaP`. Each target video is processed sequentially using temporal windows, and the resulting representation is used to construct a personalized sampled cache through class-conditioned SGLD initialized from the current target embedding. Evidence from the sampled, positive, and negative caches is combined with the CLIP prediction to obtain the refined class scores. The sampled cache is reconstructed for each temporal window, whereas the positive and negative target caches accumulate subject-specific evidence across videos and are reset when personalization begins for a new subject. No source samples, labels, or source-derived prototypes are retained during test-time adaptation, and no model parameters are updated at inference time.

C TTA BASELINE IMPLEMENTATION DETAILS

Several CLIP-based test-time adaptation methods are included as baselines. *TPT* (Shu et al., 2022) performs test-time prompt tuning by optimizing learnable text prompt tokens through entropy minimization over multiple strongly augmented views, while leaving the CLIP image and text encoders unchanged. *TDA* (Karmanov et al., 2024) avoids test-time optimization and instead maintains an online feature–label memory containing positive and negative pseudo-label information. Predictions retrieved from this dynamic cache are combined with the original CLIP zero-shot scores.

Algorithm 1 Online Test-Time Adaptation with EB-CaP

Require: Target video stream, frozen encoders E_{I}, E_{V}, and E_{T}, class prompts $\{p_c\}_{c\in\mathcal{C}}$
Ensure: Refined predictions $\widehat{y}$

$\mathbf{e}_c \leftarrow \mathrm{Normalize}(E_{\mathrm{T}}(p_c)), \forall c \in \mathcal{C}$
for each target subject **do**
  Initialize $\mathcal{K}^p, \mathcal{K}^n \leftarrow \emptyset$
  for each target video $\mathcal{X}^{\mathrm{tgt}}$ **do**
    Reset video-level entropy statistics
    for each temporal window $\mathbf{X}_r^{\mathrm{tgt}}$ **do**
      $\mathbf{z}^{\mathrm{tgt}} \leftarrow \mathrm{Normalize}\big(E_{\mathrm{V}}\big(E_{\mathrm{I}}(\mathbf{X}_r^{\mathrm{tgt}})\big)\big)$
      Compute CLIP scores $\boldsymbol{s}^{\mathrm{tgt}}$ using $\mathbf{z}^{\mathrm{tgt}}$ and $\{\mathbf{e}_c\}_{c\in\mathcal{C}}$
      Generate $\mathcal{K}^s$ using class-conditioned SGLD initialized from $\mathbf{z}^{\mathrm{tgt}}$
      Compute cache scores $\boldsymbol{s}^s$, $\boldsymbol{s}^p$, and $\boldsymbol{s}^n$
      $\boldsymbol{s}^{\mathrm{fuse}} \leftarrow \boldsymbol{s}^{\mathrm{tgt}} + \boldsymbol{s}^s + \boldsymbol{s}^p - \boldsymbol{s}^n$
      $\widehat{y} \leftarrow \arg\max_{c\in\mathcal{C}} s^{\mathrm{fuse}}(c)$
      Compute entropy H and update thresholds τ_p and τ_n
      if $H < \tau_p$ **then**
        Update $\mathcal{K}^p$ with $\mathbf{z}^{\mathrm{tgt}}$ if $\Delta D > 0$
      else if $H \leq \tau_n$ **then**
        Update $\mathcal{K}^n$ with $\mathbf{z}^{\mathrm{tgt}}$ if $\Delta D > 0$
      end if
      Discard $\mathcal{K}^s$
    end for
  end for
end for

DPE (Zhang et al., 2024a) jointly adapts visual and textual class prototypes during inference, augments them with lightweight instance-dependent residuals, and imposes cross-modal consistency to preserve agreement between the two prototype spaces under target-domain shifts. *PromptAlign* (Abdul Samadh et al., 2023) updates textual prompts by matching test-time feature statistics to statistics estimated from the source domain, providing an additional alignment objective beyond entropy minimization. *ReTA* (Liang et al., 2025) improves cache reliability through consistency-aware entropy reweighting for selective memory updates and diversity-driven distribution calibration, where each class is represented by a Gaussian distribution over dynamically evolving text embeddings.

MCP (Chen et al., 2025b) employs multiple complementary caches to improve the reliability and compactness of class prototypes. Specifically, an entropy cache retains low-entropy samples to initialize prototypes, an alignment cache selects representations that better agree with the class prototype structure, and a negative cache stores high-entropy samples as complementary evidence for prediction calibration. The retrieved information from these caches is combined with visual–textual prototype information to produce the adapted prediction. *CRG* (Zhai et al., 2025) explicitly addresses noisy pseudo-labels in cache-based adaptation. It maintains positive and negative visual prototypes and introduces residual parameters to better align them with the corresponding text prototypes. In addition, Gaussian discriminant analysis is used to dynamically estimate class-conditional feature distributions, reducing the sensitivity of similarity-based cache retrieval to noisy or atypical cached samples.

Finally, *T3AL* (Liberatori et al., 2024), originally designed for zero-shot temporal action localization, is modified for video classification. The adapted implementation derives video-level pseudo-labels from aggregated frame representations, refines frame-level predictions through self-supervision, and uses textual guidance to suppress irrelevant temporal regions. Its encoders remain frozen, and the adaptation state is reinitialized for every input video.

D Subject-Wise Results

Tables 8, 9, and 10 report the WAR obtained for each of the ten target subjects. The results provide a detailed assessment of performance under subject-specific distribution shifts and reveal substantial variability across subjects and adaptation methods.

Table 8: WAR per subject on the BioVid dataset for EB-CaP and competing TTA methods. Best results are shown in **bold**.

Method	Sub-1	Sub-2	Sub-3	Sub-4	Sub-5	Sub-6	Sub-7	Sub-8	Sub-9	Sub-10	Avg.
TPT (Shu et al., 2022)	91.9	53.0	50.0	61.5	79.0	88.9	76.0	50.0	98.9	62.0	71.1
TDA (Karmanov et al., 2024)	93.0	52.6	50.0	63.9	79.5	87.5	76.3	50.0	100.0	61.6	71.4
DPE (Zhang et al., 2024a)	91.7	55.0	50.0	85.0	73.0	84.7	80.0	50.0	100.0	61.8	73.1
PromptAlign (Abdul Samadh et al., 2023)	93.8	65.2	50.0	70.0	85.0	92.0	80.0	50.0	100.0	67.6	75.3
ReTA (Liang et al., 2025)	93.8	65.2	50.0	69.5	85.0	90.2	80.0	50.0	100.0	67.6	75.1
MCP (Chen et al., 2025b)	93.9	65.8	50.0	72.0	85.5	92.5	81.5	50.0	100.0	66.0	75.7
T3AL (Liberatori et al., 2024)	94.0	66.5	50.0	70.0	85.2	94.2	83.9	50.0	100.0	67.6	76.1
CRG Zhai et al. (2025)	94.2	67.0	50.0	73.0	86.0	94.0	83.0	50.0	100.0	66.8	76.4
EB-CaP (ours)	**94.8**	**71.9**	**50.0**	**95.4**	**92.1**	**97.5**	**90.8**	**50.0**	**100.0**	**68.0**	**81.0**

Table 9: WAR per subject on StressID for EB-CaP and other TTA methods.

Method	Sub-1	Sub-2	Sub-3	Sub-4	Sub-5	Sub-6	Sub-7	Sub-8	Sub-9	Sub-10	Avg.
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., 2024a)	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
MCP (Chen et al., 2025b)	72.5	61.0	81.5	90.5	55.0	80.0	65.0	78.0	81.8	89.7	75.5
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
CRG (Zhai et al., 2025)	74.0	62.0	82.0	**90.5**	57.0	80.0	67.0	79.0	82.0	91.5	76.5
EB-CaP (ours)	**81.1**	**66.9**	**91.2**	89.0	**69.9**	**81.2**	**69.5**	**83.8**	**84.6**	**95.7**	**81.2**

Table 10: WAR per subject on BAH for EB-CaP and other TTA methods.

Method	Sub-1	Sub-2	Sub-3	Sub-4	Sub-5	Sub-6	Sub-7	Sub-8	Sub-9	Sub-10	Avg.
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., 2024a)	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
MCP (Chen et al., 2025b)	61.0	59.5	74.0	84.5	83.5	73.0	59.0	55.5	71.5	54.5	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
CRG (Zhai et al., 2025)	61.5	60.0	75.0	85.0	84.0	74.0	59.5	56.0	71.0	55.0	68.1
EB-CaP (ours)	**62.1**	**60.4**	**79.9**	**85.6**	**84.5**	**74.7**	**61.1**	**56.5**	**72.0**	53.0	**68.9**

On BioVid, EB-CaP achieves the best or tied-best performance for all ten target subjects and obtains an average WAR of 81.0%. Compared with CRG, the strongest competing method with an average of 76.4%, this corresponds to a gain of 4.6 percentage points. The improvements are particularly pronounced for Sub-4, Sub-5, Sub-6, and Sub-7, where EB-CaP provides clear gains over the competing methods. In contrast, Sub-3, Sub-8, and Sub-9 show similar performance across several methods. In particular, Sub-9 reaches nearly perfect performance for all approaches, suggesting that this subject exhibits a relatively easier separation between the considered expression classes.

A similar trend is observed on StressID, where EB-CaP achieves the highest performance for nine of the ten subjects and improves the average WAR to 81.2%. The strongest competing method, CRG, obtains 76.5%, corresponding to a gain of 4.7 percentage points for EB-CaP. The improvements are especially notable for Sub-1, Sub-5, Sub-8, and Sub-9, indicating that the proposed personalization strategy remains effective across subjects with substantially different baseline performance. Sub-4 is the only subject for which EB-CaP does not achieve the highest WAR.

On BAH, the differences between methods are smaller. Nevertheless, EB-CaP achieves the best or tied-best result for nine of the ten subjects and obtains the highest average WAR of 68.9%, exceeding CRG by 0.8 percentage points. The smaller improvement and frequent ties among methods indicate that the benefit of personalization is more limited on this dataset. Sub-10 remains the main exception, where EB-CaP does not improve over the strongest competing methods.

Overall, the subject-wise results show that the gains of EB-CaP are distributed across most target subjects rather than being driven by a small subset. The personalized sampled cache, together with the positive and negative target caches and the adaptive entropy and diversity gates, provides more reliable adaptation to subject-specific characteristics without updating the recognition model at inference time.

Table 11: Analysis of energy-based sampling on BioVid. Left: contribution of energy-guided refinement and stochastic perturbation. Right: SGLD chain behavior.

Strategy	Grad.	Noise	WAR
Noise only	✗	✓	74.5
Deterministic	✓	✗	78.8
SGLD	✓	✓	**81.0**

Chain statistic	Value
Success rate	94.2%
Avg. iterations	6.7
Max-iter. rate	5.8%

E Analysis of Energy-Based Sampling

We analyze the contribution and behavior of the proposed energy-based sampling procedure in Table 11. Removing the stochastic term ($\sigma = 0$), corresponding to deterministic energy-guided refinement, decreases WAR from 81.0% to 78.8%. Conversely, removing the energy-gradient term while retaining only the stochastic perturbation yields 74.5% WAR. These results show that neither deterministic class-directed refinement nor random perturbation alone accounts for the performance of the proposed sampler; combining energy-guided refinement with stochastic exploration achieves the best performance. We further examine the convergence behavior of the SGLD chains. Overall, 94.2% of the chains reach the conditioning class before the maximum of 20 iterations, requiring 6.7 iterations on average, while 5.8% reach the iteration limit. Samples that do not reach the conditioning class within the maximum number of iterations are discarded and are not inserted into the personalized sampled cache.

F Cold-Start Analysis

Table 12 analyzes the contribution of the PSC at different stages of online adaptation. At the beginning of each target subject, the positive and negative target caches are empty, whereas the PSC immediately provides class-wise evidence from the incoming temporal window. The PSC provides its largest gain on the first video, with the improvement gradually decreasing as the target caches accumulate subject-specific evidence. This trend demonstrates that the PSC is particularly beneficial during early adaptation, mitigating the cold-start limitation of dynamic target caches while remaining beneficial throughout the subject sequence.

G Scaling with the Number of Target Subjects

To evaluate personalization beyond the original ten-subject protocol, Table 13 progressively increases the number of held-out target subjects on BioVid. Target subjects are randomly selected from the 87 available subjects and are kept disjoint from source training. EB-CaP maintains consistent performance as the target cohort grows from 10 to 40 subjects and continues to outperform both the source-only model and the variant without PSC. Importantly, the improvement provided by PSC remains stable across target-cohort sizes, indicating that its benefit is not restricted to the original ten target subjects and extends to a substantially larger target population.

H Analysis of Diversity Criteria

We further analyze the cache admission criterion by comparing the proposed variance-gain strategy with alternative diversity measures. All variants retain the same entropy gate and differ only in the criterion used to admit a candidate representation into the target cache. In addition to removing the diversity criterion entirely, we consider cosine-distance-based admission and a confidence-weighted diversity criterion. This comparison examines whether increasing cache variance provides useful complementary representations rather than simply favoring outlying samples.

For cosine-distance-based admission, we compute the distance between a candidate representation $\mathbf{z}$ and its most similar representation in the corresponding class-wise cache:

$$D_{\cos}(\mathbf{z}, \mathcal{K}_c) = 1 - \max_{\mathbf{z}_i \in \mathcal{K}_c} \cos(\mathbf{z}, \mathbf{z}_i). \tag{4}$$

Table 12: Cold-start analysis on BioVid. Performance is reported at different stages of online adaptation for each target subject.

Adaptation Stage	w/o PSC	EB-CaP	Gain
Video 1	73.2	**79.8**	+6.6
Videos 2–5	76.5	**80.9**	+4.4
Videos > 5	79.1	**81.2**	+2.1
Overall	78.6	**81.0**	+2.4

Table 13: Large-scale target-subject personalization analysis on BioVid. We report WAR (%) as the number of held-out target subjects increases.

Method	Number of target subjects			
	10	20	30	40
Source-only	67.7	67.1	66.8	66.5
w/o PSC	78.6	78.1	77.8	77.5
EB-CaP	**81.0**	**80.7**	**80.5**	**80.2**

Table 14: Comparison of diversity criteria for target-cache admission on BioVid. All variants use the same entropy gate and differ only in the diversity criterion.

Diversity criterion	WAR (%)
None	79.1
Cosine distance	80.0
Confidence-weighted diversity	80.4
Variance gain ($\Delta D > 0$)	**81.0**

Table 15: Comparison with temporally augmented MCP on BioVid. MCP is progressively equipped with temporal modeling and the personalized sampled cache used in EB-CaP to control for the additional video-specific information available to our method.

Method	WAR (%)	F1 (%)
MCP	75.7	73.0
MCP + Temporal	75.9	73.4
MCP + Temporal + Sampled Cache	76.9	75.0
EB-CaP	**81.0**	**78.2**

The candidate is admitted when $D_{\cos}(\mathbf{z}, \mathcal{K}_c) > \delta$.

For confidence-weighted diversity, we additionally account for the reliability of the candidate by weighting its cosine diversity by prediction confidence:

$$D_{\text{conf}}(\mathbf{z}, \mathcal{K}_c) = \big(1 - \bar{H}(\mathbf{p})\big)\, D_{\cos}(\mathbf{z}, \mathcal{K}_c), \tag{5}$$

where $\bar{H}(\mathbf{p}) \in [0, 1]$ denotes normalized predictive entropy. Candidates that are diverse but associated with uncertain predictions therefore receive lower admission scores. The comparison of these diversity criteria is reported in Table 14.

As shown in Table 14, removing the diversity criterion decreases WAR from 81.0% to 79.1%. Cosine-distance-based admission improves performance to 80.0%, while confidence-weighted diversity further increases WAR to 80.4%. The proposed variance-gain criterion achieves the highest performance at 81.0%. These results show that explicitly controlling cache diversity is beneficial and that the proposed variance-gain criterion is more effective than the considered distance- and confidence-based alternatives.

I Comparison with Temporally Augmented TTA

Most existing CLIP-based TTA methods are designed for image-level recognition and therefore do not explicitly exploit the temporal structure available in our video setting. To examine whether the advantage of EB-CaP can be attributed simply to this additional temporal information, we progressively augment MCP with the same temporal modeling and personalized sampled cache used in EB-CaP, while keeping the remaining experimental protocol unchanged. The results are reported in Table 15.

Equipping MCP with temporal modeling yields only a marginal improvement, from 75.7% to 75.9% in WAR and from 73.0% to 73.4% in F1. Further incorporating the sampled cache improves its performance to 76.9% WAR and 75.0% F1, confirming that subject-specific information is beneficial in the video setting. Nevertheless, even after providing MCP with both temporal modeling and sampled information, EB-CaP achieves 81.0% WAR and 78.2% F1. This comparison indicates that the improvement of EB-CaP cannot be explained solely by access to temporal information or sampled representations, but also stems from the proposed test-time adaptation strategy.

Table 16: Comparison of class-prompt templates for the CLIP text encoder on BioVid with a CLIP ViT-B/32 backbone. The choice of prompt influences recognition accuracy, and *"a person with an expression of [CLS]"* provides the best WAR and F1 results.

Prompt Template	**WAR**	**F1 score**
a photo of a [CLS]	68.7	63.3
a_photo_of_the_[CLS]_face	66.5	59.3
a_photo_of_one_[CLS]_face	64.2	55.8
a_close-up_photo_of_the_[CLS]_face	68.2	62.6
a_low_resolution_photo_of_a_[CLS]_face	68.7	64.2
a_good_photo_of_a_[CLS]_face	69.2	65.4
a_photo_of_my_[CLS]_face	66.2	60.3
a_cropped_photo_of_the_[CLS]_face	64.5	57.5
a_photo_of_a_person_with_[CLS]_face	67.2	61.4
a person with an expression of [CLS]	**69.7**	**66.6**
a_portrait_of_a_person_in_[CLS]	58.2	46.7
a_face_showing_signs_of_[CLS]	68.7	64.1
a close up portrait of a [CLS] expression	69.2	63.0
a realistic photo of a person experiencing [CLS]	53.5	40.0
a cropped image of a person in [CLS]	69.0	64.5
a photo of a [CLS] person	69.5	64.3
a person with a facial expression of [CLS]	69.5	64.2
a high quality photo of a [CLS] expression	69.5	64.3
a photo of a face showing [CLS]	69.5	64.3
a photo of a face in [CLS]	69.0	61.8

Table 17: Sensitivity of EB-CaP to the cache-fusion weights on BioVid. Each weight is varied independently while the remaining two are fixed to 1.

λ_s	λ_p	λ_n	WAR (%)
0.5	1.0	1.0	80.1
1.0	0.5	1.0	80.4
1.0	1.0	0.5	80.2
1.0	**1.0**	**1.0**	**81.0**
1.5	1.0	1.0	80.3
1.0	1.5	1.0	80.5
1.0	1.0	1.5	80.4

J CLASS PROMPT SELECTION

The influence of class-prompt formulation is evaluated using several templates commonly adopted in CLIP-based FER and prompt-driven recognition methods (Foteinopoulou & Patras, 2024; Ni et al., 2022; Zhao et al., 2025). This experiment examines whether different textual descriptions of the expression categories lead to more discriminative class embeddings and improved recognition accuracy. All prompt variants are assessed with CLIP ViT-B/32 trained on the source subjects of BioVid and tested on the corresponding target subjects. The results in Table 16 show that performance varies noticeably across prompt formulations in terms of both WAR and F1. The template *"a person with an expression of [CLS]"* achieves the strongest overall results and is consequently used for all prompt-based methods in the experiments.

Two energy-based variants are also evaluated. In *energy-based sampling with noise initialization*, each class-conditioned SGLD chain is initialized from a randomly sampled feature rather than from the current target representation. The energy defined by the frozen recognition model then guides the sample toward the corresponding class embedding. This variant requires neither stored source prototypes nor target labels, but the resulting samples are not explicitly personalized to the current video. In contrast, *energy-based sampling with target initialization*, corresponding to EB-CaP, initializes every class-conditioned chain from the current unlabeled target-video representation. Con-

sequently, the sampled embeddings retain target-specific visual characteristics while being progressively guided toward different expression classes.

The results show that the quality and personalization of the generated embeddings substantially influence adaptation. The GMM-, classifier-, and generator-based strategies obtain 74.0%, 77.6%, and 79.1% WAR, respectively. Energy-based sampling initialized from noise reaches 78.2%, indicating that class conditioning alone can generate useful cache entries but does not fully capture the current subject. The source-derived prototype strategy achieves the highest result of 81.5%, although it requires labeled source information at test time. Without retaining source samples, labels, or prototypes, the proposed target-initialized strategy achieves a competitive 81.0% WAR. This result demonstrates that initializing the sampling process from the current target video provides effective class-specific and subject-adaptive cache representations.

K Fusion Weight Sensitivity

Table 17 evaluates the sampled-, positive-, and negative-cache weights λ_s, λ_p, and λ_n. Each weight is varied from 0.5 to 1.5 while the remaining two are fixed to 1. Performance remains relatively stable across the evaluated settings, indicating limited sensitivity to moderate changes in the fusion weights. The balanced configuration $(\lambda_s, \lambda_p, \lambda_n) = (1, 1, 1)$ achieves the highest WAR and is therefore used in all experiments.