AURA: VISUALLY INTERPRETABLE AFFECTIVE UNDERSTANDING VIA ROBUST ARCHETYPES

Anonymous authors
Paper under double-blind review

A EXPERIMENTS

A.1 DATASETS

AffectNet-7/8 & AffectNet-VA Mollahosseini et al. (2017): AffectNet is an in-the-wild database that contains around 400K images manually annotated for 6 basic expressions, as well as neutral and contempt. For our work, we utilize the manually annotated images with the 7/8 expressions category to ensure alignment with other expression datasets. AffectNet-VA provides VA annotations in the range of [-1, 1], making it suitable for dimensional affect analysis. The training set of this database consists of around 321K images and the validation of 5K. The validation set is balanced across the different expression categories. **RAF-DB (Real-world Affective Faces Database)** Li et al. (2017): RAF-DB is an in-the-wild database that contains approximately 15,000 facial images, manually annotated for 7 basic expressions. **DISFA (Denver Intensity of Spontaneous Facial Action)** Mavadati et al. (2013): DISFA is a lab controlled database consisting of videos from 27 subjects, each with approximately 5000 frames. Each frame is annotated with AU intensities on a six-point discrete scale (0–5). For consistency in AU detection tasks, we binarize the annotations, assigning a value of 1 to AU intensities greater than 2 and a value of 0 otherwise. The dataset includes annotations for 8 AUs (1, 2, 4, 6, 9, 12, 25, 26). **EmotioNet**Fabian Benitez-Quiroz et al. (2016) consists of over 45K in-the-wild facial images, where we follow the official split and use the 11 most frequent AUs for training and evaluation.

A.2 IMPLEMENTATION DETAILS

Our AURA framework is implemented in PyTorch and trained on an NVIDIA A100 GPU. For data preprocessing, all input images are first cropped to facial regions and then resized to the CLIP-supported resolution. The CLIP visual encoder is a frozen, pre-trained model from OpenAI. Image or video frame features are extracted once using this encoder, after which all training and inference are performed purely at the feature level, eliminating the need to repeatedly invoke CLIP during optimization. We adopt the AdamW optimizer with a learning rate of 1×10^{-4} across all datasets. To enhance generalization, a dropout rate of 0.2 is applied to both the global-level and patch-level visual projectors. For all datasets, the loss weights are set as $\lambda_{\text{Proj}} = \lambda_{\text{Arc}} = \lambda_{\text{Contx}} = 1$, ensuring balanced contributions from projection, visual archetype optimization, and refinement terms. Similarly, we set $\beta = 1$ to assign equal importance to the archetype update and commitment penalty in the vector quantization loss.

A.3 EVALUATION PROTOCOLS

We adopt task-specific evaluation metrics to ensure fair and meaningful performance comparisons.

Facial Expression Recognition (FER). For FER, we report the classification accuracy (ACC), defined as:

$$\text{ACC} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\hat{y}_i = y_i\right), \tag{1}$$

where N is total number of samples, y_i is the ground-truth label, $\hat{y}_i$ is predicted label, and $\mathbb{I}(\cdot)$ is the indicator function.

Valence-Arousal (VA) Estimation. For VA estimation, we use the Concordance Correlation Coefficient (CCC) for both valence (v) and arousal (a), defined as:

$$\mathrm{CCC}(x, y) = \frac{2\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{2}$$

where ρ_{xy} is the Pearson correlation coefficient between the predicted values x and ground truth y, μ_x and μ_y are the means, and σ_x and σ_y are the standard deviations. The final CCC score is computed as the average of CCC_v and CCC_a:

$$\mathrm{CCC}_{\text{VA}} = \frac{\mathrm{CCC}_v + \mathrm{CCC}_a}{2}. \tag{3}$$

Action Unit Detection (AUD). For AUD, we compute the F1-score for each Action Unit (AU) independently:

$$\mathrm{F1}_k = \frac{2 \cdot \mathrm{Precision}_k \cdot \mathrm{Recall}_k}{\mathrm{Precision}_k + \mathrm{Recall}_k}, \tag{4}$$

where $\mathrm{Precision}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k}$ and $\mathrm{Recall}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k}$ for AU k. We further report the average F1-score across all K AUs:

$$\mathrm{F1}_{\text{avg}} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{F1}_k. \tag{5}$$

B Archetype Reset Mechanism.

To avoid archetype collapse, we introduce a usage-aware reset mechanism that periodically reinitializes underutilized archetypes based on their global selection frequency.

***Global Usage Tracking*:** Let the codebook be denoted as $\mathcal{C} = \{\mathbf{e}_1, \ldots, \mathbf{e}_N\}$, where each $\mathbf{e}_i \in \mathbb{R}^d$ is a learnable archetype. During training, we record the global usage count vector $\mathbf{u} = [u_1, \ldots, u_N] \in \mathbb{N}^N$, where u_i counts the total number of times $\mathbf{e}_i$ was selected as the nearest archetype over all training steps. We define the normalized usage ratio for each code vector as: $\alpha_i = \frac{u_i}{\sum_{j=1}^{N} u_j}, \quad \forall i = 1, \ldots, N$, We then define a fixed threshold $\tau \in (0, 1)$ (e.g., $\tau = 0.01$), and identify the set of underutilized codes: $\mathcal{P}_{\text{reset}} = \{i \mid \alpha_i < \tau\}$.

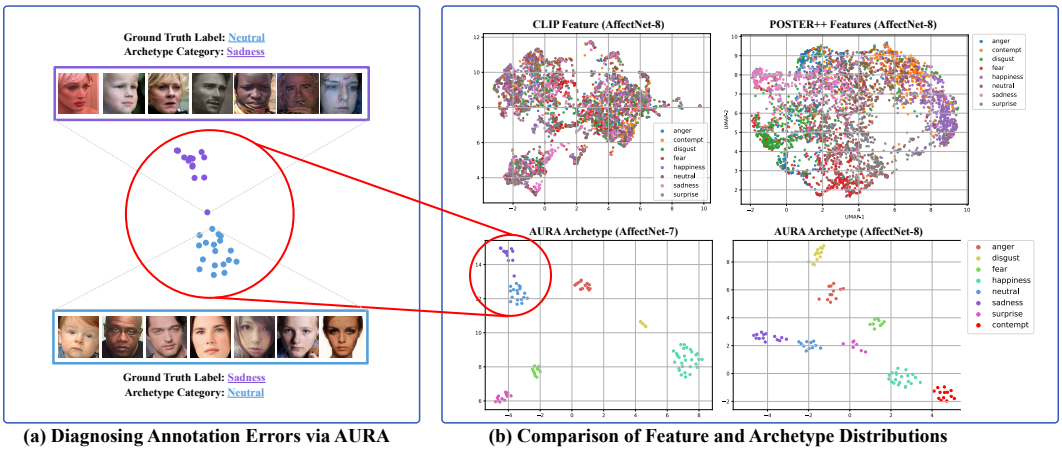
***Archetype Reset*:** For each $i \in \mathcal{P}_{\text{reset}}$, we sample a new feature vector $\mathbf{f}_i \in \mathbb{R}^d$ from the current training batch and reinitialize the archetype as:

$$\mathbf{e}_i \leftarrow \mathbf{f}_i + \boldsymbol{\xi}_i, \quad \text{where } \boldsymbol{\xi}_i \sim \mathcal{N}(0, \sigma^2\mathbf{I}),$$

with $\sigma > 0$ denoting a small Gaussian noise level used to encourage diversity. Alongside the update of $\mathbf{e}_i$, we reset all related accumulators: $u_i \leftarrow 0, \quad \mathbf{c}_i \leftarrow \mathbf{e}_i, \quad n_i \leftarrow 0$, where $\mathbf{c}_i$ denotes the accumulated cluster mean for archetype i and n_i is the cluster size (i.e., the count of features assigned to $\mathbf{e}_i$). Once all underutilized archetypes are updated, we reset the entire usage counter to zero: $\mathbf{u} \leftarrow \mathbf{0}$. This reset mechanism ensures that the codebook dynamically adapts to the evolving data distribution and avoids stagnation due to unused or outdated archetypes.

C Analysis For Learned Archetypes

We analyze the archetypes learned by AURA to understand their structure, distribution, and interpretability across multiple affective tasks. Our study covers (i) comparison with conventional classification model, (ii) error diagnosis and taxonomy refinement, (iii) spatial organization in arousal–valence space, (iv) allocation patterns in Action Unit spaces, and (v) quantitative cross-task statistics. The results show that AURA adaptively allocates representational capacity according to data distribution and emotional complexity, yielding both higher performance and more interpretable affective representations.



(a) Diagnosing Annotation Errors via AURA **(b) Comparison of Feature and Archetype Distributions**

Figure 1: UMAP visualization of archetypes, original CLIP visual features, and POSTER++ features for the AffectNet-7/-8 facial expression recognition task. (a) Diagnosis of annotation errors using AURA; (b) Visualization of feature distributions.

C.1 Emotion Representation Advantage of AURA

AURA vs. Conventional Classification Models: To assess the advantages of AURA over ***conventional label-supervised classification***, which optimizes representations directly under ground-truth labels, we visualize and compare three types of learned features on the AffectNet-8 test set (Fig. 3 (b)): AURA archetypes, original CLIP features, and POSTER++ features. Our observations are as follows: **(i) Original CLIP features** are highly entangled in the affective space, yielding poor emotional separability. **(ii) POSTER++** alleviates some entanglement and improves separability, but many samples remain intertwined and the learned features still lack semantic interpretability. **(iii) AURA archetypes**, in contrast, produce highly distinct and disentangled clusters with strong semantic coherence. These results demonstrate that AURA not only surpasses conventional objectives quantitatively but also yields qualitatively more interpretable and cognitively consistent affective representations.

C.2 Diagnosing Annotation Errors and Refining Emotion Taxonomy via AURA

We conducted an in-depth examination of the learned AURA archetypes and their associated emotion images, and found that, beyond offering inter- and intra-class interpretability (as illustrated in Fig. 3 of the main paper), AURA also serves as an effective tool for diagnosing annotation errors (as illustrated in **Fig. 3 (a)**). Upon thorough inspection, we observe that the AffectNet dataset contains a substantial number of compound expressions, which are inherently challenging to differentiate during the annotation process and therefore susceptible to mislabeling. Thanks to our semantic interpretability of AURA, we are able to systematically probe the samples assigned to each archetype, enabling precise ***analysis, explanation, and error diagnosis***.

As illustrated in **Fig. 3(a)**, we identify two closely related archetypes corresponding to "sadness" and "neutral". Closer inspection of the images assigned to the "**sadness**" archetype, despite being labeled as "neutral" in the ground truth, reveals consistently sorrowful expressions characterized by knitted brows with pronounced glabellar lines, drooping eyelids, a dull gaze, and downward-turned, compressed lips. Conversely, the images mapped to the "**neutral**" archetype, though annotated as "sadness", clearly exhibit neutral facial cues, including level eyebrows, relaxed eyelids, a steady forward gaze, and lips at rest without curvature.

Notably, ***AURA refines the conventional seven-class emotion taxonomy into finer, semantically coherent subsets***, enabling more accurate grouping of visually similar expressions. Such refinement allows AURA to capture subtle variations within a single emotion class, distinguishing, for example, between mild and intense expressions or between pure and compound emotions. This finer-grained partitioning not only improves the structural organization of the affective space but also facilitates the identification of borderline or ambiguous cases that are often misclassified under rigid categorical

schemes. By transcending the limitations of hard class boundaries, AURA provides a more continuous and interpretable representation of emotions, thereby enhancing both the semantic clarity of the learned features and the reliability of emotion annotations in large-scale datasets.

C.3 Archetype Analysis in Arousal–Valence Space

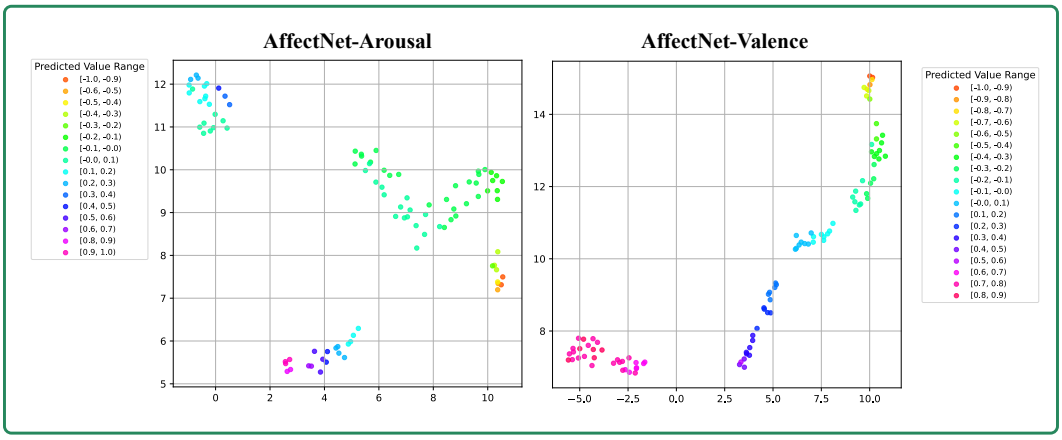


Figure 2: Valence and Arousal Prototye distribution visualization for AffectNet-VA.

AURA For Arousal: A detailed examination of the AURA archetypes in the arousal dimension reveals a distinct spatial clustering pattern that aligns well with the underlying distribution of emotional intensities in the dataset. Specifically, the archetypes aggregate into four primary clusters: those corresponding to arousal values between -1.0 and -0.5 are concentrated in the lower right region (depicted by orange to yellow-green hues) comprising 9 archetypes; the range -0.5 to -0.1 forms a cluster in the mid-right region (yellow-green to cyan) containing 48 archetypes; arousal values from -0.1 to 0.3 cluster in the upper left area (cyan to deep blue) with 23 archetypes; finally, values from 0.0 to 1.0 group near the central bottom area, comprising 20 archetypes.

This distribution reflects the natural emotional landscape captured in the dataset, where the majority of arousal values fall within the moderate range of approximately $[-0.3, 0.3]$. Emotions beyond this range correspond to intensely high or low arousal states, which are less frequently represented in the data and therefore require fewer archetypes for effective modeling. Conversely, the $[-0.3, 0.3]$ interval encompasses typical human emotional intensity, exhibiting rich intra-class variability that necessitates a denser population of archetypes to capture subtle distinctions. For instance, within this moderate arousal range, expressions may vary from calm attentiveness to mild agitation, each distinguished by nuanced facial cues that AURA archetypes effectively encode.

AURA For Valence: Turning to the valence dimension, the archetypes are distributed almost uniformly across the entire $[-1, 1]$ spectrum, with notable concentration in the intervals $[0.6, 1.0]$ and $[-0.5, 0.0]$, which are represented by 31 and 23 archetypes respectively. This allocation corresponds closely with the empirical distribution of valence in the dataset, where highly positive and mildly negative emotional states are more prevalent. The uniform spread and selective densification of archetypes indicate that AURA adapts dynamically to the data's statistical properties, providing finer granularity in emotionally significant regions while maintaining coverage across the full valence range.

Collectively, these findings underscore AURA's capacity to model the continuous valence-arousal affective space with both granularity and efficiency. By allocating archetypes in accordance with the natural distribution and complexity of emotional expressions, AURA achieves a balance between representational compactness and discriminative power, thereby enhancing interpretability and supporting nuanced emotion analysis.

Table 1: Statistics of assigned archetypes across different datasets and tasks. Each row reports the number of assigned archetypes (**Assigned Prot.**), their usage range (min–max, **Prot. Usage**), the total number of samples matched to these archetypes (**Prot. Sample Num.**), and the total number of samples in the dataset (**Sample Num.**).

RAF-DB				
Expression	**Assigned Prot.**	**Prot. Usage**	**Prot. Sample Num.**	**Sample Num.**
anger	11	12–566	710	705
disgust	9	17–288	751	717
fear	6	18–169	287	281
happiness	28	19–1788	4735	4772
neutral	19	8–1288	2417	2524
sadness	12	11–999	2009	1982
surprise	15	14–841	1362	1290
AffectNet-VA (Arousal / Valence)				
Bin Range	**Assigned Prot.**	**Prot. Usage**	**Prot. Sample Num.**	**Sample Num.**
$-1.0 - -0.7$	1 / 3	2341 / 1739–6783	2341 / 10522	3716 / 17989
$-0.7 - -0.4$	3 / 8	1013–5795 / 6741–7719	9206 / 36362	12372 / 31838
$-0.4 - -0.1$	28 / 13	1733–8666 / 1258–7383	63836 / 54243	52069 / 36753
$-0.1 - 0.1$	34 / 20	1389–7014 / 949–6734	71999 / 36927	99781 / 50891
$0.1 - 0.4$	22 / 17	1113–12569 / 154–3948	92903 / 20537	74212 / 29866
$0.4 - 0.7$	7 / 15	1178–3846 / 747–8513	21585 / 45156	30415 / 66280
$0.7 - 1.0$	5 / 24	5370–9079 / 1255–5565	28220 / 86343	21845 / 60793
DISFA (AU12 / AU25)				
Bin Range	**Assigned Prot.**	**Prot. Usage**	**Prot. Sample Num.**	**Sample Num.**
$0.0 - 0.3$	32 / 29	471–5819 / 538–7391	62970 / 54047	65819 / 55054
$0.3 - 0.5$	1 / 2	1226 / 141–2440	1226 / 2581	
$0.5 - 0.7$	2 / 0	111–963 / 0	1074 / 0	21391 / 32156
$0.7 - 1.0$	9 / 13	411–6869 / 769–4588	21940 / 30582	

C.4 Analysis of Archetypes in AU Spaces

In this section, we visualize the learned AURA archetypes for four representative Action Units: AU4, AU12, AU25, and AU26. Across these AUs, a consistent pattern emerges whereby strong activations are associated with relatively few archetypes, while weak or absent activations correspond to a larger number of archetypes. This distribution aligns well with established domain knowledge: strongly activated AUs tend to exhibit more distinctive facial patterns, warranting compact and focused archetype representation, whereas weakly activated or inactive AUs reflect greater variability in appearance, thus requiring a more diverse set of archetypes to capture the underlying heterogeneity.

Despite this general trend, notable differences arise among the four AUs. For AU25, archetypes corresponding to strong activation levels (0.8–1.0) cluster densely in the upper-right region of the latent space, whereas weaker activations (0.1–0.4) concentrate in the lower-right region. This clear spatial segregation validates the discriminative power of AURA archetypes, as AU25's strong activation typically signifies expressions of happiness, while its weaker activation corresponds to distinct emotional states such as disgust or contempt, underscoring AURA's capacity to capture fine-grained affective differences.

Similarly, for AU4 and AU26, strongly activated archetypes (activation levels between 0.6 and 1.0) are tightly clustered in the upper-left region, contrasting with other activation levels aggregated in the lower-right region. This spatial dichotomy reflects AURA's robust ability to sharply distinguish between active and inactive AU states.

In the case of AU12, the archetypes corresponding to moderate (0.4–0.7) and strong (0.7–1.0) activations form a contiguous cluster. This pattern is consistent with the known physiological characteristics of AU12, which often manifests with subtle gradations of activation due to the underlying facial muscle movements involved. Such nuanced clustering illustrates AURA's sensitivity to the fine-scale variations inherent in AU12 activation levels.

Overall, these findings demonstrate that AURA archetypes effectively model the complex distribution of AU activations, balancing compactness for strongly activated units with diversity for weaker activations, thereby capturing both the discriminative and variable nature of facial action units in a semantically meaningful manner.

C.5 Quantitative Analysis of Archetype Distribution Across Tasks

We present a quantitative analysis of archetype allocation patterns across three representative affective tasks: categorical facial expression recognition (RAF-DB), continuous arousal–valence estimation (AffectNet-VA), and action unit detection (DISFA). The statistics in Table 1 summarize the number

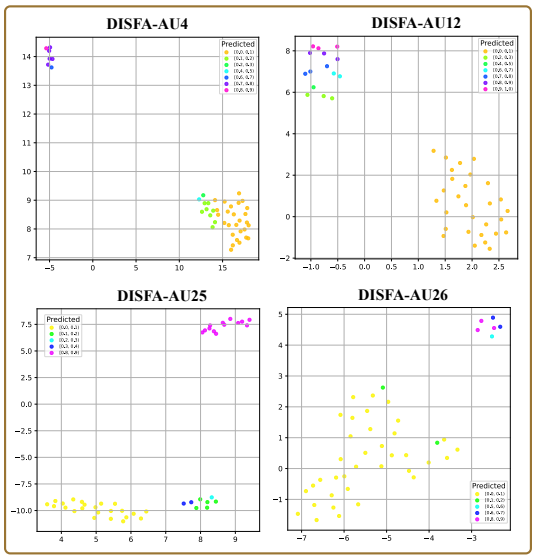


Figure 3: DISFA Archetypes distribution visualization.

of assigned archetypes (**Assigned Prot.**), their usage ranges (**Prot. Usage**), the total number of samples matched to these archetypes (**Prot. Sample Num.**), and the total dataset sample counts (**Sample Num.**). This quantitative view allows us to interpret how the AURA mechanism distributes representational capacity across different affective states, intensities, and data densities.

RAF-DB (Expression Recognition): Archetype allocation varies substantially across the seven expression categories. High-frequency and visually diverse categories such as *happiness* (28 archetypes, usage range: 19–1788) and *neutral* (19 archetypes, 8–1288) receive a larger number of archetypes with broad usage spans, indicating high intra-class variability. Conversely, categories such as *fear* (6 archetypes, 18–169) and *disgust* (9 archetypes, 17–288) have fewer archetypes and narrower ranges, reflecting lower diversity and sample counts. *Sadness* and *surprise* fall in between, with moderate archetype counts but concentrated usage, suggesting more homogeneous visual patterns.

AffectNet-VA (Arousal / Valence Estimation): In the continuous affective space, archetype allocation strongly correlates with data density. Extreme affective regions (e.g., -1.0–-0.7, 0.7–1.0) exhibit fewer archetypes (1–5 for arousal, 3–24 for valence) and lower matched sample counts, due to the scarcity of highly polarized emotions in the dataset. In contrast, the central regions (e.g., -0.1–0.1, 0.1–0.4) receive the largest number of archetypes (up to 34 for arousal, 20 for valence) and significantly higher sample counts, capturing subtle variations in near-neutral affective states. This aligns with AffectNet's known bias toward mild or mixed emotions.

DISFA (Action Unit Detection): For AU-based modeling, archetype allocation distinguishes between inactive/low-intensity and highly active facial muscle states. In AU12, the 0.0–0.3 range dominates with 32 archetypes and 62,970 matched samples, while the mid-intensity range (0.3–0.5) is covered by only a single archetype, indicating rare occurrences. High-intensity activations (0.7–1.0) have fewer archetypes (9 for AU12, 13 for AU25) but disproportionately high sample counts, suggesting these expressions, while less visually diverse, are relatively frequent in the dataset. Notably, AU25 has no archetypes in the 0.5–0.7 range, implying low occurrence or ambiguity in this activation intensity.

This quantitative view highlights that archetype allocation in AURA is inherently data-adaptive. Tasks and affective states with high visual diversity or dense sample distributions receive more archetypes with wider usage ranges, while homogeneous or rare states are represented by fewer archetypes with concentrated usage. This property ensures both representation efficiency and strong discriminative capacity across heterogeneous affective modeling scenarios.

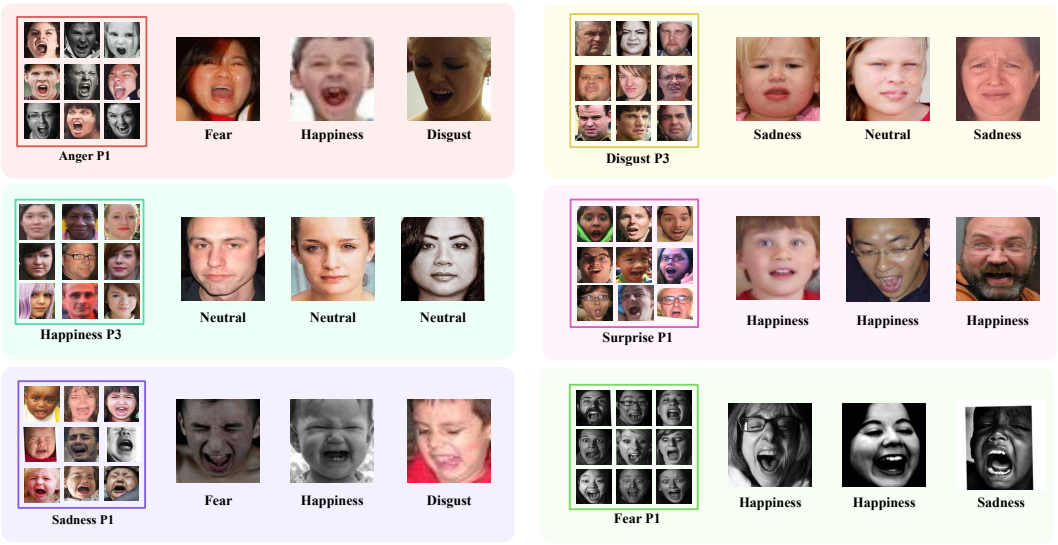


Figure 4: Failure Case Analysis. For each panel, the **Left** column shows correctly labeled samples are aligned with the corresponding expression archetype, while the **Right** column shows samples assigned to this archetype despite belonging to other expression classes.

D Failure Case Analysis

To further evaluate the behavioral characteristics of AURA, we conduct a dedicated failure-case analysis, with results shown in Figure 4. For each panel, the **Left** block presents the support images that are aligned with a particular expression archetype, while the **Right** side displays misclassified samples that were assigned to this archetype despite belonging to different ground-truth classes. This setup enables simultaneous examination of **(i) how the model behaves on individual samples and (ii) what semantic structures lead to these error**s. By comparing each mispredicted sample with the defining archetype, we can clearly identify which facial configuration the model relied on (e.g., mouth shape, eye tension, or brow contraction), thereby revealing the semantic basis of the model's mistakes.

Across all panels, a clear and consistent pattern emerges: although the categorical prediction is wrong, the assigned archetype remains semantically justified because the misclassified faces share the **same fine-grained facial-muscle configuration** as the archetype's support set. Crucially, these archetypes do not simply correspond to **"open-mouth"** expressions in general; each archetype encodes a distinct sub-pattern of facial activation. For example, **Anger P1** captures an expression characterized by a maximally forceful, ***lower-face*** (dominant contraction, a wide, tense mouth opening with strong jaw engagement), so samples with an equally forceful lower-face stretch are naturally drawn to this archetype, even if their categorical label is Fear or Disgust. In contrast, **Sadness P1** reflects a different structural signature: although the mouth is open, the expression is dominated by ***upper-face*** tension (furrowed brows, nasal contraction, and a drooping eye region). Misassigned samples in this panel exhibit precisely this upper-face configuration, explaining why the model anchors them to this archetype. **Fear P1**, on the other hand, is defined by a high-intensity, mouth-stretched configuration ***without pronounced brow contraction***, often manifested in ***grayscale*** images within the dataset; misclassified Happiness or Sadness samples assigned to Fear P1 share this same "wide-mouth, minimal-brow-movement" structure. These observations show that AURA's archetypes capture fine-grained, physiologically meaningful facial patterns, and that mispredictions occur not because the model is confused arbitrarily, but because the sample's *local facial configuration* aligns more closely with a specific archetypal mode than with its discrete ground-truth label.

In the **Disgust P3** panel, the archetype is characterized by nose wrinkling, ***raised upper lip***, and narrowed eyes. Samples labeled as Sadness or Neutral are assigned to this archetype because they exhibit a very similar nose–mouth configuration, suggesting label ambiguity or overlapping expression cues in the dataset. In **Happiness P3**, the archetype reflects subtle, ***low-intensity smiles***

close to Neutral faces. The misclassified Neutral samples assigned to this archetype have very ***mild lip corners*** raised and relaxed upper faces, revealing that the boundary between low-intensity Happiness and Neutral is intrinsically fuzzy. Finally, **Surprise P1** collects wide-eyed, mouth-open faces; the mispredicted Happiness examples assigned to it also display strong ***"wow"-like*** configurations, again showing that the model is grouping samples by a coherent facial pattern rather than arbitrary noise.

Across all panels, AURA's failure cases reveal a consistent and meaningful pattern: even when the discrete class prediction is incorrect, the assigned archetype remains semantically well-aligned with the facial structure of the input. This is because the misclassified faces share highly similar AU configurations, intensity patterns, and local geometry with the archetype's support set.

The misassigned samples naturally fall into these archetypes because their fine-grained facial structure more closely matches the archetype's learned configuration than their categorical label suggests. Importantly, such mismatches often arise from dataset statistics: if a structural pattern appears predominantly in one class (e.g., scream-like faces in Anger) and is underrepresented in others (e.g., scream-like Happiness), the model will gravitate toward the archetype that best captures that structure. Rather than being a weakness, this demonstrates the strength of our weakly supervised archetype formulation: **AURA prioritizes genuine facial-structure similarity over noisy or ambiguous labels, offering per-sample insight into which semantic mode the model activated and why.**

E SENSITIVITY TO THE CHOICE OF $K_{\max}$

Table 2: Sensitivity of AURA to the choice of $K_{\max}$ on RAF-DB.

$K_{\max}$	K_{stable}	ACC (%)	Convergence Epoch	Archetype Distribution	GFLOPs
150	100	94.1	153	Anger 11 / Disgust 9 / Fear 6 / Happiness 28 / Neutral 19 / Sadness 12 / Surprise 15	0.26
200	99	94.0	147	Anger 11 / Disgust 9 / Fear 6 / Happiness 26 / Neutral 20 / Sadness 12 / Surprise 15	0.29
300	102	94.2	158	Anger 11 / Disgust 10 / Fear 6 / Happiness 29 / Neutral 18 / Sadness 13 / Surprise 15	0.44
400	101	94.1	155	Anger 12 / Disgust 10 / Fear 6 / Happiness 29 / Neutral 19 / Sadness 13 / Surprise 12	0.48

This section clarifies the distinction between the predefined upper bound of archetypes and the effective number that AURA ultimately employs. AURA differentiates between two quantities: the predefined upper bound $K_{\max}$, typically set between 150 and 400, and the stable number of active archetypes K_{stable} that emerges automatically during training. The value of K_{stable} is not determined by $K_{\max}$; instead, $K_{\max}$ is chosen to be over-complete to ensure sufficient representational capacity, while the model identifies a much smaller and semantically meaningful subset of archetypes according to the task. Empirically, AURA converges to approximately 100 active archetypes for expression recognition on AffectNet and RAF-DB, and around 40 for AU recognition on EmotioNet, regardless of the initial choice of $K_{\max}$.

The adaptivity of AURA arises from two key components: the Adaptive Archetype Regularization and the Archetype Contextualization Module. Together, these mechanisms encourage informative archetypes to receive substantial assignments while suppressing redundant ones. Let the archetype codebook be $\mathcal{C} = \{e_1, \ldots, e_{K_{\max}}\}$. During training, AURA maintains a global usage counter u_k for each archetype, recording how often e_k is selected as the primal archetype. A normalized usage ratio is computed as $\alpha_k = u_k / \sum_{j=1}^{K_{\max}} u_j$. Archetypes with usage ratio below a threshold τ (e.g., $\tau = 0.01$) are considered under-utilized. During the early stage of optimization (first 20–30 epochs), an Archetype Reset Mechanism reinitializes the embeddings of under-utilized archetypes to avoid premature collapse of dictionary capacity. After this warm-up stage, the usage stabilize, producing a consistent K_{stable} that remains largely invariant to $K_{\max}$. At inference time, archetypes with $\alpha_k < \tau$ are pruned, yielding a compact and interpretable dictionary.

To examine robustness with respect to $K_{\max}$, we conduct a sensitivity analysis on RAF-DB with $K_{\max} \in \{150, 200, 300, 400\}$. As summarised in Table 2, the resulting number of active archetypes remains within a narrow range (99-102) for all settings. Model accuracy varies within only 0.2%, convergence epochs remain comparable, and the class-wise archetype distribution exhibits high consistency. The computational cost increases moderately for larger $K_{\max}$ due to the expanded over-complete dictionary, but this does not affect the effective number of utilized archetypes. These results demonstrate that AURA automatically identifies a suitable number of archetypes and is robust

and insensitive to the initial value of $K_{\max}$ in terms of performance, convergence behaviour, and the granularity–efficiency trade-off.

F Encoder-Agnostic Behavior of AURA

Table 3: Evaluating AURA with CLIP and DINO encoders. "Official FT" denotes full fine-tuning; "AURA (ours)" denotes frozen encoder + AURA.

Method	Encoder	RAF-DB Acc	AffectNet-VA CCC	EmotioNet AU-F1
CLIP FT	CLIP (finetuned)	89.1	66.4	62.3
AURA-CLIP	**CLIP (official frozen)**	**94.0 (+4.9)**	**74.1 (+7.7)**	**67.3 (+5.0)**
DINO FT	DINO (finetuned)	88.3	65.4	60.7
AURA-DINO	**DINO (official frozen)**	**92.6 (+4.3)**	**72.0 (+6.6)**	**65.4 (+4.7)**

To isolate the contribution of the proposed archetype mechanism from the representational priors of CLIP, we extend our study by conducting a dedicated evaluation using both **CLIP** (Radford et al., 2021) and **DINO** (Caron et al., 2021). In particular, DINO serves as a purely self-supervised vision backbone without any text–image alignment, providing a controlled testbed to determine whether AURA's improvements originate from CLIP's semantically aligned visual space or from the archetype modeling itself. This analysis allows us to rigorously disentangle the effects of encoder pretraining and the structural advantages introduced by AURA.

To ensure a fair and comprehensive comparison, we evaluate each encoder under two parallel configurations. For DINO, we first consider **DINO FT (full fine-tuning)**, where the entire DINO student network is optimized jointly with task-specific heads following official training protocols. All backbone parameters are trainable in this setting. We then construct a second configuration, **AURA-DINO**, in which the CLIP backbone in our main AURA model is replaced by the pretrained DINO encoder, but the encoder is kept *entirely frozen*. DINO is used strictly as a feature extractor, and only AURA's archetype modules are updated during training. This frozen setting cleanly isolates the effect of archetype modeling by removing any benefits from encoder adaptation.

For completeness, we perform the same two configurations using CLIP. In the **CLIP FT** condition, the full CLIP visual encoder is fine-tuned end-to-end along with the task heads. In the **AURA-CLIP** setting, the official pretrained CLIP encoder is kept *frozen*, and AURA operates solely on top of its fixed visual embeddings. By aligning CLIP and DINO under identical experimental protocols, this cross-encoder evaluation enables a controlled investigation of whether AURA's gains are tied to CLIP's vision–language alignment or generalize across fundamentally different pretraining paradigms.

The results, summarized in Table **??**, show that AURA delivers *strong and consistent performance gains* across both CLIP and DINO. Notably, despite relying on *frozen* DINO features without any fine-tuning of backbone parameters, AURA-DINO achieves substantial improvements over the fully fine-tuned DINO baseline: +4.3% accuracy on RAF-DB, +6.6 CCC on AffectNet-VA, and +4.7 F1 on EmotioNet. While absolute performance under DINO is naturally lower than under CLIP—reflecting CLIP's multimodal supervision and richer semantic priors—the *relative* gains contributed by AURA remain highly consistent across FER, VA, and AU tasks.

An additional observation further strengthens this conclusion: although CLIP outperforms DINO in absolute terms, owing to its stronger semantic alignment, *the performance gains introduced by AURA are even larger on CLIP than on DINO*. This pattern reveals two important insights. First, the benefits of AURA do not arise from CLIP's multimodal alignment alone, as similar gains appear with a purely visual self-supervised encoder. Second, AURA is capable of fully exploiting the representational capacity of stronger pretrained models—particularly CLIP's semantically aligned embedding space—yielding more structured latent geometries, reduced intra-class ambiguity, and improved predictive accuracy.

Overall, these findings demonstrate that AURA is fundamentally **encoder-agnostic**. Its improvements stem from the archetype-based feature structuring itself rather than any encoder-specific advantage. AURA consistently enhances a wide range of pretrained vision models by discovering canonical semantic anchors and decomposing fine-grained intra-class variability, independently of whether the underlying backbone is multimodally aligned (CLIP) or purely visual (DINO).

G Sensitivity Analysis of Training Hyperparameters

This section provides a comprehensive and detailed analysis of the sensitivity of AURA to its major training hyperparameters. Although AURA contains several components in its loss formulation, the overall training process is intentionally designed to remain simple, stable, and fully reproducible across datasets and tasks.

G.1 Sensitivity to the Loss-Weight Coefficients λ

This appendix provides a detailed analysis of the sensitivity of AURA with respect to the loss-weight coefficients used in the total objective

$$\mathcal{L} = \lambda_{\text{Proj}}\mathcal{L}^{\text{VAS}} + \lambda_{\text{Arc}}\mathcal{L}^{\text{Arc}} + \lambda_{\text{Contx}}\mathcal{L}^{\text{Contx}}.$$

Across all experiments, we adopt the uniform setting $\lambda_{\text{Proj}} = \lambda_{\text{Arc}} = \lambda_{\text{Contx}} = 1$, without any tuning. This simple configuration consistently produces strong results on RAF-DB, EmotioNet, and AffectNet for FER, AU detection, and VA regression, indicating that AURA does not rely on delicate loss balancing.

The reason equal weighting works is rooted in the architectural decomposition of AURA: the three loss terms act on **disjoint parameter subsets**, preventing gradient competition. The projection-supervision loss $\mathcal{L}^{\text{VAS}}$ governs only the projection head and aligns archetype mixtures with task semantics. The archetype-regularization loss $\mathcal{L}^{\text{Arc}}$ acts exclusively on the archetype dictionary, shaping geometry, sparsity, and separation. The contextual-interaction loss $\mathcal{L}^{\text{Contx}}$ applies only to the attention module that mediates cross-archetype message passing. Since each component optimizes an independent representational layer, the gradients naturally remain compatible in scale even without explicit balancing.

Within $\mathcal{L}^{\text{Arc}}$, the components $\mathcal{L}^{\text{Assign}}$, $\mathcal{L}^{\text{Dis}}$, and $\mathcal{L}^{\text{Reg}}$ operate on bounded similarity measures (cosine similarity or simplex-normalized assignments), which keeps their magnitudes comparable. These forces **act in complementary directions**—assignment encourages confident usage of archetypes, the distance term promotes geometric separation, and the regularization term stabilizes class- or score-conditioned structure. Maintaining equal weights ensures a stable and unbiased equilibrium: no component overwhelms the archetype geometry, and the system avoids collapse or overspreading. Although the form of $\mathcal{L}^{\text{Reg}}$ differs between FER, AU, and VA tasks, all variants enforce inter-class distinction, intra-class compactness, and diversity; each is designed to operate within similar numerical ranges, further supporting robust behavior under $\lambda = 1$.

To empirically verify sensitivity, we vary each coefficient within $\lambda \in \{0.5, 1.0, 1.5\}$ and evaluate all combinations. Across all datasets, the resulting performance remains remarkably stable, with only minor fluctuations attributable to convergence speed rather than model quality. The comprehensive results are summarized in Table 4. The negligible variation confirms that AURA's training dynamics are inherently well-balanced and do not require hyperparameter tuning for loss weights.

G.2 Sensitivity to the Margin Parameter m

This appendix provides a detailed analysis of the sensitivity of AURA to the choice of the margin m used in the archetype regularization terms. Conceptually, the margin m plays a unified role across tasks: it specifies the threshold that separates pairs that *should be close* from pairs that *should be separated*. As long as m is chosen within a semantically meaningful range that is consistent with the underlying similarity or distance scale, the model remains stable and does not require fine-grained tuning.

Table 4: Comprehensive sensitivity analysis of the loss-weight coefficients λ across FER (RAF-DB), AU (EmotioNet), and VA (AffectNet-VA).

$(\lambda_{\text{Proj}}, \lambda_{\text{Arc}}, \lambda_{\text{Contx}})$	RAF-DB ACC	Epoch	AU-F1	Epoch	VA-CCC	Epoch
(1.0, 1.0, 1.0) **(Default)**	94.0	152	67.3	136	74.1	188
(0.5, 1.0, 1.0)	93.8	187	67.1	165	74.1	203
(1.5, 1.0, 1.0)	94.0	149	66.8	138	74.2	184
(1.0, 0.5, 1.0)	94.2	176	67.1	148	73.9	189
(1.0, 1.5, 1.0)	94.3	142	67.2	136	74.0	196
(1.0, 1.0, 0.5)	93.8	180	66.9	164	74.0	207
(1.0, 1.0, 1.5)	94.1	146	67.2	122	74.2	175

In the classification setting, the inter-class separation loss penalizes pairs of class centers whose cosine similarity exceeds the margin m. Formally, the loss takes the form

$$\mathcal{L}_{\text{inter}} = \frac{1}{|\mathcal{S}|} \sum_{(c,c') \in \mathcal{S}} \max\big(0,\ \cos(\tilde{\mu}_c, \tilde{\mu}_{c'}) - m\big),$$

where $\mathcal{S}$ indexes pairs of distinct classes and $\tilde{\mu}_c$ denotes the normalized center of class c. In this formulation, m directly defines the maximum allowable similarity between different classes: pairs with cosine similarity below m incur no penalty, whereas pairs with similarity above m are pushed apart. Since cosine similarities on normalized centers lie in $[0, 1]$, choosing m in the range $[0.2, 0.4]$ yields a natural trade-off between enforcing sufficient separation and avoiding overly aggressive repulsion. Empirically, this interval provides a stable operating region for both FER and AU classification.

In the regression setting, the margin appears in the score-aware attraction and repulsion losses, which jointly control how archetypes with similar or dissimilar predicted scores are arranged in the feature space. Let Δ^{ij} denote the score difference between two archetypes and d^{ij} their cosine distance. The attraction loss encourages archetypes with similar predicted scores to be close, using a hinge on $(d^{ij} - m)$, while the repulsion loss enforces separation for archetypes with divergent scores, using a hinge on $(m - d^{ij})$. In this case, m plays a dual role: it defines the boundary between "similar-score pairs" and "dissimilar-score pairs", and it sets the tolerance radius within which attraction is active or beyond which repulsion becomes necessary. Because valence–arousal targets span a broader semantic range (after normalization to $[-1, 1]$), meaningful score gaps tend to be larger, and slightly smaller margins in the interval $m \in [0.1, 0.3]$ yield stable and effective behavior.

To empirically validate the above analysis, we conduct a controlled sweep over a range of margin values and evaluate performance on all three tasks. The results are reported in Table 5. As expected, performance degrades only when m is set too small, which leads to excessively strong separation forces and can fragment the representation, or when m is set too large, which weakens the separation and reduces discriminability. Within the intermediate ranges described above, the accuracy, F1, and CCC metrics remain stable, confirming that AURA is insensitive to moderate changes in m and does not rely on careful tuning of this hyperparameter.

Table 5: Sensitivity analysis of the margin m across FER (RAF-DB), AU detection (EmotioNet), and VA regression (AffectNet-VA).

m	RAF-DB (ACC)	EmotioNet (AU-F1)	AffectNet-VA (CCC)
0.1	93.2	66.7	73.9
0.2	93.9	67.2	**74.1**
0.3	**94.1**	**67.3**	74.0
0.4	94.0	67.3	73.8
0.5	93.6	66.9	73.4

REFERENCES

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5562–5570, 2016.

Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 2584–2593. IEEE, 2017.

S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *Affective Computing, IEEE Transactions on*, 4(2):151–160, 2013.

Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *arXiv preprint arXiv:1708.03985*, 2017.

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.