CONTEXT OR FACE? LOCATING EMOTION REPRESENTATIONS IN CLIP WITH SPARSE AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models such as CLIP can associate emotions with both facial expressions and contextual cues, but their predictions do not reveal which visual evidence contributes to these judgments. We investigate how joy and sadness are represented across facial and contextual evidence in CLIP using sparse autoencoders (SAEs), causal feature ablation, spatial localization, and cross-carrier linear probing. Using FindingEmo decision-factor annotations and cross-fitted feature discovery, we identify context-associated SAE features for both emotions with distinct evidential profiles. Sadness feature 9960 shows reproducible held-out context selectivity, with 95% bootstrap confidence intervals excluding zero across all three folds. Joy feature 7406 shows consistent contextual enrichment and positive held-out selectivity, although its confidence intervals include zero. Both features rank above all 200 randomly sampled SAE control features in context selectivity. At the dense-representation level, a linear probe trained to distinguish joy from sadness using only face-driven images transfers strongly to context-driven images, achieving 0.851 ± 0.011 balanced accuracy and 0.930 ± 0.006 AUROC. These results show that context-associated behavior at the level of individual sparse features can coexist with emotion-discriminative information that remains linearly accessible across facial and contextual evidence.

1 INTRODUCTION

Visual emotion recognition has traditionally emphasized facial expressions, yet emotions in natural scenes may also be conveyed through social interactions, body cues, objects, and the surrounding environment Kosti et al. (2017); Mertens et al. (2024). Context can provide important evidence when facial information is ambiguous or absent, but can also introduce spurious associations between scenes and emotion labels Yang et al. (2023). This raises an interpretability question beyond recognition accuracy: when a model associates an image with an emotion, what visual evidence supports that association?

Vision-language models (VLMs) provide a useful setting for studying this question. CLIP learns aligned image–text representations from large-scale image–text pairs Radford et al. (2021), and recent work has extended VLMs to emotion-oriented representations, contextual emotion recognition, and fine-grained affective understanding Zhang et al. (2023); Etesam et al. (2024); Lian et al. (2025). However, successful emotion recognition does not reveal which internal visual representations contribute to the resulting emotion association.

Mechanistic interpretability provides tools for examining this internal structure. Sparse autoencoders (SAEs) decompose dense neural activations into sparse features that can be inspected and intervened upon. In language models, SAEs have recovered interpretable features with causal effects on model behavior Huben et al. (2024), while recent mechanistic studies have begun to investigate emotion representations specifically in LLMs Tak et al. (2025); Sofroniew et al. (2026). SAE-based interpretability has also extended to vision-language representations, where sparse features can capture interpretable visual concepts with different spatial scope Zaigrajew et al. (2025); Ro et al. (2026). However, considerably less is known about the internal mechanisms supporting emotion representations in VLMs, particularly when the same emotion is conveyed through different visual evidence.

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.

We investigate whether emotion-related representations in CLIP reflect the emotional concept, the visual source through which it is conveyed, or a combination of both. Using FindingEmo decision-factor annotations Mertens et al. (2024), we distinguish images whose joy or sadness judgments are associated primarily with facial expressions from those associated with contextual information. With a pretrained Matryoshka Sparse Autoencoder (MSAE), we examine sparse features through activation, spatial localization, and causal ablation. Context-associated candidates are separated from evaluation through cross-fitting and assessed on held-out images using bootstrap confidence intervals, random SAE controls, and, where eligible, pre-specified activation-matched controls.

We complement this analysis with cross-carrier linear probing of CLIP's dense representation. A classifier trained to distinguish joy from sadness using only face-driven images is evaluated without adaptation on context-driven images. This separates two questions: whether individual sparse features exhibit context-associated causal behavior and whether emotion-discriminative information remains accessible across visual carriers.

Our results show that the visual evidence underlying an emotion-related output can be distinguished even when the output emotion itself is the same. This matters because emotion recognition is typically evaluated by whether the predicted label is correct, treating correct predictions as equivalent even when they may be supported by different visual evidence. For example, a sadness-related output associated with a feature responding to funeral and memorial scenes can involve different internal visual evidence from one associated with a feature responding to a crying face, although output-level evaluation alone does not expose that distinction. Our framework provides a way to ask which visual evidence contributes to an emotion-related model output, enabling evaluation of not only what emotion information a model represents, but also how that information is supported by the image.

2 Related Work

2.1 Contextual Emotion Recognition and Vision-Language Models

Emotion recognition in natural scenes increasingly incorporates evidence beyond facial expressions. EMOTIC introduced annotations for people embedded in unconstrained visual contexts and demonstrated the value of jointly modeling people and surrounding scenes Kosti et al. (2017). CAER-Net explicitly separates facial and contextual representations Lee et al. (2019), while EmotiCon integrates facial, gait, scene, and socio-dynamic information Mittal et al. (2020). Context can nevertheless introduce spurious correlations, motivating context-deconfounding and debiasing approaches Yang et al. (2023; 2024). Recent work has further modeled interactions and conflicts among head, body, and contextual evidence Shen et al. (2025). FindingEmo provides complementary decision-factor annotations describing the visual evidence underlying human emotion judgments Mertens et al. (2024); we use these annotations to distinguish face-driven and context-driven conditions without assuming that the presence of a face makes an image face-driven.

VLMs extend emotion recognition by aligning visual representations with natural language. CLIP provides a general image–text representation through contrastive pretraining Radford et al. (2021), while EmotionCLIP learns emotion-oriented representations through contextual encoding and sentiment-guided contrastive learning Zhang et al. (2023). Subsequent work has explored contextual emotion recognition with large VLMs Etesam et al. (2024), alignment of facial, body, and situational information through multi-level prompting Gao et al. (2024), and richer affective reasoning with models such as EmoVIT and AffectGPT Xie et al. (2024); Lian et al. (2025). General-purpose VLMs nevertheless show characteristic errors and robustness limitations in evoked emotion recognition Bhattacharyya & Wang (2025). These studies primarily evaluate model outputs; our work instead asks which internal visual representations contribute to emotion-related CLIP outputs as the source of emotional evidence changes.

2.2 Mechanistic Interpretability of Emotion

Mechanistic interpretability seeks to characterize model behavior through internal representations and computations rather than input–output behavior alone. Sparse autoencoders decompose dense neural activations into overcomplete sparse representations. In language models, SAEs have recovered interpretable features and components causally involved in model behavior Huben et al.

(2024). Subsequent work has examined SAE scaling and evaluation Gao et al. (2025), causal feature circuits Marks et al. (2025), features involved in in-context learning Kharlapenko et al. (2025), classification and transfer Gallifant et al. (2025), and targeted activation steering Goyal et al. (2025).

Mechanistic work has also begun to investigate emotion specifically in language models. Tak et al. (2025) identify emotion-related representations in autoregressive LLMs and use causal interventions to examine their functional role. Recent work from Anthropic similarly investigates internal representations associated with emotion concepts and their functional effects on language-model behavior Sofroniew et al. (2026). These studies establish that emotion can be studied at the level of internal mechanisms rather than only through output-level classification. Our work brings this question to visual representations and asks how emotion-related mechanisms in a VLM depend on the source of visual evidence.

A related methodological question is how sparse features should be interpreted. Movva et al. (2025) frame SAE features as sources of interpretable hypotheses whose relationship to a target can subsequently be evaluated. This distinction between discovery and validation motivates our analysis, in which candidate features are identified from activation and intervention patterns and subsequently evaluated on held-out images.

2.3 Sparse Autoencoders for Vision-Language Models

SAE-based interpretability has recently expanded to visual and multimodal representations. PatchSAE extracts interpretable patch-level concepts from CLIP and spatially attributes their evidence within images Lim et al. (2025). Matryoshka Sparse Autoencoders provide hierarchical sparse representations for VLMs, including CLIP, and expose semantically meaningful concepts useful for downstream analysis Zaigrajew et al. (2025). Related evaluations have examined sparse visual representations for applications including out-of-distribution analysis and model steering Olson et al. (2025).

Recent work has moved beyond assigning semantic labels to visual SAE features. Ro et al. (2026) characterize the information scope of CLIP SAE features, distinguishing localized from more globally aggregated visual information and relating these differences to downstream behavior. Yamauchi et al. (2026) similarly investigate decision-critical visual features using sparse internal representations. Together, these studies suggest that understanding a visual SAE feature requires considering not only what concept it appears to encode, but also where its evidence occurs and whether it functionally influences model behavior.

Our work builds on this direction but asks how sparse features associated with emotion-related model behavior depend on the visual source through which emotional evidence is conveyed. We combine SAE activation, spatial localization, and causal ablation across facial and contextual decision-factor conditions with cross-carrier probing of CLIP's dense representation. This distinguishes context-associated behavior of individual sparse features from emotion-discriminative information that remains accessible across visual carriers.

3 Methodology

3.1 CLIP and Sparse Autoencoder Representation

We use CLIP ViT-L/14-336 Radford et al. (2021) and extract post-MLP residual-stream activations from transformer layer 22. After removing the class token, the 576 spatial patch representations ($d = 1024$) are encoded using a pretrained Matryoshka BatchTopK Sparse Autoencoder (MSAE) Zaigrajew et al. (2025) into 65,536-dimensional sparse representations.

For causal intervention, selected feature j is set to zero across all patches, $z_{p,j}^{(-j)} = 0$, while all other latent dimensions remain unchanged. The modified representation is decoded and substituted into CLIP's residual stream, after which the remaining transformer computation proceeds normally. Because SAE reconstruction differs from the original CLIP activation, all interventions are evaluated relative to the corresponding unablated SAE reconstruction, isolating feature removal from SAE reconstruction error.

3.2 FINDINGEMO POPULATIONS

We use FindingEmo Mertens et al. (2024) to construct mutually exclusive face-driven, context-driven, and neutral decision-factor populations for joy and sadness. The same emotion and decision-factor definitions are used throughout the SAE and probing analyses. Complete emotion-group definitions, filtering criteria, exclusions, and population counts are provided in Appendix B.

3.3 SPARSE FEATURE DISCOVERY

Following the hypothesis-generation perspective of Movva et al. (2025), we separate feature discovery from validation and do not treat activation or semantic coherence alone as evidence of functional importance. Facial-expression-associated features are first characterized through activation, localization, and causal ablation as reference features. Context-associated candidates are then discovered independently for joy and sadness using only context-driven FindingEmo images.

For each context image, we retain 300 unique candidate features from the 10,000 largest positive patch–feature activations. Candidates are ablated individually across all patches and ranked primarily by mean target-logit drop relative to the unablated SAE reconstruction, using the prompts *"a photo of a joyful scene"* and *"a photo of a sad scene"*.

Discovery uses three-fold cross-fitting: approximately two thirds of context images are used for feature discovery in each fold and the remaining third is held out, such that every context image contributes once to held-out evaluation. Primary candidates are retained based on reproducible recovery across discovery folds together with coherent pre-held-out contextual activation patterns. Candidate selection is completed before examining held-out context-selectivity results.

3.4 CAUSAL ABLATION AND CONTEXT SELECTIVITY

For feature j and image x, causal contribution to the target emotion is defined as

$$\Delta_j(x) = \ell_{\text{recon}}(x) - \ell_{-j}(x), \tag{1}$$

where ℓ_{recon} and ℓ_{-j} denote the target-emotion logits before and after feature ablation, respectively. Positive Δ_j therefore indicates that feature removal decreases the target output.

For context (C), face (F), and neutral (N) conditions, context selectivity is

$$S_j = \min\left(\bar{\Delta}_j^C - \bar{\Delta}_j^F,\, \bar{\Delta}_j^C - \bar{\Delta}_j^N\right), \tag{2}$$

so $S_j > 0$ requires a larger mean causal effect for context-driven images than for both comparison conditions.

Evaluation is performed separately on each fold's held-out context images; face and neutral populations are independent of context-feature discovery. We estimate 95% percentile confidence intervals for S_j using 5,000 nonparametric bootstrap resamples within each condition.

3.5 CONTROL ANALYSES

For each selected feature, we compare context selectivity with 200 uniformly sampled SAE features using the same reconstruction-based ablation procedure. We report the target's rank and the add-one empirical upper-tail probability

$$p_{\text{emp}} = \frac{1 + \#\{S_{\text{control}} \geq S_{\text{target}}\}}{1 + N_{\text{control}}}. \tag{3}$$

This feature-level comparison is distinct from the image-level bootstrap analysis of held-out selectivity.

We additionally use pre-specified activation-matched controls. Eligible features must have context-population image-level activation frequency within 0.10 of the target, patch-frequency ratio within $[0.75, 1.25]$, and mean maximum-activation ratio within $[0.75, 1.25]$. All qualifying features are retained and the criteria are not relaxed based on the number of matches. When no feature satisfies all criteria, no matched comparison is performed.



Figure 1: Spatial localization of facial-expression-associated SAE features on FindingEmo. The left three columns show feature 2163, which concentrates around smiling facial regions, while the right three columns show feature 5284, which responds to facial regions in sad and crying examples.

3.6 Spatial Localization

We map the strongest positive feature activations among CLIP's 24×24 patch grid back to image space and use the ten strongest active patches for qualitative localization. Additional activation statistics and image-ranking procedures are provided in Appendix D. Localization is used to characterize where a feature responds and is not treated as evidence of causal importance.

3.7 Cross-Carrier Linear Probing

To test whether joy–sadness information remains accessible across visual carriers, we extract projected CLIP ViT-L/14-336 image embeddings and fit class-balanced logistic-regression probes with feature standardization.

For Face→Context transfer, probes are trained on stratified 80% subsets of face-driven images and evaluated both on the corresponding 20% held-out face subsets and, without adaptation, on the complete context-driven population. We additionally evaluate Context→Context using stratified 80/20 context-only splits. Both analyses are repeated using ten fixed random seeds; standardization is fitted only on the corresponding training subset.

4 Results

4.1 Facial-Expression-Associated Features

SAE features associated with facial expressions showed preferential activation on the corresponding face-driven FindingEmo populations. Joy feature 2163 was active in 98.7% of face-driven images, compared with 88.1% of context-driven and 87.5% of neutral-decision-factor images. Sadness feature 5284 showed the same ordering (78.8%, 66.8%, and 51.3%, respectively). Both features also showed greater activation magnitude and spatial extent in the face-driven condition (Appendix C.1).

Spatial localization was consistent with these aggregate patterns (Fig. 1): feature 2163 concentrated around smiling facial regions, whereas feature 5284 responded to facial regions in sad and crying examples. Neither feature was exclusive to face-driven images, consistent with FindingEmo decision factors describing the primary basis of the image-level judgment rather than the absence of other visual cues.

4.2 Context-Associated Features and Held-Out Evaluation

Cross-fitted discovery identified reproducible context-associated candidates for both emotions. Joy feature 7406 was independently recovered at ranks 12, 7, and 6 across the three discovery folds and localized to collective celebrations, sports victories, and other positive social events (Fig. 2). For sadness, feature 9960 was recovered at ranks 11, 11, and 10 and localized strongly to funeral, memorial,



Figure 2: Spatial localization of context-associated candidates. The top row shows joy feature 7406, which responds to collective celebrations and positive social events, while the bottom row shows sadness feature 9960, which responds strongly to funeral, memorial, and burial scenes. Red boxes mark the ten strongest positive patch activations.

Table 1: Held-out context selectivity of the selected SAE features. Confidence intervals are 95% bootstrap intervals for the selectivity score within each held-out fold.

Emotion	Feature	Fold	Δ_C	S	95% CI
Joy	7406	1	0.0286	0.0120	[-0.0026, 0.0275]
		2	0.0271	0.0104	[-0.0026, 0.0249]
		3	0.0268	0.0102	[-0.0048, 0.0282]
Sadness	9960	1	0.0464	0.0328	[0.0138, 0.0516]
		2	0.0567	0.0430	[0.0251, 0.0594]
		3	0.0401	0.0264	[0.0098, 0.0415]

and burial scenes. Although other sadness features produced larger absolute logit reductions, their activation patterns were broad or face-associated rather than context-associated. Candidate selection was completed before examining held-out context-selectivity results.

On held-out images, joy feature 7406 showed positive context selectivity in all three folds, with $S = 0.0120$, 0.0104, and 0.0102 (Table 1). Its mean context-driven causal effect across folds was 0.0275, compared with 0.0166 for face-driven and 0.0069 for neutral images. However, the 95% bootstrap confidence interval crossed zero in every fold. Thus, feature 7406 showed consistent contextual enrichment, but its held-out context selectivity remained statistically uncertain.

Sadness feature 9960 showed stronger held-out evidence. Its context-selectivity scores were 0.0328, 0.0430, and 0.0264, with 95% bootstrap confidence intervals excluding zero in all three folds. Its mean context-driven causal effect was 0.0477, compared with 0.0137 for face-driven and 0.0104 for neutral images. These results provide consistent held-out evidence that feature 9960 contributes more strongly to sadness outputs for context-driven images than for either comparison condition.

4.3 Control Analyses

Both selected features showed context-enriched activation prevalence, spatial extent, and magnitude; complete activation statistics are reported in Appendix C.2.

We next compared their full-population context-selectivity scores with 200 uniformly sampled SAE features (Fig. 3). Joy feature 7406 obtained $S = 0.01086$ and sadness feature 9960 $S = 0.03406$; neither was equaled or exceeded by any random feature. Both therefore ranked 1/201, corresponding to an add-one empirical upper-tail probability of $p = 0.00498$. This feature-level comparison is distinct from the held-out bootstrap analysis: it evaluates whether the selected feature is unusual relative to arbitrary SAE features rather than estimating uncertainty in selectivity across images.

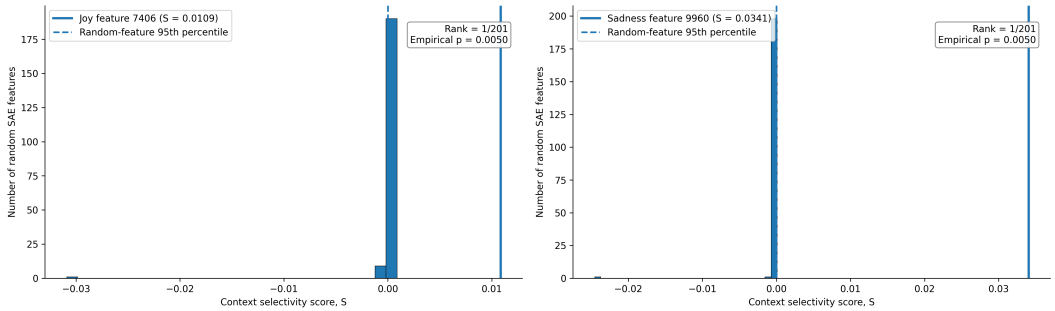

Figure 3: Random-feature control analysis. Joy feature 7406 and sadness feature 9960 each exceeded all 200 randomly sampled SAE controls in context selectivity (rank 1/201; $p = 0.00498$).

Table 2: Linear probing across ten stratified splits (mean $\pm$ sample SD). Face$\rightarrow$Context evaluates face-trained probes on the complete context population without adaptation.

Evaluation	Accuracy	Balanced Acc.	AUROC	F1
Face $\rightarrow$ Face	0.907 ± 0.014	0.852 ± 0.037	0.951 ± 0.013	0.942 ± 0.009
Face $\rightarrow$ Context	0.852 ± 0.010	0.851 ± 0.011	0.930 ± 0.006	0.867 ± 0.009
Context $\rightarrow$ Context	0.880 ± 0.033	0.879 ± 0.036	0.942 ± 0.022	0.894 ± 0.028

For the stricter activation-matched analysis, only three SAE features satisfied all pre-specified criteria for joy feature 7406. Their selectivity scores were 0.00068, -0.00013, and -0.00198, compared with 0.01086 for feature 7406. The target therefore ranked 1/4 ($p = 0.25$); given only three eligible controls, we treat this comparison as descriptive rather than inferential. No feature satisfied all matching criteria for sadness feature 9960, so no activation-matched comparison was performed and the criteria were not relaxed.

4.4 Cross-Carrier Transfer of Emotion Information

To test whether emotion information remains accessible when the visual carrier changes, we trained linear probes exclusively on face-driven images and evaluated them without adaptation on context-driven images (Table 2). Face$\rightarrow$Face evaluation achieved balanced accuracy of 0.852 ± 0.037 and AUROC of 0.951 ± 0.013. Under Face$\rightarrow$Context transfer, balanced accuracy remained 0.851 ± 0.011 and AUROC was 0.930 ± 0.006.

Context$\rightarrow$Context probing yielded balanced accuracy of 0.879 ± 0.036 and AUROC of 0.942 ± 0.022. Thus, joy–sadness information is linearly accessible within context-driven images and a decision boundary learned exclusively from face-driven images transfers strongly to context-driven images.

These results do not establish identical mechanisms across carriers or rule out scene-level confounds. In particular, joy and sadness context images may differ in properties such as illumination or scene composition. Face$\rightarrow$Context transfer provides complementary evidence because the classifier is learned exclusively from face-driven images, but correlated low-level or scene-level properties cannot be fully excluded.

5 Discussion

5.1 Sparse and Cross-Carrier Emotion Evidence

Our results support a distinction between context-associated sparse features and emotion information that remains accessible across visual carriers. Cross-fitted discovery identified reproducible candidates for contextual joy and sadness, but with different evidential profiles. Sadness feature 9960 showed the strongest evidence: its held-out context-selectivity confidence intervals excluded zero in all three folds, and it ranked above all 200 random-feature controls. Joy feature 7406 was

also reproducibly discovered, context-enriched, and ranked above all random controls and the three eligible activation-matched controls, but its held-out confidence intervals crossed zero. We therefore interpret 7406 as a context-associated joy candidate with convergent but statistically uncertain evidence of preferential causal effect.

These analyses provide complementary rather than interchangeable evidence. The held-out bootstrap quantifies uncertainty in context selectivity across images, whereas the random-feature comparison asks whether the observed selectivity is unusual relative to arbitrary SAE features. The latter therefore does not substitute for the former, particularly for feature 7406.

The feature analyses also distinguish activation and localization from causal function. Facial-expression-associated features 2163 and 5284 preferentially responded to facial regions, whereas 7406 and 9960 localized to positive social events and funeral or memorial scenes, respectively. Such patterns describe where and to what visual content a feature responds, but do not alone establish context-selective causality. Indeed, some features encountered during discovery produced larger absolute logit effects without preferential context association. Combining cross-fitted discovery, spatial characterization, and held-out ablation therefore provides a more stringent basis for interpreting sparse features.

At the broader representation level, a linear boundary learned exclusively from face-driven images transferred strongly to context-driven images, while joy and sadness were also linearly separable within context-driven images. Thus, context-associated sparse-feature behavior can coexist with emotion-discriminative information that remains accessible when the visual carrier changes. This does not imply identical mechanisms across carriers; rather, it suggests that carrier-associated sparse structure and broader cross-carrier emotion information need not be mutually exclusive.

5.2 Limitations and Implications

The distinction between sparse feature behavior and cross-carrier accessibility may help characterize whether similar emotion-related outputs are supported by similar or different visual evidence. This is particularly relevant for contextual cues—including social interactions, events, objects, clothing, and scene structure—whose interpretation may vary across datasets and populations. Our experiments do not evaluate demographic or cultural bias, and the identified features should not be interpreted as universal markers of emotion.

Several limitations constrain the conclusions. We examine only joy and sadness, one CLIP architecture, one intermediate layer, and one pretrained SAE decomposition. FindingEmo decision factors identify the primary basis of image-level judgments, but individual images may contain multiple visual cues; Face, Context, and Neutral therefore do not imply the absence of other cue types within an image.

The causal estimates are also conditional on the learned SAE decomposition: interventions operate on SAE-reconstructed activations, with the unablated reconstruction serving as the baseline to control for reconstruction effects. The two context-associated candidates additionally differ in statistical support: 9960 showed held-out confidence intervals excluding zero, whereas 7406 did not. Only three features met the pre-specified activation-matching criteria for 7406, and none did for 9960, limiting the matched-control analysis without post hoc relaxation of those criteria.

Finally, cross-carrier probing establishes linear accessibility rather than mechanistic equivalence. Context-only separability and Face→Context transfer do not rule out correlated scene-level properties such as illumination or composition. Broader validation across models, layers, emotions, datasets, and cultural settings is therefore needed to determine how stable these sparse and cross-carrier patterns are.

6 Conclusion

This work investigated how CLIP represents emotion when the same emotional concept is conveyed through different visual sources. Combining sparse autoencoder analysis, spatial localization, causal ablation, and cross-carrier linear probing, we found complementary evidence at sparse and dense representational levels. Sadness feature 9960 showed statistically supported held-out context selectivity, whereas joy feature 7406 showed reproducible contextual enrichment and positive held-

out selectivity but greater statistical uncertainty. Both were unusually context-selective relative to random SAE features.

At the same time, emotion-discriminative information remained linearly accessible across visual carriers: a boundary learned exclusively from face-driven images transferred strongly to context-driven images. These findings suggest that context-associated sparse-feature behavior can coexist with broader emotion information accessible across facial and contextual evidence.

More broadly, identifying an emotion-related prediction is not equivalent to identifying the visual evidence underlying it. Activation, spatial localization, causal contribution, and visual-carrier association provide distinct forms of evidence. Their joint analysis offers a framework for investigating both what emotion-related information is represented and which visual evidence contributes to model behavior.

Reproducibility and Code Availability

The FindingEmo resources used in this work are publicly available under the CC BY-NC-SA 4.0 license Mertens et al. (2024). The license applies to the annotations and source-image URL list; underlying images are not redistributed by the dataset authors and are obtained from the referenced URLs. We use the publicly released pretrained Matryoshka BatchTopK SAE of Zaigrajew et al. (2025) with CLIP ViT-L/14-336 layer-22 post-MLP residual activations. The exact SAE checkpoint and loading configuration are provided in the accompanying repository.

Context-feature evaluation uses three-fold cross-fitting and 5,000 nonparametric bootstrap resamples per held-out fold. Random-feature controls use 200 uniformly sampled SAE features. Linear-probe experiments use ten fixed seeds $\{42, 44, 123, 256, 512, 1024, 2024, 2025, 2026, 7777\}$ with stratified 80/20 splits, fitting feature standardization only on the corresponding training subset. Additional dataset-construction and analysis details are provided in the Methodology and Appendix.

Code for the complete analysis pipeline is available in an anonymized repository for review at `https://anonymous.4open.science/r/context-or-face-D43A/`. The repository includes dataset construction, CLIP activation extraction, SAE feature discovery and characterization, causal ablation, control analyses, spatial localization, and linear probing, together with experiment configurations, random seeds, software dependencies, and implementation details required to reproduce the reported analyses. Following double-blind review, the code will be released through the corresponding public repository.

AI-Use Disclosure

Generative AI tools were used to assist with drafting and editing manuscript text, improving readability, and suggesting section organization. They were also used to generate the two illustrative images in the qualitative conflicting-cue analysis (Appendix F); these images are explicitly identified as generated illustrations and are not experimental data or FindingEmo samples. Generative AI was not used to generate experimental data, formulate mathematical claims, design experiments, implement methods or code, or interpret experimental results. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final manuscript.

References

Sree Bhattacharyya and James Z. Wang. Evaluating vision-language models for emotion recognition. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1798–1820. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.97.

Yasaman Etesam, Özge Nilay Yalçın, Chuxuan Zhang, and Angelica Lim. Contextual emotion recognition using large vision language models. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4769–4776, 2024.

Jack Gallifant, Shan Chen, Kuleen Sasse, Hugo Aerts, Thomas Hartvigsen, and Danielle Bitterman. Sparse autoencoder features for classifications and transferability. In *Proceedings of the 2025*

Conference on Empirical Methods in Natural Language Processing, pp. 29939–29963. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1521.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, 2025.

Yu Gao, Weihong Ren, Xinglong Xu, Yan Wang, Zhiyong Wang, and Honghai Liu. MLPER: Multi-level prompts for adaptively enhancing vision-language emotion recognition. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 12640–12647, 2024. doi: 10.1109/IROS58592.2024.10801375.

Agam Goyal, Vedant Rathi, William Yeh, Yian Wang, Yuen Chen, and Hari Sundaram. Breaking bad tokens: Detoxification of llms using sparse autoencoders. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12691–12709. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.641.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024.

Dmitrii Kharlapenko, Stepan Shabalin, Arthur Conmy, and Neel Nanda. Scaling sparse feature circuits for studying in-context learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29819–29851. PMLR, 2025.

Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. EMOTIC: Emotions in context dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2309–2317, 2017. doi: 10.1109/CVPRW.2017.285.

Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10143–10152, 2019.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36993–37014, 2025.

Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. In *International Conference on Learning Representations*, 2025.

Samuel Marks, Can Rager, Eric Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*, 2025.

Laurent Mertens, Elahe' Yargholi, Hans Op de Beeck, Jan Van den Stock, and Joost Vennekens. Findingemo: An image dataset for emotion recognition in the wild. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. EmotiCon: Context-aware multimodal emotion recognition using frege's principle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14234–14243, 2020.

Rajiv Movva, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. Sparse autoencoders for hypothesis generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 44997–45023. PMLR, 2025.

Matthew Lyle Olson, Musashi Hinck, Neale Ratzlaff, Changbai Li, Phillip Howard, Vasudev Lal, and Shao-Yen Tseng. Probing the representational power of sparse autoencoders in vision models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 6226–6236, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Yusung Ro, Jaehyun Choi, and Junmo Kim. Beyond semantics: Disentangling information scope in sparse autoencoders for CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 3252–3261, 2026.

Xuli Shen, Hua Cai, Weilin Shen, Qing Xu, Dingding Yu, Weifeng Ge, and Xiangyang Xue. CocoER: Aligning multi-level feature by competition and coordination for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29591–29600, 2025.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model. *Transformer Circuits Thread*, 2026. URL `https://transformer-circuits.pub/2026/emotions/index.html`.

Ala N. Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. Mechanistic interpretability of emotion inference in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13090–13120, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.679.

Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, and Wen-Huang Cheng. EmoVIT: Revolutionizing emotion insights with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26596–26605, 2024.

Toshinori Yamauchi, Hiroshi Kera, and Kazuhiko Kawamoto. Zero-shot textual explanations via translating decision-critical features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 3282–3292, 2026.

Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, and Lihua Zhang. Context de-confounded emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19005–19015, 2023.

Dingkang Yang, Kun Yang, Mingcheng Li, Shunli Wang, Shuaibing Wang, and Lihua Zhang. Robust emotion recognition in context debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12447–12457, 2024.

Vladimir Zaigrajew, Hubert Baniecki, and Przemyslaw Biecek. Interpreting CLIP with hierarchical sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 73918–73956. PMLR, 2025.

Sitao Zhang, Yimu Pan, and James Z. Wang. Learning emotion representations from verbal and nonverbal communication. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18993–19004, 2023.

A Appendix

This appendix provides dataset construction details, activation statistics, and additional analyses supporting the main text.

B Dataset Construction

FindingEmo Mertens et al. (2024) provides both emotion labels and decision-factor annotations describing the visual evidence contributing to human emotion judgments. We focus on joy and sadness. The joy group comprises *Serenity*, *Joy*, and *Ecstasy*, while the sadness group comprises *Pensiveness*, *Sadness*, and *Grief*.

Images labeled *FacialExpression* form the face-driven condition, those labeled *Context* form the context-driven condition, and those labeled *Neutral* are retained as a comparison condition. *Neutral* here refers to the FindingEmo decision factor rather than the emotional label.

The same emotion and decision-factor definitions are used throughout the SAE and probing analyses. Images are required to have a resolvable local path and to be readable by the image-processing pipeline. The resulting joy populations contain 236 context-driven, 700 face-driven, and 160 neutral-decision-factor images. The sadness populations contain 187 context-driven, 170 face-driven, and 39 neutral-decision-factor images. The three conditions are mutually exclusive within each emotion analysis.

C Sparse-Feature Activation Statistics

C.1 Facial-Expression-Associated Features

Table 3 reports activation prevalence, spatial extent, and activation magnitude for facial-expression-associated features 2163 (joy) and 5284 (sadness). Both features showed their strongest aggregate activation in the corresponding face-driven populations, while remaining active in subsets of context-driven and neutral-decision-factor images.

Table 3: Activation statistics for facial-expression-associated SAE features across FindingEmo decision-factor populations. Image activation denotes the percentage of images containing at least one positive feature activation; active patches denotes the mean number of patches with positive activation per image; maximum activation is averaged across images.

Feature	Condition	Image (%)	Active Patches	Max Act.
2163 (Joy)	Context	88.14	9.56	4.96
	Face	98.71	20.09	9.93
	Neutral	87.50	9.30	5.52
5284 (Sadness)	Context	66.84	3.53	2.00
	Face	78.82	13.05	4.11
	Neutral	51.28	2.51	1.49

C.2 Context-Associated Features

Table 4 reports the corresponding activation statistics for context-associated features 7406 (joy) and 9960 (sadness). Both candidates showed greater activation prevalence, spatial extent, and magnitude in context-driven images than in their respective face-driven and neutral-decision-factor populations.

Table 4: Activation statistics for context-associated SAE features across FindingEmo decision-factor populations.

Feature	Condition	Image (%)	Patch (%)	Max Act.
7406 (Joy)	Context	46.19	1.57	3.65
	Face	40.43	0.98	2.91
	Neutral	30.00	0.68	2.34
9960 (Sadness)	Context	79.14	7.04	4.95
	Face	67.06	4.11	2.80
	Neutral	56.41	1.88	2.19

D Additional Spatial Localization

For qualitative localization, feature activations are mapped from CLIP's 24×24 spatial patch grid back to image space. For each image, patches are ranked by positive activation magnitude and the ten strongest active patches are highlighted. Image-level examples are ranked using their feature activation statistics to identify informative qualitative cases.

These visualizations are used only to characterize the spatial and semantic patterns associated with individual SAE features. They are not used as evidence of causal importance; causal conclusions are based on the reconstruction-controlled ablation analyses described in the main paper.

E Activation-Matched Control Analysis

For joy feature 7406, three SAE features satisfied all pre-specified activation-matching criteria. Their context-selectivity scores were 0.000675, -0.000133, and -0.001980, compared with 0.010863 for feature 7406. Feature 7406 therefore ranked first among the four features, corresponding to an add-one empirical upper-tail probability of $p = 0.25$. Because only three eligible controls were available, this comparison is treated as descriptive rather than inferential.

For sadness feature 9960, no SAE feature satisfied all pre-specified matching criteria. We therefore did not perform an activation-matched comparison and did not relax the matching criteria after observing the number of eligible features.

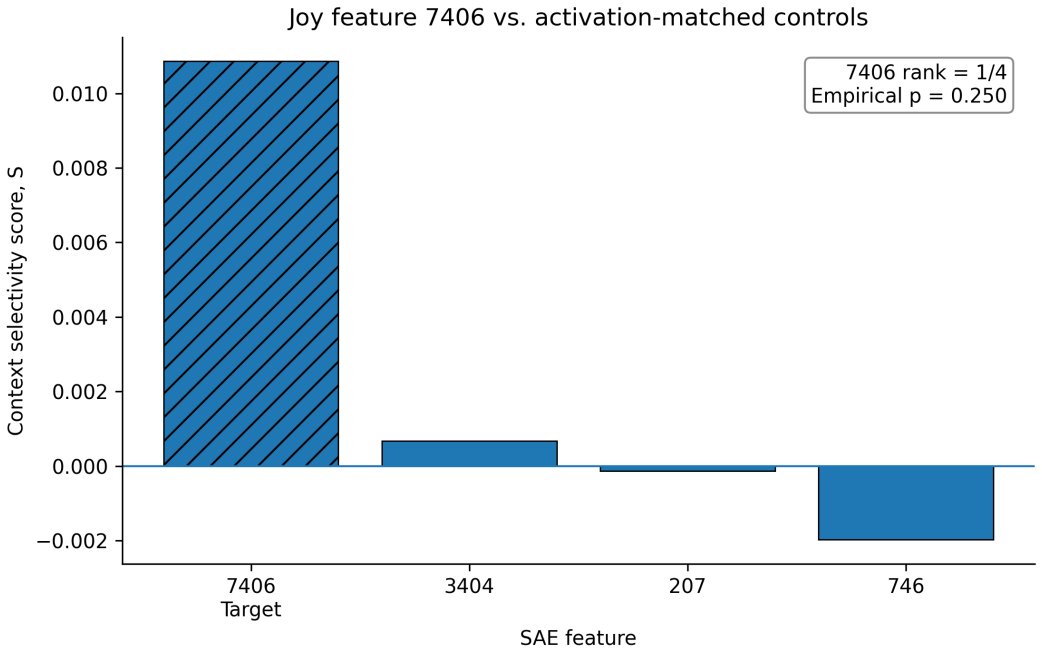

Figure 4: Activation-matched control analysis for joy feature 7406. Only three SAE features satisfied all pre-specified matching criteria. Feature 7406 showed greater context selectivity than all three eligible controls, but the small control set makes this comparison descriptive rather than inferential.

F Qualitative Conflicting-Cue Experiment

We additionally examined two illustrative scenarios containing potentially conflicting facial and contextual emotion cues (Fig. 5). These examples probe whether interventions on distinct sparse features can shift CLIP's emotion evidence in different directions when facial and contextual cues coexist.



Figure 5: Illustrative conflicting-cue scenarios: (left) positive contextual evidence with facial distress and (right) a positive facial cue with distressing social context. Images were generated using ChatGPT (OpenAI) for illustration and are not experimental outputs or FindingEmo samples.

In the sports scenario, ablating joy-context feature 7406 decreased both the emotion and cue margins (-0.160 and -0.170), whereas ablating sad-face-associated feature 5284 shifted them in the opposite direction ($+0.079$ and $+0.213$). Thus, interventions on distinct sparse features shifted CLIP's evidence in opposing directions within the same image.

In the airport scenario, face-associated feature 2163 produced the largest changes (-0.372 in the emotion margin and -0.461 in the cue margin), while feature 7406 had a smaller effect and sadness-context candidate 9960 produced essentially no change.

Table 5: Feature-ablation effects in the conflicting-cue analyses. Values are changes relative to the unablated SAE reconstruction.

Scenario	**Feature**	**Role**	ΔE	ΔC
Sports	7406	Joy context	-0.160	-0.170
	9960	Sad-context cand.	$+0.030$	-0.000
	2163	Facial	-0.260	-0.088
	5284	Facial	$+0.079$	$+0.213$
Airport	7406	Joy context	-0.143	-0.026
	9960	Sad-context cand.	0.000	0.000
	2163	Facial	-0.372	-0.461
	5284	Facial	$+0.024$	$+0.056$

These image-dependent examples illustrate that multiple emotion-related cues can coexist within a representation and that feature interventions can produce distinct shifts within the same image. They are qualitative probes, however, and do not establish that facial or contextual information generally dominates CLIP's emotion judgments.