EmoGist: Efficient In-Context Learning for Visual Emotion Understanding

Ronald Seoh and Dan Goldwasser
Purdue University
{bseoh, dgoldwas}@purdue.edu

Abstract

In this paper, we introduce EmoGist, a training-free, in-context learning method for performing visual emotion classification with LVLMs. The key intuition of our approach is that context-dependent definition of emotion labels could allow more accurate predictions of emotions, as the ways in which emotions manifest within images are highly context dependent and nuanced. EmoGist pre-generates multiple descriptions of emotion labels, by analyzing the clusters of example images belonging to each label. At test time, we retrieve a version of description based on the cosine similarity of test image to cluster centroids, and feed it together with the test image to a fast LVLM for classification. Through our experiments, we show that EmoGist allows up to 12 points improvement in micro F1 scores with the multi-label Memotion dataset, and up to 8 points in macro F1 in the multi-class FI dataset.

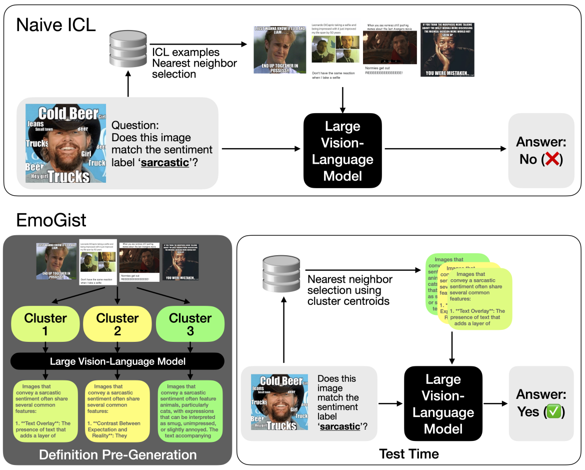

Figure 1: For visual emotion classification, naive in-context learning (ICL) struggles as providing multiple nuanced visual examples often lead LVLMs to make incorrect predictions. EmoGist guides LVLMs using pre-generated multiple descriptions of emotion labels obtained by analyzing the clusters of example images.

1 Introduction

Automated classification of visual emotion (Ekman, 1993; Lang et al., 1999; Mikels et al., 2005) is an extremely challenging problem, as the ways in which emotions are embedded within images are inherently nuanced. Hence, even large vision-language models (LVLMs) that are extensively trained for reasoning over visual inputs struggle in detecting these emotions (Bhattacharyya and Wang, 2025), as their training may not necessarily involve the ability to understand such nuanced patterns.

In this paper, we introduce EmoGist, a training-free, in-context learning method for performing visual emotion classification with LVLMs. The key intuition of our approach is that the real meaning of different emotion labels could be dependent on the image's context. For example, we could intuitively imagine that the way the emotion of 'excitement' for sporting events could be significantly different from the 'excitement' of the academics for an upcoming conference. Hence, guiding LVLMs with such context dependent definition of emotion labels could allow the models to better focus on the nuanced patterns of the image.

EmoGist automatically pre-generates nuanced, context-specific descriptions of emotion labels, by analyzing the clusters of example images belonging to each label. At test time, we retrieve a version of description based on the cosine similarity of test image to cluster centroids, and feed it together with the test image to a fast LVLM for classification. Through our experiments, we show that EmoGist allows up to 12 points improvement in micro F1 scores with the multi-label emotion classification, and up to 8 points improvement in macro F1 for the multi-class case. We also demonstrate that EmoGist could achieve improvements in smaller LVLMs with 2 billion parameters.[1]

[1]All the program codes used to produce results presented in this paper are available at `https://tinyurl.com/emo-gist`.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 2171–2182
November 4-9, 2025

2 Related Work

While visual emotions has been extensively studied in the many related fields including computer vision and psychology (Mikels et al., 2005; Machajdik and Hanbury, 2010; Peng et al., 2015; You et al., 2016; Yang et al., 2023), we are only starting to see the efforts to exploit large vision-language models for automated understanding of visual emotions (Xie et al., 2024; Etesam et al., 2024; Xenos et al., 2024; Lei et al., 2025; Bhattacharyya and Wang, 2025).

Visual in-context learning (ICL) has seen considerable amount of interest in recent literature (Zhang et al., 2023; Zhou et al., 2024; Zhang et al., 2024), where many work have investigated effective strategies for choosing visual ICL examples given the test instance. However, we believe that our work is first to investigate ICL strategies with LVLMs in detail for evoked emotion classification.

3 EmoGist

We describe the major components of EmoGist, our in-context learning method with LVLMs for emotion classification. Instead of naively retrieving individual examples based on the visual similarity of the image, the key idea is to obtain an nuanced description of emotion labels, which could effectively serve as the decision boundary for the LVLM to make its predictions on.

Because the same emotion could manifest in many different ways for across different images, we develop a strategy where we utilize stronger LVLMs to pre-generate multiple descriptions of different emotion labels, by analyzing the clusters of example images belonging to each category.

Embedding and storing the pool of emotion label examples In order to generate multiple descriptions of emotion labels, we begin by embedding the pool of example images with an embedding model. We use the MM-E5 model, the state-of-the-art multimodal embedding model by Chen et al. (2025). Then we store the embeddings into a HNSWLIB vector database (Malkov and Yashunin, 2018).

Clustering After creating the vector database of example images, we run the k-means clustering algorithm (Lloyd, 1982) against the set of embeddings to get different clusters. Because we are interested in creating multiple versions of descriptions for specific labels, clustering is done separately for each emotion label. We tune the hyperparameter k by setting a portion of example images aside as a validation set, and tune by evaluating the end task performance on them. Due to the limited computational resources available, we only experiment with the k values of 2, 4, and 6.[2]

Generating label descriptions With the cluster information, we provide a strong LVLM with images from each cluster, and prompt them to explain why the given images belong to the emotion label. For our experiments, we use the Qwen2.5-VL 72B model (Bai et al., 2025) for generating descriptions. As it is not possible to provide the LVLM with all images from the cluster due to GPU memory limits and context length, we select 4 images from each cluster to create one version of label description. Figure 2 shows the prompt used and examples of generated descriptions.

Selection of clusters at test time We note that EmoGist addresses both multi-label and multi-class classification cases. For multi-label classification, we assume that all test instances are binarized, and perform predictions for each candidate label given the image. For each candidate label, we retrieve the closest cluster among the candidates with the corresponding label.

For multi-class classification where the classes are exclusive to each other, we perform the classification only once by providing the list of all candidate classes to the model. We perform the search across the entire set of clusters regardless of their classes, and use the closest cluster to the test image in terms of its distance to the centroid.

Once the cluster and associated label description has been chosen, we prepend the description to the test image and classification prompt.[3]

Ensembles As we only select a subset of images from each cluster for generating an description, it may be the case that the selected images and descriptions may not sufficiently match the test image. In order to mitigate this issue, we introduce a simple ensemble scheme, where we generate multiple versions of descriptions for each cluster, perform multiple predictions against a single example and take the majority vote. We generate multiple descriptions for the cluster by ranking all images within the cluster by their distance to the centroid,

[2]Please see Appendix C for hyperparameter tuning procedures and sensitivity analysis.

[3]Please see Appendix E for the full prompts used for classification.

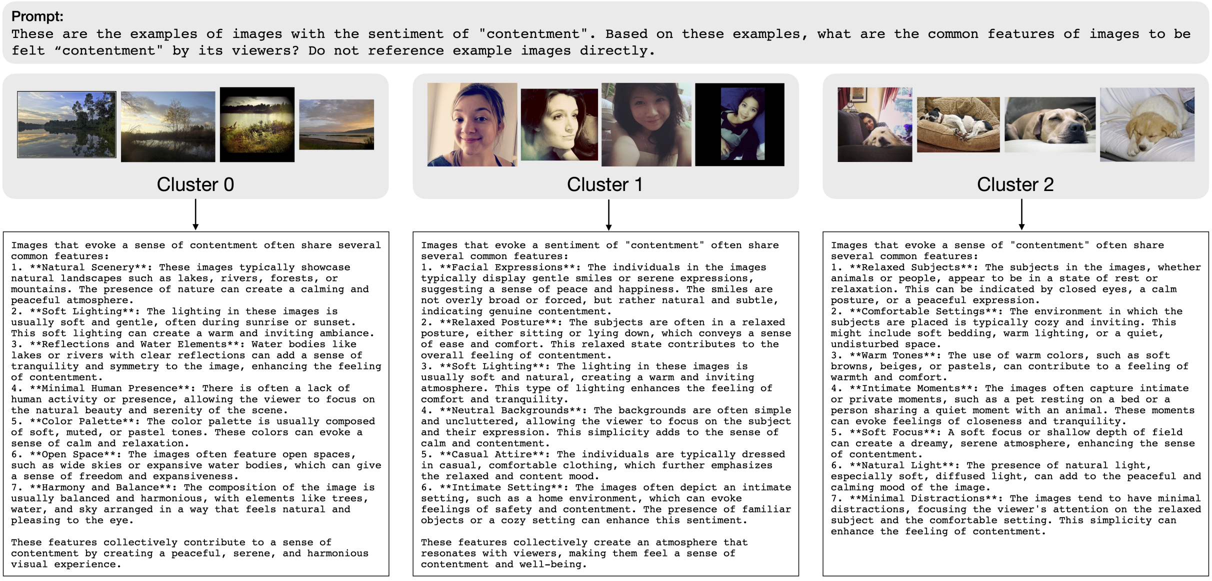

Figure 2: Example label descriptions for clusters within the `contentment` class of the FI dataset, as generated by the Qwen2.5-VL 72B model.

and generate descriptions for every top 4 images in the ranked list.

4 Experiments

We used two datasets: one is the Memotion 1.0 dataset (Sharma et al., 2020; Jin et al., 2024), a collection of meme images from social media where we perform a multi-label classification across 4 labels: `sarcastic`, `humorous`, `offensive`, and `motivational`. The second dataset is the FI dataset (You et al., 2016), a collection of everyday images across the internet tagged based on the Ekman model (Ekman, 1993) of 8 classes: `amusement`, `anger`, `awe`, `contentment`, `disgust`, `excitement`, `fear`, and `sadness`.[4]

4.1 Baselines

In order to save computational resources and make our discussions more clear, we choose three different SOTA LVLMs of similar sizes for our experiments: Qwen2.5-VL 7B (Bai et al., 2025), Aya Vision 8B (Dash et al., 2025), and InternVL2.5 8B-MPO (Wang et al., 2024). To ensure that our findings are not tied to particular pretraining, these three VLMs were chosen to ensure that they do not share the same image encoder or backbone LLM. We additionally provide a subset of results for smaller LVLMs in Table 2.

We compare EmoGist with the following comparable in-context learning methods:

- Zero-Shot: We simply prompt a LVLM with the test image and prediction prompt.
- Global Exp: Instead of providing example images for a description, we prompt a large LVLM for “global" description, where we ask the model to describe the common features of the images with the candidate emotion label, without providing any references.
- ICL_{sim}: We retrieve 4 images from the pool of images based on cosine similarity, regardless of their labels. This is closest to $\text{EmoGist}_{\text{n}}$, in terms of the number of examples.
- ICL_{all}: We also test another case of performing ICL for multiclass classification, where we provide one image each for all classes. Note that the FI dataset have 8 classes, resulting in 8 example images to be provided to a LVLM.

We also test two variants of EmoGist:

- $\text{EmoGist}_{\text{n}}$: This variant of EmoGist uses 4 images from the cluster for description generation.
- $\text{EmoGist}_{\text{e}}$: This variant of EmoGist performs ensembling, where we generate 3 versions of label description, each of them using 4 images without any overlap between them.

4.2 Results

EmoGist achieves robust performance gains over all the baselines. In Table 1, we can see that both $\text{EmoGist}_{\text{n}}$ and $\text{EmoGist}_{\text{e}}$ achieve consistent improvements over all the baselines. For FI,

[4]Please see Appendix A for more detailed dataset statistics.

Method	Model: Qwen2.5-VL 7B Precision	Qwen2.5-VL 7B Recall	Qwen2.5-VL 7B F1	Aya Vision 8B Precision	Aya Vision 8B Recall	Aya Vision 8B F1	InternVL2.5 8B-MPO Precision	InternVL2.5 8B-MPO Recall	InternVL2.5 8B-MPO F1
Zero-Shot	47.121 $\pm$0.065	50.381 $\pm$0.090	40.089 $\pm$0.065	50.094 $\pm$0.285	47.181 $\pm$0.188	44.188 $\pm$0.102	47.237 $\pm$0.297	50.379 $\pm$0.171	42.939 $\pm$0.223
Global Exp	30.503 $\pm$0.325 -16.619	30.156 $\pm$0.439 -20.226	23.480 $\pm$0.404 -16.609	26.416 $\pm$1.670 -23.678	20.782 $\pm$0.468 -26.398	15.405 $\pm$0.322 -28.783	32.242 $\pm$0.316 -14.996	35.224 $\pm$0.452 -15.155	30.583 $\pm$0.312 -12.355
ICL_{sim}	48.656 $\pm$0.029 +1.534	49.898 $\pm$0.022 -0.483	42.733 $\pm$0.016 +2.643	48.649 $\pm$0.592 -1.445	42.444 $\pm$0.122 -4.736	35.257 $\pm$0.185 -8.931	42.613 $\pm$0.215 -4.624	43.658 $\pm$0.171 -6.721	37.653 $\pm$0.161 -5.286
ICL_{all}	23.464 $\pm$0.948 -23.657	14.633 $\pm$0.024 -35.749	5.664 $\pm$0.081 -34.425	13.223 $\pm$3.242 -36.871	12.609 $\pm$0.042 -34.572	1.511 $\pm$0.102 -42.677	16.682 $\pm$0.517 -30.555	16.107 $\pm$0.259 -34.272	13.454 $\pm$0.298 -29.485
$\text{EmoGist}_{\text{n}}$	**52.944** $\pm$0.364 +5.822	**52.163** $\pm$0.265 +1.782	**48.497** $\pm$0.572 +8.408	52.579 $\pm$0.381 +2.485	**51.389** $\pm$0.271 +4.208	**47.906** $\pm$0.565 +3.718	**52.592** $\pm$0.261 +5.354	**51.767** $\pm$0.220 +1.388	**48.094** $\pm$0.303 +5.155
$\text{EmoGist}_{\text{e}}$	52.772 $\pm$0.377 +5.650	51.704 $\pm$0.235 +1.323	48.118 $\pm$0.552 +8.029	**52.632** $\pm$0.371 +2.538	51.146 $\pm$0.316 +3.965	47.579 $\pm$0.599 +3.391	52.449 $\pm$0.158 +5.211	51.712 $\pm$0.202 +1.333	47.795 $\pm$0.167 +4.856

(a) FI

Method	Qwen2.5-VL 7B Precision	Qwen2.5-VL 7B Recall	Qwen2.5-VL 7B F1	Aya Vision 8B Precision	Aya Vision 8B Recall	Aya Vision 8B F1	InternVL2.5 8B-MPO Precision	InternVL2.5 8B-MPO Recall	InternVL2.5 8B-MPO F1
Zero-Shot	77.343 $\pm$0.013	48.260 $\pm$0.104	59.434 $\pm$0.080	75.799 $\pm$0.023	63.814 $\pm$0.065	69.292 $\pm$0.035	73.520 $\pm$0.195	63.377 $\pm$0.358	68.070 $\pm$0.254
Global Exp	77.134 $\pm$0.180 -0.209	35.206 $\pm$2.182 -13.054	48.159 $\pm$1.976 -11.276	73.430 $\pm$0.702 -2.369	68.159 $\pm$0.654 +4.345	70.665 $\pm$0.116 +1.373	74.885 $\pm$0.191 +1.365	62.160 $\pm$0.360 -1.217	67.928 $\pm$0.234 -0.142
ICL_{sim}	75.061 $\pm$0.025 -2.282	64.642 $\pm$0.107 +16.381	69.462 $\pm$0.071 +10.028	**76.721** $\pm$0.043 +0.922	60.915 $\pm$0.077 -2.900	67.910 $\pm$0.062 -1.382	72.601 $\pm$0.101 -0.918	55.476 $\pm$0.432 -7.901	62.890 $\pm$0.290 -5.181
$\text{EmoGist}_{\text{n}}$	75.610 $\pm$0.183 -1.732	62.693 $\pm$0.492 +14.432	68.540 $\pm$0.237 +9.105	68.693 $\pm$0.548 -7.107	84.265 $\pm$1.958 +20.451	75.611 $\pm$0.747 +6.319	72.162 $\pm$0.204 -1.358	72.438 $\pm$0.708 +9.061	72.289 $\pm$0.338 +4.219
$\text{EmoGist}_{\text{e}}$	**78.682** $\pm$0.067 +1.339	**65.374** $\pm$0.315 +17.114	**71.411** $\pm$0.188 +11.977	70.263 $\pm$0.314 -5.536	**87.165** $\pm$0.647 +23.350	**77.795** $\pm$0.171 +8.502	**75.898** $\pm$0.257 +2.379	**74.596** $\pm$0.173 +11.219	**75.240** $\pm$0.170 +7.170

(b) Memotion

Table 1: Results of our methods and baselines. We report macro scores for FI and micro scores for Memotion. All scores are averaged over six random seeds. We show standard errors as their confidence intervals. Boldfaces indicate the best performance for each metric across all methods for each model. Green and red numbers indicate the performance changes over the zero-shot baseline.

$\text{EmoGist}_{\text{n}}$ gains 8.41 points in terms of macro F1 score over Zero-Shot, and 5.76 points over ICL_{sim}. The trend is largely similar for Memotion, where $\text{EmoGist}_{\text{e}}$ gains 11.98 points in terms of micro F1 score over Zero-Shot, and 1.95 points over ICL_{sim}.

It is interesting to note that Global Exp, which is essentially providing the strong LVLM's general knowledge about emotion labels, is considerably worse than the Zero-Shot baseline. Therefore, we could see that having EmoGist's localized, cluster-specific label description makes a substantial difference. Lastly, adding ensembling $\text{EmoGist}_{\text{e}}$ shows consistent improvements over $\text{EmoGist}_{\text{n}}$, with notable gains in precision for Memotion.

Naively performing visual ICL could be detrimental. Another observation from Table 1 is that our ICL baselines, ICL_{sim} and ICL_{all}, are either under-performing, or marginally better than the Zero-Shot baselines. In the most extreme case, ICL_{all} for FI sees over 42 points drop in F1 score, way below random guessing. In addition, while ICL achieves considerable performance with Qwen2.5-VL 7B on Memotion, Aya Vision 8B and InternVL 2.5 8B-MPO failed to achieve comparable scores. As reasoning over multiple image inputs is still an area of active research in pretraining and post-training for LVLMs (Li et al., 2024), it is likely the case that not all publicly available LVLMs are equal in terms of their ability to utilize ICL examples for classification.

Small LVLMs could also become decent emotion reasoners with EmoGist. As many practical uses of visual emotion understanding often take place within resource-constrained systems with low latency requirements such as web applications or personal computing devices, even relatively small

7 billion models may be beyond typical computing budget under such scenarios. In Table 2, we test 2 small LVLMs with the same number of 2 billion parameters, SmolVLM2 2.2B (Marafioti et al., 2025) and InternVL2.5 2B-MPO (Wang et al., 2024), to examine whether EmoGist performance benefits hold for these smaller models.

We could see that EmoGist$_e$ achieve similar levels of performance gains over the Zero-Shot and ICL baselines, with SmolVLM2.2 achieving performances similar to the 7B models in Table 1. Given that EmoGist only requires storing cluster centroids and text descriptions at test time, we believe that EmoGist shows some interesting future directions for implementing visual emotion understanding into a wide variety of applications.

Model	Method	FI			Memotion		
		Precision	Recall	F1	Precision	Recall	F1
InternVL2.5 2B-MPO	Zero-Shot	38.039 ±0.285	34.813 ±0.320	30.499 ±0.318	74.272 ±0.164	54.972 ±0.129	63.181 ±0.113
	ICL$_{sim}$	45.376 ±0.587 +7.337	11.215 ±0.221 -23.598	16.078 ±0.309 -14.421	64.674 ±0.104 -9.599	75.157 ±0.090 +20.184	69.522 ±0.074 +6.341
	EmoGist$_e$	50.706 ±0.412 +12.667	41.421 ±0.442 +6.608	37.442 ±0.675 +6.943	70.873 ±0.142 -3.400	60.658 ±0.840 +5.685	65.353 ±0.504 +2.173
SmolVLM2 2.2B	Zero-Shot	39.567 ±0.056	30.762 ±0.019	20.703 ±0.016	73.124 ±0.011	67.617 ±0.012	70.263 ±0.010
	ICL$_{sim}$	47.478 ±0.030 +7.912	48.463 ±0.029 +17.700	41.798 ±0.028 +21.096	74.896 ±0.097 +1.773	12.113 ±0.087 -55.505	20.852 ±0.130 -49.411
	EmoGist$_e$	52.358 ±0.542 +12.791	50.641 ±0.098 +19.879	46.713 ±0.445 +26.010	67.455 ±0.381 -5.668	96.311 ±0.540 +28.694	79.329 ±0.118 +9.066

Table 2: Results on small VLMs with 2B parameters.

Knowledge transfer across different domains with EmoGist. Lastly, we explore whether the knowledge about different emotion labels we acquire from example images could be used for predictions against the images from the domains different from example images. Using the label descriptions obtained from the example images of the FI dataset, we evaluate EmoGist$_e$ on the ArtPhoto dataset (Machajdik and Hanbury, 2010), a collection of artistically photographed images annotated with the same class labels as FI.

In Table 3, we can see that while EmoGist achieves slightly better scores over naive ICL, the overall performance is actually worse than the Zero-Shot baselines. As EmoGist captures more context-specific, nuanced knowledge of emotion labels, there seems to be a significant semantic gap between the images from FI and ArtPhoto that most of the FI clusters do not adequately explain the test images from ArtPhoto. Making EmoGist to capture both general and context-specific knowledge of emotions across different subject domains and visual compositions could be an interesting direction for future research.

Model	Method	ArtPhoto		
		Precision	Recall	F1
Qwen2.5-VL 7B	Zero-Shot	52.594	42.306	41.178
	ICL$_{sim}$	43.332	33.165	33.745
	EmoGist$_e$	46.535	38.555	39.227
Aya Vision 8B	Zero-Shot	55.602	43.706	43.473
	ICL$_{sim}$	43.328	26.902	26.742
	EmoGist$_e$	46.216	38.307	38.921
InternVL2.5 8B-MPO	Zero-Shot	48.824	42.365	43.518
	ICL$_{sim}$	39.606	32.083	33.018
	EmoGist$_e$	47.739	39.104	39.993

Table 3: Results on ArtPhoto with the emotion label descriptions from FI.

5 Conclusion and Future Work

In this paper, we introduced EmoGist, a training-free in-cotext learning method for visual emotion understanding with large vision-language models. We observe a significant amount of improvements over the zero-shot and naive ICL baselines across SOTA LVLMs of 2 and 7 billion parameters. In particular, we find that EmoGist$_e$, the variant of our method with simple ensembling, achieves robust performance improvements with higher precision. In future work, we'd like to explore more deeply into the label descriptions generated by the strong LVLMs and investigate various reasoning strategies for obtaining emotion label descriptions that are more transferrable across different domains.

Acknowledgments

The work was supported by NSF CAREER award IIS2048001 and the DARPA CCU program. Contents do not necessarily represent the official views of, nor an endorsement by, DARPA, or the US Government.

Limitations

Due to limited computational resources available to the authors, we perform our experiments on a limited subset of publicly available large vision-language models with 2 and 7 billion parameters. While we anticipate that our findings would hold overall for other model sizes, we do not provide any direct evidence. We also note that we only tried one model each for the embedding model and the description generation LVLM, as stated in section 3.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Sree Bhattacharyya and James Z. Wang. 2025. Evaluating vision-language models for emotion recognition. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1798–1820, Albuquerque, New Mexico. Association for Computational Linguistics.

Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. 2025. mme5: Improving multimodal multilingual embeddings via high-quality synthetic data. *Preprint*, arXiv:2502.08468.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, Jeremy Pekmez, Jason Ozuzu, Pierre Richemond, Acyr Locatelli, Nick Frosst, Phil Blunsom, Aidan Gomez, Ivan Zhang, Marzieh Fadaee, and 6 others. 2025. Aya vision: Advancing the frontier of multilingual multimodality. *Preprint*, arXiv:2505.08751.

Paul Ekman. 1993. Facial expression and emotion. *American psychologist*, 48(4):384.

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

Awni Hannun, Jagrit Digani, Angelos Katharopoulos, and Ronan Collobert. 2023. MLX: Efficient and flexible machine learning on apple silicon.

Yiqiao Jin, Minje Choi, Gaurav Verma, Jindong Wang, and Srijan Kumar. 2024. MM-SOC: Benchmarking multimodal large language models in social media platforms. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6192–6210, Bangkok, Thailand. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, and 1 others. 1999. International affective picture system (iaps): Instruction manual and affective ratings. *The center for research in psychophysiology, University of Florida*.

Yuxuan Lei, Dingkang Yang, Zhaoyu Chen, Jiawei Chen, Peng Zhai, and Lihua Zhang. 2025. Large vision-language models as emotion recognizers in context awareness. In *Proceedings of the 16th Asian Conference on Machine Learning*, volume 260 of *Proceedings of Machine Learning Research*, pages 111–126. PMLR.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

S. Lloyd. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.

Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 83–92, New York, NY, USA. Association for Computing Machinery.

Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2025. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*.

Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. 2005. Emotional category data on images from the international affective picture system. *Behavior research methods*, 37:626–630.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew Gallagher. 2015. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–868.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. 2024. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Alexandros Xenos, Niki Maria Foteinopoulou, Ioanna Ntinou, Ioannis Patras, and Georgios Tzimiropoulos. 2024. Vllms provide better context for emotion understanding through common sense reasoning. *Preprint*, arXiv:2404.07078.

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

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. 2023. Emoset: A large-scale visual emotion dataset with rich attributes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20326–20337.

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Jiahao Zhang, Bowen Wang, Liangzhi Li, Yuta Nakashima, and Hajime Nagahara. 2024. Instruct me more! random prompting for visual in-context learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2597–2606.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. 2023. What makes good examples for visual in-context learning? In *Advances in Neural Information Processing Systems*, volume 36, pages 17773–17794. Curran Associates, Inc.

Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. 2024. Visual in-context learning for large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15890–15902, Bangkok, Thailand. Association for Computational Linguistics.

A Dataset Information

For each dataset, we treat the training set as the pool of example images for description generation, use the validation set for hyperparameter tuning, and test our method and baselines against the test set.

A.1 Memotion

We use the split introduced by Jin et al. (2024) for our experiments.[5]. Dataset statistics are provided below:

Split	Number of unique images	sarcastic	humorous	offensive	motivational
Training	5593	4367	4259	3417	1972
Validation	699	538	539	437	253
Test	700	543	543	425	242

Table 4: Memotion dataset statistics.

A.2 FI

For the test set, we use the FI hard set introduced by (Bhattacharyya and Wang, 2025). As there is no separate training and validation set provided, we filter all the images includes in the FI hard set from the original FI set, and randomly split the set of remaining images into the training and validation set. Dataset statistics are provided below:

Split	amusement	anger	awe	contentment	disgust	excitement	fear	sadness
Training	3608	842	2698	4309	1391	2592	852	2634
Validation	190	45	142	259	74	137	45	139
Test	1125	368	293	188	192	185	149	128

Table 5: FI dataset statistics.

B Experimental Settings

B.1 Hardwares and Softwares Used

To fully utilize all the GPU resources available to us, we ran all of our validation and test predictions, and embedding generations using multiple NVIDIA V100, GTX TITAN Xp and GTX TITAN X GPUs.

We use the version 4.50.0 of HuggingFace Transformers library (Wolf et al., 2020), the version 0.7.3 of vLLM (Kwon et al., 2023) alongside PyTorch version 2.5.1 (Paszke et al., 2019).

For our label description generation, we used a Mac Studio hardware with M1 Ultra CPU and 128GB of RAM, using the version 0.25.0 of mlx (Hannun et al., 2023) and 0.1.23 of mlx-vlm.

[5]`https://huggingface.co/datasets/Ahren09/MMSoc_Memotion`

B.2 Random Seeds

For the results shown in Table 1 and Table 2, we run each test for 6 different random seeds: `21, 42, 63, 84, 105, 126`.

C Sensitivity analysis and hyperparameter tuning for the number of clusters k

We determine the optimal number of clusters for each dataset and test-time LVLM combination. Initially, we apply k-means clustering to both FI and Memotion datasets for k values of 2, 4, and 6. Subsequently, we generate cluster-based label descriptions for each k using the Qwen2.5-VL 72B model. We then classify the validation set based on these clusters and text descriptions, 3 times for each 7B/2B LVLM using the following random seeds: 21, 42, 63. The optimal k is ultimately identified by identifying k that achieved the most number of highest validation F1 scores across the three seeds. If there's no winner, we ran the validation for additional seeds (84, 105, 126) until the clear winner is found.

Please see Figure 9 and Figure 10 for the validation set results of EmoGist_n and EmoGist_e with 7B models on FI, Figure 11 and Figure 12 for the validation set results of EmoGist_n and EmoGist_e with 7B models on Memotion, and Figure 13 and Figure 14 for the validation set results of EmoGist_e with 2B models on FI and Memotion.

D Prompts used for label description generation

```
**Images from the cluster go here**
These are the examples of images with the sentiment of "{s_label}". Based on these examples, what are the common features of images to be felt "{s_label}" by its viewers? Do not reference example images directly.
```

Figure 3: Prompts used for label description generation.

E Prompts used for zero-shot and EmoGist

```
**Test image goes here**
**For EmoGist, label description goes here**
Question: Does this image match the sentiment label '{s_label}'? Answer:

Answer with 'Yes' or 'No'.
```

Figure 4: Prompts used for multi-label zero-shot classification and EmoGist.

Test image goes here
For EmoGist, label description goes here
Question: Which of the sentiment labels in the following list does this image belong to? List: ["amusement", "anger", "awe", "contentment", "disgust", "excitement", "fear", "sadness"] Answer:

Answer with the exact sentiment label as it appears in the list.

Figure 5: Prompts used for multi-class zero-shot classification and EmoGist.

F Prompts used for ICL baselines

```
**ICL example images 1 to 4 goes here**
**Test image goes here**

## Image 1
Question: Does this image matches the sentiment label
'{s_label}'? Answer: Yes

## Image 2
Question: Does this image matches the sentiment label
'{s_label}'? Answer: Yes

## Image 3
Question: Does this image matches the sentiment label
'{s_label}'? Answer: Yes

## Image 4
Question: Does this image matches the sentiment label
'{s_label}'? Answer: Yes

## Image 5
Question: Does this image matches the sentiment label
'{s_label}'? Answer:
```

Answer with 'Yes' or 'No'.

Figure 6: Prompts used for multi-label $\mathrm{ICL}_{\mathrm{sim}}$.

```
**ICL example images 1 to 4 goes here**
**Test image goes here**

## Image 1
Question: Which of the sentiment labels in the following
list does this image belong to? List: ["amusement",
"anger", "awe", "contentment", "disgust", "excitement",
"fear", "sadness"] Answer: {s_label}

## Image 2
Question: Which of the sentiment labels in the following
list does this image belong to? List: ["amusement",
"anger", "awe", "contentment", "disgust", "excitement",
"fear", "sadness"] Answer: {s_label}

## Image 3
Question: Which of the sentiment labels in the following
list does this image belong to? List: ["amusement",
"anger", "awe", "contentment", "disgust", "excitement",
"fear", "sadness"] Answer: {s_label}

## Image 4
Question: Which of the sentiment labels in the following
list does this image belong to? List: ["amusement",
"anger", "awe", "contentment", "disgust", "excitement",
"fear", "sadness"] Answer: {s_label}

## Image 5
```

Question: Which of the sentiment labels in the following list does this image belong to? List: ["amusement", "anger", "awe", "contentment", "disgust", "excitement", "fear", "sadness"] Answer:

Answer with the exact sentiment label as it appears in the list.

Figure 7: Prompts used for multi-class $\mathrm{ICL}_{\mathrm{sim}}$.

```
**ICL example images 1 to 8 goes here**
**Test image goes here**

## Image 1
Question: Which of the sentiment labels in the following
list does this image belong to? List: [“amusement”,
“anger”, “awe”, “contentment”, “disgust”, “excitement”,
“fear”, “sadness”] Answer: amusement

## Image 2
Question: Which of the sentiment labels in the following
list does this image belong to? List: [“amusement”,
“anger”, “awe”, “contentment”, “disgust”, “excitement”,
“fear”, “sadness”] Answer: anger

## Image 3
Question: Which of the sentiment labels in the following
list does this image belong to? List: [“amusement”,
“anger”, “awe”, “contentment”, “disgust”, “excitement”,
“fear”, “sadness”] Answer: awe

## Image 4
Question: Which of the sentiment labels in the following
list does this image belong to? List: [“amusement”,
“anger”, “awe”, “contentment”, “disgust”, “excitement”,
“fear”, “sadness”] Answer: contentment

## Image 5
Question: Which of the sentiment labels in the following
list does this image belong to? List: [“amusement”,
“anger”, “awe”, “contentment”, “disgust”, “excitement”,
“fear”, “sadness”] Answer: disgust

## Image 6
Question: Which of the sentiment labels in the following
list does this image belong to? List: [“amusement”,
“anger”, “awe”, “contentment”, “disgust”, “excitement”,
“fear”, “sadness”] Answer: excitement

## Image 7
Question: Which of the sentiment labels in the following
list does this image belong to? List: [“amusement”,
“anger”, “awe”, “contentment”, “disgust”, “excitement”,
“fear”, “sadness”] Answer: fear

## Image 8
Question: Which of the sentiment labels in the following
list does this image belong to? List: [“amusement”,
“anger”, “awe”, “contentment”, “disgust”, “excitement”,
“fear”, “sadness”] Answer: sadness

## Image 9
```

Question: Which of the sentiment labels in the following list does this image belong to? List: [“amusement", “anger", “awe", “contentment", “disgust", “excitement", “fear", “sadness"] Answer:

Answer with the exact sentiment label as it appears in the list.

Figure 8: Prompts used for multi-class, ICL_{all}.

Qwen2.5-VL 7B: FI validation set (1031 examples) - Seed 21

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	69.032%	69.920%	67.758%
+ k-means n_clusters=4	Macro	72.865%	69.499%	70.279%
+ k-means n_clusters=6	Macro	73.395%	69.957%	**70.545%**

Aya Vision 8B: FI validation set (1031 examples) - Seed 21

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	67.409%	68.413%	65.939%
+ k-means n_clusters=4	Macro	70.595%	67.732%	68.223%
+ k-means n_clusters=6	Macro	71.948%	69.044%	**69.484%**

InternVL2.5 8B-MPO: FI validation set (1031 examples) - Seed 21

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	68.627%	68.100%	66.318%
+ k-means n_clusters=4	Macro	71.786%	67.964%	68.942%
+ k-means n_clusters=6	Macro	72.140%	69.071%	**69.657%**

Qwen2.5-VL 7B: FI validation set (1031 examples) - Seed 42

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	71.758%	68.817%	69.161%
+ k-means n_clusters=4	Macro	73.190%	69.687%	70.295%
+ k-means n_clusters=6	Macro	72.642%	71.265%	**70.945%**

Aya Vision 8B: FI validation set (1031 examples) - Seed 42

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	69.464%	67.008%	66.976%
+ k-means n_clusters=4	Macro	71.665%	68.810%	69.073%
+ k-means n_clusters=6	Macro	71.278%	70.286%	**69.741%**

InternVL2.5 8B-MPO: FI validation set (1031 examples) - Seed 42

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	69.944%	67.033%	67.323%
+ k-means n_clusters=4	Macro	71.631%	68.661%	69.066%
+ k-means n_clusters=6	Macro	71.591%	70.343%	**70.012%**

Qwen2.5-VL 7B: FI validation set (1031 examples) - Seed 63

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	73.046%	69.215%	**69.802%**
+ k-means n_clusters=4	Macro	72.401%	67.529%	68.147%
+ k-means n_clusters=6	Macro	72.080%	69.147%	69.568%

Aya Vision 8B: FI validation set (1031 examples) - Seed 63

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	70.628%	67.879%	**68.083%**
+ k-means n_clusters=4	Macro	71.767%	67.099%	67.576%
+ k-means n_clusters=6	Macro	70.298%	67.914%	67.990%

InternVL2.5 8B-MPO: FI validation set (1031 examples) - Seed 63

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	71.901%	68.071%	**68.694%**
+ k-means n_clusters=4	Macro	72.597%	67.204%	68.031%
+ k-means n_clusters=6	Macro	70.030%	67.291%	67.594%

Figure 9: Validation set results for EmoGist_n with 7B models on FI.

Qwen2.5-VL 7B: FI validation set (1031 examples) - Seed 21-1

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	68.167%	68.870%	66.595%
+ k-means n_clusters=4	Macro	71.651%	68.779%	69.271%
+ k-means n_clusters=6	Macro	72.627%	69.668%	**70.104%**

Aya Vision 8B: FI validation set (1031 examples) - Seed 21-1

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	67.409%	68.413%	65.939%
+ k-means n_clusters=4	Macro	71.603%	68.653%	69.147%
+ k-means n_clusters=6	Macro	72.369%	69.482%	**69.882%**

InternVL2.5 8B-MPO: FI validation set (1031 examples) - Seed 21-1

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	65.562%	66.897%	64.388%
+ k-means n_clusters=4	Macro	70.913%	68.242%	68.744%
+ k-means n_clusters=6	Macro	71.458%	68.824%	**69.233%**

Qwen2.5-VL 7B: FI validation set (1031 examples) - Seed 42-1

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	70.314%	67.950%	67.994%
+ k-means n_clusters=4	Macro	71.938%	69.185%	69.439%
+ k-means n_clusters=6	Macro	71.764%	70.873%	**70.303%**

Aya Vision 8B: FI validation set (1031 examples) - Seed 42-1

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	69.765%	67.347%	67.309%
+ k-means n_clusters=4	Macro	71.724%	68.878%	69.146%
+ k-means n_clusters=6	Macro	71.758%	70.656%	**70.128%**

InternVL2.5 8B-MPO: FI validation set (1031 examples) - Seed 42-1

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	69.871%	67.414%	67.517%
+ k-means n_clusters=4	Macro	72.246%	68.310%	69.164%
+ k-means n_clusters=6	Macro	71.460%	70.017%	**69.719%**

Qwen2.5-VL 7B: FI validation set (1031 examples) - Seed 63-1

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	71.980%	69.102%	**69.442%**
+ k-means n_clusters=4	Macro	72.360%	67.612%	68.129%
+ k-means n_clusters=6	Macro	71.124%	68.635%	68.774%

Aya Vision 8B: FI validation set (1031 examples) - Seed 63-1

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	71.631%	68.652%	**68.871%**
+ k-means n_clusters=4	Macro	71.845%	67.192%	67.694%
+ k-means n_clusters=6	Macro	70.712%	67.362%	67.648%

InternVL2.5 8B-MPO: FI validation set (1031 examples) - Seed 63-1

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	71.772%	68.683%	**69.160%**
+ k-means n_clusters=4	Macro	71.663%	66.832%	67.615%
+ k-means n_clusters=6	Macro	70.099%	66.754%	67.339%

Figure 10: Validation set results for EmoGist_e with 7B models on FI.

Qwen2.5-VL 7B: Memotion 1.0 - validation set - Seed 21

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	54.328%	74.819%	64.403%	**69.221%**
+ k-means n_clusters=4	Micro	54.018%	75.712%	63.158%	68.868%
+ k-means n_clusters=6	Micro	53.875%	75.836%	62.875%	68.750%

Aya Vision 8B: Memotion 1.0 - validation set - Seed 21

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	60.479%	67.365%	85.512%	**75.362%**
+ k-means n_clusters=4	Micro	58.858%	69.793%	78.325%	73.813%
+ k-means n_clusters=6	Micro	58.488%	69.101%	78.721%	73.598%

InternVL2.5 8B-MPO: Memotion 1.0 - validation set - Seed 21

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	58.453%	72.485%	73.797%	**73.135%**
+ k-means n_clusters=4	Micro	54.995%	72.434%	68.704%	70.520%
+ k-means n_clusters=6	Micro	56.092%	73.029%	69.723%	71.338%

Qwen2.5-VL 7B: Memotion 1.0 - validation set - Seed 42

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	53.092%	76.288%	61.177%	67.902%
+ k-means n_clusters=4	Micro	53.469%	76.303%	62.139%	68.497%
+ k-means n_clusters=6	Micro	53.529%	75.737%	62.535%	**68.506%**

Aya Vision 8B: Memotion 1.0 - validation set - Seed 42

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	56.521%	70.054%	73.741%	71.850%
+ k-means n_clusters=4	Micro	59.382%	68.562%	81.211%	**74.352%**
+ k-means n_clusters=6	Micro	59.108%	68.296%	81.437%	74.290%

InternVL2.5 8B-MPO: Memotion 1.0 - validation set - Seed 42

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	56.235%	74.459%	68.138%	71.158%
+ k-means n_clusters=4	Micro	56.032%	72.432%	70.628%	71.519%
+ k-means n_clusters=6	Micro	56.533%	71.811%	71.364%	**71.587%**

Qwen2.5-VL 7B: Memotion 1.0 - validation set - Seed 63

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	54.495%	76.078%	63.894%	**69.456%**
+ k-means n_clusters=4	Micro	54.137%	76.074%	63.158%	69.017%
+ k-means n_clusters=6	Micro	53.743%	75.559%	63.158%	68.804%

Aya Vision 8B: Memotion 1.0 - validation set - Seed 63

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	62.625%	70.249%	84.720%	**76.809%**
+ k-means n_clusters=4	Micro	59.239%	71.182%	77.023%	73.987%
+ k-means n_clusters=6	Micro	59.287%	68.094%	81.890%	74.358%

InternVL2.5 8B-MPO: Memotion 1.0 - validation set - Seed 63

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	56.009%	72.722%	70.006%	71.338%
+ k-means n_clusters=4	Micro	56.319%	73.088%	69.779%	71.395%
+ k-means n_clusters=6	Micro	56.676%	72.923%	71.024%	**71.961%**

Figure 11: Validation set results for EmoGist_n with 7B models on Memotion.

Qwen2.5-VL 7B: Memotion 1.0 - validation set - Seed 21-1

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	57.743%	79.213%	64.912%	71.353%
+ k-means n_clusters=4	Micro	58.233%	79.589%	65.761%	**72.017%**
+ k-means n_clusters=6	Micro	58.107%	79.209%	65.761%	71.861%

Aya Vision 8B: Memotion 1.0 - validation set - Seed 21-1

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	63.805%	68.701%	89.813%	**77.851%**
+ k-means n_clusters=4	Micro	61.722%	70.577%	82.400%	76.031%
+ k-means n_clusters=6	Micro	63.379%	69.804%	86.474%	77.250%

InternVL2.5 8B-MPO: Memotion 1.0 - validation set - Seed 21-1

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	61.973%	76.355%	74.929%	**75.636%**
+ k-means n_clusters=4	Micro	60.806%	76.193%	73.175%	74.654%
+ k-means n_clusters=6	Micro	60.705%	76.314%	73.118%	74.682%

Qwen2.5-VL 7B: Memotion 1.0 - validation set - Seed 42-1

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	55.823%	79.651%	62.026%	69.742%
+ k-means n_clusters=4	Micro	57.856%	79.406%	65.025%	71.500%
+ k-means n_clusters=6	Micro	58.107%	79.642%	65.535%	**71.903%**

Aya Vision 8B: Memotion 1.0 - validation set - Seed 42-1

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	57.380%	69.778%	74.873%	72.236%
+ k-means n_clusters=4	Micro	60.178%	68.851%	82.060%	74.877%
+ k-means n_clusters=6	Micro	62.977%	71.174%	83.701%	**76.931%**

InternVL2.5 8B-MPO: Memotion 1.0 - validation set - Seed 42-1

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	59.827%	73.935%	73.684%	73.810%
+ k-means n_clusters=4	Micro	59.689%	76.162%	71.420%	73.715%
+ k-means n_clusters=6	Micro	60.228%	75.541%	73.062%	**74.281%**

Qwen2.5-VL 7B: Memotion 1.0 - validation set - Seed 63-1

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	58.446%	79.644%	65.761%	72.040%
+ k-means n_clusters=4	Micro	58.434%	79.085%	66.553%	72.280%
+ k-means n_clusters=6	Micro	58.886%	79.408%	66.780%	**72.548%**

Aya Vision 8B: Memotion 1.0 - validation set - Seed 63-1

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	63.215%	69.526%	87.153%	77.348%
+ k-means n_clusters=4	Micro	62.011%	69.444%	84.890%	76.394%
+ k-means n_clusters=6	Micro	64.194%	72.311%	84.097%	**77.760%**

InternVL2.5 8B-MPO: Memotion 1.0 - validation set - Seed 63-1

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	60.730%	75.216%	73.854%	74.529%
+ k-means n_clusters=4	Micro	60.191%	75.319%	73.401%	74.348%
+ k-means n_clusters=6	Micro	61.709%	77.143%	73.345%	**75.196%**

Figure 12: Validation set results for EmoGist_e with 7B models on Memotion.

InternVL2.5 2B-MPO: FI validation set (1031 examples) - Seed 21-1 (mmE5)

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	64.377%	54.339%	**56.048%**
+ k-means n_clusters=4	Macro	63.459%	50.837%	51.327%
+ k-means n_clusters=6	Macro	67.200%	52.979%	55.212%

InternVL2.5 2B-MPO: FI validation set (1031 examples) - Seed 42-1 (mmE5)

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	65.462%	51.573%	53.279%
+ k-means n_clusters=4	Macro	63.992%	50.652%	51.759%
+ k-means n_clusters=6	Macro	64.640%	50.546%	**53.390%**

InternVL2.5 2B-MPO: FI validation set (1031 examples) - Seed 63-1 (mmE5)

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	67.341%	52.527%	52.746%
+ k-means n_clusters=4	Macro	65.558%	49.219%	50.590%
+ k-means n_clusters=6	Macro	64.328%	54.485%	**55.939%**

SmolVLM2 2.2B: FI validation set (1031 examples) - Seed 21-1 (mmE5)

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	66.498%	66.774%	64.864%
+ k-means n_clusters=4	Macro	70.155%	67.322%	**67.791%**
+ k-means n_clusters=6	Macro	70.012%	67.109%	67.407%

SmolVLM2 2.2B: FI validation set (1031 examples) - Seed 42-1 (mmE5)

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	68.687%	66.411%	66.385%
+ k-means n_clusters=4	Macro	67.389%	63.990%	63.941%
+ k-means n_clusters=6	Macro	70.584%	69.679%	**69.100%**

SmolVLM2 2.2B: FI validation set (1031 examples) - Seed 63-1 (mmE5)

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	69.334%	67.021%	**66.945%**
+ k-means n_clusters=4	Macro	69.938%	65.412%	65.973%
+ k-means n_clusters=6	Macro	68.853%	66.546%	66.578%

SmolVLM2 2.2B: FI validation set (1031 examples) - Seed 84 (mmE5)

		Precision	Recall	F1
+ k-means n_clusters=2	Macro	71.375%	66.280%	66.562%
+ k-means n_clusters=4	Macro	69.236%	65.712%	65.962%
+ k-means n_clusters=6	Macro	70.106%	68.678%	**68.442%**

Figure 13: Validation set results for EmoGist$_e$ with 2B models on FI.

InternVL2.5 2B-MPO: Memotion 1.0 - validation set - Seed 21-1-1-1 (mmE5)

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	49.372%	69.955%	61.404%	65.401%
+ k-means n_clusters=4	Micro	48.343%	70.746%	59.536%	64.659%
+ k-means n_clusters=6	Micro	49.209%	70.968%	61.007%	**65.612%**

InternVL2.5 2B-MPO: Memotion 1.0 - validation set - Seed 42-1-1-1 (mmE5)

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	50.289%	70.978%	62.422%	66.426%
+ k-means n_clusters=4	Micro	50.188%	71.353%	61.460%	66.038%
+ k-means n_clusters=6	Micro	51.092%	71.208%	62.705%	**66.687%**

InternVL2.5 2B-MPO: Memotion 1.0 - validation set - Seed 63-1-1-1 (mmE5)

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	48.494%	70.113%	59.875%	64.591%
+ k-means n_clusters=4	Micro	44.139%	66.689%	56.650%	61.261%
+ k-means n_clusters=6	Micro	49.034%	71.070%	60.894%	**65.590%**

SmolVLM2 2.2B: Memotion 1.0 - validation set - Seed 21-1-1-1-1 (mmE5)

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	66.529%	66.529%	100.000%	**79.901%**
+ k-means n_clusters=4	Micro	64.759%	68.048%	92.926%	78.565%
+ k-means n_clusters=6	Micro	65.951%	68.462%	94.228%	79.305%

SmolVLM2 2.2B: Memotion 1.0 - validation set - Seed 42-1-1-1-1 (mmE5)

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	56.890%	66.541%	79.796%	72.568%
+ k-means n_clusters=4	Micro	60.906%	66.883%	87.550%	75.833%
+ k-means n_clusters=6	Micro	64.684%	66.588%	96.095%	**78.666%**

SmolVLM2 2.2B: Memotion 1.0 - validation set - Seed 63-1-1-1-1 (mmE5)

		Accuracy (by label)	Precision	Recall	F1
+ k-means n_clusters=2	Micro	65.374%	66.615%	97.453%	79.136%
+ k-means n_clusters=4	Micro	65.223%	66.667%	97.114%	79.060%
+ k-means n_clusters=6	Micro	66.504%	68.936%	94.567%	**79.742%**

Figure 14: Validation set results for EmoGist$_e$ with 2B models on Memotion.