Acknowledgements This research was partially supported by NSERC, Musée National des Beaux-Arts du Québec and Strateolab, as well as FRQ-NT and NSERC MS fellowships to E. Bergeron. Computing resources were provided by the Digital Research Alliance Canada. The authors thank our lab members for help with proofreading.

References

[1] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. Artemis: Affective language for visual art. *CoRR*, abs/2101.07396, 2021. 1, 2, 3

[2] Panos Achlioptas, Maks Ovsjanikov, Leonidas J. Guibas, and S. Tulyakov. Affection: Learning affective explanations for real-world visual data. *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2022. 3, 8

[3] Margaret M. Bradley and Peter J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Jour. of Behav. Ther. and Exp. Psy.*, 25(1):49–59, 1994. 2

[4] Margaret M. Bradley and Peter J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. In *Technical report C-1*, 1999. 2, 5

[5] Filipe Galvao, Soraia M. Alarcão, and Manuel José Fonseca. Predicting exact valence and arousal values from eeg. *Sensors*, 21, 2021. 3

[6] Lancheng Gao, Ziheng Jia, Yunhao Zeng, Wei Sun, Yiming Zhang, Wei Zhou, Guangtao Zhai, and Xiongkuo Min. Eemobench: A benchmark for multi-modal large language models on image evoked emotion assessment. In *ACM Int. Conf. Multimedia*, 2025. 3, 6, 1

[7] Tao He and Xiaoming Jin. Image emotion distribution learning with graph convolutional networks. In *Int. Conf. Multim. Retr.*, 2019. 3

[8] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938. 5

[9] Seoyun Kim, ChaeHee An, Junyeop Cha, Dongjae Kim, and Eunil Park. D-visa: A dataset for detecting visual sentiment from art images. In *IEEE/CVF Int. Conf. Comput. Vis. Worksh.*, 2023. 2, 3

[10] Ronak Kosti, José Manuel Álvarez, Adrià Recasens, and Àgata Lapedriza. Emotic: Emotions in context dataset. *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2017. 3, 8

[11] SangEun Lee, Seoyun Kim, Yubeen Lee, Jufeng Yang, and Eunil Park. Enhancing dimensional image emotion detection with a low-resource dataset via two-stage training. *IEEE Trans. Cogni. Dev. Systems*, 17(3):455–464, 2025. 3

[12] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2022. 4

[13] 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 *Adv. Neural Inform. Process. Syst.*, 2024. 3

[14] J.A. Mikels, B.L. Fredrickson, and G.R. et al Larkin. Emotional category data on images from the international affective picture system. In *Behavior Research Methods*, 2005. 2, 3

[15] Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Church, and Mohamed Elhoseiny. ArtELingo: A million emotion annotations of WikiArt with emphasis on diversity over language and culture. In *Empi. Meth. in Nat. Lang. Proc.*, 2022. 2, 3

[16] Youssef Mohamed, Faizan Farooq Khan, Kilichbek Haydarov, and Mohamed Elhoseiny. It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2022. 2, 3

[17] Saif M. Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Ann. Conf. of the Ass. for Comp. Lingu.*, 2018. 2, 3, 5, 1

[18] Saif M. Mohammad and Svetlana Kiritchenko. An annotated dataset of emotions evoked by art. In *Lang. Res. and Eval. Conf*, 2018. 3, 6

[19] Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, and Alice Oh. Dimensional emotion detection from categorical emotion. In *Empi. Meth. in Nat. Lang. Proc.*, 2021. 3

[20] Egon S. Pearson. The test of significance for the correlation coefficient. *Jour. Amer. Stat. Asso.*, 26(174):128–134, 1931. 6

[21] Robert Plutchik. The Nature of Emotions. *American Scientist*, 89(4):344, 2001. 7

[22] James Russell. A circumplex model of affect. *Jour. Pers. Soc. Psy.*, 39:1161–1178, 1980. 2

[23] A. Savchenko. Emotieffnets for facial processing in video-based valence-arousal prediction, expression classification and action unit detection. *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 5716–5724, 2023. 3

[24] David W. Scott. Multivariate density estimation: Theory, practice, and visualization. In *Wiley Series in Probability and Statistics*, 1992. 4

[25] J. Striegl, JW. Richter, L. Grossmann, B. Bråstad, M. Gotthardt, C. Rück, J. Wallert, and C. Loitsch. Deep learning-based dimensional emotion recognition for conversational agent-based cognitive behavioral therapy. In *PeerJ Comp. Sci.*, 2024. 3, 5

[26] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Florian Eyben, and Björn Schuller. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45:10745–10759, 2022. 3

[27] Haitao Xiong, Hongfu Liu, Bineng Zhong, and Yun Fu. Structured and sparse annotations for image emotion distribution learning. *Assoc. Adv. of Art. Int.*, 2019. 3

[28] Jufeng Yang, Dongyu She, and Ming Sun. Joint image emotion classification and distribution learning via deep convolutional neural network. In *IJCAI*, 2017. 3

[29] Jingyuan Yang, Jie Li, Leida Li, Xiumei Wang, and Xinbo Gao. A circular-structured representation for visual emotion distribution learning. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2021. 1, 3, 2

[30] Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Emoset: A large-scale

visual emotion dataset with rich attributes. In *IEEE/CVF Int. Conf. Comput. Vis.*, 2023. 2

[31] M. S. M. Yik, J. A. Russell, and L. F Barrett. Structure of self-reported current affect: Integration and beyond. In *Jour. Pers. Soc. Psy.*, 1999. 2

[32] Cheng Zhang, Hongxia Xie, Bin Wen, Songhan Zuo, Ruoxuan Zhang, and Wen-Huang Cheng. Emoart: A multidimensional dataset for emotion-aware artistic generation. In *ACM Int. Conf. Multimedia*, 2025. 3

[33] Jing Zhang, Liang Zheng, Meng Wang, and Dan Guo. Training a small emotional vision language model for visual art comprehension. In *Eur. Conf. Comput. Vis.*, 2024. 1, 3

[34] Sicheng Zhao, Hongxun Yao, Yue Gao, R. Ji, and Guiguang Ding. Continuous probability distribution prediction of image emotions via multitask shared sparse regression. *IEEE Trans. Multimedia*, 19:632–645, 2017. 3

Dimensional Distribution Emotion State: Leveraging Valence and Arousal as a Common Embedding Space for Visual Emotion Analysis

Supplementary Material

6. Adapting large pre-trained VLMs for emotion prediction

In the main section of this paper, we train models specifically for the task of predicting the emotional response evoked by artworks. In this section, drawing inspiration from EEmo-Bench [6], our goal is to determine whether large pretrained VLMs can be adapted for predicting the different emotion representations identified in this paper without retraining.

6.1. VLM adaptation

We devise different ways of mapping the model's internal representation to the three emotion representations from sec. 3. To do so, we extract the predicted probabilities of the next token given a context prompt using the following methods.

CES representation To obtain a CES representation, we give the model a contextual prompt: "Assume you are an expert in emotional psychology. Choose one word from the following list: (`<emotion-list-str>`). Which emotion is primarily expressed in this image? The primary emotion is". The model then chooses the most likely token in its vocabulary to start its answer. At this moment, we extract the probabilities of each emotion label using their precomputed token IDs, before undergoing a normalization to obtain the final distribution. The list of emotions `<emotion-list-str>` is defined according to the emotion labels present in the corresponding dataset. We sometimes run into the case where a single emotion label is split into multiple tokens by the tokenizer. To circumvent this issue, we establish a mapping from single-token, rarely used symbols to our target emotion set, and place it into the prompt.

DES representation To obtain a DES representation, we follow the same methodology as [6], where we prompt the model with the following text for each of valence/arousal: "How would you rate the valence/arousal this image evokes in the viewer? The level of valence/arousal this image evokes in the viewer is". The query tokens are "Negative", "Neutral", "Positive", and probabilities for each token are then weighted with [-1.0, 0, 1.0] respectively, before being normalized using softmax.

DDES representation To obtain a DDES representation, we first identified some emotional keywords from the NRC VAD lexicon [17] such that their valence-arousal coordinates spanned the 2D space as thoroughly and equally as possible. Each symbol is mapped to its corresponding keyword in the following prompt: "Assume you are an expert in emotional psychology. Analyze the viewer's emotion evoked by this image. The emotion codes are mapped as follows: ♢ *is boredom,* ⊙ *is hopeless,* △ *is unhappy,* ▽ *is humiliated,* ♠ *is rage,* ♡ *is shy,* ♢ *is timid,* ♣ *is beholden,* ★ *is perplexed,* ⋆ *is flustered,* ♦ *is sleepy,* ♢ *is contemplation,* • *is reverent,* ○ *is forceful,* □ *is fanatical,* ■ *is mellow,* ▲ *is reflective,* ▼ *is yearning,* † *is expectant,* ‡ *is elation,* § *is calming,* ¶ *is grace,* ¤ *is prized,* € *is proud,* £ *is surprise*. Which single code from the set (`<token-labels>`) best represents the highest probability emotion? The code is"
Then, we extract the probabilities for every keyword, and compute a KDE on the resulting weighted point cloud. The density function is then discretized and normalized, following our method. This is perhaps not the emotion representation most adapted to a VLM's representation, but we wanted to include it for comparison purposes.

6.2. Experimental results

In all of our experiments, we use the open-source Qwen-2.5-7B VLM model. We performed the same experiments as in sec. 4. Note that we could not however replicate the seen/unseen scenarios, as we have to assume that a model of this scale has seen ArtEmis, D-ViSA, EmoSet, EEmo-Bench and WikiArt Emotions during training.

6.3. Observations

Querying a VLM leads to similar performance as supervised training on seen datasets. When compared the models trained on the combined datasets in tab. 3, we can observe in tab. 4 that the VLM leads to very similar performance on the seen datasets. Qwen performs slightly better than the best supervised model on ArtEmis and D-ViSA, and considerably worse on EmoSet.

VLMs perform much better on benchmark datasets, but can not be considered zero-shot generalization. On the benchmark datasets, Qwen considerably outperforms the supervised models in every metric. On EEmo-Bench, the difference is stark, with 0.757 against 0.59 Pearson correlation on the valence axis, and 0.421 against 0.32 on the arousal axis. The same applies for accuracy, at 52.1 against 21.2. Again, on WikiArt Emotions, we observe a considerable gap with 34.2 versus 10.5 for top-1 accuracy. This is a much bigger difference in performance than for the three

Table 4. Comparison of different methods of extracting emotional comprehension from VLMs (QWEN-2.5-7B).

Mapping Method	ArtEmis		D-ViSA			EmoSet	EEmo-Bench			WikiArt Emotions	
	$\text{Acc}_{\uparrow}$	$\tau_{\uparrow}$	$r_{v\uparrow}$	$r_{a\uparrow}$	$\text{RMSE}_{\downarrow}$	$\text{Acc}_{\uparrow}$	$r_{v\uparrow}$	$r_{a\uparrow}$	$\text{Acc}_{\uparrow}$	$\text{Acc}_{\uparrow}$	$\tau_{\uparrow}$
Logits to CES	**39.9**	0.138	**0.477**	0.009	0.648	**49.8**	0.750	**0.421**	**52.1**	**34.2**	0.121
Logits to DES	16.3	**0.249**	0.325	**0.105**	**0.644**	19.7	0.682	0.358	19.9	10.9	**0.127**
Logits to DDES	15.4	-0.008	0.297	-0.067	0.870	24.4	**0.757**	0.400	21.1	2.83	-0.144

seen datasets, which is to be expected as it is not a fair comparison. The VLM has seen the datasets during its training, and not only the test splits, but the train splits as well. The very large amount of data it has ingested prevents it from remembering every training example, but it still gives it an innate advantage over the specialized models.

Categorical emotion states are the representation most suited to extract a VLM's emotional understanding. The logits to CES method yields the best results for most metrics in tab. 4; it is best or second best in all cases. This representation is also the one that most closely matches the usual inference process of a VLM, which is based on predicting probabilities for text tokens.

7. Additional results

7.1. Experiments with DDES-NET training

Additional experiments with DDES-NET under the single-dataset training and combined training settings outlined in sec. 4. The results are reported in tab. 5. These experiments serve to explore research directions orthogonal to the choice of emotion representation.

Single-dataset We assess DDES-NET's zero-shot generalization capabilities by testing it on our two benchmark datasets, EEmo-Bench and WikiArt Emotions. We find that the model trained solely on ArtEmis actually significantly outperforms the model trained on the combined datasets when looking at top-1 accuracy for both datasets, but exhibits lower performance for valence-arousal regression on EEmo-Bench.

Multi-dataset We test DDES-NET on the combined training setting, with 3 variations. The first one adds an auxiliary loss inspired by [29]. Here, the predictions and ground truth are first converted to DES representation using our equations from sec. 3.2, and a position, norm and orientation loss are computed on the resulting vectors. These loss terms are then simply added to the original loss. This additional loss slightly increases performance on some metrics, while decreasing it on others. We chose not to use this as baseline. Other loss variations could be explored, such as Earth Mover's Distance (EMD), or regularization terms could be added, for instance to penalize more heavily zones with no mass in the ground truth.

The balanced sampling experiment consists of sampling each dataset an equal amount of time during training. In practice, this heavily upsamples D-ViSA, which is a much smaller size than ArtEmis and EmoSet. This leads to slightly increased performance on D-ViSA, ArtEmis and WikiArt Emotions, but much lower performance on EmoSet and EEmo-Bench. If the combined training procedure is expanded, and more affective datasets are added to the pool, sampling might become a more important concern, but for now didn't prove to be necessary.

Lastly, we substituted the enhanced ArtEmis with the base version in the combined training pipeline. We observe drastically worse results, as reported in sec. 4.5

8. Implementation details

8.1. Full emotion sets

The following are the full emotion sets used in datasets mentioned in sec. 4.2.

EEmo-Bench The full set of emotions is: *joy, surprise, fear, disgust, sadness, anger and neutral*.

WikiArt Emotions The full set of emotions is: *agreeableness, anger, anticipation, arrogance, disagreeableness,disgust, fear, gratitude, happiness, humility, love, optimism, pessimism, regret, sadness, shame, shyness, surprise, trust, and neutral*.

Table 5. Additional experiments conducted on the DDES-NET model. In the single dataset training scenario, we evaluate a DDES-NET model trained solely on ArtEmis on the additional benchmark datasets, EEmo-Bench and WikiArt Emotions. In the combined training section, we experiment with adding an auxiliary loss, sampling the datasets samples equally each epoch instead of proportionally to their size, and with using the base version of ArtEmis that does not incorporate the emotion points from the affective explanations.

Experiment	**Seen Datasets**						**Unseen Datasets**				
	ArtEmis		**D-ViSA**			**EmoSet**	**EEmo-Bench**			**WikiArt Emotions**	
	$\text{Acc}_{\uparrow}$	$\tau_{\uparrow}$	$r_{v\uparrow}$	$r_{a\uparrow}$	$\text{RMSE}_{\downarrow}$	$\text{Acc}_{\uparrow}$	$r_{v\uparrow}$	$r_{a\uparrow}$	$\text{Acc}_{\uparrow}$	$\text{Acc}_{\uparrow}$	$\tau_{\uparrow}$
Single-Dataset Training											
ArtEmis only	40.3	0.34	0.34	-0.03	0.58	18.6	0.35	0.18	31.7	13.2	0.21
Combined Training											
Auxiliary loss	39.1	0.30	0.37	0.06	0.53	48.1	0.55	0.13	26.8	9.8	0.19
Balanced sampling	40.1	0.25	0.47	0.10	0.45	14.0	0.38	0.06	6.7	10.0	0.16
Base ArtEmis	13.9	0.17	0.184	-0.058	0.49	15.1	0.299	0.058	17.1	6.8	0.136