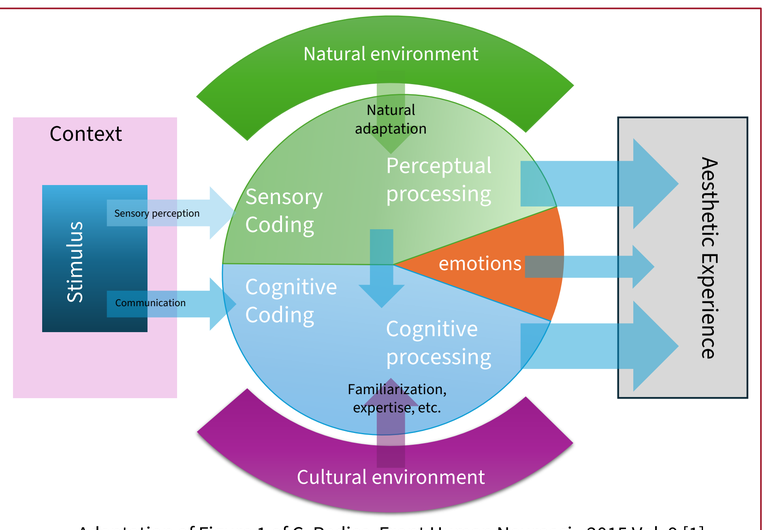
Arslan Javed, Bogdan Raducanu, C. Alejandro Parraga, Xim Cerdà
Computer Vision Center / Dept. Ciències de la Computació,
Universitat Autònoma de Barcelona (UAB), Spain.

Aesthetics with and without Semantics

1. Motivation

- **Two-channel Framework (Redies, 2015):**
  - Perceptual channel: bottom-up fast processing of low-level features such as colour, symmetry, texture, etc.
  - Semantic channel: top-down slow integration of meaning, context, etc.
  - Quantitatively separating these channels can clarify their individual contributions to aesthetic experience.
- **Strong confounds in Computational Aesthetics:**
  - Most research evaluate aesthetics without specifically addressing the confound between low-level features and high-level semantics (eg. the highest voted image in the AVA aesthetics dataset is a stylized depiction of the US flag).
  - How well semantic content alone can replicate human aesthetic judgements, and how strongly perceptual features alone drive preference when semantics are absent?
- **Impact:**
  - Our research reveals that, for everyday images, semantic content alone can reproduce human aesthetic judgements.
  - We also Identify occasions where low-level (perception-based) features remain essential.



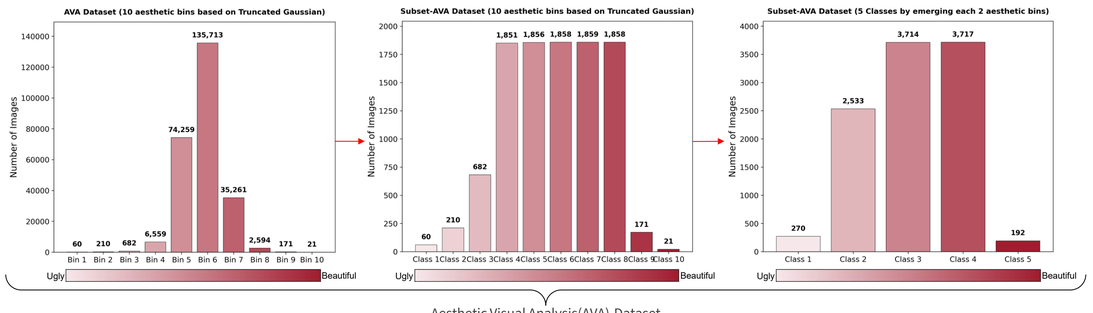
Adaptation of Figure 1 of C. Redies, Front Human Neurosci., 2015 Vol. 9 [1]

2. Methodology

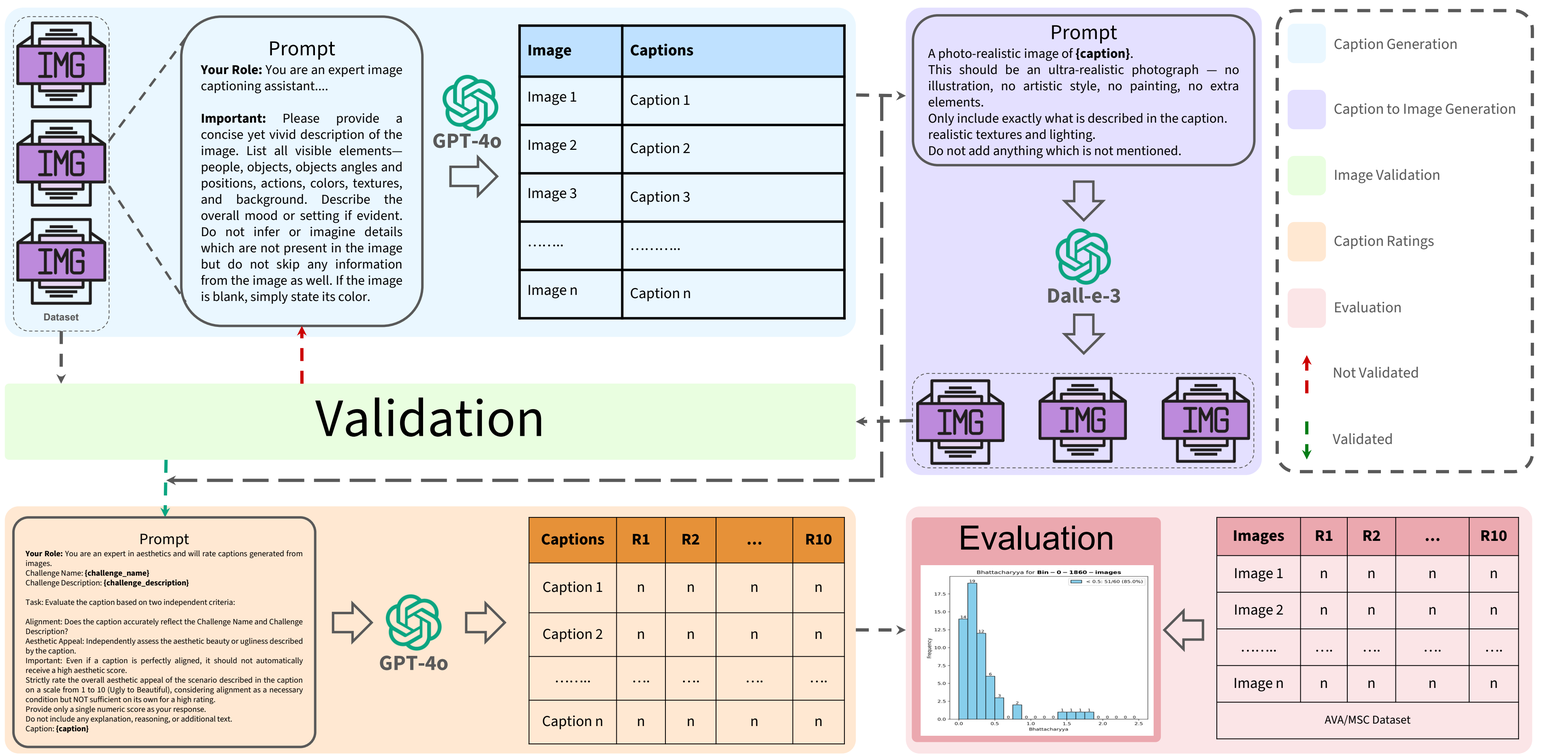
- **Two image datasets:**
  We took advantage of two existing image datasets:
  - AVA image dataset: created from samples of DPChallenge.com, an online image contest where users vote to select the winner of many “challenges”.
  - MSC image dataset: created to remove “semantic content”. Only images of natural objects were collected. No man-made objects, people or animals
  - Images in both datasets were assessed in terms of beauty by large numbers of observers(>= 10000 in total, 100-200 per image in average)

- **Two comparison methods:**
  We evaluated the human aesthetic judgements using two assessment methods:
  - **Method 1:** Text descriptions were extracted for each image and LLMs(GPT-4o, [2]) assessed the beauty of each image based solely on this text description.
  - **Method 2:** A set of low-level features (colour, lightness, contrast, etc.) was extracted using a ready-made image processing toolbox (QIP, [3]). A classifier was then trained to produce similar ratings as the observers.

  Each database was evaluated using each of the methods.



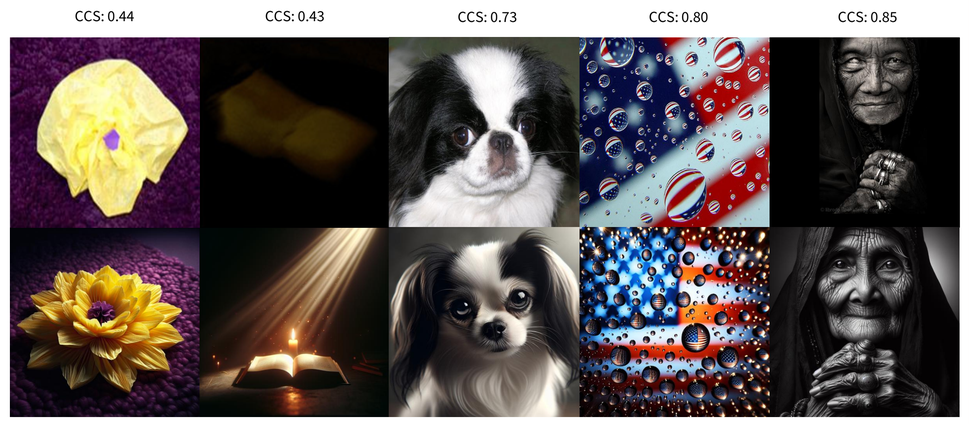
Workflow:



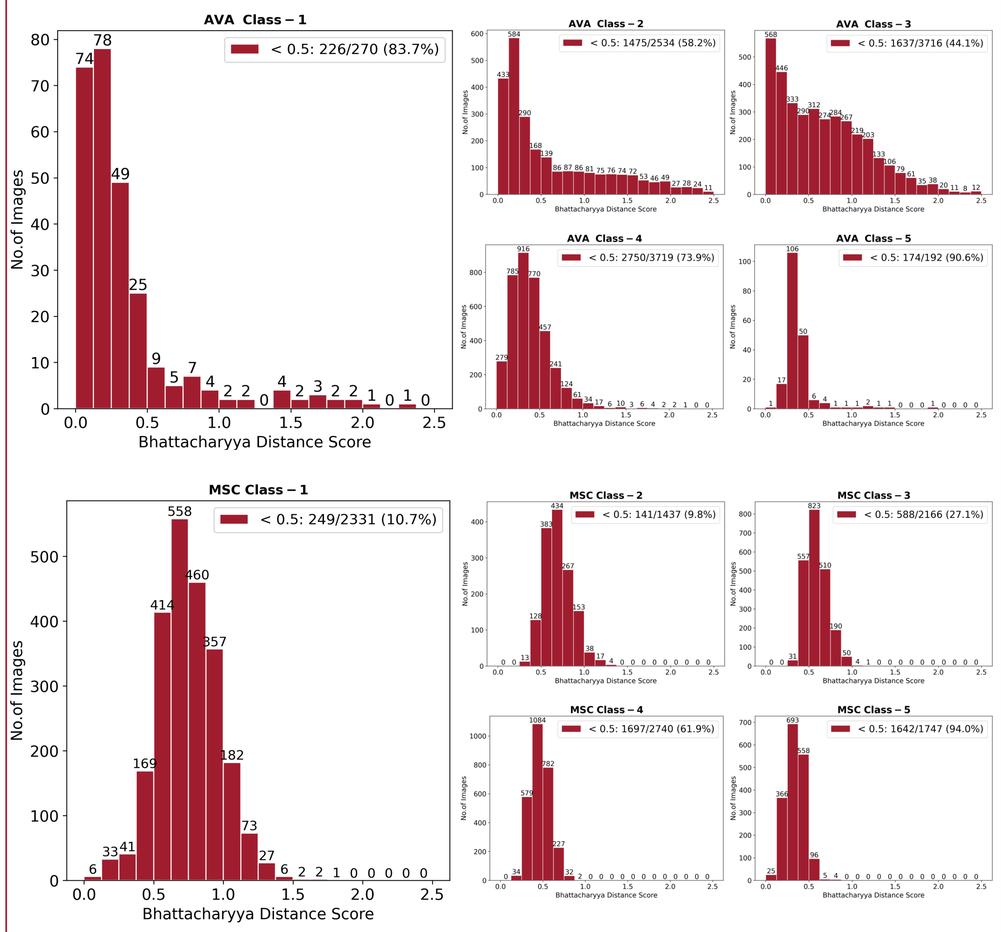
3. Experiments and Results

CLIP Cosine Similarity(CCS):

We designed psychophysics experiment to get the threshold of CLIP cosine similarity score for validation.

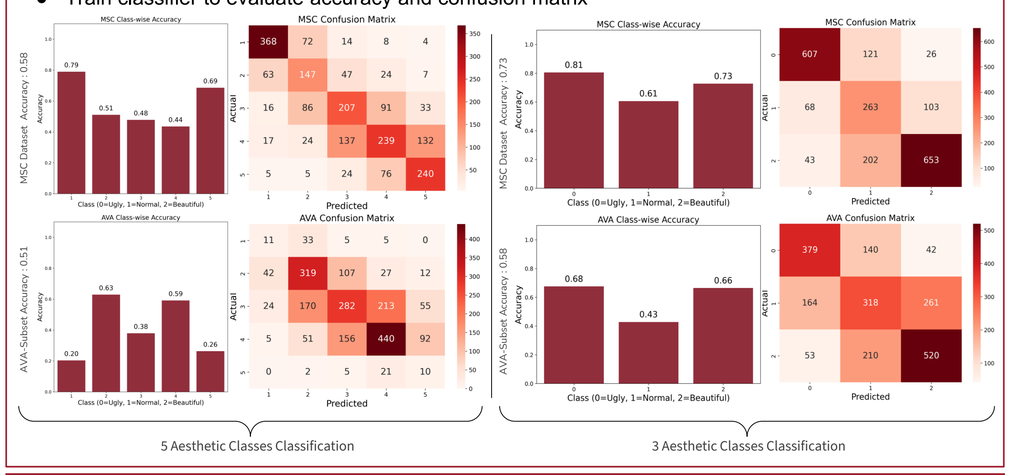


Method 1: AVA vs MSC



Method 2: AVA VS MSC

- Extract low-level features with QIP toolbox
- Train classifier to evaluate accuracy and confusion matrix



4. Discussion

- **Method 1:**
  - High alignment on AVA highlights that rich contextual information allows language models to approximate human aesthetics especially for the images at the aesthetic extremes.
  - Agreement drops for mid-range images, indicating that textual semantics alone do not fully capture nuanced human preference.
  - Lower alignment on MSC confirms that, with minimal semantic information, text-only models lack sufficient information for accurate prediction.

- **Method 2:**
  - **High accuracy on MSC** shows that, when semantic content is minimal, low-level visual features dominate aesthetic preference.
  - **Reduced accuracy on AVA** indicates that rich semantic content reduces the predictive power of purely perceptual features.

LLM AVA (High)	LLF AVA (Low)
LLM MSC (Low)	LLF MSC (High)

5. Conclusion and Future Work

- Accurate aesthetic prediction depends on recognising which channel is most informative: semantic information is crucial for images rich in context, whereas low-level visual features matter more when little semantic content is present.
- The results of AVA and MSC show that the amount of semantic information in a dataset has a major impact on how well each model works; therefore, using only an LLM or only low-level features is not sufficient for every type of image.
- Human preference study: Analyse the newly collected 2AFC similarity data to identify which visual features (e.g., colour–texture, composition, object shape, etc) most influence observers’ similarity decisions. Insights from this analysis will guide subsequent studies on perceptual criteria and their role in aesthetic evaluation.
- Develop adaptive hybrid models that combine semantic descriptions with low-level visual features promises more robust, human-aligned aesthetic assessment across diverse image types.

Acknowledgments:
- This work is supported by the Cátedra ENIA UAB-Cruïlla Chair on Research and Artificial Intelligence in the field of Music/Arts (TSI-100929-2023-2).
- Arslan Javed acknowledges the FI fellowship AGAUR 2022 FI-SDUR 00248 (Secretaria d’Universitats i Recerca, Generalitat de Catalunya, and Fons Social Europeu)

References:

[1] Redies, C., 2015. Combining universal beauty and cultural context in a unifying model of visual aesthetic experience. Frontiers in human neuroscience, 9, p.218.
[2] OpenAI. 2023. Gpt-4 technical report. Preprint, arXiv:2303.08774.
[3] Redies, C., Bartho, R., Koßmann, L., Spehar, B., Hübner, R., Wagemans, J. and Hayn-Leichsenring, G.U., 2025. A toolbox for calculating quantitative image properties in aesthetics research. Behavior Research Methods, 57(4), p.117.