Empirical Article

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Advances in Methods and
Practices in Psychological Science
January-March 2026, Vol. 9, No. 1,
pp. 1–28
© The Author(s) 2026
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/25152459251415336
www.psychologicalscience.org/AMPPS

Sage

Using Artificial Intelligence to Generate Affective Images: Methodology and Initial Library

Maciej Behnke[1,2], Maciej Kłoskowski[1], Michał Klichowski[2,3], Wadim Krzyżaniak[1], Kacper Szymański[1], Patryk Maciejewski[1], Patrycja Chwiłkowska[1], Marta Kowal[4], Rafał Jończyk[2,5], Jan Nowak[6], Szymon Kupiński[6], Dominika Kunc[7], Stanisław Saganowski[7], Aakash A. Chowkase[8], Farida Guemaz[9], Kevin S. Kertechian[10], Ameer I. M. T. Maadal[11], Leonardo A. Aguilar[12], Barnabas T. Alayande[13], Vimala Balakrishnan[14,15], Dana M. Basnight-Brown[16], Jordane Boudesseul[17,18], Tomás A. D'Amelio[19], Jovi C. Dacanay[20], Abhishek Dedhe[21,22], Shan Gao[23], Joao F. G. B. Takayanagi[24], Md. Rohmotul Islam[25], Alvaro Mailhos[26], Christine M. Mpyangu[27], Moises Mebarak[28], Arooj Najmussaqib[29], Ju Hee Park[30], Ekaterine Pirtskhalava[31], Eli Rice[32], Sohrab Sami[33], Yuki Yamada[34], Jan Baczyński[35], Lilianna Dera[36], Szymon Jęśko-Białek[37], Jakub Łączkowski[35], Hubert Marciniak[35], Filip Nowicki[35], Bartosz Wilczek[35], James J. Gross[38], and Nicholas A. Coles[24]

[1]Faculty of Psychology and Cognitive Science, Adam Mickiewicz University, Poznań, Poland; [2]Cognitive Neuroscience Center, Adam Mickiewicz University, Poznań, Poland; [3]Faculty of Educational Studies, Adam Mickiewicz University, Poznań, Poland; [4]IDN Being Human Lab – Institute of Psychology, University of Wrocław, Wrocław, Poland; [5]Faculty of English, Adam Mickiewicz University, Poznań, Poland; [6]Network Services Division, Poznan Supercomputing and Networking Center, Poznań, Poland; [7]Department of Artificial Intelligence, Wroclaw University of Science and Technology, Wrocław, Poland; [8]Yale Center for Emotional Intelligence, Child Study Center, Yale University, New Haven, Connecticut; [9]Department of Psychology and Educational Department, Mohamed Lamine Debaghine University Setif Setif, Algeria; [10]Organization, Management and Human Resource, ESSCA School of Management, Boulogne-Billancourt, France; [11]Department of Psychology, La Trobe University, Melbourne, Australia; [12]School of Psychology, Central University of Venezuela, Caracas, Venezuela; [13]Center for Equity in Global Surgery, University of Global Health Equity, Kigali, Rwanda; [14]Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia; [15]Department of Computer Science and Engineering, Korea University, Seoul, Korea; [16]Department of Psychology, United States International University-Africa, Nairobi, Kenya; [17]Laboratoire Parisien de Psychologie Sociale, University of Paris Nanterre, Paris, France; [18]Instituto de Investigación Científica, Universidad de Lima, Peru; [19]Centre de Recerca Matemàtica, Bellaterra, Spain; [20]School of Economics, University of Asia and the Pacific, Pasig City, Philippines; [21]Department of Psychology, Carnegie Mellon University, Pittsburgh, Pennsylvania; [22]Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, Pennsylvania; [23]School of Foreign Languages, University of Electronic Science and Technology of China, Chengdu, China; [24]Department of Psychology, University of Florida, Gainesville, Florida; [25]Department of Psychology, University of Chittagong, Chattogram, Bangladesh; [26]Facultad de Psicología, Universidad de la República, Montevideo, Uruguay; [27]Department of Religion

Corresponding Author:
Maciej Behnke, Faculty of Psychology and Cognitive Science, Adam Mickiewicz University, Poznań, Poland
Email: macbeh@amu.edu.pl

& Peace Studies, Makerere University, Kampala, Uganda; [28]Departamento de Psicología, Universidad del Norte, Barranquilla, Colombia; [29]Department of Professional Psychology, Bahria University Islamabad, Pakistan; [30]Department of Child and Family Studies, Yonsei University, Seoul, Korea; [31]Department of Psychology, Ivane Javakhishvili Tbilisi State University, Tbilisi, Georgia; [32]Department of Psychology, University of Pittsburgh, Pittsburgh, Pennsylvania; [33]Department of Psychology, University of Windsor, Windsor, Canada; [34]Faculty of Arts and Science, Kyushu University, Fukuoka, Japan; [35]Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland; [36]Faculty of Anthropology and Cultural Studies, Adam Mickiewicz University, Poznań, Poland; [37]Faculty of Medicine, Prince Mieszko I Medical Academy in Poznan, Poznań, Poland; and [38]Department of Psychology, Stanford University, Stanford, California

Abstract

We introduce a human-in-the-loop pipeline for creating context-aware (e.g., culture, sex, and age) affect-induction images and the initial Library of AI-Generated Affective Images. Current limitations in image-based research include weak to moderate emotional-elicitation effects, limited image diversity, and minimal cultural tailoring of images. Using generative artificial intelligence (AI) guided by existing data sets and emotion taxonomies, we generated 847 images and their corresponding descriptions across 12 discrete emotions and then iteratively refined them with local cultural experts. We validated the library through six studies ($N = 2{,}470$; 58 countries). Participants rated five types of images: (a) images from existing affective databases, (b) AI-generated images without cultural adjustments, (c) AI-generated images adjusted to specific cultural contexts, (d) AI-generated images adjusted by sex (male, female), and (e) AI-generated images adjusted by age group (childhood, adulthood, older age). The AI-generated images were as effective in eliciting affective responses as the images from existing affective databases. Culturally adjusted images were slightly more effective than unadjusted counterparts in targeting intended emotions. Sex- and age-adjusted variants produced comparable responses with their base images, demonstrating controllability without loss of affective impact. Furthermore, we calculated the smallest subjectively experienced difference for affect-induction research (ds = 0.05–0.29). This work demonstrates that researchers can now generate high-quality affect-induction stimuli cost-effectively and at scale and tailor them to diverse contexts—overcoming long-standing barriers and laying the groundwork for future AI-driven methodologies in affective science.

Keywords

affect, emotion, image database, affect-induction procedures, generative artificial intelligence, open data, open materials, preregistration

Received 6/25/25; Revision accepted 12/2/25

To conduct experimental studies in affective science, researchers need reliable and valid methods to elicit affective states. Thus, scientists often start by selecting stimuli for eliciting specific affective states, also known as "affect-induction procedures" (Joseph et al., 2020). Multiple databases of affective stimuli are widely available to researchers (Diconne et al., 2022) and include databases of auditory (e.g., International Affective Digitized Sounds, Bradley & Lang, 1999), verbal (e.g., Affective Norms for English Words, Stevenson et al., 2007), and visual materials (e.g., pictures of the affective content—International Affective Picture System [IAPS], Lang et al., 2008; Geneva Affective Picture Database [GAPED], Dan-Glauser & Scherer, 2011; Nencki Affective Picture System [NAPS], Marchewka et al., 2014). These tools significantly aid in the elicitation and measurement of affective states.

The rise of generative artificial intelligence (AI) offers an unprecedented opportunity to expand the methods available to elicit affective responses (Demszky et al., 2023). Here, we focus on two tasks generative AI excels at: generating and describing images. Our goal is to provide a pipeline—and an initial image library—that enables researchers to generate, validate, and adapt a versatile collection of affective images along with their descriptive prompts.

Existing Affect-Induction Images

Images of people, nature, and everyday objects are often the most prevalent stimuli used for affect elicitation (Joseph et al., 2020). Researchers typically use a wide variety of image databases, each designed to elicit specific affective reactions or represent particular themes (for a review, see Table 1).

Many of these databases focus on stimuli related to food and alcohol. For instance, databases such as the

Table 1. List of Databases With Affective Images

ID	Set name	Reference	Content	Dimensional ratings	Discrete	License	Participants	Rates per stimulus	Countries
1	ABPS	Pronk et al. (2015)	Alcohol and non-alcohol-containing pictures	Familiarity, valence, arousal, urge to drink, recognizability, control	—	CC BY NC 4.0	291	70	1
2	AFES_C	Scaini et al. (2017)	Images eliciting fear and anger plus neutral images	Intensity, valence, arousal, dominance	Fear, anger	ND	160	55	1
3	Animal.ID	Possidónio et al. (2019)	Animals spanning categories	Valence, arousal, familiarity, humanlike, palatability, cuteness, threat, acceptability, care	—	ND	500	51	1
4	BAPS_Ado	Szymanska et al. (2015)	Images related to attachment emotions	Valence, arousal, dominance, distress, comfort, complicity	Distress, comfort, joy, complicity	ND	140	70	1
5	BAPS_Adult	Szymanska et al. (2019)	Images representing distress, comfort, complicity, and neutrality with diverse people	Valence, arousal, dominance, horror, distinct emotion, complicity, aversiveness	Happiness, distress, comfort, complicity	CC BY NC 4.0	315	n/a	1
6	CaTIS	Noon et al. (2019)	Neutral, threat, and crime content	—	Threat, fear	Own	24	176	1
7	COMPASS	Weierich et al. (2019)	Social and nonsocial naturalistic scenes	Valence, arousal	—	CC BY 4.0	847	724	+2
8	CROCUFID	Toet et al. (2019)	Cross-cultural food-image database	Valence, arousal, healthiness, food craving	—	CC BY 4.0	805	100	3
9	DIRTI	Haberkamp et al. (2017)	Disgust-eliciting images and neutral images	Valence, arousal	Disgust, fear	CC BY 4.0	200	200	1
10	EmoMadrid	Carretié et al. (2019)	Emotional scenes	Valence, arousal	—	ND	146	146	1
11	EXCEED	de Sousa Magalhaes et al. (2018)	Neutral, drought, and flood stimuli	Valence, arousal	—	CC BY 4.0	50	50	1
12	Food_Cal	Shankland et al. (2019)	High- and low-calorie food images paired with nonfood images	Attractiveness, arousal, palatability	—	Own	264	264	1
13	Food_Pics	Blechert et al. (2014)	Food and nonfood images	Valence, arousal, palatability, visual complexity, food craving, recognizability	—	ND	1327	48	4
14	FRIDa	Foroni et al. (2013)	Food and nonfood images	Valence, arousal, familiarity, perceived calorie content, typicality, food craving, ambiguity	—	ND	86	18	1

(continued)

Table 1. (continued)

ID	Set name	Reference	Content	Dimensional ratings	Discrete	License	Participants	Rates per stimulus	Countries
15	GAPED	Dan-Glauser and Scherer (2011)	Snakes, spiders, human-concerns, animal-mistreatment, neutral, and positive pictures	Valence, arousal, acceptability	—	CC BY NC 3.0	60	61	1
16	GBPS	López-Caneda and Carbia (2018)	Alcohol and nonalcohol images in real-life scenarios	Valence, arousal, visual complexity	—	ND	201	n/a	1
17	IAPS	Bradley and Lang (2007)	Photographs covering a wide range of categories	Valence, arousal, distinctiveness, familiarity consequentiality, memorability, meaningfulness	Happiness, surprise, sadness, anger, disgust, fear	ND	1302	32	1
18	MAPS	Goodman et al. (2016)	Scenes common among military populations	Valence, arousal, dominance	—	Own	377	50	1
19	MONS	Schomaker et al. (2017)	Objects in natural scenes	Motivation, valence, arousal, recognizability, aversiveness	—	CC BY NC 4.0	439	35	1
20	NAPS	Marchewka et al. (2014)	People, faces, animals, objects, and landscapes	Valence, arousal, approach/avoidance	—	CC0 1.0 Universal	204	55	+2
21	OASIS	Kurdi et al. (2017)	Humans, animals, objects, and scenes	Valence, arousal	—	ND	822	102	ND
22	Objects_on_Hands	Fernandes et al. (2019)	Objects alone or held in hand	Familiarity, arousal, disgust, valence	Disgust	CC BY 4.0	78	25	2
23	OLAF	Miccoli et al. (2016)	Food in natural scenes (with nonfood emotional control stimuli)	Pleasure, arousal, dominance, food craving	—	ND	424	106	1
24	OLAF_adolescence	Miccoli et al. (2014)	Low-/high-calorie food images	Valence, arousal, dominance, food craving	—	Own	612	139	1
25	PiSCES	Teh et al. (2018)	Line drawings of social contexts and emotional scenes	Valence, social engagement, arousal	—	Own	62	62	1
26	SFIP	Michałowski et al. (2017)	Fear-inducing pictures (social exposure, blood/injection, small animals, angry faces, and neutral images)	Valence, arousal	Fear	ND	1671	58	1
27	SMID	Crone et al. (2018)	Morally and emotionally positive, negative, and neutral content	Valence, arousal, moral wrongness, moral relevance	—	CC BY NC 4.0	1812	28	1

(continued)

Table 1. (continued)

ID	Set name	Reference	Content	Dimensional ratings	Discrete	License	Participants	Rates per stimulus	Countries
28	SNED	Bendall et al. (2025)	Natural environments across seasons	Valence, arousal, familiarity, approach/avoidance	Neutral	ND	?	160	+2
29	ToMenovela	Herbort et al. (2016)	Daily life scenes with two or more individuals	Clarity, perspective, content	Happiness, anger, disgust, fear, sadness, surprise	CC BY NC 4.0	61	66	1
30	WFAIS	Peterson et al. (2019)	Alcohol-related images	Valence, arousal, relation to alcohol, alcohol categorization	.	Own	ND	147	1

Note: Em dashes ("—") indicate that the dimensional ratings or discrete emotions were not reported for that database; Own = database is published on license defined on their website; ND = no data because the researcher did not provide details on the type of license; CC = Creative Commons licenses; for details, see "Data, Analysis Code & Outputs OSF Component" (Behnke et al., 2025b), "SupplementaryData.xlsx" spreadsheet, and "sheet: datasets, column: License," where we link to specific licenses. Countries = +2 indicates that images were tested in more than two countries but the exact number is unknown. SMID = Socio-Moral Image Database; GAPED = Geneva Affective Picture Database; EXCEED = Extreme Climate Event Database; MAPS = Military Affective Picture System; DIRTI = DIsgust-RelaTed-Images; GBPS = Galician Beverage Picture Set; OASIS = Open Affective Standardized Image Set; WFAIS = Wake Forest Alcohol Imagery Set; FRIDa = Food cast Research Image Database; NAPS = Nencki Affective Picture System; OLAF = Open Library of Affective Foods; SFIP = Set of Fear Inducing Pictures; BAPS_Adult = Besançon Affective Picture Set-Adult; SNED = Salford Nature Environments Database; Animal.ID = Animal Images Database; ABPS = Amsterdam Beverage Picture Set; IAPS = International Affective Picture System; MONS = Motivational Objects in Natural Scenes; CaTIS = Crime and Threat Image Set; AFES_C = Anger- and Fear-Eliciting Stimuli for Children; BAPS_Ado = Besançon Affective Picture Set-Adolescents; PiSCES = Pictures with Social Context and Emotional Scenes; CROCUFID = CROss-CUltural Food Image Database; COMPASS = COMPlex Affective Scene Set.

Amsterdam Beverage Picture Set (Pronk et al., 2015), CROss-CUltural Food Image Database (Toet et al., 2019), Open Library of Affective Foods (Miccoli et al., 2016), and Food_Pics (Blechert et al., 2014) provide images of food in different contexts, often contrasting highly palatable and unpalatable food images with nonfood stimuli, whereas the Galician Beverage Picture Set (López-Caneda & Carbia, 2018) and Wake Forest Alcohol Imagery Set (Peterson et al., 2019) focus on alcohol-related content. These sets are valuable for studying affective responses related to eating and drinking behaviors, cravings, and food-related decision-making. Other databases, such as the IAPS (Bradley & Lang, 2007; Lang et al., 2008; Libkuman et al., 2007), COMPlex Affective Scene Set (Weierich et al., 2019), Open Affective Standardized Image Set (Kurdi et al., 2017), and NAPS (Marchewka et al., 2014), which include scenes of people, animals, objects, and landscapes, focus on broader emotional and social contexts. Sets such as the Pictures with Social Context and Emotional Scenes (Teh et al., 2018), Besançon Affective Picture Set-Adult (Szymanska et al., 2019), and ToMenovela (Herbort et al., 2016) explore human interactions in everyday scenarios, often depicting social and emotional situations. Other databases, such as the DIsgust-RelaTed-Images (Haberkamp et al., 2017) and Set of Fear Inducing Pictures (Michałowski et al., 2017), are designed to elicit discrete emotions such as disgust (e.g., images showing food, body fluids/excretions, injuries/infections, death, and lack of hygiene) or fear (e.g., images showing phobia-related content, including blood, injections, spiders, social exposure, or angry faces). The Animal Images Database (Possidónio et al., 2019) and GAPED (Dan-Glauser & Scherer, 2011) present a variety of animal images, some of which (e.g., snakes and spiders) are commonly used to elicit fear or disgust. Despite offering diverse, valuable content for affective research, existing image databases still have notable limitations.

First, although many studies confirm their effectiveness in eliciting affective responses (for reviews, see Behnke et al., 2022; Joseph et al., 2020; Siegel et al., 2018), the absolute effects are small (e.g., 2.72 before manipulation to 3.01 after manipulation on the scale for positive affect with ratings from 1 to 5; Diener et al., 2023). Second, the finite assortment of images and videos in specific thematic categories may result in participants encountering the same materials more than once, particularly in longitudinal studies, potentially leading to a repetition bias. Third, the technical quality (e.g., low resolution) can be inconsistent across stimuli. Given that the databases feature stimuli created in earlier times (e.g., IAPS presenting images from the 1980s and 1990s), they may not meet current technical standards or the quality of stimuli participants are used to. This issue is particularly relevant when media quality varies significantly between emotions (e.g., eliciting joy with images from one database and sadness from another). Fourth, depicted people and objects from earlier times might provoke unintended emotional responses because of changes in societal norms and fashions (e.g., eliciting nostalgia through images of outdated fashion styles). Fifth, there is an asymmetry in the databases; they often contain a larger volume of negative than positive stimuli, and the negative content is more varied (Dan-Glauser & Scherer, 2011). Furthermore, data sets are often published under different, sometimes outdated licenses with varying copyright restrictions, which can limit or prohibit their use in certain ways. For instance, some databases are stored locally on university servers and are not maintained once the authors leave academia (e.g., the official IAPS page no longer exists). Finally, most available stimuli target Western contexts, hindering cross-cultural research—yet identical images can elicit different emotional responses elsewhere (Berezina et al., 2024).

The Promise of Generative AI

Generative AI has rapidly advanced in its ability to produce both text and images, enabling the creation of highly customized, expressive, and context-aware content. One of the core strengths of generative AI lies in its ability to generate rich, context-sensitive descriptions that capture subtle emotional, cultural, and situational nuances. When paired with image-generation models, it is possible to create affective stimuli that are not only visually compelling but also semantically aligned with specific psychological definitions. This synergy enables the development of more valid, tailored emotional content than is typically possible with static, precurated image sets.

Thus, the limitations of the existing databases of affective stimuli could be addressed by leveraging generative AI. First, we believe that images generated and tailored to specific contexts will enhance the effectiveness of affect induction by increasing personal and cultural relevance, which may, in turn, elicit stronger emotional responses than standardized stimuli from existing affective databases. Second, generative models can produce highly specific affective stimuli on demand, potentially resolving the issue of limited variety in existing databases. For instance, once researchers have a detailed idea or concept for the image, they might use a generative-AI model (e.g., ChatGPT) to craft a precise and descriptive prompt that outlines the key elements, composition, and style of the image. Next, the prompt would serve as input into another (or the same) model (e.g., MidJourney) to generate the image. Thus, novel, goal-aligned images prevent repeated exposure, vital for longitudinal studies. Third, AI-generated images can be produced in high resolution, meeting modern technical standards and matching the up-to-date quality of media. Researchers using AI can also create contemporary stimuli that are more culturally relevant, reducing the likelihood of unintended responses. Moreover, using AI, researchers can help balance the asymmetry between negative and positive stimuli. AI also allows for iterative improvement in the generated content, so the less evocative (or offensive) images can be adjusted accordingly. Images can be released under open-access licenses from the outset, helping to avoid these legal and practical barriers and facilitate broader reuse, transparency, and reproducibility in affective science. To ensure long-term availability, we plan to store the generated library in established open-access repositories, which offer persistent storage and stable access independent of individual researchers' affiliations. Finally, generative-AI models can be used to adapt the stimuli to diverse social and cultural contexts, developmental stages, and sex, among others, enabling a wide variety of comparisons. The flexibility of AI-generated content combined with its availability on open-access platforms, could also resolve issues around licensing and accessibility, providing researchers with a sustainable, adaptable, and legally compliant resource.

The Present Study

We leveraged generative AI to create the pipeline and build a publicly accessible database of affect-induction images, referred to as the Library of AI-Generated Affective Images (LAI-GAI). By using generative AI, we sought to address the current limitations of existing databases and generated detailed, adaptable descriptions and images (847 of them annotated) across 12 discrete emotion categories. The chosen images were adapted to represent African, Arabic, Asian, Indian, South/Central American, and European/North American contexts. Thus, each image would have six different cultural versions. The chosen images were also adapted to depict male and female individuals and independently, individuals at childhood, adulthood, and older age. We validated the library through six studies involving 2,470 participants from 58 countries, who rated five types of images: (a) images from existing affective databases, (b) AI-generated images without cultural adjustments, (c) AI-generated images adjusted to specific cultural contexts, (d) AI-generated images adjusted by sex (male, female), and (e) AI-generated images adjusted by age group (childhood, adulthood, older age). In Study 1, participants (n = 589) evaluated images from existing affective databases and AI-generated images (images generated based on the descriptions of stimuli from existing affective databases). In Study 2, participants (n = 296) evaluated AI-generated images (images generated based on our ideas = 74%, images generated based on the descriptions of stimuli from existing affective databases = 26%). In Studies 3 and 4, participants (n = 726 and n = 275, respectively) evaluated selected AI-generated images from Studies 1 and 2 adjusted to different cultural contexts. In Study 5, participants (n = 224) evaluated sex-adjusted variants (male vs. female) of AI-generated images to test whether demographic tailoring preserves affective responses. In Study 6, participants (n = 360) evaluated age-adjusted variants (childhood, adulthood, older age) to assess affect equivalence across developmental depictions.

Using collected data, we aimed to benchmark the performance of generated images and to estimate the smallest effect size of interest (SESOI) for affective studies that use affective stimuli, defined as the smallest subjectively experienced difference in reactions to two stimuli. Thus, in Studies 1 and 3, we included direct comparisons between the images using adapted anchor items (Kamper et al., 2009). In Study 1, we included the images from existing affective databases and their AI-generated counterparts to benchmark the effectiveness of AI-generated images against their original counterparts. In Study 3, we included the culturally adjusted and unadjusted stimuli to benchmark the effectiveness of matched images against their unmatched counterparts.

We hypothesized the following:

> *Hypothesis 1:* Generated images would be at least as effective as the images from existing affective databases in affect elicitation.
>
> *Hypothesis 2:* Images adjusted to participants' cultural context would elicit stronger affective responses than their unadjusted counterparts.
>
> *Hypothesis 3:* Sex- and age-adjusted variants would produce affect responses comparable with their corresponding base images.

Because we did not recruit minors as raters, the sex/age studies did not use a matched-unmatched design. Instead, they served as a generalizability test, assessing whether demographic variants could be generated while preserving affective impact. We aimed to deliver a transparent pipeline for creating robust, validated affect-induction images freely available for future research.

Method

Materials: creating the image database

We created the LAI-GAI database with the following steps: (a) collecting evocative images from available databases, (b) generating descriptions for each image,

(c) generating new evocative images from these descriptions, (d) grouping images, and (e) creating culture-, sex- and age-adjusted variants.

In the first step, we leveraged existing data sets to create descriptions of images. We explored the data sets included in the KAPODI—the searchable database of free emotional-stimuli sets (Diconne et al., 2022)—and extracted images that elicited strong affective reactions in data-set-validation studies (for the list of studies, see Table 1). In total, we selected 284 images that elicited strong affective reactions and 78 neutral images. For details on how we selected images for each database and their ratings, see the Supplementary Information in the Supplemental Material available online.

In the second step, we used large language models (LLMs), mainly ChatGPT 4o (OpenAI, 2024a), with the Mj prompt Generator V6 chatbot to describe the images using the following prompt: "Provide me a prompt for the picture." All descriptions of the images are included in the "Data, Analysis Code & Outputs OSF Component" (Behnke et al., 2025b), "SupplementaryData.xlsx" spreadsheet, sheet: "generated & not_generated," column: "Prompt GPT."

In the third step, we used text-to-image models, such as Midjourney (Midjourney, Inc., 2024) and Freepik (Freepik, 2024), to generate new pictures based on the descriptions of the images using the following prompt: "/imagine[prompt generated in the previous step]"; we sometimes had to adjust the generated images to ensure the image quality and its fit to the definition of the emotion. Corrections were made using the *Vary (Region)* function, in which after selecting a part of the image, the researchers entered what should be changed, which in many cases included correcting hands and faces. Often, it was also necessary to remove an element or change its position to make it look more natural, such as removing the second medal from the winner's neck. Using this procedure, we generated 310 images. We also brainstormed ideas for other images that would elicit strong emotions. We asked ChatGPT 4o to describe each picture using our idea with this prompt: "I will provide you with a description of an emotion. I want you to give me prompts that will create photorealistic images reflecting that emotion." Using this procedure, we generated an additional 136 images in Midjourney and Freepik. In sum, we generated 446 images.

In the fourth step, we grouped the pictures into categories based on definitions of discrete emotions (Behnke et al., 2022; Ekman & Cordaro, 2011). We aimed to generate at least 20 pictures for 12 discrete emotional categories: amusement, awe, anger, attachment love, craving, disgust, excitement, fear, joy, neutral, nurturant love, and sadness. These categories were selected to cover a broad range of affective experiences, including basic emotions, neutral states, and an expanded representation of positive emotions. Among other typical choices, we chose to exclude surprise because we believe static images are unlikely to elicit this emotion effectively. We also omitted sexual desire because of potential ethical and cultural concerns associated with distributing such stimuli online to a global audience. After grouping the images, the research team evaluated the image quality and its alignment with the definition of a given emotion. Based on these decisions, we chose 20 images from each emotion category for further testing and validating our database. All images used and unused for database validation are presented in the "Images OSF Component" (Behnke et al., 2025c) and the "Data, Analysis Code & Outputs OSF Component" (Behnke et al., 2025b), "SupplementaryData.xlsx" spreadsheet, sheet: "generated." We also released a web application for browsing LAI-GAI. For a stable pointer to the current address, see the OSF project page (Behnke et al., 2025a).

In the fifth step, we selected six images out of the 20 categorized images. M. Behnke and M. Kłoskowski subjectively chose six images, evaluating their potential for cultural adjustments and emotion elicitation. We adjusted them to different cultural contexts, resulting in four similar pictures, each adjusted for Asian, African, Latin American, and European/North American contexts. Because most images from existing affective databases (83%) represented the European/North American context, these images also typically served as the baseline for other cultural adaptations. To adjust the pictures, we used the following prompt:

> Hello, my task is to adapt the photos to three different cultures: African, Asian, and Central/South American culture. I will upload photos, and you will create prompts for AI that will be customized under these 3 cultures. Create 3 prompts for each culture (9 prompts total).

Using this procedure, we generated 216 additional images.

At the end of the fifth step, we shared the images with researchers from the target cultural contexts to ensure they are appropriate for those contexts. We asked the reviewers to evaluate whether the image was well adjusted to their culture and to provide feedback when the image needed some additional adjustments. Based on the reviewers' feedback, we corrected 81 images (34%). For further testing and validation, we chose 20 from each emotion category (five images across four cultural contexts).

After additional input from outside experts, we refined the "Asian" category by adding two additional contexts—India-specific and Arabic-speaking regions—and adapted the 60 images from Study 3 accordingly to produce 120 new images. Incorporating expert feedback, we then revised 45 images (37.5%). We also expanded our tool set to include newer text-to-image models suitable for affect-elicitation tasks, including Imagen 4 (Google Cloud, 2025), OpenAI's GPT/DALL·E 3 image generation (OpenAI, 2024b), Seedream (Team Seedream et al., 2025), and FLUX.1 (Black Forest Labs et al., 2025). The new models also allowed us to use a reference image together with an additional prompt, which helped match the affective content while changing contextual cues.

In the sixth step, we created sex-adjusted variants of selected AI-generated images. For each base image, we produced male and female versions while holding scene semantics, composition, and the target emotion constant. We used an LLM to rewrite the prompt to depict the opposite sex of the original image, generated the variant, and iterated until nondemographic features matched the base image. In some cases, to achieve close similarity, we provided the base image as a reference. We excluded two emotion categories—craving and awe—because the canonical images did not depict humans. To enable sex adjustment where needed, we also introduced new human-depicting images in the amusement, anger, and disgust categories. These procedures yielded a set of 51 sex-matched pairs that were advanced to testing.

In the seventh step, we created age-adjusted variants to depict minors, middle-aged adults, and older adults while holding scene semantics, composition, and the target emotion constant. We used an LLM to rewrite the prompt to the target age group, generated the variants, and iterated until nonage features matched the base image. Again, in some cases, to achieve close similarity, we provided the base image as a reference. For attachment love, every image in this category necessarily depicts a minor. Therefore, we did not create a separate "minors" variant. Instead, we generated only the middle-aged-adult-minor and older-adult-minor variants for this category. This procedure yielded 45 age-matched triplets (all three age groups) and five age-matched pairs (two age groups, i.e., for attachment love), which were advanced to testing.

To openly document the disagreement between the researchers (Coles et al., 2024), we provide the evaluations of the final set of images, in which each researcher was allowed to state their opinion on each image ("Data, Analysis Code & Outputs OSF Component," "SupplementaryData.xlsx," sheet: "authors_disagreements"). For an overview of the LAI-GAI generation procedure, see Figure 1.

Procedure

Initially, participants received an overview of the study's structure. All six studies had a very similar structure and were run in English. Participants were recruited via Prolific, where they chose to participate in the study aimed at developing a new image library for emotion elicitation. They were informed that the images might elicit strong positive and negative responses and were advised to report any discomfort immediately and pause the experiment. After consenting to their participation, participants read the definitions of each emotional dimension (for details on emotion definitions, see Supplementary Information in the Supplemental Material), and before seeing the images, they were asked to report their baseline emotions. Next, participants watched a series of images. Images were displayed full size for 4 s each, followed by rating scales on a new screen to the right and a smaller version of the image to the left. Before each image, the fixation cross (+) was displayed in the center of the screen for 1,000 ms, followed by a blank screen for 1,000 ms. The smaller version of the image remained visible until all ratings were completed, and then the next image appeared.

In Study 1, participants rated 12 images from existing affective databases alongside 12 AI-generated images created based on textual descriptions of those stimuli from existing affective databases. In Study 2, participants rated 36 only-AI-generated images, the majority of which (74%) were based on original image concepts developed by the research team and the remaining (26%) derived from descriptions of stimuli from existing affective databases. In Studies 3 and 4, participants rated 12 and 15 culturally adapted AI-generated images (respectively) selected from Studies 1 and 2 that matched their region and 12 and 15 that did not match their region. In Study 5, participants rated 17 images depicting females and 17 images depicting males. In Study 6, participants rated nine images depicting minors, 10 images depicting adults, and 10 images depicting older adults to test whether demographic tailoring preserved affective responses. Thus, each participant viewed 24 images in Study 1, 36 images in Study 2, 24 images in Study 3, 30 images in Study 4, 34 images in Study 5, and 29 images in Study 6. The order of images was randomized, and each image was presented individually.

In Studies 1 and 3, once participants rated all images, they evaluated pairs of matched images—AI generated and those from existing data sets in Study 1 and images from their and other cultural contexts in Study 3—and were asked to compare the reactions elicited by them. In this phase, image pairs were displayed side by side, and the left/right assignment of each image was counterbalanced across trials.

At the end of each study, all participants were fully informed about the study's purpose and completed a newly developed scale assessing their attitudes toward emotional AI—specifically, its ability to detect, understand,

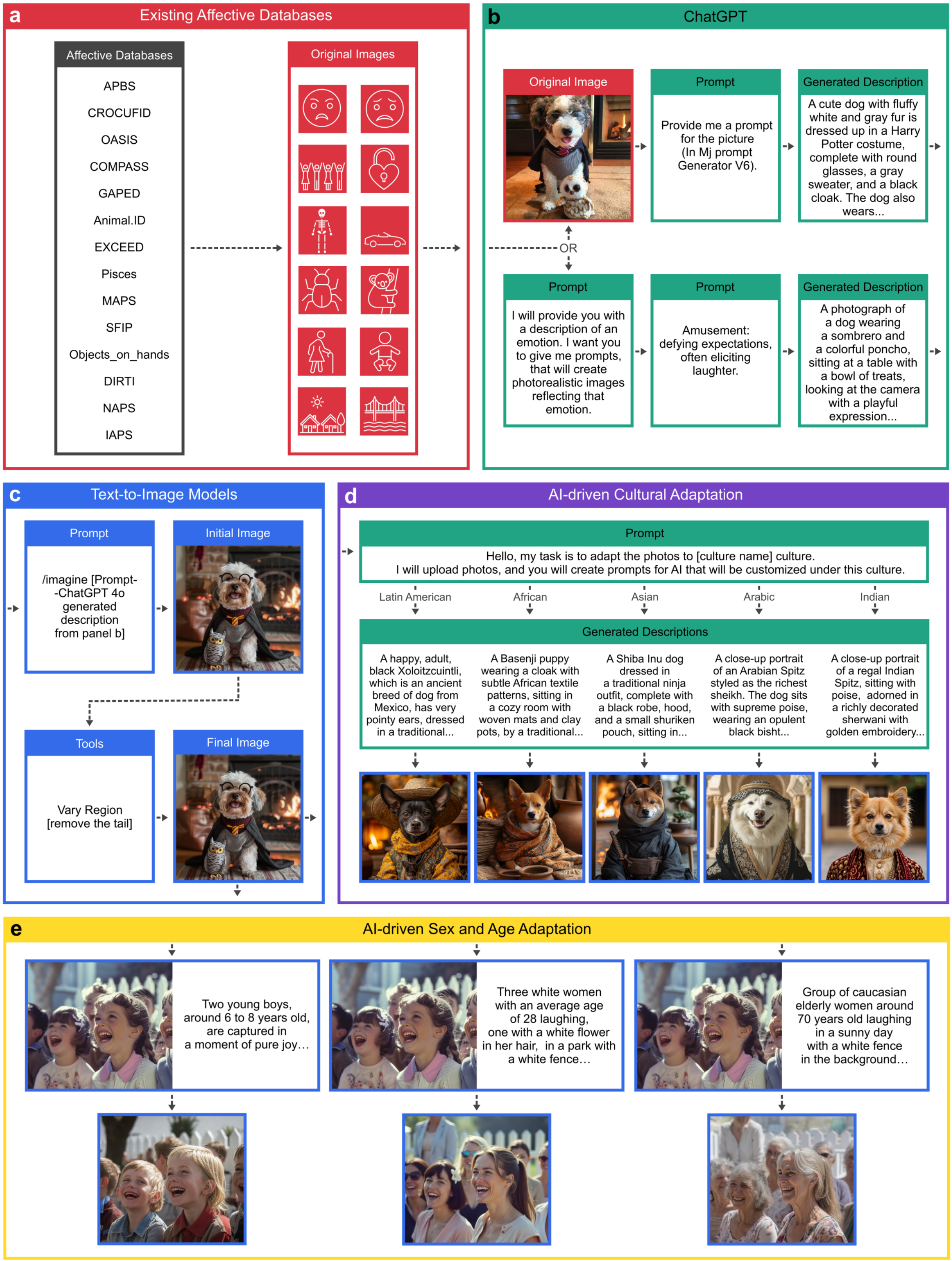
a
Existing Affective Databases
Affective Databases
APBS
CROCUFID
OASIS
COMPASS
GAPED
Animal.ID
EXCEED
Pisces
MAPS
SFIP
Objects_on_hands
DIRTI
NAPS
IAPS
Original Images
b
ChatGPT
Original Image
Prompt
Provide me a prompt for the picture (In Mj prompt Generator V6).
Generated Description
A cute dog with fluffy white and gray fur is dressed up in a Harry Potter costume, complete with round glasses, a gray sweater, and a black cloak. The dog also wears...
OR
Prompt
I will provide you with a description of an emotion. I want you to give me prompts, that will create photorealistic images reflecting that emotion.
Prompt
Amusement: defying expectations, often eliciting laughter.
Generated Description
A photograph of a dog wearing a sombrero and a colorful poncho, sitting at a table with a bowl of treats, looking at the camera with a playful expression...
c
Text-to-Image Models
Prompt
/imagine [Prompt--ChatGPT 4o generated description from panel b]
Initial Image
Tools
Vary Region [remove the tail]
Final Image
d
AI-driven Cultural Adaptation
Prompt
Hello, my task is to adapt the photos to [culture name] culture.
I will upload photos, and you will create prompts for AI that will be customized under this culture.
Latin American
African
Asian
Arabic
Indian
Generated Descriptions
A happy, adult, black Xoloitzcuintli, which is an ancient breed of dog from Mexico, has very pointy ears, dressed in a traditional...
A Basenji puppy wearing a cloak with subtle African textile patterns, sitting in a cozy room with woven mats and clay pots, by a traditional...
A Shiba Inu dog dressed in a traditional ninja outfit, complete with a black robe, hood, and a small shuriken pouch, sitting in...
A close-up portrait of an Arabian Spitz styled as the richest sheikh. The dog sits with supreme poise, wearing an opulent black bisht...
A close-up portrait of a regal Indian Spitz, sitting with poise, adorned in a richly decorated sherwani with golden embroidery...
e
AI-driven Sex and Age Adaptation
Two young boys, around 6 to 8 years old, are captured in a moment of pure joy…
Three white women with an average age of 28 laughing, one with a white flower in her hair, in a park with a white fence…
Group of caucasian elderly women around 70 years old laughing in a sunny day with a white fence in the background…

Fig. 1. Images-generation pipeline. The workflow consisted of five main phases: (a) selection of evocative images from existing affective databases; (b) generation of textual descriptions for each image using large language models, such as ChatGPT 4o; (c) creation of new images based on these descriptions with text-to-image models (e.g., Midjourney); (d) AI-driven cultural adaptation of images using both ChatGPT 4o and Midjourney or Freepik and other text-to-image models, such as Imagen 4, OpenAI's GPT/DALL·E 3 image generation, Seedream, and FLUX.1; (e) AI-driven sex and age adaptation of images using multiple models. AI = artificial intelligence.

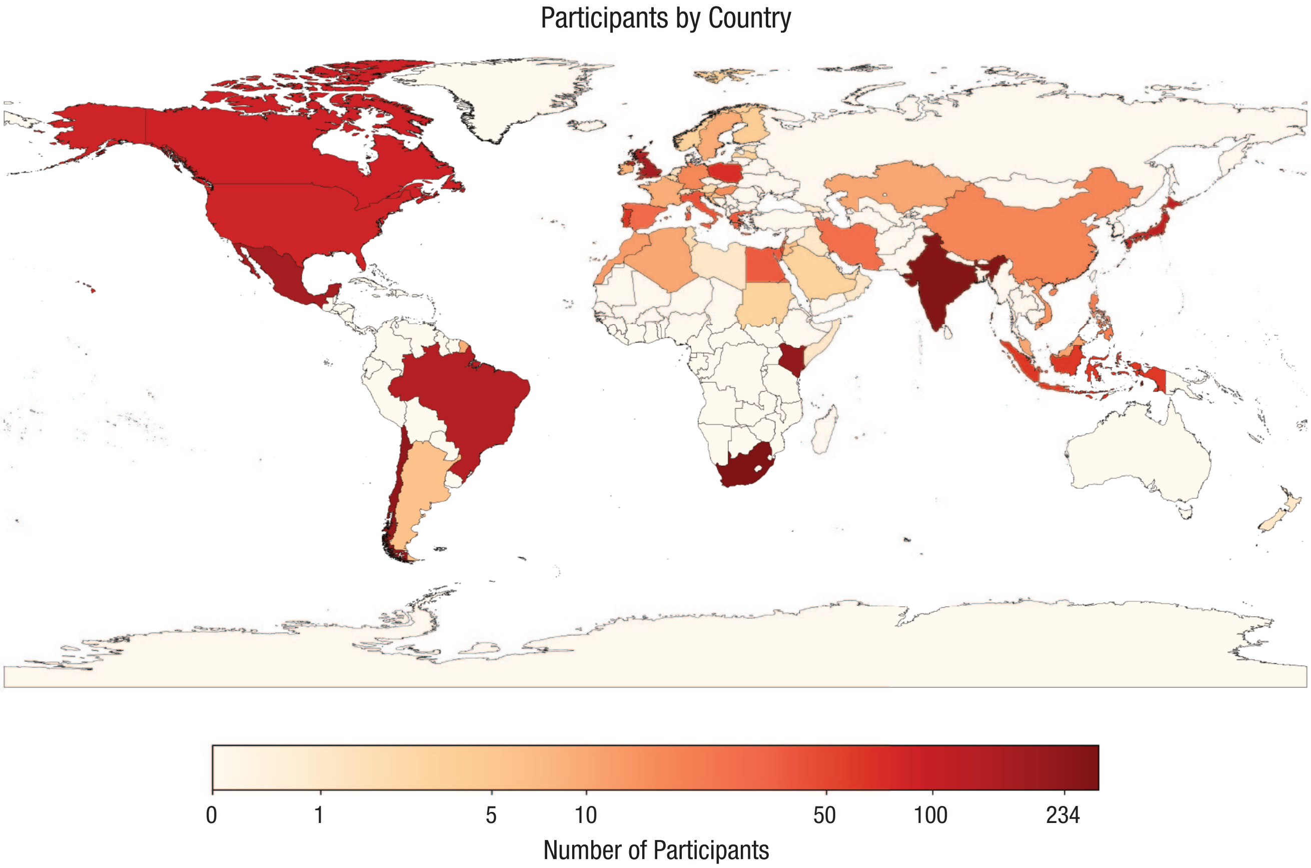

Fig. 2. Participants by country. Data were collected from 2,470 participants in 58 countries. In each of the three studies, participants came, on average, from a similar number of different countries (between 26 and 38). Darker shades of red denote larger country-specific sample sizes. The color scale uses $\log_e(1+n)$ internally to visualize variation across countries better (for the map in colors without the log transformation, see the Supplementary Information in the Supplemental Material available online).

and enhance human emotions (see Supplementary Information in the Supplemental Material). The study lasted approximately 25 to 40 min (Study 1: *Mdn* = 30.79; Study 2: *Mdn* = 35.88; Study 3: *Mdn* = 31.17; Study 4: *Mdn* = 32.23; Study 5: *Mdn* = 33.05; Study 6: *Mdn* = 31.47). The names of images used in each study are presented in the "Data, Analysis Code & Outputs OSF Component" (Behnke et al., 2025b), "SupplementaryData.xlsx" spreadsheet, sheet: "generated."

Participants

We used Prolific's gender-balance option to ensure an even distribution across genders. Study 1's (n = 589) sample included 48.56% female participants and 50.08% male participants (age: M = 33.04 years, SD = 10.38, range = 18–77) from 38 countries. Study 2's (n = 296) sample included 46.96% female participants and 50.68% male participants (age: M = 32.83 years, SD = 11.41, range = 18–77) from 33 countries. Study 3's (n = 726) sample included 46.97% female participants and 51.93% male participants (age: M = 32.12 years, SD = 10.01, range = 19–73) from 36 countries. Study 4's (n = 275) sample included 50.18% female participants and 49.45% male participants (age: M = 31.08 years, SD = 7.84, range = 18–65) from 26 countries. Study 5's (n = 224) sample included 49.55% female participants and 49.55% male participants (age: M = 34.78 years, SD = 11.16, range = 19–74) from 30 countries. Study 6's (n = 360) sample included 49.72% female participants and 50.00% male participants (age: M = 35.14 years, SD = 11.73, range = 19–77) from 32 countries. For the countries represented, see Figure 2.

In Study 3, we split the target sample into four regional groups—Africa, Asia, Europe/North America, and South/Central America—and selected specific countries within each region for data collection. Study 4 included Indian participants and participants of Arabic-speaking nationalities (Algeria, Bahrain, Egypt, Iran, Iraq, Jordan, Kuwait, Lebanon, Libya, Mauritania, Morocco, Oman, Qatar, Saudi Arabia, Somalia, Sudan, Tunisia, United Arab Emirates, Yemen).

Sample-size determination. To determine the sample size for our studies, we aimed for 70 ratings (participants) per image. In anticipation of potential data loss and expecting 10% to 20% of the sample to be reduced because of careless responding, we recruited additional participants up to *n* = 660 in Study 1, *n* = 330 in Study 2, *n* = 710 in Study 3, *n* = 300 in Study 4, *n* = 230 in Study 5, and *n* = 370 in Study 6. We expected a reduction in the sample based on our experience with similar projects (Coles et al., 2022). The number of ratings per image was based on the study budget (allowing for 2,500 participants in total), the number of images that could be evaluated within 30 min of the study, and the median of the rates per image in databases presented in Table 1 (*Mdn* = 64).

Furthermore, the sensitivity analysis using the *simr* (Green & MacLeod, 2016) package in R showed that our sample allowed us to detect the range of *d*s = 0.16 to 0.31, depending on the outcome measure with α = .05 and power (1 − β) = 0.95 for the difference between the AI-generated images and images from existing affective databases; the range of *d*s = 0.06 to 0.08 for the difference between images matched and unmatched to the cultural contexts; and the range of *d*s = 0.29–0.37 for the difference between sex variants. For age variants, convergence issues and inconsistent results across alternative model specifications (including simplified, single-outcome versions) precluded reliable simulation-based power estimates, so we do not report effects.

Exclusion. We excluded participants identified as careless responders based on the following criteria: (a) selecting answers other than "7" for the photo with instructions to "Please select '7' for all items for this photo to show that you are paying attention" and (b) answering the last item on the last study page—"In your honest opinion, should we use your data in our analyses in this study?"—with "No, I responded carelessly" (Meade & Craig, 2012). We decided to classify the participants as outliers if they skipped two or more answers other than "7" for the attention-check image, resulting in excluding 21 participants in Study 1, six in Study 2, 28 in Study 3, 11 in Study 4, 14 in Study 5, and 12 in Study 6. Furthermore, we excluded participants if they admitted to responding carelessly, resulting in the exclusion of an additional three participants in Study 1, six in Study 2, 11 in Study 3, two in Study 4, one in Study 5, and two in Study 6.

Ethics information

The Ethical Committee for Research Projects at the Faculty of Psychology and Cognitive Science at Adam Mickiewicz University (May 11, 2024) approved the study. Each participant provided written informed consent. Participants received four GBP for completing each study. The project was also consulted and approved by the Legal Advisors' Office at Adam Mickiewicz University.

Measures

In addition to the measures reported in the main text, we also collected data on participants' attitudes toward AI (for details, see the Supplementary Information in the Supplemental Material).

Emotions. Participants used single items to report the intensity of discrete emotions they experienced while watching the stimuli. Participants rated the intensity of amusement, awe, anger, attachment love, craving, disgust, excitement, fear, joy, neutral, nurturant love, and sadness. Furthermore, participants used two independent unipolar items to report the intensity of valence (positive and negative), arousal (arousing and calming), and motivation (motivating to approach the situation and motivating to avoid the situation). All responses were measured on a scale from 1 (*not at all*) to 7 (*extremely*).

Comparison. In Studies 1 and 3, participants compared pictures on targeted emotion, valence (positive and negative), and arousal (arousing and calming) choosing one of the following options: "Photo no. 1 compared to Photo no. 2 elicited/was: much less [emotion]"; "slightly less [emotion]"; "no difference, slightly more [emotion]"; or "much more [emotion]." We adapted this anchor item for affective measures from a previous study that aimed to establish the SESOI for the difference in effect across 2 or 5 days (Anvari & Lakens, 2021).

Analysis

We preregistered our analysis for Studies 1 through 3 (for the OSF preregistration, see https://doi.org/10.17605/OSF.IO/72RJ3). The deviations for the preregistration plan are explained in the Supplementary Information in the Supplemental Material. We did not preregister the analysis for Studies 4 through 6 because they were a direct continuation of the previous analysis. First, we calculated the means and standard deviations for the targeted emotional measures. Next, we examined whether the images elicited targeted affective responses. We explored whether the targeted emotion received the highest ratings.

Furthermore (not preregistered), to support future use of the images in our library, we calculated a "success index" based on the approach introduced by Gross and Levenson (1995). This index combines two key components: the intensity and discreteness of the emotional response elicited by each image. We operationalized intensity as the mean rating for the target emotion. Discreteness was defined using an idiographic hit rate, calculated as the percentage of participants who rated the target emotion at least 1 point higher than any of the six nontarget emotions. These two components were each standardized (*z*-scored) within the emotion category and then summed to form a composite success index for each image. This index provides a practical metric to identify the most effective images for eliciting specific emotional responses.

Benchmarking the library. Next, we compared the AI-generated images and images from existing affective databases using multivariate multilevel modeling in R (R Core Team, 2025) using R packages *brms* (Bürkner, 2017) and *lme4* (Bates et al., 2015). The data from Study 1 were conceptualized as a two-level structure in which repeated emotional ratings were nested in individuals. Images were grouped into two categories: AI-generated images and images from existing affective databases. We focused our analyses on the extent to which the stimuli elicited the targeted affective states. For example, we focus our analyses on whether sadness is elicited by images designed to elicit sadness—not images designed to elicit anger, happiness, or disgust. Thus, we fitted a multivariate multilevel Bayesian model using the *brms* package (Bürkner, 2017) to simultaneously examine differences in targeted discrete emotion, targeted valence, targeted arousal, and targeted motivation as a function of image type (AI-generated vs. images from existing affective databases). We ran four No-U-Turn Sampler chains (2,000 iterations each, 1,000 warm-up) and inspected trace and pairs plots. We also used posterior predictive checks based on 100 replicated data sets. We based our inferences primarily on 95% credible intervals (CrIs); thus, if the entire 95% CrI for the coefficient excludes 0, we interpreted this as evidence that image type influenced the corresponding emotional measure. Each measure was modeled with random intercepts for participants and images, random slopes for image type across both participants and images, and correlated residuals to account for interdependencies among the four outcomes. By default, *brms* allows these slopes to be correlated with the random intercept at each grouping level; thus, in our final model, random intercepts and slopes are freely estimated as correlated, providing a more flexible approach to capturing between-participants and between-images variability. The code for the analysis is available in the "Data, Analysis Code & Outputs OSF Component" (Behnke et al., 2025b), "files MMA_model_SX.Rmd."

Furthermore, we compared the images adjusted to different cultural contexts. Using a similar approach, we conceptualized the data from Studies 3 and 4 as a two-level structure in which repeated emotional ratings were nested within individuals. Images were grouped into two categories: matched and unmatched cultural context. To test the difference, we used targeted discrete-emotion measures: valence, arousal, and motivation. We applied the same analytic strategy to the sex-adjusted and age-adjusted variants.

We determined the targeted valence, arousal, and motivation based on data from another study in which participants assigned the film clips to discrete-emotion categories and evaluated them using affective dimensions (e.g., valence, arousal, approach; Cowen & Keltner, 2017). Thus, for our analysis, we used (a) positive valence for amusement, awe, craving, excitement, joy, attachment love, neutral, and nurturant love; (b) negative valence for anger, disgust, fear, and sadness; (c) arousal for all categories; (d) approach motivation for amusement, attachment love, awe, craving, excitement, joy, neutral, and nurturant love; and (e) avoidance motivation for anger, disgust, fear, and sadness.

As an exploratory (and not preregistered) analysis, we used direct comparison items to calculate the percentage of responses indicating whether the AI- or non-AI-generated (or culturally matched vs. unmatched) images elicited *much weaker*, *a little weaker*, *the same*, *a little stronger*, or *much stronger* affective responses. Next, we also performed two-sided binomial tests on the direct comparison items. Ratings of 1 (much weaker) and 2 (a little weaker) were grouped as "less," and ratings of 4 (a little stronger) and 5 (much stronger) were grouped as "more;" ties (ratings of 3) were excluded from directional tests. For each comparison item, we tested whether the proportion of "more" responses differed from 50% using an exact, two-sided binomial test, thus assessing whether participants showed a significant imbalance—either toward more or less—in judging which image elicited the stronger affective response.

Calculating SESOI. Affective scientists are often interested in effects that are large enough to be subjectively experienced and deemed meaningful by individuals. Thus, we defined the SESOI as the smallest change in an outcome measure that individuals consider to be meaningful enough to rate themselves as feeling different, following Anvari and Lakens's (2021) methodology. We calculated the SESOI for the difference between affective ratings for two pictures. First, the pairs of images were subcategorized based on responses to the anchor item. Because there was a symmetry between the subcategories (e.g., "a little less" and "a little more"), we rescored them so the categories would indicate the difference between the AI-generated image and non-AI-generated image or culturally matched or unmatched images (e.g., the AI-generated image elicited much less amusement than the non-AI-generated image).

Next, we calculated the difference score by subtracting the second image rating from the first image rating so that the positive value would indicate the stronger affective reaction elicited by the first image. Finally, the mean difference score for categories a little less and a little more on the outcome of interest was taken as the estimate of the smallest subjectively experienced difference. As recommended, we present SESOI as raw differences and standardized effect sizes (Anvari & Lakens, 2021). For standardized effect sizes for paired observations, we used Cohen's d_z because it takes the correlation between observations into account.

Results

The raw data, means, and standard deviations for affective measures for all images are reported in "Data, Analysis Code & Outputs OSF Component" (Behnke et al., 2025b). The mean levels of elicited responses are presented in Figure 3 (for radar plot for each image, see "Data, Analysis Code & Outputs OSF Component," Behnke et al., 2025b, file: "radar_by_category_with_image.zip"). For each image (AI-generated and images from existing affective databases), we had an a priori expectation regarding the target emotion it was intended

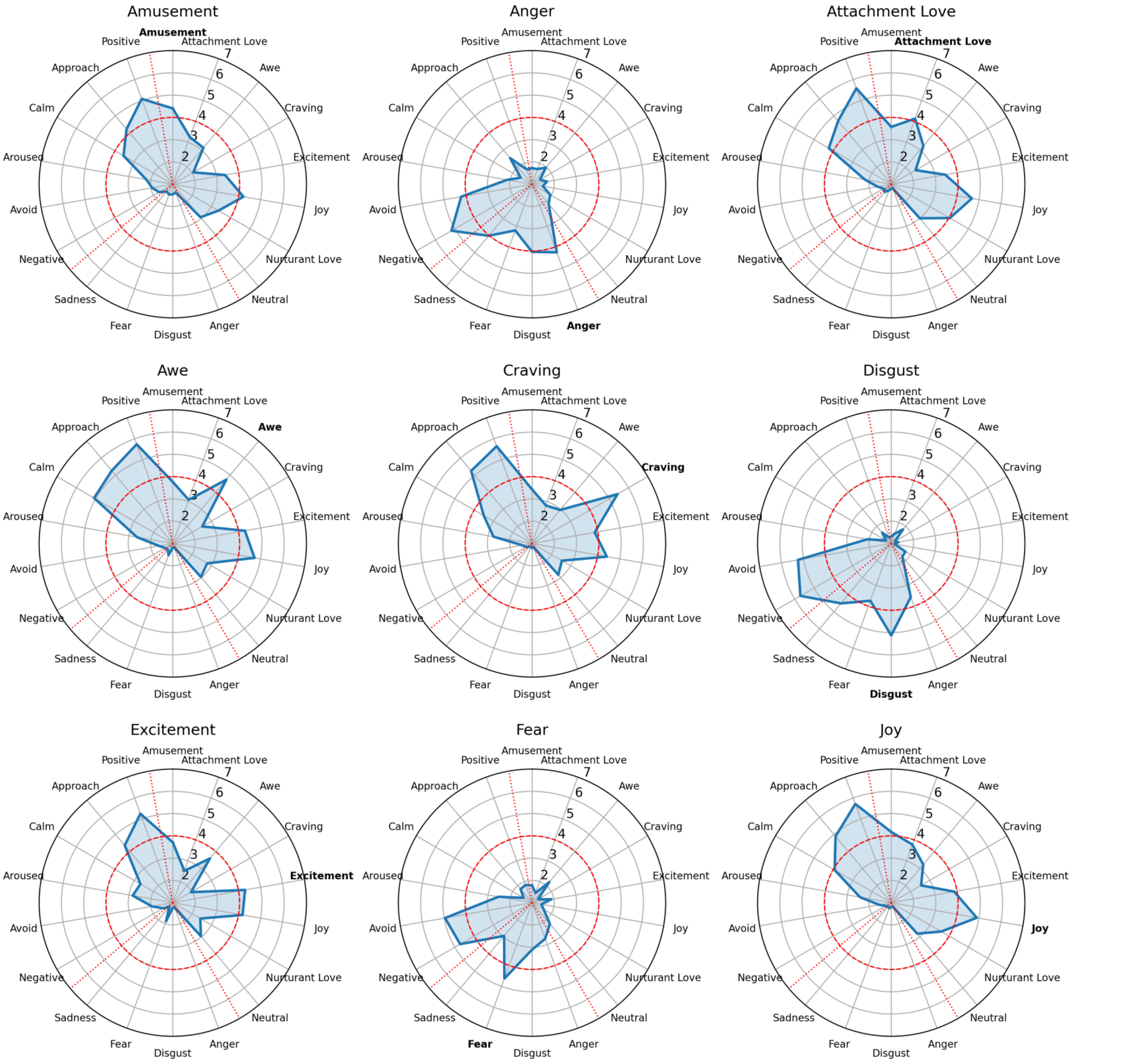

Fig. 3. *(continued on next page)*

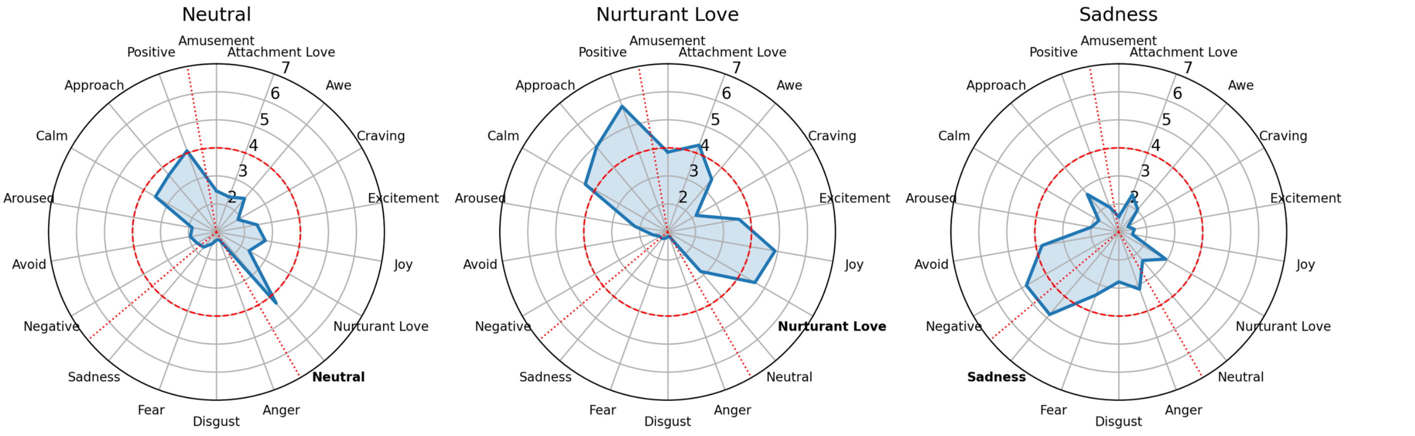

Fig. 3. Affective ratings. Radar charts visualize the strength of affective responses. Mean levels of elicited reactions are aggregated by targeted emotional category. Each spoke on the radar chart represents one of the rated emotions (e.g., joy, fear, surprise). Along each spoke, the distance from the center shows the average intensity with which participants rated that emotion in response to an image set. The red dashed concentric rings represent the midpoint of the scale, making it easier to interpret whether a rating was above or below the middle point of the scale from *not at all* to *extremely*.

to elicit. As a preliminary check, we examined whether the anticipated target emotion received, on average, a higher rating than the other nontarget emotions. Of the 847 tested AI-generated images, 544 (64%) received the highest ratings for their intended targeted emotion. Likewise, of the 96 tested images from existing affective databases, 52 (54%) received the highest ratings for targeted emotion. Thus, for both images from existing affective databases and AI-generated images, in most cases, the intended emotion aligned with participants' ratings. A complete list of success index values for each image, along with intensity and discreteness scores, is provided in the "Data, Analysis Code & Outputs OSF Component" (Behnke et al., 2025b), file: "image_emotion_means_S123456.csv." This resource can be used to select the most effective stimuli for future studies.

Benchmarking the library

For the descriptive statistics for univariate comparisons, see Tables 2 to 5. For detailed results for both analyses, see the "Data, Analysis Code & Outputs OSF Component" (Behnke et al., 2025b), files: "MMA_model_SX .Rmd."

The model for AI-generated images versus images from existing affective databases converged successfully (R^2: range = .32–.46). Although some parameters initially showed poor mixing (maximum R-hat = 1.34; bulk effective sample size [ESS] as low as 41), our primary fixed-effect estimates converged cleanly after raising the sampler's target acceptance probability to 99% (all fixed-effect R-hat = 1.00; bulk ESS ≥ 692). Posterior predictive checks showed that the model consistently reproduced the observed means and variances for all four outcomes (Bayesian *p* values = .49 to .51), and inspection of trace plots did not reveal any sampling issues. A summary of the fixed effects showed that on average, AI-generated images produced stronger reactions than images from existing affective databases for arousal (b = 0.12, 95% CrI = [0.01, 0.22]). In contrast, targeted emotion, targeted valence, and targeted motivation did not differ across image types (b = 0.06, 95% CrI = [–0.17, 0.28]; b = 0.06, 95% CrI = [–0.21, 0.31]; b = 0.12, 95% CrI = [–0.14, 0.38], respectively) because CrIs slightly overlapped zero.

We used the same approach to fit the multivariate model for Studies 3 and 4. The model for the images matched and unmatched to cultural context converged successfully (R^2: range = .33–.46). Convergence diagnostics were generally acceptable (most R-hat ≈ 1.00; key fixed effects and variances well identified). Posterior predictive checks showed that the model consistently reproduced the observed means and variances for all four outcomes (Bayesian *p* values = .50–.52), indicating good overall fit. A summary of the fixed effects showed that on average, matched images produced stronger reactions than unmatched images for targeted emotion (b = 0.09, 95% CrI = [0.02, 0.15]). In contrast, targeted valence, arousal, and motivation did not differ across image types (b = 0.04, 95% CrI = [–0.01, 0.09]; b = 0.04, 95% CrI = [–0.00, 0.09]; b = 0.04, 95% CrI = [–0.02, 0.10], respectively) because CrIs slightly overlapped zero.

The model for the female versus male variants of the images converged successfully (R^2: range = .31–.48). Convergence diagnostics were generally acceptable (most R-hat ≈ 1.00; key fixed effects and variances well identified). Posterior predictive checks showed that the

Table 2. Descriptive Statistics for Images From Study 1 and Study 2

	Study 1				Study 2
Measure	AI-generated *M* (*SD*)	Non-AI-generated *M* (*SD*)	Mean difference [95% CI]	Cohen's *d*	AI-generated *M* (*SD*)
Targeted emotion	4.75 (2.08)	4.67 (2.09)	0.07 [0.00, 0.14]	0.03	4.78 (2.06)
Amusement	5.08 (1.85)	4.53 (1.93)	0.55 [0.18, 0.93]	0.29	4.89 (1.90)
Anger	4.80 (1.93)	5.06 (1.87)	−0.26 [−0.48, −0.05]	−0.14	4.19 (2.22)
Attachment love	3.73 (2.27)	3.52 (2.18)	0.21 [−0.01, 0.43]	0.09	4.66 (2.21)
Awe	4.97 (1.93)	4.59 (1.97)	0.38 [0.20, 0.56]	0.20	4.73 (2.08)
Craving	5.38 (1.84)	5.13 (1.98)	0.26 [−0.04, 0.55]	0.13	5.48 (1.86)
Disgust	5.10 (2.03)	5.39 (1.97)	−0.28 [−0.48, −0.08]	−0.14	5.27 (1.98)
Excitement	4.12 (1.96)	4.14 (2.03)	−0.02 [−0.50, 0.46]	−0.01	4.40 (2.06)
Fear	4.76 (2.04)	4.66 (2.09)	0.10 [−0.11, 0.31]	0.05	4.85 (1.97)
Joy	4.78 (2.02)	4.55 (2.17)	0.23 [−0.19, 0.64]	0.11	4.83 (1.94)
Neutral	4.63 (2.15)	4.61 (2.16)	0.02 [−0.21, 0.25]	0.01	4.30 (2.14)
Nurturant love	4.70 (2.03)	4.39 (2.04)	0.31 [0.11, 0.51]	0.15	4.82 (2.09)
Sadness	4.96 (2.04)	5.17 (1.95)	−0.20 [−0.38, −0.03]	−0.10	4.88 (2.06)
Targeted valence					
Positive	5.39 (1.85)	5.11 (1.92)	0.28 [0.20, 0.36]	0.15	5.32 (1.84)
Negative	5.18 (1.89)	5.35 (1.88)	−0.17 [−0.26, −0.07]	−0.09	5.20 (2.01)
Targeted arousal					
Arousing	2.29 (1.79)	2.17 (1.71)	0.12 [0.06, 0.18]	0.07	2.49 (1.93)
Targeted motivation					
Approach	4.79 (2.13)	4.50 (2.15)	0.30 [0.20, 0.39]	0.14	4.59 (2.20)
Avoidance	4.58 (2.31)	4.65 (2.31)	−0.07 [−0.18, 0.05]	−0.03	4.71 (2.35)

Note: None of the discrete emotions were identified as aiming to calm the participants. AI = artificial intelligence; CI = confidence interval.

model consistently reproduced the observed means and variances for all four outcomes (Bayesian *p* values: .50–.52), indicating good overall fit. Fixed-effect estimates for differences between sex variants were small, and their 95% CrIs all included zero for each outcome (targeted emotion, valence, arousal, and motivation, respectively: *b* = 0.01, 95% CrI = [−0.21, 0.30]; *b* = 0.09, 95% CrI = [−0.30, 0.45]; *b* = 0.07, 95% CrI = [−0.07, 0.21]; *b* = 0.03, 95% CrI = [−0.33, 0.38]). It provides no clear evidence that subtly adjusting the apparent sex of depicted individuals systematically changes targeted emotion, valence, arousal, or motivational ratings.

The model converged successfully (R^2: range = .36–.49). Convergence diagnostics were generally acceptable (all reported R-hat ≈ 1.00 and key fixed effects and variances well identified). Posterior predictive checks showed that the model consistently reproduced the observed means and variances for all four outcomes (Bayesian *p* values = .49–.51), indicating good overall fit. Fixed-effect estimates for adults and older adults were small, and their 95% CrIs all included zero for each outcome (adults for targeted emotion, valence, arousal, and motivation, respectively: *b* = 0.12, 95% CrI = [−0.21, 0.43]; *b* = −0.10, 95% CrI = [−0.44, 0.24]; *b* = 0.07, 95% CrI = [−0.06, 0.20]; *b* = 0.15, 95% CrI = [−0.14, 0.45]; older adults for targeted emotion, valence, arousal, and motivation, respectively: *b* = −0.01, 95% CrI = [−0.34, 0.32]; *b* = −0.08, 95% CrI = [−0.43, 0.25]; *b* = −0.05, 95% CrI = [−0.17, 0.07]; *b* = 0.02, 95% CrI = [−0.28, 0.31]). It provides no clear evidence that subtly adjusting the apparent age of depicted individuals systematically changes targeted emotion, valence, arousal, or motivational ratings.

Unregistered frequency analysis

Using the direct-comparison items, we calculated the percentage of responses indicating whether the AI-generated images or images from existing affective databases elicited much weaker, a little weaker, the same, a little stronger, or much stronger affective responses (Table S1 in the Supplemental Material). Across all measures, "the same" was the most frequently selected response, ranging from 21% to 71%. In addition, for amusement, attachment love, awe, craving, excitement, fear, neutral, nurturant love, positive, arousing, and calming items, the proportion of participants indicating stronger responses (combined a little stronger and much stronger) exceeded those indicating weaker responses

Table 3. Descriptive Statistics for Images From Study 3 and Study 4

Measure	Culturally matched *M* (*SD*)	Culturally unmatched *M* (*SD*)	Mean difference [95% CI]	Cohen's *d*
Targeted emotion	4.70 (2.04)	4.61 (2.06)	0.09 [0.04, 0.14]	0.05
Amusement	4.64 (1.97)	4.43 (2.05)	0.20 [0.03, 0.38]	0.10
Anger	4.45 (2.17)	4.34 (2.17)	0.11 [−0.08, 0.30]	0.05
Attachment love	4.53 (2.11)	4.31 (2.13)	0.22 [0.03, 0.40]	0.10
Awe	4.74 (2.06)	4.69 (2.03)	0.06 [−0.12, 0.23]	0.03
Craving	5.54 (1.73)	5.37 (1.83)	0.17 [0.02, 0.32]	0.10
Disgust	4.82 (2.03)	4.80 (2.07)	0.02 [−0.16, 0.20]	0.01
Excitement	4.37 (2.01)	4.19 (2.03)	0.18 [0.01, 0.35]	0.09
Fear	4.74 (2.01)	4.56 (2.04)	0.17 [0.00, 0.34]	0.09
Joy	5.01 (1.88)	5.02 (1.8)	−0.01 [−0.16, 0.15]	0.00
Neutral	4.01 (1.96)	4.01 (2.03)	0.00 [−0.17, 0.17]	0.00
Nurturant love	4.68 (2.04)	4.75 (1.99)	−0.07 [−0.24, 0.1]	−0.04
Sadness	4.88 (2.06)	4.80 (2.11)	0.09 [−0.09, 0.26]	0.04
Valence				
Positive	5.49 (1.69)	5.43 (1.72)	0.06 [0.00, 0.11]	0.03
Negative	5.03 (1.99)	5.01 (2.02)	0.02 [−0.06, 0.11]	0.01
Arousal				
Arousing	2.38 (1.83)	2.34 (1.80)	0.04 [0.00, 0.09]	0.02
Motivation				
Approach	4.75 (2.06)	4.73 (2.09)	0.02 [−0.05, 0.08]	0.01
Avoidance	4.45 (2.31)	4.37 (2.32)	0.08 [−0.02, 0.18]	0.04

Note: Positive values for the mean difference and Cohen's *d* indicate stronger reactions elicited by the culturally matched images than by the unmatched images. CI = confidence interval.

(combined a little weaker and much weaker). Most of the differences between combined more and combined less were significant, except for disgust and joy.

Furthermore, we ran the same analysis for culturally matched and unmatched images from Study 3 (Table S2 in the Supplemental Material). Across all measures, "the

Table 4. Descriptive Statistics for Images From Study 5

Measure	Female variants *M* (*SD*)	Male variants *M* (*SD*)	Mean difference [95% CI]	Cohen's *d*
Targeted emotion	4.30 (2.06)	4.29 (2.10)	0.01 [−0.08, 0.11]	0.01
Amusement	4.22 (1.95)	3.79 (2.14)	0.43 [0.14, 0.72]	0.21
Anger	3.77 (2.11)	3.82 (2.15)	−0.05 [−0.36, 0.26]	−0.02
Attachment love	3.85 (2.09)	3.59 (2.15)	0.26 [−0.05, 0.56]	0.12
Disgust	5.12 (1.95)	5.33 (1.92)	−0.21 [−0.49, 0.06]	−0.11
Excitement	4.38 (1.85)	4.21 (1.94)	0.16 [−0.11, 0.43]	0.09
Fear	4.21 (2.06)	4.14 (2.06)	0.06 [−0.23, 0.36]	0.03
Joy	4.72 (1.78)	4.62 (1.82)	0.10 [−0.16, 0.36]	0.05
Neutral	4.54 (2.13)	4.83 (2.09)	−0.29 [−0.57, −0.02]	−0.14
Nurturant love	4.00 (2.15)	4.06 (2.08)	−0.06 [−0.37, 0.25]	−0.03
Sadness	4.16 (2.17)	4.35 (2.02)	−0.19 [−0.49, 0.11]	−0.09
Valence				
Positive	5.10 (1.80)	4.86 (1.89)	0.24 [0.13, 0.35]	0.13
Negative	4.90 (2.03)	5.03 (1.97)	−0.13 [−0.27, 0.01]	−0.07
Arousal				
Arousing	2.08 (1.60)	2.01 (1.56)	0.07 [0.00, 0.14]	0.05
Motivation				
Approach	4.28 (2.11)	4.06 (2.17)	0.22 [0.10, 0.35]	0.10
Avoidance	4.36 (2.34)	4.62 (2.26)	−0.26 [−0.43, −0.10]	−0.11

Note: Positive values for the mean difference and Cohen's *d* indicate stronger reactions elicited by the female variants of the images than by the male variants. CI = confidence interval.

Table 5. Descriptive Statistics for Images From Study 6

Measure	Minor variants *M* (*SD*)	Adult variants *M* (*SD*)	Older-adult variants *M* (*SD*)	Cohen's *d* Minors vs. adults	Cohen's *d* Minors vs. older adults	Cohen's *d* Adults vs. older adults
Targeted emotion	4.44 (2.11)	4.55 (2.09)	4.44 (2.11)	−0.05	0.00	0.05
Amusement	3.87 (2.08)	3.75 (2.11)	3.42 (2.13)	0.06	0.21	0.16
Anger	3.63 (2.07)	3.77 (2.05)	3.62 (2.15)	−0.06	0.00	0.07
Attachment love	4.25 (2.31)	4.18 (2.24)	4.23 (2.30)	0.03	0.01	−0.02
Disgust	5.30 (1.87)	5.72 (1.73)	5.49 (1.86)	−0.23	−0.10	0.13
Excitement	4.12 (2.08)	4.40 (1.99)	4.17 (2.02)	−0.14	−0.03	0.11
Fear	4.47 (2.12)	4.80 (2.09)	4.65 (2.05)	−0.15	−0.08	0.07
Joy	5.09 (1.78)	5.01 (1.75)	4.89 (1.82)	0.05	0.11	0.06
Neutral	4.13 (2.00)	4.47 (2.18)	4.33 (1.93)	−0.17	−0.10	0.07
Nurturant love	—	4.54 (2.00)	4.81 (1.93)	—	—	−0.14
Sadness	5.06 (1.96)	4.85 (1.98)	4.85 (2.01)	0.11	0.11	0.00
Valence						
Positive	5.34 (1.81)	5.20 (1.87)	5.21 (1.84)	0.08	0.07	−0.01
Negative	5.14 (1.92)	5.10 (1.93)	5.10 (1.89)	0.02	0.02	0.00
Arousal						
Arousing	2.19 (1.75)	2.27 (1.79)	2.15 (1.70)	−0.04	0.02	0.07
Motivation						
Approach	4.59 (2.13)	4.52 (2.16)	4.46 (2.15)	0.03	0.06	0.03
Avoidance	4.46 (2.31)	4.94 (2.19)	4.72 (2.25)	−0.21	−0.11	0.10

Note: Positive values for the mean difference and Cohen's *d* indicate stronger reactions elicited by the first category variants of the images than by the second category variants.

same" was the most frequently selected response, ranging from 25% to 71%. In addition, for amusement, anger, attachment love, craving, excitement, joy, nurturant love, negative, and calming items, the proportion of participants indicating stronger responses (combined a little stronger and much stronger) exceeded those indicating weaker responses (combined a little weaker and much weaker). However, the differences between combined more and combined less were significant only for amusement, nurturant love, and sadness.

Calculating SESOI

We present the mean differences between the images divided into subcategories (much less, little less, the same, little more, and much more) in Table 6. For the comparison between affective images, we first noticed that images rated as eliciting "the same" response demonstrated slight differences in affective responses (ranging from $d = -0.04$ to $d = 0.18$). Images that elicited a little more affective response, in fact, elicited stronger reactions (ranging from $d = 0.07$ to $d = 0.51$), and images that elicited a little less affective response, in fact, elicited weaker reactions (ranging from $d = -0.38$ to $d = 0.03$). Combining the estimates for the a little more and a little less categories yielded overall SESOI estimates (from absolute $d = 0.05$ to $d = 0.29$). These values represent the smallest subjectively experienced differences in affective response, providing an essential reference point for affective studies. For the separate results for the SESOI analysis for Study 1 and Study 3, see the Supplementary Information in the Supplemental Material.

Discussion

In this project, we focused on leveraging generative AI to advance affect-induction procedures. First, we proposed a pipeline demonstrating how generative AI can be used to develop affective images by building on existing data sets and emotion definitions. Second, we created and validated a library of AI-generated images and accompanying descriptions, spanning 12 discrete emotion categories and annotated with affective ratings. Our goal was to address key limitations of existing image databases by providing a proof of concept: that generative AI can produce emotionally evocative images that are both efficient and adaptable to specific contexts (i.e., culture, sex, and age). For efficiency, the results showed that the AI-generated images elicited comparable emotional responses from those from existing affective databases. To demonstrate the potential for adaptability, we tailored images to six broad cultural regions: African, Arabic, Asian, Indian, Latin American, and European/

Table 6. Mean Difference Between the Images Subcategorized by the Direct-Comparison Category

	Much less		Little less		The same		Little more		Much more		SESOI	
	M (*SD*)	d_z	*M* (*SD*)	d_z	*M* (*SD*)	d_z	*M* (*SD*)	d_z	*M* (*SD*)	d_z	Absolute *M* (*SD*)	Absolute d_z
Targeted Emotion	–0.72 (2.01)	–0.36	–0.24 (1.51)	–0.16	0.10 (1.54)	0.07	0.40 (1.50)	0.27	0.74 (1.95)	0.38	0.32 (1.51)	0.21
Amusement	–0.65 (1.79)	–0.36	–0.08 (1.74)	–0.05	0.24 (1.37)	0.18	0.76 (1.49)	0.51	1.12 (1.82)	0.61	0.42 (1.62)	0.28
Anger	–0.96 (2.27)	–0.42	–0.46 (1.55)	–0.30	0.10 (1.76)	0.06	0.35 (1.68)	0.21	0.73 (2.21)	0.33	0.41 (1.61)	0.25
Attachment Love	0.05 (1.60)	0.03	0.03 (1.34)	0.03	0.17 (1.24)	0.14	0.37 (1.26)	0.29	0.79 (1.29)	0.61	0.20 (1.30)	0.16
Awe	–0.52 (1.78)	–0.29	–0.16 (1.44)	–0.11	0.24 (1.51)	0.16	0.49 (1.52)	0.32	0.65 (1.77)	0.37	0.32 (1.48)	0.21
Craving	–0.90 (2.31)	–0.39	–0.35 (1.28)	–0.27	–0.04 (1.10)	–0.03	0.37 (1.15)	0.32	1.14 (1.75)	0.65	0.36 (1.21)	0.29
Disgust	–0.99 (2.27)	–0.44	–0.57 (1.51)	–0.38	0.02 (1.86)	0.01	0.33 (1.69)	0.20	0.59 (2.02)	0.29	0.45 (1.60)	0.29
Excitement	–0.32 (1.51)	–0.21	–0.25 (1.47)	–0.17	0.17 (1.28)	0.13	0.31 (1.28)	0.24	0.69 (1.57)	0.44	0.28 (1.38)	0.20
Fear	–0.58 (1.90)	–0.30	–0.26 (1.62)	–0.16	0.16 (1.36)	0.12	0.40 (1.63)	0.24	0.88 (2.01)	0.44	0.33 (1.62)	0.20
Joy	–0.79 (1.50)	–0.53	–0.19 (1.25)	–0.15	0.06 (1.31)	0.05	0.53 (1.22)	0.44	0.73 (1.68)	0.44	0.36 (1.23)	0.29
Neutral	–0.93 (2.30)	–0.40	0.04 (1.83)	0.02	0.05 (1.90)	0.03	0.12 (1.76)	0.07	0.30 (3.04)	0.10	0.08 (1.79)	0.05
Nurturant Love	–0.38 (1.97)	–0.19	–0.22 (1.47)	–0.15	0.11 (1.35)	0.08	0.40 (1.43)	0.28	0.52 (1.74)	0.30	0.31 (1.45)	0.21
Sadness	–0.99 (1.93)	–0.51	–0.31 (1.44)	–0.22	–0.06 (1.63)	–0.04	0.34 (1.61)	0.21	0.57 (2.05)	0.28	0.33 (1.52)	0.22
Valence												
Positive	–0.53 (1.73)	–0.31	–0.17 (1.24)	–0.13	0.11 (1.19)	0.09	0.39 (1.22)	0.32	0.74 (1.59)	0.47	0.28 (1.23)	0.23
Negative	–0.79 (1.96)	–0.40	–0.37 (1.40)	–0.26	0.03 (1.48)	0.02	0.27 (1.45)	0.18	0.37 (1.97)	0.19	0.32 (1.42)	0.22
Arousal												
Arousing	–0.13 (1.73)	–0.08	–0.14 (1.55)	–0.09	0.06 (1.18)	0.05	0.30 (1.52)	0.20	0.24 (1.75)	0.14	0.22 (1.54)	0.14
Calming	–0.29 (1.88)	–0.15	–0.20 (1.63)	–0.12	0.10 (1.31)	0.07	0.41 (1.60)	0.25	0.57 (1.90)	0.30	0.30 (1.61)	0.19

Note: d_z = Cohen's d_z. Positive values for the mean difference and Cohen's *d* indicate stronger reactions elicited by the AI-generated images than by the non-AI-generated images. AI = artificial intelligence.

North American. When testing matched versus unmatched images across different cultures, the images adjusted to participants' cultural context elicited slightly stronger emotional responses than the unadjusted stimuli. Sex- and age-adjusted variants produced comparable responses with their base images, demonstrating controllability without loss of affective impact. Finally, we calculated the smallest detectable differences in people's affective states that can be used in future studies using affect-induction procedures.

Generating the affective images using AI

Affective scientists are increasingly embracing AI to support various aspects of research, from modeling emotion dynamics (Coles et al., 2025) and classifying affective behavior (Cowen et al., 2021) to developing robots (Saad et al., 2024) and analyzing physiological data (Perz & Kazienko, 2025; Saganowski et al., 2022). Our study adds to that growing body of work by demonstrating another promising application: creating efficient and culturally adaptable affective images.

One of the key contributions of this project is the creation of a pipeline that can be used to generate affective images with AI. We transparently documented each step that led to the creation of the final image set. The process of creating the AI-generated images proved to be both rewarding and challenging. Overall, researchers found the experience to be highly positive, with the AI platforms being efficient and enjoyable to use. The ability to describe and generate new stimuli was particularly impressive, with the realism of the generated photos often surpassing expectations. However, we noted that achieving the desired image frequently required additional effort and tweaking because it was not always a straightforward task. Thus, the iterative process of refining prompts and generating multiple variations helped optimize the results.

One significant challenge was generating detailed and natural-looking faces, particularly when the image included more than four or five people. In addition, issues with image quality persisted, including low-resolution photos and the occasional generation of disturbing or inappropriate content, such as graphic depictions of violence, wounds, or controversial symbols. Another limitation was the inclusion of written text in some images, which was not always grammatically correct. Furthermore, inconsistencies in context occasionally arose, such as mismatches between a person's appearance and the setting (e.g., dirty hands paired with a clean shirt) or issues with bodily anatomy, such as poorly rendered hands. Adjusting elements such as color balance and focus also presented challenges. We also noticed that the generated images presented an unrealistic standard of beauty for all depicted people. Every face appeared highly attractive, and it proved challenging to generate faces that were not physically attractive.

During the image-generation period from April 2024 to August 2025, the researchers observed a noticeable improvement in the quality of the AI-generated images. The evolution of the models over time and the introduction of new models became evident, with certain models proving more adept at generating specific types of images. This variation highlighted the importance of selecting the right model for the task at hand. For example, although some models excelled at creating highly realistic portraits, others were better suited for generating scenes with complex backgrounds. In our qualitative review of outputs, Midjourney appeared more consistent at producing awe-evoking scenes, whereas Google Imagen often handled larger group depictions more cleanly. These are observational patterns rather than results of a controlled benchmark. The models also differed in their content restrictions. For example, some (e.g., Seedream) allowed the generation of highly realistic disgust-eliciting images, whereas others (e.g., Google Imagen and Midjourney) imposed stricter limits that constrained such content. Initially, only Midjourney allowed for the effective correction of small errors using the "Vary (Region)" option, whereas over time, other models also acquired similar capabilities.

Importantly, this experience highlighted a broader challenge in working with generative AI: the rapid pace of technological advancement. Tools are evolving continuously, and the capabilities, limitations, and even interfaces of these platforms can change significantly within months. This dynamic landscape has important implications for affective science. Newer models may quickly outperform previous image sets, making older stimuli less relevant or useful. Furthermore, if the tools change, benchmarks and validation results may lose their relevance quickly. Affective researchers may need to routinely revalidate stimuli or shift toward frameworks that validate generation pipelines instead of static image sets. To address this, we not only validated the images themselves but also generated detailed, human-written image descriptions that serve as an "immortal" layer of annotation—interpretable across time, theories, and tools. These descriptions ensure that even as AI evolves, the emotional meaning and context of the stimuli remain accessible and comparable. On the positive side, the pace of development also opens doors for real-time adaptation—for instance, tailoring stimuli to individuals or dynamically generating stimuli during an experiment.

Furthermore, generative images enable ethically controlled exposure when real photographs are scarce, restricted, or inappropriate to collect—for example, scenes involving minors, medical procedures, disaster aftermath, illicit activities, or culturally sensitive taboos (e.g., alcohol). They also help study low-base-rate or logistically inaccessible phenomena (e.g., rare hazards, remote settings) without endangering participants or violating privacy/consent norms. Crucially, the pipeline supports institutional-review-board-compliant tailoring, allowing researchers to generate only content that aligns with local ethics and study aims.

Finally, our image-generation pipeline integrated generative AI with careful human oversight. Throughout the process, we consulted with cultural raters to ensure that the images not only matched their intended emotional targets but also fit appropriately within specific cultural contexts. Despite the advantages of AI, the involvement of human reviewers remained essential—each image was viewed, discussed, and iteratively refined based on qualitative feedback. Note that we observed meaningful disagreement among raters from the same cultural group, which we interpreted as a valuable reminder of within-cultures diversity. These differences were transparently documented and included in the "Data, Analysis Code & Outputs OSF Component" (Behnke et al., 2025b), "SupplementaryData.xlsx" spreadsheet, sheet: "authors_disagreements," to support future research.

The LAI-GAI

With LAI-GAI, we showed that it is possible to overcome several key limitations outlined in the introduction, ensuring relevance for current research needs and showing that AI-generated images could serve as a valuable complement to traditional image databases. We were able to generate images intended to elicit the 12 discrete emotions. In terms of emotion elicitation, we found that our AI-generated positive stimuli often elicited stronger affective responses than their counterparts from existing affective databases—a finding consistent across multiple statistical approaches. However, similar improvements were not observed for negative stimuli. The negative images were consistently worse at eliciting sadness, anger, and general negative affect than images from existing affective databases.

One likely reason for this discrepancy lies in the limitations encountered during the image generation. We faced challenges in creating images that depict disturbing or emotionally intense content, such as visible injuries, acts of violence, or controversial symbols. These difficulties do not appear to reflect limitations in the generative model's capabilities per se but rather stem from embedded filtering mechanisms designed to prevent the production of sensitive or potentially harmful material. Consequently, we were unable to generate negative stimuli that might elicit stronger affective responses. However, it is important to emphasize that in our early attempts, these filters not only blocked some extreme images depicted in existing data sets but also prevented the creation of more moderate "negative" scenes—such as street fights or violence against animals and minors—representative of everyday norm violations. With the emergence of newer, less restrictive open models, generating such moderate negative content became feasible. For example, in the most recent iteration of age variants, the models successfully generated strong disgust elicitors. Further validation and refinement—especially for negative stimuli—are needed to realize their potential as complementary resources that enhance existing approaches. Importantly, given that existing data sets tend to overrepresent negative content while offering fewer validated positive stimuli, our study addresses a significant gap by expanding the availability and quality of positive emotional materials.

There is growing recognition in the literature of the need to move beyond convenience sampling—not only of participants but also of stimuli, situations, and contexts (Barrett, 2022). Traditionally, stimulus selection has often been constrained by the availability of validated materials, limiting the ability to explore stimulus variation and generalization systematically. One of the major contributions of this project is that it makes such exploration far more feasible. Generative AI allows systematic sampling of large, diverse, context-specific stimuli, yielding robust tests of emotional generalizability, supporting within-categories variation, and reducing overinterpretation from narrow sets.

Crucially, the present library is a proof of concept, not a comprehensive endpoint. Its coverage—limited to six broad, simplified cultural groupings—is necessarily incomplete; we had to start somewhere, and these categories offered a pragmatic first step. Even when affective responses are comparable with existing stimuli, our approach enables purpose-specific adjustments (e.g., reducing overreliance on Caucasian faces in non-Western settings) while preserving affective impact, demonstrating novelty through controllability rather than stronger elicitation alone. Accordingly, our primary contribution is the pipeline: a transparent, human-in-the-loop process that others can reuse and extend to build country- or community-specific sets, update content as models evolve, and integrate local expertise to mitigate bias.

Adjusting stimuli to the context

We also demonstrate that generative AI can on-demand produce customized affective stimuli tailored to cultural contexts and to demographic characteristics, including sex and age. We further show that it is feasible to adapt stimuli to a single country context—here, India (despite its size and heterogeneity)—illustrating a concrete use case for more fine-grained, country-specific applications.

This opens new possibilities for tailoring affective stimuli to particular populations or cultural settings, addressing a major limitation of many existing static-image databases. We showed that images can be adapted to reflect different cultural contexts and observed some advantages of using culturally matched versus unmatched images, although they were smaller than expected. One possible explanation is that globalization has diminished distinct cultural differences because people worldwide are increasingly exposed to similar images. Another explanation might be that affective reactions are not as strongly culturally driven as initially hypothesized but instead reflect more universal emotional responses. We also note that culture may matter in ways not captured by minor adjustments: In some settings, entire concepts (e.g., companion animals) can be unacceptable, so "adjusted" versions may be ineffective or inappropriate; in such cases, redesign—not cosmetic modification—is warranted. However, because this study was not designed to test theoretical claims about the cultural versus universal nature of emotions, we refrain from drawing strong conclusions and encourage future research to address these questions directly.

We extend our approach by generating matched variants depicting male and female individuals and minors, middle-aged adults, and older adults while preserving the underlying scene and targeted emotion. This allows researchers to align stimuli more closely with the populations and designs of their studies (e.g., age-congruent stimuli for older adults, sex-consistent depictions to avoid confounds) without compromising affective validity.

We demonstrated that by using generative AI, researchers can now generate stimuli tailored to their precise research needs, allowing for greater control over the emotional content of the images. However, it also raises important considerations for cumulative science. If each research group creates its own set of tailored stimuli, comparability across studies may be reduced unless reporting standards are followed. To address this, we provide detailed documentation of our generation pipeline, including AI prompts, model names, selection procedures, and open access to all stimuli and annotations. We recommend that future studies adopt similar approaches, including this level of transparency, to support reproducibility and ensure that generative methods strengthen rather than fragment the field.

Estimating the smallest difference participants can distinguish

Our study also offered a comparative estimate of the smallest subjectively experienced difference in the context of affective image evaluation using the anchor-based approach proposed by Anvari and Lakens (2021). When we compared responses with images judged as eliciting a little more or a little less intense emotional responses, we found consistent shifts in the expected directions. Images classified as eliciting a little more emotion produced stronger responses (up to d = 0.51), and those judged as a little less yielded weaker reactions (down to d = −0.38). These results parallel Anvari and Lakens's findings that even small self-reported changes in affect correspond to measurable differences on validated scales. By combining these little-more and little-less categories, we derived the estimates between Ms = 0.20 and 0.45 for affective images (with neutral notably smaller at M = 0.08), values close to the estimates reported by Anvari and Lakens (M = 0.31 for positive affect and M = 0.25 for negative affect) for the changes in affect across 2 weeks. Furthermore, similarly to the Anvari and Lakens study, we observed that images rated as eliciting the same affective responses still yielded small but nonzero differences in emotion ratings, indicating that even subjectively equivalent affective experiences may be associated with measurable variance.

Note that our design applied the anchor-based method to cross-context comparisons (i.e., image pairs) rather than longitudinal self-assessments. This suggests that the global rating of difference can also be fruitfully used in between-stimuluses designs, as proposed in the extended discussion by Anvari and Lakens (2021). Importantly, we found no single, consistent threshold: Some dimensions (e.g., neutral, attachment love) showed smaller estimates, variability was substantial (wide standard deviations), and in the a-little-less category, the sign occasionally reversed. We therefore present these values as descriptive benchmarks—not fixed cutoffs—to guide study design, power analysis, and the interpretation of future equivalence tests.

Limitations and future directions

Despite the promising results of this study, several limitations warrant consideration. One limitation of the present study is that it was conducted solely in English. Although this is the case for most available research, future studies should account for the linguistic diversity of affective expression (Jackson et al., 2019) and explore conducting research in other languages. This is particularly important given evidence that emotional responses—especially to negative stimuli—may be attenuated when processed in a second language (e.g., Jończyk et al., 2016; Wu & Thierry, 2012; Zhang et al., 2023). In addition, even though we used stimuli adapted for four distinct cultures in the study, we did not fully capture the diversity of all these cultures. The images may not be universally applicable in a given region (e.g., Asia), and future research should include a wider range of cultures to better understand cross-cultural emotional responses.

In terms of diversity, although the current data set includes a variety of emotional images, it predominantly features heteronormative gender representations. To broaden the scope and reflect a more inclusive range of experiences, future versions of the stimuli should consider adding images that represent diverse sexual orientations and gender identities. For example, incorporating images of same-sex couples alongside the more traditional heteronormative representations could further enhance the diversity and inclusivity of the library.

To further optimize the generation of affective stimuli, future studies should consider controlling additional factors that may affect the emotional responses elicited by the images. For example, factors such as camera type, time of day, and lighting conditions can influence the way emotions are conveyed in visual stimuli. Although it may not be possible to alter these factors in the current library, acknowledging their potential impact in future discussions and studies would be an important consideration for improving consistency in emotional responses.

Another limitation is the challenge of effectively eliciting certain emotions, such as anger, using still visual stimuli. Although many "anger" stimuli appeared ambiguous and coactivated sadness, fear, or disgust, we do not treat this coactivation as evidence of induction failure. Rather, it likely reflects that negative-emotion reports tend to be naturally correlated, which can blur distinctions between closely related emotions in static visual elicitation. In addition, although generative AI is currently limited in its ability to generate certain types of affective content, such as films, this is an area in which affective research may evolve in the near future. These limitations are partly due to restricted access to some of the most advanced generative video models and partly because the generated videos often still appear odd or unrealistic—reminiscent of the early stages of generative image models. As generative AI continues to advance, researchers should prepare for the potential of generating not only static images but also dynamic stimuli, such as video clips (e.g., by preparing the descriptions of the videos). These could offer new possibilities for studying emotional responses in more complex contexts, adding a layer of depth to the research.

The cultural adjustments made during the image-generation process revealed the challenges of accurately representing diverse cultural contexts in a single cluster. For instance, images adapted to the Asian context risk ignoring the enormous diversity in physical appearance, cultural practices, and religious norms across and within East, South, and Southeast Asia. Likewise, images that were considered acceptable in some countries (e.g., Japan) could be viewed as inappropriate in others (e.g., Bangladesh) because of religious taboos, such as depictions of pigs or alcohol. Crucially, continuous, iterative feedback from cultural experts on clothing cues, food presentation, gesture norms, and landscape/architectural details was indispensable for detecting such issues early, adjudicating ambiguous cases, and guiding revisions toward locally appropriate depictions. In addition, stereotypical portrayals (e.g., often emphasizing poverty in Africa) were noted by cultural experts as misrepresenting the continent's diversity. Some authors also pointed out the forced nature of certain cultural adjustments, particularly when traditional elements, such as clothing or flowers, were added in an attempt to fit the cultural context. Although these cues were helpful in signaling the intended culture, they sometimes exaggerated stereotypes, which could compromise the ecological validity of the stimuli. Moreover, some traditional cultures may still treat generative AI with suspicion, which can hinder adoption and limit the adaptability of these tools.

Importantly, our goal is not to present the current image library as a finished, globally representative resource. Rather, we see our main contribution as the methodological pipeline. We demonstrate that generative image creation can be effectively integrated with iterative, local-expert review in a human-in-the-loop process. This pipeline is designed to be extensible: For cases in which specific use cases warrant it (and relevant expertise is available), it can be adapted to generate country-specific or even subcultural image sets. Because generative models often encode Global North-centric priors that can surface in subtle ways (e.g., high-arousal smiling), it is essential that images be reviewed by local cultural experts and revised iteratively before inclusion, with ongoing monitoring for residual biases. In this way, future contributors can build additional, finer-grained subsets without overclaiming cultural coverage and while maintaining ecological validity. Thus, future work should continue to refine these methods, focusing on enhancing the ecological validity of the generated images and exploring deeper cultural nuances to avoid oversimplification and stereotyping.

As AI-generated images become more prevalent, a broader risk arises from training new AI systems on these images, which could compromise their ability to accurately reflect real-world pictures and experiences. Our data set is not meant for training such general systems. That said, there may be value in purpose-built tools that estimate the emotional impact of AI-generated images—for example, tools that could hide or down-rank images likely to elicit strong negative affect. If used this way, images should be clearly labeled as AI-generated/AI-made, and human-made images should be analyzed separately; evaluations should be planned in advance and include checks for bias and for whether results hold up when the images come from different sources.

Conclusion

The use of generative AI to create affective stimuli marks a meaningful step forward in affective science, providing greater flexibility, contextual adaptability, and technical quality than traditional data sets. This project introduced a reproducible pipeline for generating and validating AI-based images. Our results demonstrate that such AI-generated images can reliably elicit affective responses and be tailored to specific cultural and demographic contexts. As generative tools continue to evolve, the presented approach offers a scalable and customizable solution for advancing affective research across diverse domains and scenarios. More broadly, this work lays a robust foundation for future AI-driven investigations in affective science, emphasizing transparency and openness in such scientific inquiry. Importantly, the outcome of this project is the LAI-GAI—the first publicly available database of AI-generated affective images, validated across international and diverse samples, ready for immediate use in future research.

Transparency

Action Editor: Pamela Davis-Kean
Editor: David A. Sbarra
Author Contributions

Maciej Behnke: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.
Maciej Kłoskowski: Methodology; Resources; Writing – review & editing.
Michał Klichowski: Conceptualization; Funding acquisition; Resources; Visualization; Writing - original draft; Writing – review & editing.
Wadim Krzyżaniak: Methodology; Resources; Writing – review & editing.
Kacper Szymański: Methodology; Resources; Writing – review & editing.
Patryk Maciejewski: Methodology; Resources; Writing – review & editing.
Patrycja Chwiłkowska: Methodology; Resources; Writing – review & editing.
Marta Kowal: Methodology; Resources; Writing – original draft; Writing - review & editing.
Rafał Jończyk: Methodology; Resources; Writing – review & editing.
Jan Nowak: Methodology; Resources; Software; Writing – review & editing.
Szymon Kupiński: Methodology; Resources; Software; Writing – review & editing.
Dominika Kunc: Formal analysis; Resources; Writing – review & editing.
Stanisław Saganowski: Methodology; Resources; Writing – review & editing.
Aakash A. Chowkase: Methodology; Validation; Writing – review & editing.
Farida Guemaz: Methodology; Validation; Writing – review & editing.
Kevin S. Kertechian: Methodology; Validation; Writing – review & editing.
Ameer I.M.T. Maadal: Methodology; Validation; Writing – review & editing.
Leonardo A. Aguilar: Methodology; Validation; Writing – review & editing.
Barnabas T. Alayande: Methodology; Validation; Writing – review & editing.
Vimala Balakrishnan: Methodology; Validation; Writing – review & editing.
Dana M. Basnight-Brown: Methodology; Validation; Writing – review & editing.
Jordane Boudesseul: Methodology; Validation; Writing – review & editing.
Tomás A. D'Amelio: Methodology; Validation; Writing – review & editing.
Jovi C. Dacanay: Methodology; Validation; Writing – review & editing.
Abhishek Dedhe: Methodology; Validation; Writing – review & editing.
Shan Gao: Methodology; Validation; Writing – review & editing.
Joao F. G. B. Takayanagi: Methodology; Validation; Writing – review & editing.
Md. Rohmotul Islam: Methodology; Validation; Writing – review & editing.
Alvaro Mailhos: Methodology; Validation; Writing – review & editing.
Christine M. Mpyangu: Methodology; Validation; Writing – review & editing.
Moises Mebarak: Methodology; Validation; Writing – review & editing.
Arooj Najmussaqib: Methodology; Validation; Writing – review & editing.
Ju Hee Park: Methodology; Validation; Writing – review & editing.
Ekaterine Pirtskhalava: Methodology; Validation; Writing – review & editing.
Eli Rice: Methodology; Validation; Writing – review & editing.
Sohrab Sami: Methodology; Validation; Writing – review & editing.
Yuki Yamada: Methodology; Validation; Writing – review & editing.
Jan Baczyński: Methodology; Resources; Writing – review & editing.
Lilianna Dera: Methodology; Resources; Writing – review & editing.
Szymon Jęśko-Białek: Methodology; Resources; Writing – review & editing.
Jakub Łączkowski: Methodology; Resources; Writing – review & editing.
Hubert Marciniak: Methodology; Resources; Writing – review & editing.

Filip Nowicki: Methodology; Resources; Software; Writing – review & editing.

Bartosz Wilczek: Methodology; Resources; Writing – review & editing.

James J. Gross: Conceptualization; Methodology; Writing – original draft; Writing – review & editing.

Nicholas A. Coles: Conceptualization; Methodology; Writing – original draft; Writing – review & editing.

Declaration of Conflicting Interests

The authors declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

The National Science Centre in Poland (UMO-2020/39/B/HS6/00685; 2023/49/B/HS5/02139) and Excellence Initiative—Research University program at Adam Mickiewicz University, Poznan (140/04/POB5/0001, 151/12/POB5/0005, 174/12/POB5/0001, 177/02/UAM/0012, 181/13/SNS/0003, 198/12/POB5/0007) supported preparing this article with a research grant.

Open Practices

This article has received the badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

Maciej Behnke https://orcid.org/0000-0002-2455-4556

Michał Klichowski https://orcid.org/0000-0002-1614-926X

Marta Kowal https://orcid.org/0000-0001-9050-1471

Szymon Kupiński https://orcid.org/0000-0002-4704-6802

Aakash Chowkase https://orcid.org/0000-0001-6990-4698

Leonardo A. Aguilar https://orcid.org/0000-0001-9516-0557

Joao F. G. B. Takayanagi https://orcid.org/0000-0001-5982-6585

Ju Hee Park https://orcid.org/0000-0003-3031-0272

Ekaterine Pirtskhalava https://orcid.org/0000-0003-0345-0560

Yuki Yamada https://orcid.org/0000-0003-1431-568X

James J. Gross https://orcid.org/0000-0003-3624-3090

Nicholas A. Coles https://orcid.org/0000-0001-8583-5610

Acknowledgments

All data used for analysis and materials are available on OSF: https://doi.org/10.17605/OSF.IO/V8DKM. The preregistered analysis plan is available on OSF: https://doi.org/10.17605/OSF.IO/72RJ3. All analysis code (completed in Python and R) is available on OSF: https://doi.org/10.17605/OSF.IO/8P572. During the preparation of this work, the authors used Grammarly and ChatGPT models to improve readability and language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content. The funders had no role in study design, data collection, analysis, publishing decisions, or manuscript preparation.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/25152459251415336

References

Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, *96*, Article 104159. https://doi.org/10.1016/j.jesp.2021.104159

Barrett, L. F. (2022). Context reconsidered: Complex signal ensembles, relational meaning, and population thinking in psychological science. *American Psychologist*, *77*(8), 894–920. https://doi.org/10.1037/amp0001054

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Behnke, M., Kreibig, S. D., Kaczmarek, L. D., Assink, M., & Gross, J. J. (2022). Autonomic nervous system activity during positive emotions: A meta-analytic review. *Emotion Review*, *14*(2), 132–160. https://doi.org/10.1177/17540739211073084

Behnke M., Kłoskowski, M., Klichowski, M., Krzyżaniak, W., Szymański, K., Maciejewski, P., Chwiłkowska, P., Kowal, M., Jończyk, R., Nowak, J., Kupiński, S., Kunc, D., Saganowski, S., Chowkase, A., Guemaz, F., Kertechian, K. S., Maadal, A. I. M. T., Aguilar, L. A., Alayande, B. T., . . . Coles, N. A. (2025a). *Library of AI-Generated Affective Images (LAI-GAI)*. OSF. https://doi.org/10.17605/OSF.IO/V8DKM

Behnke M., Kłoskowski, M., Klichowski, M., Krzyżaniak, W., Szymański, K., Maciejewski, P., Chwiłkowska, P., Kowal, M., Jończyk, R., Nowak, J., Kupiński, S., Kunc, D., Saganowski, S., Chowkase, A., Guemaz, F., Kertechian, K. S., Maadal, A. I. M. T., Aguilar, L. A., Alayande, B. T., . . . Coles, N. A. (2025b). *Library of AI-Generated Affective Images (LAI-GAI), analysis code & outputs component*. OSF. https://doi.org/10.17605/OSF.IO/8P572

Behnke M., Kłoskowski, M., Klichowski, M., Krzyżaniak, W., Szymański, K., Maciejewski, P., Chwiłkowska, P., Kowal, M., Jończyk, R., Nowak, J., Kupiński, S., Kunc, D., Saganowski, S., Chowkase, A., Guemaz, F., Kertechian, K. S., Maadal, A. I. M. T., Aguilar, L. A., Alayande, B. T., . . . Coles, N. A. (2025c). *Library of AI-Generated Affective Images (LAI-GAI), images component*. OSF. https://doi.org/10.17605/OSF.IO/K8XVH

Bendall, R. C., Royle, S., Dodds, J., Watmough, H., Gillman, J. C., Beevers, D., Cassidy, S., Short, B., Metcalfe, P., Lomas, M. J., Graham-Kevan, D., & Gregory, S. E. (2025). The Salford Nature Environments Database (SNED): An open-access database of standardized high-quality pictures from natural environments. *Behavior Research Methods*, *57*(1), Article 21. https://doi.org/10.3758/s13428-024-02556-4

Berezina, E., Lee, A. S., Gill, C. M. H. D., & Chua, J. Y. (2024). Is a picture worth the same emotions everywhere? Validation of images from the Nencki Affective Picture System in Malaysia. *Discover Mental Health*, *4*(1), Article 61. https://doi.org/10.1007/s44192-024-00116-y

Black Forest Labs, Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., . . . Smith, L. (2025). *FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space*. arXiv. https://doi.org/10.48550/arXiv.2506.15742

Blechert, J., Meule, A., Busch, N. A., & Ohla, K. (2014). Food-pics: An image database for experimental research on eating and appetite. *Frontiers in Psychology*, *5*, Article 617. https://doi.org/10.3389/fpsyg.2014.00617

Bradley, M. M., & Lang, P. J. (1999). *International Affective Digitized Sounds (IADS): Stimuli, instruction manual and affective ratings* (Technical Report No. B-2). University of Florida, Center for Research in Psychophysiology.

Bradley, M. M., & Lang, P. J. (2007). The International Affective Picture System (IAPS) in the study of emotion and attention. In J. A. Coan, & J. J. B. Allen (Eds.), *Handbook of emotion elicitation and assessment* (pp. 29–46). Oxford University Press. https://doi.org/10.1093/oso/9780195169157.003.0003

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28, https://doi.org/10.18637/jss.v080.i01

Carretié, L., Tapia, M., López-Martín, S., & Albert, J. (2019). EmoMadrid: An emotional pictures database for affect research. *Motivation and Emotion*, *43*(6), 929–939. https://doi.org/10.1007/s11031-019-09780-y

Coles, N. A., March, D. S., Marmolejo-Ramos, F., Larsen, J. T., Arinze, N. C., Ndukaihe, I. L., Willis, M. L., Foroni, F., Reggev, N., Mokady, A., Forscher, P. S., Hunter, J. F., Kaminski, G., Yüvrük, E., Kapucu, A., Nagy, T., Hajdu, N., Tejada, J., Freitag, R. M., . . . Liuzza, M. T. (2022). A multi-lab test of the facial feedback hypothesis by the Many Smiles Collaboration. *Nature Human Behaviour*, *6*(12), 1731–1742. https://doi.org/10.1038/s41562-022-01458-9

Coles, N. A., Perz, B., Behnke, M., Eichstaedt, J., Kim, S. H., Vu, T. N., Raman, C., Tejada, J., Huynh, V.-T., Zhang, G., Cui, T., Podder, S., Chavda, R., Pandey, S., Upadhyay, A., Padilla-Buritica, J. I., Causil, C. J. B., Ji, L., Dollack, F., . . . Saganowski, S. (2025). Big team science reveals promises and limitations of machine learning efforts to model physiological markers of affective experience. *Royal Society of Open Science*, *12*(6), Article 241778. https://doi.org/10.1098/rsos.241778

Coles, N. A., Tenney, E. R., Chin, J. M., Friedrich, J. C., O'Dea, R. E., & Holcombe, A. O. (2024). Team scientists should normalize disagreement. *Science*, *384*(6700), 1076–1077. https://doi.org/10.1126/science.ado7070

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, *114*(38), E7900–E7909. https://doi.org/10.1073/pnas.1702247114

Cowen, A. S., Keltner, D., Schroff, F., Jou, B., Adam, H., & Prasad, G. (2021). Sixteen facial expressions occur in similar contexts worldwide. *Nature*, *589*(7841), 251–257. https://doi.org/10.1038/s41586-020-3037-7

Crone, D. L., Bode, S., Murawski, C., & Laham, S. M. (2018). The Socio-Moral Image Database (SMID): A novel stimulus set for the study of social, moral and affective processes. *PLOS ONE*, *13*(1), Article e0190954. https://doi.org/10.1371/journal.pone.0190954

Dan-Glauser, E. S., & Scherer, K. R. (2011). The Geneva Affective Picture Database: A new 730-picture database focusing on valence and normative significance. *Behavior Research Methods*, *43*(2), 468–477. https://doi.org/10.3758/s13428-011-0064-1

de Sousa Magalhaes, S., Miranda, D. K., de Miranda, D. M., Malloy-Diniz, L. F., & Romano-Silva, M. A. (2018). The Extreme Climate Event Database (EXCEED): Development of a picture database composed of drought and flood stimuli. *PLOS ONE*, *13*(9), Article e0204093. https://doi.org/10.1371/journal.pone.0204093

Demszky, D., Guntuku, S. C., & Ungar, L. H. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(9), 688–701. https://doi.org/10.1038/s44159-023-00234-9

Diconne, K., Kountouriotis, G. K., Paltoglou, A. E., Parker, A., & Hostler, T. J. (2022). Presenting KAPODI–The searchable database of emotional stimuli sets. *Emotion Review*, *14*(1), 84–95. https://doi.org/10.1177/17540739211072803

Diener, E., Cha, Y., & Oishi, S. (2023). Reinterpreting mood induction experiments. *The Journal of Positive Psychology*, *18*(3), 339–349. https://doi.org/10.1080/17439760.2022.2106544

Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion Review*, *3*(4), 364–370. https://doi.org/10.1177/1754073911410740

Fernandes, N. L., Pandeirada, J. N., & Nairne, J. S. (2019). Presenting new stimuli to study emotion: Development and validation of the Objects-on-Hands Picture Database. *PLOS ONE*, *14*(7), Article e0219615. https://doi.org/10.1371/journal.pone.0219615

Foroni, F., Pergola, G., Argiris, G., & Rumiati, R. I. (2013). The FoodCast Research Image Database (FRIDa). *Frontiers in Human Neuroscience*, *7*, Article 51. https://doi.org/10.3389/fnhum.2013.00051

Freepik. (2024). *Freepik image library and tools*. https://www.freepik.com/

Goodman, A. M., Katz, J. S., & Dretsch, M. N. (2016). Military Affective Picture System (MAPS): A new emotion-based stimuli set for assessing emotional processing in military populations. *Journal of Behavior Therapy and Experimental Psychiatry*, *50*, 152–161. https://doi.org/10.1016/j.jbtep.2015.07.006

Google Cloud. (2025, May 21). *Announcing Veo 3, Imagen 4, and Lyria 2 on Vertex AI*. https://cloud.google.com/blog/products/ai-machine-learning/announcing-veo-3-imagen-4-and-lyria-2-on-vertex-ai

Green, P., & MacLeod, C. J. (2016). simr: An R package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Gross, J. J., & Levenson, R. W. (1995). Emotion elicitation using films. *Cognition & Emotion*, *9*(1), 87–108. https://doi.org/10.1080/02699939508408966

Haberkamp, A., Glombiewski, J. A., Schmidt, F., & Barke, A. (2017). The DIsgust-RelaTed-Images (DIRTI) database: Validation of a novel standardized set of disgust pictures. *Behaviour Research and Therapy*, *89*, 86–94. https://doi.org/10.1016/j.brat.2016.11.010

Herbort, M. C., Iseev, J., Stolz, C., Roeser, B., Großkopf, N., Wüstenberg, T., Hellweg, R., Walter, H., Dziobek, I., & Schott, B. H. (2016). The ToMenovela – A photograph-based stimulus set for the study of social cognition with high ecological validity. *Frontiers in Psychology*, *7*, Article 1883. https://doi.org/10.3389/fpsyg.2016.01883

Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Forkel, R., Mucha, P. J., Greenhill, S. J., Gray, R. D., & Lindquist, K. A. (2019). Emotion semantics show both cultural variation and universal structure. *Science*, *366*(6472), 1517–1522. https://doi.org/10.1126/science.aaw8160

Jończyk, R., Boutonnet, B., Musiał, K., Hoemann, K., & Thierry, G. (2016). The bilingual brain turns a blind eye to negative statements in the second language. *Cognitive, Affective, & Behavioral Neuroscience*, *16*(3), 527–540. https://doi.org/10.3758/s13415-016-0411-x

Joseph, D. L., Nathan, D., & MacKinnon, S. (2020). The manipulation of affect: A meta-analysis of affect induction procedures. *Psychological Bulletin*, *146*(4), 355–375. https://doi.org/10.1037/bul0000224

Kamper, S. J., Maher, C. G., & Mackay, G. (2009). Global rating of change scales: A review of strengths and weaknesses and considerations for design. *Journal of Manual & Manipulative Therapy*, *17*(3), 163–170. https://doi.org/10.1179/jmt.2009.17.3.163

Kurdi, B., Lozano, S., & Banaji, M. R. (2017). Introducing the Open Affective Standardized Image Set (OASIS). *Behavior Research Methods*, *49*(2), 457–470. https://doi.org/10.3758/s13428-016-0715-3

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International Affective Picture System (IAPS): Affective ratings of pictures and instruction manual* (Technical Report No. A-8). University of Florida, Center for Research in Psychophysiology.

Libkuman, T. M., Otani, H., Kern, R., Viger, S. G., & Novak, N. (2007). Multidimensional normative ratings for the international affective picture system. *Behavior Research Methods*, *39*, 326–334. https://doi.org/10.3758/BF03193164

López-Caneda, E., & Carbia, C. (2018). The Galician Beverage Picture Set (GBPS): A standardized database of alcohol and non-alcohol images. *Drug and Alcohol Dependence*, *184*, 42–47. https://doi.org/10.1016/j.drugalcdep.2017.11.022

Marchewka, A., Żurawski, Ł., Jednoróg, K., & Grabowska, A. (2014). The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, realistic picture database. *Behavior Research Methods*, *46*(3), 596–610. https://doi.org/10.3758/s13428-013-0379-1

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Miccoli, L., Delgado, R., Guerra, P., Versace, F., Rodríguez-Ruiz, S., & Fernández-Santaella, M. C. (2016). Affective pictures and the Open Library of Affective Foods (OLAF): Tools to investigate emotions toward food in adults. *PLOS ONE*, *11*(8), Article e0158991. https://doi.org/10.1371/journal.pone.0158991

Miccoli, L., Delgado, R., Rodríguez-Ruiz, S., Guerra, P., García-Mármol, E., & Fernández-Santaella, M. C. (2014). Meet OLAF, a good friend of the IAPS! The Open Library of Affective Foods: A tool to investigate the emotional impact of food in adolescents. *PLOS ONE*, *9*(12), Article e114515. https://doi.org/10.1371/journal.pone.0114515

Michałowski, J. M., Droździel, D., Matuszewski, J., Koziejowski, W., Jednoróg, K., & Marchewka, A. (2017). The Set of Fear Inducing Pictures (SFIP): Development and validation in fearful and nonfearful individuals. *Behavior Research Methods*, *49*(4), 1407–1419. https://doi.org/10.3758/s13428-016-0797-y

Midjourney, Inc. (2024). *Midjourney text-to-image model*. https://www.midjourney.com/

Noon, M. S., Beaudry, J. L., & Knowles, A. (2019). The Crime and Threat Image Set (CaTIS): A validated stimulus set to experimentally explore fear of crime. *Journal of Experimental Criminology*, *15*(2), 227–242. https://doi.org/10.1007/s11292-017-9314-2

OpenAI. (2024a). *ChatGPT-4*. https://openai.com/

OpenAI. (2024b). *Improving image generation with better captions (DALL·E 3)* [White paper]. https://cdn.openai.com/papers/dall-e-3.pdf

Perz, B., & Kazienko, P. (2025). Personalization of affective state recognition from physiological signals: A review. In N. T. Nguyen, T. Matsuo, F. L. Gaol, Y. Manolopoulos, H. Fujita, T.-P. Hong, & K. Wojtkiewicz (Eds.), *Recent challenges in intelligent information and database systems* (pp. 143–158). Springer. https://doi.org/10.1007/978-981-96-5881-7_12

Peterson, H., Simpson, S. L., & Laurienti, P. J. (2019). Wake Forest Alcohol Imagery Set: Development and validation of a large standardized alcohol imagery dataset. *Alcoholism: Clinical and Experimental Research*, *43*(12), 2559–2567. https://doi.org/10.1111/acer.14214

Possidónio, C., Graça, J., Piazza, J., & Prada, M. (2019). Animal Images Database: Validation of 120 images for human-animal studies. *Animals*, *9*(8), Article 475. https://doi.org/10.3390/ani9080475

Pronk, T., van Deursen, D. S., Beraha, E. M., Larsen, H., & Wiers, R. W. (2015). Validation of the Amsterdam Beverage Picture Set: A controlled picture set for cognitive bias measurement and modification paradigms. *Alcoholism: Clinical and Experimental Research*, *39*(10), 2047–2055. https://doi.org/10.1111/acer.12853

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Saad, E., Broekens, J., & Neerincx, M. A. (2024). A little chit-chat goes a long way: Design and evaluation of task-and person-oriented styles for social robots. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)* (pp. 1712–1719). IEEE.

Saganowski, S., Perz, B., Polak, A. G., & Kazienko, P. (2022). Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review. *IEEE Transactions on Affective Computing*, *14*(3), 1876–1897. https://doi.org/10.1109/TAFFC.2022.3176135

Scaini, S., Rancoita, P. M., Martoni, R. M., Omero, M., Ogliari, A., & Brombin, C. (2017). Integrating dimensional and discrete theories of emotions: A new set of anger- and fear-eliciting stimuli for children. *The Journal of Genetic Psychology*, *178*(5), 253–261. https://doi.org/10.1080/00221325.2017.1351416

Schomaker, J., Rau, E. M., Einhäuser, W., & Wittmann, B. C. (2017). Motivational Objects in Natural Scenes (MONS): A database of >800 objects. *Frontiers in Psychology*, *8*, Article 1669. https://doi.org/10.3389/fpsyg.2017.01669

Shankland, R., Favre, P., Corubolo, D., Méary, D., Flaudias, V., & Mermillod, M. (2019). Food-Cal: Development of a controlled database of high and low calorie food matched with non-food pictures. *Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity*, *24*, 1041–1050. https://doi.org/10.1007/s40519-019-00687-8

Siegel, E. H., Sands, M. K., Dy, J., & Barrett, L. F. (2018). Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories. *Psychological Bulletin*, *144*(4), 343–393. https://doi.org/10.1037/bul0000128

Stevenson, R. A., Mikels, J. A., & James, T. W. (2007). Characterization of the affective norms for English words by emotional categories. *Behavior Research Methods*, *39*(4), 1020–1024. https://doi.org/10.3758/BF03192999

Szymanska, M., Comte, A., Tio, G., Vidal, C., Monnin, J., Smith, C. C., Nezelof, S., & Vulliez-Coady, L. (2019). The Besançon Affective Picture Set-Adult (BAPS-Adult): Development and validation. *Psychiatry Research*, *271*, 31–38. https://doi.org/10.1016/j.psychres.2018.11.005

Szymanska, M., Monnin, J., Noiret, N., Tio, G., Galdon, L., Laurent, E., Nezelof, S., & Lauriane Vulliez-Coady. (2015). The Besançon Affective Picture Set-Adolescents (the BAPS-Ado): Development and validation. *Psychiatry Research*, *228*(3), 576–584. https://doi.org/10.1016/j.psychres.2015.04.055

Team Seedream, Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., Jian, X., Kuang, H., Lai, Z., Li, F., Li, L., Lian, X., Liao, C., Liu, L., Liu, W., . . . Zhu, W. (2025). *Seedream 4.0: Toward next-generation multimodal image generation*. arXiv. https://doi.org/10.48550/arXiv.2509.20427

Teh, E. J., Yap, M. J., & Liow, S. J. R. (2018). PiSCES: Pictures with social context and emotional scenes with norms for emotional valence, intensity, and social engagement. *Behavior Research Methods*, *50*(5), 1793–1805. https://doi.org/10.3758/s13428-017-0947-x

Toet, A., Kaneko, D., de Kruijf, I., Ushiama, S., van Schaik, M. G., Brouwer, A.-M., Kallen, V., & van Erp, J. B. F. (2019). CROCUFID: A cross-cultural food image database for research on food elicited affective responses. *Frontiers in Psychology*, *10*, Article 58. https://doi.org/10.3389/fpsyg.2019.00058

Weierich, M. R., Kleshchova, O., Rieder, J. K., & Reilly, D. M. (2019). The Complex Affective Scene Set (COMPASS): Solving the social content problem in affective visual stimulus sets. *Collabra: Psychology*, *5*(1), Article 53. https://doi.org/10.1525/collabra.256

Wu, Y. J., & Thierry, G. (2012). How reading in a second language protects your heart. *The Journal of Neuroscience*, *32*(19), 6485–6489. https://doi.org/10.1523/jneurosci.6119-11.2012

Zhang, W., Jończyk, R., Wu, Y. J., Lan, Y., Gao, Z., Hu, J., Thierry, G., & Gao, S. (2023). Brain potentials reveal how emotion filters native language access when bilinguals read words in their second language. *Cerebral Cortex*, *33*(13), 8783–8791. https://doi.org/10.1093/cercor/bhad161