MOTIVATION.

[question1] For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

[answer1] The authors describe the dataset as the first attempt to associate visual art with human affective judgements and natural language explanations of the rationale behind each affective judgement.
They motivate the development of this dataset by identifying "the formation of linguistic affective explanations grounded on visual stimuli" as an under-explored problem in computer vision.

[question2] Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?

[answer2] The dataset is authored by 5 researchers distributed across 3 institutions: Panos Achlioptas (Stanford
University), Maks Ovsjanikov (LIX, Ecole Polytechnique, IP Paris), Kilichbek Haydarov (King
Abdullah University of Science and Technology), Mohamed Elhoseiny (Stanford University; King
Abdullah University of Science and Technology), Leonidas Guibas (Stanford University).

[question3] Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

[answer3] The work was funding by a Vannevar Bush Faculty Fellowship, a KAUST BAS/1/1685-01-01, a CRG-2017-3426, the ERC Starting Grant No. 758800 (EXPROTEA) and the ANR AI Chair AIGRETTE,
and gifts from the Adobe, Amazon AWS, Autodesk, and Snap corporations.

[question4] Any other comments?

[answer4] No.

COMPOSITION.

[question5] What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

[answer5] The data instances are images of artworks in various media—such as paintings, sculptures, and photographs—although not all artistic media are represented. The ArtEmis developers claim that at the time of their download from WikiArt in order to build the new dataset, the instances were drawn
from 80,031 unique entries representing the work of 1,119 artists.

[question6] How many instances are there in total (of each type, if appropriate)?

[answer6] Wikiart contains 81,446 artworks from 1,119 artists (as downloaded in 2015). The artworks cover
27 art-styles (abstract, baroque, cubism, impressionism, etc.) and 45 genres (cityscape, landscape,
portrait, still life, etc.). Details can be found in [79].

[question7] Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).

[answer7] The ArtEmis dataset is built on a download fromWikiArt that occurred in 2015. [32, 79] The ArtEmis developers state that, at the time of download, WikiArt contained 81,446 data instances of which they identified 1,415 as exact duplicates. To find and eliminate redundancies, they used the fdupes program [80] and performed manual inspection on pairs of nearest-neighbors artworks by using features of a ResNet-32, pretrained on ImageNet. Afterward, 80,031 instances remained for annotation and analysis. The ArtEmis dataset therefore represents a non-random sample of the current version of Wikiart, which at the time of this writing in 2021 contains some 250,000 data instances. [78]
Under a broader historical lens, the ArtEmis dataset represents multiple layers of curation, operating cumulatively, that resulted in the makeup of WikiArt at the moment in 2015 when it was used as a basis for the new dataset. ArtEmis does not reach beyond WikiArt for sources; hence, it reiterates environmental, sociopolitical, and professional factors that shaped WikiArt as of 2015. Numbering about 80,000 works, WikiArt is predictably a reduction of the total corpus of artistic creation, i.e., all works made globally throughout human history. Both WikiArt and ArtEmis necessarily exclude artworks that were destroyed or lost before the invention of photography. The datasets focus on a limited number of artistic media, styles, historical periods, and geographies. A high-level review suggests that their composition centers on modern Western artistic traditions and preferences, prioritizing certain media and styles from the six most recent centuries and only two continents (North America and Europe). The precise reasons for this reductive result are difficult to document due to the low availability of public information regarding the historical architecture of WikiArt. Generally, a series of choices made by its development team and external contributors determined which artworks are included versus not. In turn, such choices were constrained by prior factors that affected the composition of art repositories, such as those of GLAMs, from which WikiArt derived its data.

[question8] What data does each instance consist of? "Raw" data (for example, unprocessed text or images) oer features? In either case, please provide a description.

[answer8] Each data instance includes natural
language labels provided by annotators who were sourced by the ArtEmis developers to contribute descriptions of their personal emotional responses when they viewed the image derived from WikiArt.

[question9] Is there a label or target associated with each instance? If so, please provide a description.

[answer9]

[question10] Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.

[answer10]

[question11] Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

[answer11]

[question12] Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

[answer12]

[question13] Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

[answer13]

[question14] Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

[answer14]

[question15] Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

[answer15]

[question16] Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

[answer16] The original Wikiart does contain sensitive information such as those concerning religious beliefs and political opinions, and artworks that could potentially trigger anxiety such as scenes from wars.
ArtEmis, also includes artworks that contain such sensitive information, and those that cause anxiety.
For example, see artworks corresponding to categories labeled as "disgust", and "fear" in Figure 6 of the paper

If the dataset does not relate to people, you may skip the remaining questions in this section.

[SkipA] NO

[question17] Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

[answer17] The dataset does not consider textual inscriptions associated with the artworks, and therefore does
not identify any sub-populations that could have been potentially identified by virtue of such textual
metadata. However, the annotation process involved in curating ArtEmis does identify sub-populations (e.g., by
profession, social status, gender, etc.). For example, one annotation reads "The peaceful look of the
aristocratic individuals makes you wonder about their lives."
The annotations contained in the dataset are not necessarily backed by their historical context, instead
they reflect the subjective opinions of the AMT workers with regards to the emotion triggered in
them, looking at the artwork.

[question18] Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.

[answer18]

[question19] Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

[answer19]

[question20] Any other comments?

[answer20] No.

COLLECTION PROCESS.

[question21] How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

[answer21] The individual visual artworks were obtained from the publicly available Wikiart dataset. The original Wikiart dataset contains about 250000 artworks (as of 2021) by 3000 artists across 8 languages (https://www.wikiart.org/en/about). The artworks in Wikiart are from museums, universities, town halls, private collections, and other civic buildings of more than 100 countries. The annotations of the artworks in ArtEmis were obtained by AMT workers– at least 5 annotators were asked to annotate each artwork to express their dominant emotional reaction to the visual artwork along with an utterance explaining the reason behind their response.

[question22] What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

[answer22]

[question23] If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?

[answer23]

[question24] Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?

[answer24] It seems the authors and the AMT workers are the only stakeholders involved in the data collection process.

[question25] Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

[answer25]

[question26] Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

[answer26]

If the dataset does not relate to people, you may skip the remaining questions in this section.

[SkipB] NO

[question27] Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?

[answer27]

[question28] Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

[answer28]

[question29] Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

[answer29]

[question30] If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

[answer30]

[question31] Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

[answer31]

[question32] Any other comments?

[answer32] No.

PREPROCESSING/CLEANING/LABELING.

[question33] Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so,

please provide a description. If not, you may skip the remaining questions in this section.

[answer33] When displaying the image of an artwork in AMT, the authors scaled down the largest size of the image to 600 pixels, keeping the original aspect-ratio (they did not apply any scaling if the largest size was less than 600 pixels). The authors mention that this "scaling was done to homogenize the presentation of our visual stimuli, and crucially to also reduce the loading and scrolling time required with higher resolution images." Please see https://www.artemisdataset.org/materials/ artemis_supplemental.pdf [79] for details. The average length of the captions of ArtEmis is 15.8 words. On average, each caption contained 4
nouns, 0.9 pronouns, 1.6 adjectives, 3 verbs, and 1.9 adpositions.
With regards to the diversity of captions per image, the number of unique nouns (normalized averages) in an image caption was 3.4, unique pronouns (normalized averages) was 0.6, unique adjectives (normalized averages) was 1.5, unique verbs (normalized averages) was 2.4 , and unique adpositions (normalized averages) was 1.2
To the best of our knowledge, the AMT workers were specifically not instructed to write captions satisfying certain criteria such as minimum word count, parts of speech, etc.
The AMT workers were not instructed to specifically include historical, cultural, social, religious or other such contexts in labeling the artworks. Instead, they were asked to identify the dominant emotion triggered by the artwork (in their opinion), and also to provide an explanation for the same. Detailed rater guidelines are provided in the supplementary material [79].
The labels were not obtained from existing GLAMs.

[question34] Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
[answer34]

[question35] Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.
[answer35]

[question36] Any other comments?
[answer36] No.

USES.

[question37] Has the dataset been used for any tasks already? If so, please provide a description.
[answer37]

[question38] Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
[answer38]

[question39] What (other) tasks could the dataset be used for?
[answer39]

[question40] Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example,

stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

[answer40]

[question41] Are there tasks for which the dataset should not be used? If so, please provide a description.

[answer41]

[question42] Any other comments?

[answer42] No.

DISTRIBUTION.

[question43] Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

[answer43] The following are the terms and conditions for the use and distribution of the ArtEmis dataset https://www.artemisdataset.org/materials/artemis_terms_of_use.txt
The code is released under MIT license https://www.artemisdataset.org/materials/MIT_
license.txt

[question44] How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

[answer44] All the artworks are sourced from Wikiart. According to the copyright terms of Wikiart, copyrighted items are also included , and are showcased under the fair use principles as listed here https://www.wikiart.org/en/terms-of-use

[question45] When will the dataset be distributed?

[answer45]

[question46] Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

[answer46] WikiArt presents both public domain and copyright protected artworks. Copyright protected artworks
are showcased in accordance with fair use principles. Additional details on copyright policy and
responses to copyright violations can be found at: https://www.wikiart.org/en/about.
The ArtEmis dataset is licensed under the ArtEmis Terms of Use: https://www.artemisdataset.
org/materials/artemis_terms_of_use.txt

[question47] Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

[answer47] There is no explicit information available about IP rights and cost of acquisition of individual artworks.
https://www.wikiart.org/en/terms-of-use lists the terms of use of the originalWikiart from

which the artworks in ArtEmis were sourced. The terms of use of ArtEmis can be found at https://www.artemisdataset.org/materials/artemis_terms_of_use.txt. Several groups of stakeholders are affected by the authors' choice to make the dataset and code
available for distribution under terms of use and a MIT license. Each adoption of the dataset by new parties introduces greater distance between the end-users and people attached to the artworks' creation, cultural significance, and care. The artworks' communities of origin—including artists, audiences, and supporters—are not ensured of continual connection as named individuals or groups with the data. Instead, human associations with the artworks are permitted to be stripped away by new users' design and development decisions about which labels to retain versus discard. The dataset distribution presents further opportunity to loosen ties between the artworks and stakeholders by inserting optionality into whether information is kept for all data instances about their histories with specific art collectors and owners, groups associated with the artworks' reception and location, and caretakers (such as curators and conservators), to name a few examples. In such scenarios, human connections with artworks are potentially forfeited, a foreseeable risk inherent in the authors' decision to share the dat Artaset while releasing considerable responsibility for how the data and metadata will be used, interpreted, and distributed again. By contrast, the authors of thet and their institutions stand to benefit through the mechanism of citation, whether in scholarly literature or broader channels.

[question48] Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

[answer48]

[question49] Any other comments?

[answer49] No.

MAINTENANCE.

[question50] Who will be supporting/hosting/maintaining the dataset?

[answer50] The dataset is maintained by researchers Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas of Stanford University, King Abdullah University of Science and Technology (KAUST), and Ecole Polytechnique (the "Universities")

[question51] How can the owner/curator/manager of the dataset be contacted (for example, email address)?

[answer51] The authors can be contacted at artemis:dataset@gmail:com

[question52] Is there an erratum? If so, please provide a link or other access point.
[answer52]

[question53] Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?

[answer53] The dataset was recently curated; currently there is no information regarding dataset update.

[question54] If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period

of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

[answer54]

[question55] Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

[answer55]

[question56] If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

[answer56] It looks like others cannot contribute to the ArtEmis dataset at this time.

[question57] Any other comments?

[answer57] No.