Position: Measure Dataset Diversity, Don't Just Claim It

Dora Zhao [* 1] **Jerone T. A. Andrews** [2] **Orestis Papakyriakopoulos** [* 3] **Alice Xiang** [4]

Abstract

Machine learning (ML) datasets, often perceived as neutral, inherently encapsulate abstract and disputed social constructs. Dataset curators frequently employ value-laden terms such as diversity, bias, and quality to characterize datasets. Despite their prevalence, these terms lack clear definitions and validation. Our research explores the implications of this issue by analyzing "diversity" across 135 image and text datasets. Drawing from social sciences, we apply principles from measurement theory to identify considerations and offer recommendations for conceptualizing, operationalizing, and evaluating diversity in datasets. Our findings have broader implications for ML research, advocating for a more nuanced and precise approach to handling value-laden properties in dataset construction.

1. Introduction

Cloaked under the guise of objectivity, machine learning (ML) datasets are portrayed as impartial entities, giving the illusion of reflecting an "unbiased look" at the world (Torralba & Efros, 2011). Yet, beneath this veneer, datasets are not neutral—they are infused with values, bearing the indelible imprints of social, political, and ethical ideologies woven into their fabric by their curators (Raji et al., 2021; Blili-Hamelin & Hancox-Li, 2023; Malevé, 2021).

This inherent value-laden nature becomes glaringly apparent in the perpetuation of social stereotypes and the stark underrepresentation of marginalized communities within the lifecycle of ML datasets (Wang et al., 2022; Buolamwini & Gebru, 2018; Zhao et al., 2021; Birhane et al., 2021; Denton et al., 2020). From inception to release, datasets emerge as political artifacts, etched with the signature of their creators'

---

[*]Work done at Sony AI [1]Stanford University, Stanford, USA [2]Sony AI, London, UK [3]Technical University of Munich, Munich, Germany [4]Sony AI, Seattle, USA. Correspondence to: Dorothy Zhao <dorothyz@stanford.edu>.

perspectives, organizational priorities, and the broader cultural zeitgeist, making them potent instruments in shaping narratives and reinforcing power structures (Winner, 2017; Hanna & Park, 2020; Birhane et al., 2022).

This politicization of datasets is particularly conspicuous in the criteria set by curators. Terms related to diversity, bias, quality, realism, difficulty, and comprehensiveness are frequently invoked (Scheuerman et al., 2021), despite a glaring lack of consensus regarding their precise definitions. For instance, diversity dimensions can encompass a multitude of concepts, spanning "dressing styles" (Bai et al., 2021), "weather" (Diaz et al., 2022), and "ethnic[ity]" (Fu et al., 2021) to "verbs" (Sadhu et al., 2021), "sentential contexts" (Culkin et al., 2021), and "conversation forms" (Fabbri et al., 2021a). Diversity can also refer to part of the collection process, such as recruiting annotators with "diversity in gender, age, occupation/background (linguistic and ethnographic knowledge), region (spoken dialects)" (Zeinert et al., 2021b), and "psychological personality" (Chawla et al., 2021).

Recognizing this ambiguity, the need for precise and unambiguous definitions becomes paramount to ascertain whether datasets genuinely embody the proclaimed qualities. Treating value-laden constructs, such as diversity, bias, and quality, as self-evident perpetuates the fallacious belief that datasets are inherently neutral. Instead, we posit that datasets serve as tools wielded by curators to quantify abstract social constructs. Interrogating these values demands critical questions: *How are these constructs defined and operationalized? And how do we validate that datasets genuinely encapsulate the values they claim to represent?*

In this position paper, we leverage *measurement theory*, a framework widely employed in the social sciences, to develop precise numerical representations of abstract constructs (Bandalos, 2018). This application is integral to our focused analysis of *diversity*—a frequently touted trait in ML datasets (Scheuerman et al., 2019)—providing a structured approach to conceptualizing, operationalizing, and evaluating claimed dataset qualities. Our scrutiny extends to 135 text and image datasets, where we uncover key considerations and offer recommendations for applying measurement theory to their collection. We underscore the imperative for transparency in articulating how diversity

is defined (Section 4) and how the data collection process aligns with this definition (Section 5). Further, we present methodologies for evaluating diversity, scrutinized through the lenses of reliability (Section 6) and validity (Section 7). Our work culminates with a case study on the Segment Anything dataset (Kirillov et al., 2023), illustrating the practical application of our recommendations (Section 8).

In summary, this position paper advocates for clearer definitions and stronger validation methods in creating diverse datasets, proposing measurement theory as a scaffolding framework for this process.

2. Background

Measurement is a fundamental aspect of ML systems (Jacobs & Wallach, 2021; Jacobs, 2021). These systems employ observable, real-world instances to quantify abstract constructs, including "moral foundations" (Johnson & Goldwasser, 2018), "emotion" (Wei et al., 2020), and "gender" (Wang et al., 2019a). However, given the unobservable nature of these constructs, researchers rely on proxies and inference. For instance, the identification and prevalence of specific linguistic features, like derogatory language, serve as proxies to deduce the existence of misogynistic content in text (Zeinert et al., 2021a). Assessing the quality of resulting datasets requires careful consideration of the validity of chosen proxies and the reliability of assumptions underlying ML systems. This is crucial, as proxies may normalize inadequacies without acknowledging limitations (Andrus et al., 2021). Such considerations lead us to measurement theory in the social sciences, which offers methodologies for quantifying and encapsulating theoretical constructs that resist direct measurement (Bandalos, 2018).

Conceptually, measurement theory provides a structured approach to move from latent, abstract constructs to observable, real-world variables. *Conceptualization* involves precisely defining constructs using agreed-upon terms. Researchers then *operationalize* these concepts by translating them into observable indicators that can be empirically measured in the real world (Babbie, 2020; Check & Schutt, 2011). Finally, the measurements undergo *evaluation*, considering *reliability* and *validity* (Bandalos, 2018).

Recent ML research explores the application of measurement modeling to refine the conceptualization and operationalization of constructs, including fairness (Jacobs & Wallach, 2021), bias (Jacobs et al., 2020) and intelligence (Blili-Hamelin & Hancox-Li, 2023). This trend underscores a growing emphasis on leveraging measurement theory to refine evaluation metrics and benchmarks (Xiao et al., 2023; Zhou et al., 2022; Subramonian et al., 2023). Principles from measurement theory have also been applied to datasets, where Mitchell et al. (2022) suggest measuring various

facets to facilitate dataset curation and meaningful comparisons.

Complementarily, we leverage measurement theory to enhance ML datasets by transforming implicit, value-laden properties into measurable constructs. In contrast to previous work (Blodgett et al., 2021) that concentrated on assessing the validity of natural language processing benchmark datasets to evaluate stereotyping, we adopt a different approach. Here, we view the data collection process *itself* as the measurement to validate. This involves questioning the reliability of the collection process and exploring methods for researchers to validate claimed dataset properties.

Zooming in on the intricacies of diversity, we unearth inconsistencies in its definitions, a pervasive challenge that our paper addresses head-on. As in other fields with broad definitions of diversity (Xu et al., 2020), there is no consensus on what diversity means for ML datasets. Rather than offering a prescriptive definition, we adopt the various interpretations provided by dataset collectors and develop an inductive taxonomy based on these definitions (see Table 5). While many ML datasets claim to be more diverse, neither what researchers mean by the term diversity nor how this property is achieved are clearly specified. Our aim, by shedding light on these disparities, is not only to guide dataset creators but also to equip reviewers with the necessary insights to critically evaluate authors' claims. Our work extends beyond the confines of dataset creation, presenting a broader contribution to the enhancement of ML and scientific practices. We summarize our considerations and recommendations in Figure 1.

3. Method

To inform our position, we conducted a systematic literature review encompassing 135 datasets presented as being more "diverse". We choose to focus on image and text datasets, aligning with the prevailing emphasis on data collection practices in these domains (Scheuerman et al., 2021; Paullada et al., 2021; Raji et al., 2021). Datasets were identified through searches across well-established venues in computer vision (ICCV, ECCV, CVPR), natural language processing (*CL), fairness (FAccT, AIES), and ML (NeurIPS, NeurIPS Datasets and Benchmarks Track), concluding with publications available until September 2022. Inclusion criteria involved retaining papers featuring both a diversity-related keyword ("divers*", "bias*") and dataset-related keyword ("dataset", "annotation") within their abstracts.

Our comprehensive analysis traverses the entire dataset collection pipeline, ranging from the motivation behind dataset creation to its eventual release and ongoing maintenance. To accomplish this, we consulted established guidelines for responsible data collection (Andrews et al., 2023; Scheuerman
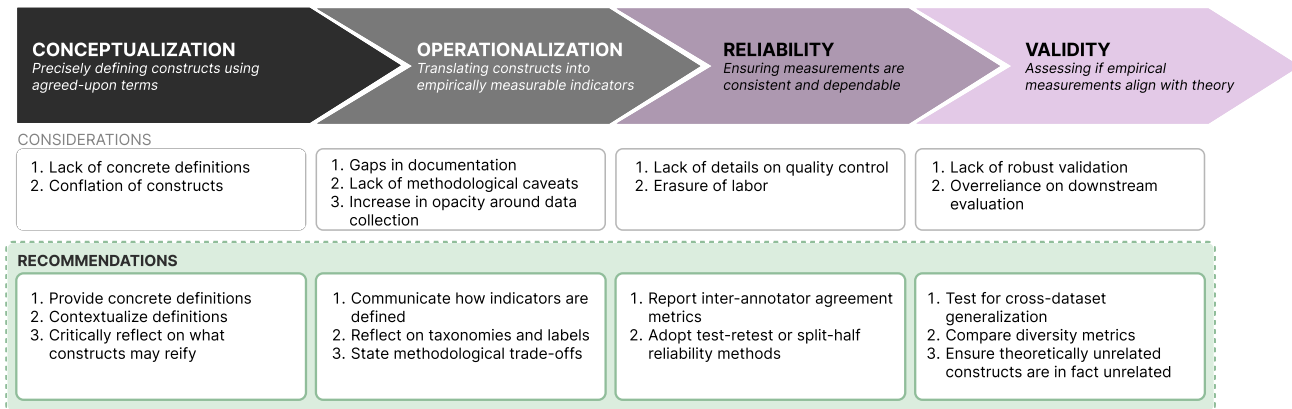
Figure 1. Overview of considerations and recommendations for conceptualizing (Section 4), operationalizing (Section 5), and evaluating the reliability (Section 6) and validity (Section 7) of diverse datasets.

et al., 2021; Peng et al., 2021; Google PAIR, 2019; Gebru et al., 2021; Holland et al., 2020; Bender & Friedman, 2018; Pushkarna et al., 2022; Blodgett et al., 2022). The evolution of our codebook variables occurred iteratively during the coding process, driven by ongoing discussions among the authors. Refer to Appendix A for more details.

4. Conceptualization

Conceptualization plays a pivotal role in the research process, involving the definition and specification of clear constructs to be measured (Check & Schutt, 2011; Babbie, 2020). For dataset creators, this phase resembles the translation of abstract values, such as diversity, into tangible and concrete definitions. Despite the widely acknowledged importance of diversity as a value for datasets (Van Horn et al., 2021; Yang et al., 2021; Sugawara et al., 2022; Derczynski et al., 2016), there is a notable lack of consistency and clarity among creators regarding its practical interpretation. This section examines conceptual challenges and provides recommendations.

4.1. Considerations

Lack of concrete definitions. Recognizing the inherently abstract nature of diversity, a well-defined concept not only clarifies the significance of a diverse dataset but also lays the groundwork for operationalizing the collection process. Significantly, only 52.9% of datasets explicitly justify the need for diverse data. For example, Rojas et al. (2022) underscore the need for geodiverse data, driven by observations that previous datasets "suffer from amerocentric and eurocentric representation bias that impacts the performance of classification tasks on images from other regions". In this context, *geodiversity* serves as a clearly defined construct, specifying the criteria for diversity based on geographical

locations. Such constructs play a crucial role in streamlining the operationalization of data collection (e.g., employing photographers across 63 countries), facilitating a more straightforward validation process against the intended motivation. In essence, unambiguous definitions empower both consumers and reviewers to assess a dataset's suitability for specific tasks with heightened confidence.

Conflation of constructs. The lack of a standardized conceptualization approach can lead to unintended consequences, such as the paradoxical increase in offensive content observed in larger datasets (Birhane et al., 2023). This stems from misinterpretations where curators conflate *scale* with diversity, assuming increased size inherently leads to more diverse data (Bender et al., 2021). Scale may help in some contexts, such as achieving composition diversity, but relying solely on "scale-thinking" might overlook crucial nuances and idiosyncrasies inherent in specific tasks or setups (Diaz & Madaio, 2023; Hanna & Park, 2020).

There is also a tendency to conflate *bias* with diversity (Scheuerman et al., 2021), exemplified by attempts to mitigate bias in datasets for text summarization. As Kim et al. (2019) report, prior works "alleviate [extractive] bias by collecting articles from diverse news publications." Establishing clear conceptual frameworks is not only a methodological necessity but also a crucial means to navigate complex relationships, recognizing that surface-level connections between diversity and other constructs, such as scale and bias, are not causally related.

4.2. Recommendations

Provide concrete definitions. Following the conceptualization phase, curators are urged to establish a precise definition of diversity, ensuring alignment and consistency throughout the dataset collection process (Check & Schutt, 2011; Bab-

bie, 2020). Emphasizing the importance of clear definitions, consider the subtle distinction between seemingly similar claims, such as representing "diverse scenarios" (Miao et al., 2022b) and "diverse social scenes" (Fan et al., 2018). The former is defined in terms of things, stuff, and scene location (i.e., "indoor and outdoor"), whereas the latter in terms of people, cultures, and scene location (i.e., "living room, kitchen, restaurant, ..."). Definitional disagreements naturally arise, but providing an explicit definition signals the interpretation of diversity within the context of a dataset, enhancing assessment and enabling meaningful cross-dataset comparisons.

Contextualize definitions. When crafting definitions, curators must align interpretations with existing literature, evaluating theoretical underpinnings (Blodgett et al., 2020) and building on prior scholarship. In addressing *skin tone* diversity, Hazirbas et al. (2021) critically assess the drawbacks of ethnicity labeling practices in a previous facial attribute dataset (Karkkainen & Joo, 2021), highlighting its subjectivity and potential to cause conceptual confusion. This scrutiny extends to citing "unconscious biases" linked to the "other-race effect" identified in psychology (O'Toole et al., 1996). In response, Hazirbas et al. (2021) opt to annotate apparent skin tone, using the Fitzpatrick scale (Fitzpatrick, 1975), acknowledging limitations regarding the scale's reliability and validity (Howard et al., 2021). Such contextualization plays a crucial role in positioning how a dataset *contributes* to or *challenges* prevailing notions.

Critically reflect on constructs. Definitions hold power. The choices we make in defining constructs, or even the decision to define certain constructs, bestow legitimacy upon specific beliefs. Thus, prior to advancing with the collection and release of a dataset, it is crucial to engage in a thoughtful reflection on the potential for *reification*—the act of treating an abstract concept as something concrete (Bhattacherjee, 2012). This tendency is frequently illustrated by the use of tangible test scores, such as those from IQ tests, to represent abstract concepts like "intelligence" (Blili-Hamelin & Hancox-Li, 2023).

This concern gains particular significance in the realm of demographic diversity. For example, casting gender as binary reaffirms the normative notion of a binary gender system (Hamidi et al., 2018), and using racial categories may inadvertently endorse the fallacy that these groups are natural rather than social (Hanna et al., 2020; Khan & Fu, 2021).

Parallel considerations extend to text datasets. Nine surveyed text datasets (Orbach et al., 2021; Abdul-Mageed et al., 2021; Fabbri et al., 2021a; Peskov et al., 2019; Durmus et al., 2019; Derczynski et al., 2016; Ide et al., 2008; Rahman et al., 2021; Sugawara et al., 2022) underscore *source diversity* by drawing from various topic domains or websites. It is crucial to recognize that different source

media inherently introduce their own values. Take, for example, the Broad Twitter Corpus (Derczynski et al., 2016), a diverse collection of English-speaking social media content sourced from the US, UK, New Zealand, Ireland, Canada, and Australia. Although the authors "were constrained by the number of local crowd workers" accessible, the exclusion of English-speaking countries, "such as Botswana and Singapore", confines the dataset to a Western-centric view.

5. Operationalization

Operationalization involves the meticulous development of methodologies to empirically measure abstract concepts (Check & Schutt, 2011; Babbie, 2020). In the context of ML datasets, this manifests as the tangible process of accumulating instances for a dataset. Within our corpus, we identify five primary dataset types categorizable by collection methodology: derivatives, "real-world" sampled, synthetically generated, web scraped, and crowdsourced (refer to Appendix C for further details). In this section, we spotlight deficiencies in current collection processes and present recommendations to surmount these issues.

5.1. Considerations

Gaps in documentation. A significant concern centers on the insufficiency of information provided regarding datasets. Consistent with prior work (Scheuerman et al., 2021), we observe that most papers introduce not only a dataset but also a new model, task, or algorithm. In particular, out of the 135 papers analyzed, only 40 are standalone dataset papers, or those where the dataset is positioned as the *primary* contribution. Consequently, limited space is allocated for dataset creators to furnish detailed insights into collection strategies or the rationale behind methodological choices. This lack of (thorough) documentation results in unclear definitions from curators, creating uncertainty about the type of diversity sought and how these concepts are operationalized. For instance, among the 21 "real-world" sampled datasets, information on the location and time of data capture, as well as the identity of the data collectors (e.g., authors, researchers), is missing in 13 instances. Similarly, for web-scraped datasets, papers often omit collection criteria, such as the keyword search queries, crucial for understanding the dataset's sampled distribution and potential biases, be they social or methodological.

This *documentation gap* is, in part, indicative of cultural attitudes prevalent in the ML research. As echoed previously, there exists a tendency to undervalue data work compared to model work or algorithmic contributions (Sambasivan et al., 2021). Papers exclusively dedicated to datasets typically do not find their way into "top-tier" research venues (Scheuerman et al., 2021; Heinzerling, 2022), potentially dissuading authors from allocating sufficient time and space to elabo-

rate on their dataset creation processes.

Lack of methodological caveats. Even when datasets offer insights into their collection processes, methodological considerations or limitations are seldom addressed, accounting for 74.8% of cases. Nonetheless, it is essential to acknowledge that every data collection method has inherent drawbacks. For example, web scraping offers a fast and cost-effective means to amass large quantities of data (Ramaswamy et al., 2024; Raji & Fried, 2021; Li, 2023). However, concerns arise that this method might only capture "canonical" perspectives, particularly when utilizing Internet search engines (Barbu et al., 2019; Zhu et al., 2017). This issue may be exacerbated by, e.g., consensus-based quality filtering, which tends to *exclude* noncanonical examples (Mayo et al., 2023).

In contrast, synthetic generation, offering similar benefits to web scraping with increased systematic control (Johnson et al., 2017; Wood et al., 2021; Ros et al., 2016; Wei et al., 2020), may introduce a domain gap between generated and real-world data (Wei et al., 2020; Ros et al., 2016). Inclusion of these considerations is crucial as it justifies why the chosen method is best suited for capturing the proposed definition of diversity.

Increase in opacity. We posit that the identified issues will worsen as data collection processes become more opaque. A clear manifestation of this trend is the increasing reliance on third-party assistants in data collection efforts (Orbach et al., 2021; Dave et al., 2020; Miceli et al., 2022b). While this approach may improve quality and efficiency, it introduces a layer of separation between those commissioning a dataset and those collecting it. This separation results in the loss of detailed knowledge, such as participant recruitment practices and quality assessment criteria, necessary for evaluating the validity of the collection process and the utility of the dataset. Furthermore, closed-source models and datasets raise additional concerns (Bommasani et al., 2023). Examples include closed- and open-source models such as GPT-4 (Achiam et al., 2023), CLIP (Radford et al., 2021), Gemini (Team et al., 2023), and Llama 2 (Touvron et al., 2023). This growing opacity can lead to serious issues, as model consumers are unable to audit training datasets for potential biases or critique their construction methodology.

5.2. Recommendations

Clearly define variables of interest. As highlighted by Raji et al. (2021), there is a prevailing inclination toward generality in ML research. This is evident in benchmarks claiming to evaluate "general-purpose" capabilities and datasets attempting to encompass diversity across numerous axes. However, just as it is impractical for a dataset to capture all the nuances and complexities of the world, it is equally

implausible to collect diversity across every conceivable dimension (Raji et al., 2021). For instance, defining diversity as "variety across writing styles" encompasses aspects such as, but not limited to, "genre" (Soldan et al., 2022), "narrative elements" (Xu et al., 2020), or "passage source, length, and readability" (Sugawara et al., 2022).

Rather than pursuing generality, we advocate for curators to distinctly select and communicate specific variables that are most relevant to the task their dataset is meant to serve. Their definition of diversity, insights from prior datasets, considerations of the limitations in their collection practices, as well as rationale for excluding plausible aspects of diversity can inform this selection. For example, Van Horn et al. (2021) justify the decision to restrict the number of Animalia, Plantae and Fungi species in iNat2021 to those "observed 'enough' times by 'enough' people" during a one-year time period, while transparently acknowledging the arbitrariness of their enumeration of "enough".

Communicate how indicators are defined. We encourage researchers to explicitly define the empirical indicators they use to measure diversity, specifying the scale, inclusion/exclusion criteria, and other relevant parameters. The variability in measurements across datasets makes it challenging to understand specific indicators without clear definitions. For example, Ithaca365 (Diaz et al., 2022), Mapillary Traffic Sign (Ertler et al., 2020), JHU-Crowd (Sindagi et al., 2019), and MOTSynth (Fabbri et al., 2021b) all claim to include images representing diverse weather conditions. However, the measurement of weather diversity varies, ranging from employing nine categories (MOTSynth) to using only three (JHU-Crowd).

Clear definitions are paramount to prevent misinterpretations arising from diverse operationalizations, especially given that different operationalizations of the same construct can yield markedly different results. For example, indicators can be categorized as either *objective* or *subjective*. Objective indicators rely on explicit criteria and are assessed by external observers, while subjective indicators involve personal perception or evaluation. The YASO dataset (Orbach et al., 2021) exemplifies this variability, where sentiment, typically considered subjective, is labeled by 7–10 crowdworkers. Although sentiment is inherently subjective, the YASO creators operationalize it as an objective measurement by reporting the majority sentiment among crowdworkers. Operationalizing sentiment as an objective measurement can obfuscate some of the nuances of sentiment that exist within the data. Such an approach fails to capture the full spectrum of sentiment, leading to limited understanding of sentiment diversity.

The approach by Orbach et al. (2021) also introduces potential concerns related to *selection bias*. Under-performing workers are excluded based on "test questions with an a

priori known answer" to ensure annotation quality. This practice may inadvertently filter out diverse perspectives or sentiments misaligned with the authors' subjective views, impacting the overall representativeness and objectivity of the YASO sentiment annotations.

Critically reflect on taxonomies and labels. Dataset curators should apply the same care during operationalization as in conceptualization, giving thoughtful consideration to label taxonomies, their suitability for the task, and potential implications. As an illustrative example, consider the operationalization of "offensive" across several offensive language datasets (Warner & Hirschberg, 2012; Rosenthal et al., 2021; Nobata et al., 2016), where authors employ a keyword-based approach, selecting texts containing slurs or swears (Waseem et al., 2017). Defining "offensiveness" based on keywords presents challenges, as it may overlook *implicitly* offensive text (Wiegand et al., 2021; Waseem et al., 2017) or language where the true offensive nature is obscured (Hada et al., 2021). Further, this operationalization ignores the context of the text, leading to potential mislabeling. For example, depending on the text's author and context, the use of an identity term can be either innocuous or pejorative (Dixon et al., 2018; Davidson et al., 2019).

Evaluate trade-offs for data collection. Curators should thoroughly assess various data collection strategies and justify their chosen method. Trade-offs are inevitable, and while demonstrating robustness to different measurement types is ideal, it can be challenging, especially in resource-intensive dataset collection. At a minimum, transparent communication of decision-making processes facilitates the evaluation of operationalization and collection process validity. For example, Zhao et al. (2021) revealed a demographically imbalanced annotator pool, a limitation when crowdsourcing skin color and gender expression annotations on Amazon Mechanical Turk. Sharing insights into both successful and unsuccessful methods can prevent the duplication of null or negative results by future curators.

6. Reliability

Measurement evaluation involves assessing two critical qualities: reliability and validity. In this section, we specifically delve into the concept of reliability, which concerns the consistency and dependability of measurement results. For diverse dataset collection, ensuring the reliability of the dataset is pivotal, forming the bedrock for the validity of the entire collection process—that is, how effectively the gathered dataset captures the essence of diversity.

To assess reliability in diverse datasets, we illuminate two evaluation methods drawn from the literature on measurement theory: *inter-annotator agreement* and *test-retest reliability*. We conceptualize dataset reliability as analogous

to concerns about the quality and consistency of the collection process. Much like with an unreliable measurement, a dataset that yields inconsistent results when applied to the same case erodes trust in drawn conclusions, making the reliability of data collection essential (Bandalos, 2018).

6.1. Considerations

Lack of details on quality control. In our corpus analysis, a notable theme that emerged is the limited information available on quality control measures for datasets. Only 56.3% of the datasets provided specifics about their quality control processes. This deficiency, as discussed in Section 5, is closely tied to the broader issue of inadequate details in the dataset collection process.

Erasure of labor. While datasets occasionally offer information regarding quality, the focus tends to be on human annotators rather than the data instances. Surprisingly, only 36.8% of datasets include crucial details about annotation quality, such as annotator training processes, attention checks, compensation methods, and work rejection policies. This omission underscores the undervaluation of the labor invested in dataset creation. Crowdworkers are often seen as costs to be minimized, overlooking their significant contributions to datasets (Williams et al., 2022; Shmueli et al., 2021; Miceli et al., 2022a). Similarly, the substantial effort required to ensure high-quality datasets is frequently ignored.

6.2. Recommendations

Inter-annotator agreement. An established method for assessing reliability, particularly in crowdsourcing, is through inter-annotator agreement. This method often entails multiple annotators labeling an instance, with the final label determined by a majority vote (Davani et al., 2022). Another method to gauge inter-annotator reliability is by employing statistical measures of agreement. We find that some text datasets provide quantitative metrics (Sun et al., 2021; Castro et al., 2022; Cao & Daumé III, 2020; Peskov et al., 2019; Johnson & Goldwasser, 2018; Angelidis & Lapata, 2018; Ide et al., 2008; Webster et al., 2018; Zhong et al., 2021), such as Fleiss's κ (Fleiss, 1971) or Cohen's κ (Cohen, 1960), to quantify inter-annotator agreement. While consensus methods are employed in both text and image datasets, quantitative metrics for inter-annotator agreement are reported exclusively in text datasets. We recommend that image dataset curators also incorporate these statistical measures when evaluating crowdsourced labels.

The suitability of inter-annotator agreement is contingent on how diversity is defined and operationalized in the specific dataset. While measuring agreement enhances reliability, in certain cases, it might conflict with the diversity goal. For

example, prior research has demonstrated that "integrating dissenting voices into machine learning models" (Gordon et al., 2022) can be beneficial, conferring "robustness to adversarial attacks" (Peterson et al., 2019). Furthermore, relying on aggregation methods, such as majority voting, may systematically erase certain groups' perspectives from datasets altogether (Davani et al., 2022; Sachdeva et al., 2022). This tension between reliability and diversity is fundamentally linked to the chosen construct of diversity. If diversity involves a range of annotator demographics or perspectives, traditional reliability measures may not be appropriate. However, in tasks aiming for singular, congruous "objective" labels, incorporating inter-annotator agreement metrics can offer valuable insights into reliability.

Test-retest reliability. Another approach that dataset collectors can adopt is the test-retest method. In education, this method involves administering the same test twice over a period, with consistent results indicating reliability (Guttman, 1945). This principle is particularly relevant when assessing the reliability of collection methods like web scraping. For instance, curators can reapply the same methodology to recollect instances, validating whether the recollected dataset maintains the same diversity properties. Nonetheless, as emphasized by Jacobs & Wallach (2021), a lack of reliability from these tests does not necessarily imply that the collection methodology inadequately captures diversity. Changes in the underlying data distribution over time (Chen & Zou, 2024; Yao et al., 2022) can influence the results. For example, when evaluating linguistic diversity using data scraped from Reddit, major societal events, such as elections (Waller & Anderson, 2021), can unexpectedly alter the distribution. Even in such cases, measuring test-retest reliability remains valuable for gaining insights into potential shifts in data distributions.

7. Validity

Conceptually, construct validity ensures that a measure *aligns* with theoretical hypotheses. Adapting this definition, we explore the construct validity of diversity in ML datasets, aiming to determine whether the final dataset aligns with theoretical definitions. To assess validity in diverse datasets, in this section, we apply two commonly used subtypes of construct validity (Campbell & Fiske, 1959) for evaluation: *convergent validity* and *discriminant validity*.

7.1. Considerations

Lack of robust validation. We observe a lack of robust validation for diversity claims made by dataset creators. This issue is partly rooted in the ambiguity surrounding the term. When the construct lacks a clear definition, it becomes challenging to empirically assess whether the collected dataset genuinely adheres to the specified standards. Even when

validation is attempted, it may often center around incorrect constructs. While dataset authors may present metadata frequency or other summary statistics about their datasets, these metrics do not consistently align with the diversity dimensions described during the dataset's motivation

Overreliance on downstream evaluation. Another prevalent evaluation method involves benchmarking the downstream performance of a newly proposed model. This approach is observed in 48.5% of datasets. However, it may assess the wrong construct by primarily focusing on model performance rather than the intrinsic characteristics of the dataset. Model performance, for instance, may increase due to the learning of "shortcuts" (Geirhos et al., 2020) rather than genuine increases in dataset diversity.

7.2. Recommendations

Convergent validity. Convergent validity ensures correlated results for different measurements of the same construct (Campbell & Fiske, 1959). One approach is to compare a newly collected dataset to existing ones. Striking the right balance involves establishing the novelty of a new dataset while demonstrating its similarity to existing work (Quinn et al., 2010; Jacobs & Wallach, 2021). We provide recommendations for evaluating convergent validity.

Cross-dataset generalization. Commonly employed to evaluate "dataset bias" (Torralba & Efros, 2011), cross-dataset generalization enables researchers to compare datasets. By utilizing existing datasets with similar structures (e.g., label taxonomy, modality) and constructs of diversity, collectors can train on their dataset and test on existing datasets or vice versa, comparing relevant metrics such as accuracy. Model performance can also be assessed against standard train-test splits from the same dataset. If the models perform similarly in both cross-dataset and same-dataset scenarios, it suggests that the datasets have similar distributions for the target variable, indicating correlated constructs of diversity. Prior work (Khan & Fu, 2021) adopted Fleiss's κ to measure the consistency of predictions among models. However, a constraint of employing cross-dataset generalization is the necessity for congruent taxonomies (for the target variable) and comparable distributions across datasets.

Comparing existing diversity metrics. Dataset collectors can leverage established metrics for measuring data diversity (Mitchell et al., 2022). For instance, Friedman & Dieng (2023) introduced the Vendi Score, drawing inspiration from ecology and quantum statistical mechanics, as a measure of diversity within image and text dataset categories. Curators can demonstrate how their collection process aligns with such recognized diversity metrics. Given that diversity metrics depend on the embedding space employed (Friedman & Dieng, 2023), datasets should be benchmarked across a mul-

tiplicity of spaces optimized for the definition of diversity selected by the dataset curators.

Discriminant validity. Discriminant validity assesses whether measurements for theoretically unrelated constructs yield unrelated results (Campbell & Fiske, 1959). Consider the initial Visual Question Answer (VQA) dataset (Antol et al., 2015), which aimed to collect diverse and interesting questions and answers, encompassing question types such as "What is ...", "How many ...", and "Do you see a ...". If diversity is defined by the types of questions asked, it should have no relation to other factors, such as gender distribution.[1]

Prior works (Goyal et al., 2017; Agrawal et al., 2018; Hudson & Manning, 2019; Johnson et al., 2017) identified language biases in how questions and answers are formulated in the VQA dataset. For instance, based on the dataset construction, a model predicting "Yes" whenever the question begins with "Do you see a ..." can achieve high accuracy without considering the image in question (Goyal et al., 2017). This suggests potential low discriminant validity for the given measure, highlighting the importance of applying discriminant validity to mitigate construction biases during dataset creation.

8. Case Study

To demonstrate the practical application of measurement theory in ML data collection, we examine the Segment Anything dataset (SA-1B) (Kirillov et al., 2023) as a case study. We choose SA-1B due to its recent release and its alignment with large-scale datasets aimed at advancing foundation model research (Bommasani et al., 2021). Moreover, SA-1B benefits from clear documentation (Gebru et al., 2021) provided by the authors. Through this case study, we investigate the conceptualization, operationalization, and evaluation of diversity, showcasing how our recommendations can enhance the dataset collection process.

SA-1B comprises "11M diverse, high-resolution, licensed, and privacy protecting images and 1.1B high-quality segmentation masks". According to the datasheet, the authors emphasize the dataset's enhanced geographical diversity compared to existing datasets, driven by the goal of fostering "fairer and more equitable models". The images are sourced from a third-party and captured by photographers. Notably, the masks are generated using a novel data engine detailed in the paper, and they are not semantically labeled.

Conceptualization. Although the datasheet mentions how diversity is defined, we can enhance this conceptualization

by providing a more concrete and specific definition. First, the term "geographic diversity" is clarified to encompass a variety in both the country where the image is taken and the socioeconomic status of that country. Second, beyond geographic diversity, the dataset also defines diversity in terms of the variety in object appearance, including factors such as object size and complexity of object shape, as well as the number of objects per image. Finally, these definitions can be contextualized by referencing existing geodiverse datasets (Rojas et al., 2022; Shankar et al., 2017).

Operationalization. Transitioning from conceptualization to operationalization, the different variables are implemented as outlined below. We highlight a strength of this work, which lies in the clear and well-defined indicators.

- *Country of origin*: Inferred from a caption describing the content in the image and a named-entity recognition model (NER) (Peters et al., 2017) to identify location names.
- *Socioeconomic status*: Used the World Bank's (WDI, 2022) income level categorization for the country.
- *Object size*: Calculated image-relative mask size (i.e., square root of mask area divided by image area).
- *Object complexity*: Calculated mask concavity (i.e., 1 - mask area divided by area of masks' convex hull).
- *Number of objects*: Calculated as the count of masks.

A critical examination of these operationalizations brings forth two noteworthy considerations. First, as acknowledged by the authors, relying on an NER model for inferring the country of origin introduces the potential for errors due to, e.g., social bias or ambiguity (e.g., "Georgia" can refer to both a US state and a country). Second, opting to operationalize geolocation at the country level overlooks intra-national differences, potentially leading to the presentation of stereotypical representations of individuals within that country (Naggita et al., 2023).

While the indicators are well-defined, there is room for improvement in the transparency surrounding the dataset collection process. The dataset, collected through a third party, reflects increasing opacity in collection processes (Section 5.1). Several important details about the collection process are omitted. For instance, the instructions given to photographers to enhance variation in object appearance and location are undisclosed. Further, while SA-1B does achieve diversity in objects, it is unclear whether this is a result of explicit instructions and deliberate sampling or simply a byproduct of scale.

Evaluation. The authors substantiate their diversity claims by comparing object masks with those of similar datasets, such as COCO (Lin et al., 2014) and OpenImages (Kuznetsova et al., 2020). For example, they demonstrate alignment with respect to the distribution of object complexity. SA-1B distinguishes itself, however, by exhibiting more variety in mask sizes and the number of masks per

---

[1]Note that comparing *gender distribution* may not be reasonable for testing discriminant validity when the interest is in *demographic diversity*, as these are related constructs.

image, showcasing convergent validity.

A recommended enhancement involves providing validation for geographic diversity. Although the authors present figures illustrating the distribution of country of origin, emphasizing geodiversity is crucial as object appearances vary globally (De-Arteaga et al., 2019; Shankar et al., 2017; Ramaswamy et al., 2024). To bolster the validation of this diversity aspect, the authors could compare distributions inferred from object labels or visualize object mask representations across different geographic regions.

9. Discussion

Finally, we dive into additional considerations regarding three key points: the inherent tensions between measurement and scale, the documentation burden imposed on dataset creators, and changing constructs over time.

Measurement and scale. One potential tension within our proposed framework revolves around the interplay between diversity and scale. Currently, there is a prevalent belief among dataset curators that diversity will organically emerge as a consequence of dataset scale. For instance, Sindagi et al. (2019) argue that having "such a large number of images", in their dataset, JHU-Crowd, "results in increased diversity in terms of count, background regions, scenarios, etc." Contrary to this notion, we posit that diversity, along with other constructs dataset curators aim to capture, is not an automatic byproduct of scale. Instead, it requires careful conceptualization, operationalization, and subsequent evaluation. This may impede scalability by introducing an additional need for curation and explicit control on the curator's side. Nevertheless, as previous studies have advocated (Diaz & Madaio, 2023; Hanna & Park, 2020), challenging the concept of "scale thinking" can be advantageous, not only ethically but also for downstream performance. For example, Byrne et al. (2021) emphasize the importance of "carefully curated, annotated datasets that cover all the idiosyncrasies of a single task or transaction" as a key factor in enhancing model performance, countering the prevailing notion solely focused on scale.

Documentation burden. In Section 5.1, we shed light on the *documentation gap* in ML research. To tackle this challenge, we advocate for enhanced clarity and explicit communication throughout the dataset collection pipeline. However, we acknowledge the significant and often underestimated burden that documentation places on dataset creators—a task that is frequently undervalued within research communities and organizations (Heger et al., 2022). This burden is further exacerbated by the necessity to address existing "documentation debt" (Bandy & Vincent, 2024), referring to commonly used datasets with inadequate or absent documentation. Overcoming these challenges

demands systemic changes in how data work is perceived within the ML community (Sambasivan et al., 2021). We highlight ongoing efforts in this area, including the establishment of academic venues such as the Journal of Data-centric Machine Learning Research and NeurIPS Datasets and Benchmarks Track. These initiatives mark a crucial initial step toward tackling this pervasive problem.

Constructs changing over time. Our framework does not account for potential shifts over time, such as discrepancies between training and testing data caused by temporal changes (Yao et al., 2022; Bergman et al., 2023). In Section 6.2, we acknowledge this when discussing how the reliability of webscraped data can be impacted by changes to Internet trends or current events. Moreover, the very nature of the construct being measured may undergo transformations. For instance, the racial categories employed in the US Census have undergone significant changes over the years (Lai & Medina, 2023). Since these categories often serve as a taxonomy for race in ML datasets, it is evident how such changes can influence and date the underlying construct. Recognizing that construct definitions should be contextualized, datasets inevitably become tied to the temporal setting of their collection. Instead of striving for time-invariant datasets, we advocate for directing our efforts towards the development of algorithms capable of withstanding such distribution shifts (Koh et al., 2021; Yao et al., 2022).

10. Conclusion

We explored the application of measurement theory principles as a framework for enhancing ML datasets. Present data collection practices often treat value-laden constructs within datasets, like diversity, as implicit or self-explanatory. This approach gives rise to subsequent challenges in validating and replicating the assertions made by authors in their work. Significantly, the lack of a standardized framework in dataset creation exacerbates issues related to the reproducibility crisis in science. Our analysis discerns that the absence of clear definitions and quantification by dataset authors amounts to *selective reporting*, hindering standardization. By leveraging measurement theory principles, we advocate for a more nuanced and precise approach to conceptualizing, operationalizing, and evaluating value-laden properties in dataset construction. This approach not only fosters transparency, reliability, and reproducibility in ML research but also contributes to addressing the pervasive issue of the reproducibility crisis in science.

Impact Statement

Our paper introduces strategies for enhancing ML dataset collection by leveraging measurement theory principles. By fostering a more deliberate and thoughtful approach among researchers, our work stands to positively influence the societal impact of ML dataset curation. Moreover, our recommendations extend beyond mere dataset collection, offering insights into improving transparency and reproducibility within ML research more broadly.

Acknowledgements

References

The world by income and region, 2022. URL https://datatopics.worldbank.org/world-development-indicators/the-world-by-income-and-region.html.

Abdul-Mageed, M., Elmadany, A., Nagoudi, E. M. B., Pabbi, D., Verma, K., and Lin, R. Mega-cov: A billion-scale dataset of 100+ languages for covid-19. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Agrawal, A., Batra, D., Parikh, D., and Kembhavi, A. Don't just assume; look and answer: Overcoming priors for visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Aly, R., Guo, Z., Schlichtkrull, M. S., Thorne, J., Vlachos, A., Christodoulopoulos, C., Cocarascu, O., and Mittal, A. Feverous: Fact extraction and verification over unstructured and structured information. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

An, J., Kim, J., Lee, H., Kim, J., Kang, J., Shin, S., Kim, M., Hong, D., and Woo, S. S. Vfp290k: A large-scale benchmark dataset for vision-based fallen person detection. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

Andrews, J. T., Zhao, D., Thong, W., Modas, A., Papakyriakopoulos, O., Nagpal, S., and Xiang, A. Ethical considerations for collecting human-centric image datasets. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2023.

Andrus, M., Spitzer, E., Brown, J., and Xiang, A. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 249–260, 2021.

Angelidis, S. and Lapata, M. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015.

Asaadi, S., Mohammad, S., and Kiritchenko, S. Big bird: A large, fine-grained, bigram relatedness dataset for examining semantic composition. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (NAACL)*, 2019.

Asano, Y. M., Rupprecht, C., Zisserman, A., and Vedaldi, A. Pass: An imagenet replacement for self-supervised pretraining without humans. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

Babbie, E. R. *The practice of social research*. Cengage AU, 2020.

Bagher Zadeh, A., Cao, Y., Hessner, S., Liang, P. P., Poria, S., and Morency, L.-P. Cmu-moseas: A multimodal language dataset for spanish, portuguese, german and french. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Bai, Y., Jiao, J., Ce, W., Liu, J., Lou, Y., Feng, X., and Duan, L.-Y. Person30k: A dual-meta generalization network for person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Bandalos, D. L. *Measurement theory and applications for the social sciences*. Guilford Publications, 2018.

Bandy, J. and Vincent, N. Addressing" documentation debt" in machine learning research: A retrospective datasheet for bookcorpus. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2024.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. Objectnet: A

large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Bender, E. M. and Friedman, B. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics (TACL)*, 6:587–604, 2018.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big?. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

Bergman, A. S., Hendricks, L. A., Rauh, M., Wu, B., Agnew, W., Kunesch, M., Duan, I., Gabriel, I., and Isaac, W. Representation in ai evaluations. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.

Bhattacherjee, A. *Social science research: Principles, methods, and practices*. USA, 2012.

Birhane, A., Prabhu, V. U., and Kahembwe, E. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., and Bao, M. The values encoded in machine learning research. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

Birhane, A., Prabhu, V., Han, S., and Boddeti, V. N. On hate scaling laws for data-swamps. *arXiv preprint arXiv:2306.13141*, 2023.

Blili-Hamelin, B. and Hancox-Li, L. Making intelligence: Ethical values in iq and ml benchmarks. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. Language (technology) is power: A critical survey of "bias" in NLP. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., and Wallach, H. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (ACL-IJCNLP)*, 2021.

Blodgett, S. L., Liao, Q. V., Olteanu, A., Mihalcea, R., Muller, M., Scheuerman, M. K., Tan, C., and Yang, Q. Responsible language technologies: Foreseeing and mitigating harms. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2022.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., and Liang, P. The foundation model transparency index. *arXiv preprint arXiv:2310.12941*, 2023.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2018.

Byrne, B., Krishnamoorthi, K., Sankar, C., Neelakantan, A., Goodrich, B., Duckworth, D., Yavuz, S., Dubey, A., Kim, K.-Y., and Cedilnik, A. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Byrne, B., Krishnamoorthi, K., Ganesh, S., and Kale, M. Tickettalk: Toward human-level performance with end-to-end, transaction-based dialog systems. In *Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (ACL-IJCNLP)*, 2021.

Campbell, D. T. and Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, 56(2):81, 1959.

Cao, Y., He, Z., Wang, L., Wang, W., Yuan, Y., Zhang, D., Zhang, J., Zhu, P., Van Gool, L., Han, J., Hoi, S., Hu, Q., Liu, M., Cheng, C., Liu, F., Cao, G., Li, G., Wang, H., He, J., Wan, J., Wan, Q., Zhao, Q., Lyu, S., Zhao, W., Lu, X., Zhu, X., Liu, Y., Lv, Y., Ma, Y., Yang, Y., Wang, Z., Xu, Z., Luo, Z., Zhang, Z., Zhang, Z., Li, Z., and Zhang, Z. Visdrone-det2021: The vision meets drone object detection challenge results. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021.

Cao, Y. T. and Daumé III, H. Toward gender-inclusive coreference resolution. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Castro, S., Wang, R., Huang, P., Stewart, I., Ignat, O., Liu, N., Stroud, J., and Mihalcea, R. Fiber: Fill-in-the-blanks as a challenging video understanding evaluation framework. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*, 2022.

Celis, L. E., Deshpande, A., Kathuria, T., and Vishnoi, N. K. How to be fair and diverse? *arXiv preprint arXiv:1610.07183*, 2016.

Chawla, K., Ramirez, J., Clever, R., Lucas, G., May, J., and Gratch, J. Casino: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2021.

Check, J. and Schutt, R. K. *Research methods in education*. Sage publications, 2011.

Chen, Y. and Zou, J. Twigma: A dataset of ai-generated images with metadata from twitter. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2024.

Chuang, C.-Y., Li, J., Torralba, A., and Fidler, S. Learning to act properly: Predicting and explaining affordances from images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Culkin, R., Hu, J. E., Stengel-Eskin, E., Qin, G., and Durme, B. V. Iterative paraphrastic augmentation with discriminative span alignment. *Transactions of the Association for Computational Linguistics (TACL)*, 9:494–509, 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00380.

Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

Davani, A. M., Díaz, M., and Prabhakaran, V. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics (TACL)*, 10:92–110, 2022.

Dave, A., Khurana, T., Tokmakov, P., Schmid, C., and Ramanan, D. Tao: A large-scale benchmark for tracking any object. In *European Conference on Computer Vision (ECCV)*, 2020.

Davidson, T., Bhattacharya, D., and Weber, I. Racial bias in hate speech and abusive language detection datasets. In *Workshop on Abusive Language Online*, 2019.

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., and Kalai, A. T. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2019.

De Kock, C. and Vlachos, A. Leveraging wikipedia article evolution for promotional tone detection. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*, 2022.

Denton, E., Hanna, A., Amironesei, R., Smart, A., Nicole, H., and Scheuerman, M. K. Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv preprint arXiv:2007.07399*, 2020.

Derczynski, L., Bontcheva, K., and Roberts, I. Broad twitter corpus: A diverse named entity recognition resource. In *International Conference on Computational Linguistics: Technical Papers (COLING)*, 2016.

Diaz, C. A., Xia, Y., You, Y., Nino, J., Chen, J., Monica, J., Chen, X., Luo, K. Z., Wang, Y., Emond, M., et al. Ithaca365: Dataset and driving perception under repeated and challenging weather conditions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Diaz, F. and Madaio, M. Scaling laws do not scale. *arXiv preprint arXiv:2307.03201*, 2023.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. Measuring and mitigating unintended bias in text classification. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2018.

Drawzeski, K., Galassi, A., Jablonowska, A., Lagioia, F., Lippi, M., Micklitz, H. W., Sartor, G., Tagiuri, G., and Torroni, P. A corpus for multilingual analysis of online terms of service. In *Natural Legal Language Processing Workshop 2021*, 2021.

Dumitrescu, S. D., Rebeja, P., Lorincz, B., Gaman, M., Avram, A., Ilie, M., Pruteanu, A., Stan, A., Rosia, L., Iacobescu, C., et al. Liro: Benchmark and leaderboard for romanian language tasks. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

Durmus, E., Ladhak, F., and Cardie, C. The role of pragmatic and discourse context in determining argument impact. In *Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Ertler, C., Mislej, J., Ollmann, T., Porzi, L., Neuhold, G., and Kuang, Y. The mapillary traffic sign dataset for detection and classification on a global scale. In *European Conference on Computer Vision (ECCV)*, 2020.

Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., Chai, Y., Sapp, B., Qi, C. R., Zhou, Y., et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Fabbri, A., Rahman, F., Rizvi, I., Wang, B., Li, H., Mehdad, Y., and Radev, D. Convosumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. In *Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (ACL-IJCNLP)*, 2021a.

Fabbri, M., Brasó, G., Maugeri, G., Cetintas, O., Gasparini, R., Ošep, A., Calderara, S., Leal-Taixé, L., and Cucchiara, R. Motsynth: How can synthetic data help pedestrian detection and tracking? In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021b.

Fan, D.-P., Wang, W., Cheng, M.-M., and Shen, J. Shifting more attention to video salient object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019a.

Fan, D.-P., Lin, Z., Ji, G.-P., Zhang, D., Fu, H., and Cheng, M.-M. Taking a deeper look at co-salient object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Fan, L., Chen, Y., Wei, P., Wang, W., and Zhu, S.-C. Inferring shared attention in social scene videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Fan, L., Wang, W., Zhu, S.-C., Tang, X., and Huang, S. Understanding human gaze communication by spatio-temporal graph reasoning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019b.

Farha, I. A. and Magdy, W. From arabic sentiment analysis to sarcasm detection: The arsarcasm dataset. In *Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, 2020.

Fitzpatrick, T. Sun and skin. *Journal de Medecine Esthetique*, 2:33–34, 1975.

Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

Friedman, D. and Dieng, A. B. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research (TMLR)*, 2023.

Fu, D., Chen, D., Bao, J., Yang, H., Yuan, L., Zhang, L., Li, H., and Chen, D. Unsupervised pre-training for person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., and Crawford, K. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Gillani, N. and Levy, R. Simple dynamic word embeddings for mapping perceptions in the public sphere. In *Workshop on Natural Language Processing and Computational Social Science*, 2019.

Goldman, O. and Tsarfaty, R. Well-defined morphology is sentence-level morphology. In *Workshop on Multilingual Representation Learning*, 2021.

Google PAIR. People + AI Guidebook. https://pair.withgoogle.com/guidebook/, 2019.

Gordon, M. L., Lam, M. S., Park, J. S., Patel, K., Hancock, J., Hashimoto, T., and Bernstein, M. S. Jury learning: Integrating dissenting voices into machine learning models. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2022.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Guan, J., Feng, Z., Chen, Y., He, R., Mao, X., Fan, C., and Huang, M. Lot: A story-centric benchmark for evaluating chinese long text understanding and generation. *Transactions of the Association for Computational Linguistics (TACL)*, 10:434–451, 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00469.

Guttman, L. A basis for analyzing test-retest reliability. *Psychometrika*, 10(4):255–282, 1945.

Hada, R., Sudhir, S., Mishra, P., Yannakoudakis, H., Mohammad, S. M., and Shutova, E. Ruddit: Norms of offensiveness for english reddit comments. In *Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (ACL-IJCNLP)*, 2021.

Hamidi, F., Scheuerman, M. K., and Branham, S. M. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2018.

Hanna, A. and Park, T. M. Against scale: Provocations and resistances to scale thinking. *arXiv preprint arXiv:2010.08850*, 2020.

Hanna, A., Denton, E., Smart, A., and Smith-Loud, J. Towards a critical race methodology in algorithmic fairness. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.

Hazirbas, C., Bitton, J., Dolhansky, B., Pan, J., Gordo, A., and Ferrer, C. C. Towards measuring fairness in ai: the casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):324–332, 2021.

Heger, A. K., Marquis, L. B., Vorvoreanu, M., Wallach, H., and Wortman Vaughan, J. Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–29, 2022.

Heinzerling, B. NLP's clever hans moment has arrived, Jan 2022. URL https://thegradient.pub/nlps-clever-hans-moment-has-arrived/#fn1.

Hemani, M., Patel, A., Shimpi, T., Ramesh, A., and Krishnamurthy, B. What ails one-shot image segmentation: A data perspective. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K. The dataset nutrition label. *Data Protection and Privacy*, 12(12):1, 2020.

Howard, J. J., Sirotin, Y. B., Tipton, J. L., and Vemury, A. R. Reliability and validity of image-based and self-reported skin phenotype metrics. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(4):550–560, 2021.

Huang, Q., Hao, S., Ye, Y., Zhu, S., Feng, Y., and Zhao, D. Does recommend-revise produce reliable annotations? an analysis on missing instances in docred. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*, 2022.

Hudson, D. A. and Manning, C. D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Ide, N., Baker, C., Fellbaum, C., Fillmore, C., and Passonneau, R. Masc: The manually annotated sub-corpus of american english. In *International Conference on Language Resources and Evaluation (LREC)*, 2008.

Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., and Shah, M. Composition loss for counting, density map estimation and localization in dense crowds. In *European Conference on Computer Vision (ECCV)*, 2018.

Inoue, N., Stenetorp, P., and Inui, K. R4c: A benchmark for evaluating rc systems to get the right answer for the right reason. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Jacobs, A. Z. Measurement as governance in and for responsible ai. *arXiv preprint arXiv:2109.05658*, 2021.

Jacobs, A. Z. and Wallach, H. Measurement and fairness. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

Jacobs, A. Z., Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. The meaning and measurement of bias: lessons from natural language processing. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., and Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Johnson, K. and Goldwasser, D. Classification of moral foundations in microblog political discourse. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*, 2018.

Karazija, L., Laina, I., and Rupprecht, C. Clevrtex: A texture-rich benchmark for unsupervised multi-object segmentation. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

Karkkainen, K. and Joo, J. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

Khalid, H., Tariq, S., Kim, M., and Woo, S. S. Fakeavceleb: A novel audio-video multimodal deepfake dataset. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

Khan, Z. and Fu, Y. One label, one billion faces: Usage and consistency of racial categories in computer vision.

In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

Kim, B., Kim, H., and Kim, G. Abstractive summarization of reddit posts with multi-level memory networks. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (NAACL)*, 2019.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

Ku, A., Anderson, P., Patel, R., Ie, E., and Baldridge, J. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision (IJCV)*, 128(7):1956–1981, 2020.

Lai, K. R. and Medina, J. An american puzzle: Fitting race in a box. *The New York Times*, Oct 2023.

Lebret, R., Grangier, D., and Auli, M. Neural text generation from structured data with application to the biography domain. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

Lei, J., Yu, L., Berg, T., and Bansal, M. What is more likely to happen next? video-and-language future event prediction. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020a.

Lei, J., Yu, L., Berg, T. L., and Bansal, M. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision (ECCV)*, 2020b.

Lei, J., Berg, T., and Bansal, M. mtvr: Multilingual moment retrieval in videos. In *Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (ACL-IJCNLP)*, 2021a.

Lei, J., Berg, T. L., and Bansal, M. Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021b.

Lepp, H. and Levow, G.-A. Pardon the interruption: An analysis of gender and turn-taking in u.s. supreme court oral arguments. In *Interspeech 2020*, 2020.

Li, X. "There's no data like more data" automatic speech recognition and the making of algorithmic culture. *Osiris*, 38(1):165–182, 2023.

Lin, B. Y., Lee, S., Qiao, X., and Ren, X. Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning. In *Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (ACL-IJCNLP)*, 2021.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., and Yi, L. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Lu, P., Qiu, L., Chen, J., Xia, T., Zhao, Y., Zhang, W., Yu, Z., Liang, X., and Zhu, S.-C. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

Malevé, N. On the data set's ruins. *AI & Society*, 36:1117–1131, 2021.

Marino, K., Rastegari, M., Farhadi, A., and Mottaghi, R. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Martin, M., Roitberg, A., Haurilet, M., Horne, M., ReiB, S., Voit, M., and Stiefelhagen, R. Drive&act: A multimodal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Martyniuk, T., Kupyn, O., Kurlyak, Y., Krashenyi, I., Matas, J., and Sharmanska, V. Dad-3dheads: A large-scale dense, accurate and diverse dataset for 3d head alignment from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Mayo, D., Cummings, J., Lin, X., Gutfreund, D., Katz, B., and Barbu, A. How hard are computer vision datasets? calibrating dataset difficulty to viewing time. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Miao, J., Wei, Y., Wu, Y., Liang, C., Li, G., and Yang, Y. Vspw: A large-scale dataset for video scene parsing in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Miao, J., Wang, X., Wu, Y., Li, W., Zhang, X., Wei, Y., and Yang, Y. Large-scale video panoptic segmentation in the wild: A benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a.

Miao, J., Wang, X., Wu, Y., Li, W., Zhang, X., Wei, Y., and Yang, Y. Large-scale video panoptic segmentation in the wild: A benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b.

Miceli, M., Posada, J., and Yang, T. Studying up machine learning data: Why talk about bias when we mean power? *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):1–14, 2022a.

Miceli, M., Yang, T., Alvarado Garcia, A., Posada, J., Wang, S. M., Pohl, M., and Hanna, A. Documenting data production processes: A participatory approach for data work. *Proceedings of the ACM Human Computer Interaction*, 6 (CSCW2), 2022b.

Mitchell, M., Luccioni, A. S., Lambert, N., Gerchick, M., McMillan-Major, A., Ozoani, E., Rajani, N., Thrush, T., Jernite, Y., and Kiela, D. Measuring data. *arXiv preprint arXiv:2212.05129*, 2022.

Mohamed, Y., Khan, F. F., Haydarov, K., and Elhoseiny, M. It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Mundhenk, T. N., Konjevod, G., Sakla, W. A., and Boakye, K. A large contextual dataset for classification, detection and counting of cars with deep learning. In *European Conference on Computer Vision (ECCV)*, 2016.

Müller, M., Bibi, A., Giancola, S., Alsubaihi, S., and Ghanem, B. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *European Conference on Computer Vision (ECCV)*, 2018.

Naggita, K., LaChance, J., and Xiang, A. Flickr africa: Examining geo-diversity in large-scale, human-centric visual data. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2023.

Ni, J., Li, J., and McAuley, J. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. Abusive language detection in online user content. In *International Conference on World Wide Web (TheWebConf)*, 2016.

Orbach, M., Toledo-Ronen, O., Spector, A., Aharonov, R., Katz, Y., and Slonim, N. Yaso: A targeted sentiment analysis evaluation dataset for open-domain reviews. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

O'Toole, A. J., Peterson, J., and Deffenbacher, K. A. An 'other-race effect' for categorizing faces by sex. *Perception*, 25(6):669–676, 1996.

O'Neill, J., Rozenshtein, P., Kiryo, R., Kubota, M., and Bollegala, D. I wish i would have loved this one, but i didn't – a multilingual dataset for counterfactual detection in product review. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., and Hanna, A. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

Peng, K. L., Mathur, A., and Narayanan, A. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

Peskov, D., Clarke, N., Krone, J., Fodor, B., Zhang, Y., Youssef, A., and Diab, M. Multi-domain goal-oriented dialogues (multidogo): Strategies toward curating and annotating large scale dialogue data. In *Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Peters, M. E., Ammar, W., Bhagavatula, C., and Power, R. Semi-supervised sequence tagging with bidirectional language models. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*, 2017.

Peterson, J. C., Battleday, R. M., Griffiths, T. L., and Russakovsky, O. Human uncertainty makes classification more robust. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015.

Pont-Tuset, J., Uijlings, J., Changpinyo, S., Soricut, R., and Ferrari, V. Connecting vision and language with localized narratives. In *European Conference on Computer Vision (ECCV)*, 2020.

Pratapa, A., Gupta, R., and Mitamura, T. Multilingual event linking to wikidata. In *Workshop on Multilingual Information Access (MIA)*, 2022.

Pushkarna, M., Zaldivar, A., and Kjartansson, O. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., and Radev, D. R. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228, 2010.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Rahman, M. M., Balakrishnan, D., Murthy, D., Kutlu, M., and Lease, M. An information retrieval approach to building datasets for hate speech detection. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

Raji, I. D. and Fried, G. About face: A survey of facial recognition evaluation. *arXiv preprint arXiv:2102.00813*, 2021.

Raji, I. D., Denton, E., Bender, E. M., Hanna, A., and Paullada, A. Ai and the everything in the whole wide world benchmark. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

Ramaswamy, V. V., Lin, S. Y., Zhao, D., Adcock, A. B., van der Maaten, L., Ghadiyaram, D., and Russakovsky, O. Geode: a geographically diverse evaluation dataset for object recognition. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2024.

Rojas, W. A. G., Diamos, S., Kini, K. R., Kanter, D., Reddi, V. J., and Coleman, C. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2022.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. M. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Rosenthal, S., Atanasova, P., Karadzhov, G., Zampieri, M., and Nakov, P. Solid: A large-scale semi-supervised dataset for offensive language identification. In *Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (ACL-IJCNLP)*, 2021.

Rudinger, R., Naradowsky, J., Leonard, B., and Van Durme, B. Gender bias in coreference resolution. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (NAACL)*, 2018.

Sachdeva, P. S., Barreto, R., von Vacano, C., and Kennedy, C. J. Assessing annotator identity sensitivity via item response theory: A case study in a hate speech corpus. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

Sadhu, A., Gupta, T., Yatskar, M., Nevatia, R., and Kembhavi, A. Visual semantic role labeling for video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L. M. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2021.

Scheuerman, M. K., Paul, J. M., and Brubaker, J. R. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–33, 2019.

Scheuerman, M. K., Hanna, A., and Denton, E. Do datasets have politics? disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–37, 2021.

Schops, T., Schonberger, J. L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., and Geiger, A. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Schulz, C., Meyer, C. M., Kiesewetter, J., Sailer, M., Bauer, E., Fischer, M. R., Fischer, F., and Gurevych, I. Analysis of automatic annotation suggestions for hard discourse-level tasks in expert domains. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., and Sculley, D. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

Sharma, R., Allen, J., Bakhshandeh, O., and Mostafazadeh, N. Tackling the story ending biases in the story cloze test. In *Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018.

Shen, Z., Lo, K., Wang, L. L., Kuehl, B., Weld, D. S., and Downey, D. Vila: Improving structured content extraction from scientific pdfs using visual layout groups. *Transactions of the Association for Computational Linguistics*, 10:376–392, 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00466.

Shmueli, B., Fell, J., Ray, S., and Ku, L.-W. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2021.

Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., and Gupta, A. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision (ECCV)*, 2016.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision (ECCV)*, 2012.

Sindagi, V. A., Yasarla, R., and Patel, V. M. Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Soldan, M., Pardo, A., Alcazar, J. L., Heilbron, F. C., Zhao, C., Giancola, S., and Ghanem, B. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Song, S., Lichtenberg, S. P., and Xiao, J. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Steed, R. and Caliskan, A. Image representations learned with unsupervised pre-training contain human-like biases. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

Subramonian, A., Yuan, X., Daumé III, H., and Blodgett, S. L. It takes two to tango: Navigating conceptualizations of nlp tasks and measurements of performance. *arXiv preprint arXiv:2305.09022*, 2023.

Sugawara, S., Nangia, N., Warstadt, A., and Bowman, S. What makes reading comprehension questions difficult? In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*, 2022.

Sun, K., Moon, S., Crook, P., Roller, S., Silvert, B., Liu, B., Wang, Z., Liu, H., Cho, E., and Cardie, C. Adding chit-chat to enhance task-oriented dialogues. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2021.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., and Anguelov, D. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Thames, Q., Karpur, A., Norris, W., Xia, F., Panait, L., Weyand, T., and Sim, J. Nutrition5k: Towards automatic nutritional understanding of generic food. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Torralba, A. and Efros, A. A. Unbiased look at dataset bias. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Tseng, G., Zvonkov, I., Nakalembe, C. L., and Kerner, H. Cropharvest: A global dataset for crop-type classification. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

Valverde, F. R., Valeria Hurtado, J., and Valada, A. There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., and MacAodha, O. Benchmarking representation learning for natural world image collections. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Waller, I. and Anderson, A. Quantifying social organization and political polarization in online platforms. *Nature*, 600 (7888):264–268, 2021.

Wan, R., Shi, B., Duan, L.-Y., Tan, A.-H., and Kot, A. C. Crrn: Multi-scale guided concurrent reflection removal network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Wang, A., Liu, A., Zhang, R., Kleiman, A., Kim, L., Zhao, D., Shirai, I., Narayanan, A., and Russakovsky, O. Revise: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision (IJCV)*, 130(7): 1790–1810, 2022.

Wang, M., Deng, W., Hu, J., Tao, X., and Huang, Y. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019a.

Wang, Q., Gao, J., Lin, W., and Yuan, Y. Learning from synthetic data for crowd counting in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019b.

Wang, R., Zhang, Q., Fu, C.-W., Shen, X., Zheng, W.-S., and Jia, J. Underexposed photo enhancement using deep illumination estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019c.

Warner, W. and Hirschberg, J. Detecting hate speech on the world wide web. In *Workshop on Language in Social Media*, 2012.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. Understanding abuse: A typology of abusive language detection subtasks. 2017.

Webster, K., Recasens, M., Axelrod, V., and Baldridge, J. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics (TACL)*, 6:605–617, 2018. ISSN 2307-387X. doi: 10.1162/tacl_a_00240.

Wei, Z., Zhang, J., Lin, Z., Lee, J.-Y., Balasubramanian, N., Hoai, M., and Samaras, D. Learning visual emotion representations from web data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Wiegand, M., Geulig, M., and Ruppenhofer, J. Implicitly abusive comparisons – a new dataset and linguistic analysis. In *European Chapter of the Association for Computational Linguistics: Main Volume (EACL)*, 2021.

Williams, A., Miceli, M., and Gebru, T. The exploited labor behind artificial intelligence. *Noema Magazine*, 13, 2022.

Winner, L. Do artifacts have politics? In *Computer ethics*, pp. 177–192. Routledge, 2017.

Wood, E., Baltrušaitis, T., Hewitt, C., Dziadzio, S., Cashman, T. J., and Shotton, J. Fake it till you make it: face analysis in the wild using synthetic data alone. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Wu, W., Guo, Z., Zhou, X., Wu, H., Zhang, X., Lian, R., and Wang, H. Proactive human-machine conversation with explicit conversation goal. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Xiao, Z., Zhang, S., Lai, V., and Liao, Q. V. Evaluating evaluation metrics: A framework for analyzing nlg evaluation metrics using measurement theory. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Xie, E., Wang, W., Wang, W., Ding, M., Shen, C., and Luo, P. Segmenting transparent objects in the wild. In *European Conference on Computer Vision (ECCV)*, 2020.

Xu, J., Rao, Y., Yu, X., Chen, G., Zhou, J., and Lu, J. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a.

Xu, S., Böttcher, L., and Chou, T. Diversity in biology: definitions, quantification and models. *Physical Biology*, 17(3):031001, 2020.

Xu, X., Qi, Z., Ma, J., Zhang, H., Shan, Y., and Qie, X. Bts: A bi-lingual benchmark for text segmentation in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b.

Xu, Y., Wang, D., Yu, M., Ritchie, D., Yao, B., Wu, T., Zhang, Z., Li, T., Bradford, N., Sun, B., Hoang, T., Sang, Y., Hou, Y., Ma, X., Yang, D., Peng, N., Yu, Z., and Warschauer, M. Fantastic questions and where to find them: Fairytaleqa – an authentic dataset for narrative

comprehension. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*, 2022c.

Yadav, K., Ramrakhya, R., Ramakrishnan, S. K., Gervet, T., Turner, J., Gokaslan, A., Maestre, N., Chang, A. X., Batra, D., Savva, M., Clegg, A. W., and Chaplot, D. S. Habitat-matterport 3d semantics dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Yang, A., Miech, A., Sivic, J., Laptev, I., and Schmid, C. Just ask: Learning to answer questions from millions of narrated videos. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Yang, Y., Gupta, A., Feng, J., Singhal, P., Yadav, V., Wu, Y., Natarajan, P., Hedau, V., and Joo, J. Enhancing fairness in face detection in computer vision systems by demographic bias mitigation. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2022.

Yao, H., Choi, C., Cao, B., Lee, Y., Koh, P. W. W., and Finn, C. Wild-time: A benchmark of in-the-wild distribution shift over time. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

You, Y., Lou, Y., Li, C., Cheng, Z., Li, L., Ma, L., Lu, C., and Wang, W. Keypointnet: A large-scale 3d keypoint dataset aggregated from numerous human annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., and Tao, D. Ap-10k: A benchmark for animal pose estimation in the wild. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

Yuan, A., Ippolito, D., Nikolaev, V., Callison-Burch, C., Coenen, A., and Gehrmann, S. Synthbio: A case study in faster curation of text datasets. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

Yuan, X., Wang, T., Meng, R., Thaker, K., Brusilovsky, P., He, D., and Trischler, A. One size does not fit all: Generating and evaluating variable number of keyphrases. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Zeinert, P., Inie, N., and Derczynski, L. Annotating online misogyny. In *Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (ACL-IJCNLP)*, 2021a.

Zeinert, P., Inie, N., and Derczynski, L. Annotating online misogyny. In *Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (ACL-IJCNLP)*, 2021b.

Zellers, R., Bisk, Y., Farhadi, A., and Choi, Y. From recognition to cognition: Visual commonsense reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Zendel, O., Schorghuber, M., Rainer, B., Murschitz, M., and Beleznai, C. Unifying panoptic segmentation for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Zhang, K., Luo, W., Zhong, Y., Ma, L., Stenger, B., Liu, W., and Li, H. Deblurring by realistic blurring. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Zhang, P., Zhao, J., Wang, D., Lu, H., and Ruan, X. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Zhang, X., Ma, X., Shi, P., and Lin, J. Mr. tydi: A multilingual benchmark for dense retrieval. In *Workshop on Multilingual Representation Learning*, 2021.

Zhao, D., Wang, A., and Russakovsky, O. Understanding and evaluating racial biases in image captioning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (NAACL)*, 2018.

Zhong, Y., Yang, J., Xu, W., and Yang, D. Wikibias: Detecting multi-span subjective biases in language. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. Scene parsing through ade20k dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Zhou, K., Blodgett, S. L., Trischler, A., Daumé III, H., Suleman, K., and Olteanu, A. Deconstructing nlg evaluation: Evaluation practices, assumptions, and their implications. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2022.

Zhu, Z., Xie, L., and Yuille, A. Object recognition with and without objects. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M. J., and Brox, T. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Zou, C., Yu, Q., Du, R., Mo, H., Song, Y.-Z., Xiang, T., Gao, C., Chen, B., and Zhang, H. Sketchyscene: Richly-annotated scene sketches. In *European Conference on Computer Vision (ECCV)*, 2018.

A. Methodology

We provide additional details on our methodology, including how we selected datasets to review, our coding procedure, and categorization details.

A.1. Literature Review

We surveyed a total of 135 image and text datasets using a defined process outlined in Figure 2. Initially, we collected papers from key conferences and journals in computer vision (CVPR, ECCV, ICCV), natural language processing (*CL released on ACL Anthology), machine learning (NeurIPS, NeurIPS Datasets and Benchmarks Track), and fairness (FAccT, AIES) venues. Using regular expressions, we identified papers containing keywords related to "diversity" or "bias" in their titles or abstracts. Subsequently, we filtered these papers based on abstract content, focusing on those introducing new datasets or annotations. Manual screening ensured that the discussions on "diversity" or "bias" were pertinent to dataset creation, the articles were in English, and primarily research-focused. Finally, we randomly selected 135 datasets from this pool for review.
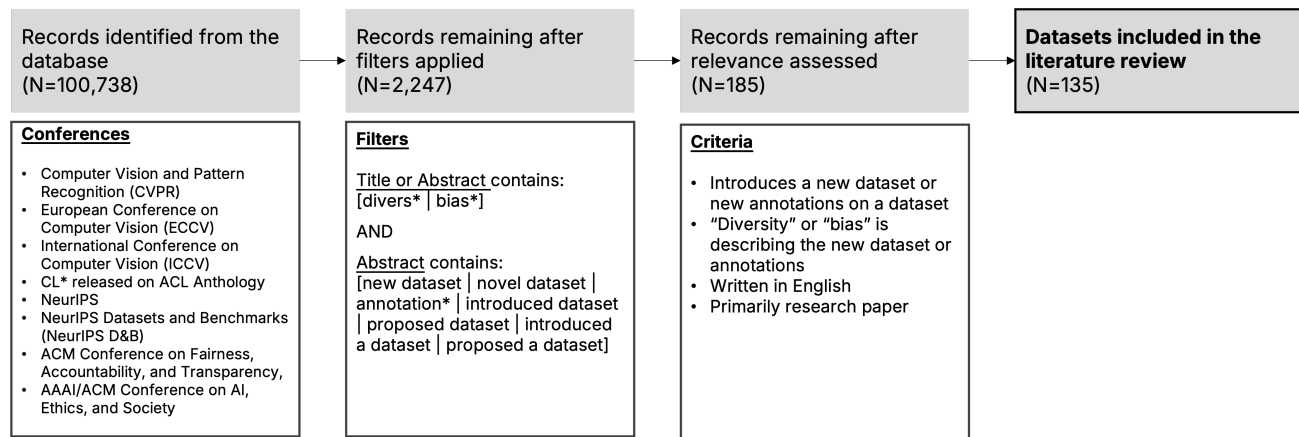
Records identified from the database
(N=100,738)

Conferences

- Computer Vision and Pattern Recognition (CVPR)
- European Conference on Computer Vision (ECCV)
- International Conference on Computer Vision (ICCV)
- CL* released on ACL Anthology
- NeurIPS
- NeurIPS Datasets and Benchmarks (NeurIPS D&B)
- ACM Conference on Fairness, Accountability, and Transparency,
- AAAI/ACM Conference on AI, Ethics, and Society

Records remaining after filters applied
(N=2,247)

Filters

Title or Abstract contains:
[divers* | bias*]

AND

Abstract contains:
[new dataset | novel dataset | annotation* | introduced dataset | proposed dataset | introduced a dataset | proposed a dataset]

Records remaining after relevance assessed
(N=185)

Criteria

- Introduces a new dataset or new annotations on a dataset
- "Diversity" or "bias" is describing the new dataset or annotations
- Written in English
- Primarily research paper

Datasets included in the literature review
(N=135)

Figure 2. Overview of the search strategy used to identify diverse datasets for inclusion in our corpus.

Dataset	Diversity Definition
Ithaca365 (Diaz et al., 2022)	"[D]ata is repeatedly recorded along a 15 km route under diverse scene (urban, highway, rural, campus), weather (snow, rain, sun), time (day/night), and traffic conditions (pedestrians, cyclists and cars)"
Vehicle Re-Identification for Aerial Image (VRAI) (Wang et al., 2019c)	"The images are taken by two moving UAVs in real urban scenarios, and the flight altitude ranges from 15m to 80m. It results in a large diversity of view-angles and pose variations, and so increases the difficulty of the corresponding ReID task."
GTA5 Crowdcounting (Wang et al., 2019b)	"Diverse Scenes. GCC dataset consists of 400 different scenes, which includes multiple types of locations. For example, indoor scenes: convenience store, pub, etc. outdoor scenes: mall, street, plaza, stadium and so on ... Diverse Environments. In order to construct the data that are close to the wild, the images are captured at a random time in a day and under a random weather conditions. In GTA5, we select seven types of weathers: clear, clouds, rain, foggy, thunder, overcast and extra sunny."
Bilingual Text Separation (BTS) (Xu et al., 2020)	"The diversity of BTS can be described at three levels: (1) scene-level diversity: it covers common life scenes including street signs, shop signs, plaques, attractions, book covers, banners, and couplets; (2) image-level diversity: appearances and geometric variances caused by camera-captured settings and background distractions such as perspective, illumination, resolution, partly blocking, blur and so on, in total including 14,250 fine-annotated text images; (3) character-level diversity: variances of character categories, up to 3,985 classes including Chinese characters, English letters, digits, common punctuation with varied fonts and sizes."

Table 1. Example quotes illustrating concrete definitions of diversity from selected surveyed datasets.

A.2. Coding Procedure

Coding was performed by DZ, JTAA, and OP. We initially coded a set of three datasets independently. After coding these initial datasets, we collaboratively discussed any discrepancies in our codes and refined the codebook definitions. Subsequently, we divided the remaining 132 datasets among the three of us for individual coding. To identify overarching themes, each of us individually generated themes, which we later grouped into larger categories during synchronous discussions.

A.3. Categorization

Definition of standalone dataset. A paper was considered to present a standalone dataset if its primary contributions were the introduction of a novel dataset, benchmark, or set of annotations. For example, if a paper introduces a new model, task, or algorithm, it would *not* be classified as presenting a standalone dataset.

Criteria for concrete definition. We applied the following criteria to determine whether a concrete definition of diversity was provided in a dataset paper. We checked whether the dataset authors explicitly stated a consistent and complete definition of diversity. For example, if a dataset was described as "diverse" in multiple ways throughout the paper, it would not be considered to have a concrete definition. Additionally, we ensured that the dataset authors listed out the specific aspects that were more diverse (e.g., different illumination conditions, sources of news). Examples of concrete diversity definitions identified are provided in Table 1.

B. Additional Dataset Collection Details

In Section 5, we categorize dataset collection methodologies into five main types. Further details on this taxonomy are provided in Table 2, with a list of datasets in our corpus categorized accordingly in Table 3.

C. Dataset Details

Here, we offer additional insights into the 135 image and text datasets coded during our literature review. Table 4 provides a taxonomy for different types of diversity we observed and Table 5 outlines the diversity type for each dataset. Furthermore, Tables 6 to 8 summarize our analysis across all datasets.

Method	Definition	Example
Webscraping ($N = 47$)	Instances are sourced and downloaded from content that is posted to the Internet, typically for purposes other than being used for creating a machine learning dataset	• **Text**: Mega-COV (Abdul-Mageed et al., 2021) consists of multi-lingual tweets sourced using Twitter's streaming API. • **Image**: Idrees et al. (2018) combine three web sources, "Flickr, Web Search, and the Hajj footage," to create UCF-QNRF, an image dataset used for crowd counting.
Crowdsourcing ($N = 9$)	Instances are sourced from a group of people that are not directly members of the research team for the purposes of creating a machine learning dataset	• **Text**: To create dialogue data for MultiDoGO, Peskov et al. (2019) ask Amazon Mechanical Turk workers to engage with an "agent" in an Wizard-of-Oz approach. • **Image**: EPIC-KITCHENS (Damen et al., 2018) contains video footage collected by 32 participants across four different countries. To note, participants contribute this footage voluntarily and without compensation.
Direct Collection ($N = 20$)	Instances are created or collected by members of the research team in the real-world	• **Text**: While there are no text datasets from our corpus in which the instances are directly collected, the GICoref (Cao & Daumé III, 2020) is an example where annotations are manually generated (i.e., directly collected) by the paper authors. • **Image**: The Waymo Open Dataset (Sun et al., 2020) is collected using specialized equipment and vehicles across three different locations: Phoenix, Mountain View, and San Francisco.
Derivative ($N = 47$)	Instances are sourced from an already existing dataset	• **Text**: Ni et al. (2019) combine review datasets from Amazon Clothing and Yelp Reviews dataset to generate a novel dataset for recommendation justification. • **Image**: The Localized Narratives dataset (Pont-Tuset et al., 2020) provides new annotations on images sourced from COCO (Lin et al., 2014), Flickr30k (Plummer et al., 2015), ADE20K (Zhou et al., 2017), and Open Images (Kuznetsova et al., 2020).
Synthetic Generation ($N = 13$)	Instances are artificially manufactured or procedurally generated rather than capturing real-world events	• **Text**: SynthBio (Yuan et al., 2021) is generated using language models which involve creating attribute lists (e.g., notabilities, nationality, birth date) and synthesizing biographies for people who possess these attributes. • **Image**: Karazija et al. (2021) create the image scenes in Clevr-Tex using a selection of photo-mapped materials, objects, and backgrounds that are procedurally generated with the Blender API.

Table 2. Taxonomy of the commonly used dataset collection methodologies identified through our literature review. We provide a definition of the collection methodology, the number of datasets (N) in our corpus that use this methodology, as well as an example of an image and text dataset from our corpus.

Method	Dataset
Webscraping	VideoCoAtt (Fan et al., 2018), VCR (Zellers et al., 2019), StockEmotion (Wei et al., 2020), FineDiving (Xu et al., 2022a), iNat2021 (Van Horn et al., 2021), Natural World Tasks (NeWT) (Van Horn et al., 2021), FEVEROUS (Aly et al., 2021), Semantic Scholar Visual Layoutenhanced Scientific Text Understanding Evaluation (S2-VLUE) (Shen et al., 2022), Mega-COV (Abdul-Mageed et al., 2021), Bajer (Zeinert et al., 2021a), GICOREF (Cao & Daumé III, 2020), StackEx (Yuan et al., 2020), Gillani & Levy (2019), DuConv (Wu et al., 2019), Durmus et al. (2019), Broad Twitter Corpus (BTC) (Derczynski et al., 2016), VPS (Miao et al., 2021), JHU-Crowd (Sindagi et al., 2019), PASS (Asano et al., 2021), LiRo (Dumitrescu et al., 2021), IconQA (Lu et al., 2021), Pratapa et al. (2022), FairytaleQA (Xu et al., 2022b), Celis et al. (2016), Ruddit (Hada et al., 2021), CMU-MOSEAS (CMU Multimodal Opinion Sentiment, Emotions and Attributes) (Bagher Zadeh et al., 2020), Video-and-Language Event Prediction (VLEP) (Lei et al., 2020a), Big BiRD: A Large, Fine-Grained, Bigram Relatedness Dataset for Examining Semantic Composition (Asaadi et al., 2019), Reddit TIFU (Kim et al., 2019), MAD (Soldan et al., 2022), Fu et al. (2021), Mapillary Traffic Sign Dataset (MTSD) (Ertler et al., 2020), TrackingNet (Müller et al., 2018), UCF-QNRF (Idrees et al., 2018), SketchyScene (Zou et al., 2018), Video gAze CommunicATION (VACATION) (Fan et al., 2019a), Query-based Video Highlights (QVHighlights) (Lei et al., 2021b), Rahman et al. (2021), Chinese LOng Text understanding andgeneration (LOT) (Guan et al., 2022), WikiEvolve (De Kock & Vlachos, 2022), Sugawara et al. (2022), Multi-lingual retrieval Typologically Diverse (Mr. TyDi) (Zhang et al., 2021), WIKIBIAS (Zhong et al., 2021), Gender Shades (Buolamwini & Gebru, 2018), iEAT (Steed & Caliskan, 2021), GAP Coreference (Webster et al., 2018), Bias in Bios (De-Arteaga et al., 2019)
Crowdsourcing	EPIC-KITCHENS (Damen et al., 2018), ObjectNet (Barbu et al., 2019), MultiDoGO (Peskov et al., 2019), SCT v1.5 (Sharma et al., 2018), Charades (Sigurdsson et al., 2016), Wiegand et al. (2021), TicketTalk (Byrne et al., 2021), Taskmaster-1 (Byrne et al., 2019), Camp Site Negotiation (CaSiNo) (Chawla et al., 2021)
Direct Collection	ETH3D (Schops et al., 2017), Cityscapes (Cordts et al., 2016), SUNRGB-D (Song et al., 2015), RID (Wan et al., 2018), Wang et al. (2019c), Waymo Open Dataset (Sun et al., 2020), Zhang et al. (2020), HOI4D (Liu et al., 2022), Person30k (Bai et al., 2021), VisDrone-DET 2018 (Cao et al., 2021), Vehicle Re-Identification for Aerial Image (VRAI) (Wang et al., 2019c), Drive&Act (Martin et al., 2019), Wilddash2 (WD2) (Zendel et al., 2022), VTUAV (Zhang et al., 2022), Nutrition5k (Thames et al., 2021), Habitat-Matterprot 3D (HM3D) (Yadav et al., 2023), Ithaca365 (Diaz et al., 2022), Multimodal Audio-Visual Detection (MAVD) (Valverde et al., 2021), FreiHAND (Zimmermann et al., 2019), Vision-based Fallen Person (VFP290K) (An et al., 2021)
Derivative	SUNRGB-D (Song et al., 2015), ADE-Affordance (Chuang et al., 2018), OK-VQA (Marino et al., 2019), GQA (Hudson & Manning, 2019), VCR (Zellers et al., 2019), DAVSOD (Fan et al., 2019a), CoSOD3k (Fan et al., 2020), KeypointNet (You et al., 2020), ArtEmis v2.0 (Mohamed et al., 2022), VidSitu (Sadhu et al., 2021), Tracking Any Object (TAO) (Dave et al., 2020), HowToVQA69M (Yang et al., 2021), iVQA (Yang et al., 2021), CropHarvest (Tseng et al., 2021), FIBER (Castro et al., 2022), DocRED (Huang et al., 2022), Sun et al. (2021), YASO (Orbach et al., 2021), Mickey Corpus (Lin et al., 2021), X-CSQA (Lin et al., 2021), ConvoSumm (Fabbri et al., 2021a), Room-Across-Room (RxR) (Ku et al., 2020), Ni et al. (2019), Johnson & Goldwasser (2018), OpoSum (Angelidis & Lapata, 2018), Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008), WIDERFACE-DEMO (Yang et al., 2022), TVR (Lei et al., 2020a), PASS (Asano et al., 2021), LiRo (Dumitrescu et al., 2021), AP-10K (Yu et al., 2021), Drawzeski et al. (2021), MightyMorph (Goldman & Tsarfaty, 2021), Amazon Customer Reviews (O'Neill et al., 2021), mTVR (Lei et al., 2021a), ArSarcasm (Farha & Magdy, 2020), R4C (Inoue et al., 2020), Lepp & Levow (2020), Schulz et al. (2019), MAD (Soldan et al., 2022), VIdeo Panoptic Segmentation in the Wild (VIPSeg) (Miao et al., 2022a), TrackingNet (Müller et al., 2018), Localized Narratives (Pont-Tuset et al., 2020), Zhao et al. (2021), Hemani et al. (2021), Culkin et al. (2021), Racial Faces in the Wild (RFW) (Wang et al., 2019c)
Synthetic Generation	CLEVR (Johnson et al., 2017), Synthia (Ros et al., 2016), GTA5 Crowd Counting (Wang et al., 2019c), FakeAVCeleb (Khalid et al., 2021), MotSynth (Fabbri et al., 2021b), ClevrTex (Karazija et al., 2021), SketchyScene (Zou et al., 2018), 3D Furnished Rooms with layOuts and semaNTics (3D-FRONT) (Fu et al., 2021), SynthBio (Yuan et al., 2021), Chinese LOng Text understanding andgeneration (LOT) (Guan et al., 2022), Culkin et al. (2021), Winobias (Zhao et al., 2018), Winogender (Rudinger et al., 2018)

Table 3. Collection methodologies employed in the datasets encompass a range of approaches: *webscraping*, involving the retrieval of instances from the Internet; *crowdsourcing*, which entails soliciting contributions from individuals not affiliated with the research team; *direct collection*, where data instances are gathered by members of the research team themselves; *derivative*, which leverages one or multiple existing datasets; and *synthetic generation*, wherein instances are artificially generated rather than derived from real-world events. Note that each dataset may utilize multiple collection methodologies.

Diversity	Definition	Example
Composition (N=58)	Variety in what a dataset instance contains, such as linguistic properties (e.g., language, vocabulary used), scene or background, objects, viewpoint, image properties (e.g., resolution, focal length), or pose	• **Text**: X-CSQA (Lin et al., 2021) is collected to extend question-answering text evaluation beyond just English into multiple other languages • **Image**: Nutrition5k (Thames et al., 2021) contains a "wide variety of ingredients, portion sizes, and dish complexities."
Source (N=26)	Variety in where data instances are collected from, such as web source online or geographic origin	• **Text**: ConvoSumm (Fabbri et al., 2021a) draws from multiple online sources (New York Times comments, Reddit, StackExchange, and email threads) • **Image**: Tseng et al. (2021) sample their crop dataset across several countries
Domain (N=18)	Variety in the "topic area" of the data instances, such as what disciplines the text is drawn from or what artistic style is represented in the image	• **Text**: S2-VLUE (Shen et al., 2022) consists of scientific papers from 19 different academic disciplines • **Image**: TVR (Lei et al., 2020b) contains videos from six TV shows across diverse genres
Subject (N=16)	Representation of human subjects in the dataset, such as by protected attributes (e.g., gender, race, age), physical characteristics (e.g., skin tone, weight, height), nationality, socioeconomic status, and language	• **Text**: SynthBio (Yuan et al., 2021) is a synthetically generated evaluation set for WikiBio (Lebret et al., 2016) which is balanced with respect to the gender and nationality of biography subjects • **Image**: An et al. (2021) recruit participants with different "skin color and age" as data subjects
Annotator (N=2)	Representation of annotator backgrounds, such as demographic background, domain expertise, or political affiliation	• **Text**: Zeinert et al. (2021a) recruit annotators with diversity "in gender, age, occupation / background (linguistic and ethnographic knowledge), region (spoken dialects)" to label misogyny online. • **Image**: N/A

Table 4. Taxonomy of the definitions of diversity identified through our literature review. We provide a definition of diversity, the number of datasets (N) in our corpus that use the respective definition of diversity, as well as an example of an image and text dataset from our corpus. We do not find any image datasets in our corpus that seek to use a more diverse annotator pool.

Category	Dataset
Composition Diversity	ETH3D (Schops et al., 2017), Cityscapes (Cordts et al., 2016), Synthia (Ros et al., 2016), ADE-Affordance (Chuang et al., 2018), RID (Wan et al., 2018), VideoCoAtt (Fan et al., 2018), OK-VQA (Marino et al., 2019), Wang et al. (2019c), GTA5 Crowd Counting (Wang et al., 2019c), DAVSOD (Fan et al., 2019a), CoSOD3k (Fan et al., 2020), Zhang et al. (2020), HOI4D (Liu et al., 2022), Person30k (Bai et al., 2021), Tracking Any Object (TAO) (Dave et al., 2020), VisDrone-DET 2018 (Cao et al., 2021), Vehicle Re-Identification for Aerial Image (VRAI) (Wang et al., 2019c), ObjectNet (Barbu et al., 2019), Mickey Corpus (Lin et al., 2021), X-CSQA (Lin et al., 2021), Room-Across-Room (RxR) (Ku et al., 2020), Broad Twitter Corpus (BTC) (Derczynski et al., 2016), Wilddash2 (WD2) (Zendel et al., 2022), DAD-3DHeads (Martyniuk et al., 2022), Nutrition5k (Thames et al., 2021), VPS (Miao et al., 2021), Charades (Sigurdsson et al., 2016), Trans10K (Xie et al., 2020), Silberman et al. (2012), Waymo Open Motion Dataset (Ettinger et al., 2021), MotSynth (Fabbri et al., 2021b), JHU-Crowd (Sindagi et al., 2019), Habitat-Matterprot 3D (HM3D) (Yadav et al., 2023), AP-10K (Yu et al., 2021), ClevrTex (Karazija et al., 2021), Drawzeski et al. (2021), Mighty-Morph (Goldman & Tsarfaty, 2021), CMU-MOSEAS (CMU Multimodal Opinion Sentiment, Emotions and Attributes) (Bagher Zadeh et al., 2020), Taskmaster-1 (Byrne et al., 2019), MAD (Soldan et al., 2022), Bilingual Text Segmentation (BTS) (Xu et al., 2022b), VIdeo Panoptic Segmentation in the Wild (VIPSeg) (Miao et al., 2022a), Ithaca365 (Diaz et al., 2022), Fu et al. (2021), Multimodal Audio-Visual Detection (MAVD) (Valverde et al., 2021), Mapillary Traffic Sign Dataset (MTSD) (Ertler et al., 2020), TrackingNet (Müller et al., 2018), Localized Narratives (Pont-Tuset et al., 2020), UCF-QNRF (Idrees et al., 2018), Cars Overhead with Context (COWC) (Mundhenk et al., 2016), 3D Furnished Rooms with layOuts and semaNTics (3D-FRONT) (Fu et al., 2021), FreiHAND (Zimmermann et al., 2019), Video gAze CommunicATION (VACATION) (Fan et al., 2019a), Query-based Video Highlights (QVHighlights) (Lei et al., 2021b), Vision-based Fallen Person (VFP290K) (An et al., 2021), Culkin et al. (2021), Camp Site Negotiation (CaSiNo) (Chawla et al., 2021), Multi-lingual retrieval Typologically Diverse (Mr. TyDi) (Zhang et al., 2021)
Source Diversity	VideoCoAtt (Fan et al., 2018), Waymo Open Dataset (Sun et al., 2020), iNat2021 (Van Horn et al., 2021), EPIC-KITCHENS (Damen et al., 2018), Drive&Act (Martin et al., 2019), CropHarvest (Tseng et al., 2021), Semantic Scholar Visual Layoutenhanced Scientific Text Understanding Evaluation (S2-VLUE) (Shen et al., 2022), YASO (Orbach et al., 2021), Mega-COV (Abdul-Mageed et al., 2021), ConvoSumm (Fabbri et al., 2021a), MultiDoGO (Peskov et al., 2019), Durmus et al. (2019), Broad Twitter Corpus (BTC) (Derczynski et al., 2016), Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008), Wilddash2 (WD2) (Zendel et al., 2022), VTUAV (Zhang et al., 2022), The Dollar Street Dataset (Rojas et al., 2022), PASS (Asano et al., 2021), Ithaca365 (Diaz et al., 2022), Multimodal Audio-Visual Detection (MAVD) (Valverde et al., 2021), Mapillary Traffic Sign Dataset (MTSD) (Ertler et al., 2020), Cars Overhead with Context (COWC) (Mundhenk et al., 2016), Video gAze CommunicATION (VACATION) (Fan et al., 2019a), Query-based Video Highlights (QVHighlights) (Lei et al., 2021b), Rahman et al. (2021), Sugawara et al. (2022)
Domain Diversity	OK-VQA (Marino et al., 2019), StockEmotion (Wei et al., 2020), FineDiving (Xu et al., 2022a), VidSitu (Sadhu et al., 2021), Tracking Any Object (TAO) (Dave et al., 2020), EPIC-KITCHENS (Damen et al., 2018), Drive&Act (Martin et al., 2019), OpoSum (Angelidis & Lapata, 2018), VPS (Miao et al., 2021), TVR (Lei et al., 2020a), Waymo Open Motion Dataset (Ettinger et al., 2021), GQA (Hudson & Manning, 2019), HowToVQA69M (Yang et al., 2021), IconQA (Lu et al., 2021), Video-and-Language Event Prediction (VLEP) (Lei et al., 2020a), MAD (Soldan et al., 2022), Fu et al. (2021), Localized Narratives (Pont-Tuset et al., 2020)
Subject Diversity	Martin et al. (2019); Khalid et al. (2021); Cao & Daumé III (2020); Yang et al. (2022); Rojas et al. (2022); Fabbri et al. (2021b); Byrne et al. (2021); Zimmermann et al. (2019); An et al. (2021); Yuan et al. (2021); Chawla et al. (2021); Wang et al. (2019c); Buolamwini & Gebru (2018); Zhao et al. (2018); Rudinger et al. (2018); Webster et al. (2018)
Annotator Diversity	Bajer (Zeinert et al., 2021a), mTVR (Lei et al., 2021a)
Reduce Dataset Bias	CLEVR (Johnson et al., 2017), VCR (Zellers et al., 2019), ArtEmis v2.0 (Mohamed et al., 2022), iVQA (Yang et al., 2021), FEVEROUS (Aly et al., 2021), FIBER (Castro et al., 2022), DocRED (Huang et al., 2022), SCT v1.5 (Sharma et al., 2018), Salient Objects in Clutter (SOC) (Fan et al., 2018), Wiegand et al. (2021), ArSarcasm (Farha & Magdy, 2020), R4C (Inoue et al., 2020), Big BiRD: A Large, Fine-Grained, Bigram Relatedness Dataset for Examining Semantic Composition (Asaadi et al., 2019), Reddit TIFU (Kim et al., 2019), Hemani et al. (2021), Chinese LOng Text understanding andgeneration (LOT) (Guan et al., 2022), WikiEvolve (De Kock & Vlachos, 2022), WIKIBIAS (Zhong et al., 2021)
Promote Diversity (or Fairness) in Downstream Applications	Natural World Tasks (NeWT) (Van Horn et al., 2021), Sun et al. (2021), StackEx (Yuan et al., 2020), Gillani & Levy (2019), DuConv (Wu et al., 2019), Ni et al. (2019), Johnson & Goldwasser (2018), LiRo (Dumitrescu et al., 2021), FairytaleQA (Xu et al., 2022b), Celis et al. (2016), Lepp & Levow (2020), Zhao et al. (2021), iEAT (Steed & Caliskan, 2021), Bias in Bios (De-Arteaga et al., 2019)
Not specified	SUNRGB-D (Song et al., 2015), KeypointNet (You et al., 2020), Pratapa et al. (2022), Schulz et al. (2019)

Table 5. Surveyed datasets cover various categories of diversity, alongside additional identified factors such as *reducing dataset bias* and *promoting diversity (or fairness) in downstream applications*. Note that each dataset may encompass multiple categories of diversity.

Dataset	Standalone	Concret defn.	Justification	Collection trade-offs	Quality info	Annot. info
3D Furnished Rooms with layOuts and semaNTics (3D-FRONT) (Fu et al., 2021)						NA
ADE-Affordance (Chuang et al., 2018)			✓			
AP-10K (Yu et al., 2021)	✓	✓	✓		✓	✓
Amazon Customer Reviews (O'Neill et al., 2021)	✓	NA	NA	✓	✓	✓
ArSarcasm (Farha & Magdy, 2020)		NA	NA		✓	✓
ArtEmis v2.0 (Mohamed et al., 2022)	✓	NA	NA	✓	✓	
Bajer (Zeinert et al., 2021a)	✓	✓	✓	✓	✓	✓
Bias in Bios (De-Arteaga et al., 2019)		NA	NA	✓		NA
Big BiRD: A Large, Fine-Grained, Bigram Relatedness Dataset for Examining Semantic Composition (Asaadi et al., 2019)	✓				✓	
Bilingual Text Segmentation (BTS) (Xu et al., 2022b)		✓	✓		✓	
Broad Twitter Corpus (BTC) (Derczynski et al., 2016)	✓	✓	✓	✓	✓	✓
CLEVR (Johnson et al., 2017)	✓	NA	NA			
CMU-MOSEAS (CMU Multimodal Opinion Sentiment, Emotions and Attributes) (Bagher Zadeh et al., 2020)	✓	✓			✓	✓
Camp Site Negotiation (CaSiNo) (Chawla et al., 2021)				✓		
Cars Overhead with Context (COWC) (Mundhenk et al., 2016)						
Charades (Sigurdsson et al., 2016)	✓					
Chinese LOng Text understanding andgeneration (LOT) (Guan et al., 2022)		✓	✓		✓	
Cityscapes (Cordts et al., 2016)	✓	✓	✓	✓	✓	✓
ClevrTex (Karazija et al., 2021)	✓	✓	✓			NA
CoSOD3k (Fan et al., 2020)	✓	✓			✓	✓
ConvoSumm (Fabbri et al., 2021a)					✓	✓
CropHarvest (Tseng et al., 2021)	✓		✓	✓		NA
DAD-3DHeads (Martyniuk et al., 2022)		✓			✓	
DAVSOD (Fan et al., 2019a)		✓				✓
DocRED (Huang et al., 2022)		NA	NA		✓	✓
Drive&Act (Martin et al., 2019)	✓		✓			NA
DuConv (Wu et al., 2019)		NA	NA			NA
EPIC-KITCHENS (Damen et al., 2018)	✓	✓		✓	✓	
ETH3D (Schops et al., 2017)		✓			✓	NA
FEVEROUS (Aly et al., 2021)		NA	NA	✓	✓	✓
FIBER (Castro et al., 2022)		NA	NA	✓	✓	✓
FairytaleQA (Xu et al., 2022c)	✓				✓	✓
FakeAVCeleb (Khalid et al., 2021)	✓	✓	✓		✓	NA
FineDiving (Xu et al., 2022a)			✓			
FreiHAND (Zimmermann et al., 2019)		NA	NA			
GAP Coreference (Webster et al., 2018)		NA	NA		✓	
GICOREF (Cao & Daumé III, 2020)		NA	NA		✓	
GQA (Hudson & Manning, 2019)		✓	✓			
GTA5 Crowd Counting (Wang et al., 2019b)		✓	✓	✓		NA
Gender Shades (Buolamwini & Gebru, 2018)		NA	NA	✓	✓	✓
HOI4D (Liu et al., 2022)	✓	✓	✓			
Habitat-Matterprot 3D (HM3D) (Yadav et al., 2023)	✓				✓	
HowToVQA69M (Yang et al., 2021)					✓	NA
IconQA (Lu et al., 2021)			✓		✓	
Ithaca365 (Diaz et al., 2022)	✓	✓	✓			
JHU-Crowd (Sindagi et al., 2019)		✓				NA
KeypointNet (You et al., 2020)						

Table 6. (Table continues in Tables 7 and 8). Summary of our analysis across all 135 image and text datasets. Column labels: *standalone*, indicating whether the dataset paper exclusively presents a dataset without introducing novel models, tasks, or algorithms; *concrete definition*, indicating whether the dataset paper includes a clear definition of diversity (*NA* for datasets focusing solely on bias, not diversity); *justification*, denoting whether the dataset paper discusses the rationale behind incorporating diversity into the dataset (*NA* for datasets focusing solely on bias, not diversity); *collection trade-offs*, indicating whether the dataset paper discusses drawbacks of the chosen collection methodology or provides considerations for alternative methodologies; *quality information*, signifying whether the dataset paper includes details on how dataset quality was verified (e.g., manual inspection by creators, inter-annotator agreement scores); and *annotation information*, specifying whether additional details about annotators were provided (e.g., training they underwent, qualification criteria). We rely solely on the description of the dataset provided in the body of the accompanying paper when completing this table.

Dataset	Standalone	Concrete defn.	Justification	Collection trade-offs	Quality info	Annot. info
LiRo (Dumitrescu et al., 2021)		NA	NA		✓	
Localized Narratives (Pont-Tuset et al., 2020)		✓			✓	
MAD (Soldan et al., 2022)					✓	✓
Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008)	✓				✓	
Mapillary Traffic Sign Dataset (MTSD) (Ertler et al., 2020)	✓	✓	✓		✓	
Mega-COV (Abdul-Mageed et al., 2021)		✓	✓			NA
Mickey Corpus (Lin et al., 2021)		✓	✓	✓	✓	NA
MightyMorph (Goldman & Tsarfaty, 2021)		NA	NA			NA
MotSynth (Fabbri et al., 2021b)	✓		✓	✓		NA
Multi-lingual retrieval Typologically Diverse (Mr. TyDi) (Zhang et al., 2021)			✓	✓		
MultiDoGO (Peskov et al., 2019)		✓			✓	✓
Multimodal Audio-Visual Detection (MAVD) (Valverde et al., 2021)		✓				NA
Natural World Tasks (NeWT) (Van Horn et al., 2021)	✓		✓		✓	NA
Nutrition5k (Thames et al., 2021)		✓	✓	✓		NA
OK-VQA (Marino et al., 2019)	✓				✓	
ObjectNet (Barbu et al., 2019)	✓	NA	NA	✓	✓	NA
OpoSum (Angelidis & Lapata, 2018)					✓	
PASS (Asano et al., 2021)	✓	✓		✓	✓	NA
Person30k (Bai et al., 2021)			✓			
Query-based Video Highlights (QVHighlights) (Lei et al., 2021b)					✓	
R4C (Inoue et al., 2020)		NA	NA		✓	✓
RID (Wan et al., 2018)		✓				NA
Racial Faces in the Wild (RFW) (Wang et al., 2019a)			✓			
Reddit TIFU (Kim et al., 2019)		✓			✓	NA
Room-Across-Room (RxR) (Ku et al., 2020)		✓	✓	✓	✓	✓
Ruddit (Hada et al., 2021)	✓				✓	✓
SCT v1.5 (Sharma et al., 2018)		NA	NA			NA
SUNRGB-D (Song et al., 2015)	✓	NA	NA		✓	✓
Salient Objects in Clutter (SOC) (Fan et al., 2018)	✓	✓	✓			
Semantic Scholar Visual Layoutenhanced Scientific Text Understanding Evaluation (S2-VLUE) (Shen et al., 2022)			✓			
SketchyScene (Zou et al., 2018)		✓			✓	
StackEx (Yuan et al., 2020)				✓	✓	NA
StockEmotion (Wei et al., 2020)		✓	✓		✓	
SynthBio (Yuan et al., 2021)		NA	NA	✓	✓	✓
Synthia (Ros et al., 2016)	✓				✓	NA
TVR (Lei et al., 2020b)		✓	✓		✓	✓
Taskmaster-1 (Byrne et al., 2019)				✓	✓	✓
The Dollar Street Dataset (Rojas et al., 2022)	✓	✓	✓	✓	✓	NA
TicketTalk (Byrne et al., 2021)			✓			
Tracking Any Object (TAO) (Dave et al., 2020)	✓		✓		✓	
TrackingNet (Müller et al., 2018)	✓				✓	
Trans10K (Xie et al., 2020)		✓	✓			
UCF-QNRF (Idrees et al., 2018)		✓			✓	
VCR (Zellers et al., 2019)				✓	✓	✓
VIdeo Panoptic Segmentation in the Wild (VIPSeg) (Miao et al., 2022a)		✓	✓	✓	✓	
VPS (Miao et al., 2021)				✓		
VTUAV (Zhang et al., 2022)		✓			✓	NA
Vehicle Re-Identification for Aerial Image (VRAI) (Wang et al., 2019c)		✓	✓			

Table 7. (Continued from Table 6). Summary of our analysis across all 135 image and text datasets. Column labels: *standalone*, indicating whether the dataset paper exclusively presents a dataset without introducing novel models, tasks, or algorithms; *concrete definition*, indicating whether the dataset paper includes a clear definition of diversity (*NA* for datasets focusing solely on bias, not diversity); *justification*, denoting whether the dataset paper discusses the rationale behind incorporating diversity into the dataset (*NA* for datasets focusing solely on bias, not diversity); *collection trade-offs*, indicating whether the dataset paper discusses drawbacks of the chosen collection methodology or provides considerations for alternative methodologies; *quality information*, signifying whether the dataset paper includes details on how dataset quality was verified (e.g., manual inspection by creators, inter-annotator agreement scores); and *annotation information*, specifying whether additional details about annotators were provided (e.g., training they underwent, qualification criteria). We rely solely on the description of the dataset provided in the body of the accompanying paper when completing this table.

Dataset	Standalone	Concret defn.	Justification	Collection trade-offs	Quality info	Annot. info
VidSitu (Sadhu et al., 2021)	✓				✓	✓
Video gAze CommunicATION (VACATION) (Fan et al., 2019b)		✓	✓		✓	
Video-and-Language Event Prediction (VLEP) (Lei et al., 2020a)		✓			✓	
VideoCoAtt (Fan et al., 2018)		✓	✓			
VisDrone-DET 2018 (Cao et al., 2021)		✓				
Vision-based Fallen Person (VFP290K) (An et al., 2021)	✓	✓	✓		✓	✓
WIDERFACE-DEMO (Yang et al., 2022)			✓			
WIKIBIAS (Zhong et al., 2021)		NA	NA		✓	
Waymo Open Dataset (Sun et al., 2020)	✓	✓	✓		✓	
Waymo Open Motion Dataset (Ettinger et al., 2021)			✓			NA
WikiEvolve (De Kock & Vlachos, 2022)		NA	NA			NA
Wilddash2 (WD2) (Zendel et al., 2022)		✓		✓		
Winobias (Zhao et al., 2018)		NA	NA			NA
Winogender (Rudinger et al., 2018)		NA	NA			
X-CSQA (Lin et al., 2021)		✓	✓	✓		NA
YASO (Orbach et al., 2021)	✓		✓	✓	✓	
Culkin et al. (2021)					✓	
Drawzeski et al. (2021)			✓	✓		
Durmus et al. (2019)		✓	✓			
Fu et al. (2021)		✓	✓			NA
Gillani & Levy (2019)		NA	NA			NA
Johnson & Goldwasser (2018)		NA	NA		✓	✓
Lepp & Levow (2020)		NA	NA		✓	✓
Ni et al. (2019)		NA	NA			NA
Pratapa et al. (2022)		✓				NA
Schulz et al. (2019)		NA	NA		✓	✓
Silberman et al. (2012)						
Sugawara et al. (2022)	✓		✓		✓	✓
Sun et al. (2021)			✓		✓	
Wang et al. (2019c)		✓	✓			NA
Wiegand et al. (2021)		NA	NA		✓	✓
Zhang et al. (2020)						NA
Zhao et al. (2021)		NA	NA	✓	✓	✓
Celis et al. (2016)		✓	✓			NA
Hemani et al. (2021)		NA	NA			NA
Rahman et al. (2021)			✓		✓	✓
iEAT (Steed & Caliskan, 2021)		NA	NA			NA
iNat2021 (Van Horn et al., 2021)	✓	✓	✓	✓	✓	
iVQA (Yang et al., 2021)					✓	
mTVR (Lei et al., 2021a)		✓			✓	✓

Table 8. (Continued from Table 7). Summary of our analysis across all 135 image and text datasets. Column labels: *standalone*, indicating whether the dataset paper exclusively presents a dataset without introducing novel models, tasks, or algorithms; *concrete definition*, indicating whether the dataset paper includes a clear definition of diversity (*NA* for datasets focusing solely on bias, not diversity); *justification*, denoting whether the dataset paper discusses the rationale behind incorporating diversity into the dataset (*NA* for datasets focusing solely on bias, not diversity); *collection trade-offs*, indicating whether the dataset paper discusses drawbacks of the chosen collection methodology or provides considerations for alternative methodologies; *quality information*, signifying whether the dataset paper includes details on how dataset quality was verified (e.g., manual inspection by creators, inter-annotator agreement scores); and *annotation information*, specifying whether additional details about annotators were provided (e.g., training they underwent, qualification criteria). We rely solely on the description of the dataset provided in the body of the accompanying paper when completing this table.