MEASURING DIVERSITY IN DATASETS

Dorothy Zhao
Stanford University

Jerone T.A. Andrews
Sony AI, Tokyo

Orestis Papakyriakopoulos
Technical University of Munich

Alice Xiang
Sony AI, Seattle

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 in datasets. Our research explores the implications of this issue, specifically analyzing "diversity" across 135 image and text datasets. Drawing from social sciences, we leverage principles from measurement theory to pinpoint considerations and offer recommendations on conceptualization, operationalization, and evaluation of diversity in ML datasets. Our recommendations extend to broader implications for ML research, advocating for a more nuanced and well-defined 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., 2021a; Birhane et al., 2021; Denton et al., 2020). From inception to release, datasets emerge as political artifacts, etched with the signature of their creators' 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., 2021a) 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) or "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, or 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 assert?

In this position paper, we leverage *measurement theory*, a framework widely employed in the social sciences, to develop 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).

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). There is a growing emphasis on using measurement theory to improve the precision of 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) suggests 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 focused on assessing the validity of natural language processing benchmark datasets for evaluating stereotyping, our approach treats the data collection process as the measurement to validate *itself*. This involves questioning the reliability of the collection process and exploring methods for researchers to validate claimed properties in the resulting dataset.

Zooming in on the intricacies of diversity, we unearth inconsistencies in its definitions, a pervasive challenge that our paper addresses head-on. 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.

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 1: 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.

3 METHOD

To inform our position, we conducted a systematic literature review encompassing 135 datasets presented as being more "diverse". We 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), 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. See Appendix A for 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. See Table 1 for a taxonomy of diversity definitions.

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". Here, *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. 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 throughout the dataset collection process (Check & Schutt, 2011; Babbie, 2020). Highlighting the significance of clear definitions, examine the nuanced distinction between seemingly similar claims such as representing "diverse scenarios" (Miao et al., 2022b) and "diverse social scenes" (Fan et al., 2018b). 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, the authors 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 the 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 instance, 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 – for detailed information, refer to Appendix D. Within this section, we spotlight deficiencies in current collection processes and present recommendations to surmount these limitations.

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 algorithmic contribution. In particular, out of the 135 papers analyzed, only 38 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. 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-related efforts in comparison to the development of models or algorithmic contributions (Sambasivan et al., 2021). Papers exclusively dedicated to datasets typically do not find their way into "top-tier" research publications (Scheuerman et al., 2021; Heinzerling, 2022), potentially dissuading authors from allocating sufficient time and space to elaborate 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 87.4% 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., 2023; 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., 2016). This issue may be exacerbated by, e.g., consensus-based quality filtering, which tends to exclude noncanonical examples (Mayo et al., 2022). Synthetic generation, offering similar benefits to web scraping with increased systematic control (Johnson et al., 2017a; 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 aligns best with 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 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 methods and quality assessment criteria, necessary for evaluating the validity of the collection process and the utility of the dataset. Furthermore, closed models and datasets raise additional concerns (Bommasani et al., 2023). Examples include closed- and open-sourced 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 the 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., 2020b), 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 MOTSynth employing nine categories to JHU-Crowd using three.

Clear definitions are paramount to prevent misinterpretations arising from diverse operationalizations, especially given that different operationalizations of the same construct can yield markedly distinct results. For example, indicators can be categorized as either *objective* or *subjective*—indicating reliance on explicit criteria and assessment by external observers or involving personal perception or evaluation, respectively. 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 authors operationalize it as an objective measurement by reporting the majority sentiment among crowdworkers. This choice results in a measurement distinct from what would be obtained if sentiment were operationalized as subjective.

The approach by Orbach et al. (2021) also introduces potential concerns related to *selection bias*. The authors exclude under-performing workers based on a "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 the decision-making process facilitates the evaluation of operationalization and collection process validity. For exam-

ple, Zhao et al. (2021b) revealed a demographically imbalanced annotator pool, a limitation when collecting skin color and gender expression annotations on the crowdsourcing platform 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 borrowed from measurement theory literature: *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 is 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. However, 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 neglect of the labor invested in dataset creation. Crowdworkers are often viewed as costs to be optimized, failing to acknowledge their substantial contributions to the dataset (Williams et al., 2022; Shmueli et al., 2021; Miceli et al., 2022a). Likewise, the significant effort required to ensure a high-quality dataset is frequently disregarded.

6.2 RECOMMENDATIONS

Inter-annotator agreement. One area where methods for assessing reliability are already in practice is the measurement of inter-annotator agreement. In crowdsourcing, a common approach involves having multiple annotators label an instance, and the majority label is then adopted (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. Prior research has demonstrated that capturing diverse annotator perspectives (Gordon et al., 2022) can be beneficial. Relying on aggregation methods, such as majority voting, may systematically erase certain groups' perspectives from the dataset 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 applicable. 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, and consistent results indicate reliability (Guttman, 1945). This principle is particularly applicable 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, 2023; 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 the data distribution.

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 papers may present metadata frequency or other summary statistics about the dataset, these metrics do not consistently align with diversity dimensions described when motivating a dataset.

Overreliance on downstream evaluation. Another prevalent evaluation method involves benchmarking the downstream performance of a newly proposed model. This approach is observed in 49% 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 indicating an actual increase 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 the *dataset bias* phenomenon (Torralba & Efros, 2011), this technique enables researchers to compare datasets. By utilizing existing datasets with similar 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 predicted label, indicating correlated constructs of diversity. Prior work (Khan & Fu, 2021) adopted Fleiss's κ to measure prediction consistency across models; however, a limitation of using cross-dataset generalization is that it requires similar taxonomies (for the predicted label) and similar 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 (2022) introduced the

Vendi Score, inspired by ecology and quantum statistical mechanics, to assess diversity within categories of image and text datasets. 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, 2022), datasets should be benchmarked across a multiplicity 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 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 answer complexity. (Comparing *answer complexity* may not be reasonable for testing discriminant validity if the interest is in *answer diversity*, as these are related constructs.) Nonetheless, prior research (Goyal et al., 2017; Agrawal et al., 2018; Hudson & Manning, 2019; Johnson et al., 2017a) identified language biases in how questions and answers are formulated. 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 Conclusion

This paper 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. This, coupled with underlying methodological issues, such as flawed experimental or dataset designs, inadequate statistical methods, and improper data analysis, collectively contributes to the pervasive challenge of irreproducibility. By highlighting these systemic issues, our paper advocates for a more robust and standardized approach in dataset creation, fostering transparency, reliability, and reproducibility in the broader scientific landscape.

References

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

Muhammad Abdul-Mageed, AbdelRahim Elmadany, El Moatez Billah Nagoudi, Dinesh Pabbi, Kunal Verma, and Rannie Lin. Mega-cov: A billion-scale dataset of 100+ languages for covid-19. In *Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.298. URL http://dx.doi.org/10.18653/v1/2021.eacl-main.298.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 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.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. Feverous: Fact extraction and verification over unstructured and structured information. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS DB)*, 2021.

Jaeju An, Jeongho Kim, Hanbeen Lee, Jinbeom Kim, Junhyung Kang, Saebyeol Shin, Minha Kim, Donghee Hong, and Simon S Woo. Vfp290k: A large-scale benchmark dataset for vision-based fallen person detection. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS DB)*, 2021.

Jerone TA Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, Shruti Nagpal, and Alice Xiang. Ethical considerations for collecting human-centric image datasets. In *Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS DB)*, 2023.

McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 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.

Stefanos Angelidis and Mirella Lapata. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1403. URL http://dx.doi.org/10.18653/v1/d18-1403.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Shima Asaadi, Saif Mohammad, and Svetlana Kiritchenko. 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)*, pp. 505–516, 2019.

Yuki M Asano, Christian Rupprecht, Andrew Zisserman, and Andrea Vedaldi. Pass: An imagenet replacement for self-supervised pretraining without humans. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS DB)*, 2021.

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

AmirAli Bagher Zadeh, Yansheng Cao, Simon Hessner, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. Cmu-moseas: A multimodal language dataset for spanish, portuguese, german and french. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.141. URL http://dx.doi.org/10.18653/v1/2020.emnlp-main.141.

Yan Bai, Jile Jiao, Wang Ce, Jun Liu, Yihang Lou, Xuetao Feng, and Ling-Yu Duan. Person30k: A dual-meta generalization network for person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021. doi: 10.1109/cvpr46437.2021.00216. URL http://dx.doi.org/10.1109/cvpr46437.2021.00216.

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

Jack Bandy and Nicholas Vincent. Addressing" documentation debt" in machine learning research: A retrospective datasheet for bookcorpus. *arXiv preprint arXiv:2105.05241*, 2021.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

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

A Stevie Bergman, Lisa Anne Hendricks, Maribeth Rauh, Boxi Wu, William Agnew, Markus Kunesch, Isabella Duan, Iason Gabriel, and William Isaac. Representation in ai evaluations. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.

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

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The values encoded in machine learning research. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

Abeba Birhane, Vinay Prabhu, Sang Han, and Vishnu Naresh Boddeti. On hate scaling laws for data-swamps. *arXiv preprint arXiv:2306.13141*, 2023.

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

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, 2020.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1004–1015, 2021.

Su Lin Blodgett, Q Vera Liao, Alexandra Olteanu, Rada Mihalcea, Michael Muller, Morgan Klaus Scheuerman, Chenhao Tan, and Qian Yang. Responsible language technologies: Foreseeing and mitigating harms. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2022.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. The foundation model transparency index. *arXiv preprint arXiv:2310.12941*, 2023.

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

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/d19-1459. URL http://dx.doi.org/10.18653/v1/d19-1459.

Bill Byrne, Karthik Krishnamoorthi, Saravanan Ganesh, and Mihir Kale. Tickettalk: Toward human-level performance with end-to-end, transaction-based dialog systems. In *Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.55. URL http://dx.doi.org/10.18653/v1/2021.acl-long.55.

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

Yang Trista Cao and Hal Daumé III. Toward gender-inclusive coreference resolution. In *Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.418. URL http://dx.doi.org/10.18653/v1/2020.acl-main.418.

Yaru Cao, Zhijian He, Lujia Wang, Wenguan Wang, Yixuan Yuan, Dingwen Zhang, Jinglin Zhang, Pengfei Zhu, Luc Van Gool, Junwei Han, Steven Hoi, Qinghua Hu, Ming Liu, Chong Cheng, Fanfan Liu, Guojin Cao, Guozhen Li, Hongkai Wang, Jianye He, Junfeng Wan, Qi Wan, Qi Zhao, Shuchang Lyu, Wenzhe Zhao, Xiaoqiang Lu, Xingkui Zhu, Yingjie Liu, Yixuan Lv, Yujing Ma, Yuting Yang, Zhe Wang, Zhenyu Xu, Zhipeng Luo, Zhimin Zhang, Zhiguang Zhang, Zihao Li, and Zixiao Zhang. Visdrone-det2021: The vision meets drone object detection challenge results. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE, October 2021. doi: 10.1109/iccvw54120.2021.00319. URL http://dx.doi.org/10.1109/iccvw54120.2021.00319.

Santiago Castro, Ruoyao Wang, Pingxuan Huang, Ian Stewart, Oana Ignat, Nan Liu, Jonathan Stroud, and Rada Mihalcea. 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)*. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.209. URL http://dx.doi.org/10.18653/v1/2022.acl-long.209.

L Elisa Celis, Amit Deshpande, Tarun Kathuria, and Nisheeth K Vishnoi. How to be fair and diverse? *arXiv preprint arXiv:1610.07183*, 2016.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. 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*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.254. URL http://dx.doi.org/10.18653/v1/2021.naacl-main.254.

Joseph Check and Russell K Schutt. *Research methods in education*. Sage publications, 2011.

Yiqun Chen and James Zou. Twigma: A dataset of ai-generated images with metadata from twitter. *arXiv preprint arXiv:2306.08310*, 2023.

Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2018. doi: 10.1109/cvpr.2018.00108. URL http://dx.doi.org/10.1109/cvpr.2018.00108.

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

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016. doi: 10.1109/cvpr.2016.350. URL http://dx.doi.org/10.1109/cvpr.2016.350.

Ryan Culkin, J. Edward Hu, Elias Stengel-Eskin, Guanghui Qin, and Benjamin Van Durme. Iterative paraphrastic augmentation with discriminative span alignment. *Transactions of the Association for Computational Linguistics*, 9:494–509, 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00380. URL http://dx.doi.org/10.1162/tacl_a_00380.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018a.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 720–736, 2018b.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022.

Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. *TAO: A Large-Scale Benchmark for Tracking Any Object*, pp. 436–454. Springer International Publishing, 2020. ISBN 9783030585587. doi: 10.1007/978-3-030-58558-7_26. URL http://dx.doi.org/10.1007/978-3-030-58558-7_26.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pp. 25–35, 2019.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *ACM Conference on Fairness, Accountability, and Transparency FAccT*, pp. 120–128, 2019.

Christine De Kock and Andreas Vlachos. Leveraging wikipedia article evolution for promotional tone detection. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.384. URL http://dx.doi.org/10.18653/v1/2022.acl-long.384.

Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv preprint arXiv:2007.07399*, 2020.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. Broad twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1169–1179, 2016.

Carlos A Diaz, Youya Xia, Yurong You, Jose Nino, Junan Chen, Josephine Monica, Xiangyu Chen, Katie Z Luo, Yan Wang, Marc Emond, et al. Ithaca365: Dataset and driving perception under repeated and challenging weather conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

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

Carlos A. Diaz-Ruiz, Youya Xia, Yurong You, Jose Nino, Junan Chen, Josephine Monica, Xiangyu Chen, Katie Luo, Yan Wang, Marc Emond, Wei-Lun Chao, Bharath Hariharan, Kilian Q. Weinberger, and Mark Campbell. Ithaca365: Dataset and driving perception under repeated and challenging weather conditions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022. doi: 10.1109/cvpr52688.2022.02069. URL http://dx.doi.org/10.1109/cvpr52688.2022.02069.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67–73, 2018.

Kasper Drawzeski, Andrea Galassi, Agnieszka Jablonowska, Francesca Lagioia, Marco Lippi, Hans Wolfgang Micklitz, Giovanni Sartor, Giacomo Tagiuri, and Paolo Torroni. A corpus for multilingual analysis of online terms of service. In *Proceedings of the Natural Legal Language Processing Workshop 2021*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.nllp-1.1. URL http://dx.doi.org/10.18653/v1/2021.nllp-1.1.

Stefan Daniel Dumitrescu, Petru Rebeja, Beata Lorincz, Mihaela Gaman, Andrei Avram, Mihai Ilie, Andrei Pruteanu, Adriana Stan, Lorena Rosia, Cristina Iacobescu, et al. Liro: Benchmark and leaderboard for romanian language tasks. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS DB)*, 2021.

Esin Durmus, Faisal Ladhak, and Claire Cardie. The role of pragmatic and discourse context in determining argument impact. In *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/d19-1568. URL http://dx.doi.org/10.18653/v1/d19-1568.

Christian Ertler, Jerneja Mislej, Tobias Ollmann, Lorenzo Porzi, Gerhard Neuhold, and Yubin Kuang. *The Mapillary Traffic Sign Dataset for Detection and Classification on a Global Scale*, pp. 68–84. Springer International Publishing, 2020a. ISBN 9783030585921. doi: 10.1007/978-3-030-58592-1_5. URL http://dx.doi.org/10.1007/978-3-030-58592-1_5.

Christian Ertler, Jerneja Mislej, Tobias Ollmann, Lorenzo Porzi, Gerhard Neuhold, and Yubin Kuang. The mapillary traffic sign dataset for detection and classification on a global scale. In *European Conference on Computer Vision*. Springer, 2020b.

Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *IEEE/CVF International Conference on Computer Vision*, pp. 9710–9719, 2021.

Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. Convosumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. In *Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021a. doi: 10.18653/v1/2021.acl-long.535. URL http://dx.doi.org/10.18653/v1/2021.acl-long.535.

Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Orcun Cetintas, Riccardo Gasparini, Aljoša Ošep, Simone Calderara, Laura Leal-Taixé, and Rita Cucchiara. Motsynth: How can synthetic data help pedestrian detection and tracking? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10849–10859, 2021b.

Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019a. doi: 10.1109/cvpr.2019.00875. URL http://dx.doi.org/10.1109/cvpr.2019.00875.

Deng-Ping Fan, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Huazhu Fu, and Ming-Ming Cheng. Taking a deeper look at co-salient object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020. doi: 10.1109/cvpr42600.2020.00299. URL http://dx.doi.org/10.1109/cvpr42600.2020.00299.

Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. Inferring shared attention in social scene videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2018a. doi: 10.1109/cvpr.2018.00676. URL http://dx.doi.org/10.1109/cvpr.2018.00676.

Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. Inferring shared attention in social scene videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6460–6468, 2018b.

Lifeng Fan, Wenguan Wang, Song-Chun Zhu, Xinyu Tang, and Siyuan Huang. Understanding human gaze communication by spatio-temporal graph reasoning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2019b. doi: 10.1109/iccv.2019.0w0582. URL http://dx.doi.org/10.1109/iccv.2019.00582.

Ibrahim Abu Farha and Walid Magdy. From arabic sentiment analysis to sarcasm detection: The arsarcasm dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pp. 32–39, 2020.

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

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

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021a. doi: 10.1109/cvpr46437.2021.01451. URL http://dx.doi.org/10.1109/cvpr46437.2021.01451.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3d-front: 3d furnished rooms with layouts and semantics. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2021b. doi: 10.1109/iccv48922.2021.01075. URL http://dx.doi.org/10.1109/iccv48922.2021.01075.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Nabeel Gillani and Roger Levy. Simple dynamic word embeddings for mapping perceptions in the public sphere. In *Workshop on Natural Language Processing and Computational Social Science*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/w19-2111. URL http://dx.doi.org/10.18653/v1/w19-2111.

Omer Goldman and Reut Tsarfaty. Well-defined morphology is sentence-level morphology. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.mrl-1.23. URL http://dx.doi.org/10.18653/v1/2021.mrl-1.23.

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

Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2022.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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.

Jian Guan, Zhuoer Feng, Yamei Chen, Ruilin He, Xiaoxi Mao, Changjie Fan, and Minlie Huang. Lot: A story-centric benchmark for evaluating chinese long text understanding and generation. *Transactions of the Association for Computational Linguistics*, 10:434–451, 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00469. URL http://dx.doi.org/10.1162/tacl_a_00469.

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

Rishav Hada, Sohi Sudhir, Pushkar Mishra, Helen Yannakoudakis, Saif M. Mohammad, and Ekaterina Shutova. Ruddit: Norms of offensiveness for english reddit comments. In *Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.210. URL http://dx.doi.org/10.18653/v1/2021.acl-long.210.

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

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

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

Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. Towards measuring fairness in ai: the casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):324–332, 2021.

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

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

Mayur Hemani, Abhinav Patel, Tejas Shimpi, Anirudha Ramesh, and Balaji Krishnamurthy. What ails one-shot image segmentation: A data perspective. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS DB)*, 2021.

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label. *Data Protection and Privacy*, 12(12):1, 2020.

John J Howard, Yevgeniy B Sirotin, Jerry L Tipton, and Arun R Vemury. 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.

Quzhe Huang, Shibo Hao, Yuan Ye, Shengqi Zhu, Yansong Feng, and Dongyan Zhao. 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)*. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.432. URL http://dx.doi.org/10.18653/v1/2022.acl-long.432.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019. doi: 10.1109/cvpr.2019.00686. URL http://dx.doi.org/10.1109/cvpr.2019.00686.

Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. Masc: The manually annotated sub-corpus of american english. In *International Conference on Language Resources and Evaluation, LREC 2008*, pp. 2455–2460. European Language Resources Association (ELRA), 2008.

Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. *Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds*, pp. 544–559. Springer International Publishing, 2018. ISBN 9783030012168. doi: 10.1007/978-3-030-01216-8_33. URL http://dx.doi.org/10.1007/978-3-030-01216-8_33.

Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. 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*. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.602. URL http://dx.doi.org/10.18653/v1/2020.acl-main.602.

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

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

Abigail Z Jacobs, Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. The meaning and measurement of bias: lessons from natural language processing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 706–706, 2020.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017a.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017b. doi: 10.1109/cvpr.2017.215. URL http://dx.doi.org/10.1109/cvpr.2017.215.

Kristen Johnson and Dan Goldwasser. Classification of moral foundations in microblog political discourse. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018. doi: 10.18653/v1/p18-1067. URL http://dx.doi.org/10.18653/v1/p18-1067.

Laurynas Karazija, Iro Laina, and Christian Rupprecht. Clevrtex: A texture-rich benchmark for unsupervised multi-object segmentation. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS DB)*, 2021.

Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1548–1558, 2021.

Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS DB)*, 2021.

Zaid Khan and Yun Fu. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. Abstractive summarization of reddit posts with multi-level memory networks. In *Proceedings of NAACL-HLT*, pp. 2519–2531, 2019.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.356. URL http://dx.doi.org/10.18653/v1/2020.emnlp-main.356.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

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

Rémi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020a. doi: 10.18653/v1/2020.emnlp-main.706. URL http://dx.doi.org/10.18653/v1/2020.emnlp-main.706.

Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. *TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval*, pp. 447–463. Springer International Publishing, 2020b. ISBN 9783030585891. doi: 10.1007/978-3-030-58589-1_27. URL http://dx.doi.org/10.1007/978-3-030-58589-1_27.

Jie Lei, Tamara Berg, and Mohit Bansal. mtvr: Multilingual moment retrieval in videos. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, 2021a. doi: 10.18653/v1/2021.acl-short.92. URL http://dx.doi.org/10.18653/v1/2021.acl-short.92.

Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021b.

Haley Lepp and Gina-Anne Levow. Pardon the interruption: An analysis of gender and turn-taking in u.s. supreme court oral arguments. In *Interspeech 2020*, interspeech$_2$020.$ISCA$, $October$2020.doi : . URL http://dx.doi.org/10.21437/interspeech.2020-2964.

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

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning. In *Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021. 10.18653/v1/2021.acl-long.102. URL http://dx.doi.org/10.18653/v1/2021.acl-long.102.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022. 10.1109/cvpr52688.2022.02034. URL http://dx.doi.org/10.1109/cvpr52688.2022.02034.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS DB)*, 2021.

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

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019. 10.1109/cvpr.2019.00331. URL http://dx.doi.org/10.1109/cvpr.2019.00331.

Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon ReiB, Michael Voit, and Rainer Stiefelhagen. Driveamp;act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2019. 10.1109/iccv.2019.00289. URL http://dx.doi.org/10.1109/iccv.2019.00289.

Tetiana Martyniuk, Orest Kupyn, Yana Kurlyak, Igor Krashenyi, Jiri Matas, and Viktoriia Sharmanska. 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)*. IEEE, June 2022. 10.1109/cvpr52688.2022.02027. URL http://dx.doi.org/10.1109/cvpr52688.2022.02027.

David Mayo, Jesse Cummings, Xinyu Lin, Dan Gutfreund, Boris Katz, and Andrei Barbu. How hard are computer vision datasets? calibrating dataset difficulty to viewing time. 2022.

Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021. 10.1109/cvpr46437.2021.00412. URL http://dx.doi.org/10.1109/cvpr46437.2021.00412.

Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022a. 10.1109/cvpr52688.2022.02036. URL http://dx.doi.org/10.1109/cvpr52688.2022.02036.

Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b.

Milagros Miceli, Julian Posada, and Tianling Yang. 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.

Milagros Miceli, Tianling Yang, Adriana Alvarado Garcia, Julian Posada, Sonja Mei Wang, Marc Pohl, and Alex Hanna. Documenting data production processes. *PACM HCI*, 2022b.

Margaret Mitchell, Alexandra Sasha Luccioni, Nathan Lambert, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, Tristan Thrush, Yacine Jernite, and Douwe Kiela. Measuring data. *arXiv preprint arXiv:2212.05129*, 2022.

Youssef Mohamed, Faizan Farooq Khan, Kilichbek Haydarov, and Mohamed Elhoseiny. It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022. 10.1109/cvpr52688.2022.02058. URL http://dx.doi.org/10.1109/cvpr52688.2022.02058.

T. Nathan Mundhenk, Goran Konjevod, Wesam A. Sakla, and Kofi Boakye. *A Large Contextual Dataset for Classification, Detection and Counting of Cars with Deep Learning*, pp. 785–800. Springer International Publishing, 2016. ISBN 9783319464879. 10.1007/978-3-319-46487-$9_48.URL$.

Matthias Müller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. *TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild*, pp. 310–327. Springer International Publishing, 2018. ISBN 9783030012465. 10.1007/978-3-030-01246-$5_19.URL$.

Keziah Naggita, Julienne LaChance, and Alice Xiang. Flickr africa: Examining geo-diversity in large-scale, human-centric visual data. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 520–530, 2023.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019. 10.18653/v1/d19-1018. URL http://dx.doi.org/10.18653/v1/d19-1018.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pp. 145–153, 2016.

Matan Orbach, Orith Toledo-Ronen, Artem Spector, Ranit Aharonov, Yoav Katz, and Noam Slonim. Yaso: A targeted sentiment analysis evaluation dataset for open-domain reviews. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021. 10.18653/v1/2021.emnlp-main.721. URL http://dx.doi.org/10.18653/v1/2021.emnlp-main.721.

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

James O'Neill, Polina Rozenshtein, Ryuichi Kiryo, Motoko Kubota, and Danushka Bollegala. I wish i would have loved this one, but i didn't – a multilingual dataset for counterfactual detection in product review. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021. 10.18653/v1/2021.emnlp-main.568. URL http://dx.doi.org/10.18653/v1/2021.emnlp-main.568.

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

Kenneth L Peng, Arunesh Mathur, and Arvind Narayanan. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. In *Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS DB)*, 2021.

Denis Peskov, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang, Adel Youssef, and Mona Diab. 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 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019. 10.18653/v1/d19-1460. URL http://dx.doi.org/10.18653/v1/d19-1460.

Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1756–1765, 2017.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. *Connecting Vision and Language with Localized Narratives*, pp. 647–664. Springer International Publishing, 2020. ISBN 9783030585587. 10.1007/978-3-030-58558-$7_38.URL$.

Adithya Pratapa, Rishubh Gupta, and Teruko Mitamura. Multilingual event linking to wikidata. In *Proceedings of the Workshop on Multilingual Information Access (MIA)*. Association for Computational Linguistics, 2022. 10.18653/v1/2022.mia-1.5. URL http://dx.doi.org/10.18653/v1/2022.mia-1.5.

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

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Md Mustafizur Rahman, Dinesh Balakrishnan, Dhiraj Murthy, Mucahid Kutlu, and Matthew Lease. An information retrieval approach to building datasets for hate speech detection. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS DB)*, 2021.

Inioluwa Deborah Raji and Genevieve Fried. About face: A survey of facial recognition evaluation. *arXiv preprint arXiv:2102.00813*, 2021.

Inioluwa Deborah Raji, Emily Denton, Emily M Bender, Alex Hanna, and Amandalynne Paullada. Ai and the everything in the whole wide world benchmark. In *Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS DB)*, 2021.

Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron B Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Beyond web-scraping: Crowd-sourcing a geographically diverse image dataset. *arXiv preprint arXiv:2301.02560*, 2023.

William A Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. 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)*. IEEE, June 2016. 10.1109/cvpr.2016.352. URL http://dx.doi.org/10.1109/cvpr.2016.352.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. Solid: A large-scale semi-supervised dataset for offensive language identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 915–928, 2021.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 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)*. Association for Computational Linguistics, 2018. 10.18653/v1/n18-2002. URL http://dx.doi.org/10.18653/v1/n18-2002.

Pratik S Sachdeva, Renata Barreto, Claudia von Vacano, and Chris J Kennedy. Assessing annotator identity sensitivity via item response theory: A case study in a hate speech corpus. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1585–1603, 2022.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling for video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021. 10.1109/cvpr46437.2021.00554. URL http://dx.doi.org/10.1109/cvpr46437.2021.00554.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "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.

Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 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.

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

Thomas Schops, Johannes L. Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017. 10.1109/cvpr.2017.272. URL http://dx.doi.org/10.1109/cvpr.2017.272.

Claudia Schulz, Christian M. Meyer, Jan Kiesewetter, Michael Sailer, Elisabeth Bauer, Martin R. Fischer, Frank Fischer, and Iryna Gurevych. Analysis of automatic annotation suggestions for hard discourse-level tasks in expert domains. In *Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019. 10.18653/v1/p19-1265. URL http://dx.doi.org/10.18653/v1/p19-1265.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

Rishi Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh. Tackling the story ending biases in the story cloze test. In *Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2018. 10.18653/v1/p18-2119. URL http://dx.doi.org/10.18653/v1/p18-2119.

Zejiang Shen, Kyle Lo, Lucy Lu Wang, Bailey Kuehl, Daniel S. Weld, and Doug Downey. 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. 10.1162/tacl$_{a0}$0466.URL.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. Beyond fair pay: Ethical implications of nlp crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3758–3769, 2021.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. *Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding*, pp. 510–526. Springer International Publishing, 2016. ISBN 9783319464480. 10.1007/978-3-319-46448-0$_3$1.URL.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. *Indoor Segmentation and Support Inference from RGBD Images*, pp. 746–760. Springer Berlin Heidelberg, 2012. ISBN 9783642337154. 10.1007/978-3-642-33715-4$_5$4.URL.

Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Mattia Soldan, Alejandro Pardo, Juan Leon Alcazar, Fabian Caba Heilbron, Chen Zhao, Silvio Giancola, and Bernard Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022. 10.1109/cvpr52688.2022.00497. URL http://dx.doi.org/10.1109/cvpr52688.2022.00497.

Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015. 10.1109/cvpr.2015.7298655. URL http://dx.doi.org/10.1109/cvpr.2015.7298655.

Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21. ACM, March 2021. 10.1145/3442188.3445932. URL http://dx.doi.org/10.1145/3442188.3445932.

Arjun Subramonian, Xingdi Yuan, Hal Daumé III, and Su Lin Blodgett. It takes two to tango: Navigating conceptualizations of nlp tasks and measurements of performance. *arXiv preprint arXiv:2305.09022*, 2023.

Saku Sugawara, Nikita Nangia, Alex Warstadt, and Samuel Bowman. What makes reading comprehension questions difficult? In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022. 10.18653/v1/2022.acl-long.479. URL http://dx.doi.org/10.18653/v1/2022.acl-long.479.

Kai Sun, Seungwhan Moon, Paul Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. Adding chit-chat to enhance task-oriented dialogues. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2021. 10.18653/v1/2021.naacl-main.124. URL http://dx.doi.org/10.18653/v1/2021.naacl-main.124.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020. 10.1109/cvpr42600.2020.00252. URL http://dx.doi.org/10.1109/cvpr42600.2020.00252.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Quin Thames, Arjun Karpur, Wade Norris, Fangting Xia, Liviu Panait, Tobias Weyand, and Jack Sim. Nutrition5k: Towards automatic nutritional understanding of generic food. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021. 10.1109/cvpr46437.2021.00879. URL http://dx.doi.org/10.1109/cvpr46437.2021.00879.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Gabriel Tseng, Ivan Zvonkov, Catherine Lilian Nakalembe, and Hannah Kerner. Cropharvest: A global dataset for crop-type classification. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS DB)*, 2021.

Francisco Rivera Valverde, Juana Valeria Hurtado, and Abhinav Valada. 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)*. IEEE, June 2021. 10.1109/cvpr46437.2021.01144. URL http://dx.doi.org/10.1109/cvpr46437.2021.01144.

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin MacAodha. Benchmarking representation learning for natural world image collections. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021. 10.1109/cvpr46437.2021.01269. URL http://dx.doi.org/10.1109/cvpr46437.2021.01269.

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

Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C. Kot. Crrn: Multi-scale guided concurrent reflection removal network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2018. 10.1109/cvpr.2018.00502. URL http://dx.doi.org/10.1109/cvpr.2018.00502.

Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision (IJCV)*, 130(7):1790–1810, 2022.

Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *IEEE/CVF International*

Conference on Computer Vision (ICCV). IEEE, October 2019a. 10.1109/iccv.2019.00078. URL http://dx.doi.org/10.1109/iccv.2019.00078.

Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019b. 10.1109/cvpr.2019.00839. URL http://dx.doi.org/10.1109/cvpr.2019.00839.

Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019c. 10.1109/cvpr.2019.00701. URL http://dx.doi.org/10.1109/cvpr.2019.00701.

William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pp. 19–26, 2012.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*, 2017.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617, December 2018. ISSN 2307-387X. 10.1162/tacl$_{a}0$0240.URL.

Zijun Wei, Jianming Zhang, Zhe Lin, Joon-Young Lee, Niranjan Balasubramanian, Minh Hoai, and Dimitris Samaras. Learning visual emotion representations from web data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13106–13115, 2020.

Michael Wiegand, Maja Geulig, and Josef Ruppenhofer. Implicitly abusive comparisons – a new dataset and linguistic analysis. In *Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 2021. 10.18653/v1/2021.eacl-main.27. URL http://dx.doi.org/10.18653/v1/2021.eacl-main.27.

Adrienne Williams, Milagros Miceli, and Timnit Gebru. The exploited labor behind artificial intelligence. *Noema Magazine*, 13, 2022.

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

Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3681–3691, 2021.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. Proactive human-machine conversation with explicit conversation goal. In *Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019. 10.18653/v1/p19-1369. URL http://dx.doi.org/10.18653/v1/p19-1369.

Ziang Xiao, Susu Zhang, Vivian Lai, and Q Vera Liao. Evaluating evaluation metrics: A framework for analyzing nlg evaluation metrics using measurement theory. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10967–10982, 2023.

Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. *Segmenting Transparent Objects in the Wild*, pp. 696–711. Springer International Publishing, 2020. ISBN 9783030586010. 10.1007/978-3-030-58601-0$_{4}$1.URL.

Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2949–2958, 2022a.

Song Xu, Lucas Böttcher, and Tom Chou. Diversity in biology: definitions, quantification and models. *Physical Biology*, 17(3):031001, 2020.

Xixi Xu, Zhongang Qi, Jianqi Ma, Honglun Zhang, Ying Shan, and Xiaohu Qie. Bts: A bi-lingual benchmark for text segmentation in the wild. In *IEEE/CVF Conference on Computer Vision and*

Pattern Recognition (CVPR). IEEE, June 2022b. 10.1109/cvpr52688.2022.01856. URL http://dx.doi.org/10.1109/cvpr52688.2022.01856.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 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)*. Association for Computational Linguistics, 2022c. 10.18653/v1/2022.acl-long.34. URL http://dx.doi.org/10.18653/v1/2022.acl-long.34.

Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, Alexander William Clegg, and Devendra Singh Chaplot. Habitat-matterport 3d semantics dataset. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023. 10.1109/cvpr52729.2023.00477. URL http://dx.doi.org/10.1109/cvpr52729.2023.00477.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2021. 10.1109/iccv48922.2021.00171. URL http://dx.doi.org/10.1109/iccv48922.2021.00171.

Yu Yang, Aayush Gupta, Jianwei Feng, Prateek Singhal, Vivek Yadav, Yue Wu, Pradeep Natarajan, Varsha Hedau, and Jungseock Joo. Enhancing fairness in face detection in computer vision systems by demographic bias mitigation. In *AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22. ACM, July 2022. 10.1145/3514094.3534153. URL http://dx.doi.org/10.1145/3514094.3534153.

Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei W Koh, and Chelsea Finn. Wild-time: A benchmark of in-the-wild distribution shift over time. *Advances in Neural Information Processing Systems*, 35:10309–10324, 2022.

Yang You, Yujing Lou, Chengkun Li, Zhoujun Cheng, Liangwei Li, Lizhuang Ma, Cewu Lu, and Weiming Wang. Keypointnet: A large-scale 3d keypoint dataset aggregated from numerous human annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13647–13656, 2020.

Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS DB)*, 2021.

Ann Yuan, Daphne Ippolito, Vitaly Nikolaev, Chris Callison-Burch, Andy Coenen, and Sebastian Gehrmann. Synthbio: A case study in faster curation of text datasets. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS DB)*, 2021.

Xingdi Yuan, Tong Wang, Rui Meng, Khushboo Thaker, Peter Brusilovsky, Daqing He, and Adam Trischler. One size does not fit all: Generating and evaluating variable number of keyphrases. In *Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. 10.18653/v1/2020.acl-main.710. URL http://dx.doi.org/10.18653/v1/2020.acl-main.710.

Philine Zeinert, Nanna Inie, and Leon Derczynski. Annotating online misogyny. In *Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021a. 10.18653/v1/2021.acl-long.247. URL http://dx.doi.org/10.18653/v1/2021.acl-long.247.

Philine Zeinert, Nanna Inie, and Leon Derczynski. Annotating online misogyny. In *Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3181–3197, 2021b.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019. 10.1109/cvpr.2019.00688. URL http://dx.doi.org/10.1109/cvpr.2019.00688.

Oliver Zendel, Matthias Schorghuber, Bernhard Rainer, Markus Murschitz, and Csaba Beleznai. Unifying panoptic segmentation for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022. 10.1109/cvpr52688.2022.02066. URL http://dx.doi.org/10.1109/cvpr52688.2022.02066.

Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020. 10.1109/cvpr42600.2020.00281. URL http://dx.doi.org/10.1109/cvpr42600.2020.00281.

Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022. 10.1109/cvpr52688.2022.00868. URL http://dx.doi.org/10.1109/cvpr52688.2022.00868.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. Mr. tydi: A multi-lingual benchmark for dense retrieval. In *Workshop on Multilingual Representation Learning*. Association for Computational Linguistics, 2021. 10.18653/v1/2021.mrl-1.12. URL http://dx.doi.org/10.18653/v1/2021.mrl-1.12.

Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2021a. 10.1109/iccv48922.2021.01456. URL http://dx.doi.org/10.1109/iccv48922.2021.01456.

Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14830–14840, 2021b.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 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)*. Association for Computational Linguistics, 2018. 10.18653/v1/n18-2003. URL http://dx.doi.org/10.18653/v1/n18-2003.

Yang Zhong, Jingfeng Yang, Wei Xu, and Diyi Yang. Wikibias: Detecting multi-span subjective biases in language. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, 2021. 10.18653/v1/2021.findings-emnlp.155. URL http://dx.doi.org/10.18653/v1/2021.findings-emnlp.155.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. Deconstructing nlg evaluation: Evaluation practices, assumptions, and their implications. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 314–324, 2022.

Zhuotun Zhu, Lingxi Xie, and Alan L Yuille. Object recognition with and without objects. *arXiv preprint arXiv:1611.06596*, 2016.

Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max J. Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2019. 10.1109/iccv.2019.00090. URL http://dx.doi.org/10.1109/iccv.2019.00090.

Changqing Zou, Qian Yu, Ruofei Du, Haoran Mo, Yi-Zhe Song, Tao Xiang, Chengying Gao, Baoquan Chen, and Hao Zhang. *SketchyScene: Richly-Annotated Scene Sketches*, pp. 438–454. Springer International Publishing, 2018. ISBN 9783030012670. 10.1007/978-3-030-01267-$0_2$6.URL.

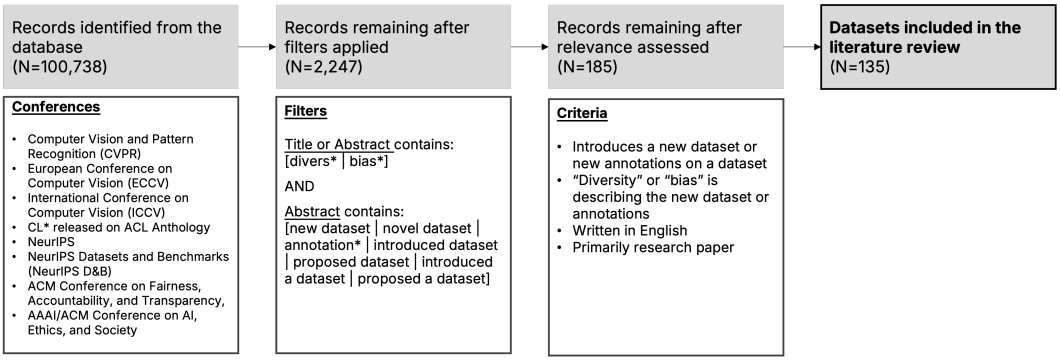

Figure 1: Overview of the search strategy employed to identify diverse datasets to include in our corpus.

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 survey a total of 135 image and text datasets. To define this corpus of datasets, we follow the process shown in Figure 1. First, we sample all papers from computer vision (CVPR, ECCV, ICCV), natural language processing (*CL released on ACL Anthology), machine learning (NeurIPS, including the Datasets and Benchmarks track), and fairness (FAccT, AIES) venues. From these papers, we used regular expression matching to find those that contained a keyword related to diversity or bias in either the title or abstract. We further sub-selected those that had phrases in the abstract indicating that the paper presented a new dataset or annotations. Next, we manually filtered the remaining papers to ensure that the descriptor of "diversity" or "bias" was related to the new dataset or annotations, the article was written in English, and that the article was primarily a research paper. Finally, from this pool, we sub-sampled 135 datasets to review.

A.2 CODING PROCEDURE

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 et al., 2021; Peng et al., 2021; Google PAIR, 2019; Gebru et al., 2021; Holland et al., 2020; Bender

Paper	Definition of Diversity
Ithaca365 (Diaz-Ruiz 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 2: Example quotes from select surveyed datasets where diversity is concretely defined.

& 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.

To code the 135 datasets, three of the co-authors separately coded an initial set of three datasets. After coding the initial set, the authors synchronously discussed when their codes for the datasets diverged and iterating on codebook definitions. The remaining 132 datasets were then split between the three co-authors and individually coded. To identify the higher-level themes, the authors first individually generated themes and then synchronously met to group the themes into larger categories.

A.3 Categorization

Defining standalone dataset. We count a paper as presenting a standalone dataset if the core contributions of the work are the novel dataset, benchmark, or set of annotations. For example, if a paper presents a novel model, task, or metric, we would *not* count the paper as presenting a standalone dataset.

Concrete definition. We used the following criteria in determining whether there was a concrete definition of diversity in the paper. We check whether the authors explicitly state a consistent and complete definition of diversity. For example, a dataset that is described as "diverse" in multiple ways across the paper would not be considered to have a concrete definition. In addition, we ensure that the authors list out what aspects are more diverse (e.g., the dataset presents different illumination conditions; the dataset draws from different news sources). In Table 2, we provide examples of concrete diversity definitions we identified.

B 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 the 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 validation 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 name 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 a NER model for inferring the country of origin introduces the potential for errors due to social bias or ambiguity (e.g., "Georgia" being both a state in the US 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 the increasing opacity in the collection process (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 or location are undisclosed. Further, it remains unclear whether diversity is a result of explicit instructions and deliberate sampling or an 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 image, showcasing convergent validity.

A recommended enhancement involves providing validation for geographic diversity. While the authors currently 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., 2023). 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.

C Additional details on dataset collection

In Sec. 5, we identify five main collection methodologies used to create machine learning datasets. We provide further detail on this taxonomy in Table 3 and enumerate the datasets in our corpus that utilize the respective methodologies in Table 4.

D Dataset details

We provide additional details on the 135 image and text datasets that were coded as a part of our literature review. In Table 5, we enumerate the diversity type for each dataset. Finally, in Tables 6 and 8, we provide a summary of our analysis across all datasets.

E DISCUSSION

Here, we delve into additional considerations regarding two key points: the inherent tensions between measurement and scale, and the documentation burden imposed on dataset creators.

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 results in increased diversity in terms of count, background regions, scenarios, etc." in their dataset JHU-Crowd. 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 or organizations (Heger et al., 2022). This burden is further exacerbated by

Collection Method	Definition	Example
Web-scraping (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 multilingual 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., 2018b) 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 ClevrTex using a selection of photo-mapped materials, objects, and backgrounds that are procedurally generated with the Blender API.

Table 3: 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.

	Papers
Web-scraping	Fan et al. (2018a); Zellers et al. (2019); Wei et al. (2020); Xu et al. (2022a); Van Horn et al. (2021); Aly et al. (2021); Shen et al. (2022); Abdul-Mageed et al. (2021); Zeinert et al. (2021a); Cao & Daumé III (2020); Yuan et al. (2020); Gillani & Levy (2019); Wu et al. (2019); Durmus et al. (2019); Derczynski et al. (2016); Miao et al. (2021); Sindagi et al. (2019); Asano et al. (2021); Dumitrescu et al. (2021); Lu et al. (2021); Pratapa et al. (2022); Xu et al. (2022b); Celis et al. (2016); Hada et al. (2021); Bagher Zadeh et al. (2020); Lei et al. (2020a); Asaadi et al. (2019); Kim et al. (2019); Soldan et al. (2022); Fu et al. (2021a); Ertler et al. (2020a); Müller et al. (2018); Idrees et al. (2018); Zou et al. (2018); Fan et al. (2019b); Lei et al. (2021b); Rahman et al. (2021); Guan et al. (2022); De Kock & Vlachos (2022); Sugawara et al. (2022); Zhang et al. (2021); Zhong et al. (2021); Buolamwini & Gebru (2018); Steed & Caliskan (2021); Webster et al. (2018); De-Arteaga et al. (2019)
Crowdsourcing	Damen et al. (2018a); Barbu et al. (2019); Peskov et al. (2019); Sharma et al. (2018); Sigurdsson et al. (2016); Wiegand et al. (2021); Byrne et al. (2021; 2019); Chawla et al. (2021)
Direct Collection	Schops et al. (2017); Cordts et al. (2016); Song et al. (2015); Wan et al. (2018); Wang et al. (2019c); Sun et al. (2020); Zhang et al. (2020); Liu et al. (2022); Bai et al. (2021); Cao et al. (2021); Wang et al. (2019c); Martin et al. (2019); Zendel et al. (2022); Zhang et al. (2022); Thames et al. (2021); Yadav et al. (2023); Diaz-Ruiz et al. (2022); Valverde et al. (2021); Zimmermann et al. (2019); An et al. (2021)
Derivative	Song et al. (2015); Chuang et al. (2018); Marino et al. (2019); Hudson & Manning (2019); Zellers et al. (2019); Fan et al. (2019a; 2020); You et al. (2020); Mohamed et al. (2022); Sadhu et al. (2021); Dave et al. (2020); Yang et al. (2021); Tseng et al. (2021); Castro et al. (2022); Huang et al. (2022); Sun et al. (2021); Orbach et al. (2021); Lin et al. (2021); Fabbri et al. (2021a); Ku et al. (2020); Ni et al. (2019); Johnson & Goldwasser (2018); Angelidis & Lapata (2018); Ide et al. (2008); Yang et al. (2022); Lei et al. (2020b); Asano et al. (2021); Dumitrescu et al. (2021); Yu et al. (2021); Drawzeski et al. (2021); Goldman & Tsarfaty (2021); O'Neill et al. (2021); Lei et al. (2021a); Farha & Magdy (2020); Inoue et al. (2020); Lepp & Levow (2020); Schulz et al. (2019); Soldan et al. (2022); Miao et al. (2022a); Müller et al. (2018); Pont-Tuset et al. (2020); Zhao et al. (2021a); Hemani et al. (2021); Culkin et al. (2021); Wang et al. (2019a)
Synthetic Generation	Johnson et al. (2017b); Ros et al. (2016); Wang et al. (2019b); Khalid et al. (2021); Fabbri et al. (2021b); Karazija et al. (2021); Zou et al. (2018); Fu et al. (2021b); Yuan et al. (2021); Guan et al. (2022); Culkin et al. (2021); Zhao et al. (2018); Rudinger et al. (2018)

Table 4: Collection methodology used in the datasets include **web-scraping** – downloading instances available on the Internet; **crowdsourcing**– asking a group of people that are not directly members of the research team to collect instances; **direct collection** – members of the research team collect the data instances themselves; **derivative** – uses one or multiple existing datasets; and **synthetic** – instances are artificially manufactured or procedurally generated rather than capturing real-world events. Each dataset can involve multiple collection methodologies.

	Papers
Composition Diversity	Schops et al. (2017); Cordts et al. (2016); Ros et al. (2016); Chuang et al. (2018); Wan et al. (2018); Fan et al. (2018a); Marino et al. (2019); Wang et al. (2019c;c); Fan et al. (2019a; 2020); Zhang et al. (2020); Liu et al. (2022); Bai et al. (2021); Dave et al. (2020); Cao et al. (2021); Wang et al. (2019c); Barbu et al. (2019); Lin et al. (2021); Ku et al. (2020); Derczynski et al. (2016); Zendel et al. (2022); Martyniuk et al. (2022); Thames et al. (2021); Miao et al. (2021); Sigurdsson et al. (2016); Xie et al. (2020); Silberman et al. (2012); Ettinger et al. (2021); Fabbri et al. (2021b); Sindagi et al. (2019); Yadav et al. (2023); Yu et al. (2021); Karazija et al. (2021); Drawzeski et al. (2021); Goldman & Tsarfaty (2021); Bagher Zadeh et al. (2020); Byrne et al. (2019); Soldan et al. (2022); Xu et al. (2022b); Miao et al. (2022a); Diaz-Ruiz et al. (2022); Fu et al. (2021a); Valverde et al. (2021); Ertler et al. (2020a); Müller et al. (2018); Pont-Tuset et al. (2020); Idrees et al. (2018); Mundhenk et al. (2016); Fu et al. (2021b); Zimmermann et al. (2019); Fan et al. (2019b); Lei et al. (2021b); An et al. (2021); Culkin et al. (2021); Chawla et al. (2021); Zhang et al. (2021)
Source Diversity	Fan et al. (2018a); Sun et al. (2020); Van Horn et al. (2021); Damen et al. (2018a); Martin et al. (2019); Tseng et al. (2021); Shen et al. (2022); 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); Zendel et al. (2022); Zhang et al. (2022); Rojas et al. (2022); Asano et al. (2021); Diaz-Ruiz et al. (2022); Valverde et al. (2021); Ertler et al. (2020a); Mundhenk et al. (2016); Fan et al. (2019b); Lei et al. (2021b); Rahman et al. (2021); Sugawara et al. (2022)
Domain Diversity	Marino et al. (2019); Wei et al. (2020); Xu et al. (2022a); Sadhu et al. (2021); Dave et al. (2020); Damen et al. (2018a); Martin et al. (2019); Angelidis & Lapata (2018); Miao et al. (2021); Lei et al. (2020b); Ettinger et al. (2021); Hudson & Manning (2019); Yang et al. (2021); Lu et al. (2021); Lei et al. (2020a); Soldan et al. (2022); Fu et al. (2021a); 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	Zeinert et al. (2021a); Lei et al. (2021a)
Reduce Dataset Bias	Johnson et al. (2017b); Zellers et al. (2019); Mohamed et al. (2022); Yang et al. (2021); Aly et al. (2021); Castro et al. (2022); Huang et al. (2022); Sharma et al. (2018); Fan et al. (2018a); Wiegand et al. (2021); Farha & Magdy (2020); Inoue et al. (2020); Asaadi et al. (2019); Kim et al. (2019); Hemani et al. (2021); Guan et al. (2022); De Kock & Vlachos (2022); Zhong et al. (2021)
Promote Diversity (or Fairness) in Downstream Applications	Van Horn et al. (2021); Sun et al. (2021); Yuan et al. (2020); Gillani & Levy (2019); Wu et al. (2019); Ni et al. (2019); Johnson & Goldwasser (2018); Dumitrescu et al. (2021); Xu et al. (2022c); Celis et al. (2016); Lepp & Levow (2020); Zhao et al. (2021a); Steed & Caliskan (2021); De-Arteaga et al. (2019)
Not specified	Song et al. (2015); You et al. (2020); Pratapa et al. (2022); Schulz et al. (2019)

Table 5: Categories of diversity that the surveyed datasets include, as well as other identified categories such as reducing dataset bias or promoting diversity (or fairness) in downstream applications. Each dataset can contain multiple categories of diversity.

the necessity to address existing "documentation debt" (Bandy & Vincent, 2021), referring to commonly used datasets with inadequate or no documentation. Overcoming these challenges demands systemic changes in how data work is perceived within the ML community (Sambasivan et al.,

Dataset	Standalone	Concrete defn.	Justification	Collection trade-offs	Quality info	Annot. info
3D Furnished Rooms with layOuts and semaNTics (3D-FRONT) (Fu et al., 2021a)						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., 2017b)	✓	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., 2018a)	✓	✓		✓	✓	
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-Ruiz et al., 2022)	✓	✓	✓			
JHU-Crowd (Sindagi et al., 2019)		✓				NA
KeypointNet (You et al., 2020)						
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., 2020a)	✓	✓	✓		✓	
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)		✓			✓	✓

Table 6: **Condensed key for Table 6**: *Standalone* – if the paper presents only a dataset (i.e., no novel models, metrics, or tasks); *Concrete defn* – if the paper includes a concrete definition of diversity (NA for datasets that focused on bias, not diversity); *Justification* – if the paper discusses why diversity is needed in the dataset (NA for datasets that focused on bias, not diversity); *Collection trade-offs* – if the paper discusses downsides to the selected collection methodology or lists considerations for other potential collection methodologies; *Quality info* – if the paper includes information about how dataset quality is validated (e.g., manually inspected by dataset creators, use of inter-annotator agreement scores); *Annot. info* – if additional information about annotators are supplied (e.g., what training they went through, what the qualification criteria were).

Dataset	Standalone	Concret defn.	Justification	Collection trade-offs	Quality info	Annot. info
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., 2018a)	✓	✓	✓			
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)		✓	✓			
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., 2018a)		✓	✓			
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)	✓		✓	✓	✓	

Table 7: Continuation of surveyed papers with the same key as Table 6

2021). We highlight ongoing initiatives in this realm, such as the establishment of academic venues (e.g., the Journal of Data-centric Machine Learning Research, NeurIPS Datasets and Benchmarks), as a pivotal initial step in addressing this pervasive problem.

Constructs changing over time. Our framework does not safeguard against potential changes in *temporal shifts* to constructs—for example, discrepancies between train and time time data that may arise due to the passage of time (Yao et al., 2022; Bergman et al., 2023). In Section 6.2, we acknowledge this when discussing how the reliability of web-scraped data can be impacted by changes to

Dataset	Standalone	Concrete defn.	Justification	Collection trade-offs	Quality info	Annot. info
Culkin et al. (2021)					✓	
Drawzeski et al. (2021)			✓	✓		
Durmus et al. (2019)		✓	✓			
Fu et al. (2021a)		✓	✓			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. (2021a)		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: Continuation of surveyed papers with the same key as Table 6

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).