

The Casual Conversations v2 Dataset

A diverse, large benchmark for measuring fairness and robustness in audio/vision/speech models

Bilal Porgali

porgali@meta.com

Vítor Albiero

valbiero@meta.com

Jordan Ryda

jryda@meta.com

Cristian Canton Ferrer

ccanton@meta.com

Caner Hazirbas

hazirbas@meta.com

Meta AI

Abstract

This paper introduces a new large consent-driven dataset aimed at assisting in the evaluation of algorithmic bias and robustness of computer vision and audio speech models in regards to 11 attributes that are self-provided or labeled by trained annotators. The dataset includes 26,467 videos of 5,567 unique paid participants, with an average of almost 5 videos per person, recorded in Brazil, India, Indonesia, Mexico, Vietnam, Philippines, and the USA, representing diverse demographic characteristics. The participants agreed for their data to be used in assessing fairness of AI models and provided self-reported age, gender, language/dialect, disability status, physical adornments, physical attributes and geo-location information, while trained annotators labeled apparent skin tone using the Fitzpatrick Skin Type and Monk Skin Tone scales, and voice timbre. Annotators also labeled for different recording setups and per-second activity annotations.

1. Introduction

Ethical considerations on dataset construction [4, 15] has become more significant in AI shortly after several number of studies carried out by researchers to identify fairness concerns of biometrics and facial processing technologies were published [2, 5, 7, 10, 13, 18, 21], and discussions of data protection and related regulation have continued to evolve [17, 30].

One of the biggest challenges of identifying fairness issues has been a lack of clean, and most importantly responsibly-constructed, benchmarks. Casual Conversations [16], Dollar Street [27], Open Images MIAP [29], FairFace [19], UTK Faces [33], RFW [31], and MORPH [26], are some of the most widely used datasets to



Figure 1. *The Casual Conversations v2* dataset includes a total of 11 attributes that are self-provided or annotated.

identify fairness gaps [8, 12]. However, each of these has their own limitations. Casual Conversations (CCv1) is not geographically diverse (U.S. only), Dollar Street does not have person attributes such as age & gender, Open Images MIAP uses perceived gender instead of self-provided, FairFace, UTK Faces, RFW are made up of images collected from the internet. MORPH has binary gender with limited number of attributes that are age, gender, race, height, weight, and eye coordinates. Moreover, except Casual Conversations, all other aforementioned datasets are designed to measure only computer vision models.

To make progress against aforementioned limitations, we propose a large, diverse and consent-driven audio/vision/speech dataset with many attributes, *i.e.* *Casual Conversations v2*¹. Our dataset is composed of 26,467 videos of

¹<https://ai.facebook.com/datasets/casual-conversations-v2-dataset>

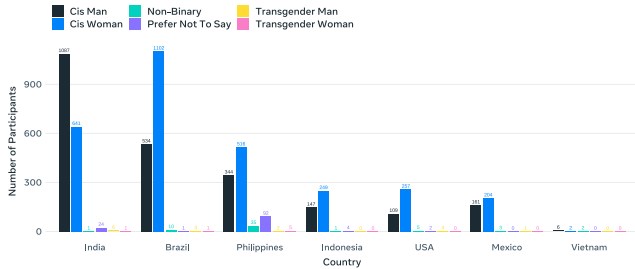


Figure 2. **Gender** distribution of participants by country. India contains a considerably higher participant ratio of cis men to cis women compared to the other countries.

5,567 unique paid participants recorded in seven countries and include 11 self-provided and annotated attributes. Participants directly provided us their data to be used in AI and provided self-identified *age*, *gender*, *language/dialect*, *disability*, *physical adornments & attributes* and *geo-location* information. In addition, trained annotators labeled participants’ *apparent skin tones & voice timbre* and annotated videos for *different recording setup* and per-second *activity* on videos. Subcategories of these attributes were selected and defined based on the literature survey that is carried out by Hazirbas *et al.* [15].

To our knowledge, Casual Conversations v2 (CCv2) is the most comprehensive dataset that is constructed with ethical considerations in mind and may be used in many AI tasks from computer vision and audio/speech recognition to deepfake detection — not only for measuring fairness but also evaluating robustness of the models. Furthermore, this dataset may also be used for model training (note that certain labels age, gender, disability, physical adornments & attributes may not be used for training). Please see the license agreement for further use of the data².

2. Related Work

In this section, we review a few datasets that were assembled for the purpose of or were later used to measure AI fairness. The UTKFace dataset [33] contains over 20,000 face images that were assembled from either other existing datasets or the web, and it contains a wide span of ages (0-116), however, binary gender and five ethnicity groups. This dataset was primarily assembled to train an age progression/regression model, however, it has been used for evaluating fairness of age, gender, and race detectors [20, 23]. FairFace [19] is a crowd-sourced dataset containing 108,501 images, that are labelled with age, gender and race. The seven race groups used are White, Black, Indian, East Indian, Southeast Asian, Middle East, and Latino, and the dataset is reasonably balanced across these groups.

²<https://ai.facebook.com/datasets/casual-conversations-v2-downloads>

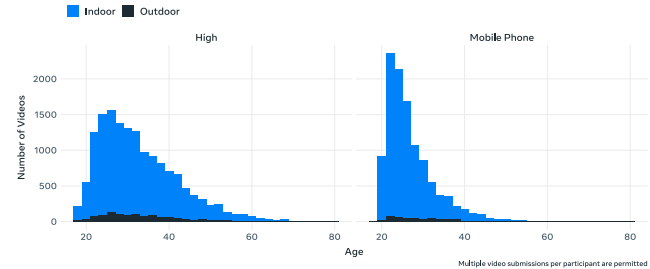


Figure 3. **Age** distribution by video quality. Most videos are recorded indoors. High quality videos are greater than or equal to 720p.

However, race is seen as a social construct [28] and its use in categorization exercises may be problematic [15]. FairFace also contains age, but non-inclusive binary gender labels. The authors examined models on four datasets, and reported that models trained on FairFace are more accurate on other datasets, and the accuracy is consistent across race and gender groups.

The MIAP (More Inclusive Annotations for People) [29] is a subset of the OpenImages dataset [22], and contains additional labels for 100,000 images. The additional labels provide bounding boxes and demographic attributes for everyone present in an image. The demographic attributes include perceived gender and age range. However, the groups for gender (feminine, masculine, unknown) and age range (young, middle, older, unknown) are quite limited in number. These annotations allow fairness analysis on models performing vision tasks. RFW [31] is a face recognition dataset composed of four race groups (Caucasian, Indian, Asian, African) each of which contains around 10,000 images. In addition to race being a social construct, the number of race groups (4) is also extremely limited. This dataset was created by pulling images from the MS-Celeb-1M [14] dataset. The only annotation provided is race, which is done semi-automatically, where authors check low confidence score of the tool used to predict race. The authors report that there is a significant bias in verification accuracy towards Caucasians, when compared to the other three race groups. And as previously mentioned, our dataset is consent-driven.

MORPH [26] is a dataset composed of 55,134 mug-shot photos of 13,617 subjects, where the race, gender (binary), and age attributes are manually assigned for each individual. This dataset also has the recurring issues with race and binary gender. MORPH was initially assembled to study age progression across several types of facial analysis, including animation, face modeling, and face recognition, however, as it contains manually annotated demographic attributes, this dataset has become commonly used to study fairness of AI models [1, 3, 10, 25, 32].

Differing from the aforementioned datasets, the pre-

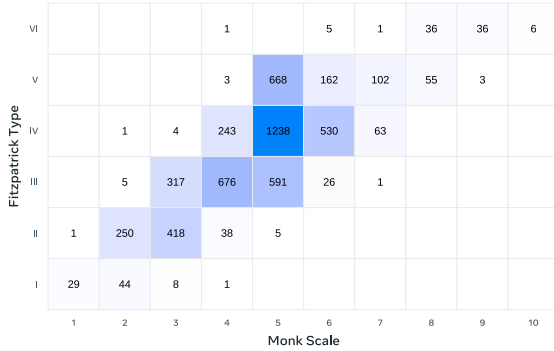


Figure 4. **Fitzpatrick** vs **Monk** skin tone scales in CCv2. More than 90% of the annotations have medium to high confidence in either scale.

cursor of the proposed dataset, Casual Conversations (CCv1) [16] was created with full consent of the participants. CCv1 is composed of around 45,000 videos and 3,011 subjects, which were collected across five U.S. cities. This dataset provides self-provided annotations for age, and gender, and annotated labels for apparent skin tone, and low ambient lighting. Where previous datasets only use binary genders (with/without unknown as option), which can be seen as discriminative, CCv1 has a multi-choice for gender that are “male”, “female” and “other”. This dataset was primarily used in the DeepFake Detection Challenge (DFDC) [9], where videos were augmented with deepfakes, and competitors designed algorithms to detect fake videos. Alongside fairness for age, gender and apparent skin tone detectors, Casual Conversations [16] also report on fairness of DeepFake detectors.

DollarStreet [27] dataset is designed to ensure computer vision fairness across different populations, and is composed of images of everyday household items from 63 different countries around the globe. This dataset is composed of 38,479 images with manual annotations for the objects present, containing 289 categories. Along with these annotations, the dataset also provides country, region, and monthly income as demographic features. In their experiments with current vision models, the authors show that there is a significant performance bias to household items coming from higher incomes. They also show that by fine-tuning models on their dataset, fairness issues can be mitigated.

3. Casual Conversations v2 Dataset

CCv2 is a dataset that has been carefully curated to improve the fairness and robustness of audio, vision, and speech models. The dataset is diverse across multiple axes, including geographically, demographically, and linguistically, which aids in ensuring that the models that are bench-

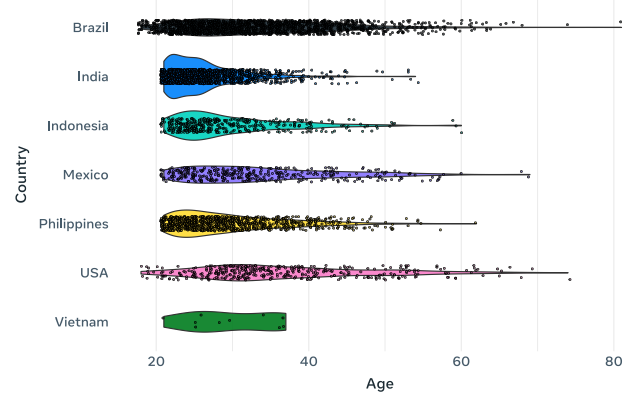


Figure 5. **Age** distribution by country. The median age of participants in India is 25 compared to 34 in the USA.

marked or trained on it are representative of a wide range of subgroups. CCv2 is constructed with the recordings of paid participants that have explicitly provided their consent for their videos’ use in research. This ensures that the dataset aligns with ethical standards *e.g.* [4], for data collection, respects the privacy and autonomy of the participants, but also promotes transparency and other key ethical considerations in responsible data collection practices. The dataset offers 7 self-provided and 4 annotated attributes. Categories and subcategories in our dataset were informed by the literature review presented by Hazirbas *et al.* [15]. We collected our dataset in 7 countries, *i.e.* Brazil, India, Indonesia, Mexico, Vietnam, Philippines and the United States of America. The dataset is composed of 26,467 videos of 5,567 unique paid participants. There are approximately 674 hours of recordings in total, with 354 hours of videos with *nonscripted* and 320 hours of videos with *scripted* text (see Appendix A.2).

Participants self-provided their *age, gender, language/dialect, disability, physical adornments & attributes and geo-location*. For geo-location, we release only “country” and “state/region” information. All self-provided fields were optional during recordings and participants are allowed to withdraw their data anytime after the collection.

In addition to the self-provided categories, trained annotators labeled each participant for *Fitzpatrick Skin Type* [11] and *Monk Skin Tone* [24] scales, *voice timbre* as well as annotated videos for *different recording setup* and *per-second activity* on videos. By providing both skin tone scales, we believe researchers will be able to also compare these two scales in their studies. *Voice timbre* is usually used in music industry for categorizing singing voice using voice types [6]. In this dataset, we only annotate for *low* (bass), *average* (alto, tenor) and *high* (soprano) pitch. During labelling, we provided an example of voice timbre classification video³ and only presented audio to the raters to

³<https://www.youtube.com/watch?v=1IfxH2119cU>

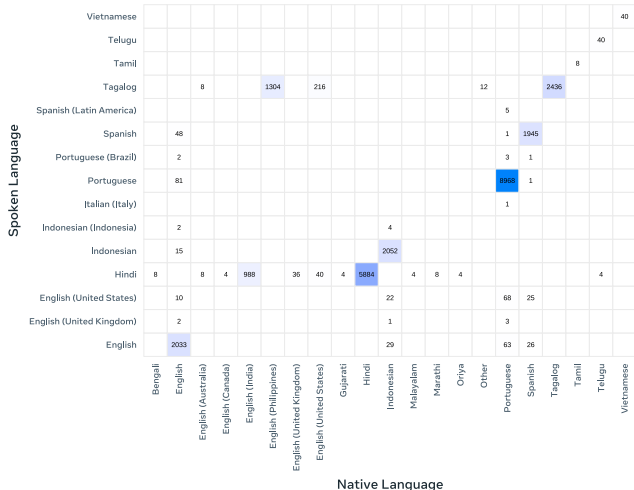


Figure 6. **Spoken vs Native language/dialect** distribution. Portuguese, Hindi, Indonesian, Tagalog and English are most preferred spoken languages/dialects. Participants may speak in more than one language across their video submissions.

make the annotation task apparent gender-blind and therefore remove gender-based voice classification bias during labelling.

As different recording setups, we include *video setup: nonscripted* refers to one of the five questions in Appendix A.2.1 selected by participants in each *nonscripted* video. Participants may have multiple videos of different nonscripted questions. *Scripted* refers to three paragraphs that are read by participants. Text is sampled from *The Idiot* book by *F. Dostoyevsky* and is translated from English (U.S.) to all “spoken languages/dialects” in the videos (see Appendix A.2.2); *video quality: high* if resolution is higher than or equal to 720 pixels or *mobile phone*; *background noise: boolean* if there is any background noise in the video; *capture environment: indoor* or *outdoor*; *hemisphere: north* or *south*; *weather*⁴: *cloudy, dark, rainy, sunny* and *video duration in seconds*.

Activity per-second is annotated for *three* categories: 1) *action*: standing, walking, sitting, laying, waving, 360 rotation; 2) *gesture*: stretching body, raising hand/leg, moving head; and *appearance*: full body visible, upper body visible, lower body visible, only head visible. We asked some of the participants to complete a full rotation (360 degrees) at the end of the video and it is marked under *action*. In some videos, rotation may not be complete and we left it to users’ discretion to remove these parts of the videos. Annotations are completed only for the first frame of each second in videos.

In addition to annotated labels, we also share annotators’ label confidence for *skin tone, voice timbre* and *activity* in

⁴We include weather also for videos recorded indoor.

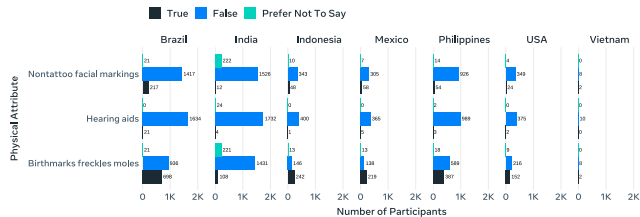


Figure 7. **Binary physical attributes** distributions by country.

the scale of *low, medium* and *high*.

Figure 2 shows the gender distribution over countries. We have more participants in Brazil and India than the rest of the countries and most of the participants in the dataset identified themselves as “cis man” and “cis woman”.

The age distribution across video quality is shown in Figure 3. The majority of videos are indoors, and with younger participants. For mobile phone videos, we see a significant peak around the age of twenty. Furthermore, we show the age distribution by country in Figure 5. While there is a wide range of ages, most participants are less than 40 years old.

Figure 6 shows a contingency table between spoken language and native language. In this figure, we can see the most correlation between the native language and the spoken language, as expected. The language spoken by most different native language groups is Hindi, while the native language group that speaks most different languages are English and Portuguese.

Figure 4 compares skin tones annotated with Fitzpatrick Type and Monk Scale. We can see some correlation between both types of skin tone, however, as expected skin tone is more spread in Monk Scale.

Furthermore, we also show the binary physical attributes by country in Figure 7, disability by country in Figure 8, US state distribution of participants in Figure 9 and also the frequency of uploads per participant in Figure 10.

4. Conclusion

We have presented The Casual Conversations v2 (CCv2) dataset, a diverse, large benchmark for measuring fairness and robustness in audio, vision, and speech models. To the best of our knowledge, CCv2 is the largest and most diverse consent-driven dataset for fairness and robustness benchmarking. It contains 7 self-provided and 4 annotated attributes, that enable fairness and robustness measurements across multi-modalities. It is our hope that CCv2 will boost the development of AI models that are more fair and robust across the proposed attributes.

Acknowledgements. We would like to extend our heartfelt gratitude to the following teams for their invaluable partnership and support throughout this research: the Civil Rights team, the Accessibility team, the Responsible AI team, the

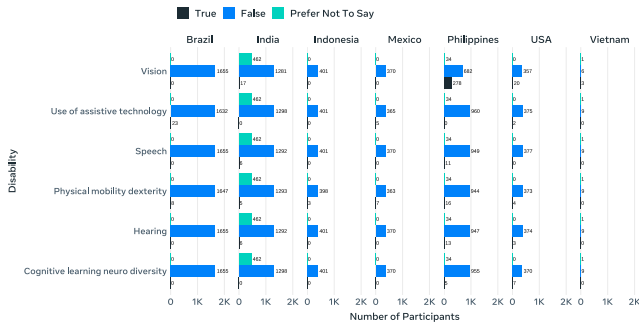


Figure 8. Disability distributions by country.

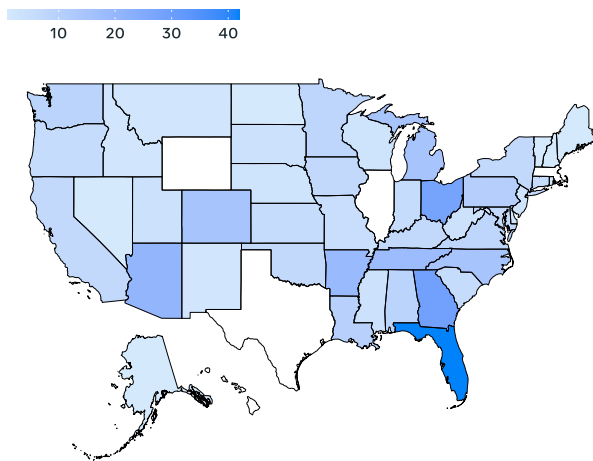


Figure 9. US state distribution of participants.

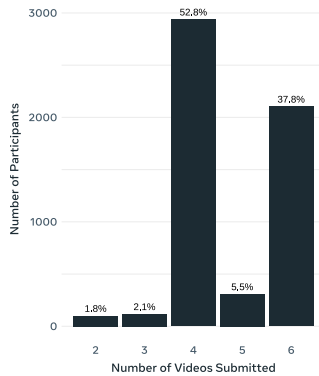


Figure 10. Frequency distribution of uploads per participant.

AI Analytics team, the Assistant team, the Speech Recognition team, and the FAIR team. Without their collaboration, this project would not have been a success.

We are particularly grateful for the exceptional work and invaluable insights provided by Parisa Assar, who served as the program manager and played a critical role in the project. Her attention to detail, expertise, and dedication were essential to the success and efficient execution of this

research.

We also extend our appreciation to Ida Cheng for her diligent work. Her commitment to quality has been instrumental in ensuring high standards for the accuracy and reliability of the data we collected.

Additionally, we would like to acknowledge the partnership and support provided by Miranda Bogen and Lauren Cohen. Their contributions and insights were instrumental in shaping this dataset.

Thank you all for your commitment to this project and for your tireless efforts in advancing our research. Our research has been significantly shaped by your valuable contributions.

References

- [1] Vitor Albiero, Kevin Bowyer, Kushal Vangara, and Michael King. Does face recognition accuracy get better with age? deep face matchers say no. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 261–269, 2020. 2
- [2] Vitor Albiero, S. Krishnapriya K., K. Vangara, K. Zhang, Michael C. King, and K. Bowyer. Analysis of gender inequality in face recognition accuracy. *IEEE Winter Applications of Computer Vision Workshops*, 2020. 1
- [3] Vitor Albiero, Kai Zhang, Michael C King, and Kevin W Bowyer. Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. *IEEE Transactions on Information Forensics and Security*, 17:127–137, 2021. 2
- [4] Jerone T. A. Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, Shruti Nagpal, and Alice Xiang. Ethical considerations for collecting human-centric image datasets. *arXiv*, 2023. 1, 3
- [5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, 2018. 1
- [6] The Kennedy Center. Understanding different voice types. <https://www.kennedy-center.org/education/resources-for-educators/classroom-resources/media-and-interactives/media/opera/understanding-different-voice-types>. Accessed: 2023-02-27. 3
- [7] Cynthia M. Cook, John J. Howard, Yevgeniy B. Sirotnin, Jerry L. Tipton, and Arun R. Vemury. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2019. 1
- [8] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, June 2019. 1
- [9] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset, 2020. 3
- [10] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva,

- Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 2020. 1, 2
- [11] Thomas B. Fitzpatrick. “Soleil et peau” [Sun and skin]. *Journal de Médecine Esthétique (in French)*, 2:33–34, 1975. 3
- [12] Priya Goyal, Adriana Romero Soriano, Caner Hazirbas, Levent Sagun, and Nicolas Usunier. Fairness indicators for systematic assessments of visual feature extractors. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022. 1
- [13] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Ongoing face recognition vendor test (frvt) part 2: Identification, 2018. 1
- [14] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, 2016. 2
- [15] Caner Hazirbas, Yejin Bang, Tiezheng Yu, Parisa Assar, Bilal Porgali, Vítor Albiero, Stefan Hermanek, Jacqueline Pan, Emily McReynolds, Miranda Bogen, Pascale Fung, and Cristian Canton Ferrer. Casual conversations v2: Designing a large consent-driven dataset to measure algorithmic bias and robustness. *arXiv*, 2022. 1, 2, 3
- [16] 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*, 2022. 1, 3
- [17] The White House. Blueprint for an AI bill of rights. <https://www.whitehouse.gov/ostp/ai-bill-of-rights>. Accessed: 2023-02-16. 1
- [18] John J. Howard, Yevgeniy B. Sirotin, and Arun R. Vemury. The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance. In *International Conference on Biometrics Theory, Applications and Systems*, 2019. 1
- [19] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021. 1, 2
- [20] Anoop Krishnan, Ali Almadan, and Ajita Rattani. Understanding fairness of gender classification algorithms across gender-race groups. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1028–1035. IEEE, 2020. 2
- [21] K. S. Krishnapriya, Vítor Albiero, Kushal Vangara, Michael C. King, and Kevin W. Bowyer. Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society*, 2020. 1
- [22] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, 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. 2
- [23] Xiaofeng Lin, Seungbae Kim, and Jungseock Joo. Fairgrape: Fairness-aware gradient pruning method for face attribute classification. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 414–432. Springer, 2022. 2
- [24] Ellis P Monk Jr. Skin tone stratification among black americans, 2001–2003. *Social Forces*, 92(4):1313–1337, 2014. 3
- [25] Ying Qiu, Vítor Albiero, Michael C King, and Kevin W Bowyer. Does face recognition error echo gender classification error? In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021. 2
- [26] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *7th international conference on automatic face and gesture recognition (FGRO6)*, pages 341–345. IEEE, 2006. 1, 2
- [27] 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 *NeurIPS*, 2022. 1, 3
- [28] Wendy D. Roth. The multiple dimensions of race. *Ethnic and Racial Studies*, 39(8):1310–1338, 2016. 2
- [29] Candice Schumann, Caroline Rebecca Pantofaru, Susanna Ricco, Utsav Prabhu, and Vittorio Ferrari. A step towards more inclusive people annotations for fairness. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2021. 1, 2
- [30] The European Union. The Artificial Intelligence Act. <https://artificialintelligenceact.eu>. Accessed: 2023-02-23. 1
- [31] 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 *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1, 2
- [32] Haiyu Wu, Vítor Albiero, KS Krishnapriya, Michael C King, and Kevin W Bowyer. Face recognition accuracy across demographics: Shining a light into the problem. *arXiv preprint arXiv:2206.01881*, 2022. 2
- [33] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

A. Casual Conversations v2

In this appendix, we provide details of the categories and subcategories.

A.1. Annotations File

We provide annotations in a JSON file. For each video we provide all attributes except activity that is stored in a separate JSON file to allow fast loading of the data into memory. Subject ids are numbered from 0000 to 5566 and video names are constructed from *subject id*, *spoken language*, *video setup* and *video number*.

```
{
  "video_name": "3828_spanish_nonscripted_1.mp4",
  "subject_id": "3828",
  "age": 22,
  "gender": "transgender man",
  "native_language": "spanish",
  "secondary_languages": [
    "english (united states)",
    "english (united kingdom)",
    "spanish (latin america)",
    "japanese (japan)",
    "italian (italy)"
  ],
  "disabilities": {
    "vision": false,
    "hearing": false,
    "physical_mobility_dexterity": false,
    "speech": false,
    "cognitive_learning_neuro_diversity": false,
    "use_of_assistive_technology": false
  },
  "physical_adornments": {
    "have_hair_cover": false,
    "hair_color": "blue",
    "have_beard_mustache": false,
    "have_face_covering": false,
    "have_face_mask": false,
    "have_make_up": false,
    "have_eye_wear": true,
    "have_ear_wear": false,
    "have_visible_tattoos": false,
    "have_bindi": false,
    "have_visible_piercings": false
  },
  "physical_attributes": {
    "hair_type": "curly",
    "hair_color": "brunette",
    "have_hearing_aids": false,
    "eye_color": "brown",
    "have_birthmarks_freckles_moles": true,
    "have_nontattoo_facial_markings": false
  },
  "voice_timbre": {
    "type": "high pitch",
    "confidence": "high"
  },
  "fitzpatrick_skin_tone": {
    "type": "type iii",
    "confidence": "medium"
  },
  "monk_skin_tone": {
    "scale": "scale 5",
    "confidence": "medium"
  },
  "geo_location": {
    "country": "mexico",
    "state_region": "baja california"
  },
  "video_setup": {
    "type": "nonscripted",
    "speech_topic":
      "tell us how you usually spend your weekends"
  }
}
```

```
},
"video_quality": "high",
"background_noise": false,
"spoken_language": "spanish",
"capture_environment": "indoor",
"hemisphere": "northern_hemisphere",
"weather": "sunny",
"video_duration_secs": "91.895122"
}
```

We recommend users to load the JSON files into memory using Pandas library in Python:

```
# import Pandas
import pandas as pd

# load annotations into memory
annotations = pd.read_json("CasualConversationV2.json",
                           dtype={'subject_id': object})

# load activity annotations in memory
activity_annotations = pd.read_json("
  CasualConversationV2_activity.json", dtype={'
    subject_id': object})

# serialize the dictionary fields
geoloc = annotations.geo_location.apply(pd.Series)

# concatenate country with the data frame
annotations = pd.concat([annotations, geoloc.country],
                        axis=1)
```

A.2. Video Setup

In video setup, for each participant we have at least one nonscripted and one scripted video. Most of the participants recorded more than two videos. Exact total duration of recordings is 674 hrs 22 min 45 sec. Out of this, 354 hrs 28 min 8 sec of it with nonscripted and 319 hrs 54 min 37 sec of it with scripted text.

A.2.1 Nonscripted Text

In nonscripted videos, we asked participants to choose one of the following five questions and provide a monologue in about 1 minute. For each nonscripted video, we store the corresponding question.

- * tell us how you connect with family and friends
- * tell us how you usually spend your weekends
- * tell us if you would rather spend your leisure time in nature or in a city and why
- * tell us what activities you like to do in summer
- * tell us what you think about the weather in the city you live in

A.2.2 Scripted Text

The Idiot book by F. Dostoyevsky

“Toward the end of November, during a thaw, at 9 o’clock one morning, a train on the Warsaw and Petersburg railway was approaching the latter city at full speed. The morning was so damp and misty that it was only with great difficulty that the day succeeded in breaking; and it was impossible to distinguish anything more than a few yards away from the rail car windows.

Some of the passengers by this particular train were returning from abroad; but the third-class carriages were the most filled up, mainly with insignificant persons of various occupations and degrees, picked up at the different stations nearer town. All of them seemed weary, and most of them had sleepy eyes and a shivering expression, while their complexions generally appeared to have taken on the color of the fog outside.”

One of them was a young man of about twenty-seven, not tall, with black curling hair, and small, gray, fiery eyes. He wore a large fur—or rather astrakhan—overcoat, which had kept him warm all night, while his neighbor had been obliged to bear the full severity of a Russian November night entirely unprepared. The wearer of this cloak was a young man, also of about twenty-six or twenty-seven years of age, slightly above average height, very fair, with a thin, pointed and very light-colored beard; his eyes were large and blue, and had an intent look about them.

“Cold?”

“Very,” said his neighbor, readily, “and this is a thaw, too. Imagine if it had been a hard frost! I never thought it would be so cold in the old country. I’ve gotten quite unaccustomed to it.”

“What, been abroad, I suppose?”

“Yes, straight from Switzerland.”

“Wow! My goodness!” The young, black-haired man whistled, and then laughed.

are released as part of the annotations. All “spoken languages/dialects” are English (U.S.), English (U.K.), Hindi, Indonesian, Italian, Portuguese, Portuguese (Brazil), Spanish, Spanish (Latin America), Tagalog, Tamil, Telugu and Vietnamese.

The scripted text is taken from *The Idiot* book by F. Dostoyevsky. The text is translated into all other *twelve* spoken languages/dialects present in the videos and translations