Words are all you need? Capturing human sensory similarity with textual descriptors

Raja Marjieh[1,*] **Pol van Rijn**[2,*] **Ilia Sucholutsky**[3,*]

Theodore R. Sumers[3] **Harin Lee**[2,4] **Thomas L. Griffiths**[1,3,**] **Nori Jacoby**[2,**]

[*/**]**Equal contribution.**
[1]Department of Psychology, Princeton University, USA
[2]Max Planck Institute for Empirical Aesthetics, Germany
[3]Department of Computer Science, Princeton University, USA
[4]Max Planck Institute for Human Cognitive and Brain Sciences, Germany

Abstract

Recent advances in multimodal training use textual descriptions to significantly enhance machine understanding of images and videos. Yet, it remains unclear to what extent language can fully capture sensory experiences across different modalities. A well-established approach for characterizing sensory experiences relies on similarity judgments, namely, the degree to which people perceive two distinct stimuli as similar. We explore the relation between human similarity judgments and language in a series of large-scale behavioral studies ($N = 1,823$ participants) across three modalities (images, audio, and video) and two types of text descriptors: simple word tags and free-text captions. In doing so, we introduce a novel adaptive pipeline for tag mining that is both efficient and domain-general. We show that our prediction pipeline based on text descriptors exhibits excellent performance, and we compare it against a comprehensive array of 611 baseline models based on vision-, audio-, and video-processing architectures[2]. We further show that the degree to which textual descriptors and models predict human similarity varies across and within modalities. Taken together, these studies illustrate the value of integrating machine learning and cognitive science approaches to better understand the similarities and differences between human and machine representations.

1 Introduction

Whether playing an instrument or deciding when to cross a street, human experience is based on complex multimodal sensory information. Language is an extremely efficient way for humans to communicate information about their sensory environment [1–4]. However, a long-standing problem in cognitive science concerns the limitations of language as a tool for describing the full extent of sensory experiences [5, 6]. Recent advances in machine learning (ML) suggest that, like humans, machines can benefit greatly from language [7–9]. Traditional supervised learning models are trained on massive amounts of labeled data, often exhibiting human-level performance on target tasks [10]. Nevertheless, these models are criticized for failing to capture certain aspects of scene understanding,

[*]Correspondence to: {raja.marjieh, is2961}@princeton.edu, pol.van-rijn@@ae.mpg.de

[2]We present an interactive visualization https://words-are-all-you-need.s3.amazonaws.com/index.html for exploring the similarity between stimuli as experienced by humans and different methods reported in the paper.

Preprint. Under review.

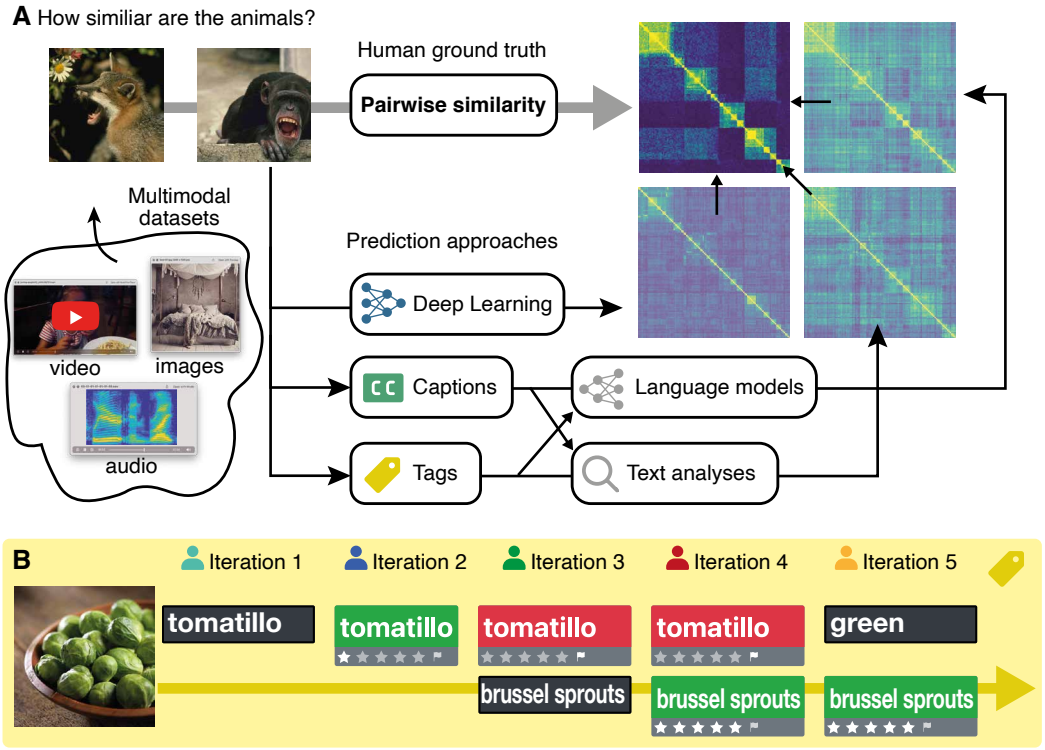

Figure 1: Comparing human similarity scores gathered through crowdsourcing with ML pipelines. **A**: We used data from three modalities: images, audio, and video. For each modality, we extracted deep model embeddings and gathered human captions and tags. Word- and language-embedding models, as well as simple text analysis, were used to predict human similarity judgments. **B**: Novel tag-mining paradigm. We ran an adaptive process in which results of one iteration are used as inputs for subsequent iterations. In every iteration, participants can add a new tag, rate the relevance of existing tags or flag tags that are inappropriate.

such as physics [11] or compositionality [12]. Modern multimodal models seem to push scene understanding to a new frontier by jointly training on multimodal datasets (images/video/sounds) along with detailed textual descriptions [7–9, 13, 14]. But are there any limitations for learning from supervised textual descriptions? And to what degree can task-relevant sensory information be obtained from supervised linguistic descriptors?

One common way to investigate human sensory experiences is using similarity judgments, namely gathering perceived similarity ratings on pairs of distinct stimuli. Given a set of N stimuli, similarity-based techniques begin by constructing an $N \times N$ similarity matrix, whereby each entry s_{ij} corresponds to the degree of similarity between stimuli i and j. Then, a suitable algorithm (e.g., multidimensional scaling [15] or UMAP [16]) finds a low-dimensional embedding such that similar stimuli are mapped to nearby points in space. In psychology, this process has been used to shed light on the underlying structure of the perceptual space and its associated mental representation [15]. It has also inspired ML techniques that use similarity between batches of stimuli as part of their objective function [7]. Here, our aim is to investigate the extent to which textual descriptions can predict human similarity judgments across the visual, audio, and audio-visual modalities.

We considered datasets from three domains, namely images, sounds, and videos (Figure 1A). For each of these datasets, we collected two types of text descriptors that serve complementary roles: a) free-text captions that emulate the unconstrained form in which humans describe objects, and b) concise semantic tags consisting of single words or small combinations of several words that capture the important aspects of complex stimuli. Free-text descriptions are easy to crowdsource online and are part of the common practice in modern ML pipelines. For tag collection, we used a novel paradigm for tag-mining to gather high-quality responses (Figure 1B). Our paradigm extends existing crowdsourcing text-mining techniques [17–20] by integrating ideas from transmission chain

experiments [21, 22]. In this *tag-mining* paradigm, participants adaptively create annotations (or tags) for a set of target stimuli and rate the annotations of others. In each trial, each participant inspects a target stimulus (e.g., an image) and is asked to rate the relevance of tags that were created by other participants or flag a tag that is inappropriate (with tags removed if they were flagged more than twice). Participants are also given the opportunity to add new tags if desired. The results of the annotation procedure of one participant then propagate to the next participant. Thus, as the process unfolds over iterations, it quickly converges on top tags for each target stimulus (additional details about the paradigm, and screenshots are provided in Supplementary Section B.6).

To generate predictions and quantify semantic similarity, we used word- and language-embedding models that are capable of generating vector representations for both individual words and freely generated text (e.g., CNNB [23], SimCSE [24] and CLIP [7]). In addition, we complemented this approach with embedding-free methods based on word co-occurrence that rely on bag-of-words. Our objective here is to determine how well we can predict human similarity judgments across the three modalities (images, sounds, and video) using a combination of ML and human text-mining.

As a baseline, in each modality, we tested a wide range of pre-trained ML models that do not rely on text (overall we tested 611 models) and compared their internal representations to human similarity judgments and text-based predictions (Figure 1A). We also examined whether there are specific types of architectures that are better suited for predicting human similarity, and whether modern multimodal training indeed makes a difference in producing human-like representations. Moreover, we wanted to investigate whether there are systematic variations in human similarity not explained by contemporary machine representations, and if so, how large the gap is between human and machine representations.

The contributions of this paper are as follows: a) We compare human similarity judgments to a comprehensive set of pre-trained ML models in three different modalities: images (Section 3.1; including state-of-the-art multimodal models), sounds (Section 3.2), and videos (Section 3.3). b) These baseline models are compared to embedding-based and embedding-free approaches applied to textual descriptors in order to show that the performance of text descriptors is comparable to image-, audio-, and video-processing architectures. c) We introduce a novel adaptive tag-mining procedure that is domain-general, efficient, and can predict human sensory similarity well (Section 3.1). d) We present two new large descriptor datasets of audio and video to compare the performance of ML algorithms with human similarity judgments (Section 3.4).

2 Comparison to previous research

In the past few years, significant effort has been put into the comparison of human internal representations and brain signals to ML representations [25–32]. In this context, different notions of similarity are often used as a tool for studying internal representations [15, 33–40]. Despite the success of similarity-based techniques, the quadratic scaling of the number of human comparisons as a function of the number of stimuli limits their applicability to large naturalistic and machine-learning datasets. This problem stimulated recent research which attempts to predict similarity judgments using cheaper data [36–38, 41–43]. For example, Peterson et al. [36] estimated perceptual similarity over images by leveraging the latent representations of convolutional neural networks (CNNs). Here we extend this comparison to significantly more models ranging across diverse architectures. Others have developed custom active-sampling pipelines for finding maximally informative comparisons [42]. These studies, however, are domain-specific and focus specifically on images, which potentially over-rely on low-level features to guide prediction. Language, on the other hand, provides an attractive alternative as it is well-suited for use across modalities, and text descriptions scale only linearly in the number of stimuli, which makes crowdsourcing much more feasible. More recently, we explored [38] the use of text captions to predict human similarity, however, the work focused solely on images, it did not collect tag data that were not available for the datasets, and was limited to the datasets already analyzed by Peterson et al. [36]. In addition, the work did not include a comprehensive comparison of ML architectures, nor consider combining visual and textual representations to explore the sources of variance in human judgments. Our current work seeks to remedy these limitations and extend the study of representational similarity between humans and machines (e.g., [25–28, 32]), by comprehensively studying text-based proxies jointly with a large baseline of modern architectures across different modalities, and providing an efficient pipeline for crowdsourcing textual descriptors.

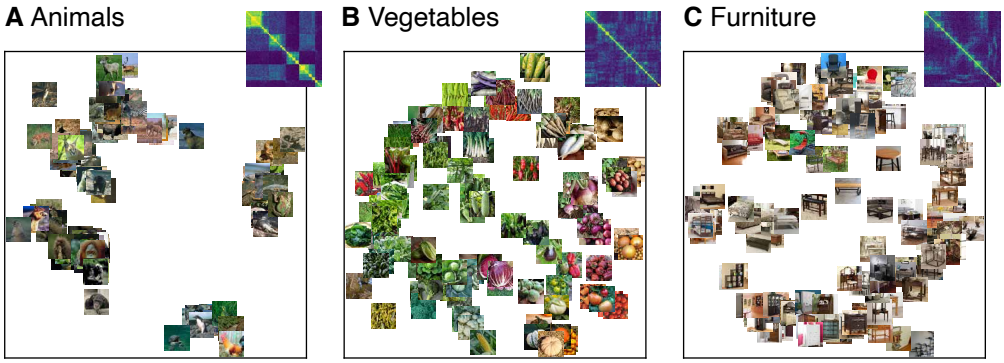

Figure 2: Similarity matrices and MDS embeddings for the three image datasets.

3 Studies

Experiment details: participants and compute time. We collected data from $N = 1,823$ US participants for the new behavioral experiments reported in this paper. Participants were recruited anonymously from Amazon Mechanical Turk and provided informed consent under an approved protocol by either the Institutional Review Board (IRB) at Princeton University (application 10859) or the Max Planck Ethics Council (application 2021_42) before taking part. Participants earned 9-12 USD per hour, and each session lasted less than 30 minutes. The total amount spent on participant compensation was 6,319 USD. The recruitment criteria were participating in more than 5,000 previous tasks with a 99 % approval rate (see Supplementary Section B for additional details about the behavioral experiments). All experiments were implemented with the Dallinger and PsyNet frameworks for large-scale behavioral research [44]. In Supplementary Section A.2, we include the data that was collected, instructions, and code for replication of the behavioral experiments. We also provide the code for computational experiments and analysis. Our computational experiments cumulatively took about 2 weeks of continuous run-time on a x1.16xlarge AWS instance with 64 vCPUs and 976 GiB of memory (see Supplementary Section C for additional details about the computational experiments).

3.1 Images

Images are both high-dimensional and naturalistic, and at the same time benefit from a large array of available embedding models, such as convolutional neural networks and vision transformers [45]. To link our work to previous research, we considered three image datasets of common objects (animals, furniture, and vegetables) introduced in [36] and further studied in [38], each comprising 120 images. The similarity matrices (obtained from [36] and used with permission from the authors) are shown in Figure 2 along with their MDS embeddings (for interactive maps see[3]). Captions (but not tags) for these images were collected in [38]. As can be seen, the organization of the stimuli in MDS space is interpretable, capturing semantic groups that can be easily recognized such as predator animals, or vegetables that grow above or below ground.

To collect semantic information on images, we first applied our novel tag-mining pipeline to each of the three datasets to get concise semantic labels that were not available in the original datasets (see Figure 1B; overall, we recruited $N = 171$ participants for this modality). Then, given the captions and the newly collected tags, we generated similarity predictions based on an array of approaches. These approaches can be split into two main groups, namely, embedding-based (e.g. language models) and embedding-free methods (text analysis), depending on whether they used pre-trained machine embeddings or not (Figure 1A). The embedding-based methods can be further split into three groups based on the kinds of input data they process, namely, image [45], text [23, 24] or image-text [7] (denoted respectively as "unimodal baseline", "captions/tags embeddings" and "multimodal baseline" in Figure 3). To compare against the multimodal model CLIP [7], we introduced a set of additional "stacked" representations, which we produce by concatenating the best performing (see below) pure-textual and pure-visual representations into a single long embedding. Similarity predictions for pairs

[3] `https://words-are-all-you-need.s3.amazonaws.com/index.html`

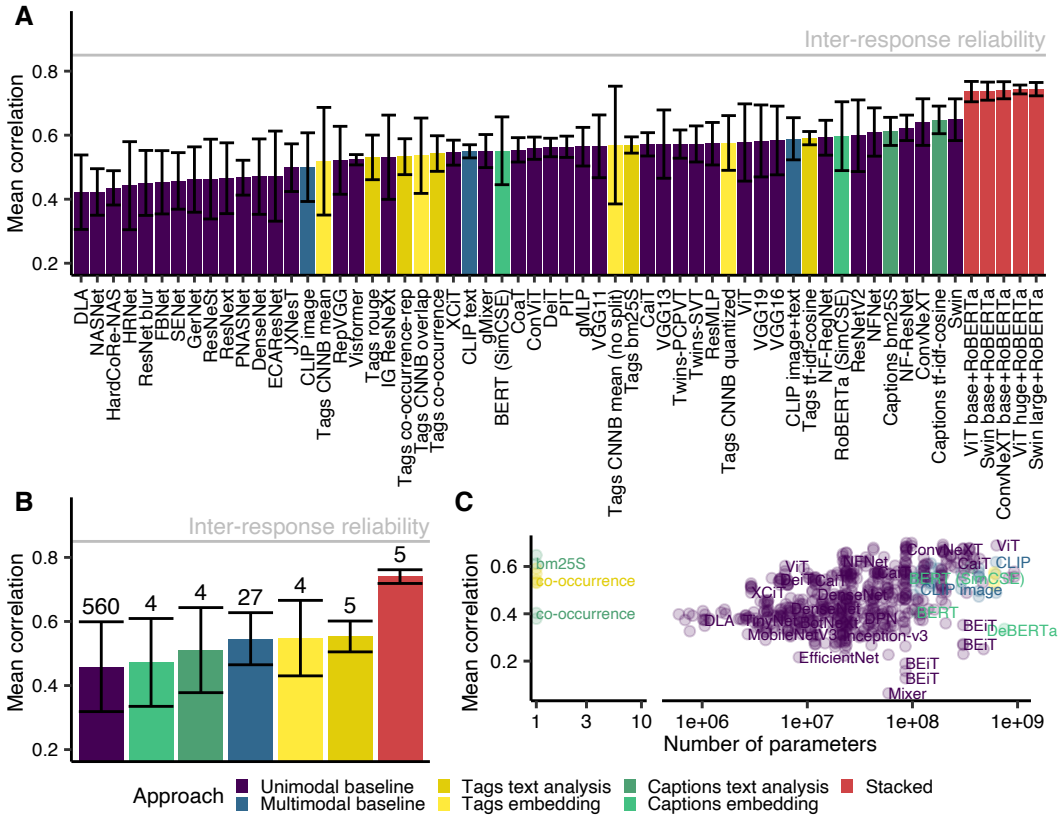

Figure 3: Model evaluation on images. **A**: Correlation between top models and human similarity scores averaged over datasets and architecture variants. Error bars denote one standard deviation (SD). **B**: Mean correlation per prediction approach (the small numbers in the top represent the number of models averaged to produce each bar). **C**: Correlation between models and human similarity as a function of number of model parameters.

of stimuli in the embedding-based group are generated by encoding the stimuli's associated input data (i.e., the image itself or its text descriptors) into a pair of latent vectors and then computing a similarity score by applying a suitable metric such as cosine similarity (see Supplementary Section C). In particular, to embed tags we used ConceptNet Numberbatch (CNNB) which is a word-embedding model trained on the ConceptNet knowledge graph that leverages other popular word embedding models such as word2vec and GloVe [23]. To embed free text, we used variants of SimCSE [24], a recent contrastive language-embedding model that builds on BERT [46] and is tuned for semantic similarity. As for the embedding-free approaches, the similarity was instead computed based on tokenization and co-occurrence counts directly in word space (denoted as "text analysis" in Figure 3). Throughout, we quantified performance by computing the Pearson correlation r between the model predictions and ground truth similarity scores. We compared this simple method with two other criteria in which we introduce a supervised re-weighting for the different embedding dimensions and estimate performance on held-out stimuli, as suggested in the previous literature [36, 38]. We discuss this in Supplementary Section D.1 and validate that they overall give consistent results.

Figure 3 summarizes the performance of the various techniques averaged over the three image datasets and architecture variants (see Supplementary Section D.2). The top-performing models and their approach-level averages are shown in increasing order in Figure 3A-B, relative to the average inter-response reliability (IRR) which serves as an approximated prediction upper bound. We see that the best performing models are the stacked representations which combine visual embeddings such as ConvNext [47] or the vision transformers Swin [48] and ViT [49] with the textual representations of SimCSE [24], outperforming the multimodal baseline CLIP [7]. These are then followed by the unstacked Swin transformer and the embedding-free TF-IDF-cosine method, which computes cosine similarity between the bag-of-words representations of tags/captions scaled with TF-IDF

scores applied to captions (see Supplementary Section C.2.1). Crucially, while the caption-based method performs on par with the best pure-vision model, the boosted performance of the stacked representations suggests that language-based and vision-based representations capture different sources of variance. The remaining gap between the models and the IRR bound suggests that there are aspects of human similarity judgments that are not captured by the models. Possibly, this is due to score dilution that arises from the contributions of irrelevant dimensions in embedding-based methods. To test this, we applied the reweighting procedure discussed above, whereby the contribution of each latent dimension is fine-tuned based on a subset of the similarity scores and then evaluated on held-out-images. We found that, while this procedure does reduce the gap (particularly for caption embeddings), it does not eliminate it (see additional analysis with two types of fine-tuning methods in Supplementary Section D.1). To further investigate the effect of architecture, Figure 3C plots model performance against the number of model parameters on a log scale. Overall, we found a positive correlation between similarity performance and the number of model parameters ($r = 0.41, p < 0.001$). We also found that for baseline models, performance on ImageNet [50] correlated positively with human similarity ($r = 0.26, p < 0.001$), though there were some exceptions with high ImageNet performance but low similarity performance, such as the image transformer BEiT [51] (see Supplementary Section C.1.1 and Figure 13).

3.2 Sounds

We next consider the domain of sounds and specifically focus on emotional prosody. Speech prosody is characterized by variations in pitch, loudness, timing, and voice quality which can communicate the emotional state of the speaker [52]. We selected 1,000 recordings from the RAVDESS corpus ([53], released under a CC Attribution license), which consists of emotionally-neutral sentences spoken by 24 US American actors to convey a specific target emotion. In this corpus, the same sentence is recorded for all emotions by all speakers.

We collected two judgment batches of different respects for similarity on the same subset of 100 recordings. In one batch ($N = 252$ participants), participants were instructed to focus on the emotional similarity of each pair of recordings, whereas in the other ($N = 257$) they were instructed to focus on the speaker's voice irrespective of their emotions (see Supplementary Section B). We elicited an average of 85 judgments per participant, collected over 4,950 unique pairs. The resulting similarity matrices as well as their MDS embeddings are shown in Figure 4 and show commonalities but also substantial differences (the correlation between the similarity matrices corresponding to the identical stimuli under different respects for similarity was $r = 0.33$). In the case of the emotion-respect, a block-diagonal structure emerges in the data corresponding to the underlying target emotions, as well as off-diagonal patterns. The MDS embedding of this matrix reveals a clear valence-arousal distribution (for interactive audio map see[4]) suggesting that this space is indeed semantically rich for participants. As for the speaker respect, the emotion blocks disappear and the MDS map reveals two clear clusters corresponding to the speaker's sex. This finding highlights the idea that perceived similarity can have many respects which can vary in qualitative ways [54].

We also collected tags ($N = 217$ participants) and free-text captions ($N = 151$) for the full 1,000 recordings in the emotion respect, as well as tags ($N = 35$) and captions ($N = 39$) for the 100 recording subset in the speaker respect for comparison. In this section, we focus on text descriptors of the 100-stimulus subset from the similarity experiments (full set discussed in Section 3.4). To generate predictions based on these descriptors, we apply the same techniques as in images. As for baseline audio-processing models, we used a suite of audio models including Wav2Vec [55] and HuBERT [13] (see Supplementary Section C.1.1 for full list). The performance scores are shown in Figure 5. We see that in both cases, the embedding-based caption methods yield the best results, with the SimCSE models saturating the IRR bound for the speaker respect case. The observed gap in the emotion respect potentially reflects the richness and variability of emotional expression.

Another interesting question to ask here is why the same pair of sounds would trigger different responses under different respects for similarity. One idea is that different respects are inferred from different low-level cues. Preliminary support for this was given by the relatively high correlation ($r = 0.42$ and 0.49 for emotion and speaker, respectively) of 88 standard low-level features that are used in voice research (extracted directly from the recordings; [56]) for both respects (see the purple bar, "low-level features" in Figure 5A-B). To further investigate this, we selected pitch- (mean

[4] `https://words-are-all-you-need.s3.amazonaws.com/index.html`

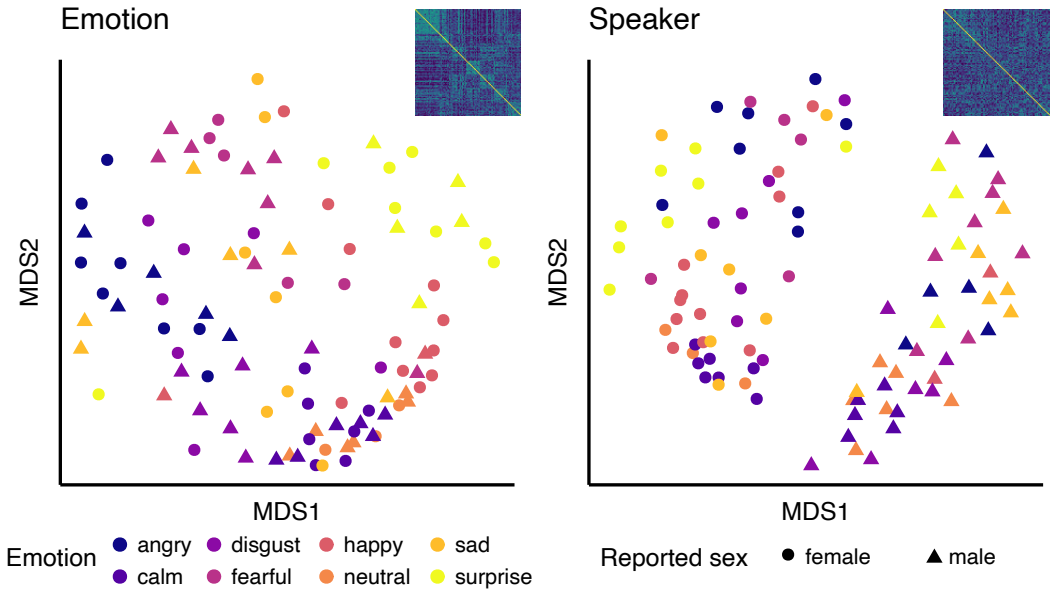

Figure 4: Similarity judgments and MDS embeddings for the emotion and speaker audio respects.

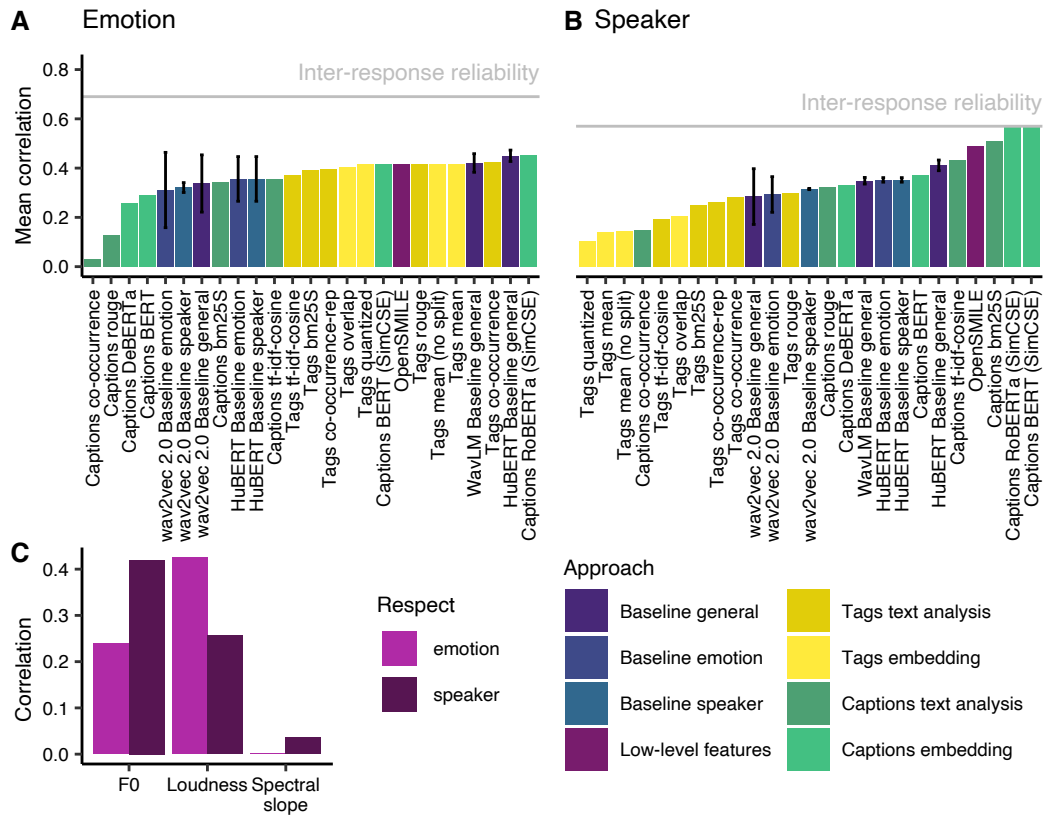

Figure 5: Top models averaged over datasets and architecture variants. **A**: Emotion respect. Error bars denote 1 SD, none if single variant. **B**: Speaker respect. **C**: Results of the 3 low-level features analysis showing an interaction between the feature and the correlation with similarity matrix.

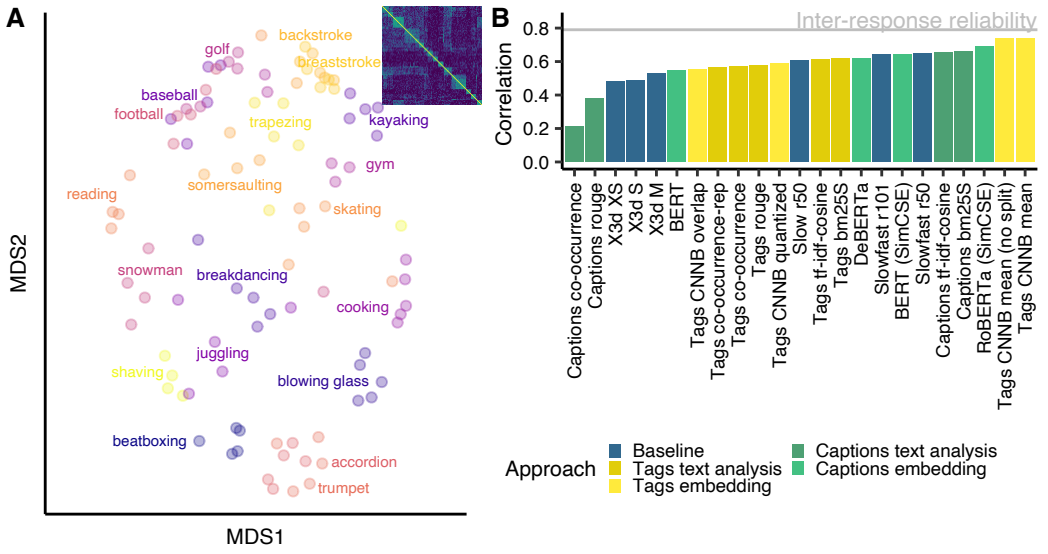

Figure 6: Video dataset. **A**: Similarity matrix and MDS embedding. Colors indicate the activity type in the dataset. **B**: Top performing models.

fundamental frequency), intensity- (loudness), and voice-quality related (spectral slope) features from the set [57] (see Supplementary Section C.2.2). We then correlated the similarity score of each pair of stimuli with the absolute difference between each of their low-level features. The results are shown in Figure 5C. We see that, indeed, there are clear strong low-level predictors in both respects, with loudness being the strongest in the case of emotions, potentially capturing the arousal dimension, whereas mean pitch being the most important in the case of speakers, potentially capturing perceived sex. The predictive power of these low-level features suggests that they can be used themselves as a predictive embedding that can be combined with the other semantic embeddings discussed earlier; we explore this option in Supplementary Section C.1.3.

3.3 Video

Moving one further step up in complexity, we consider the audio-visual domain of videos. We chose in this case to study a subset of the Mini-Kinetics-200 dataset [58] (released under a CC BY 4.0 International License). This subset contains 1,000 short video clips of diverse human activities, ranging from cooking and swimming to playing instruments and reading books (see additional details in Supplementary Section A.1). As before, we began by collecting similarity judgments ($N = 284$ participants) over a subset of 100 videos, as well as tags ($N = 221$) and captions ($N = 151$) over the full 1,000-video dataset. The similarity matrix and its MDS embedding are shown in Figure 6, we see clusters emerging such as different sports-related activities like playing golf, baseball, and football as well as music-making like beatboxing, playing accordion or a trumpet (for interactive map showing videos see[5]). As for baseline models, here we considered three modern video-processing architectures known as SlowFast and Slow [59], and X3d [60] which incorporate information from different temporal and spatial scales to capture motion dynamics as well as fine details. The performance of the various techniques is shown in Figure 6. We see that embedding-based techniques applied to tags and captions perform particularly well, reaching very close to the IRR bound, suggesting that similarity in this domain is heavily driven by high-level semantic knowledge. This is not surprising, as the interpretation of the similarity of activities like beatboxing and playing accordion involves complex knowledge that cannot be mapped into simple low-level cues. We also tried including audio embeddings from the videos in similarity predictions, but this did not increase correlation with human similarity judgments (see Supplementary Section C.1.3).

[5] `https://words-are-all-you-need.s3.amazonaws.com/index.html`

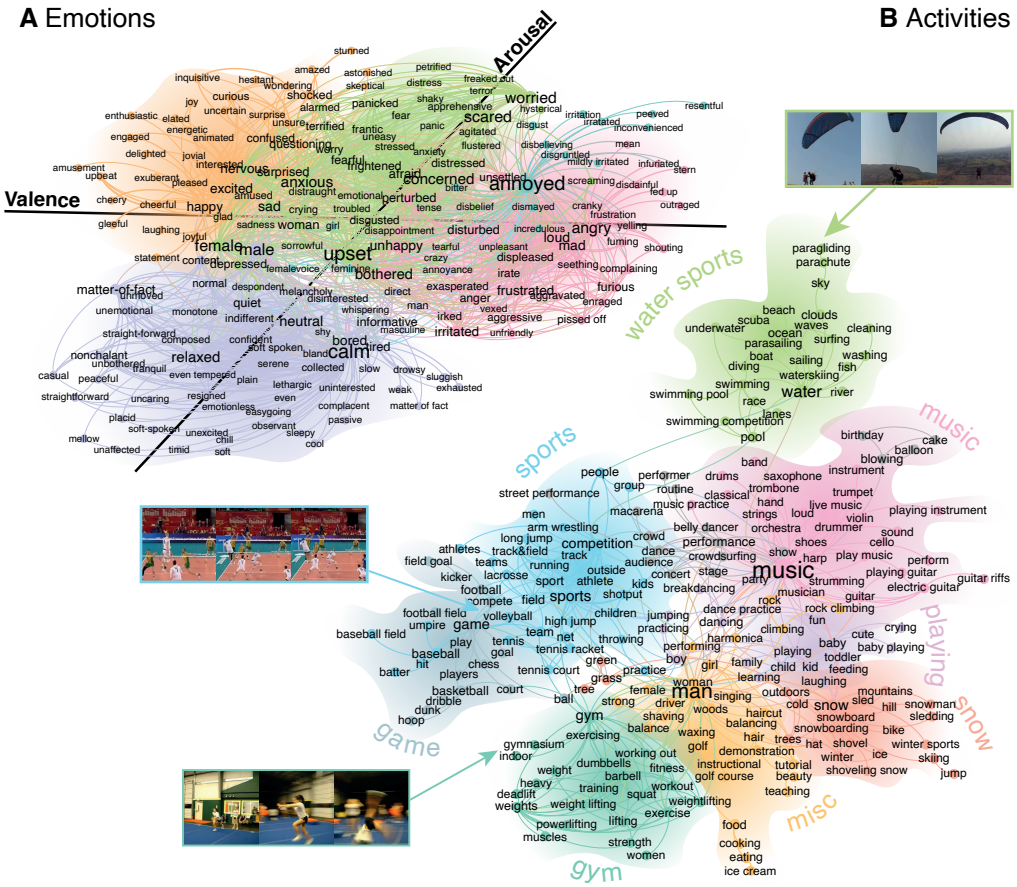

Figure 7: Semantic networks based on tag-mining for the 1,000-object audio and video datasets.

3.4 Similarity at scale

A major limitation of similarity-based approaches is the quadratic scaling of the number of human judgments as a function of the number of stimuli. Indeed, even relatively small datasets (as in this work) on the scale of 100 objects required about 20,000 judgments to produce high-quality similarity matrices and to average out inter-rater noise. This scaling quickly becomes prohibitive once we move one magnitude higher in the number of objects, requiring a variety of sophisticated methods [37, 42, 61]. Our approach suggests that text descriptors provide a much simpler route for generating such proxies, since a) they scale only linearly in the number of objects, and b) they are compatible with different modalities. To get a sense of the semantic content captured by our text descriptors, Figure 7 shows a network analysis inspired by [62] of the tags mined using our adaptive pipeline over the 1,000 emotion and video datasets (see Supplementary Section D.5). As can be seen, both semantic networks are extremely rich. In the case of videos, we see large-scale clustering emerging across the various tags, capturing a variety of activity topics such as sports, music, and playing more generally. As for emotions, the tag distribution factorizes neatly into the classic valence-arousal distribution, but at a much more detailed level. All text descriptors and their associated 1,000-by-1,000 predicted similarity matrices are made publicly available (see Supplementary Section A.2).

4 Discussion

In this work, we studied the extent to which sensory experiences can be captured by language. We quantified sensory experiences using similarity judgments and provided compelling evidence from three different modalities and large behavioral studies showing that textual descriptors, when combined with text embedding models, correlate strongly with human similarity scores and perform just as well as state-of-the-art image-, audio- and video-processing architectures. In doing so, we

introduced a novel adaptive technique for crowdsourcing high-quality semantic tags and highlighted its prospect for studying the semantic organization of large datasets and their similarity structure.

4.1 Limitations and societal implications

One limitation of our work is that our textual representations did not always capture human similarity entirely, as indicated by the observed gaps between model performance and IRR bounds. This suggests that ML algorithms still have room to improve in terms of how they represent stimuli in a human-like manner. Second, a promising avenue for improving machine representations using similarity judgments is to incorporate them in the training objectives of deep models, e.g. in the context of contrastive learning [63]. On the one hand, the proxies generated from our pipeline can support ML datasets, but they are also at risk of baking in high-level human biases that can lead to adverse societal implications, such as amplifying race and gender gaps. Researchers should devote utmost care to what they choose to incorporate in their training objective. On the positive side, we believe that our approach paves the way for the study of cross-cultural variation of human semantic representations by providing efficient tools for crowdsourcing high-quality semantic descriptors across languages. This is particularly relevant for low-resource languages, where our tag-mining techniques can work even with the absence of pre-trained ML models [64, 65].

4.2 Conclusion

Our work showcases the importance of combining machine learning and cognitive science approaches for mutually advancing both fields. In particular, we believe that the methodologies adopted in this work have the potential to greatly advance basic research on naturalistic representations in cognitive science and improve machine representations and drive them toward human alignment.

Acknowledgments and Disclosure of Funding

This work was supported by a grant from the John Templeton Foundation to TLG and an NDSEG Fellowship to TRS.

Supplementary materials

A Stimuli and data

A.1 Stimuli

Throughout this work we considered five stimulus datasets across three different modalities: images, audio, and video. For images, we considered three datasets of common objects introduced in [36], namely, animals, furniture and vegetables, and each dataset contained 120 images. As for the audio dataset, we used the RAVDESS corpus ([53], released under a CC Attribution license), which consists of emotionally-neutral sentences spoken by 24 US American actors to convey a specific target emotion. To construct the 1,000-recording dataset, we selected 3 emotions per speaker per sentence. We randomly omitted 104 emotional stimuli and included all 96 neutral recordings (the dataset only contains 2 neutral recordings per speaker per sentence). We selected another subset of 100 stimuli for the similarity study by balancing the different emotions ($\sim$13 recordings per emotion). Finally, for the video dataset, we considered the Mini-Kinetics-200 dataset [58] (released under a CC BY 4.0 International License), which contains a large set of short video clips of human activities from 200 activity classes. Specifically, we focused on the validation split, which contains 5,000 videos in total. To construct our 1,000-video dataset, we sampled 5 random videos from each of the 200 activity categories. The 100-video subset used in the similarity experiment was then generated by sampling 100 random stimuli from the 1,000 list.

A.2 Code and data availability

A view-only anonymous link is provided to the public, containing all the data collected for this project during the review stage [6]. It includes the new human behavioral data, the computational experiments with machine learning models, and all the necessary analyses scripts for producing the results. Additionally, the repository includes the Dallinger/PsyNet source codes for reproducing the behavioral experiments. Finally, we present an interactive visualization [7] for exploring the similarity between stimuli as experienced by humans and different methods reported in the paper.

B Behavioral Paradigms

B.1 Participants

All participants provided informed consent under an approved IRB protocol by either the Institutional Review Board (IRB) at Princeton University (application 10859) or the Max Planck Ethics Council (application 2021_42) prior to participating in our studies. Participants were recruited on Amazon Mechanical Turk[8] (AMT), an online crowd-sourcing platform for worker recruitment. To help recruit reliable participants, we required that participants: a) have at least 99% approval rate on prior AMT tasks, and b) have completed no less than 5,000 tasks on AMT. We further required that participants are at least 18 years of age and that they reside in the United States. Participants were paid a fair wage of 9-12 USD per hour. Overall, $N = 1,823$ participants completed our studies and the self-reported ages ranged from 20–78 ($M = 41.1, SD = 11.4$). The total amount spent on participant compensation was 6,319 USD. The exact number of participants for each of the 12 new behavioral experiments is reported in Table 1.

B.2 Implementation

All behavioral experiments were implemented using the Dallinger[9] and PsyNet [44] frameworks. Dallinger is a modern tool for experiment hosting and deployment which automates the process of participant recruitment and compensation by integrating cloud-based services such as Heroku[10] with

[6] **Code and data:** `https://osf.io/dxzg7/?view_only=18a49289a66640e8b8abb8edae378149`

[7] **Interactive plots:** `https://words-are-all-you-need.s3.amazonaws.com/index.html`

[8] `https://www.mturk.com/`

[9] `https://dallinger.readthedocs.io/en/latest/`

[10] `https://www.heroku.com/`

Table 1: Behavioral experiment summary table.

Modality	Paradigm	Respect	Total stimuli	Trials per participant	Section	N	Pre-screening
Images	Tags	Animals	120	60	3.1	56	LX
Images	Tags	Furniture	120	60	3.1	58	LX
Images	Tags	Vegetables	120	60	3.1	57	LX
Audio	Similarity	Emotions	100	85	3.2	252	HT
Audio	Captions	Emotions	1,000	50	3.2	151	HT, LX
Audio	Tags	Emotions	1,000	50	3.2	217	HT, LX
Audio	Similarity	Speaker	100	85	3.2	257	HT
Audio	Captions	Speaker	100	50	3.2	39	HT, LX
Audio	Tags	Speaker	100	50	3.2	35	HT, LX
Video	Similarity	Activities	100	85	3.3	284	HT
Video	Captions	Activities	1,000	50	3.3	196	HT, LX
Video	Tags	Activities	1,000	50	3.3	221	HT, LX

Note. 'N' denotes the number of participants included in the analysis; 'LX' denotes the LexTALE English proficiency pre-screening task; 'HT' denotes the headphone test.

online crowd-sourcing platforms such as AMT. PsyNet is a novel experiment design framework that builds on Dallinger and allows for flexible specification of experiment timelines as well as providing support for a wide array of tasks across different modalities (visual, auditory and audio-visual). Participants interact with the experiment through their web-browser, which in turn communicates with a backend Python server responsible for the experiment logic.

B.3 Pre-screening

A common technique for filtering out participants that are likely to deliver low-quality responses, as well as automated scripts (bots), is to implement pre-screening tasks prior to the main part of each experiment. Failing the pre-screening tasks results in early termination of the experiment. Nevertheless, participants are still compensated for their time regardless of whether they fail or succeed on a pre-screener to ensure fair compensation. The role of pre-screeners in our studies was to realize two main criteria for data quality, namely, a) to be able to collect high-quality text descriptors, and b) to ensure that participants are able to inspect the target stimuli properly (in particular the audio component in prosody and videos). To do this, we implemented two pre-screening tasks, an English proficiency test and a standardized headphone test (used only for audio and video experiments). Table 1 provides details on which pre-screeners were used in each of the behavioral experiments.

Figure 8: Example trial from the LexTALE pre-screening task [66].

English proficiency test. To test participants' English proficiency, we used LexTALE, a lexical decision task developed in [66]. In each trial, participants were briefly presented (1 second) with either a real English word or a made up word that do not exist. Participants were instructed to guess whether the word was real or not. A total of 12 trials (half of them being real words) were presented, and 8 of them needed to be correct for the participant to pass. The presented words were: hasty, fray, stoutly, moonlit, scornful, unkempt, mensible, kilp, plaintively, crumper, plaudate, alberation. An example trial is shown in Figure 8.

Figure 9: Example trial from the headphone pre-screening test [67].

Headphone test. We used the headphone test developed by Wood et al. [67], which is used as a standard pre-screener for high-quality auditory psychophysics data-collection procedures [68]. The test is designed to ensure that the participants are wearing headphones and are able to perceive subtle differences in volume. The task consists of a forced choice task, in which three consecutive tones are played, and the participant has to identify which of them is the quietest. Crucially, these tones are constructed to exhibit a phase cancellation effect when not using headphones, and therefore making it difficult for non-headphone users to identify the quietest tone. Participants had to answer 4 out of 6 trials correctly to pass this test. An example trial is shown in Figure 9.

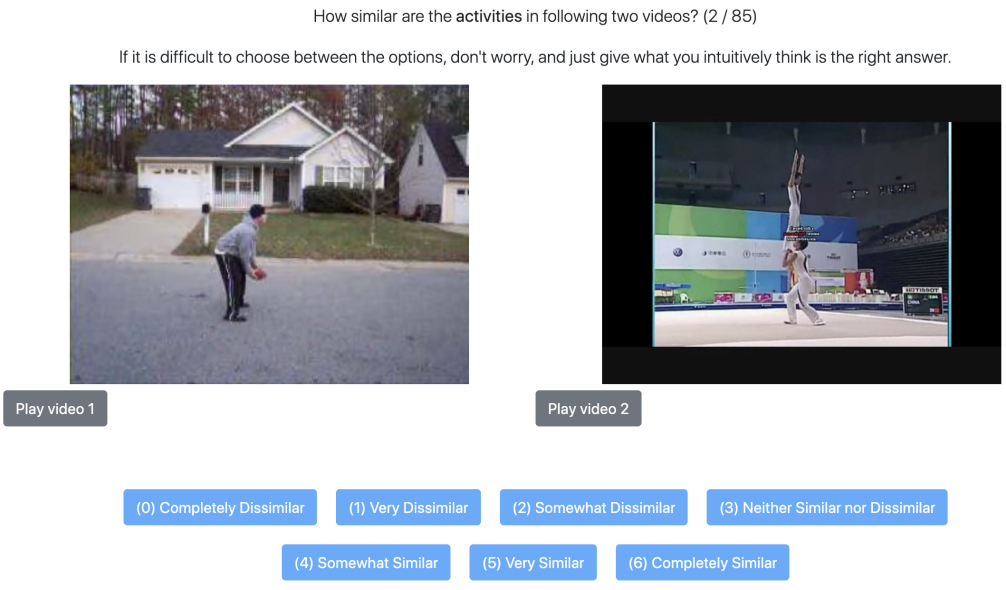

Figure 10: Screenshot from the similarity judgment task over video pairs.

B.4 Similarity judgments

In the present work, we collected similarity judgments across audio and video datasets. Each dataset comprised of 4,950 unique pairs corresponding to the number of unordered subsets that contain two distinct objects (i.e., excluding self-similarity), within a set of 100 stimuli. We did not collect similarity judgments over the three datasets of images, as these were provided in [36] (and used here with permission). The experiments proceeded as follows: upon completion of the consent form and the pre-screening tasks, participants received instructions regarding the main experiment:

> **Audio (emotion-respect).** In this experiment we are studying how people perceive emotions. In each round you will be presented with two different recordings and your task will be to simply judge how similar are the emotions of the speakers.
>
> **Audio (speaker-respect).** In this experiment we are studying how people perceive speaker voices. In each round you will be presented with two different recordings and your task will be to simply judge how similar are the speakers' voices irrespective of their emotions.
>
> **Video.** In this experiment we are studying how people perceive activities. In each round you will be presented with two different videos and your task will be to simply judge how similar are the activities in them.

The instructions then continued as follows:

> You will have seven response options, ranging from 0 ('Completely Dissimilar') to 6 ('Completely Similar'). Choose the one you think is most appropriate. Note: no prior expertise is required to complete this task, just choose what you intuitively think is the right answer.
>
> The quality of your responses will be automatically monitored, and you will receive a bonus at the end of the experiment in proportion to your quality score. The best way to achieve a high score is to concentrate and give each round your best attempt.
>
> The experiment will begin now. You will take up to 85 rounds where you have to answer this question. Remember to pay careful attention in order to get the best bonus!

As described in the instructions, in each trial, participants rated the similarity between a pair of sounds (how similar are the emotions/voices of the two speakers?) or videos (how similar are the activities in the following two videos?) on a scale ranging from 0 (completely dissimilar) to 6 (completely similar) (Figure 10). Overall, participants completed 85 trials on a random subset of the possible pairs. To further motivate participants to provide good responses, we gave them an additional performance bonus for providing consistent data. Among the 85 trials, 5 trials were repeated for consistency checking. The responses were converted into a performance score by computing the Spearman correlation between the original and repeat ratings. Perfect scores resulted in a 10 cent bonus.

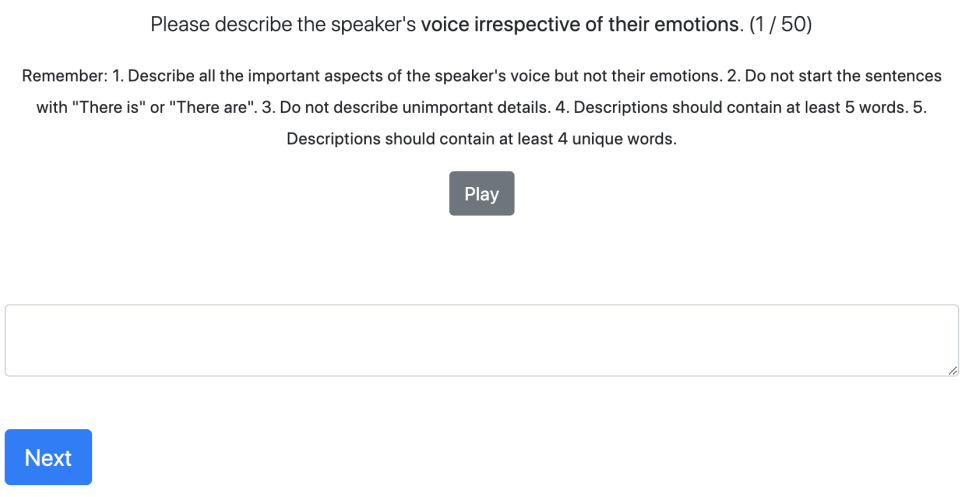

Figure 11: Screenshot from the speaker-respect audio captioning task.

B.5 Captions

We collected free-text captions for the video and audio datasets. Captions for the image datasets were already collected in [38]. After completing the consent form and pre-screening tests, participants received the following instructions:

> **Audio (emotion-respect).** In this experiment we are studying how people describe emotions. You will be presented with different recordings of speakers and your task will be to describe their emotions. In doing so, please keep in mind the following instructions
>
> - Describe all the important aspects of the recording.
>
> **Audio (speaker-respect).** In this experiment we are studying how people describe speaker voices. You will be presented with different recordings of speakers and your task will be to describe their voices irrespective of their emotions. In doing so, please keep in mind the following instructions
>
> - Describe all the important aspects of the speaker's voice but not their emotions.
>
> **Video.** In this experiment we are studying how people describe activities in videos. You will be presented with different videos of activities and your task will be to describe their content. In doing so, please keep in mind the following instructions
>
> - Describe all the important activities in the video.

As well as the following guidelines adapted from [38]:

> - Do not start the sentences with "There is" or "There are".
> - Do not describe unimportant details.
> - You are not allowed to copy and paste descriptions.
> - Descriptions should contain at least 5 words.
> - Descriptions should contain at least 4 unique words.
>
> Note: No prior expertise is required to complete this task, just describe what you intuitively think is important as accurately as possible.
>
> The quality of your captions will be monitored automatically and providing low quality and repetitive responses could result in early termination of the experiment and hence a lower bonus.
>
> You will describe up to 50 recordings.

These guidelines were enforced to ensure that participants deliver sufficiently informative captions that are not repetitive. In each trial of the main experiment, participants described a single audio (please describe the emotions/voice of the speaker) or video stimulus (please describe the activity in the video) (Figure 11). Overall, participants described up to 50 randomly presented stimuli. To filter out bad participants that tend to deliver repeated responses, in each trial (excluding the first 4 trials) we computed the mean edit distance between their current response and all previous responses that they previously provided using the `partial_ratio` function in `thefuzz`[11] Python package for fuzzy string matching. This function returns for a pair of input strings a matching score between 0 and 100 (identical strings). Early termination was enforced if the mean response matching score was above 80. The idea here was to prevent participants from copying and pasting the same response over and over again (or varying it only a little bit).

B.6 Tags

For the image, audio, and video datasets, we collected tag data, i.e., concise labels that describe the salient features of a stimulus. To do so, we developed a novel tag mining paradigm in which each stimulus was treated as a separate "chain" (see Figure 1B in the paper). When the stimulus was

[11] `https://github.com/seatgeek/thefuzz`

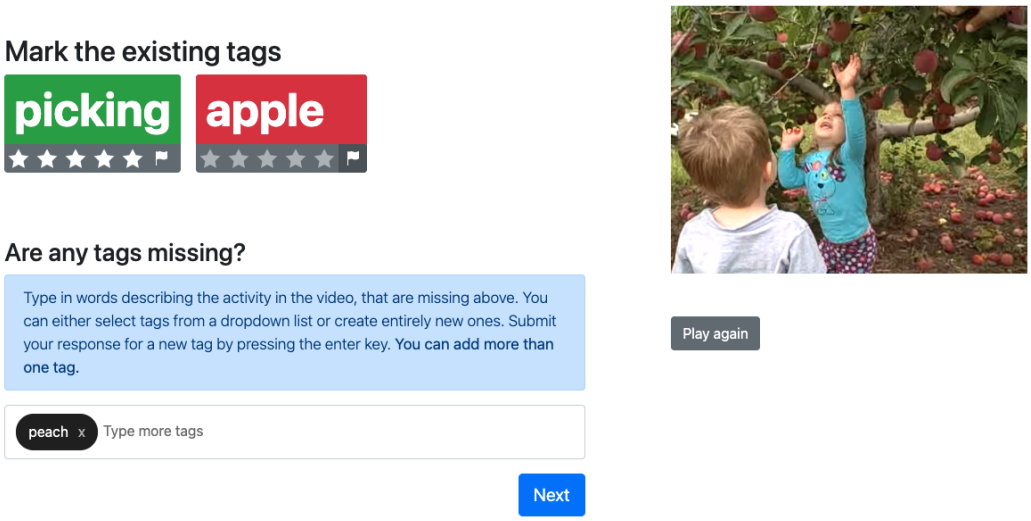

Figure 12: Screenshot from the tag mining task for videos. The tag "picking" received 5 stars (very relevant), whereas the tag "apple" is flagged (marked as irrelevant)

presented for the first time, the participant was asked to provide at least one tag. For the following iterations, we sequenced participants so that each of them had to rate the tags provided by participants from the previous iterations within the same chain. The rating was either choosing between one (not very relevant) to five stars (very relevant), or marking the tag as completely irrelevant by using the flag icon (see Figure 12). Participants could optionally introduce new tags that will subsequently be presented to other participants assigned to the same chain. Participants could only provide tags that were not already present, and they had to be in lower-case letters. If participants used two or more white spaces (i.e. three or more words), a pop-up window appeared asking if such spaces were really necessary (to discourage frequent use of long word combinations). This process continued for at least 10 iterations, after which we checked at each consequent iteration whether the chain was "full". We considered a chain to be full if its latest iteration had at least 2 tags that were rated at least 3 times and had a mean rating of 3 stars. If a chain was not full after 20 iterations, we stopped collecting further iterations. Since each experimental batch lasted for a fixed duration of less than one day, in some cases we did not complete all chains, and a few chains had fewer iterations (3 for vegetables, 6 for animals and 2 for furniture, out of 120 chains each). Our experiment incentivized participants to provide new tags by paying them a performance bonus of 0.01 USD if their tags were up-voted (i.e., not flagged) by other participants. Nevertheless, if two or more tags of the same participant were flagged by others, the participant was excluded (the participant received a warning after the first flag).

After accepting the consent form and passing the pre-screening tasks, participants received introductory instructions regarding the main experiment:

> **Images**. Rate & Tag animals/furniture/vegetables! Thanks for participating in this game! In this game you will:
>
> - Watch images of animals/furniture/vegetables.
> - Rate tags that other players have given.
> - Add new tags that you think are missing.
>
> **Audio (emotion-respect)**. Rate & Tag emotions! Thanks for participating in this game! In this game you will:
>
> - Listen to a speech fragment and focus on the emotional content of the recording.
> - Rate tags that other players have given.
> - Add new tags that you think are missing.
>
> **Audio (speaker-respect)**. Rate & Tag speakers' voice! Thanks for participating in this game! In this game you will:

- Listen to a speech fragment and focus on the speakers' voice in the recording.
- Rate tags that other players have given.
- Add new tags that you think are missing.

Video. Rate & Tag activities! In this game you will:

- Watch a video and focus on the activities happening.
- Rate tags that other players have given.
- Add new tags that you think are missing.

Participants then received further instructions regarding the rules of the game

> **Images**. After watching the animal/furniture/vegetable you will see tags given by other players that describe the animal/furniture/vegetable. You should rate the relevance of each tag by clicking the appropriate amount of stars (1 star not very relevant, 5 stars very relevant). If you think that the tag is a mistake or completely irrelevant, you should flag it by clicking the flag icon. If you are the first person seeing this animal/furniture/vegetable, you may see no previous tags. You can also add your own tag that is relevant to describe the animal/furniture/vegetable. Your tag will then be rated by other players who are playing the game simultaneously.
>
> **Audio (emotion-respect)**. After listening to the recording, you will see tags given by other players that describe the emotions in the speech fragment. You should rate the relevance of each tag by clicking the appropriate amount of stars (1 star not very relevant, 5 stars very relevant). If you think that the tag is a mistake or completely irrelevant, you should flag it by clicking the flag icon. If you are the first person listening to this speech sample, you may see no previous tags. You can also add your own tag that is relevant to describe the emotions in the speech fragment. Your tag will then be rated by other players who are playing the game simultaneously.
>
> **Audio (speaker-respect)**. After listening to the recording, you will see tags given by other players that describe the speakers' voice irrespective of their emotions. You should rate the relevance of each tag by clicking the appropriate amount of stars (1 star not very relevant, 5 stars very relevant). If you think that the tag is a mistake or completely irrelevant, you should flag it by clicking the flag icon. If you are the first person listening to this speech sample, you may see no previous tags. You can also add your own tag that is relevant to describe the speakers' voice. Your tag will then be rated by other players who are playing the game simultaneously.
>
> **Video**. After watching the video, you will see tags given by other players that describe the activities in the video. You should rate the relevance of each tag by clicking the appropriate amount of stars (1 star not very relevant, 5 stars very relevant). If you think that the tag is a mistake or completely irrelevant, you should flag it by clicking the flag icon. If you are the first person watching this video, you may see no previous tags. You can also add your own tag that is relevant to describe the activities in the video. Your tag will then be rated by other players who are playing the game simultaneously.

Finally, participants received the following guidelines regarding the tag input and the bonus scheme:

> **Keep tags short.** A word like "green grass" should rather be submitted as "green" and "grass", whereas a compound word such as "red wine" cannot be separated, since "red wine" means something different than just "red" and "wine".
>
> **Bonus rules.**
>
> - If the tag you provide gets rated as a relevant tag (i.e., not flagged) by other players
> - If your tag is unique and have not been introduced by others

Note: Simply writing many and irrelevant tags is not a good idea because other players might flag your tag. Your experiment will terminate early if there are too many red flags!

Please try to use a variety of words to describe the animal / furniture / vegetable / emotion in the speech fragment / speakers' voice / activities in the video, and use the entire star rating scale for your responses.

C Prediction methods

We used two main types of methods to predict human similarity judgments. The first class ("Embedding methods", described in section C.1) make use of pre-trained embedding models. In the second class of models ("Embedding-free methods", described in the section C.2) simple feature extraction techniques are used instead of pre-trained deep learning models. Figure 1A depicts schematically an overview of all prediction methods that we used.

C.1 Embedding models

The embedding-based methods use various embeddings and deep learning representations to predict human similarity judgments. These methods could be further split into three groups based on the kinds of input data they process, namely if they use a single sensory modality that is either image, audio or video ("unimodal baseline"; see subsection C.1.1), or use text that is either tag or captions ("text embeddings"; see subsection C.1.2), or use both ("multimodal baseline"; see subsection C.1.3). In addition, we also tested the performance of "stacked" representations (reported also in subsection C.1.3), where the sensory and textual embedding of a select number of models were concatenated into a single long embedding. Overall, the computation time of embedding methods took about two weeks on an x1.16xlarge Amazon Web Services instance with 64 vCPUs and 976 GiB of memory.

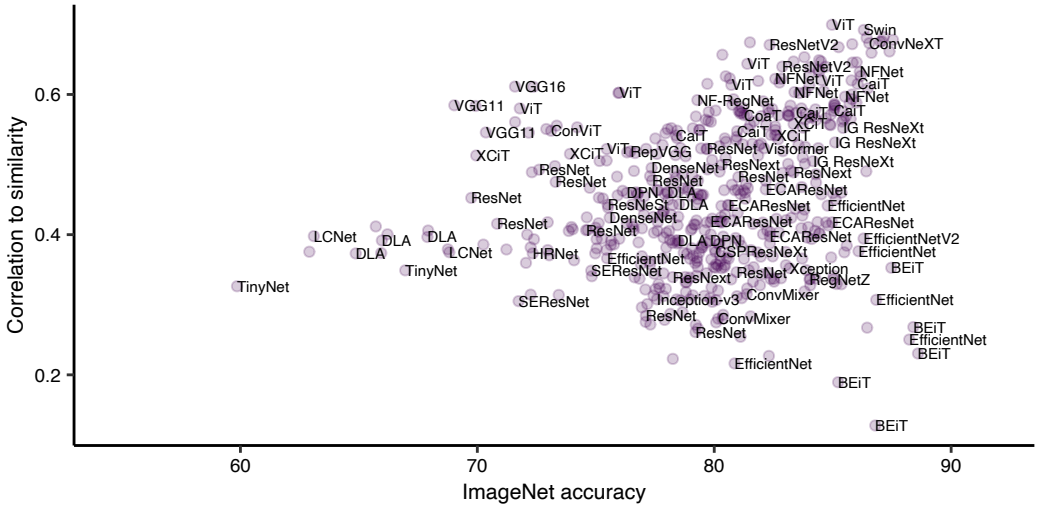

Figure 13: Correlation with human similarity judgments as a function of ImageNet accuracy for the various image baseline models.

C.1.1 Unimodal baseline methods

Unimodal baseline - Image models. We used 560 pre-trained models from the Pytorch Image Models (`timm`) repository [45]. We chose this repository as it contains an extensive and highly diverse set of pre-trained models in terms of architecture backbones, model sizes, and training sets. The repository includes models published from 2014 to 2022 that use various training sets (such as ImageNet1k, ImageNet21k, Instagram, etc.), training procedures objectives (e.g. pre-training, fine-tuning, self-supervision, weak supervision, etc.) and architectures (e.g. VGG, ResNet, Inception,

Table 2: All 42 image baseline models occurring in the top 60 best models reported in Figure 3A.

	Model name	Average score	SD score	Top 1 accuracy	Number of parameters (M)
1	Swin	0.65	0.07	81.52	23.37
2	ConvNeXT xlarge in22k	0.64	0.07	N/A	348.15
3	NF-ResNet-50	0.62	0.04	80.65	23.51
4	NFNet l0	0.61	0.08	82.75	32.77
5	ResNetV2 152x4 bitm in21k	0.60	0.11	N/A	928.34
6	NF-RegNet b1	0.59	0.05	79.29	9.26
7	CLIP RN50 (image+text)	0.59	0.07	N/A	102.01
8	VGG16 batchnorm	0.58	0.11	73.35	134.27
9	VGG19 batchnorm	0.58	0.11	74.21	139.58
10	ViT tiny r s16 p8 384	0.58	0.12	75.95	6.16
11	ResMLP big 24 distilled 224	0.57	0.07	83.59	128.37
12	Twins-SVT small	0.57	0.06	81.68	23.55
13	Twins-PCPVT small	0.57	0.04	81.09	23.59
14	VGG13 batchnorm	0.57	0.11	71.59	128.96
15	CaiT xxs36 384	0.57	0.04	82.19	17.18
16	VGG11 batchnorm	0.57	0.10	70.36	128.77
17	gMLP s16 224	0.56	0.06	79.64	19.17
18	PIT xs 224	0.56	0.03	78.19	10.23
19	DeiT tiny patch16 224	0.56	0.03	72.17	5.52
20	ConViT tiny	0.56	0.03	73.11	5.52
21	CoaT tiny	0.55	0.04	78.43	5.35
22	gMixer 24 224	0.55	0.05	78.04	24.34
23	CLIP RN50 (text)	0.55	0.02	N/A	102.01
24	XCiT tiny 24 p16 384 dist	0.55	0.04	82.57	11.92
25	IG ResNeXt 101 32x48d	0.53	0.13	85.44	826.36
26	Visformer small	0.52	0.02	82.11	39.45
27	RepVGG b3g4	0.52	0.11	80.21	81.26
28	CLIP RN50 (image)	0.50	0.11	N/A	102.01
29	JXNesT tiny	0.50	0.07	81.42	16.67
30	ECAResNet light	0.47	0.14	80.45	28.11
31	DenseNet 121	0.47	0.12	74.74	6.95
32	PNASNet 5 large	0.47	0.05	82.79	81.74
33	ResNext 50 32x4d	0.47	0.11	77.62	22.98
34	ResNeSt 269e	0.46	0.12	84.52	108.88
35	GerNet S	0.46	0.10	76.91	6.25
36	SENet 154	0.46	0.09	81.31	113.04
37	FBNet 100	0.45	0.10	75.13	3.59
38	ResNet blur 60	0.45	0.10	79.30	23.51
39	HRNet w64	0.44	0.14	79.47	126.01
40	HardCoRe-NAS f	0.44	0.05	78.10	6.92
41	NASNet large	0.42	0.07	82.63	84.72
42	DLA 169	0.42	0.12	78.69	52.36

Note. Performance accuracy on ImageNet was based on [45] and was not available for all models.

Transformer, etc.). The repository also reports various evaluation metrics for each model (e.g. their ImageNet performance).

For each model, we computed the embedding from the last layer (typically before the final softmax layer; see below and Figure 14 for a preliminary analysis for the effect of layer depth). We then computed the cosine similarity between pairs of embedding vectors to produce a similarity matrix. The entire list of the performance of all models is detailed in the OSF repository associated with this project [12]. Table 2 presents additional details for the top 42 image baseline models in Figure 3A including their average score (correlation to human judgments) across the three image datasets, the

[12] `https://osf.io/dxzg7/?view_only=18a49289a66640e8b8abb8edae378149`

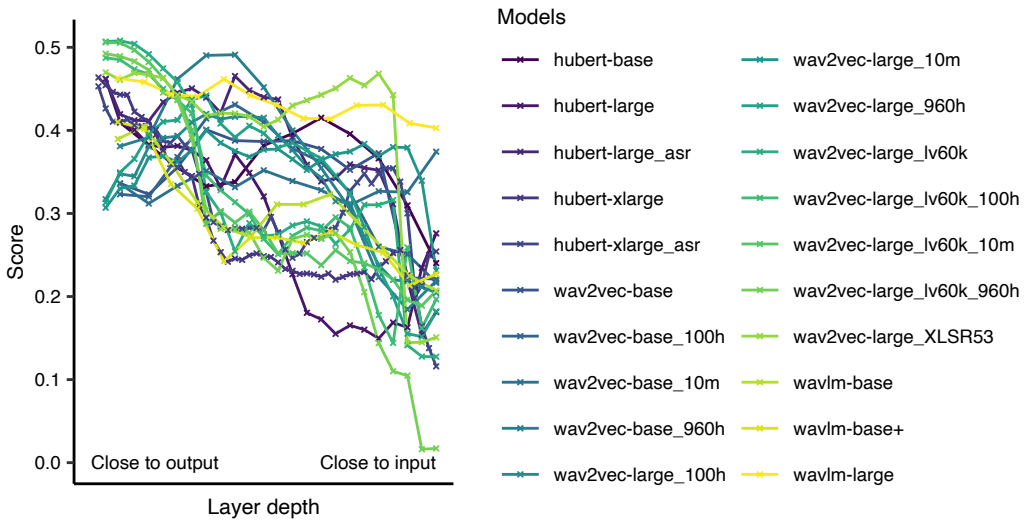

Figure 14: Scores for individual layers of audio models scaled to the total number of layers. Models are colored by their meta architecture.

standard deviation (SD) of this score (across datasets, repeated runs and available model parameters in [45]), their ImageNet accuracy, and their number of trainable parameters.

Figure 3C shows the correlation to human similarity as a function of the number of parameters for all 569 models (uni- and multimodal baselines). In general, we found that models that have more parameters perform better (Figure 3C). Plotting all the embedding technique correlations against the number of training parameters of their respective models showed statistically significant positive correlation ($r = 0.41, p < 0.001$). However, one possible explanation for this could be the improved performance of newer models, which typically have more parameters, on various computer vision tasks. To test this, we computed the performance (i.e., correlation with human similarity) of the various models as a function of their accuracy on ImageNet [50] - which was provided in [45] for all models except for CLIP (whose implementation came from a different repository; see Section C.1.3) and those architectures ending with the suffixes 'in22k' and 'in21k' (which signify that the model was pre-trained for, and comes with the classification head for, ImageNet22k and ImageNet21k, respectively). This analysis is summarized in Figure 13. We found a positive correlation between the two metrics ($r = 0.26, p < 0.001$), though with some clear exceptions. For example, the vision transformer BEiT [51] and the convolutional architecture EfficientNet [69] achieved high accuracy on ImageNet but performed poorly on human data. On the other hand, the vision transformer Swin [48] and the convolutional architecture ConvNext [47] both performed well on ImageNet and human similarity. This suggests that architecture and number of parameters are better predictors of similarity judgments than performance on ImageNet. Further analysis is required to determine what kind of architectural components actually contribute to more human-like performance [27].

Unimodal baseline - Audio models. We used all pre-trained wav2vec 2.0 [55] and HuBERT [13] models available in `torchaudio` [70]. We also extracted embeddings from WavLM [71] and data2vec audio models [72]. Furthermore, we used additional wav2vec 2.0 and HuBERT models that were either specialized on emotion recognition or speaker identification [73–75]. The performance of HuBERT, wav2vec 2.0, and WavLM models is shown in Figure 5A-B. Additional details about the models are displayed in Table 3.

In addition, we explored the correlation between the audio models and human similarity data as a function of the layer in the model. Earlier literature has suggested that similarity to human representations may depend on the layer of the model [25, 28, 76]. We expected that the layers closer to the input of the model (where the representation is more low-level) to be less predictive. In general, we found that this was the case (Figure 14). In some variants of wav2vec, however, intermediate representations performed better, possibly due to the misalignment of the training task of wav2vec with the emotion/speaker task. This analysis confirms the choice we made in the paper to mostly use the last two layers of the models. Preliminary analysis of the image and video models also explored

Table 3: All audio baseline models used in the analysis.

	Model name	Emotion correlation	Speaker correlation	Number of parameters (M)
1	wav2vec 2.0 lv60k (100h)	0.49	0.45	317
2	wav2vec 2.0 lv60k (960h)	0.49	0.42	317
3	wav2vec 2.0 lv60k	0.51	0.40	317
4	wav2vec 2.0 lv60k (10m)	0.51	0.39	317
5	HuBERT xlarge ASR	0.45	0.44	1000
6	HuBERT xlarge	0.46	0.42	1000
7	HuBERT large ASR	0.46	0.41	300
8	wav2vec 2.0 large XLSR53	0.47	0.38	317
9	HuBERT large	0.46	0.38	300
10	wav2vec 2.0 (Audeering, emotion)	0.49	0.35	317
11	HuBERT base	0.41	0.41	90
12	WavLM large	0.46	0.34	316.62
13	HuBERT base (superb, emotion)	0.42	0.36	90
14	HuBERT base (superb, speaker)	0.42	0.36	90
15	WavLM base+	0.41	0.36	94.70
16	wav2vec 2.0 base (960h)	0.38	0.37	95
17	WavLM base	0.39	0.34	94.70
18	wav2vec 2.0 base	0.34	0.33	95
19	wav2vec 2.0 base (10m)	0.34	0.32	95
20	wav2vec 2.0 base (superb, emotion)	0.34	0.31	95
21	wav2vec 2.0 base (superb, speaker)	0.34	0.31	95
22	wav2vec 2.0 base (100h)	0.32	0.32	95
23	HuBERT large (superb, emotion)	0.29	0.35	300
24	HuBERT large (superb, speaker)	0.29	0.35	300
25	wav2vec 2.0 large (100h)	0.32	0.31	317
26	wav2vec 2.0 large (superb, emotion)	0.31	0.32	317
27	wav2vec 2.0 large (superb, speaker)	0.31	0.32	317
28	wav2vec 2.0 large (960h)	0.31	0.30	317
29	wav2vec 2.0 large (10m)	0.31	0.29	317
30	data2vec audio large (960h)	0.31	0.17	313.28
31	data2vec audio base (100h)	0.23	0.17	313.28
32	data2vec audio large (100h)	0.23	0.13	313.28
33	data2vec audio large (10m)	0.21	0.14	313.28
34	wav2vec 2.0 (SpeechBrain, emotion)	0.11	0.19	95
35	data2vec audio base (960h)	0.16	0.12	93.16
36	data2vec audio base (10m)	0.15	0.12	93.16

different layers, but the results were similar to those we presented in audio, and are therefore not reported here.

Unimodal baseline - Video models. We extracted embeddings from the 'Slow' (a 3D ResNet; see [59]), Slowfast (a 2-path model with one path capturing semantics and the other capturing fine details; see [59]), and X3d (a model that initially start as a simple 2D image classifier but is expanded in several axes; see [60]) architectures implemented in `pytorchvideo` [77]. All video models were pre-trained on the Kinetics-400 dataset [78]. The performance of the models is displayed in Figure 6B. Numeric correlation values are detailed in Table 4 along with model accuracy (Top1 and Top5) on Kinetics-400, and the number of parameters in each model. The accuracies and parameter counts are listed as reported in [77]. As with previous modalities, the number of parameters appears to be positively correlated with correlation to human similarity.

C.1.2 Text embedding methods

Caption text embedding. We used four large pre-trained language models from HuggingFace [79] to compute embeddings of the captions collected for each dataset: 'bert-base-uncased', 'deberta-xlarge-mnli', 'sup-simcse-bert-base-uncased', and 'sup-simcse-roberta-large'. SimCSE is a pre-training

Table 4: All video baseline models used in the analysis.

	Model name	Correlation	Kinetics-400 Top1 Acc	Kinetics-400 Top5 Acc	Number of parameters (M)
1	Slowfast r50	0.65	76.94	92.69	34.57
2	Slowfast r101	0.64	77.90	93.27	62.83
3	Slow r50	0.61	74.58	91.63	32.45
4	X3d M	0.53	75.94	92.72	3.79
5	X3d S	0.49	73.33	91.27	3.79
6	X3d XS	0.48	69.12	88.63	3.79

procedure that uses semantic entailment in a contrastive learning objective [24]. According to BERTScore [80], the latter three models are ranked in the top 40 models by correlation with human evaluations on certain tasks, with 'deberta-xlarge-mnli' ranked first. However, in our experiments, we found that embedding similarity computed from 'sup-simcse-roberta-large' has the highest correlation with human similarity judgments out of the four models. For SimCSE-based models, we used representations from the (final) embedding layer (where the SimCSE contrastive objective is actually applied). For the other two models, we computed embeddings from every layer, but restricted the main analysis to embeddings from the penultimate layers. This was to be consistent with our procedure for modality-specific embedding models. Since there are multiple captions per stimulus, an aggregation procedure had to be applied to produce a single embedding vector for each stimulus. In our main analysis, for each stimulus, we extracted the embedding for each associated caption and averaged these embeddings together before computing cosine similarity between the mean embeddings. We also tried an alternative approach of concatenating the captions together into a single paragraph, which we then passed through the language models to compute a single embedding per stimulus. We found that this did not consistently improve performance and in many cases even decreased it, though we note that we did not experiment with different permutations of the concatenated captions, nor did we extensively study other ways to combine them together. Future work could explore other techniques for pre-processing captions and aggregating representations from multiple captions in ways that would improve correlation with human similarity judgments.

Tag text embedding. We experimented with several algorithms for computing similarity between sets (or multi-sets) of tags. The algorithms described in this section all involve using ConceptNet NumberBatch (CNNB) [23] as the embedding backbone for turning discrete tags into continuous vector representations. For each stimulus, we took the tags remaining in the final iteration, and tested whether they were found in the dictionary for our embedding model. If a tag was not found and if it contained no spaces, we tried to correct the spelling before trying to look it up in the dictionary again. If a tag contained spaces, we would split it into individual words, correct their spelling, and average together the embedded representations of those words that were found in the dictionary. Tags that were not found even after spelling correction and splitting were excluded from the set and did not contribute to the final representation. For the methods marked '(no split)' we did not split multi-word tags, instead we just excluded multi-word tags that were not found in the embedding model dictionary. In the following, we describe the different techniques used to generate predictions based on tag embeddings.

Tags CNNB overlap. For each pair of stimuli, we counted the number of 'almost identical' tag embeddings, defined as every respective element of the two embeddings being less than a certain threshold apart (in our case, this threshold was 0.1). We then set similarity for that pair of stimuli to be this count, i.e. the number of 'almost identical' tags, normalized by the total number of tags across the respective two sets.

Tags CNNB quantized. This method involves quantizing tags using cosine similarity to find the number of unique tags. For each pair of stimuli, we counted the number of tags assigned to the first stimulus that had cosine similarity greater than a certain threshold (in our case, this threshold was 0.7) to at least one tag of the second stimulus (call this value N_A) and vice-versa (N_B). The minimum of these two values is the number of unique, shared tags between the two sets ($\min(N_A, N_B)$). The total number of unique tags across the two sets is then the total number of tags in each set ($T_A + T_B$) minus the maximum number of shared tags ($\max(N_A, N_B)$). We compute similarity as the ratio of the number of unique, shared tags to the total number of unique tags, $S_{AB} = \frac{\min(N_A, N_B)}{T_A + T_B - \max(N_A, N_B)}$. For

example, suppose the two sets of tags are $A : \{a, b, c, g\}$ and $B : \{a, b, d, e\}$, so $T_A = T_B = 4$, and that a, c have cosine similarity of 0.8. The number of tags from set A found in set B is $N_A = 3$, and those from B found in A is $N_B = 2$. The number of unique, shared tags is $\min(N_A, N_B) = 2$ (since $\{a, b, c\}$ can be represented by $\{a, b\}$), and the total number of unique tags is $4 + 4 - 3 = 5$ (since $\{a, b, c, g, a, b, d, e\}$ can be represented by $\{a, b, d, e, g\}$). The assigned similarity is then $S_{AB} = \frac{2}{5}$.

Tags CNNB mean. The set of tag embeddings for each stimulus were averaged together to form a single embedding assigned to the respective stimulus. We then computed cosine similarity on the embeddings of each pair of stimuli.

Tags CNNB mean (no split). Same as above, but without splitting multi-word tags (i.e. ones that contain spaces) during the embedding process.

All spelling corrections in these algorithms were performed using the Python package `pyspellchecker`[13], taking the top corrected recommendation returned by the spell checker in each case.

As a baseline experiment, we also tried randomly selecting a single high-rated tag per stimulus and computing cosine similarity on embeddings of those. As expected, we found that correlation with human similarity judgments was significantly lower with this method than with other methods and highly variable depending on random seed.

C.1.3 Multimodal and stacked analysis

We conducted several analyses to determine whether stacking (i.e. concatenating) representations from multiple modalities of the same stimulus would improve correlation with human similarity judgments. During our analyses, we mostly focused on stacking caption embeddings with modality-specific embeddings. For the image datasets, we started with the multimodal model CLIP [7] which jointly learns representations of images and text, with stimuli from both modalities embedded into the same space. As a result, users can extract comparable embeddings both for an image and its associated captions. Our CLIP models [7] were taken from the OpenAI CLIP repository[14]. We found that the stacked CLIP embeddings outperformed both of the single modality versions (text and images) of our CLIP embeddings. This motivated additional analyses to probe whether the improvement in performance came from CLIP's multi-modal pre-training procedure, or whether it was due to the combination of the two modalities. For each dataset, we stacked the five best sets of model embeddings with the best set of caption embeddings (those coming from 'sup-simcse-roberta-large'). However, unlike in the CLIP case, the modality-specific embeddings and the caption embeddings are not from the same space and had to be normalized at the feature-level (by subtracting mean from each feature and dividing by standard deviation) before being stacked (stacking performance was generally lower if the embeddings were not pre-normalized before stacking). These results are visualized in Figure 15, and in the case of images, stacking embeddings that were individually good indeed outperformed stacking the multi-modal embeddings from CLIP. In addition, stacked embeddings generally performed better than either the caption embeddings or modality-based embeddings did on their own (except in the reweighted case for animal images and videos, i.e. LT-CCV/(norm), where pure captions yielded the best results). This suggests that while text-based predictions perform on par with baseline models that do not rely on language, they do seem to explain non-identical sources of variance in human judgments. For the video dataset, we also considered stacking embeddings of the audio from the videos along with the associated video and caption embeddings. However, we found that the cosine similarity from audio embeddings on their own had almost no correlation with human similarity judgments and, as a result, did not add audio embeddings to the embedding stacking analysis for the video dataset.

C.2 Embedding-free models

In this work, we also conducted an additional evaluation of prediction models beyond embedding-based techniques (described in the previous section). Specifically, we compared the predictions of embedding-based models, which utilize deep learning representations, with those of traditional methods of text mining and audio feature extraction.

[13] `https://pyspellchecker.readthedocs.io/en/latest/`

[14] `https://github.com/openai/CLIP`

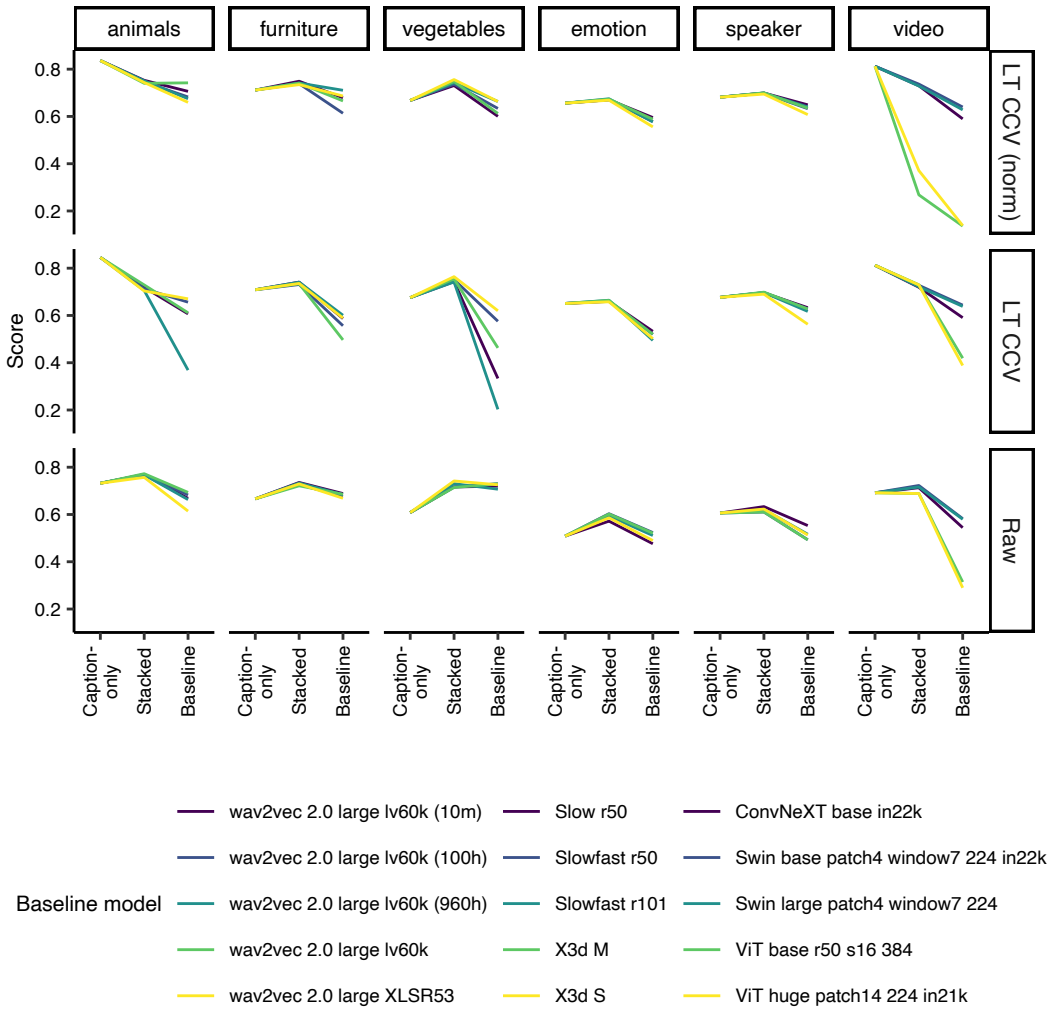

Figure 15: Scores for best baseline models, the best caption models and the stacked (i.e. concatenated) embeddings.

C.2.1 Tag/caption text analysis

The aim of the simple text analysis methods was to compare deep learning embeddings with traditional embedding-free techniques. Such techniques are particularly useful for low-resource languages or cross-cultural comparisons [62, 65], for which pre-trained models are lacking, as they work solely on the basis of the text itself.

The text analysis proceeded as follows. We first performed the following initial pre-processing steps

- For caption data, we concatenated all the captions describing the same stimulus into a single long "document."
- For tag data, we wanted to prioritize tags that appeared earlier in the tag-mining chains and were rated higher. To that end, we gathered all tags from all iterations and duplicated tags from a given iteration based on the ratings they received. For example, if the tag "tomato" received three stars, then we would add the repeated tokens "tomato, tomato, tomato" to the aggregated list ("document"). In a given iteration, flagged tags are removed, but if they are rated later, then they are included. The total number of repetitions per token is equal to the sum of all the stars they received in all iterations. As a result, each token is repeated multiple times, which we take into consideration in consequent analysis.

For the next steps, we used the Matlab text analysis toolbox [15]. Unless otherwise specified, we used default parameters for all functions. To generate similarity matrices, we applied the following methods:

Co-occurrence method. In this approach, we simply counted the number of repeated pairs of words in documents i and j and normalized by the total number of pairs. Formally, we use w_i to denote the word list of a document i. Let $w_{i,k}$ be the k-th word in the w_i list of words, and let $|w_i|$ denote the length of the list. We denote by $\delta(c, d)$ the indicator function that returns 1 if and only if the word c is identical to the word d, and 0 otherwise. We computed the co-occurrence score $S(w_i, w_j)$ according to the following formula:

$$S(w_i, w_j) = \frac{\sum_k \sum_l \delta(w_{i,k}, w_{j,l})}{|w_i||w_j|}$$

Co-occurrence-rep. This method was applied only to tags. We used an identical procedure to the *Co-occurrence* method, except that we did not separate the words within a tag as separate tokens and instead treated the entire tag (that may include multiple words) as a single token.

Rouge score. In this approach, similarity was estimated by computing the rouge score of the word lists associated with each pair of documents. The Rouge score was computed using `rougeEvaluationScore` [81].

The following methods make use of tokenized data and a pre-processing procedure that we found effective. Pre-processing was applied to both tag and caption data and tokenization was performed as follows:

- We separate all text into single words by applying the `tokenizedDocument` function.
- We added part of speech information using the `addPartOfSpeechDetails` function.
- We performed Lemmatization using the `normalizeWords` function.
- We erased punctuation from the token using the `erasePunctuation` function.
- We removed stopwords using the `removeStopWords` function.
- We removed words with less than two characters or more than 15 characters.
- We created a bag of words representation of each tokenized document using the `bagOfWords` function.
- We also removed words that were not present in more than two documents using the `InfrequentWords` function.

With the results of these pre-processing steps, we then computed similarity matrices based on the following methods:

bm25S. We used `bm25+` to compute similarity between documents [82] using Matlab's `bm25Similarity` function. This function represents TF-IDF-like retrieval functions used in document retrieval. We used a variant that has a normalization function that properly handles documents with a long list of words.

tf-idf-cosine. We computed pairwise cosine similarities between document pairs using the TF-IDF matrix derived from their word counts and Matlab's `cosineSimilarity` function.

C.2.2 Low-level audio features

We used OpenSMILE [83] to extract 88 standard low-level audio features from the eGeMAPS feature set [56] consisting of frequency, energy, and spectral parameters for the 1,000 acoustic stimuli. In Figure 5C, we used a single pitch-related (mean F0), energy-related (mean loudness), or spectral feature (spectral slope) to predict the pairwise similarity for the emotion- and speaker-respect. We show that mean F0 alone gives us a correlation of $r = 0.41$ to predict speaker-respect. This is already close to the correlation of $r = 0.49$ which we obtain for 88 z-scored features. The correlation for loudness and spectral slope is much lower, indicating that pitch plays an important role in describing the voice, potentially capturing perceived sex. For the emotion-respect data, we observe a reversed effect. Loudness correlates strongest ($r = 0.43$) with the pairwise similarity. This correlation is

[15] `https://mathworks.com/products/text-analytics.html`

similarly strong as for the 88 z-scored features ($r = 0.42$). The correlation for mean F0 and spectral slope is much lower. This indicates that loudness plays an important role in predicting emotion-respect, potentially capturing the arousal dimension. Together, the low-level analysis indicates a strong interaction between the respect of similarity and certain low-level features (but not all) that are selectively predictive of one kind of respect.

D Performance and visualization

D.1 Performance quantification

Our primary performance metric throughout this work was Pearson correlation (r) between predicted similarity matrices and ground truth human similarity judgments. In all cases, correlation was computed between the triangular off-diagonal matrices to avoid skewing results based on the diagonal (which consists of all ones) or by having duplicates (since the similarity matrix is symmetrical). To lower the risk of over-fitting, for our main results, we compared raw similarity scores produced via the methods discussed above without performing any additional transformations or optimization that would fit the predictions to the ground truth judgments. Several previous studies investigated improving correlation by applying and fine-tuning simple linear transformations to embedding vectors $z^T \mathbf{W} z$ where $\mathbf{W} = \text{diag}(w_1, \ldots, w_d)$ via a cross-validated ridge regression procedure that could be fit to ground truth similarity judgments. The parameters of the diagonal reweighting matrix $\mathbf{W}$ are fitted to a training subset of stimuli and used to predict similarity of pairs in a held-out validation set [36, 38]. To be consistent and make results comparable, here we report the results of performing this 6-fold cross-validated linear transformation (LT-CCV) on the model embeddings and datasets considered in this work. The analysis was carried out using the `RidgeCV` package from the `scikit-learn` Python library [84]. Results with both normalized ('LT CCV (norm)') and un-normalized ('LT CCV') regressors are shown in Figures 16 and 17; see `RidgeCV` documentation for details on normalization [16]. We see that the linear transformation does not consistently improve performance when applied to many of the modality-based embeddings, but it does frequently improve performance when applied to caption embeddings. As mentioned in Section 3.1, we also see that the reweighting in the case of images (Figures 16A and 17A) does not close the gap between the models and the IRR level, though it does appear to bring CLIP embeddings that incorporate text particularly close to it. As for the other modalities, we see that the audio speaker-respect and video datasets, the reweighted caption-based models seem to cross the IRR line, potentially reflecting a) the fact that IRR is only an approximate bound, and b) some degree of over-fitting.

D.2 Averaging and summary bar graphs

In the performance bar plots in Figure 3A, we averaged over the performance correlation of the three image datasets and over variants of the same architecture to reduce the total number of models in the figure. To give an overview of the different approaches, we averaged over all models using the same approaches in Figure 3B. To reduce the complexity of Figure 5A–B, we grouped the results by approach and architecture.

D.3 Inter-response reliability (IRR)

We compared the performance of the different prediction methods to the inter-response reliability (IRR) of participants, which serves as an approximate bound on performance. Following [36], we computed IRR for each human similarity matrix using the split-half correlation method. Specifically, we randomly split the similarity judgments associated with each stimulus pair into two groups and took their respective means. This process resulted in a pair of vectors with mean rating for each possible pair of stimuli. We then computed the Pearson correlation r between the two vectors to generate an estimate. We repeated this random splitting process 100 times and took the mean over all resulting values. To generate the final IRR value, we applied the Spearman-Brown correction [85] $nr/(1 + (n - 1)r)$ with $n = 2$.

[16] `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RidgeCV.html`

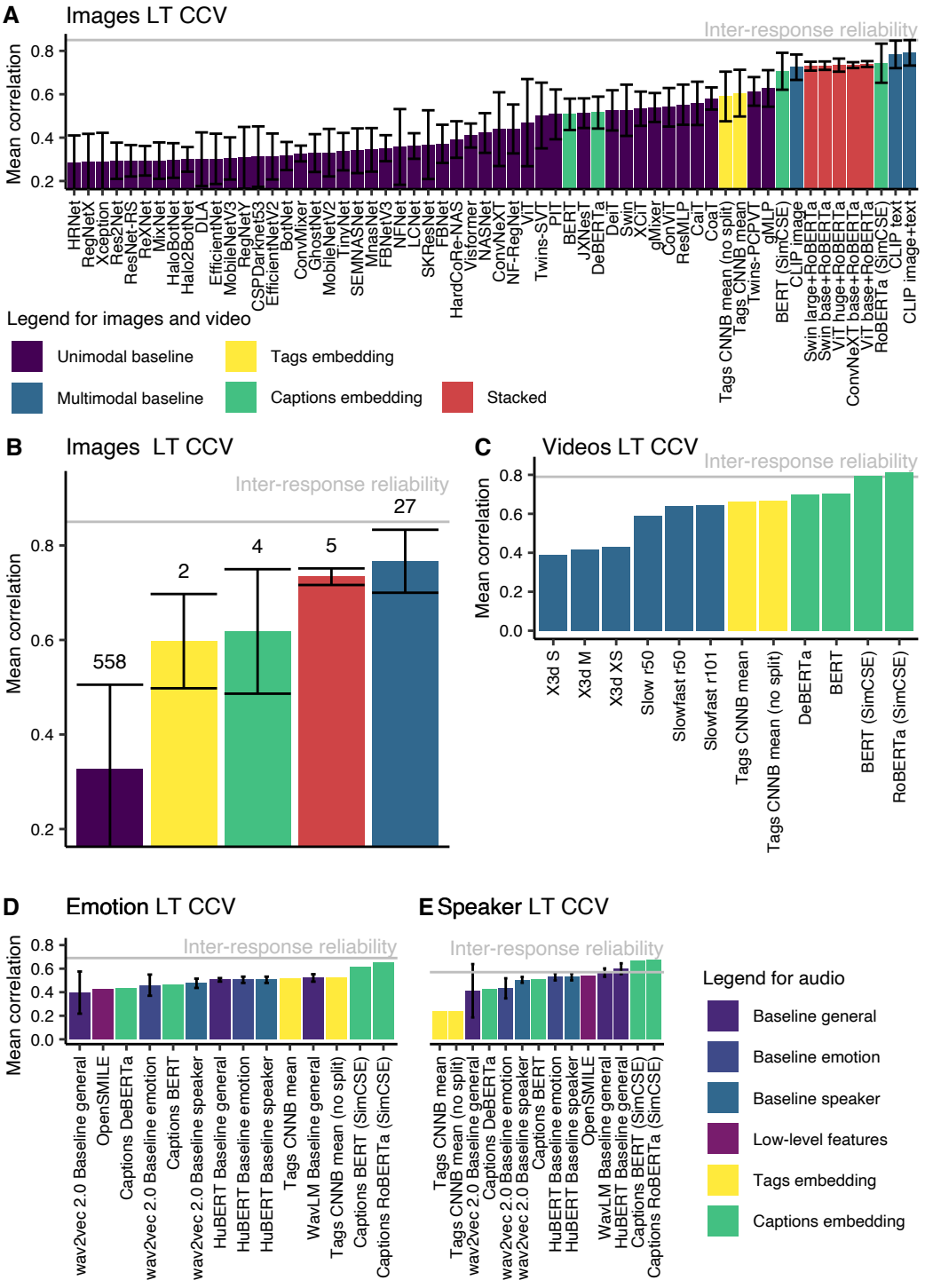

Figure 16: Model evaluations for image, audio and video datasets using the LT_CCV method. **A**. Performance of top 60 image models, averaged over the three image datasets and architecture subvariants. Here and throughout, error bars indicate 1 standard deviation (SD). Absence of error bars indicates single variant. **B**. Average model performance on images grouped by approach. **C**. Model evaluation for video. **D**. Model evaluation for audio in the emotion respect. **E**. Model evaluation for audio in the speaker respect.

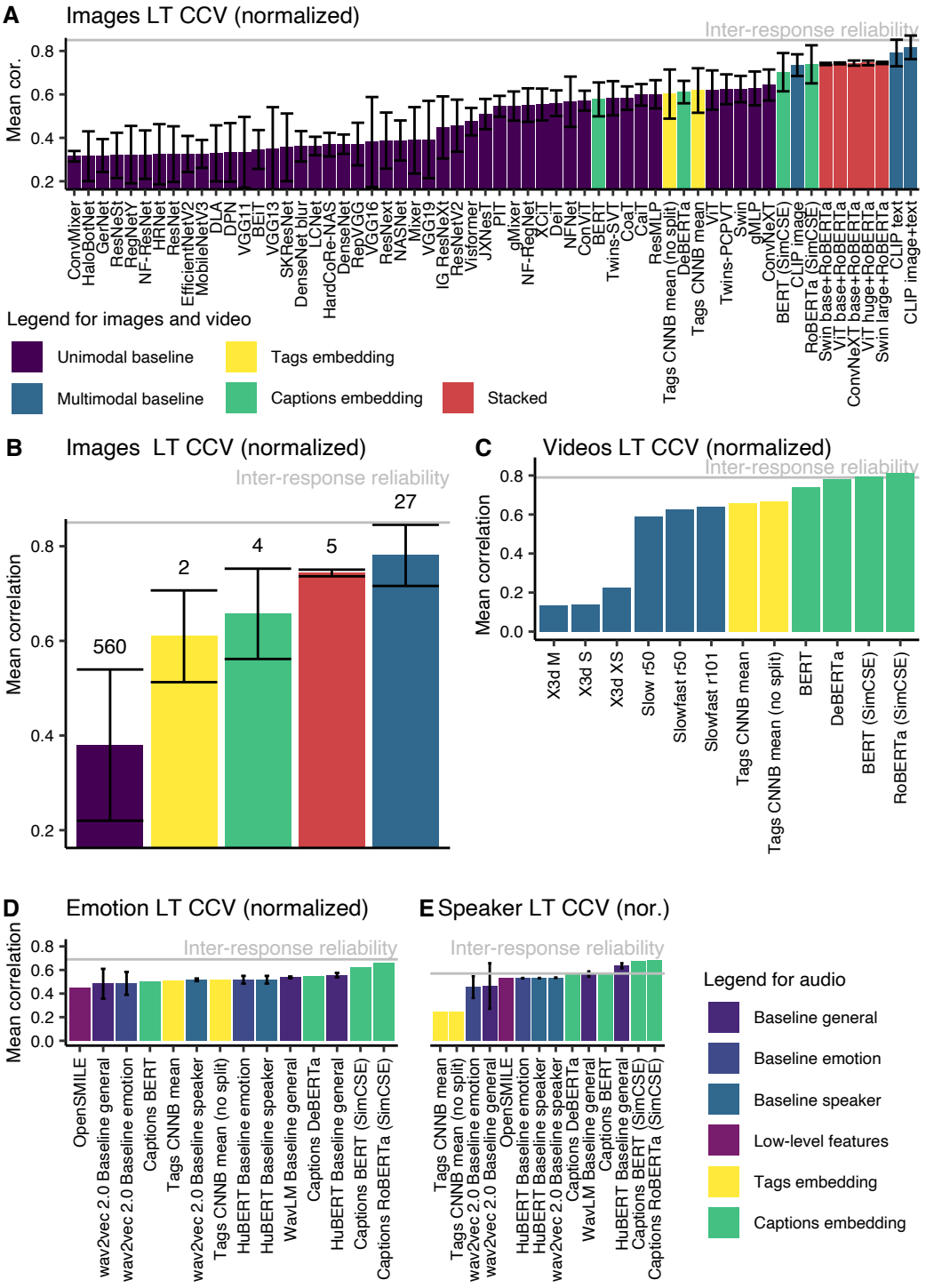

Figure 17: Model evaluations for image, audio and video datasets using the LT_CCV (norm) method. **A**. Performance of top 60 image models, averaged over the three image datasets and architecture subvariants. Here and throughout, error bars indicate 1 standard deviation (SD). Absence of error bars indicates single variant. **B**. Average model performance on images grouped by approach. **C**. Model evaluation for video. **D**. Model evaluation for audio in the emotion respect. **E**. Model evaluation for audio in the speaker respect.

D.4 Visualizing semantic analysis with multidimensional scaling

We used multidimensional scaling (MDS) embeddings to visualize the semantic content of similarity matrices in two dimensions. MDS maps were constructed using the `manifold.MDS` function in the `scikit-learn` Python library [84]. We first ran metric MDS to find a suitable initialization, and then used the resulting embeddings as the starting point for a non-metric MDS process. We used a maximum iteration limit of 10,000 and a convergence tolerance of 1e-100. You can explore all the MDS maps interactively here[17].

To create the MDS visualization for images (Figure 2) we generated a scatter plot for each image embedding and then overlaid the corresponding image on top of the scatter point. As for the audio MDS visualization (Figure 4) we colored MDS points based on the original target emotion in the speaker dataset and changed their shape (triangle vs. circle) based on the reported speaker sex. Finally, to visualize video MDS (Figure 6) we colored each point based on the original activity in the video dataset.

D.5 Visualizing semantic network with graph-based analysis

We applied semantic network analysis to explore the mined tags from the 1,000-object audio and video datasets (see Section 3.4). This method was used to investigate relationships between concepts in a more detailed way than a simple MDS could provide (section D.4). To visualize the network in Figure 7, we used each individual tag as a *node* in the network and the number of co-occurrence between each dyadic tag pairs as their *edge* weight. We observed that many of the tags were connected with many other tags, making the number of edges to be very large. To reduce the noise and make the network more interpretable, we pruned edges that had values below a certain threshold. We set this threshold differently for each dataset in proportion to the number of collected ratings and unique tags (video = 3, audio emotion = 5). Furthermore, to penalize tags that occur frequently and across many stimuli (and thus less informative), we re-weighted the edges by dividing the original edge weights by the sum of IDF (from TF-IDF approach) values of each tag pairs; note that we only consider the IDF component for calculation because the term frequency (TF) cannot be computed since the tags cannot be repeated (Note that here we did not apply the preprocessing steps of section C.1.2, in which tags that were rated more highly were repeated). This is analogous to the edge re-weighting method suggested by Newman [86].

Next, we used `Gephi` [87], an open-source software for visualizing and analyzing large network graphs, to visualize the tags for each dataset. We colored the nodes by their modularity class, a popular method in the network sciences for detecting community structures in a network [88]. 9 communities (or modularities) were detected for the video tags with a modularity score of 0.74, 5 were detected for audio emotion with a score of 0.49. The detected communities for each of the tag sets were highly interpretable. For instance, four of the modularities from the audio emotion tags distinctly separated the arousal–valence emotion space, with the detected communities being: high arousal–high valence, high arousal–low valence, low arousal–high valence, and low arousal–low valence. To illustrate this overlap, we manually added lines in Figure 7 that indicate our interpretation of the main axes for valance and arousal. Furthermore, we scaled the size of each node by their betweenness centrality value [86], which is a measure for the amount of influence a node has in the flow of the entire network. As visualized in Figure 7, the tags that had prominent roles in the video network defined distinct groups of activities well (e.g., music, game, sports). The prominent tags in the audio emotion network overlapped with emotion categories defined by previous literature [62] (e.g., scared, calm, upset, annoyed).

References

[1] G. Murphy, *The big book of concepts*. MIT press, 2004.

[2] N. Zaslavsky, C. Kemp, T. Regier, and N. Tishby, "Efficient compression in color naming and its evolution," *Proceedings of the National Academy of Sciences*, vol. 115, no. 31, pp. 7937–7942, 2018.

[3] S. T. Piantadosi, H. Tily, and E. Gibson, "Word lengths are optimized for efficient communication," *Proceedings of the National Academy of Sciences*, vol. 108, no. 9, pp. 3526–3529, 2011.

[17] `https://words-are-all-you-need.s3.amazonaws.com/index.html`

[4] T. Jaeger and R. Levy, "Speakers optimize information density through syntactic reduction," *Advances in Neural Information Processing Systems*, vol. 19, 2006.

[5] P. Kay and W. Kempton, "What is the Sapir-Whorf hypothesis?," *American Anthropologist*, vol. 86, no. 1, pp. 65–79, 1984.

[6] A. Majid, M. Bowerman, S. Kita, D. B. Haun, and S. C. Levinson, "Can language restructure cognition? the case for space," *Trends in Cognitive Sciences*, vol. 8, no. 3, pp. 108–114, 2004.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, pp. 8748–8763, PMLR, 2021.

[8] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," *arXiv preprint arXiv:2204.06125*, 2022.

[9] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "VATT: Transformers for multimodal self-supervised learning from raw video, audio and text," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[11] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum, "Simulation as an engine of physical scene understanding," *Proceedings of the National Academy of Sciences*, vol. 110, no. 45, pp. 18327–18332, 2013.

[12] G. Marcus, "Deep learning: A critical appraisal," *arXiv preprint arXiv:1801.00631*, 2018.

[13] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[14] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, *et al.*, "Flamingo: a visual language model for few-shot learning," *arXiv preprint arXiv:2204.14198*, 2022.

[15] R. N. Shepard, "Multidimensional scaling, tree-fitting, and clustering," *Science*, vol. 210, no. 4468, pp. 390–398, 1980.

[16] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[17] L. Von Ahn and L. Dabbish, "Designing games with a purpose," *Communications of the ACM*, vol. 51, no. 8, pp. 58–67, 2008.

[18] L. Von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 319–326, 2004.

[19] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[20] E. L. Law, L. Von Ahn, R. B. Dannenberg, and M. Crawford, "TagATune: A game for music and sound annotation," in *ISMIR*, vol. 3, p. 2, 2007.

[21] S. Kirby, H. Cornish, and K. Smith, "Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language," *Proceedings of the National Academy of Sciences*, vol. 105, no. 31, pp. 10681–10686, 2008.

[22] T. L. Griffiths and M. L. Kalish, "A bayesian view of language evolution by iterated learning," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 27, 2005.

[23] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *Thirty-first AAAI Conference on Artificial Intelligence*, 2017.

[24] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," *arXiv preprint arXiv:2104.08821*, 2021.

[25] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8619–8624, 2014.

[26] D. Linsley, S. Eberhardt, T. Sharma, P. Gupta, and T. Serre, "What are the visual features underlying human versus machine vision?," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 2706–2714, 2017.

[27] T. Langlois, H. Zhao, E. Grant, I. Dasgupta, T. Griffiths, and N. Jacoby, "Passive attention in artificial neural networks predicts human visual selectivity," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[28] A. J. Kell, D. L. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. McDermott, "A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy," *Neuron*, vol. 98, no. 3, pp. 630–644, 2018.

[29] B. J. Devereux, A. Clarke, A. Marouchos, and L. K. Tyler, "Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects," *Journal of Neuroscience*, vol. 33, no. 48, pp. 18906–18916, 2013.

[30] J. S. German and R. A. Jacobs, "Can machine learning account for human visual object shape similarity judgments?," *Vision Research*, vol. 167, pp. 87–99, 2020.

[31] R. M. Cichy, A. Khosla, D. Pantazis, and A. Oliva, "Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks," *NeuroImage*, vol. 153, pp. 346–358, 2017.

[32] M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, F. Geiger, *et al.*, "Brain-Score: Which artificial neural network for object recognition is most brain-like?," *BioRxiv*, p. 407007, 2020.

[33] J. Diedrichsen and N. Kriegeskorte, "Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis," *PLoS Computational Biology*, vol. 13, no. 4, p. e1005508, 2017.

[34] H. Nili, C. Wingfield, A. Walther, L. Su, W. Marslen-Wilson, and N. Kriegeskorte, "A toolbox for representational similarity analysis," *PLoS Computational Biology*, vol. 10, no. 4, p. e1003553, 2014.

[35] M. Mur, M. Meys, J. Bodurka, R. Goebel, P. A. Bandettini, and N. Kriegeskorte, "Human object-similarity judgments reflect and transcend the primate-IT object representation," *Frontiers in Psychology*, vol. 4, p. 128, 2013.

[36] J. C. Peterson, J. T. Abbott, and T. L. Griffiths, "Evaluating (and improving) the correspondence between deep neural networks and human representations," *Cognitive Science*, vol. 42, no. 8, pp. 2648–2669, 2018.

[37] M. N. Hebart, C. Y. Zheng, F. Pereira, and C. I. Baker, "Revealing the multidimensional mental representations of natural objects underlying human similarity judgements," *Nature Human Behaviour*, vol. 4, no. 11, pp. 1173–1185, 2020.

[38] R. Marjieh, I. Sucholutsky, T. R. Sumers, N. Jacoby, and T. L. Griffiths, "Predicting human similarity judgments using large language models," *arXiv preprint arXiv:2202.04728*, 2022.

[39] M. C. Iordan, T. Giallanza, C. T. Ellis, N. M. Beckage, and J. D. Cohen, "Context matters: Recovering human semantic structure from machine learning analysis of large-scale text corpora," *Cognitive Science*, vol. 46, no. 2, p. e13085, 2022.

[40] I. I. Groen, M. R. Greene, C. Baldassano, L. Fei-Fei, D. M. Beck, and C. I. Baker, "Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior," *Elife*, vol. 7, p. e32962, 2018.

[41] A. Jha, J. Peterson, and T. L. Griffiths, "Extracting low-dimensional psychological representations from convolutional neural networks," *arXiv preprint arXiv:2005.14363*, 2020.

[42] B. D. Roads and B. C. Love, "Enriching ImageNet with human similarity judgments and psychological embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3547–3557, 2021.

[43] Z. Parekh, J. Baldridge, D. Cer, A. Waters, and Y. Yang, "Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO," *arXiv preprint arXiv:2004.15020*, 2020.

[44] P. Harrison, R. Marjieh, F. Adolfi, P. van Rijn, M. Anglada-Tort, O. Tchernichovski, P. Larrouy-Maestri, and N. Jacoby, "Gibbs sampling with people," *Advances in Neural Information Processing Systems*, vol. 33, pp. 10659–10671, 2020.

[45] R. Wightman, "PyTorch image models." `https://github.com/rwightman/pytorch-image-models`, 2019.

[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[47] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," *arXiv preprint arXiv:2201.03545*, 2022.

[48] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

[49] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Ieee, 2009.

[51] H. Bao, L. Dong, and F. Wei, "BEiT: BERT pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[52] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression.," *Journal of Personality and Social Psychology*, vol. 70, no. 3, p. 614, 1996.

[53] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[54] D. L. Medin, R. L. Goldstone, and D. Gentner, "Respects for similarity," *Psychological Review*, vol. 100, no. 2, p. 254, 1993.

[55] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[56] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[57] P. Van Rijn, S. Mertes, D. Schiller, P. Harrison, P. Larrouy-Maestri, E. André, and N. Jacoby, "Exploring emotional prototypes in a high dimensional TTS latent space," *arXiv preprint arXiv:2105.01891*, 2021.

[58] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 305–321, 2018.

[59] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6202–6211, 2019.

[60] C. Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 203–213, 2020.

[61] L. Caplette and N. Turk-Browne, "Computational reconstruction of mental representations using human behavior," *PsyArXiv*, 2022.

[62] A. S. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proceedings of the National Academy of Sciences*, vol. 114, no. 38, pp. E7900–E7909, 2017.

[63] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, pp. 1597–1607, PMLR, 2020.

[64] B. Thompson, S. G. Roberts, and G. Lupyan, "Cultural influences on word meanings revealed through large-scale semantic alignment," *Nature Human Behaviour*, vol. 4, no. 10, pp. 1029–1038, 2020.

[65] H. C. Barrett, "Towards a cognitive science of the human: cross-cultural approaches and their urgency," *Trends in Cognitive Sciences*, vol. 24, no. 8, pp. 620–638, 2020.

[66] K. Lemhöfer and M. Broersma, "Introducing lextale: A quick and valid lexical test for advanced learners of english," *Behavior research methods*, vol. 44, no. 2, pp. 325–343, 2012.

[67] K. J. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, "Headphone screening to facilitate web-based auditory experiments," *Attention, Perception, & Psychophysics*, vol. 79, no. 7, pp. 2064–2072, 2017.

[68] A. E. Milne, R. Bianco, K. C. Poole, S. Zhao, A. J. Oxenham, A. J. Billig, and M. Chait, "An online headphone screening test based on dichotic pitch," *Behavior Research Methods*, vol. 53, no. 4, pp. 1551–1562, 2021.

[69] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.

[70] Y.-Y. Yang, M. Hira, Z. Ni, A. Chourdia, A. Astafurov, C. Chen, C.-F. Yeh, C. Puhrsch, D. Pollack, D. Genzel, D. Greenberg, E. Z. Yang, J. Lian, J. Mahadeokar, J. Hwang, J. Chen, P. Goldsborough, P. Roy, S. Narenthiran, S. Watanabe, S. Chintala, V. Quenneville-Bélair, and Y. Shi, "Torchaudio: Building blocks for audio and speech processing," *arXiv preprint arXiv:2110.15018*, 2021.

[71] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *arXiv preprint arXiv:2110.13900*, 2021.

[72] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," 2022.

[73] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, pp. 1194–1198, 2021.

[74] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," 2022.

[75] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021. arXiv:2106.04624.

[76] D. Yamins, "An optimization-based approach to understanding sensory systems," *The Cognitive Neurosciences*, vol. 4, no. V1, p. 381, 2020.

[77] H. Fan, T. Murrell, H. Wang, K. V. Alwala, Y. Li, Y. Li, B. Xiong, N. Ravi, M. Li, H. Yang, J. Malik, R. Girshick, M. Feiszli, A. Adcock, W.-Y. Lo, and C. Feichtenhofer, "PyTorchVideo: A deep learning library for video understanding," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. `https://pytorchvideo.org/`.

[78] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The Kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[79] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Association for Computational Linguistics, 2020.

[80] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020.

[81] L. C. ROUGE, "A package for automatic evaluation of summaries," in *Proceedings of Workshop on Text Summarization of ACL, Spain*, 2004.

[82] F. Barrios, F. López, L. Argerich, and R. Wachenchauzer, "Variations of the similarity function of textrank for automated summarization," *arXiv preprint arXiv:1602.03606*, 2016.

[83] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile," in *Proceedings of the international conference on Multimedia - MM '10*, ACM Press, 2010.

[84] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[85] W. Brown, "Some experimental results in the correlation of mental abilities 1," *British Journal of Psychology, 1904-1920*, vol. 3, no. 3, pp. 296–322, 1910.

[86] M. E. J. Newman, "Scientific collaboration networks. i. network construction and fundamental results," *Phys. Rev. E*, vol. 64, p. 016131, Jun 2001.

[87] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks," 2009.

[88] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, p. P10008, oct 2008.