scientific data

OPEN

DATA DESCRIPTOR

101 Dalmatians: a multimodal naturalistic fMRI dataset in typical development and congenital sensory loss

Francesca Setti[1], Davide Bottari[1], Andrea Leo[2], Matteo Diano[3], Valentina Bruno[4], Carla Tinti[3], Luca Cecchetti[1], Francesca Garbarini[4], Pietro Pietrini[1], Emiliano Ricciardi[1]✉ & Giacomo Handjaras[1]

In recent years, neuroscience has increasingly leveraged naturalistic stimuli, like movies and narratives, to investigate cognitive processes underlying real-world human behavior. Here, we present a functional Magnetic Resonance Imaging (fMRI) dataset featuring 50 participants, with and without congenital sensory loss (typical development, congenitally blind and deaf individuals), who were exposed to audiovisual, auditory, or visual versions of the live-action movie 101 *Dalmatians*. The dataset incorporates auditory and visual descriptors from established computational models (e.g., VGGish, VGG-19) and semantic embeddings generated by GPT-4, complemented by human-tagged annotations of movie events and content. All data are provided in a standardized BIDS format. fMRI data quality was evaluated through Inter-Subject Correlation (ISC), ensuring robust comparisons across participants and groups. The *101 Dalmatians* dataset facilitates the exploration of the effects of congenital sensory deprivation on brain functional organization, neuroplasticity, and the interplay between sensory inputs and cognitive processes. It is a valuable resource for understanding how sensory experiences —or their absence— shape human brain development and functional adaptation.

Background & Summary

Over the past decade, cognitive science has increasingly embraced naturalistic stimuli, such as movies and narratives, to deepen our understanding of the cognitive processes that underlie human behavior in real-world contexts[1–3]. While introducing novel methodological and analytical challenges, naturalistic paradigms also facilitate the collection of more ecological and multidimensional data, including a broad spectrum of real-world sensory, cognitive, emotional, and social inputs. Furthermore, due to their inherent complexity and dynamism, naturalistic stimuli provide a unique opportunity to investigate the temporal dynamics of cognitive processes. We present an fMRI dataset comprising data from three groups of typically developed individuals and two groups of participants with congenital sensory loss (i.e., congenitally blind and congenitally deaf) who were exposed to either the audiovisual, auditory, or visual versions of the edited action movie 101 *Dalmatians*. The dataset includes low- and high-level visual (i.e., motion energy and VGG-19) and auditory (i.e., sound power spectrum and VGGish) models that describe spectro- and spatio- temporal modulations in the auditory and visual streams. Additionally, we complement our dataset by providing semantic information based on a comprehensive set of annotations that characterize the categorical content of the movie's visual display and audio description, as well as GPT-4 sentence embeddings from the subtitles. A schematic overview of the workflow encompassing data collection, preprocessing, technical validation, and annotations procedure is presented in Fig. 1. Through this dataset, the characterization of the sensory-deprived brain offers a unique opportunity to understand to what extent a post-natal sensory experience - or the lack of it - shapes the development and the refinement of brain

[1]MoMiLab, IMT School for Advanced Studies Lucca, Lucca, Italy. [2]Nuclear Medicine Unit, Department of Diagnostic Imaging, N.O.P. - S. Stefano, U.S.L. Toscana Centro, Prato, Italy. [3]Department of Psychology, University of Turin, Turin, Italy. [4]Manibus Lab, Department of Psychology, University of Turin, Turin, Italy. ✉e-mail: emiliano.ricciardi@imtlucca.it

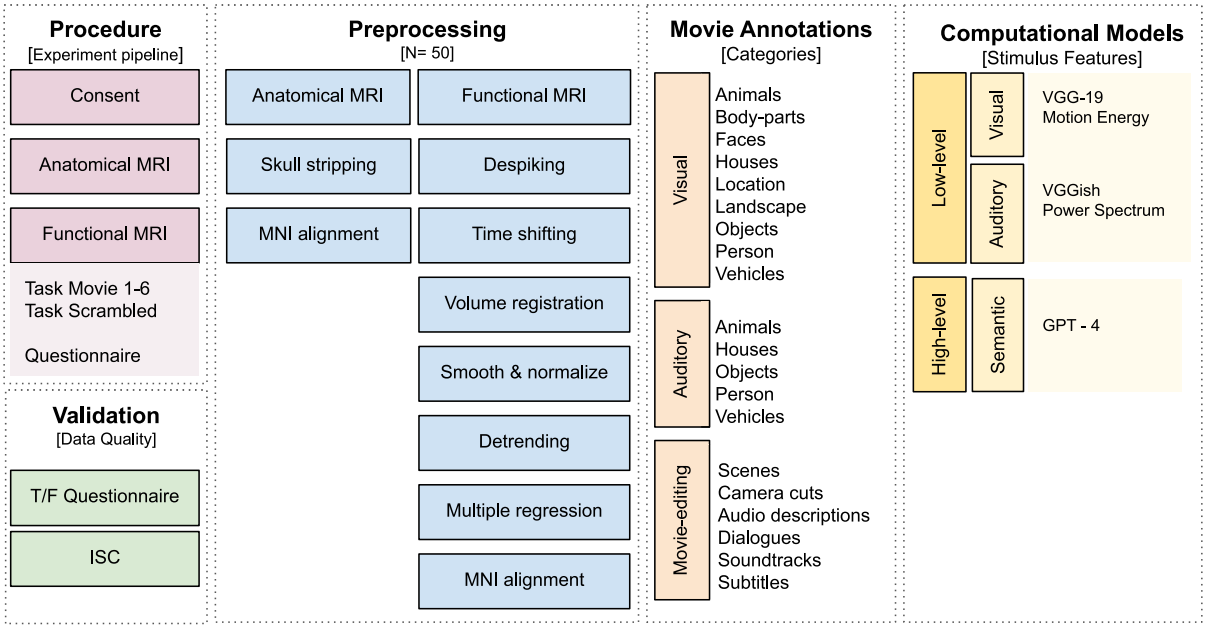

Fig. 1 Schematic overview of the *101 Dalmatians* dataset. Data collection procedures, preprocessing, technical validation and annotation are shown. Anatomical and functional MRI data were collected and analyzed. After data preprocessing (light blue), a validation step was performed by assessing the framewise displacement (FD) and the intersubject correlations (ISC) metrics. A comprehensive set of categorical annotations is provided along with an accurate computational modelling of low-level, high-level and linguistic movie features.

functional organization[4,5]. Furthermore, the present dataset allows the study of the adaptive changes resulting from sensory loss and the role of the distinct senses on human brain development, cognition and behaviour.

Methods

Participants. Fifty participants were enrolled in the present study which included three groups of individuals with typical development (TD) and two separate samples of participants with a congenital sensory-deprivation (SD, congenitally lacking visual or auditory experiences). Specifically, TD participants were assigned to one of three distinct experimental conditions and were instructed to attend the presentation of a specific version of the same movie: the complete multimodal audiovisual (AV) ($n = 10$, 35 ± 13 years, 8 females), the auditory (A) ($n = 10$, 39 ± 17 years, 7 females) or the visual (V) ($n = 10$, 37 ± 15 years, 5 females) one. Likewise, the two SD groups consisting of congenitally blind ($n = 11$, mean age 46 ± 14 years, 3 females) and deaf ($n = 9$, mean age 24 ± 4, 5 females) participants were presented with the auditory (A) and visual (V) movie conditions, respectively. All participants were right-handed, as determined by the Edinburgh Handedness Inventory. Only native Italian speakers were selected for enrollment in the study. None of the participants had a history of psychiatric or neurological disorders. Educational level was matched across groups. All participants, except for one in the TD group, held at least a high school diploma.

The blind and deaf participants were congenitally deprived, except for one deaf subject who reported sensorineural hearing loss before the age of one. We adopted the term "congenital" in line with prior literature[6,7] to refer to individuals who experienced total sensory deprivation within the first year of life, prior to the onset of structured sensory experience (e.g., patterned vision or speech perception) and during critical developmental windows for sensory cortex maturation. All deaf participants were proficient in Italian Sign Language (LIS) and did not use hearing aids during the study. TD subjects reported no hearing impairment, normal or corrected-to-normal vision and no knowledge of the LIS. For more detailed information about the SD participants' demographics, refer to Table 1. To ensure adequate levels of Italian proficiency, all deaf participants completed a screening procedure prior to inclusion. This procedure combined both comprehension questionnaires, designed to assess familiarity with the movie plot and understanding of the story content, and self-reported data on language use and comprehension in daily life. Reading behavior during the questionnaires (e.g., response consistency, completeness) was monitored to detect any potential comprehension difficulties. All participants completed the questionnaires successfully and consistently reported regular use of written Italian.

All participants received detailed information about the nature of the research. Each volunteer provided written informed consent for participation and for sharing their de-identified data publicly by the guidelines set forth by the institutional board of the Turin University Imaging Centre for Brain Research. The study was approved by the Ethical Committee of the University of Turin (protocol n. 195874, dated 05/29/19) and adheres to the principles outlined in the Declaration of Helsinki (2013).

Procedures. The fMRI data collection was preceded by an initial assessment of the participant's knowledge of the movie plot, conducted through a five-stage familiarity questionnaire ranging from 1 (never heard of it)

Subject	Gender	Age	Cause of blindness	Residual Light Perception	Age of Braille reading	Education
041	M	49	retinopathy of prematurity	NLP	6	Bachelor's Degree
039	M	42	retinitis pigmentosa	NLP	6	High school diploma
036	M	48	optic nerve atrophy	NLP	6	High school diploma
035	F	37	Leber congenital amaurosis	NLP	6	Bachelor's Degree
038	M	32	retinal detachment	NLP	6	High school diploma
043	F	41	bilateral retinoblastoma	NLP	6	High school diploma
042	M	19	retinal detachment	NLP	6	High school diploma
053	M	57	retinopathy of prematurity	NLP	6	High school diploma
033	F	68	optic nerve atrophy	NLP	6	High school diploma
Subject	**Gender**	**Age**	**Cause of deafness**	**First Language**	**Hearing aid use**	**Education**
044	M	24	hereditary	sign	used during childhood	High school diploma
045	F	21	hereditary	sign	used during childhood	High school diploma
046	M	24	hereditary	sign	used during childhood	High school diploma
047	M	26	hereditary	sign	used during childhood	High school diploma
048	M	22	hereditary	sign	used during childhood	High school diploma
049	F	18	hereditary	sign	used during childhood	High school diploma
050	F	28	Sensorineural hearing loss	sign	currently used	High school diploma
051	F	32	hereditary	sign	used during childhood	High school diploma
052	F	22	hereditary	sign	used during childhood	High school diploma

Table 1. Characteristics of congenitally blind (upper table) and congenitally deaf (lower table) participants. NLP, No Light Perception; M, male; F, female. The last column reports the highest level of education self-reported by participants at the time of data acquisition.

to 5 (know it very well). Then, each subject was instructed to attend one of the three edited versions (e.g., the visual -V-, auditory -A-, or audiovisual -AV-), of the movie, according to the characteristics of the group they belong to, while undergoing the fMRI recordings. Both structural and functional data acquisition took place on a single scanning day. After the scanning session, participants completed an ad hoc two-alternative forced-choice questionnaire to evaluate their engagement with the story and compliance during the experiment. Audio and visual stimulation were provided using MR-compatible LCD goggles and headphones (VisualStim Resonance Technology) featuring a video resolution of 800 × 600 at 60 Hz, a visual field 30° × 22°, and a 5-inch display. The audio system offered 30 dB noise attenuation and a frequency response ranging from 40 Hz to 40 kHz. All participants utilized both the goggles and headphones regardless of the experimental condition or group membership. In the audio-only condition, all participants, including those without sensory loss, the goggles were turned off, resulting in a completely dark screen and the absence of any visual input from the experimental stimuli. Importantly, the use of these goggles also effectively prevented any incidental visual stimulation from the surrounding environment, ensuring a strictly auditory sensory context. In addition to this, in the audio-only condition, participants were explicitly instructed to keep their eyes closed for the entire duration of the fMRI session. The visual clips and auditory narratives were delivered through the Presentation® 16.5 software package (Neurobehavioral System, Berkeley, CA, USA - http://www.neurobs.com).

Stimuli. Three edited (e.g., audiovisual, visual, and auditory) and shortened (≃54 minutes) versions of the live-action movie 101 *Dalmatians* (S. Herek, Great Oaks Entertainment & Walt Disney, 1996) were employed. The movie was presented in six runs of variable length ranging from approximately eight to ten minutes. Additionally, a scrambled run was created by randomly combining discarded scenes from the original movie intentionally disrupting the narrative coherence. A voice-over was integrated with the movie's original soundtrack to describe elements typically provided by the visual scenery that were not captured either by character dialogues or music valence yet were crucial for the narrative comprehension. The voice-over was recorded by a professional Italian actor in a soundproof studio equipped with high-quality hardware (Neumann U87 ai microphone, Universal Audio LA 610 mk2 preamplifier, Apogee Rosetta converter, Apple MacOS) and software (Logic Pro 10.4), which included various microphones and sound manipulation filters. The voice track was subsequently mixed with the movie's original audio, and fade-in and fade-out effects were applied to enhance auditory transitions, followed by a final remix of the music and voice components. We meticulously transcribed the soundscape of the movie including human voices, the narrator's voice-over, and environmental and natural sounds into subtitles. These subtitles were designed in various styles and colors according to the speaking voice to enhance speech segmentation and comprehension, and adjusted for both one-line and two-line displays. Video editing was conducted using iMovie (version 10.1.10) on an Apple MacBook Air Pro Retina, while subtitles generation was performed with the open-source software Aegisub 3.2.2 (http://www.aegisub.org/). In both visual and audiovisual conditions, a small red fixation cross was centrally displayed, with subtitles positioned at the bottom of the screen.

MR Acquisition. Brain activity was recorded with a Philips 3 T Ingenia scanner equipped with a 32-channel head coil. Functional images were acquired using Gradient-Recalled Echo Echo-Planar Imaging (GRE-EPI;

TR = 2000 ms; TE = 30 ms; FA = 75°; FOV = 240 mm; acquisition matrix (in-plane resolution) = 80 × 80; acquisition slice thickness = 3 mm; acquisition voxel size = 3 × 3 × 3 mm; reconstruction voxel size = 3 × 3 × 3 mm; 38 sequential axial ascending slices; total volumes 1,614 for the six runs of the movie, plus 256 for the scrambled run). In the same session, three-dimensional high-resolution anatomical images of the brain were also acquired using a Magnetization-Prepared RApid Gradient Echo (MPRAGE) sequence (TR = 7 ms; TE = 3.2 ms; FA = 9°; FOV = 224, acquisition matrix = 224 × 224; slice thickness = 1 mm; voxel size = 1 × 1 × 1 mm; 156 sagittal slices), and a T2-weighted sequence (TR = 2.5 ms; TE = 250 ms; FA = 90°; FOV = 224, acquisition matrix = 224 × 224; slice thickness = 1 mm; voxel size = 1 × 1 × 1 mm; 152 sagittal slices).

Preprocessing. *Anatomical.* First, we removed the skull from the anatomical/structural MRI scans using the brain extraction script (antsBrainExtraction.sh) provided within the Advanced Normalization Tools (ANTs-v2.3.5–126) software[8]. The resulting anatomical images were nonlinearly warped to the ICBM 152 Nonlinear Symmetrical template version 2009c.

Functional. Shared fMRI data was processed using a minimal pipeline following the standard steps with the AFNI_17.1.12 software package[9]. First, we removed scanner-related noise by correcting the data by spike removal (*3dDespike*). Then, all volumes comprising a run were temporally aligned (*3dTshift*) and successively corrected for head motion using as a base the first run (*3dvolreg*). Spatial smoothing with a Gaussian kernel (*3dBlurToFWHM*, 6 mm, Full Width at Half Maximum) was applied and then data of each run underwent percentage normalization. Aside, detrending applying Savitzky-Golay filtering (*sgolayfilt*, polynomial order: 3, frame length: 200 timepoints) in MATLAB R2019b (MathWorks Inc., Natick, MA, USA) was performed onto the normalized runs to smooth the corresponding time series and clean them from unwanted trends and outliers. Runs were then concatenated, and multiple regression analysis was performed (*3dDeconvolve*) to remove signals related to head motion parameters and movement spike regressors (framewise displacement above 0.3). Afterward single subject fMRI volumes were nonlinearly (*3dQWarp*) registered to the MNI-192 standard space[10]. Four blind participants (i.e, sub-034, sub-036, sub-037, sub-053), and one audiovisual participants (i.e., sub-016) retained >20% of timepoints with relatively high framewise displacement (>0.3 mm), and may require additional procedures to remove artifacts related to head motion[11].

Computational models. The following sections will outline the models used in this study to extract both low- and high-level (visual and auditory) and the semantic and categorical movie's properties. In line with the theoretical framework of hierarchical sensory processing[12–14], we made use of the complexity of naturalistic stimuli to identify low-level, higher-level, and categorical features that can help understand how these features modulate brain responses. To analyze the low-level properties of the early visual and auditory systems, we measured both the frequency and the temporal variations in the visual and auditory movie stimuli. Specifically, the low-level visual model was based on features derived from VGG-19[15] and motion energy[16], while the low-level acoustic model was based on the sound power spectrum[12] and VGGish[17]. Additionally, two semantic high-level models were employed. One model utilized information from the stimulus semantics through a GPT-4[18] model, while the other involved manual tagging of the visual and auditory content for event categorical discrimination. This tagging included categories such as *Animals, Houses, Objects, Person*, and *Vehicles*, as illustrated in Fig. 1. Finally, we developed a *movie-editing* model that focused on properties of the stimuli, including *Camera Cuts, Scenes, Dialogues, Audio Descriptions, and Soundtracks*. This model captured the slow-paced relationship between auditory and visual elements introduced during the editing process, which influenced both low-level and high-level semantic descriptors and whose role could be thus interesting to investigate.

Low-level visual model: motion energy feature space. As in our previous works[4,19], the total motion energy was calculated for each two-second segment of the movie videoclip using a comprehensive set of 4,715 motion energy descriptors derived from a quadrature-pair of space-time Gabor filters. These filters included Gabor wavelets with three distinct temporal frequencies: 0 Hz (static energy), 2 Hz, and 4 Hz as referenced in[16].

The MATLAB code utilized for this analysis is accessible at: https://github.com/gallantlab/motion_energy_matlab.

Each movie frame was characterized by a set of preferred spatial frequencies, orientations and temporal frequencies, effectively capturing rapidly changing visual information.

High-level visual model: VGG-19 feature space. In this study, we employed the VGG-19[15] convolutional deep neural network architecture to extract a comprehensive set of visual features from the movie frames. Specifically, we extracted the output from the ReLU3.1 (Rectified Linear Unit) activation function of one of the intermediate convolutional layers, namely Conv3_1, which has demonstrated superior performance compared to traditional V1 models[20] in capturing the properties of primary visual cortex neurons. This approach enabled us to leverage the network's non-linear properties to enhance the representation of low/mid-level spatial statistics of each movie frame[19]. Additionally, we evaluated the output of ReLu6, which follows the fully connected layer fc6 and carries high-level information about the visual scene crucial for object recognition and image classification. By processing the movie frames through the VGG-19 model, we aimed to obtain a robust set of descriptors that reflects both low-level and higher-level visual characteristics, thereby facilitating a deeper understanding of the neural mechanisms underlying visual perception in the context of dynamic stimuli. This approach not only aligns with contemporary methodologies in vision neuroscience but also provides a solid foundational model for further research in the field of computer vision and visual neuroscience.

Low-level auditory model: power spectrum feature space. Spectral features extraction was conducted following the methodology outlined by[12]. We estimated the signal power spectrum for each run using the Welch's power spectral density estimate[21] employing a Gaussian window with a standard deviation of 5 ms, a length 30 ms, and 1 ms spacing between windows, applied to segments of the signal lasting 2 seconds to match the temporal resolution of the fMRI. The resulting output is a 449-dimensional vector that describes the signal power spectrum (expressed in decibels) across the frequency range of 0 Hz to approximately 15000 Hz, computed in frequency bands of 33.5 Hz. For additional details about the parameters used please refer to[22].

High-level auditory model: VGGish features space. A set of high-level acoustic features was extracted from the auditory stream of the movie by means of the VGGish[17] model, a convolutional neural network based on the VGGNet architecture designed for image recognition and successfully adapted for audio classification[23]. This model captures a rich set of auditory features that encompass the spectral (e.g., harmonic structures and timbre), temporal (e.g., rhythm and tempo), and semantic dimensions of the sound inputs. By using the output of the ReLu5.1 layer we were able to investigate more abstract, high-level properties of the movie's audio track which include contextual information about the soundscape such as the presence of background noise, music or speech.

Compositional semantic features using GPT-4. Given the availability of the complete verbal content of the movie in English, the narrative was divided according to the subtitles, which predominantly consisted of individual sentences. Subsequently, we obtained contextual word embeddings from each sentence using the pre-trained GPT-4[18] model via the OpenAI API (https://openai.com/). GPT-4 is currently the most advanced artificial language model available. In summary, we obtained a vector of 1536 dimensions from each sentence using the model text-embedding-3-small.

Categorical model: Annotations. Here, we introduce annotations for a set of semantic descriptors designed to characterize the representation of stimulus categories through visual and auditory processing (Fig. 2). We created two distinct annotation sets corresponding to the different modalities of movie presentation. This was necessary because while the narrator's speech aims to convey visual information, the auditory content does not always align with visually defined categories. For instance, while visual scenes allow for the identification of specific features such as human or animal faces and relevant body parts during actions, the auditory narrative provides a more global representation, often merely suggesting the presence of a person. The present auditory annotations focus on non-stationary foreground sounds -those whose signal characteristics evolve over time and provide richer contextual information- and natural background sounds. The inclusion of background environmental noise is justified by its differential impact on responses in primary and non-primary auditory areas to concurrent foreground sounds[24], as well as its role as a valuable information source for individuals with visual impairments. Consequently, our classification encompasses audio signals from nature, animals, man-made objects, and human activities. Auditory and visual categories were tagged by one human rater, who recorded the presence of relevant features and their corresponding timecodes at a one-second resolution. The rater was trained to annotate the movie across multiple dimensions, including both auditory and visual features, covering 17 categories in total: 12 visual (animals, body parts, faces [human & animal], houses, locations [indoor vs outdoor], landscapes [natural vs artificial], objects, persons, and vehicles) and 5 auditory (animals, houses, objects, persons, and vehicles).

Moreover, we annotated a range of non-specific visual, auditory, and linguistic stimulus properties that cannot be distinctly attributed either to the categorical or to the computational models (see[4]) conceived in the analytical framework of the study. This choice was motivated by the fact that the synchronization of brain activity observed in the fMRI literature, primarily stems from slow-frequency fluctuations traced back to the movie's intrinsic structure. Thus, this class of descriptors encompasses features related to the editing process such as scene transitions, camera cuts, the pacing of soundtracks and dialogues, the addition of audio descriptions and subtitles. Overall, the annotation process involved approximately 200 hours of manual scoring.

Visual categories. For the visual condition, the continuous stream of information was classified in the following superordinate categories: Animals, Body-parts, Faces (animal faces and human faces), Houses, Location (inside vs outside), Landscape (artificial vs natural), Objects, Person, and Vehicles.

In each one-second time interval, the elements present in the foreground were tagged. Additionally, when applicable, supplementary items were noted based on their visual characteristics -such as color, size, and positional changes- or their narrative significance, which included essential information and main characters that could potentially engage the viewer's attention. Each recorded entry was subsequently categorized into one of the categories detailed below.

Animals: this category covers all species of animals depicted in the movie.

Body-parts: this term refers exclusively to individual parts of the body that appear in isolation and in the foreground, excluding faces. Examples include a leg, toe, or hand.

Faces: this category includes all faces presented in the foreground, irrespective of viewpoint and lighting conditions. Faces that are visible from a distance or in the presence of the trunk or entire body do not fall under this category (refer to "Person"). This class is further divided into human and animal faces.

Houses: this term pertains to both the façade of a building presented in isolation and to groups of buildings or cityscapes. It also comprises other types of structures that may not be classified as "houses" in a strict sense but still relate to the broader concept of "buildings" (e.g., farms, castles).

Location: this term refers to the context in which a scene occurs, indicating whether it is indoors (e.g., in a living room or an office) or outdoors (e.g., on a road or in a garden).

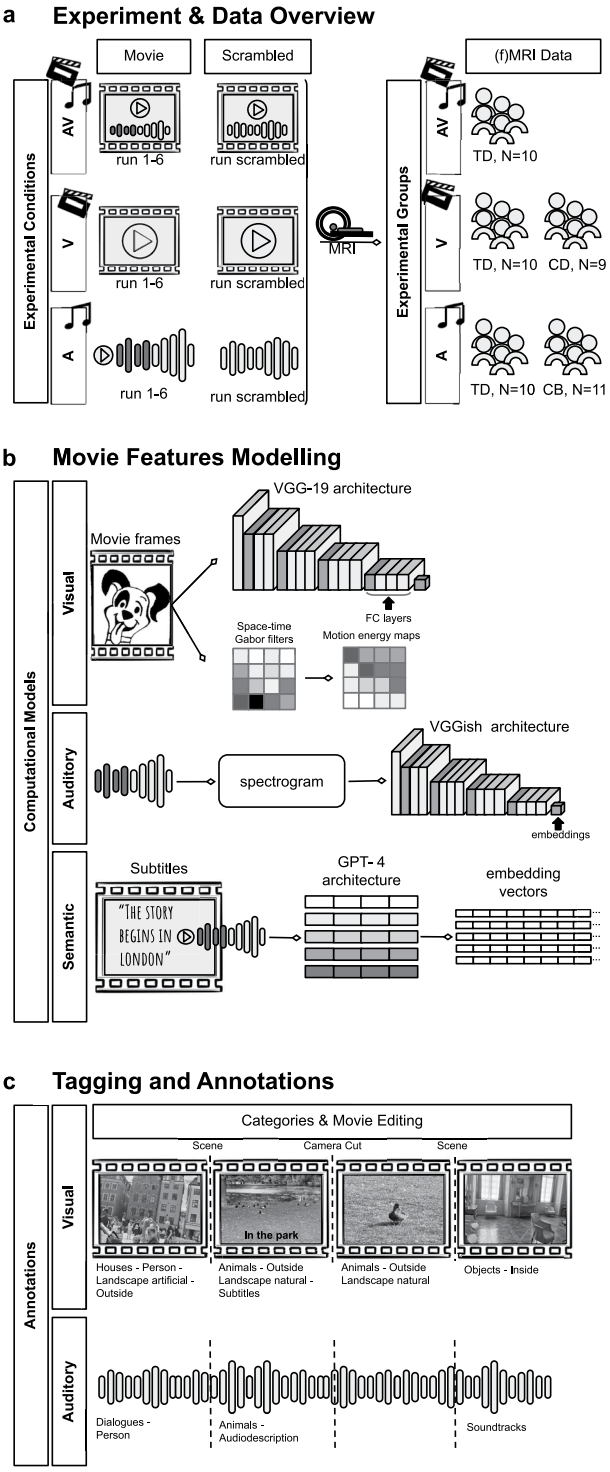

Fig. 2 Experimental protocol, movie features modelling and annotations. (**a**) The neural correlates of a full audiovisual (AV), a visual-only (V) and an auditory-only (A) versions of the same movie were presented to three distinct samples of Typically-Developed (TD) participants. Moreover, two groups of individuals with congenitally Sensory-Deprivation (SD), namely blind and deaf participants, listened and watched the two unimodal conditions (i.e., A, V) respectively. (**b**) We provide a brief description of the features extracted through computational modeling from the movie. Movie-related features fall into two categories: (i) low-level acoustic (e.g., spectral and VGGish properties) and visual features (e.g., features extracted from the VGG-19 architecture and motion energy information based on Gabor's spatiotemporal filters); and (ii) semantic descriptors captured by GPT-4 word embedding features. (**c**) Annotations for visual and auditory categories (e.g., animals, faces, objects, vehicles, body parts) and for the properties of the movie related to the editing (e.g., scenes, camera cuts, subtitles, soundtracks) are illustrated in the figure and detailed in the Methods section.

Landscape: this category describes the environment in which the action unfolds or the scene is narrated. It can be either natural (e.g., countryside) or artificial (e.g., cityscape).

Objects: this category includes man-made items and tools.

Person: this class encompasses images in which the complete silhouette of the body is visible, or where the head and the upper body are shown. Isolated faces do not fall into this category.

Vehicles: this category is intended to include all means of transportation depicted in the movie (e.g., cars, bicycles, trucks) or parts of these vehicles that are sufficiently large and detailed to be recognizable (e.g., a bicycle handlebar or tire; a car hood).

Auditory categories. The categorial content of the auditory movie was described following the same criteria employed for the visual stimuli. Therefore, the auditory track of each run was sampled at a one-second resolution, and salient sounds were categorized. To ensure consistency across models, we utilized the same categories as those applied to the visual stimuli whenever feasible. Consequently, each sound was classified according to the following labels.

Animals: this category encompasses all types of animal sounds that are audible in the foreground or are clearly distinguishable from the background noise.

Houses: this refers to the descriptions of houses, buildings, or cityscapes. Typically conveyed through the narrator's voice, these portrayals explicitly reference the presence of a building which may not be limited to the traditional "houses". Examples include phrases such as "in front of Anita's house" and "outside, from the castle gate".

Objects: this category includes explicit references to objects or sounds produced by man-made items or tools, such as the ringing of a bell, the sound of a shower, or the clicking of teacups.

Person: the presence of a person is primarily indicated through speech or dialogues. This category encompasses descriptions of a person's appearance and sounds that can be distinctly attributed to human activity, such as footsteps, coughing, background chatter, and screaming.

Vehicles: this category includes sounds produced by vehicles as well as descriptions of their appearance. It contains all onomatopoeic sounds associated with vehicles such as "wrooooom", "beep", "slam", and "screech".

Movie editing features. The stylistic choices made by the editor - such as the selection of camera cuts and the arrangement of scenes- contribute to the unique features of the movie's framework, which may influence brain activity[25,26] as well. Indeed, for instance, a change in the scene setting can correspond to modification in image luminance and be associated with corresponding adjustments in the music which thus affect perceptual and cognitive processes. To investigate whether and to what extent these formal aspects have an impact in stimulus perception, we modeled what we referred to as "movie editing" features. This term encompasses not only the editor's choices but also the significant modifications we introduced such as the addition of audio descriptions and subtitles. We modeled the movie editing features using the same approach devised for the annotation of the movie categories: four visual classes (i.e., cuts, scenes, subtitles, and text embedded in the visual frames) and three auditory classes (i.e., audio descriptions, dialogues, and soundtracks) were binary tagged within a 1-second window, as detailed below.

Scenes: this term pertains to significant alterations in the narrative context, encompassing aspects such as location, characters, actions, and temporal elements. Consequently, we define a scene as a narrative unit occurring within a particular location and timeframe. In contrast to cuts, these events unfold over a more extended temporal scale.

Camera cuts: this term refers to abrupt alterations in camera angle, position, and placement between consecutive shots. Such occurrences are prevalent throughout the narrative and can be readily identified.

Audio descriptions: all components of the movie script are presented through the narrator's voice-over which communicates the essential elements of the narrative. Specifically, audio descriptions primarily consist of scene portrayals to enhance contextual understanding, as well as depictions of characters' actions and emotional states. It is important to note that this category excludes dialogues, environmental sounds, and musical elements.

Dialogues: it pertains solely to the segment of the discourse delivered by an individual. This encompasses both conversations and "monologues" (for instance, a reporter presenting news to the audience or a priest addressing the congregation in a church).

Soundtracks: all pieces of music and songs featured in the movie

Subtitles: this descriptor indicates the presence of subtitles on the screen. It reflects the narrative script, including all the spoken elements such as the voice-over, dialogues, and ambient sound (primarily consisting of animal noises).

Data Records

Defaced structural images, as well as raw and preprocessed fMRI data, were organized according to the Brain Imaging Directory Structure (BIDS), and are available at[27] (https://figshare.com/s/de0d52bf08280cf0c4bd). Please refer to Fig. 3 for an overview of the data structure.

The present work included a set of computational models that describe the auditory, visual, and semantic properties of the movie. Furthermore, we provide a set of annotation files that identify stimulus categories for both the visual movie and the audio description. Additionally, a comprehensive characterization of the timing files is provided, detailing the entity (e.g., horse) associated with each superordinate category (e.g., Animals) for every tagged event. Each README file offers a description of the available content for each main file directory. Please be aware that the original video and audio files utilized in the study are available only upon reasonable request to the corresponding author to comply with intellectual property rights and copyright laws.

Anatomical MRI

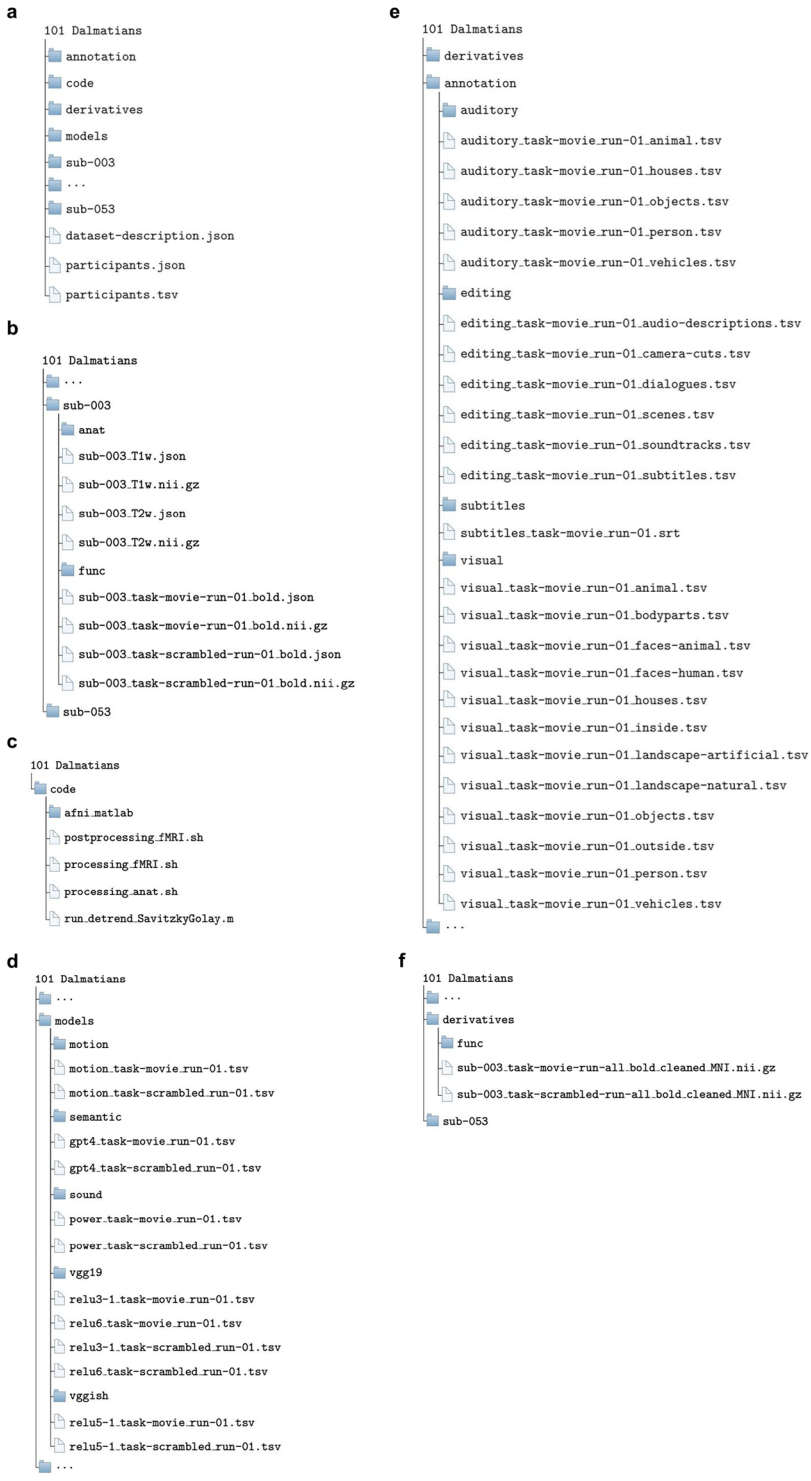

Fig. 3 Organization of the data collection. (**a**) General overview of the directory structure. (**b**) Content of subject-specific anatomical and raw data directories. (**c**) Code to preprocess (f)MRI data. (**d**) Content of computational modelling. (**e**) Content of the annotation directory. (**f**) Content of subject-specific preprocessed data directories.

Location: sub-<ID>/anat/sub-<ID>_T1w.nii.gz
File format: NIfTI, gzip-compressed.
Sequence protocol: sub-<ID>/anat/sub-<ID>_T1w.json.
The defaced raw high-resolution T1w anatomical image.
Location: sub-<ID>/anat/sub-<ID> _T2w.nii.gz
File format: NIfTI, gzip-compressed.

Sequence protocol: sub-<ID>/anat/sub-<ID>_T2w.json.
The defaced raw high-resolution T2w anatomical image.
Functional MRI, raw data
Location: sub-<ID>/func/sub-<ID>_task-movie_run-[1-6]_bold.nii.gz
File format: NIfTI, gzip-compressed.
Sequence protocol: sub-<ID>/func/sub-<ID>_task-movie_run-[1-6]_bold.json.
Location: sub-<ID>/func/sub-<ID>_task-scrambled_run-01_bold.nii.gz
File format: NIfTI, gzip-compressed.
Sequence protocol: sub-<ID>/func/sub-<ID>_task-scrambled_run-01_bold.json.
Functional MRI, preprocessed data
Location:
derivatives/sub-<ID>/func/sub-<ID>_task-movie_run-all_bold_cleaned_MNI.nii.gz
derivatives/sub-<ID>/func/sub-<ID>_task-scrambled_run-all_bold_cleaned_MNI.nii.gz
File format: NIfTI, gzip-compressed.
Annotations
Location:
annotation/auditory/auditory_task-movie_run-[1-6]_<category>.tsv
annotation/auditory/auditory_task-scrambled_run-01_<category>.tsv
File format: tab-separated values
Annotations of categories tagged in the auditory version of the movie
Location:
annotation/visual/visual_task-movie_run-[1-6]_<category>.tsv
annotation/visual/visual_task-scrambled_run-01_<category>.tsv
File format: tab-separated values
Annotations of categories tagged in the visual version of the movie
Location:
annotations/editing/editing_task-movie_run-[1-6]_audio-descriptions.tsv
annotations/editing/editing_task-movie_run-[1-6]_camera-cuts.tsv
annotations/editing/editing_task-movie_run-[1-6]_dialogues.tsv
annotations/editing/editing_task-movie_run-[1-6_scenes.tsv
annotations/editing/editing_task-movie_run-[1-6]_soundtracks.tsv
annotations/editing/editing_task-movie_run-[1-6]_subtitles.tsv
annotations/editing/subtitles/subtitles_task-movie_run-[1-6].tsv
annotations/editing/subtitles_task-scrambled.tsv
File format: tab-separated values
Annotations of editing features
Location:
annotations/subtitles/subtitles_task-movie_run-[1-6].srt
annotations/subtitles/subtitles_task-scrambled_run-01.srt
File format: SubRip Subtitle format for movie subtitles
Movie subtitles
Models
Location:
models/motion/motion_task-movie_run-[1-6].tsv
models/motion/motion_task-scrambled_run-01.tsv
File format: tab-separated values
Motion energy features[14]
Location:
models/semantic/gpt4_task-movie_run-[1-6].tsv
models/semantic/gpt4_task-scrambled_run-01.tsv
File format: tab-separated values
GPT-4 embeddings[16]
Location:
models/sound/power_task-movie_run-[1-6].tsv
models/sound/power_task-scrambled_run-01.tsv
File format: tab-separated values
Power spectrum of the auditory stream[10]
Location:
models/vgg19/relu3-1_task-movie_run-[1-6].tsv.gz
models/vgg19/relu6_task-movie_run-[1-6].tsv
models/vgg19/relu3-1_task-scrambled_run-01.tsv.gz
models/vgg19/relu6_task-scrambled_run-01.tsv
File format: tab-separated values
VGG-19 features from the visual stream[13]
Location:
models/vggish/relu5-1_task-movie_run-[1-6].tsv
models/vggish/relu5-1_task-scrambled_run-01.tsv
File format: tab-separated values
VGGish features from the auditory stream[15]

Technical Validation

The accuracy of responses to the questionnaire was evaluated to ensure story comprehension and participants' compliance. To check for the fMRI data quality, we measured framewise displacement (FD) and inter-subject correlation (ISC)[28] analysis.

Behavioral results. After the fMRI scanning session ended, participants were asked to answer a true-false list of questions concerning the major events occurring in the movie plot. The mean accuracy was 90 ± 1.2% (mean ± standard deviation), 87 ± 1.3%, and 85 ± 1.3%for TD subjects performing the audiovisual, visual and auditory movie conditions respectively. Congenitally blind and congenitally deaf individuals performed with a mean accuracy of 76 ± 3.2% and 87 ± 1.5% respectively. The questions below provide a few examples drawn from the questionnaire.

A. Cruella is the owner of a fashion house that produces silk and linen garments (T/F)
B. The dalmatians puppies are kidnapped (T/F)
C. Pongo asks for the help of the other dogs (T/F)

Framewise displacement. To control for head motion, we computed the mean framewise displacement (FD) across scans for all individuals. The average FD per run across all the participants was 0.12 ± 0.10 mm (mean ± standard deviation; see Fig. 4A), while 5.7 ± 12.6% of the timepoints across runs and participants retained an average framewise displacement greater than 0.3 mm.

Inter-subject correlation. An inter-subject correlation (ISC)[28] analysis was conducted to assess the reliability of brain responses evoked by the processing of the audiovisual, auditory, and visual versions of the live-action movie 101 *Dalmatians* in both participants with typical development and congenital sensory deprivation. For each voxel, we extracted the preprocessed time series of brain activity and calculated the average Pearson correlation coefficient (r) for all possible subject pairs. Average ISC across participants and experimental groups is reported in Fig. 4B. Our ISC analysis revealed robust and anatomically consistent synchronization across participants. We observed strong ISC in auditory, visual and multisensory cortices (e.g., bilateral STG/STS), as well as in higher-order regions such as the angular gyrus and medial prefrontal cortex. These results are consistent with findings in the ISC literature and support the functional integrity and signal quality of our data. Specifically, in our dataset, the peak ISC during the audiovisual condition (r = 0.45) was comparable to those reported in recent studies using multisensory naturalistic stimulations (e.g., r ~ 0.25 in[29]; r ~ 0.28 in[30]; r ~ 0.50 in[31]; r ~ 0.60 in[32]).

Usage Notes

The *101 Dalmatians* dataset has the potential to improve our understanding of how the brain processes and interprets complex sensory and linguistic information as it occurs in real-life situations. To this regard, the inclusion of two groups of individuals with congenital sensory loss provides the unique opportunity to gain new insights into the mechanisms governing brain functions by investigating the role of visual and auditory experiences on the development of the functional architecture of the human brain. Additionally, the comprehensive modeling of the movie properties through manual annotations and computational techniques may facilitate the investigation of the cortical representation of a wide gamut of stimulus feature spaces. Below, we summarize and discuss some important points which should be considered when working with the 101 Dalmatians dataset.

- The three conditions (i.e., AV, A, V) represent distinct versions of the same movie. In the audiovisual (AV) condition, participants are exposed to both the auditory and visual streams of the movie, which includes the narrator's voice and accompanying subtitles. The two unimodal conditions are essentially variations of the audiovisual condition, differing only in that one modality is silenced. Specifically, the audio-only (A) presentation features the original auditory track along with the narrator's voice, while the visual-only (V) presentation consists of the corresponding visual clips accompanied by subtitles.
- As far as concern the participants, we acknowledge that the deaf individuals were not age-matched with the other experimental groups, as they were, on average, younger than the other samples (i.e., deaf: N = 9, mean age 24 ± 4 years; blind: N = 11, mean age 46 ± 14 years; TD: AV condition, N = 10, mean age 35 ± 13 years; N = 10, A condition, mean age 39 ± 17 years; V condition, N = 10, mean age 37 ± 15 years). This age discrepancy originates from the unique nature of congenitally deprived samples and the challenges associated with their recruitment, which limits the possibility of age-based selection.
- SD groups were matched on educational attainment (Table 1), but not on age, reflecting demographic constraints of the target populations. Congenitally blind individuals are particularly rare in younger age groups, whereas older deaf individuals often have lower levels of formal education, reducing their eligibility for neuroimaging studies.
- In relation to the movie stimuli, it is important to note that the subtitles, speech and lip movements in the visual condition are not fully congruent, as the subtitles and speech are in Italian while the acting in the original movie is in English, a common practice in Italy for dubbed films.
- The participant groups are relatively small (n = 9–11), which is a common constraint in fMRI research with populations with a congenital sensory loss. While the dataset provides rich, naturalistic neural recordings, users should remain mindful that the limited group sizes may constrain statistical power, particularly for analyses requiring fine-grained subgroup comparisons or sensitivity to inter-individual variability.
- No formal psychometric testing was conducted as part of the current protocol. All participants successfully completed the paradigm, which required sustained attention to a feature-length movie and engagement with post-scan tasks, and no participants were excluded for comprehension or attentional difficulties. Nevertheless,

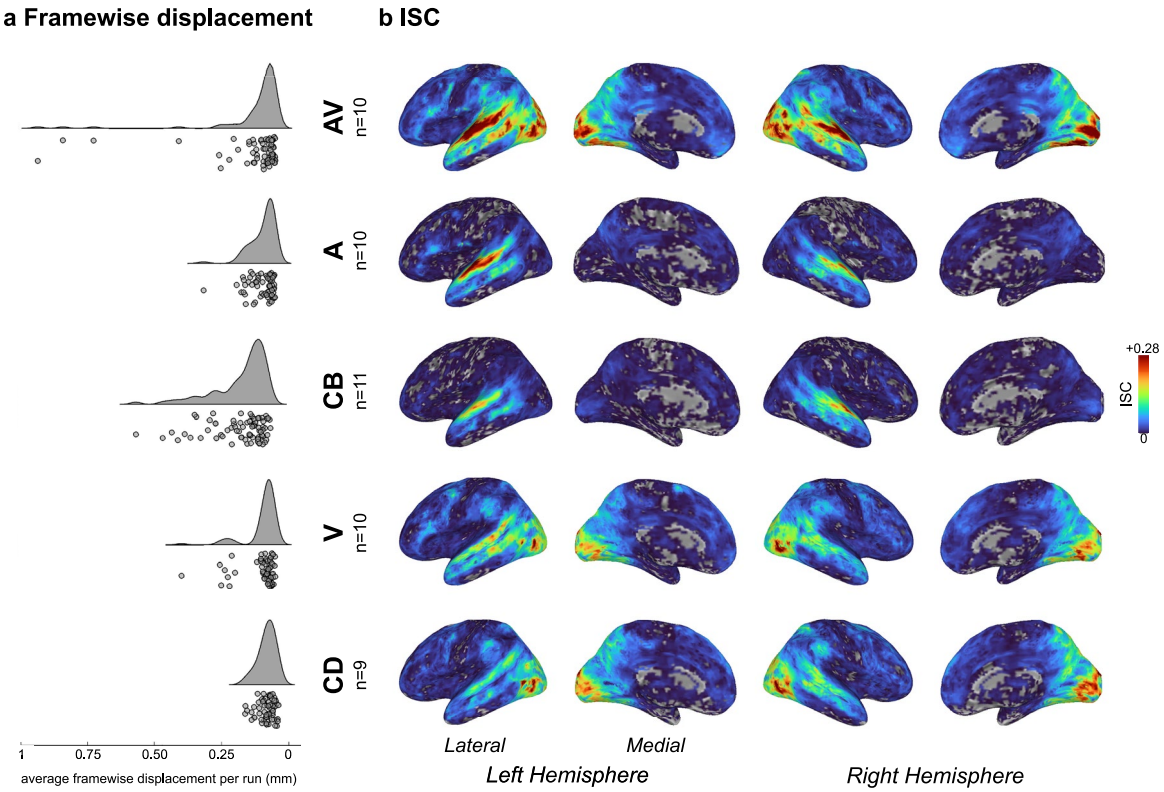

Fig. 4 Validation procedure. (**a**) Framewise displacement (FD). (**b**) Intersubject correlation (ISC). Results are shown for the five experimental groups: audiovisual (AV), audio-only (A), video-only (V), congenitally blind (CB), and congenitally deaf (CD). For typically developed groups, ISC peaks were observed in the left superior temporal gyrus (STG) for AV ($r = 0.45$; MNI_{xyz}: −65, +14, +1), in the right lateral occipital cortex (LOC) for V ($r = 0.40$; MNI_{xyz}: +52, −74, −2), and in the left STG for A ($r = 0.41$; MNI_{xyz}: −65, −11, +4). For sensory-deprived groups, ISC peaked in the right STG for CB ($r = 0.27$; MNI_{xyz}: +64, −8, −2) and in the right LOC for CD ($r = 0.40$; MNI_{xyz}: +49, −74, +1). The color scale in the figure was adjusted to match the group with the lowest ISC values.

the absence of standardized cognitive assessments remains a limitation of this dataset, and users should take this into account when conducting group comparisons.

- For what concerns the manual annotations of movie visual and auditory categories, it is worth noting that the tagging procedure was performed by a single trained human rater. This may represent a possible limitation and should be taken into account by future users.
- For the visual tagging, the distinction between "Face" and "Person" was guided by perceptual salience, functional relevance, and established neurocognitive models of visual category processing. Specifically, the "Face" label was reserved for close-up faces where internal facial features (eyes, nose, mouth) were clearly visible and visually dominant within the frame. These typically occupied a minimum of ~5–10% of the image's visual area and were free of occlusion or substantial crowding. Faces that appeared smaller, more distant, or embedded within full-body views were instead labeled "Person," reflecting a more holistic, socially integrated representation. This helps future users of the dataset more precisely model face-selective neural responses versus broader person-level representations.
- Two separate annotation tracks, one for the auditory stream and one for the visual stream, were intentionally created based on the modality-specific versions of the stimulus (i.e., the audio-only and video-only edits) that were presented during scanning. This choice reflects the ecological reality that, in natural settings, auditory and visual inputs often convey non-redundant or asynchronous information. For example, off-screen dialogue or ambient sounds may be present in the audio but not visually represented, while silent actions, facial expressions, or on-screen text may be exclusive to the visual modality. Rather than treating these discrepancies as limitations, our annotation strategy was designed to preserve them, allowing researchers to explore how the brain responds to modality-specific content as well as how it integrates across sensory streams. To ensure accurate alignment with participants' perceptual experience, annotations were created directly on the audio-only and video-only versions of the movie, not on the full audiovisual version. This approach guarantees that each annotation set faithfully reflects the information available in its respective sensory condition, as experienced during fMRI acquisition. This design not only supports analyses focused on unimodal

processing, but also opens the door to novel investigations of multisensory integration, temporal prediction, and perceptual completion under naturalistic conditions.

Data availability

Defaced structural images, as well as raw and preprocessed fMRI data, were organized according to the Brain Imaging Directory Structure (BIDS), and are available at[27] (https://figshare.com/s/de0d52bf08280cf0c4bd). Please refer to Fig. 3 for an overview of the data structure.

Code availability

The code to preprocess (f)MRI data is publicly available in the repository under code/ subdirectory[27]. The code includes bash scripts for the preprocessing of anatomical and functional data using ANTs, AFNI and FSL software. The code for the ISC analysis is available on Github (https://github.com/giacomohandjaras/101_Dalmatians). An alternative fMRI preprocessing pipeline is available here[5].

Received: 31 January 2025; Accepted: 30 September 2025;
Published online: 17 November 2025

References

1. Hasson, U., Malach, R. & Heeger, D. J. Reliability of cortical activity during natural stimulation. *Trends Cogn. Sci.* **14**, 40–48 (2010).
2. Zhang, Y., Kim, J. H., Brang, D. & Liu, Z. Naturalistic stimuli: A paradigm for multiscale functional characterization of the human brain. *Curr. Opin. Biomed. Eng.* **19**, 100298 (2021).
3. Sonkusare, S., Breakspear, M. & Guo, C. Naturalistic stimuli in neuroscience: critically acclaimed. *Trends Cogn. Sci.* **23**, 699–714 (2019).
4. Setti, F. *et al*. A modality-independent proto-organization of human multisensory areas. *Nat. Hum. Behav.* **7**, 397–410 (2023).
5. Lettieri, G. *et al*. Dissecting abstract, modality-specific and experience-dependent coding of affect in the human brain. *Sci. Adv.* **10**, eadk6840 (2024).
6. Kral, A. & Sharma, A. Developmental neuroplasticity after cochlear implantation. *Trends Neurosci.* **35**, 111–122 (2012).
7. Collignon, O. *et al*. Functional specialization for auditory–spatial processing in the occipital cortex of congenitally blind humans. *Proc. Natl Acad. Sci. USA* **108**, 4435–4440 (2009).
8. Avants, B. B., Tustison, N. & Song, G. Advanced normalization tools (ANTS). *Insight J.* **2**, 365 (2009).
9. Cox, R. W. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* **29**, 162–173 (1996).
10. Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R. & Collins, D. L. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage* **47**, S102 (2009).
11. Ciric, R. *et al*. Mitigating head motion artifact in functional connectivity MRI. *Nat. Protoc.* **13**, 2801–2826 (2018).
12. de Heer, W. A. *et al*. The hierarchical cortical organization of human speech processing. *J. Neurosci.* **37**, 6539–6557 (2017).
13. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
14. Heeger, D. J., Simoncelli, E. P. & Movshon, J. A. Computational models of cortical visual processing. *Proc. Natl Acad. Sci. USA* **93**, 623–627 (1996).
15. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. Preprint at https://arxiv.org/abs/1409.1556 (2014).
16. Nishimoto, S. *et al*. Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol.* **21**, 1641–1646 (2011).
17. Hershey, S. *et al*. CNN architectures for large-scale audio classification. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*. 131–135 (IEEE, 2017).
18. Achiam, J. *et al*. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
19. Simonelli, F. *et al*. Sensitivity and specificity of the action observation network to kinematics, target object, and gesture meaning. *Hum. Brain Mapp.* **45**, e26762 (2024).
20. Cadena, S. A. *et al*. Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Comput. Biol.* **15**, e1006897 (2019).
21. Welch, P. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Trans. Audio Electroacoust.* **15**, 70–73 (1967).
22. Martinelli, A. *et al*. Auditory features modelling reveals sound envelope representation in striate cortex. Preprint at https://doi.org/10.1101/2020.04.15.043174v2 (2020).
23. Tuckute, G., Feather, J., Boebinger, D. & McDermott, J. H. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *PLoS Biol.* **21**, e3002366 (2023).
24. Kell, A. J. E. & McDermott, J. H. Invariance to background noise as a signature of non-primary auditory cortex. *Nat. Commun.* **10**, 395 (2019).
25. Herbec, A., Kauppi, J. P., Jola, C., Tohka, J. & Pollick, F. E. Differences in fMRI intersubject correlation while viewing unedited and edited videos of dance performance. *Cortex* **71**, 341–348 (2015).
26. Cao, Z. *et al*. Exploring the combined impact of color and editing on emotional perception in authentic films: Insights from behavioral and neuroimaging experiments. *Humanit. Soc. Sci. Commun.* **11**, 1–14 (2024).
27. Setti, F. *et al*. 101 Dalmatians: a multimodal naturalistic fMRI dataset in typical development and congenital sensory loss. *figshare* https://doi.org/10.6084/m9.figshare.28264178 (2025).
28. Hasson, U., Nir, Y., Levy, I., Fuhrmann, G. & Malach, R. Intersubject synchronization of cortical activity during natural vision. *Science* **303**, 1634–1640 (2004).
29. Meliss, S., Pascua-Martin, C., Skipper, J. I. & Murayama, K. The magic, memory, and curiosity fMRI dataset of people viewing magic tricks. *Sci. Data* **11**, 1063 (2024).
30. Aliko, S., Huang, J., Gheorghiu, F., Meliss, S. & Skipper, J. I. A naturalistic neuroimaging database for understanding the brain using ecological stimuli. *Sci. Data* **7**, 347 (2020).
31. Visconti di Oleggio Castello, M., Chauhan, V., Gu, J. & Gobbini, M. I. An fMRI dataset in response to *The Grand Budapest Hotel*, a socially-rich, naturalistic movie. *Sci. Data* **7**, 383 (2020).
32. Pinho, A. L. *et al*. Individual Brain Charting dataset extension, third release for movie watching and retinotopy data. *Sci. Data* **11**, 590 (2024).

Acknowledgements
This work has been supported by a PRIN grant by the Italian Ministry of University and Research granted to E.R. (20223K8B3X) and P.P. (P20228PHN2), and by Next Generation EU, Ecosistema dell'Innovazione "the–Tuscany Health Ecosystem"—code ecS00000017 to E.R., P.P., L.C., G.H. Additionally, F.S. was supported by the Frontier Proposal Fellowship (FPF program, 2019) granted by IMT School for Advanced Studies Lucca; M.D. was supported by the PRIN (Grant ID: 2022PNJS5Z). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank the Unione Italiana Ciechi e Ipovedenti (The Italian Union of the Blind and Partially Sighted) and the Ente Nazionale Sordi Onlus (The Italian Union of the Deaf) for their support.

Author contributions
F.S., G.H., D.B., A.L., P.P. and E.R. conceived and designed the study. F.S., A.L. and G.H. curated the methodological and analytical pipeline. F.S. curated the movie editing, the creation of movie audio description and subtitles and the annotation procedure. F.S., A.L. and M.D. performed the study and collected the fMRI data. F.S., G.H., F.G., V.B. and C.T. handled participants' recruitment. F.G., D.B., L.C., P.P. and E.R. supervised the project. F.G., P.P. and E.R., took care of the project administration and funding acquisition. G.H. and F.S. prepared the data in BIDS format. All the authors took part in writing and reviewing the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-06077-3.

Correspondence and requests for materials should be addressed to E.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.