nature portfolio

Peer Review File

Mapping Functional Homologies Between Human and Marmoset Brain Networks Using Movie-Driven Ultra-High Field fMRI

Corresponding Author: Dr Alessandro Zanini

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors aimed to examine within- and between-species temporal correlations to assess functional homologies between human and common marmosets using movie-driven fMRI.
By analyzing functional networks using tensor ICA and hierarchical clustering analysis, they found strong interspecies correspondence in sensory networks, whereas networks associated with higher-order cognition were only observed in humans. Their approach allowed to identify both conserved and divergent neural dynamics, which would provide novel insights about species-specific brain dynamics such as functional differences of default mode network, which could be difficult to detect using the anatomical approach.
This approach would be valuable for cross-species comparisons in other animals, such as mice and Old World monkeys, to better understand the unique features of human brain dynamics. It may also enhance our understanding of marmosets and other species as translational models for studying brain function and investigating neurological diseases.
The study highlights the importance of examining both shared and divergent brain network architectures through precise cross-species comparisons between marmosets and humans.
While many previous studies have emphasized similarities to humans to support the utility of experimental animals, understanding the differences is equally valuable. This paper contributes to a deeper understanding of the human brain and strengthens the rationale for using marmosets and other species as preclinical models for brain disorders.

There are some questions and concerns about the paper.

1. Description of movie-driven fMRI
Movie-driven fMRI is a key method in this study, enabling interspecies comparisons of functional networks. The authors describe naturalistic movies as an ecologically valid alternative to traditional task-based paradigms. However, the rationale for why such stimuli elicit robust and reliable neural responses across sensory and cognitive systems in awake animals remains somewhat unclear. It would strengthen the manuscript to include a more detailed explanation, along with relevant references, to support this claim.
Additionally, it would be helpful if the authors could describe the criteria or considerations used when selecting the specific movie stimuli.

2. Information about human participants:
It would be helpful to include information about the native language of the human participants to better understand the extent to which they could comprehend the content of the documentary stimuli. This is particularly relevant for interpreting potential activation of language-related brain networks.

3. Details on stimulus structure:
More detailed information about the structure of the stimulus presentation would be valuable. For instance, the timing and arrangement of baseline periods and documentary segments (e.g., baseline – 1st documentary (10 min) – baseline – 2nd documentary (10 min)) should be clarified to improve reproducibility and interpretability.

4. Repeated viewing by marmosets (Line 116):

The authors mention that marmosets watched the movie five times across five sessions. It would be helpful to clarify whether the same 33-minute movie was repeated in each session. If so, were the data from the five sessions averaged or concatenated for analysis? Also, it might be worth discussing whether repeated viewings could affect the animals' attention to the stimuli, and how that was addressed.

5. Discussion of data and pipeline availability:
Although the authors state in the abstract and summary that their data and analysis pipelines are available to readers, this point is not well discussed in the main text. A more detailed explanation of how others can access and use these resources would strengthen the transparency and reproducibility of the work.

6. Abbreviations:
I noticed that some abbreviations, such as tensor ICA (tICA), are introduced multiple times throughout the manuscript. It would be more concise and conventional to define each abbreviation only once upon first mention and then use the abbreviated form consistently thereafter.

Reviewer #2

(Remarks to the Author)
In the current paper, Zanini and colleagues explore the functional networks during naturalistic stimulation in human and marmoset brains and evaluate the inter-species similarities and differences between the functional architecture. This is an extremely interesting question, and the application of naturalistic species-relevant stimuli makes the approach attractive.

Overall, the study is well presented and methodologically appropriate. However, some interpretations that are currently speculative could be strengthened with the existing data with additional analyses. Moreover, some evaluations of the temporal correlations should be supported by statistics and more consistent language. Thus, I would welcome a revision addressing the following few observations.

1. The interpretation of the component HM seems to be based on visual inspection and is used for reverse inference of its potential functions in multiple places in the manuscript. However, to me, this network looks like one of the lateralized frontoparietal “executive” networks that are often referred to in the literature. Rather, the component HB appears more similar to the typical dorsal attention network that, at least during naturalistic stimulation, usually includes the middle temporal visual areas in addition to the intraparietal sulcus and frontal eye fields.
For instance, the authors state in the discussion (Lines 619–621):

“The executive component, strongly lateralized and resembling the dorsal attention stream (50–53,77), may reflect uniquely human mechanisms for sustained attention and goal-directed processing.”

Seems like this is unnecessarily speculative based on reverse inference and subjective evaluation of the shape of the component. Could there be anything that might be inferred from the activity timecourse in relation to the stimuli here? There are various papers that evaluate the temporal behavior of independent components linked to experiment phases and stimulus features to explain the functional roles of the components, e.g. Bartels & Zeki 2005 (doi:10.1098/rstb.2005.1627), Ylipaavalniemi et al. 2009 (doi:10.1016/j.neuroimage.2009.03.056), Lahnakoski et al. 2012 (doi:10.1371/journal.pone.003521), among others.
If it seems unreasonable to include such analyses here, I would suggest to at least expand on this topic in the limitations section, where it is already touched upon, as not employing information on the stimuli limits the reliability of the interpretations that can be made.

2. The results section contains a lot of speculative functional interpretations without references. I would recommend removing such speculation from the results and leave them for the discussion.
For instance the lines 418–422 (“Despite substantial differences in cortical expansion and specialization, this cluster points to a high degree of evolutionary conservation in sensory integration…”) would be better placed in the discussion and potentially supported by some references.
Similarly e.g. lines 437–438 and 449–460 would be better placed in the discussion, but there are also other cases.

3. Also related the previous points, on lines 271–273 the authors write:

“The auditory component (AUD) showed moderate correlations with both HOV and SUV components ($r = 0.36–0.50$), indicating cross-modal interactions and integrative processing across sensory modalities.”

Arguably, these correlations could be solely driven by temporal correlations between auditory and visual features in the stimuli. Thus, I would be cautious claiming integrative processing without additional evidence. I believe the later description of the marmoset networks is more appropriately framed (377–379):

“reflecting the integrated audiovisual nature of the stimulus and the simultaneous activation of visual and auditory pathways”.

4. Line 369–372:

“The second cluster was a sensorimotor-executive-insular group, comprising the somatomotor (SOM), executive 1 (EXE1),

and insular (INS) networks, which also showed strong mutual correlations (r = 0.58–0.75), suggesting co-activation related to action observation, executive demands, or salience-driven processes."

Together, these components seem to encompass areas similar to what is often called the "salience network" or, in other instances a set of similar areas might be called the "pain matrix". I wonder if the classification of HD as an executive network is informative in this context? Arguably, naming conventions like "salience network" are also oversimplifications that may not be helpful in the long term, but may help in discussing the correlations between these components in relation to existing literature.

5. Line 385–386:
"The DMN component exhibited little to no correlation with other networks, in line with its established role in internally oriented cognition."

The DMN is, indeed, often discussed in relation to internally oriented thought, but it is also often reported to exhibit negative correlations with other functional networks, especially the dorsal attention network. However, during naturalistic stimulation, the functions of DMN subregions may be more complicated and strong anticorrelations might not occur. Perhaps rather than just referring to internally oriented cognition in the results, the topic could be brought up in some more detail in the discussion?

6. The correlations between functional networks are discussed somewhat inconsistently. Sometimes correlations ranging from approximately 0.2–0.6 are said to be moderate to strong and sometimes correlations of 0.2–0.9 are just said to be strong. Please state the thresholds for the levels used in the paper explicitly and use them consistently.

7. Relatedly, on line 495 the authors mention "significant temporal alignment", but how was the significance tested?

8. Similarly, on the following page, lines 513–514, the authors state that the "marmoset components were more temporally synchronous", but how was this tested? Explicit statistical tests of the temporal correlations would be important when making such claims and could strengthen the paper throughout.

9. Lines 593–595:

"This discrepancy may reflect reduced detectability of subcortical signals in human imaging, or fundamental differences in the degree to which subcortical regions form temporally coherent networks".

With the current analyses, it seems there would be an opportunity to test this directly. What was the tSNR in the comparable subcortical structures in this data? Or were any of the excluded noise components showing higher weights in the basal ganglia?
ICA decompositions of course can vary with the number of components extracted, initialization of the algorithm and intersubject variability in group analyses. Perhaps a sentence on specific issues with (t)ICA on component reliability would be of interest here.

10. Lines 605–611:

"In humans, no cerebellar network was extracted, consistent with known difficulties in resolving posterior fossa activity in whole-brain ICA (50–53, but see 74). However, targeted ICA approaches have demonstrated the cerebellum's rich functional architecture (75), and its absence here likely reflects limitations in sensitivity rather than true functional differences."

Here, my impression is that the (group level) ICA often returns a single component for the whole cerebellum that may often be classified as noise as it is a mixture of multiple functional subregions, is likely affected by motion and sometimes limited image coverage. Did any of the "noise" components include areas of the cerebellum? Perhaps also differences in the localization of functional regions between participants may complicate finding representative group components.

Finally, there are some minor things that could be improved.

1. Stimuli-section in Methods: It would be helpful to show a simple timeline of the stimulus schedule in figure 1. Now it is unclear e.g. what the lengths of the alternating excerpts were. Were they shown as single blocks or did they alternate back and forth?

2. The font sizes in the figures are inconsistent and sometimes too small to read. Particularly in Figures 1, 3 and 6 there are some elements that are illegible.

3. In figure 3, the inlays for the auditory cortices appear counterproductive. With the gray overlay indicating the location on the cortex makes it seem that there is no signal in the auditory regions as the gray overlay obscures the activity and the closeup inlays are entirely empty with no activity. As far as I can tell, the auditory subregions are also not referred to in the text, so the benefit of this detailed view seems unclear.

4. Lines 176–177: "To assess spatial coverage, non-noise components were binarized and averaged to generate probability

maps.”

How where the components binarized? That is, how was it defined which areas are considered to be a part of the network? In principle, all the components cover the entire brain with varying weights, so some kind of threshold (e.g. p or t-value) has to be defined for the significant part of the network, which would also affect the coverage and overlap of components.

5. Lines 381–381: “The LAN component showed minimal temporal overlap with either sensory or attentional networks”

Maybe temporal correlation rather than overlap? The timecourses are not binary.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have provided the datasets, analysis scripts, and demographic information in the Zenodo repository (https://doi.org/10.5281/zenodo.12746414). However, I was unable to access the link. Could you please confirm whether the URL is correct?

Reviewer #2

(Remarks to the Author)
In my evaluation, the authors have thoroughly addressed all the comments of both reviewers, and I am happy to indorse the publication of the revised manuscript.

Mapping Functional Homologies Between Human and Marmoset Brain Networks Using Movie-Driven Ultra-High Field fMRI

Review

Reviewer #1:

The authors aimed to examine within- and between-species temporal correlations to assess functional homologies between human and common marmosets using movie-driven fMRI. By analyzing functional networks using tensor ICA and hierarchical clustering analysis, they found strong interspecies correspondence in sensory networks, whereas networks associated with higher-order cognition were only observed in humans. Their approach allowed to identify both conserved and divergent neural dynamics, which would provide novel insights about species-specific brain dynamics such as functional differences of default mode network, which could be difficult to detect using the anatomical approach.
This approach would be valuable for cross-species comparisons in other animals, such as mice and Old World monkeys, to better understand the unique features of human brain dynamics. It may also enhance our understanding of marmosets and other species as translational models for studying brain function and investigating neurological diseases.
The study highlights the importance of examining both shared and divergent brain network architectures through precise cross-species comparisons between marmosets and humans.
While many previous studies have emphasized similarities to humans to support the utility of experimental animals, understanding the differences is equally valuable. This paper contributes to a deeper understanding of the human brain and strengthens the rationale for using marmosets and other species as preclinical models for brain disorders.

We thank the reviewer for taking the time to carefully read our manuscript and for providing such insightful and encouraging comments. We greatly appreciate the recognition of our approach and its potential for advancing cross-species comparisons and translational neuroscience.

There are some questions and concerns about the paper.

1. Description of movie-driven fMRI

Movie-driven fMRI is a key method in this study, enabling interspecies comparisons of functional networks. The authors describe naturalistic movies as an ecologically valid alternative to traditional task-based paradigms. However, the rationale for why such stimuli elicit robust and reliable neural responses across sensory and cognitive systems in awake animals remains somewhat unclear. It would strengthen the manuscript to include a more detailed explanation, along with relevant references, to support this claim. Additionally, it

would be helpful if the authors could describe the criteria or considerations used when selecting the specific movie stimuli.

We appreciate the reviewer's request for clarification. We have now expanded the Introduction to explain why naturalistic movies are increasingly used in fMRI research as an ecologically valid alternative to conventional paradigms. Such stimuli provide continuous, multisensory input that reliably engages a broad range of sensory and cognitive systems in both humans and nonhuman primates, while remaining reproducible across repeated presentations and participants. We also clarify the criteria used for stimulus selection: we chose excerpts from nature documentaries that contain rich visual and auditory information, including social interactions, animal behavior, and diverse environments, which are engaging for both humans and marmosets and have been shown to elicit reliable neural responses across species. Moreover, the documentaries we selected include numerous scenes depicting marmoset or macaque monkeys, stimuli triggering a strong engagement in our marmoset sample. Relevant references have been added to support these points.

Given the importance of this aspect, we have also added a sentence to the Discussion to highlight the possible connection between the suggested stronger tuning of the brain to naturalistic (rather than "traditional") stimuli and our finding of robust interspecies convergence in sensory networks.

The modified parts of the Introduction, Methods, and Discussion now read as:

Lines 65-76:

Movie-driven fMRI (md-fMRI) has emerged as a promising alternative. Naturalistic movie stimuli offer an ecologically valid alternative to conventional task paradigms by providing continuous, multisensory input that more closely resembles real-life experience. Prior work has shown that movies elicit robust and reliable neural responses across visual, auditory, and higher-order cognitive systems both in humans (27–33) and non-human primates (14,34,35), making them particularly suitable for comparative and translational research. Importantly, movie stimuli allow for the simultaneous engagement of multiple brain systems while preserving experimental control and repeatability. Crucially, such stimuli also maintain the engagement of non-human primates during long fMRI sessions. Finally, recent literature suggests that the brain may be more strongly 'tuned' to naturalistic than to artificial stimuli, such that movies and other real-world narratives can evoke more reliable and representative patterns of neural activity across sensory and cognitive systems (for review, see (36)).

Lines 109-116:

The movie contained a wide variety of visual (e.g., marmosets, humans, animals, cityscapes, landscapes) and auditory stimuli (e.g., speech, vocalizations, music, environmental sounds), including dynamic social and ecological content. These features were selected to ensure engagement of multiple sensory modalities and higher-order networks in both humans and marmosets, while maintaining relevance and interest across species. Moreover, the documentaries were selected based on their frequent inclusion of scenes depicting marmoset or macaque monkeys, able to trigger the engagement of our marmoset participants.

Lines 669-671:

Importantly, emerging evidence suggests that the brain is more strongly tuned to naturalistic than to artificial stimuli (for review, see (35), which supports and is consistent with our observation of robust interspecies convergence in sensory networks using movie-driven fMRI.

2. Information about human participants

It would be helpful to include information about the native language of the human participants to better understand the extent to which they could comprehend the content of the documentary stimuli. This is particularly relevant for interpreting potential activation of language-related brain networks.

We agree with the reviewer that information about the participants' language background is important for interpreting potential involvement of language-related networks. We have therefore added to Zenodo an Excel table including detailed demographic information for each participant (sex, age, prior scanning experience, and languages spoken), which complements the data and code already provided. The Methods section of the manuscript has been updated to reflect this addition:

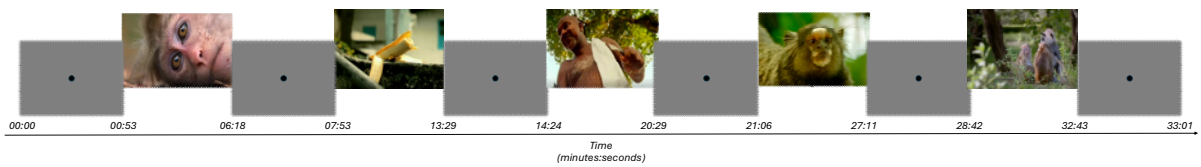
Lines 99-100:

Demographic information including sex, age, prior experience with MRI scanning, and languages spoken are available in the Zenodo repository (https://doi.org/10.5281/zenodo.12746414).

3. Details on stimulus structure

More detailed information about the structure of the stimulus presentation would be valuable. For instance, the timing and arrangement of baseline periods and documentary segments (e.g., baseline – 1st documentary (10 min) – baseline – 2nd documentary (10 min)) should be clarified to improve reproducibility and interpretability.

We appreciate the reviewer's suggestion to provide more detail on the stimulus structure. We have expanded the Methods section to introduce the Supplementary Figure 1, which clarifies the timing and arrangement of baseline and documentary segments. This figure illustrates the exact sequence of baseline and movie blocks, with start and end times indicated, to improve clarity and reproducibility.



Supplementary Figure 1. Structure of the naturalistic movie stimulus. The 33-minute stimulus alternated between baseline blocks and documentary excerpts. Baseline blocks consisted of a fixation target (black circle on a gray background) presented without audio. Documentary blocks consisted of excerpts from *Monkey Kingdom*

(Disneynature, Spanish narration) and *Hidden Kingdoms – Urban Jungles* (BBC, English narration). The duration of each block and its start/end times (min:s) are shown below the schematic.

4. Repeated viewing by marmosets (Line 116)

The authors mention that marmosets watched the movie five times across five sessions. It would be helpful to clarify whether the same 33-minute movie was repeated in each session. If so, were the data from the five sessions averaged or concatenated for analysis? Also, it might be worth discussing whether repeated viewings could affect the animals' attention to the stimuli, and how that was addressed.

We are grateful for the reviewer's helpful comment. We now clarify in the Methods that the same 33-minute movie was presented in each of the five sessions. To minimize possible reductions in attention due to repeated exposure, sessions were separated by at least two weeks, and the vigilance state of the animals was monitored with the MRI-compatible camera to ensure that the eyes were open during movie presentation. In addition, we specify in the Statistical Analysis section that the data from the five sessions were averaged per animal before performing tICA. The modified sections now report:

Lines 130-134:

Each marmoset watched the same 33-minute naturalistic movie across five separate sessions. To minimize possible reductions in attention caused by repeated exposure, sessions were spaced at least two weeks apart. The vigilance state of the animals was monitored throughout using the MRI-compatible camera, ensuring that the eyes remained open during stimulus presentation.

Lines 185-186:

For the marmosets, the data from the five sessions were averaged within each animal prior to performing tICA.

5. Discussion of data and pipeline availability

Although the authors state in the abstract and summary that their data and analysis pipelines are available to readers, this point is not well discussed in the main text. A more detailed explanation of how others can access and use these resources would strengthen the transparency and reproducibility of the work.

We thank the reviewer for highlighting this important point, that is also one of the main aims of our study. To strengthen transparency and reproducibility, we have added a paragraph at the end of the Discussion that provides a more detailed explanation of the resources we have shared, including both the data and the full analysis pipeline. The paragraph (lines 672-679) now reads:

To facilitate transparency and reproducibility, we have made all datasets and analysis scripts used in this study openly available on Zenodo (https://doi.org/10.5281/zenodo.12746414). The repository contains the individual- and sample-level fMRI data for both humans and marmosets, as well as the full preprocessing and analysis pipelines used in the present work. A

detailed text file is provided to guide users through the structure of the repository and the steps necessary to reproduce our results. By sharing both the data and the computational workflow, we aim to enable other researchers to replicate our analyses, adapt our pipeline to their own data, and extend cross-species comparisons in future studies.

6. Abbreviations

I noticed that some abbreviations, such as tensor ICA (tICA), are introduced multiple times throughout the manuscript. It would be more concise and conventional to define each abbreviation only once upon first mention and then use the abbreviated form consistently thereafter.

We thank the reviewer for pointing this out. We have revised the manuscript to ensure that each abbreviation, including tensor ICA (tICA), is defined only upon its first occurrence and that the abbreviated form is used consistently thereafter.

Reviewer #2:

In the current paper, Zanini and colleagues explore the functional networks during naturalistic stimulation in human and marmoset brains and evaluate the inter-species similarities and differences between the functional architecture. This is an extremely interesting question, and the application of naturalistic species-relevant stimuli makes the approach attractive.

Overall, the study is well presented and methodologically appropriate. However, some interpretations that are currently speculative could be strengthened with the existing data with additional analyses. Moreover, some evaluations of the temporal correlations should be supported by statistics and more consistent language. Thus, I would welcome a revision addressing the following few observations.

We are grateful to the reviewer for the positive assessment of our study and for recognizing the interest and value of our approach. We also appreciate the constructive feedback regarding the interpretation of results, the need for statistical support in some evaluations, and the consistency of our language. We have carefully revised the manuscript to address each of these points in detail, as outlined below.

1. The interpretation of the component HM seems to be based on visual inspection and is used for reverse inference of its potential functions in multiple places in the manuscript. However, to me, this network looks like one of the lateralized frontoparietal "executive" networks that are often referred to in the literature. Rather, the component HB appears more similar to the typical dorsal attention network that, at least during naturalistic stimulation, usually includes the middle temporal visual areas in addition to the intraparietal sulcus and frontal eye fields.

For instance, the authors state in the discussion (Lines 619–621):

"The executive component, strongly lateralized and resembling the dorsal attention stream (50–53,77), may reflect uniquely human mechanisms for sustained attention and goal-directed processing."

Seems like this is unnecessarily speculative based on reverse inference and subjective evaluation of the shape of the component. Could there be anything that might be inferred from the activity timecourse in relation to the stimuli here?

There are various papers that evaluate the temporal behavior of independent components linked to experiment phases and stimulus features to explain the functional roles of the components, e.g. Bartels & Zeki 2005 (doi:10.1098/rstb.2005.1627), Ylipaavalniemi et al. 2009 (doi:10.1016/j.neuroimage.2009.03.056), Lahnakoski et al. 2012 (doi:10.1371/journal.pone.003521), among others.

If it seems unreasonable to include such analyses here, I would suggest to at least expand on this topic in the limitations section, where it is already touched upon, as not employing information on the stimuli limits the reliability of the interpretations that can be made.

We appreciate the reviewer's thoughtful comments regarding the interpretation of components HB and HM. We agree that their spatial characteristics overlap with both dorsal attention and frontoparietal executive networks described in the literature. Our initial labeling was based on anatomical distribution: component HM included the intraparietal sulcus (LIP, VIP) and large portions of the frontal eye fields (both FEF and premotor eye field), but also engaged broader prefrontal regions while lacking superior temporal involvement. This frontal engagement was the main reason for labeling HM as an "executive 2" network. At the same time, the presence of both IPS and FEF suggested a possible resemblance to the "dorsal attention" network. In contrast, HB involved superior temporal and parietal regions, with only minimal FEF engagement, suggesting a higher-order visual/attentional role.

We recognize that our statement in the Results section — *"Notably, component HB engaged the frontal eye fields (FEF) and lateral intraparietal cortex, resembling a saccade or attentional control network"* — may have overstated the FEF involvement in this network. Moreover, the choice to represent each network with a single slice per view in Figure 5, although effective for summarizing all components at once, limits the visualization of their full spatial extent. To address this, we have made all functional components (including noise components) available with the preprocessed data in NIfTI format, allowing readers to inspect the complete spatial distributions.

We acknowledge that reliance on spatial distribution alone has limitations and that reverse inference can be speculative. We regret if our manuscript gave the impression that our labels were intended as definitive categorizations; our intention was to provide functional shorthand for interpretation, while recognizing that alternative classifications are possible. Accordingly, we have revised the Results and Discussion sections to (1) remove speculative statements from the Results, and (2) explicitly acknowledge possible alternative labeling of these components.

We fully agree with the reviewer that a deeper examination of the temporal dynamics of these components in relation to stimulus features would provide a stronger basis for functional interpretation, as demonstrated by prior work already highlighted by the reviewer (e.g., Bartels & Zeki, 2005; Ylipaavalniemi et al., 2009; Lahnakoski et al., 2012). Conducting such analyses across all identified components represents an important next step. Given the scope of the present study, we have chosen not to expand the analyses here, but we are currently preparing a follow-up manuscript using these same datasets to systematically investigate component timecourses and their relation to stimulus features. This should allow not only a better labelling of our networks, but eventually also a better understanding of the clustering within-species and (we hope) the correlation between-species. To reflect this, we have revised the limitations section to highlight that our interpretations are primarily based on spatial patterns, and that future work will address the temporal dynamics in detail.

The limitations section (lines 693-700) now reads:

Finally, our functional interpretations were primarily based on the spatial distribution of components. While this strategy is common in the literature, it involves a degree of reverse inference. For example, while we labeled component HM as an executive network, and component HB as a higher-order visual/attentional network, their respective inclusion and exclusion of regions such as the frontal eye fields and superior temporal cortex illustrate how different plausible interpretations can arise depending on the spatial emphasis. Analyses linking component timecourses to stimulus features would provide stronger evidence for functional roles (33,81,82).

2. The results section contains a lot of speculative functional interpretations without references. I would recommend removing such speculation from the results and leave them for the discussion.

For instance the lines 418–422 ("Despite substantial differences in cortical expansion and specialization, this cluster points to a high degree of evolutionary conservation in sensory integration…") would be better placed in the discussion and potentially supported by some references.
Similarly e.g. lines 437–438 and 449–460 would be better placed in the discussion, but there are also other cases.

We thank the reviewer for raising this to our attention. We agree that several paragraphs in the Results section report speculations and/or interpretations that are in most cases already discussed in the Discussion. We thus modified the Results section, removing the speculative parts, and edited the Discussion, adding the interpretations and the references that were missing.

The edited parts in the Discussion are the following:

Lines 499-505:

We observed moderate to strong cross-species correlations among core sensory systems, particularly those supporting audiovisual processing. HOV, SUV, and AUD networks formed tightly correlated clusters in both species. These findings suggest a shared temporal structure in how primate brains process dynamic audiovisual input, despite differences in cortical architecture and specialization (32,63). The high degree of within- and between-species temporal correlation among these networks reinforces the idea that sensory integration, especially in naturalistic contexts, is supported by evolutionarily conserved mechanisms.

Lines 507-515:

In contrast, components linked to higher cognitive functions — including language, executive control, and default mode processes — were uniquely observed in humans and did not correlate with any marmoset networks. The LAN network, for instance, exhibited strong left-lateralized recruitment of frontal and temporal regions traditionally associated with speech comprehension and semantic processing. Its absence in marmosets aligns with the lack of linguistic capacity in non-human primates and underscores the human specificity of this network (64). Similarly, a left-lateralized executive component resembling the dorsal attention network emerged only in humans, suggesting functional specialization in attentional control and top-down modulation not mirrored in the marmoset brain.

Lines 582-612:

An intriguing aspect of our findings lies in the interpretation of the marmoset DMN. While anatomically similar to the human DMN, the marmoset component labeled as DMN clustered with visual and occipitoparietal networks and showed strong temporal coupling with sensory and subcortical components—patterns not observed in humans. The human DMN was largely functionally segregated, exhibiting weak correlations with sensory and executive systems. Furthermore, the marmoset DMN does not show temporal coupling with the human DMN, whereas it shows moderate correlations with the human executive and insular networks, which respectively resemble known cognitive control and salience networks in humans (53–55,60). This discrepancy highlights a critical gap in our understanding of the marmoset DMN: unlike the human DMN, which is extensively characterized in the literature (53,54,56,60,61), resting-state studies in marmosets report inconsistent findings, with some identifying networks that resemble the human DMN anatomically and others reporting divergent patterns (11–13,48–51).

These observations raise the possibility that the so-called marmoset DMN may serve a different function, perhaps related to visual attention or environmental monitoring, rather than internally directed thought or mind-wandering as in humans (78). The lack of interspecies correlation between DMN components, despite periods of reduced stimulation in the movie, underscores the need to reconsider the functional equivalence of these networks across species. However, it is important to note that during naturalistic stimulation, DMN subregions may not exhibit the strong anticorrelations with sensory and attentional networks typically reported in resting-state studies (79–82). Instead, emerging evidence suggests that DMN activity in such contexts may be more dynamic and multifaceted, supporting processes such as narrative comprehension, social cognition, or integration of external and internal information (83,84). This broader functional repertoire could account for the atypical pattern of within-species correlations observed in our human sample and may help explain why the human DMN showed limited correlations with the marmoset DMN: while the human network may flexibly adapt to complex cognitive demands, the marmoset counterpart might engage in more basic or distinct functions, leading to reduced interspecies correspondence during naturalistic stimulation. Given the rich, naturalistic design and cross-species comparability of our paradigm, this work may offer a valuable framework to further clarify the architecture and role of the DMN in marmosets and its potential evolutionary divergence from the human counterpart.

Lines 512-515:

Similarly, a left-lateralized executive component resembling the frontoparietal executive network (53–56) emerged only in humans, suggesting a functional specialization not mirrored in the marmoset brain.

Lines 656-659:

The EXE2 component, strongly lateralized and resembling the frontoparietal executive network described in resting-state studies (53–56,83), may reflect uniquely human mechanisms for sustained attention, memory orgoal-directed processing consistent with evolutionary expansions in lateral prefrontal cortex.

3. Also related the previous points, on lines 271–273 the authors write:

“The auditory component (AUD) showed moderate correlations with both HOV and SUV components (r = 0.36–0.50), indicating cross-modal interactions and integrative processing across sensory modalities.”

Arguably, these correlations could be solely driven by temporal correlations between auditory and visual features in the stimuli. Thus, I would be cautious claiming integrative processing without additional evidence. I believe the later description of the marmoset networks is more appropriately framed (377–379):

“reflecting the integrated audiovisual nature of the stimulus and the simultaneous activation of visual and auditory pathways”.

We thank the reviewer for this important observation. We have removed the speculative claim about cross-modal integration from the Results section to avoid over-interpretation. Instead, we now describe these findings in more neutral terms in the Results and expand on their potential significance in the Discussion, where we carefully consider the role of audiovisual stimulus features in shaping the observed correlations. The edited paragraph in the Discussion now reads:

Lines 499-505:

We observed moderate to strong cross-species correlations among core sensory systems, particularly those supporting audiovisual processing. HOV, SUV, and AUD networks formed tightly correlated clusters in both species. These findings suggest a shared temporal structure in how primate brains process dynamic audiovisual input, despite differences in cortical architecture and specialization (32,63). The high degree of within- and between-species temporal correlation among these networks reinforces the idea that sensory integration, especially in naturalistic contexts, is supported by evolutionarily conserved mechanisms.

4. Line 369–372:

“The second cluster was a sensorimotor-executive-insular group, comprising the somatomotor (SOM), executive 1 (EXE1), and insular (INS) networks, which also showed strong mutual correlations (r = 0.58–0.75), suggesting co-activation related to action observation, executive demands, or salience-driven processes.”

Together, these components seem to encompass areas similar to what is often called the "salience network" or, in other instances a set of similar areas might be called the "pain matrix". I wonder if the classification of HD as an executive network is informative in this context? Arguably, naming conventions like "salience network" are also oversimplifications that may not be helpful in the long term, but may help in discussing the correlations between these components in relation to existing literature.

We appreciate the reviewer’s insightful suggestion. We agree that the co-activation of somatomotor, insular, and executive regions can be interpreted in light of broader systems often referred to as the *salience network* or, in other contexts, the *pain matrix*. To address this point, we have expanded the Discussion to note the overlap between these designations, the challenges of rigid network labelling, and the possibility that the observed correlations reflect a more general

salience- and context-driven mechanism rather than a purely executive function. We also highlight that this aspect may represent an interesting avenue for future work, as examining the dynamic engagement of these networks during naturalistic stimulation could provide further insight into their shared and distinct computational roles across species.

The new paragraph in the Discussion now reads:

Lines 570-580:

Otherwise, the co-activation of somatomotor, insular, and executive regions may also reflect engagement of broader integrative systems often referred to as the salience network (57,76) or, in other contexts, the so-called pain matrix (77). These overlapping designations highlight both the importance and the difficulty of assigning unique functional labels to networks that flexibly participate in multiple processes, including salience detection, interoception, and cognitive control. In our study, the relations observed between somatomotor, executive, and insular components may therefore index a more general mechanism for orienting attention and behavior toward salient, socially and biologically relevant events in the movie, rather than reflecting a single "executive" function. This underscores the need for caution in applying rigid naming conventions and supports a more dynamic view of large-scale networks as multifunctional systems recruited according to contextual demands.

5. Line 385–386:

"The DMN component exhibited little to no correlation with other networks, in line with its established role in internally oriented cognition."

The DMN is, indeed, often discussed in relation to internally oriented thought, but it is also often reported to exhibit negative correlations with other functional networks, especially the dorsal attention network. However, during naturalistic stimulation, the functions of DMN subregions may be more complicated and strong anticorrelations might not occur. Perhaps rather than just referring to internally oriented cognition in the results, the topic could be brought up in some more detail in the discussion?

We thank the reviewer for this thoughtful suggestion. We agree that DMN functions during naturalistic stimulation are more complex than a simple association with internally oriented cognition, and that anticorrelations with attentional networks may not be as prominent as in resting-state contexts. To address this point, we have expanded the Discussion to highlight the possibility that DMN subregions perform more dynamic and varied functions in naturalistic settings, which may help explain the unusual correlation patterns observed in our study.

Lines 599-609:

However, it is important to note that during naturalistic stimulation, DMN subregions may not exhibit the strong anticorrelations with sensory and attentional networks typically reported in resting-state studies (79–82). Instead, emerging evidence suggests that DMN activity in such contexts may be more dynamic and multifaceted, supporting processes such as narrative comprehension, social cognition, or integration of external and internal information (83,84). This broader functional repertoire could account for the atypical pattern of within-species correlations observed in our human sample and may help explain why the human DMN showed limited correlations with the marmoset DMN: while the

human network may flexibly adapt to complex cognitive demands, the marmoset counterpart might engage in more basic or distinct functions, leading to reduced interspecies correspondence during naturalistic stimulation.

6. The correlations between functional networks are discussed somewhat inconsistently. Sometimes correlations ranging from approximately 0.2–0.6 are said to be moderate to strong and sometimes correlations of 0.2–0.9 are just said to be strong. Please state the thresholds for the levels used in the paper explicitly and use them consistently. We thank the reviewer for pointing out this inconsistency. We have now explicitly stated the thresholds used to classify correlation strengths (weak: $r < 0.2$; moderate: 0.2–0.5; strong: $r \geq 0.5$) and have applied these definitions consistently throughout the manuscript.

Lines 188-190:

We computed within- and between-species correlation matrices of component timecourses (following conventional benchmarks, we classified correlations as weak ($r < 0.2$), moderate ($0.2 < r < 0.5$), or strong ($r > 0.5$)).

7. Relatedly, on line 495 the authors mention "significant temporal alignment", but how was the significance tested?

8. Similarly, on the following page, lines 513–514, the authors state that the "marmoset components were more temporally synchronous", but how was this tested? Explicit statistical tests of the temporal correlations would be important when making such claims and could strengthen the paper throughout.

We appreciate the reviewer's suggestion for a more rigorous assessment of the temporal correlations between human and marmoset components. In response, we implemented a non-parametric permutation procedure to statistically evaluate these correlations. Specifically, the human timecourses were kept fixed while the marmoset timecourses were randomly circularly shifted (N = 5000 permutations) to generate a null distribution of correlation values for each pair of human–marmoset components. Observed correlations were then compared against this null distribution to compute p-values, which were subsequently corrected for multiple comparisons using the FDR (Benjamini–Hochberg) procedure. Significant correlations are now indicated with asterisks in Figure 8.

We also note that our original statement that "marmoset components were more temporally synchronous" may have been misleading. Our intention was to highlight that, in general, the within-species correlations among marmoset components tended to be higher than those among human components, rather than implying formal statistical testing of synchrony. With the addition of the permutation-based analysis for the interspecies correlations, we now provide a principled framework to assess significance, clarifying the interpretation of temporal correspondence across species and thereby strengthening the robustness and reproducibility of our findings. The Methods section has been updated to describe this procedure in detail, the statement about temporal synchrony has been edited and the figure legend has been modified to clarify how significance is represented.

Lines 190-197:

To assess the statistical significance of cross-species temporal correlations, we implemented a non-parametric permutation test. For each pair of human–marmoset components, we kept the human timecourses fixed and applied random circular shifts to the marmoset timecourses (N = 1000 permutations), recomputing the correlations at each iteration. This procedure generated a null distribution of correlations for each component pair, against which the observed correlation was compared to obtain a p-value. Resulting p-values were corrected for multiple comparisons across the matrix using false discovery rate (FDR) adjustment (Benjamini–Hochberg).

Lines 517-521:

Human components exhibited greater segregation, with distinct correlation clusters reflecting sensory, cognitive, and default mode systems. In contrast, correlations among marmoset components were generally higher, reflecting a tendency toward more broadly shared temporal fluctuations across networks.

Legend Figure 8:

Figure 8. Cross-species correlation matrix*. Matrix showing temporal correlations (Pearson's r) between the 12 marmoset and 14 human functional components. Warm colors indicate positive correlation; cool colors indicate anticorrelation. Two major interspecies clusters emerge: a conserved multisensory cluster (upper left), including visual, auditory, and audiovisual networks, and a divergent cognitive cluster (lower right) comprising somatomotor, executive, and default mode components. Language and lateralized executive networks in humans show little or no correspondence in the marmoset. Only correlations with an absolute r value greater than 0.10 are displayed. Asterisks indicate correlations that reached statistical significance after FDR correction based on a permutation test (p < 0.05, p < 0.01**, p < 0.001***).*

9. Lines 593–595:

"This discrepancy may reflect reduced detectability of subcortical signals in human imaging, or fundamental differences in the degree to which subcortical regions form temporally coherent networks".

With the current analyses, it seems there would be an opportunity to test this directly. What was the tSNR in the comparable subcortical structures in this data? Or were any of the excluded noise components showing higher weights in the basal ganglia?

ICA decompositions of course can vary with the number of components extracted, initialization of the algorithm and intersubject variability in group analyses. Perhaps a sentence on specific issues with (t)ICA on component reliability would be of interest here.

We are grateful to the reviewer for this insight. We directly assessed tSNR in cortical and subcortical structures in both species. In humans, mean tSNR was comparable between cortex and subcortex, indicating that the absence of a basal ganglia network cannot be explained only by signal quality. In marmosets, despite slightly lower subcortical tSNR, a basal ganglia network

was readily extracted, supporting the idea that pure signal quality cannot explain the extraction of subcortical components. None of the human components (both selected and noise) showed clearly elevated weights in subcortical structures, although these regions are included in a more distributed manner across components. We also acknowledge that the parameters chosen for temporal ICA—such as the number of components, algorithm initialization, and intersubject variability—may contribute to differences in component extraction between species. These points have been added to the Discussion to clarify potential causes of the observed discrepancy. We also modified the Methods and Results section to include the new analysis.

Lines 209-212:

Specifically, for both cortical and subcortical regions, mean tSNR values were extracted using binary masks from the corresponding atlases. Paired t-tests were then performed to compare cortical and subcortical signal quality, to determine whether differences in tSNR could confound the detection of subcortical networks in either species.

Lines 281-283:

Quantitative comparison of mean tSNR between cortical and subcortical regions revealed significantly higher signal quality in the cortex (paired t-test: $t_{(39)} = 3.01$, $p = 0.0046$).

Lines 385-388:

Similarly, comparison of mean tSNR between cortical and subcortical regions did not reveal a significant difference (paired t-test: t(18) = -1.45, p = 0.164), suggesting that the lack of a basal ganglia network in the human sample is unlikely to be explained only by signal quality differences.

Lines 620-632:

This discrepancy is unlikely to reflect differences in signal quality, as the mean tSNR in human subcortical structures was comparable to that of cortical regions, whereas the marmoset exhibited lower subcortical tSNR yet still yielded a clear basal ganglia network. Across all human components—both those retained for analysis and those identified as noise—no component showed distinctly higher weights in basal ganglia regions, suggesting that these areas are represented more diffusely within the human functional architecture. . While some human studies do extract basal ganglia components (12,56,85,86), others do not (53–55), pointing to variability in detectability and functional coupling depending on species, analysis methods, and stimulus context. It is therefore plausible that the observed interspecies difference arises from other factors, such as differences in the intrinsic functional organization of subcortical networks or the specific parameters used for temporal ICA decomposition, including the number of components, initialization, and intersubject variability, all of which can affect component reliability.

10. Lines 605–611:

“In humans, no cerebellar network was extracted, consistent with known difficulties in resolving posterior fossa activity in whole-brain ICA (50–53, but see 74). However, targeted ICA approaches have demonstrated the cerebellum’s rich functional architecture (75), and

its absence here likely reflects limitations in sensitivity rather than true functional differences."

Here, my impression is that the (group level) ICA often returns a single component for the whole cerebellum that may often be classified as noise as it is a mixture of multiple functional subregions, is likely affected by motion and sometimes limited image coverage. Did any of the "noise" components include areas of the cerebellum? Perhaps also differences in the localization of functional regions between participants may complicate finding representative group components.

We appreciate the reviewer's insightful comment. To address this point, we carefully examined all components extracted by tICA in the human dataset. None of the selected components showed reliable recruitment of the cerebellum. A few noise components did include parts of the cerebellum, but these were not retained as meaningful networks because their spatial distributions appeared artifactual (e.g., edge-related or scattered across non-contiguous regions) and their temporal spectra were dominated by noise-like frequencies. Thus, we could not identify a robust cerebellar network in humans with the current analysis.

We agree with the reviewer that several factors could have contributed to this outcome. First, group-level ICA often fails to recover the fine-grained functional organization of the cerebellum, sometimes producing a single mixed component that is difficult to classify. Second, variability in the precise localization of cerebellar subregions across participants may further complicate group-level decomposition. Finally, technical aspects such as motion sensitivity, image coverage of the posterior fossa, and the parameter choices in tICA (e.g., number of components, initialization) may have limited our ability to recover cerebellar networks. We have now acknowledged these points as limitations in the revised manuscript.

Lines 642-645:

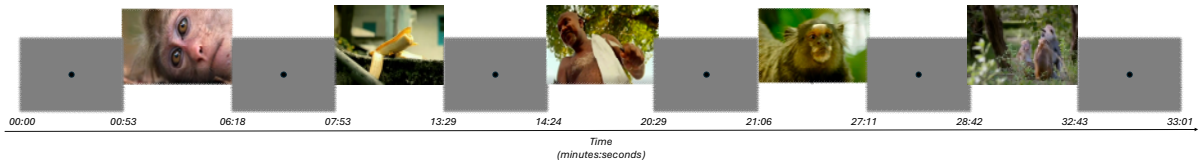
We note that although a few noise components included parts of the cerebellum, these were not reliable due to artificial spatial distributions or noise-like frequency spectra. Other factors, including tICA parameters, intersubject variability, motion, and limited image coverage, may have further limited the recovery of cerebellar networks.

Finally, there are some minor things that could be improved.

1. Stimuli-section in Methods: It would be helpful to show a simple timeline of the stimulus schedule in figure 1. Now it is unclear e.g. what the lengths of the alternating excerpts were. Were they shown as single blocks or did they alternate back and forth?

We thank the reviewer for the suggestion. Given both reviewers raised a point concerning the structure of the stimulus, we found appropriate to create a new figure (Supplementary Figure 1) describing in detail how the movie was structured.

Supplementary Figure 1 now clarifies the timing and arrangement of baseline and documentary segments. This figure illustrates the exact sequence of baseline and movie blocks, with start and end times indicated, to improve clarity and reproducibility.



Supplementary Figure 1. Structure of the naturalistic movie stimulus. The 33-minute stimulus alternated between baseline blocks and documentary excerpts. Baseline blocks consisted of a fixation target (black circle on a gray background) presented without audio. Documentary blocks consisted of excerpts from *Monkey Kingdom* (Disneynature, Spanish narration) and *Hidden Kingdoms – Urban Jungles* (BBC, English narration). The duration of each block and its start/end times (min:s) are shown below the schematic.

2. The font sizes in the figures are inconsistent and sometimes too small to read. Particularly in Figures 1, 3 and 6 there are some elements that are illegible.

We thank the reviewer for this useful suggestion. We increased the font size in Figures 1, 3, and 6 to improve readability, adjusting it as much as possible without compromising figure clarity. However, given the relatively small size and high density of several ROIs, particularly in the marmoset brain, further enlarging the labels would have resulted in overcrowding and reduced legibility of the maps. We believe the revised figures now represent the best balance between readability and clarity.

3. In figure 3, the inlays for the auditory cortices appear counterproductive. With the gray overlay indicating the location on the cortex makes it seem that there is no signal in the auditory regions as the gray overlay obscures the activity and the closeup inlays are entirely empty with no activity. As far as I can tell, the auditory subregions are also not referred to in the text, so the benefit of this detailed view seems unclear.

We thank the reviewer for this valuable insight. We have modified Figure 3 accordingly: the internal gray overlay has been removed and replaced with only the borders of the auditory cortex, ensuring that the underlying activity remains visible. Additionally, the insets now display a zoomed view of the actual activations from the flat map, providing a clearer visualization of this small region of the marmoset brain. While it is true that we do not specifically refer to individual auditory ROIs in the main text, we believe that presenting the complete set of ROIs for both species offers a useful reference for readers, particularly those less familiar with human or marmoset brain labeling, and may serve as a resource for those interested in specific regions.

4. Lines 176–177: "To assess spatial coverage, non-noise components were binarized and averaged to generate probability maps."

How where the components binarized? That is, how was it defined which areas are considered to be a part of the network? In principle, all the components cover the entire

brain with varying weights, so some kind of threshold (e.g. p or t-value) has to be defined for the significant part of the network, which would also affect the coverage and overlap of components.

We thank the reviewer for pointing this out. In our study, the component spatial maps extracted with MELODIC are expressed as z-statistic images. These z-scores represent the relative contribution of each voxel to a given independent component, based on a mixture model separating signal from noise. For binarization, we used the component-specific threshold values automatically reported by MELODIC for each independent component (all reported on Figures 2 and 5 for each specific component). These thresholds emphasize voxels most strongly associated with the component according to the model fit. While sometimes loosely interpreted as corresponding to a $p < 0.05$ criterion under Gaussian assumptions, these thresholds should be understood primarily as a heuristic provided by MELODIC to separate structured signal from background noise. We have revised the Methods to explicitly state this procedure for clarity.

Lines 202-207:

To assess spatial coverage, non-noise components were converted to z-statistic images by MELODIC and thresholded using the component-specific values automatically estimated by the mixture-model fit. These thresholds emphasize voxels most strongly associated with each component, and should be understood as a heuristic for separating structured signal from background noise rather than as a formal voxelwise statistical test. Then, thresholded maps were binarized and averaged to generate probability maps.

5. Lines 381–381: "The LAN component showed minimal temporal overlap with either sensory or attentional networks"

Maybe temporal correlation rather than overlap? The timecourses are not binary.

We thank the reviewer for the suggestion. We now edited the sentence, who reads:

Lines 416-417:

The LAN component showed minimal temporal correlation with either sensory or attentional networks