nature portfolio

Peer Review File

Modeling Macroscopic Brain Dynamics with Brain-inspired Computing Architecture

Corresponding Author: Professor Lei Deng

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)
In this paper, the authors demonstrate a computational pipeline for simulating macroscopic brain dynamics on brain-inspired computing chips and GPUs, using DMF model as the case study. They perform progressive and dynamic quantisation of the models and parallelise the model using a hierarchical mapping to leverage hardware capabilities. They analyse the characteristics of their low-precision simulations both in terms of speed and accuracy.

The key results of the paper consist of the following:

1. Simulations on both the brain-inspired chip and GPU was shown to achieve 1-2 orders of magnitude reduction in simulation time and overall inversion time.
2. The collection of quantisation approaches used in the paper was shown to achieve reasonable functional fidelity in comparisons on static and dynamic indicators.

The authors use a clever combination of ways to deal with quantisation on brain-inspired hardware. They run a warm-up period on the host for the transient dynamical phase of the system with large variations in the variables. They then use do group-wise quantisation depending on the range of different variables, and also choose different time-steps for different variables to allow their proper quantisation.
This approach is very practical, and engineering oriented, progressively addressing all the issues that arise around quantising such a simulation. While the authors show an overall advantage in simulation time, it would have been useful to see, explicitly, how much of time is spent in making decisions about quantisation.

The authors use a DMF model as a case study for their simulations. While it is reasonable to choose one model for reasons of computational effort, it would have been useful to understand how representative the DMF model is as a case study compared to other models they mention such as Wilson-Cowan Model, Kuramoto Model etc. More generally, is it often the case that the models have a transient regime and a stable fluctuation-driven regime to allow semi-dynamic quantisation? A more detailed discussion references/examples for this would be useful.

The authors perform comparisons between the quantised and full-precision simulations for both static and dynamic indicators, and provide convincing demonstrations of their qualitative similarities.
But the quantitative comparison is a bit more limited -- the only aggregate quantitative comparison that the authors provide is R^2 values using linear fitting. The implications of this were not clear to me, especially to answer the question of whether quantised simulation provides utility in the contexts it is usually used, i.e. in computational studies that are carried out using a DMF model. And whether the linear fitting results provide a reasonable measure of such utility. Discussions and examples addressing this aspect would be very useful. If possible (considering computational constraints), reproducing an existing study with the quantised model and seeing if the resulting predictions match the original studies would significantly strengthen the paper.

The hierarchical parallelism scheme described is well thought out, and the the demonstrated speed-ups on both brain-inspired hardware and GPUs are impressive. One question I had is: why not do the warmup phase of simulation on a GPU as well? It would seem that at GPU and brain-inspired chip could be combined to give the best of both worlds.
While it's clear that their method would scale well, it would be necessary to see the host time included in the statistics to truly

get an idea of the scalability of their method. I would expect that with larger model sizes, host times becomes a bottleneck (especially for the simulation on the brain inspired chip).

In summary, this is a very sound and innovative study that provides a way to engineer simulations of macroscopic brain dynamics on brain-inspired chips or GPU, providing impressive performance gains while retaining sufficient fidelity on the demonstrated use case. The paper is overall well written and clear, and results are described completely. I do not have sufficient expertise in this subfield to know if all relevant literature is discussed.
The main weaknesses of this study, in my opinion, are that (1) it looks only at one particular use case (DMF); (2) An analysis of the utility of this method in the context of existing literature is missing; and (3) the support for the scalability of the model is not fully convincing. I believe the work has the potential to be quite significant if some of these weaknesses are addressed.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The manuscript introduces a novel implementation of whole-brain mean-field simulations with neuromorphic hardware (TianjicX computing chip). This allows for faster simulations and model inversions to fit neuroimaging data by one to two orders of magnitude compared to modern CPU and GPU implementations, while still capturing dynamical aspects of more established models of whole-brain dynamics. To do so, the authors employ two strategies: (1) semi-quantization of steady-state firing rates and time leading to low-precision simulations by taking advantage of fixed points in dynamical mean-field models, and (2) massively parallel distribution of computations leveraging distributed memory storage from neuromorphic chips. The paper has potentially transformative impacts for the field of whole-brain modeling by substantially accelerating simulations and model inference, with translational implications for the development of whole-brain digital twins for diagnosis prediction and treatment optimization. However, I have several major concerns on validations and comparisons made as well as on the clarity of explanations in the manuscript.

Major points:
1. I could not understand if Figure 3 characterizes low-precision fitting of simulated dynamics from high-precision models or of empirical rsfMRI data from the HCP dataset. It is imperative to make this explicit in the text. Further, the manuscript needs to present an in-depth characterization of both. Crucially, it should be shown that the low-resolution model can capture ground-truth parameter changes in the high-resolution model and to quantify the robustness of this by varying the noise and complexity of the high-resolution model as well as the amplitude of parameter changes. It is also important to show how well the model can fit rsfMRI data - are the correlations between empirical and low-resolution model FC comparable to those between low-resolution and high-resolution FC after model fitting?
2. I am concerned about how robust and meaningful the discretization of time is. All iterative simulation methods discretize time, and it is not clear here how this differs from other implementations. Particularly, the claim that “their temporal evolution demonstrates heterogeneity characterized by distinct timescales” raises concerns. Biological processes in the brain often span temporal scales, with many having scale-invariant properties. For instance, the power spectrum of time series measured with several neuroimaging methods (M/EEG, fMRI etc) scales in 1/f. Even oscillations, that are associated with one distinct time scale, show frequencies that fluctuate through time and across brain regions even in the same individual (see for instance Donoghue et al Nat Neuro 2020). This needs to be acknowledged, with the paper’s approach to this problem much more explicitly explained, and validated. While comparisons shown between high-resolution simulations, low-resolution simulations, and rsfMRI are mainly spatial across regions, it is necessary to also show comparisons in the temporal domain such as at least the power spectral structure. Other metrics might also be used to characterize the complexity of time series such as the Lempel-Ziv complexity.
3. Relatedly, I am concerned that Euler’s method, on which the findings rely, is not adequate for simulations of chaotic systems with potentially complex attractors. Can other algorithms readily be implemented with the hardware and algorithms used here? If so, the authors should replicate at least part of the results with another integration method to demonstrate robustness.
4. I did not understand if parallelization was conducted over time at all. Parallelization over time should absolutely not be applied here, as dynamics in each time step depend on dynamics from previous time steps. More generally, how parallelization was performed here in comparison to state-of-the-art whole-brain modeling tools should be explained in much more detail. Modern implementations in languages such as Python already parallelize over regions, model realizations, etc, what is the difference here? Is it only that the memory is implemented in a distributed way instead of remotely accessed?
5. How long are the rsfMRI time series that simulations are compared to? Empirical brain dynamics are highly non-stationary as mentioned throughout the manuscript, with parameters varying and fixed points shifting through time, and I worry if quantization of firing rates can robustly account for nonstationary brain activity, when compared to high-resolution models.
6. Another important concern is scalability to larger numbers of parameters to infer. Recent work (Zhang et al, PNAS 2024) emphasizes the need for inferring more parameters, such as region-specific parameters, toward personalized whole-brain models. The manuscript should demonstrate scalability to more than 3 parameters to some extent, or acknowledge lack of scalability if this cannot be shown.
7. The manuscript methods claim that a particle swarm algorithm was adapted to perform metaheuristic inferences. How does this compare in performance and speed to state-of-the-art methods in whole-brain model inference e.g. evolutionary algorithms (Zhang et al, PNAS 2024, Cakan et al, Cog Computation 2023) and simulation-based inference (Hashemi et al,

Mach Learn Sci Technol 2024)? Are they comparable? Are other algorithms, such as those mentioned, able to be easily and readily integrated with the hardware and model presented here?
8. Especially for wider audiences unfamiliar with the details of neuromorphic hardware, applicability and flexibility of the modeling and inference methodology should be discussed to address potential concerns. For instance, are there hardware changes needed to implement a changed structural connectome? If so, how long do they take? Same question to implement different firing rate differential equations.

Minor points:
“can integrate macroscopic empirical data” - references are needed to back this point up
Ref 14-16 not about diagnosing brain disorders
“brain-inspired computing chips are increasingly oriented to intelligent computing workloads rather than scientific computing ones” - what intelligent vs scientific means is unclear
“they prioritize low-precision computing to reduce hardware resource costs and power consumption.” - references are needed here.
Fig 3b, 3d, and similar: all axes need to be labeled
From the HCP dataset, how are subjects elected? The population appears smaller than the pool of subjects for whom data is available in this large-scale dataset.
Fig S3: what is the interpretation behind higher inferred G in low-precision than high-precision models?

(Remarks on code availability)
To my knowledge, no code has been made available yet. The manuscripts mentions that code will be made available, and a link is provided to an empty Github repository.

Reviewer #3

(Remarks to the Author)
This manuscript presents a hybrid hardware-software framework for accelerating the inversion of coarse-grained whole-brain models. By introducing a dynamics-aware quantization strategy and deploying simulations on neuromorphic (TianjicX) and GPU architectures, the authors demonstrate considerable speedups (up to ×400) while maintaining key features of the underlying dynamical system.

The manuscript addresses a critical challenge in computational neuroscience — the acceleration of macroscopic brain model inversion. The proposed pipeline directly tackles this issue with a quantization framework, including semi-dynamic quantization, group-wise range adaptation, and multi-timescale integration. The implementation is both scalable and practical, leveraging TianjicX neuromorphic hardware and GPU parallelism to support large-scale parameter exploration through strong system-level integration. Notably, the work introduces a platform-enabling insight by demonstrating an unconventional yet efficient use of neuromorphic hardware (TianjicX) for non-spiking, coarse-grained dynamical modeling — thereby expanding the potential applications of such architectures beyond traditional neuron-level simulations.

The manuscript could be further improved by addressing the following points:

Generalization. The method is tested exclusively on the dynamic mean-field (DMF) model, which is smooth, low-dimensional, and relatively well-behaved. The authors do not demonstrate generalizability to models with richer or more complex dynamics (e.g., oscillatory or chaotic regimes). To broaden the impact, the authors should either include another model class or more explicitly discuss the limitations of applying the approach beyond DMF (see integration or oscillatory behavior, for example, in: 10.1201/9781003143499).

While the quantization pipeline is well-designed, the remaining components (e.g., simulation parallelism, GPU mapping) reflect well-established practices. The manuscript contributes more as a systems integration effort than a source of novel algorithmic or neuroscientific insights. The authors should highlight this point.

A critical limitation of the framework is the requirement for a warm-up phase using high-precision floating-point simulation (typically on a CPU) to determine the quantization parameters. This step breaks the standalone nature of the system and may reduce its deployability in embedded or edge applications. Moreover, it assumes convergence to a relatively stable state during warm-up — a condition that may not hold for more volatile brain dynamics. This trade-off between quantization precision and model generality deserves a clearer treatment in the manuscript.

The generalization of the proposed approach to other neuromorphic platforms, particularly, analog designs (e.g. 10.3390/app12094528). These architectures are primarily optimized for spiking neural network simulations and may not natively support the type of low-precision, non-spiking differential equation integration used in this work. Adapting the framework to these platforms would likely require substantial reengineering or conceptual reformulation. This should be discussed, particularly in light of the growing diversity in neuromorphic hardware design.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I have read the author responses to my comments and the changes made in the manuscript. All of my concerns were addressed satisfactorily.

I'm satisfied with the generalisability of this method across models. The split of times is useful, and the speed-up from the GPU is promising. The additional correlation based quantitative results support the utility of the model.

One thing the authors could optionally expand on is the analysis of the variation in the correlation with the scaling factor and bifurcation parameter in the Hopf model simulation. Does the values of scaling factors where the correlation is high also a useful regime?

Overall, my earlier comments stand on the strengths of the paper and I recommend acceptance.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors addressed my comments in a very thorough and careful manner. The clarity and quality of the manuscript has substantially improved after this round of review in my opinion. I have no more comments for the authors and believe the manuscript is ready for publication.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors addressed my concerns, and I propose Acceptance.

(Remarks on code availability)

Response Letter

Manuscript ID: NCOMMS-25-20318-T
Title: Modeling Macroscopic Brain Dynamics with Brain-inspired Computing Architecture
Authors: Zhong Zheng, Jing Wei, Yaru Xu, Chenhua Li, Tianying Lu, Qili Guo, Xueyao Ji, Hao Guo, Gang Wang, Lei Deng

We would like to thank the reviewers for spending valuable time and raising insightful comments which we feel have substantially improved our manuscript. The point-by-point responses are provided as follows. All the revisions in the manuscript and Supplementary Information (SI) are marked in blue for your convenience.

1. Responses to Reviewer 1

Overall Comment: ***In this paper, the authors demonstrate a computational pipeline for simulating macroscopic brain dynamics on brain-inspired computing chips and GPUs, using DMF model as the case study. They perform progressive and dynamic quantisation of the models and parallelise the model using a hierarchical mapping to leverage hardware capabilities. They analyse the characteristics of their low-precision simulations both in terms of speed and accuracy.***

The key results of the paper consist of the following:

1. Simulations on both the brain-inspired chip and GPU was shown to achieve 1-2 orders of magnitude reduction in simulation time and overall inversion time.

2. The collection of quantisation approaches used in the paper was shown to achieve reasonable functional fidelity in comparisons on static and dynamic indicators.

Response: We greatly appreciate the your right recognition of our contributions and insightful feedback. The responses to your comments are detailed below. We hope our responses and revisions can make you satisfied.

Comment 1: ***The authors use a clever combination of ways to deal with quantisation on brain-inspired hardware. They run a warm-up period on the host for the transient dynamical phase of the system with large variations in the variables. They then use do group-wise quantisation depending on the range of different variables, and also choose different time-steps for different variables to allow their proper quantisation. This approach is very practical, and engineering oriented, progressively addressing all the issues that arise around quantising such a simulation. While the authors show an overall advantage in simulation time, it would have been useful to see, explicitly, how much of time is spent in making decisions about quantisation.***

Response: Actually, you raise a critical point regarding the time spent on making decisions about quantization. In Table 1, we did not explicitly list the specific time cost for the quantization process. This is because it is unique to the brain-inspired computing chip, and including it in the table might make the comparison across the three architectures less straightforward. In Fig. 5a and b, we did report the time

consumed by the host in the pipeline using the brain-inspired computing chip, which includes the time for quantization. However, as you suggested, explicitly indicating the overhead of quantization would render the cost breakdown of the pipeline more transparent and understandable.

Following your valuable advice, we have compiled a more detailed breakdown of the time cost of pipeline on the brain-inspired computing chip. In the course of carefully reviewing the data and experimental code, prompted by your feedback, we identified an inadvertent overestimation of the total inversion time (i.e., $T_{overall\_inversion}$) in our initial host implementation for the GPU and brain-inspired chip pipelines. The corrected results indicate that our solutions are even faster than those reported in the previous manuscript. The detailed clarifications are as follows.

First, while we utilized multiprocessing in the CPU-based model inversion process, we had not fully leveraged the optimization for certain host-executed steps, i.e., model quantization and post-processing, in the pipelines for GPU and the brain-inspired computing chip. Second, when estimating $T_{overall\_inversion}$ on GPU and the brain-inspired computing chip, the PSO algorithm itself (including parameter updates and iterations) should consume quite little time. However, we found that our original measurements for these components inadvertently included overhead from concurrent or temporally adjacent operations.

Table R1. Computational performance of brain modeling across different platforms.

#Node	Platform	$T_{simulation}$	$T_{overall\_inversion}$
68	CPU	1978.85	2074.64
	GPU (ours)	31.10 (63.6×)	52.46 (39.5×)
	Brain-inspired (ours)	**26.32 (75.2×)**	**42.33 (49.0×)**
148	CPU	4246.34	4365.02
	GPU (ours)	109.35 (38.8×)	154.51 (28.3×)
	Brain-inspired (ours)	**47.14 (90.1×)**	**77.11 (56.6×)**
512	CPU	92426.65	93534.92
	GPU (ours)	2007.15 (46.0×)	2184.16 (42.8×)
	Brain-inspired (ours)	**217.80 (424.4×)**	**796.43 (117.4×)**

After addressing above issues, the inversion time consumed by both pipelines on GPU and the brain-inspired computing chip can be shorter (see Table R1). We sincerely appreciate your suggestion, which has enabled us to identify these issues and make the improvements.

Returning to your suggestion, we present a more detailed time breakdown for the inversion of the three representative scales of the whole-brain model in Table R2. We have divided the total time into three parts: on-chip simulation time (t_{chip}), host time (t_{host}), and data transfer time (t_{data}). Note that the on-chip simulation time and data transfer time partially overlap (leading to an over 100% ratio when summing the three parts together). The time consumed by the host can be further divided into three components: quantization warm-up time (t_{quant_warmup}), post-processing time (t_{post}), and others (t_{others}). The "others" category includes operations such as data loading, pre-processing, and the PSO algorithm, which are grouped together due to the short individual time.

When the model scale (#Node) is small, the on-chip simulation accounts for the majority of the inversion time, and the proportion of time spent on quantization is not significant. However, as the model scale becomes large, the quantization overhead increases markedly, occupying a major portion of the time. It should be noted that t_{quant_warmup} here represents the total time for the warm-up stage, the quantization parameter selection (QPS) stage, and the calculation of quantization parameters. Since the parameter calculation time is almost negligible, the primary overhead of quantization stems from the floating-point simulations conducted during the warm-up and QPS stages.

Table R2. Runtime breakdown of the inversion pipeline on the brain-inspired computing chip across three representative model scales corresponding to Table R1.

#Node	68	148	512
t_{total}	42.33 (100%)	77.11 (100%)	796.43 (100%)
t_{host}	8.581 (20.3%)	20.99 (27.2%)	578.6 (72.7%)
t_{quant_warmup}	**4.866 (11.5%)**	**11.37 (14.7%)**	**534.2 (67.1%)**
t_{post}	3.441 (8.1%)	9.296 (12.1%)	43.93 (5.5%)
t_{others}	0.2741 (0.6%)	0.3220 (0.4%)	0.5293 (0.1%)
t_{chip}	26.32 (62.2%)	47.14 (61.1%)	217.8 (27.3%)
t_{data}	19.19 (45.3%)	31.98 (41.5%)	102.3 (12.8%)

Revision: Updated Table 1 (corresponding to Table. R1) and Fig. 5 (corresponding to Fig. R1) with the time consumption of the improved pipelines on GPU and the brain-inspired computing chip.

Modified the relevant descriptions accordingly as follows.

In the revised part *Mapping hierarchical parallelism of the model inversion onto parallel computing architectures* of the *Results* section: "...This is because the semi-dynamic quantization on the brain-inspired computing chip relies on the host (CPU) to complete the warm-up stage and the QPS stage, which can only leverage the limited parallel resources of multi-core CPUs for acceleration."

In the revised part *Scalability of model inversion on different architectures* of the *Results* section: "...For brain-inspired computing chip implementation, the proportion of device-side time remains relatively stable at smaller scales (from 68 to 148 nodes), but decreases significantly for the 512-node model due to the substantial growth in host-side time."

Added a more detailed breakdown of the time consumed by the brain-inspired computing chip-based pipeline in Supplementary Table S3 (corresponding to Table. R2).

Comment 2: ***The authors use a DMF model as a case study for their simulations. While it is reasonable to choose one model for reasons of computational effort, it would have been useful to understand how representative the DMF model is as a case study compared to other models they mention such as Wilson-Cowan Model, Kuramoto Model etc. More generally, is it often the case that the models have a transient regime and a stable fluctuation-driven regime to allow semi-dynamic quantisation? A more detailed discussion references/examples for this would be useful.***

Response: The suggestions are very helpful for improving the clarity and depth of our manuscript. The detailed responses are as follows:

1. Representativeness of the DMF Model:

Our choice of the DMF model as our primary case study is based on several reasons. First, to the best of our knowledge, the DMF model is the most widely used model in the field of whole-brain dynamics in recent years. Several influential works, such as [1–3], have utilized the DMF model or its variants to construct whole-brain networks and fit empirical data. Second, the DMF model is derived from microscopic spiking neuron models [4], which allows it to bridge dynamical models at different scales within the brain to some extent. In contrast, other models, such as Hopf and Kuramoto models, are relatively more coarse-grained and phenomenological in their fundamental assumptions.

Your question motivated us to carefully re-examine the various existing models of brain dynamics.

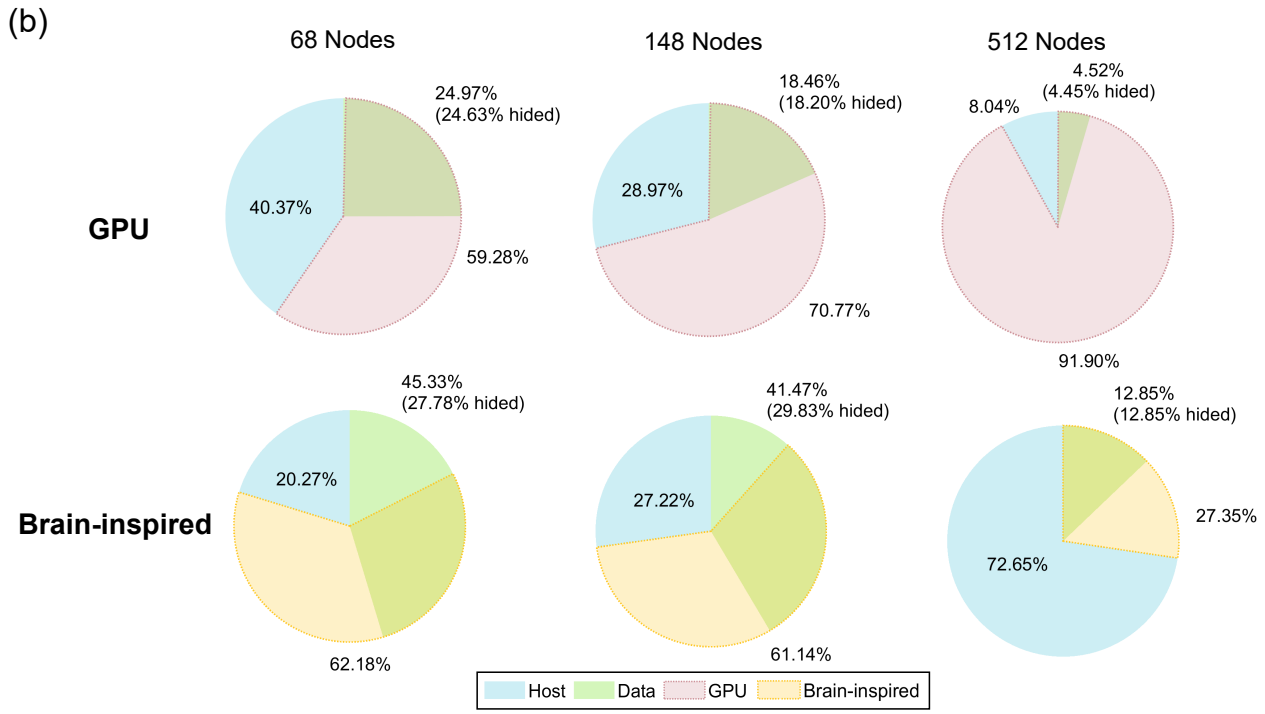

Fig. R1. Detailed runtime breakdown at three representative model scales.

Indeed, different whole-brain models employ distinct differential equations to describe brain regions, and the signals may exhibit different characteristics. For instance, in oscillator-based models such as Hopf and Kuramoto models, the variables representing individual brain regions demonstrate pronounced periodic oscillatory dynamics over time, rather than randomly fluctuation observed in the DMF model. Therefore, we have additionally tested our proposed pipeline over the Hopf model [5, 6].

We validated the low-precision simulation of the Hopf model. Since the Hopf model we employed has only two parameters (the bifurcation parameter a and the global scaling g of the connectivity), we directly computed the correlation between simulated and empirical FC results across the two-dimensional parameter space, and compared the FC correlation distributions of high-precision and low-precision models. Figure R2 demonstrates that the quantization approach can maintain strong correspondence with high-precision simulations for oscillatory dynamics. Specifically, the correlation between high-precision and low-precision FC results across the parameter space can reach $r = 0.9968$ ($p < 10^{-4}$), indicating that the low-precision model can effectively capture the parameter-dependent FC patterns observed in the high-precision model.

It is important to note that despite the differences in the underlying equations, the simulation procedures for various models when applied to resting-state brain modeling are remarkably similar. Models typically begin simulation from randomly initialized states, followed by an initial transient regime. Analysis is then conducted on the time series with the initial transient period removed after the signal behavior reaches a relatively stable state. For example, in the Hopf model, during resting-state conditions, the oscillations are relatively stable with amplitude fluctuations that remain within a bounded range. From this perspective, the DMF model is not a special case but rather representative of the common simulation workflow employed across different whole-brain dynamics models.

2. The transient regime and the stable regime for semi-dynamic quantization:

You raised an important question about the prevalence of transient and stable regimes and their signifi-

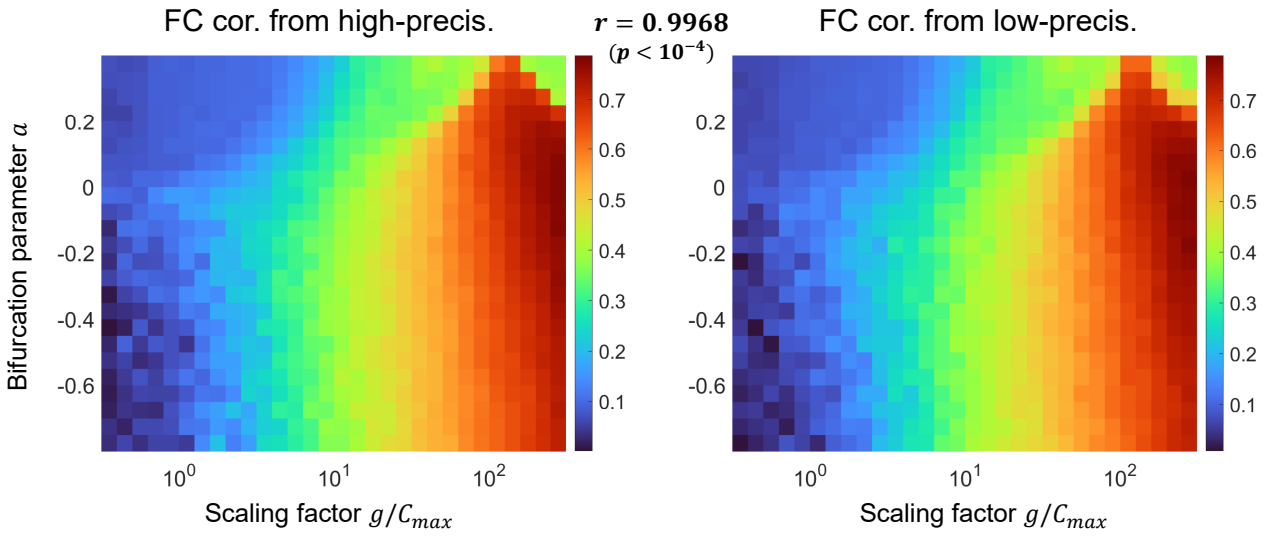

Fig. R2. Hopf model FC correlation across the parameter space. Correlation coefficients between empirical and simulated FC results from the high-precision model (left) and the low-precision model (right). Parameters a and g represent the bifurcation parameter and global connectivity scaling, respectively, while C_{max} denotes the maximum value of the structural connectivity matrix.

cance for semi-dynamic quantization. We agree that it is a crucial point that requires detailed clarification.

Regarding the representativeness of transient and stable regimes in whole-brain modeling: Most current macroscopic whole-brain models are developed based on the resting-state data. Even without employing low-precision quantization methods, the standard simulation procedure in prior studies typically begins from an arbitrarily specified initial state. After a warm-up period (e.g., a fixed duration of several seconds), when the model's signal activities reach a largely stable state, the signal characteristics are analyzed and compared with empirical resting-state data. From a computational perspective, this warm-up or transient regime effectively represents the process of numerically solving the nonlinear dynamical equations, while the stable regime is used to simulate the brain network based on the converged solution. This procedure represents a common practice in the field and constitutes the primary scenario where our proposed method is applied. Importantly, this procedure would remain consistent regardless of which specific model is employed. Our quantization method builds upon such established simulation procedure by adding a quantization parameter selection phase during the transient regime and implementing the stable state simulation using low-precision data types and computation.

Regarding the scope and limitations of semi-dynamic quantization: We wish to define the applicability and constraints of semi-dynamic quantization more explicitly. For low-precision quantization, the crucial determining factor is the numerical range of variables rather than the specific model behaviors. Low-precision data types have inherently limited the representational range and precision. Therefore, as long as the ranges of the variables and their derivatives of the model do not undergo substantial changes over time, low-precision quantization can remain effective. The DMF model converges to a fixed point with variables fluctuating within a bounded range, thus satisfying this requirement. However, this does not imply that only models converging to fluctuations around a fixed point can utilize the proposed semi-dynamic quantization method. For instance, in oscillator-based models such as the Kuramoto or Hopf model, state variables may oscillate over a wider range. Nevertheless, as long as their ranges remain relatively stable, semi-dynamic quantization can still be applicable. Previous literature empirically indicates that the resting-

state FC does not exhibit extremely drastic transient changes (although some time-varying information might be extractable) [7]. Therefore, we believe that semi-dynamic quantization should be capable of meeting the computational demands of resting-state modeling of brain dynamics across various model types.

Looking toward future extension, we can envision scenarios where the ranges of variables change significantly during model execution. If such changes are predictable, semi-dynamic quantization could still be employed. For instance, when simulating brain lesions [8], perturbations are typically introduced to a baseline resting-state brain model. These perturbations can cause the model to transfer from one stable state to another. Since the timing of these perturbations is known as a priori, high-precision floating-point numbers could be used for the transition phase, while a new low-precision quantization scheme could be employed once another stable state is reached. By iteratively applying the warm-up and simulation stages, the semi-dynamic quantization method could be potentially extended to more complex scenarios. While these scenarios are theoretically plausible, they go beyond the scope of the current paper; therefore, we just discuss them in the revised Discussion section.

We also acknowledge that the semi-dynamic quantization method has inherent limitations. When a dynamical model exhibits frequent or unpredictable large-magnitude numerical fluctuations, this method would likely become unsuitable. To our knowledge, existing macroscopic brain dynamics models do not exhibit such characteristics under typical simulation conditions, but we have explicitly mentioned this limitation in the revised Discussion section.

Revision:

Added descriptions of the DMF model's representativeness and validation of the method's generalization with the Hopf model in the revised part *Dynamics-aware quantization for macroscopic brain models* of the *Results* section: "...We specifically focus on the simulation of the DMF model and the hemodynamic model for the resting-state brain. The choice of the DMF model as the primary case study is based on its widespread use in whole-brain dynamics research [1–3] and its biophysically grounded foundation derived from microscopic spiking neuron models [4]. To further validate the generalization of our dynamics-aware quantization framework, we also conduct experiments over the Hopf model [5, 6], which represents a phenomenological approach with different oscillatory dynamics compared to fixed-point attractor behavior of the DMF model."

Added the low-precision simulation results of the Hopf model in Supplementary Fig. S10 (as shown in Fig. R2) and the acceleration results in Supplementary Table S6 (corresponding to Table R7). Added a new part *Generalization of the proposed pipeline to other brain dynamics models* in the *Methods* section, which includes the following contents:

"Different whole-brain models employ distinct differential equations to describe brain regions, and the signals may exhibit different characteristics [9]. The DMF model we primarily focus on evolves from random initial states to fixed points during simulation and exhibits random fluctuations around these fixed points driven by noise. In contrast, oscillator-based models such as Hopf and Kuramoto models demonstrate pronounced periodic oscillation over time, with variables showing regular temporal patterns rather than the stochastic fluctuations observed in the DMF model. Therefore, to demonstrate the generalization of our proposed pipeline, we conduct additional validation experiments over the Hopf model [5, 6].

The whole-brain dynamics of the Hopf model can be described by

$$\frac{dz_j}{dt} = (a + i\omega_j)z_j - |z_j|^2 z_j + g\sum_{k=1}^{N} C_{jk}(z_k - z_j) + \eta_j, \tag{R1}$$

where $z_j = x_j + iy_j$ is the complex state variable of the brain region j, $|z_j|$ is the module of z_j, the parameter a is the bifurcation parameter which we make constant across nodes, ω_j is the intrinsic angular frequency, g is a global scaling of the connectivity C, and η_j is uncorrelated white noise. From the modeling foundation perspective, the Hopf model represents a phenomenological approach to whole-brain dynamics, in contrast to the DMF model which is a biophysically grounded model derived from microscopic spiking neuron networks. From the dynamical behavior perspective, the Hopf model under certain parameter regimes exhibits oscillatory dynamics mentioned in [10], fundamentally different from the DMF model which converges to fluctuations around a fixed-point attractor. The difference allows us to validate our approach across distinct dynamical regimes.

We first validate the low-precision simulation of the Hopf model. Since the Hopf model we employ has only two parameters (the bifurcation parameter a and the global scaling parameter g), we directly computed the correlation between simulated and empirical FC results across the two-dimensional parameter space, and compared the correlation distributions of high-precision and low-precision models. Supplementary Fig. S10 demonstrates that the quantization approach can maintain strong correspondence with high-precision simulations for oscillatory dynamics. Specifically, the correlation between high-precision and low-precision FC results across the parameter space can reach $r = 0.9968$ ($p < 10^{-4}$), indicating that the low-precision model can effectively capture the parameter-dependent FC patterns observed in the high-precision model.

We also validate the acceleration potential of the Hopf model simulation on the advanced computing architectures. We conduct experiments over the Hopf model with 68, 148, and 512 nodes, and compare the performance of CPU and brain-inspired computing chip implementations. The results are summarized in Supplementary Table S6. Unlike the DMF model, the Hopf model requires smaller time steps for simulation, resulting in significantly longer simulation times on CPU, with simulation time accounting for a higher proportion of the total inversion time. Additionally, the warm-up phase in the brain-inspired computing chip pipeline also consumes more time. Despite these challenges, the brain-inspired computing chip pipeline still demonstrates significant acceleration compared to the CPU implementation, achieving speedup factors of up to 90.3× for simulation time and 30.4× for overall inversion time in the 512-node case."

Provided detailed explanations regarding the prevalence of transient and stable regimes in the part *Dynamics-aware quantization for macroscopic brain models* of the *Results* section: "We introduce semi-dynamic quantization to address this problem. Most current macroscopic whole-brain models are developed based on the resting-state data. The simulation of the brain's resting state activity can be divided into two stages. In the warm-up stage, the variables may start from an arbitrary state and experience a transient period with significant changes before converging to a relatively stable state. In the simulation stage, the model typically runs in a relatively stable state where the variables fluctuate randomly or oscillate within a certain range. These stages represent a common practice in the model simulation scenario, regardless of which specific model is employed."

Clarified the scope of applicability of the proposed method in the part *Dynamics-aware quantization for macroscopic brain models* of the *Results* section: "The applicability of semi-dynamic quantization depends on the numerical range of variables. As long as the ranges of variables and their derivatives of the model do not undergo substantial changes over time, low-precision quantization can remain effective. Previous literature empirically indicates that the resting-state FC does not exhibit extremely drastic transient changes (although some time-varying information might be extractable) [7]. Therefore, we believe that semi-dynamic quantization should be capable of meeting the computational demands of resting-state modeling of brain dynamics across various model types. It is also worth..."

Added discussions regarding the broader applicability of semi-dynamic quantization in potential sce-

narios in the revised part *Semi-dynamic quantization for brain dynamics* of the *Methods* section: "...In the subsequent simulation stage, all variables are represented as integers. We can envision scenarios where the ranges of variables change significantly during model execution. If such changes are predictable, semi-dynamic quantization could still be employed. For instance, when simulating brain lesions [8], perturbations are typically introduced to a baseline resting-state brain model. These perturbations can cause the model to transfer from one stable state to another. Since the timing of these perturbations is known as a priori, high-precision floating-point numbers could be used for the transition phase, while a new low-precision quantization scheme could be employed once another stable state is reached. By iteratively applying the warm-up and simulation stages, the semi-dynamic quantization method could be potentially extended to more complex scenarios."

Added discussions regarding the limitations of the quantization approach in the revised *Discussion* section: "...allowing broader scientific computing tasks to benefit from advanced computing architectures. However, the proposed method also suffers inherent limitations: When a dynamical model exhibits frequent or unpredictable large-magnitude fluctuations in its numerical range, the semi-dynamic quantization would likely become unsuitable. There exists a fundamental trade-off between quantization precision and generality: higher quantization precision enables support for a broader variety of dynamical models with better simulation fidelity, but may result in reduced computational efficiency. Future work should explore adaptive quantization schemes that can dynamically balance this trade-off based on the specific characteristics of the target dynamic system and the available computational resources. Additionally, the computational operators supported by existing brain-inspired computing chips are limited and may not be capable of implementing arbitrary dynamic systems. These constraints highlight the need for future research in developing more flexible quantization strategies and expanding the computational capabilities of brain-inspired computing architectures to accommodate a broader range of scientific computing applications."

References:

[1] G. Deco, J. Cruzat, and M. L. Kringelbach, "Brain songs framework used for discovering the relevant timescale of the human brain," *Nature communications*, vol. 10, no. 1, p. 583, 2019.

[2] M. Demirtaş, J. B. Burt, M. Helmer, J. L. Ji, B. D. Adkinson, M. F. Glasser, D. C. Van Essen, S. N. Sotiropoulos, A. Anticevic, and J. D. Murray, "Hierarchical heterogeneity across human cortex shapes large-scale neural dynamics," *Neuron*, vol. 101, no. 6, pp. 1181–1194, 2019.

[3] P. Wang, R. Kong, X. Kong, R. Liégeois, C. Orban, G. Deco, M. P. Van Den Heuvel, and B. Thomas Yeo, "Inversion of a large-scale circuit model reveals a cortical hierarchy in the dynamic resting human brain," *Science advances*, vol. 5, no. 1, p. eaat7854, 2019.

[4] K.-F. Wong and X.-J. Wang, "A recurrent network mechanism of time integration in perceptual decisions," *Journal of Neuroscience*, vol. 26, no. 4, pp. 1314–1328, 2006.

[5] G. Deco, M. L. Kringelbach, V. K. Jirsa, and P. Ritter, "The dynamics of resting fluctuations in the brain: metastability and its dynamical cortical core," *Scientific reports*, vol. 7, no. 1, p. 3095, 2017.

[6] A. Ponce-Alvarez and G. Deco, "The hopf whole-brain model and its linear approximation," *Scientific reports*, vol. 14, no. 1, p. 2615, 2024.

[7] T. O. Laumann, A. Z. Snyder, A. Mitra, E. M. Gordon, C. Gratton, B. Adeyemo, A. W. Gilmore, S. M. Nelson, J. J. Berg, D. J. Greene *et al.*, "On the stability of bold fmri correlations," *Cerebral cortex*, vol. 27, no. 10, pp. 4719–4732, 2017.

[8] J. Wei, B. Wang, Y. Yang, Y. Niu, L. Yang, Y. Guo, and J. Xiang, "Effects of virtual lesions on temporal dynamics in cortical networks based on personalized dynamic models," *NeuroImage*, vol. 254, p. 119087, 2022.

[9] J. Cabral, M. L. Kringelbach, and G. Deco, "Functional connectivity dynamically evolves on multiple time-scales over a static structural connectome: Models and mechanisms," *NeuroImage*, vol. 160, pp. 84–96, 2017.

[10] E. E. Tsur, *Neuromorphic Engineering: The Scientist's, Algorithms Designer's and Computer Architect's Perspectives on Brain-Inspired Computing*. CRC Press, 2021.

Comment 3: ***The authors perform comparisons between the quantised and full-precision simulations for both static and dynamic indicators, and provide convincing demonstrations of their qualitative***

similarities. But the quantitative comparison is a bit more limited – the only aggregate quantitative comparison that the authors provide is R^2 values using linear fitting. The implications of this were not clear to me, especially to answer the question of whether quantised simulation provides utility in the contexts it is usually used, i.e. in computational studies that are carried out using a DMF model. And whether the linear fitting results provide a reasonable measure of such utility. Discussions and examples addressing this aspect would be very useful. If possible (considering computational constraints), reproducing an existing study with the quantised model and seeing if the resulting predictions match the original studies would significantly strengthen the paper.

Response: We appreciate your comments regarding the quantitative comparisons between the quantized and full-precision simulations. The main scenario where we would like to employ the quantized model is to search the space for a set of parameters that yields the best fit to empirical data. Our hope is that the searched parameter set, when subsequently applied to the full-precision model, will also achieve a strong goodness-of-fit. This necessitates that the goodness-of-fit indicators (e.g., FC Corr or metastability) obtained from models of different precisions are positively correlated across the parameter space. In simpler terms, for a given parameter set, if the indicator from the quantized model is high (or low), the indicator from the full-precision model should correspondingly be high (or low).

However, as you correctly pointed out, our current method for quantitative assessment has limitations. Inspired by your insightful comment, we have enriched our quantitative validation approach for the low-precision models to address these limitations.

First, the R^2 from linear fitting, while indicative of linear dependency [1], may not serve as the most direct quantitative metric for correlation. In fact, a more appropriate quantitative metric would be the correlation coefficient. So we calculated the Pearson correlation for the indicator values (FC correlation) obtained from the models of different precisions across the sample points in the parameter space. To test the robustness of our results, we also conducted additional experiments under different noise intensities and across different model scales. Specifically, we varied the additive noise amplitude σ in the stochastic differential equations from 0.0006 to 0.003 and ran simulations with both 68-node and 148-node connectomes. The noise amplitude range encompasses the majority of scenarios encountered in the parameter search and optimization procedures. As shown in Fig. R3, the results indicate that low-precision simulations preserve significantly positive correlations with high-precision counterparts under all tested scenarios, indicating robust performance under varying noise levels and model complexities.

Meanwhile, we also observed that the positive correlation coefficient shows some degradation when the noise amplitude is high (e.g. $\sigma = 0.003$). On one hand, larger noise may render the dynamic system more unstable, and if significant range variations occur, the effectiveness of low-precision simulation would deteriorate. On the other hand, the complex dynamics may amplify the noise effects, especially in the DMF model where the functional relationship between the total input current and the population firing rate is highly sensitive to numerical variations. This accuracy degradation could potentially be mitigated by appropriately increasing the quantization precision.

Second, as previously mentioned, we recognize that the high positive correlations between the indicators obtained from the two types of models are only necessary conditions, not sufficient conditions. As you pointed out, conducting experiments directly using quantized models would be more convincing for validating the application of low-precision models in practical scenarios. We used the PSO algorithm to search for optimal parameters in the parameter space based on the low-precision model. Unlike the experiments above, here we also included the noise amplitude as an optimizable parameter. We conducted 30 repeated simulations using both high-precision and low-precision models with the searched parameters. The correlation between low-precision simulated and empirical FC results was, on average,

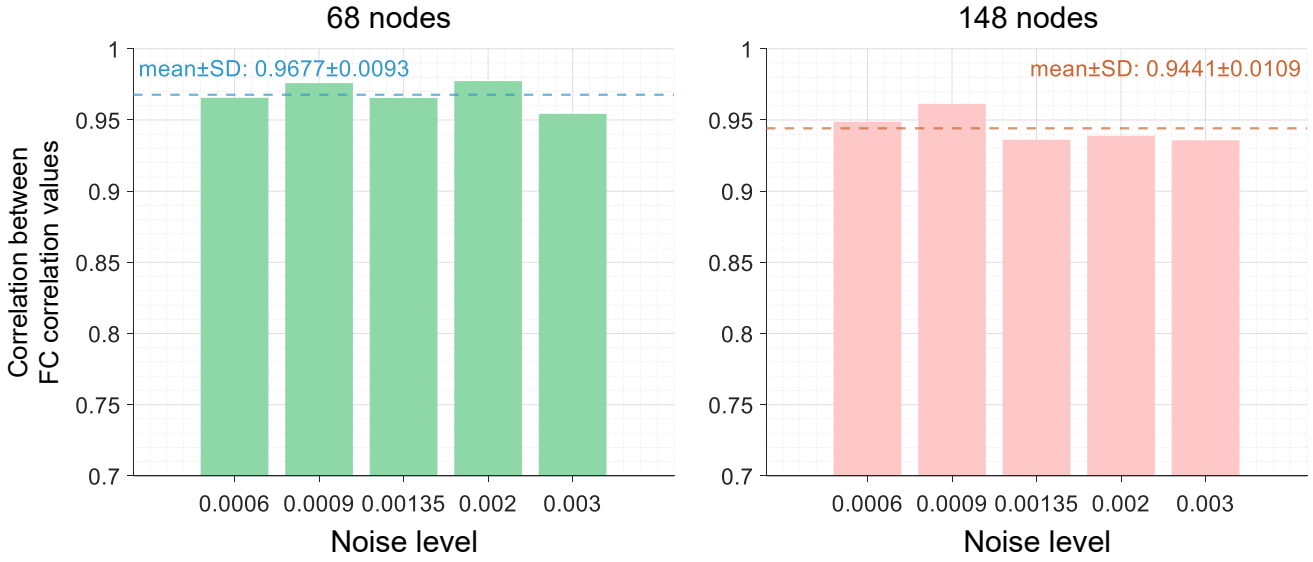

Fig. R3. Robustness analysis of the quantization method across different noise conditions and model scales. Correlation coefficients between high-precision and low-precision simulation results across different noise amplitudes for 68-node (**left**) and 148-node (**right**) connectomes. Each bar represents the Pearson correlation coefficient calculated from FC correlation values obtained from 1200 parameter sets sampled across the parameter space. All correlation results are statistically significant with $p < 10^{-4}$.

0.7018 ($p < 10^{-4}$), with an SD of 0.0103. The correlation between high-precision simulated and empirical FC results also achieved 0.6908 ($p < 10^{-4}$), with an SD of 0.0130. Meanwhile, the correlation between the simulated FC results from the two precisions reached 0.9675 on average ($p < 10^{-4}$). As a baseline, the correlation between the empirical SC and FC was 0.4545. These results demonstrate that using the low-precision model for model inversion is feasible and effective.

Additionally, we evaluated the ability of low-precision models to reproduce the spatial topology of empirical data and presented the results in Supplementary Fig. S5. Please see the part *Goodness-of-fit indicators and other quality metrics* of the *Methods* section and Supplementary Fig. S5 for more details.

Revision: Following your suggestion, we have changed the quantitative results in Fig. 3 from linear fitting to correlation (as shown in Fig. R3).

We have also revised the description of these results in the part *Consistency of goodness-of-fit indicators between low- and high-precision models* of *Results* Section: "To further quantitatively evaluate the consistency of goodness-of-fit indicators between low-precision and high-precision models, we calculate the Pearson correlation for the indicator values (FC correlation) obtained from the models across random sample points in the parameter space. To test the robustness of our results, we conduct additional experiments under different noise intensities and across different model scales. Specifically, we vary the additive noise amplitude σ in the stochastic differential equations from 0.0006 to 0.003 and run simulations with both 68-node and 148-node connectomes. The noise amplitude range encompasses the majority of scenarios encountered in the parameter search and optimization procedures. As shown in Fig. 3c, the low-precision simulations preserve significantly positive correlations with high-precision counterparts under all tested scenarios, indicating robust performance under varying noise levels and model complexities. Meanwhile, we also observe that the positive correlation coefficient shows some degradation when the noise amplitude is high (e.g. $\sigma = 0.003$). On one hand, larger noise may render the dynamic system more

unstable, and if significant range variations occur, the effectiveness of low-precision simulation would deteriorate. On the other hand, the complex dynamics may amplify the noise effects, especially in the DMF model where the functional relationship between the total input current and the population firing rate is highly sensitive to numerical variations. This accuracy degradation could potentially be mitigated by appropriately increasing the quantization precision."

We have moved the original linear fitting results and corresponding analysis to Section S1.9 in Supplementary Information.

Additionally, we have added the results of model inversion with the low-precision model in the part *Consistency of goodness-of-fit indicators between low- and high-precision models* of *Results* Section: "To directly validate the applicability of low-precision models in practical scenarios, we use the PSO algorithm to search for optimal parameters in the parameter space based on the low-precision model. Unlike the experiments above, here we also include the noise amplitude as an optimizable parameter. We conduct 30 repeated simulations using both high-precision and low-precision models with the searched parameters. The correlation between low-precision simulated and empirical FC results is, on average, 0.7018 ($p < 10^{-4}$), with an SD of 0.0103. The correlation between high-precision simulated and empirical FC results also achieves 0.6908 ($p < 10^{-4}$), with an SD of 0.0130. Meanwhile, the correlation between the simulated FC results from the two precisions reaches 0.9675 on average ($p < 10^{-4}$). As a baseline, the correlation between the empirical SC and FC is 0.4545. These results demonstrate that using the low-precision model for model inversion is effective."

References:

[1] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

Comment 4: ***The hierarchical parallelism scheme described is well thought out, and the the demonstrated speed-ups on both brain-inspired hardware and GPUs are impressive. One question I had is: why not do the warmup phase of simulation on a GPU as well? It would seem that at GPU and brain-inspired chip could be combined to give the best of both worlds. While it's clear that their method would scale well, it would be necessary to see the host time included in the statistics to truly get an idea of the scalability of their method. I would expect that with larger model sizes, host times become a bottleneck (especially for the simulation on the brain-inspired chip).***

Response: For your concerns on the hierarchical parallelism scheme and system scalability, our answers are as follows:

1. Why not perform the warm-up phase on GPU as well?

This is an excellent suggestion about combining GPU and the brain-inspired computing chip to leverage the best of both worlds. Previously, we primarily focused on demonstrating and comparing the performance of different architectures for whole-brain modeling to intuitively evaluate their respective advantages and disadvantages in this scenario. However, from a performance optimization perspective, utilizing multiple advanced computing architectures to accelerate different steps in the inversion process is indeed a worthwhile approach.

Following your suggestion, we have explored using GPU to perform the warm-up phase while maintaining the brain-inspired computing chip for the main simulation. From Table R3, it can be observed that using GPU for quantization warm-up reduces the time consumed by the quantization steps (including both the warm-up phase and the QPS phase). When the model scale is small, this approach does not provide particularly significant improvements on the total inversion performance, because in these cases the time

Table R3. Performance comparison between the original brain-inspired computing chip pipeline (B+C) and the hybrid GPU-accelerated warm-up pipeline (B+G+C). The table shows total inversion time and quantization time (t_{quant}). B: Brain-inspired computing chip, G: GPU, C: CPU.

Platform		68 nodes	148 nodes	512 nodes
B+C	t_{total}	42.33	77.11	796.4
	t_{quant}	4.866	11.37	534.2
B+G+C	t_{total}	40.32	74.79	375.9
	t_{quant}	2.858	9.054	113.6

spent on warm-up accounts for a relatively small proportion of the total time consumption. However, when the model scale is large (e.g., 512 nodes), the overhead of quantization warm-up using CPU becomes substantial, and GPU-based acceleration can significantly improve overall performance. Specifically, for the 512-node model, the hybrid pipeline (B+G+C) reduces the total inversion time from 796.4 seconds to 375.9 seconds, representing a 2.1× speedup. This substantial improvement demonstrates that the hybrid approach effectively addresses the scalability limitations of CPU warm-up.

The results validate your suggestion that combining GPU and brain-inspired computing architectures can indeed leverage the best of both worlds. It should be noted that this approach may increase the complexity of the system, but we believe that this integration strategy represents the future trend of heterogeneous computing platforms.

2. Host time as a potential bottleneck with larger model sizes:

Your prediction on the scalability is correct which can be confirmed by our experimental results. We agree that including the host time statistics can be crucial for understanding the true scalability of our method. In fact, most of the experimental results reported in our manuscript had already included the host time consumption. The $T_{overall\_inversion}$ values shown in Table 1 incorporate host time costs, and the time breakdowns presented in Fig. 5a and b explicitly show both the host time consumption and its proportion relative to total execution time.

Additionally, we further present more detailed scalability analyses with comprehensive time breakdown data across different model scales to make the scalability of the host time clearer. Related experimental settings can be found in the earlier response to Comment 1.

Table R4. Scalability analysis of the runtime breakdown on the brain-inspired computing chip. The model scales correspond to Fig. 5a.

# Node	64	128	192	256	320	384	448	512
t_{total}	30.7 (100%)	62.8 (100%)	101 (100%)	141 (100%)	185 (100%)	487 (100%)	624 (100%)	796 (100%)
t_{host}	8.11 (26.4%)	17.9 (28.5%)	33.9 (33.5%)	50.9 (36.2%)	72.1 (39.0%)	342 (70.2%)	445 (71.2%)	579 (72.7%)
t_{quant}	**4.61 (15.0%)**	**9.49 (15.1%)**	**20.5 (20.2%)**	**32.6 (23.2%)**	**48.4 (26.2%)**	**311 (64.0%)**	**408 (65.3%)**	**534 (67.1%)**
t_{post}	3.23 (10.5%)	8.11 (12.9%)	13.1 (12.9%)	17.9 (12.7%)	23.3 (12.6%)	29.9 (6.1%)	36.4 (5.8%)	43.9 (5.5%)
t_{others}	0.266 (0.9%)	0.305 (0.5%)	0.340 (0.3%)	0.378 (0.3%)	0.416 (0.2%)	0.453 (0.1%)	0.489 (0.1%)	0.529 (0.1%)
t_{chip}	17.1 (55.6%)	36.3 (57.8%)	58.7 (58.0%)	84.3 (59.9%)	113 (61.0%)	145 (29.8%)	180 (28.8%)	218 (27.3%)
t_{data}	12.8 (41.7%)	25.6 (40.7%)	38.4 (37.9%)	51.2 (36.4%)	64.0 (34.6%)	76.8 (15.8%)	89.5 (14.3%)	102 (12.8%)

As shown in Table R4, the proportion of the host time (t_{host}) indeed increases gradually with the model scale: from 26.4% for 64 nodes to 39.0% for 320 nodes, and then jumps significantly to 70.2% for 384 nodes, eventually reaching 72.7% for 512 nodes. This trend aligns perfectly with your prediction that the host time would become the dominant bottleneck as the model size increases, particularly for simulation

on the brain-inspired computing chip. Within the host time breakdown, the quantization time (t_{quant}) represents the largest component, accounting for the majority of the host overhead. This includes both the warm-up simulation time and the QPS stage time. The sudden jump in the host time when scaling from 320 to 384 nodes might be caused by factors such as cache size limitations, memory paging, or BLAS library switching, which we had mentioned in the part *Scalability of model inversion on different architectures* of *Results* section in the manuscript.

This scalability analyses demonstrate that the host-side warm-up overhead does indeed become a limiting factor for very large-scale models. However, it should be noted that even with this bottleneck, the computational cost of model inversion based on the brain-inspired computing chip is still the lowest, and it is one to two orders of magnitude better than the CPU-based implementation. The bottleneck presented by host time does not invalidate the advantages of the proposed methods, but rather highlights a significant opportunity for future optimization. Performance might be further improved through more efficient quantization algorithms or by offloading portions of the quantization process to specific hardware accelerators.

Revision: Incorporated the acceleration results of hybrid pipeline and extra discussions into the revised part *Mapping hierarchical parallelism of the model inversion onto parallel computing architectures* of *Results* section: "...including faster low-precision computing units and efficient distributed near-memory computing circuits with reduced redundancy. Additionally, we also explore using GPU to perform the warm-up phase while maintaining the main simulation on the brain-inspired computing chip. When the model scale is small, this approach does not provide particularly significant improvements as the time spent on warm-up accounts for a relatively small proportion of the total time consumption. However, when the model scale is large (e.g., 512 nodes), GPU-based warm-up acceleration can significantly improve overall performance. Specifically, for the 512-node model, the hybrid pipeline achieves a 2.1× speedup over the brain-inpired computing chip pipeline (see Supplementary Table S4). This hybrid pipeline can leverage the best of both worlds, and it is worth exploring in future work."

Added results of the hybrid GPU-accelerated warm-up pipeline in Supplementary Table S4 (corresponding to Table R3).

Pointed out the bottleneck effect of host time and discussed the potential optimization in the part *Scalability of model inversion on different architectures* of *Results* section: "We can predict that the host time would become the dominant bottleneck in the brain-inspired computing chip pipeline as the model size increases. This highlights a significant opportunity for future optimization. Performance might be further improved through more efficient quantization algorithms, or by offloading portions of the warm-up process to GPU (like what we do in Supplementary Table S4) or other specific hardware accelerators."

Added detailed scalability analysis of the runtime breakdown of the brain-inspired computing chip pipeline in Supplementary Table S5 (corresponding to Table R4).

Comment 5: ***In summary, this is a very sound and innovative study that provides a way to engineer simulations of macroscopic brain dynamics on brain-inspired chips or GPU, providing impressive performance gains while retaining sufficient fidelity on the demonstrated use case. The paper is overall well written and clear, and results are described completely. I do not have sufficient expertise in this subfield to know if all relevant literature is discussed. The main weaknesses of this study, in my opinion, are that (1) it looks only at one particular use case (DMF); (2) An analysis of the utility of this method in the context of existing literature is missing; and (3) the support for the scalability of the model is not fully convincing. I believe the work has the potential to be quite significant if some of these weaknesses are addressed.***

Response: We sincerely appreciate you for both the right recognition of the contributions and the insightful feedback on the limitations of our work. Your suggestions are extremely valuable and have substantially helped us improve the manuscript. In summary, we have made the following major improvements.

1. Addressing the limitation of a single use case (DMF): We have incorporated an additional representative model, the Hopf model, to demonstrate the generalizability of our proposed approach. The Hopf model differs significantly from the DMF model in both principles and characteristics, and our quantization method still remains applicable. This demonstrates that our proposed method can extend beyond the specific dynamics of mean-field models to a broader scope of whole-brain dynamics models.

2. Enhanced analysis of utility in practical contexts: We have significantly strengthened our quantitative validation approach. Beyond the R^2 values from linear fitting, we now use Pearson correlation coefficients between high-precision and low-precision results across the parameter space as a more direct quantitative indicator of the effectiveness of low-precision models. More importantly, we have conducted model inversion experiments directly using the low-precision model for parameter optimization on empirical data, demonstrating that optimal parameters found using low-precision models can achieve comparable goodness-of-fit with high-precision models. This validates the applicability of our proposed method in practical modeling scenarios.

3. Comprehensive scalability analysis: We have provided detailed computational time breakdown across different model scales to provide clearer insights into the scalability characteristics of the acceleration method. Your suggestions also inspired us to optimize the implementation of our pipeline, resulting in improved overall performance on both GPU and the brain-inspired computing chip. Since the host-side warm-up becomes a bottleneck for the brain-inspired pipeline as the model scale increases, we have tried to use the GPU to perform the warm-up phase. This hybrid approach further improves performance, especially for large-scale models.

We hope our responses and revisions can make you satisfied.

2. Response to Reviewer 2

Overall Comment: ***The manuscript introduces a novel implementation of whole-brain mean-field simulations with neuromorphic hardware (TianjicX computing chip). This allows for faster simulations and model inversions to fit neuroimaging data by one to two orders of magnitude compared to modern CPU and GPU implementations, while still capturing dynamical aspects of more established models of whole-brain dynamics. To do so, the authors employ two strategies: (1) semi-quantization of steady-state firing rates and time leading to low-precision simulations by taking advantage of fixed points in dynamical mean-field models, and (2) massively parallel distribution of computations leveraging distributed memory storage from neuromorphic chips. The paper has potentially transformative impacts for the field of whole-brain modeling by substantially accelerating simulations and model inference, with translational implications for the development of whole-brain digital twins for diagnosis prediction and treatment optimization. However, I have several major concerns on validations and comparisons made as well as on the clarity of explanations in the manuscript.***

Response: We sincerely appreciate your correct recognition of our contributions. For the concerns you mentioned below, we have made revisions including comparing the approaches and results with those in the mentioned references, adding explanations on how to encode non-spiking data, and adding discussions on exploring more dendritic functions. Details can be found in the following responses, which we hope can make you satisfied.

Major Comment 1: ***I could not understand if Figure 3 characterizes low-precision fitting of simulated dynamics from high-precision models or of empirical rsfMRI data from the HCP dataset. It is imperative to make this explicit in the text. Further, the manuscript needs to present an in-depth characterization of both. Crucially, it should be shown that the low-resolution model can capture ground-truth parameter changes in the high-resolution model and to quantify the robustness of this by varying the noise and complexity of the high-resolution model as well as the amplitude of parameter changes. It is also important to show how well the model can fit rsfMRI data - are the correlations between empirical and low-resolution model FC comparable to those between low-resolution and high-resolution FC after model fitting?***

Response: This comment is insightful for us to present clearer characterization of the proposed low-precision approach. To clarify, Fig. 3a-c demonstrates the comparison between high-precision and low-precision models in terms of their simulation results (i.e., goodness-of-fit to empirical data) across different parameters. The main scenario where we would like to employ the quantized model is to search the space for a set of parameters that yields the best fit to empirical data. Our hope is that this optimal parameter set, when subsequently applied to the full-precision model, will also achieve a strong goodness-of-fit to empirical data. This necessitates that, for a given parameter set, if the goodness-of-fit indicator from the quantized model is high (or low), the indicator from the full-precision model should correspondingly be high (or low).

Figure 3 had actually encompassed both aspects mentioned in your comment:

1. The fitting performance of different models to empirical data. The color value (according to the colorbar) of each data point in Fig. 3a, the y-coordinate value of each point on the curves in Fig. 3b, and the x-y coordinate values of each data point in Fig. 3c all represent the goodness-of-fit between high-precision or low-precision model results and empirical data under a given parameter set. For example, in the left panel of Fig. 3b, we can observe that when w and I_0 are fixed, the maximum correlation coefficient between FC obtained from the high-precision model and empirical FC is approximately 0.7, while the maximum correlation coefficient from the low-precision model is slightly lower.

2. The correlation of low-precision simulation results to high-precision simulation results in the parameter space. Figure 3a-b show the distributions of simulation results with different precisions across the parameter space for qualitative comparison, aiming to demonstrate the similarity in the distribution patterns of simulation results from models with different precisions. Furthermore, Fig. 3c quantitatively demonstrates the correlation between high-precision and low-precision simulation results across sampling points in the parameter space, indicating that the low-precision model has strong potential to substitute for the high-precision model during the parameter search process.

Moreover, we totally agree with your suggestions. First, the meaning of Fig. 3 needs to be more explicitly presented in both the manuscript text and the figure. Second, we need to further quantify the robustness of the low-precision model's ability to capture parameter changes in the high-precision model. Third, we should demonstrate how well the model can fit rsfMRI data more directly. We have added relevant descriptions in the revised manuscript following your first suggestion. For the latter two suggestions, we have designed corresponding experiments described below.

1) Robustness of the correlation between low-precision and high-precision simulation results. We have varied the noise and complexity of the models. Specifically, beyond the original 68-node connectome, we have incorporated a more complex 148-node model to test the robustness of our approach. For both model scales, we sampled 1200 parameter sets across the parameter space. For each parameter set, we conducted both high-precision and low-precision simulations and calculated the correlation between the simulated and empirical FC, which serves as a commonly used goodness-of-fit indicator in model inversion. We then computed the correlation coefficient between the 1200 indicator values obtained from high-precision and low-precision simulations. Additionally, we tested the robustness of the correlation under different noise intensities to validate the stability of our quantization approach across varying conditions. As shown in Fig. R4, the results indicate that low-precision simulations preserve significantly positive correlations with high-precision counterparts under all tested scenarios, indicating robust performance under varying noise levels and model complexities.

Meanwhile, we also observed that the positive correlation coefficient shows some degradation when the noise amplitude is high (e.g. $\sigma = 0.003$). On one hand, larger noise may render the dynamic system more unstable, and if significant range variations occur, the effectiveness of low-precision simulation would deteriorate. On the other hand, the complex dynamics may amplify the noise effects, especially in the DMF model where the functional relationship between the total input current and the population firing rate is highly sensitive to numerical variations. This accuracy degradation could potentially be mitigated by appropriately increasing the quantization precision.

2) Fitting performance of the low-precision model to rsfMRI data. We have also conducted experiments to directly demonstrate the fitting performance of the low-precision model to empirical rsfMRI data. Specifically, we used the PSO algorithm with low-precision model to search for optimal parameters that best fit the group-level empirical data. We then computed the FC results from both low-precision and high-precision models under these optimal parameters, and calculated the correlations between simulated and empirical FC results. Unlike the experiments above, here we also included noise amplitude as an optimizable parameter. We conducted 30 repeated simulations using both high-precision and low-precision models with the searched parameters. The correlation between low-precision simulated and empirical FC results was, on average, 0.7018 ($p < 10^{-4}$), with an SD of 0.0103. The correlation between high-precision simulated and empirical FC results also achieved 0.6908 ($p < 10^{-4}$), with an SD of 0.0130. Meanwhile, the correlation between the simulated FC results from the two precisions reached 0.9675 on average ($p < 10^{-4}$). As a baseline, the correlation between the empirical SC and FC was 0.4545. These results demonstrate that using the low-precision model for model inversion is feasible and effective.

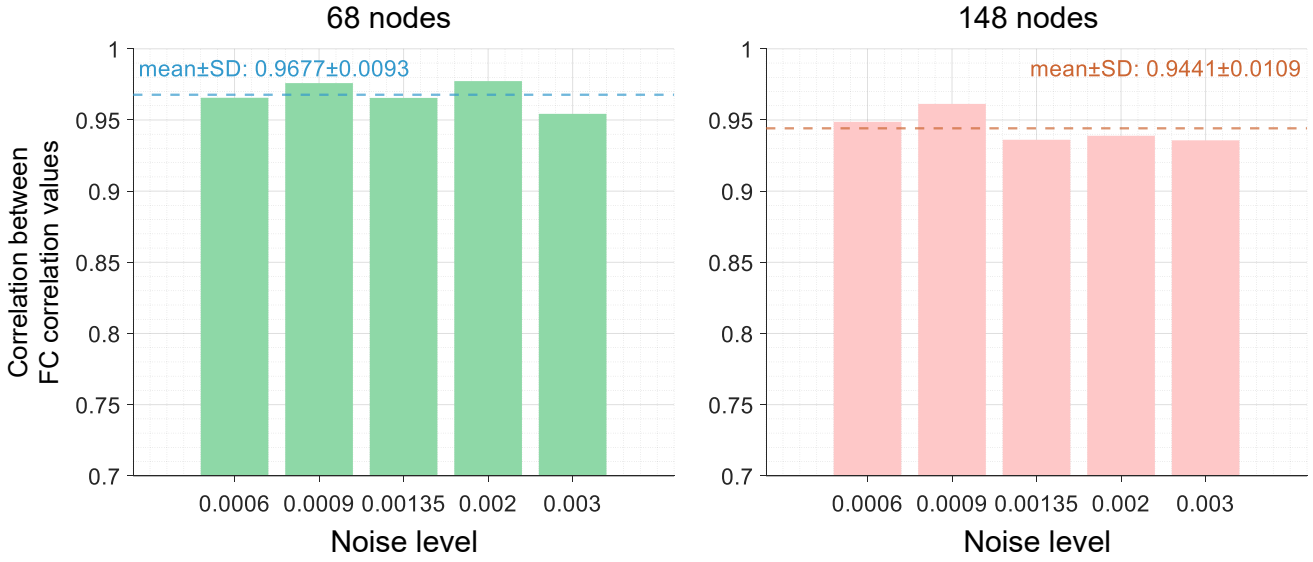

Fig. R4. Robustness analysis of the quantization method across different noise conditions and model scales. Correlation coefficients between high-precision and low-precision simulation results across different noise amplitudes for 68-node (**left**) and 148-node (**right**) connectomes. Each bar represents the Pearson correlation coefficient calculated from FC correlation values obtained from 1200 parameter sets sampled across the parameter space. All correlation results are statistically significant with $p < 10^{-4}$.

Revision: Changed the quantitative results in Fig. 3 from linear fitting to correlation (as shown in Fig. R4).

Revised the descriptions of these results in the part *Consistency of goodness-of-fit indicators between low- and high-precision models* of *Results* Section: "To further quantitatively evaluate the consistency of goodness-of-fit indicators between low-precision and high-precision models, we calculate the Pearson correlation for the indicator values (FC correlation) obtained from the models across random sample points in the parameter space. To test the robustness of our results, we conduct additional experiments under different noise intensities and across different model scales. Specifically, we vary the additive noise amplitude σ in the stochastic differential equations from 0.0006 to 0.003 and run simulations with both 68-node and 148-node connectomes. The noise amplitude range encompasses the majority of scenarios encountered in the parameter search and optimization procedures. As shown in Fig. 3c, the low-precision simulations preserve significantly positive correlations with high-precision counterparts under all tested scenarios, indicating robust performance under varying noise levels and model complexities. Meanwhile, we also observe that the positive correlation coefficient shows some degradation when the noise amplitude is high (e.g. $\sigma = 0.003$). On one hand, larger noise may render the dynamic system more unstable, and if significant range variations occur, the effectiveness of low-precision simulation would deteriorate. On the other hand, the complex dynamics may amplify the noise effects, especially in the DMF model where the functional relationship between the total input current and the population firing rate is highly sensitive to numerical variations. This accuracy degradation could potentially be mitigated by appropriately increasing the quantization precision."

Moved the original linear fitting results and corresponding analysis to Section S1.9 in Supplementary Information.

Added the results of model inversion with the low-precision model in the part *Consistency of goodness-*

of-fit indicators between low- and high-precision models of *Results* Section: "To directly validate the applicability of low-precision models in practical scenarios, we use the PSO algorithm to search for optimal parameters in the parameter space based on the low-precision model. Unlike the experiments above, here we also include the noise amplitude as an optimizable parameter. We conduct 30 repeated simulations using both high-precision and low-precision models with the searched parameters. The correlation between low-precision simulated and empirical FC results is, on average, 0.7018 ($p < 10^{-4}$), with an SD of 0.0103. The correlation between high-precision simulated and empirical FC results also achieves 0.6908 ($p < 10^{-4}$), with an SD of 0.0130. Meanwhile, the correlation between the simulated FC results from the two precisions reaches 0.9675 on average ($p < 10^{-4}$). As a baseline, the correlation between the empirical SC and FC is 0.4545. These results demonstrate that using the low-precision model for model inversion is effective."

Major Comment 2: ***I am concerned about how robust and meaningful the discretization of time is. All iterative simulation methods discretize time, and it is not clear here how this differs from other implementations. Particularly, the claim that "their temporal evolution demonstrates heterogeneity characterized by distinct timescales" raises concerns. Biological processes in the brain often span temporal scales, with many having scale-invariant properties. For instance, the power spectrum of time series measured with several neuroimaging methods (M/EEG, fMRI etc) scales in 1/f. Even oscillations, that are associated with one distinct time scale, show frequencies that fluctuate through time and across brain regions even in the same individual (see for instance Donoghue et al Nat Neuro 2020). This needs to be acknowledged, with the paper's approach to this problem much more explicitly explained, and validated. While comparisons shown between high-resolution simulations, low-resolution simulations, and rsfMRI are mainly spatial across regions, it is necessary to also show comparisons in the temporal domain such as at least the power spectral structure. Other metrics might also be used to characterize the complexity of time series such as the Lempel-Ziv complexity.***

Response: Your concerns on temporal discretization and the characterization of distinct timescales in brain dynamics are relevant. We address them through the following points.

1. Clarification of the distinction of our temporal discretization approach from other methods. As the reviewer correctly points out, nearly all iterative simulation methods discretize time. Traditional dynamical simulation implementations typically employ the same discrete time step for all differential equations in the system, such as 0.01 seconds or 0.001 seconds. Under conditions of sufficient computational precision, smaller time steps generally yield higher simulation accuracy.

However, we discovered that when using low-precision data for simulation, smaller time steps may result in very small variable changes that could be quantized to zero due to the limited precision, preventing variables from evolving properly over time and causing simulation failure. To mitigate this issue, we employ different discrete time steps for different variables (with different differential equations): we use smaller time steps for variables with larger derivative values to improve computational accuracy, while applying larger time steps for variables with smaller derivative values. More detailed methodological descriptions have been provided in the Methods Section of the manuscript.

We term this approach "multi-timescale simulation" because one of its fundamental motivations is that different variables evolve at different rates over time. This computational method may potentially be confused with the temporal scale concept in neuroscience; therefore, we have added clarifications in the revised manuscript. Additionally, we acknowledge that the time step size in integration can affect accuracy. Our experimental results demonstrate that, at least in the context of resting-state whole-brain model simulation, the discretization method is effective. Applying this discretization approach to models with higher temporal precision requirements may necessitate higher data precision. We have pointed out

the potential limitation in the Methods section of the manuscript.

2. Validation of temporal domain characteristics. We agree with your mention about the scale-invariant properties of the brain. The 1/f power spectral scaling observed in M/EEG and fMRI data, as well as the temporal and spatial fluctuations in oscillatory frequencies [1], represent fundamental characteristics of brain dynamics that need to be considered in our simulation. Although our proposed method is purely computational in nature and has no direct connection to temporal scales in the brain, it is indeed necessary to examine the impact of this approach on temporal domain characteristics. Following your suggestion, we have conducted temporal domain analyses to validate our approach.

First, we have computed and compared the power spectral density of time series from high-precision simulation, low-precision simulation, and empirical rsfMRI data. As shown in Fig. R5a, both high-precision and low-precision simulations exhibit power spectral profiles that closely match the empirical data in terms of peak locations and overall curve shapes, although specific amplitude values show some differences. The low-precision simulation results demonstrate high similarity to high-precision counterparts. Notably, the low-precision simulation shows reduced peak power intensity while exhibiting elevation in other frequency components. This phenomenon is theoretically expected, as the quantization process introduces noise-like artifacts that elevate the overall power spectrum while reducing the signal-to-noise ratio.

For temporal complexity meatures, we employed the sample entropy as our primary metric. Sample Entropy quantifies the regularity of a time series by computing the negative natural logarithm of the conditional probability that two similar sequences will maintain their similarity when extended by one data point [2]. A higher sample entropy value reflects more complex nonlinear dynamics (e.g., chaotic or stochastic behaviors). This metric is widely employed in biomedical signal analysis, particularly for characterizing dynamic properties of neuroimaging data such as fMRI and EEG [3–5]. Using the same set of optimized parameters, we conducted simulations with both high-precision and low-precision models, calculated the sample entropy for each brain region's time series, and computed correlations with the empirical regional sample entropy. As demonstrated in Fig. R5b, both precision levels show significant correlations with empirical data. Furthermore, the correlation between high-precision and low-precision sample entropy values reaches $r = 0.9482$ ($p < 10^{-4}$), indicating strong consistency.

These results demonstrate that the low-precision simulation can preserve essential temporal domain characteristics of the high-precision simulation to a considerable extent. However, due to the inherent impact of low-precision quantization on numerical precision, the preservation of temporal features is less robust compared to spatial characteristics. This represents a fundamental trade-off between computational efficiency and temporal fidelity that should be considered when applying our quantization framework.

Revision: Added clarification of "multi-timescale simulation" in the revised part *Determining the time steps for multi-timescale simulation* of *Methods* section: ", allowing each variable to update at different time intervals. The term "multi-timescale simulation" is used because the method stems from the different evolution rates among variables in the dynamic system. However, it is important to clarify that this is a computational strategy and is distinct from the neuroscientific concept of temporal scales."

Described the limitation of the proposed simulation method in the part *Determining the time steps for multi-timescale simulation* of *Methods* section: "...in our simulation. Our experimental results demonstrate that the proposed discretization method is effective, at least within the context of resting-state whole-brain model simulation. However, we acknowledge that this approach has inherent limitations: when a dynamical model cannot achieve effective low-precision simulation with uniform time steps and exhibits high sensitivity to temporal resolution, higher data precision may be the only viable solution to maintain adequate numerical accuracy."

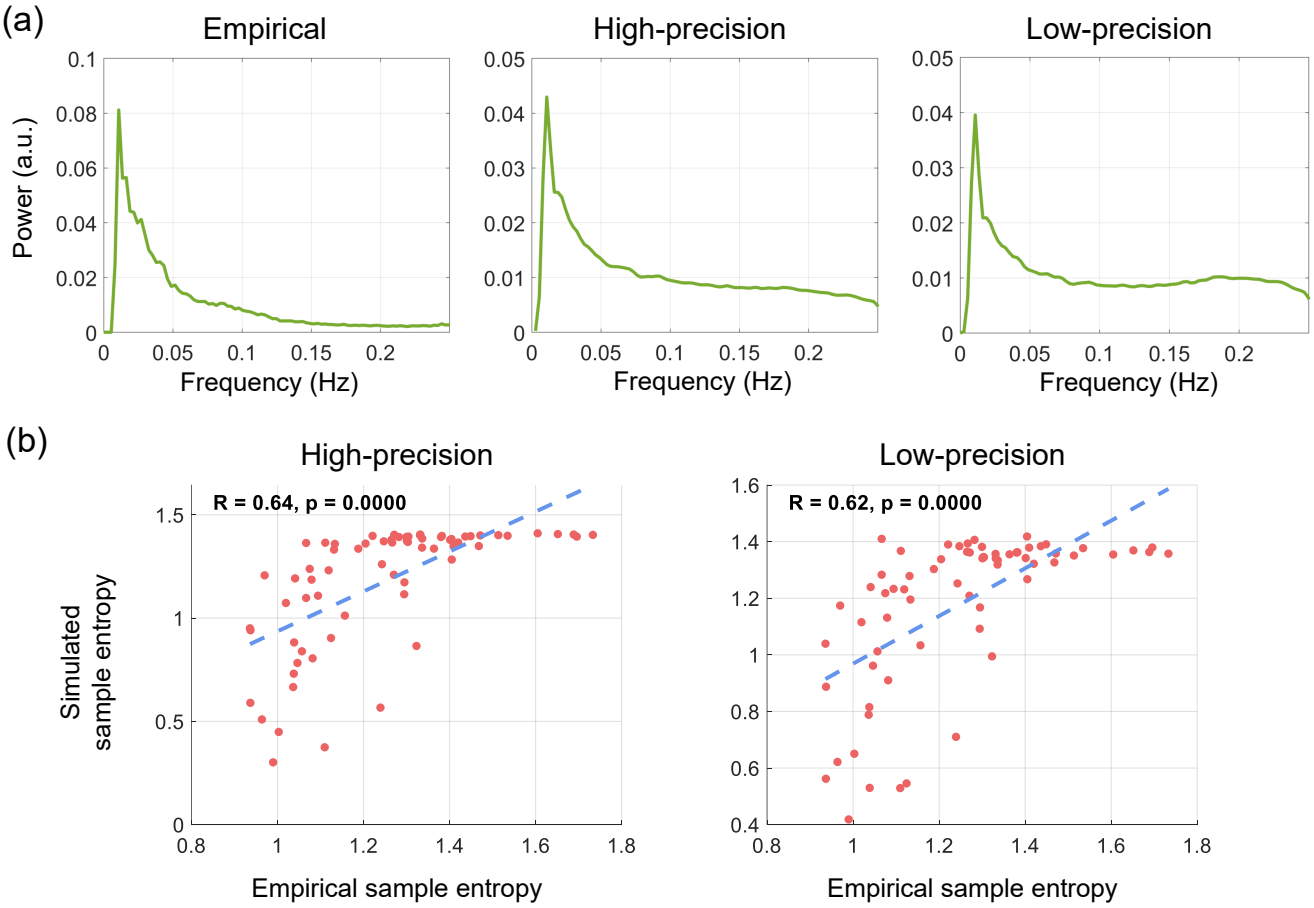

Fig. R5. Temporal characteristics comparison between different precision simulations. a. Group-averaged power spectral density computed from empirical and simulated time series across different model precisions. **b.** Correlation between the regional sample entropy values from simulated and empirical time series, comparing high-precision and low-precision simulation results. All correlations are statistically significant with $p < 10^{-4}$.

Added power spectral density of time series from high-precision simulation, low-precision simulation, and empirical rsfMRI data in Supplementary Fig. S11 (the same as Fig. R5). Added corresponding descriptions as follows.

In the part *Consistency of goodness-of-fit indicators between low- and high-precision models* of *Results* section: "Furthermore, we conduct temporal domain analyses to examine the impact of the quantization approach on temporal domain characteristics. We first compare the power spectral density of time series from high-precision simulation, low-precision simulation, and empirical rsfMRI data. As shown in Supplementary Fig. S11a, both high-precision and low-precision simulations exhibit power spectral profiles that closely match the empirical data in terms of peak locations and overall curve shapes, although specific amplitude values show slight differences. The low-precision simulation results demonstrate high similarity to high-precision counterparts. Notably, the low-precision results shows reduced peak power intensity while exhibiting elevation in other frequency components. This phenomenon is theoretically expected, as the quantization process introduces noise-like artifacts that elevate the overall power spectrum while reducing the signal-to-noise ratio. We also calculate the sample entropy for each brain region's simulated time series, and compute correlations with the empirical regional sample entropy. As demonstrated in Supplementary Fig. S11b, both precision levels show significant correlations with empirical data. Furthermore, the cor-

relation between high-precision and low-precision sample entropy values reaches $r = 0.9482$ ($p < 10^{-4}$), indicating strong consistency. The detailed description of the indicators and metrics we use are presented in Methods."

In the part *Goodness-of-fit indicators and other quality metrics* of *Methods* section: "For temporal complexity meatures, we employ the sample entropy as our primary metric. Sample Entropy quantifies the regularity of a time series by computing the negative natural logarithm of the conditional probability that two similar sequences will maintain their similarity when extended by one data point [2]. A higher sample entropy value reflects more complex nonlinear dynamics (e.g., chaotic or stochastic behaviors). This metric is widely employed in biomedical signal analysis, particularly for characterizing dynamic properties of neuroimaging data such as fMRI and EEG [3–5]."

References:

[1] T. Donoghue, M. Haller, E. J. Peterson, P. Varma, P. Sebastian, R. Gao, T. Noto, A. H. Lara, J. D. Wallis, R. T. Knight *et al.*, "Parameterizing neural power spectra into periodic and aperiodic components," *Nature neuroscience*, vol. 23, no. 12, pp. 1655–1665, 2020.

[2] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American journal of physiology-heart and circulatory physiology*, vol. 278, no. 6, pp. H2039–H2049, 2000.

[3] M. O. Sokunbi, W. Fung, V. Sawlani, S. Choppin, D. E. Linden, and J. Thome, "Resting state fmri entropy probes complexity of brain activity in adults with adhd," *Psychiatry Research: Neuroimaging*, vol. 214, no. 3, pp. 341–348, 2013.

[4] J. M. Yentes, N. Hunt, K. K. Schmid, J. P. Kaipust, D. McGrath, and N. Stergiou, "The appropriate use of approximate entropy and sample entropy with short data sets," *Annals of biomedical engineering*, vol. 41, no. 2, pp. 349–365, 2013.

[5] D. Abásolo, R. Hornero, P. Espino, D. Alvarez, and J. Poza, "Entropy analysis of the eeg background activity in alzheimer's disease patients," *Physiological measurement*, vol. 27, no. 3, p. 241, 2006.

Major Comment 3: ***Relatedly, I am concerned that Euler's method, on which the findings rely, is not adequate for simulations of chaotic systems with potentially complex attractors. Can other algorithms readily be implemented with the hardware and algorithms used here? If so, the authors should replicate at least part of the results with another integration method to demonstrate robustness.***

Response: First, to the best of our knowledge, almost all existing implementations of whole-brain coarse-grained dynamical models employ the Euler's method, including comprehensive frameworks such as neurolib that supports multiple model types. From this perspective, we believe the Euler's method represents the current standard practice for the computational scenarios we address.

Second, from an algorithmic validation standpoint, applying integration methods other than the Euler's method to the DMF model lacks established benchmarks. Even with the original high-precision model, it remains uncertain whether results obtained through alternative integration methods would be comparable to those from the Euler's method. Since our work focuses primarily on leveraging advanced computing architectures for acceleration, exploring different integration methods goes beyond the scope of this work.

Last, the field might eventually require other integration methods for specific applications. Our quantization framework is not inherently tied to the Euler's method, since the core principles could potentially be adapted to other integration schemes. The application of low-precision data types and computation to different integration schemes would indeed be a valuable direction for future research.

Revision: Added a brief discussion in the revised *Discussion* section: Furthermore, the proposed low-

precision methods are currently tested only with the Euler integration scheme. However, the field might require other integration methods for specific applications in the future, such as higher-order Runge-Kutta methods for enhanced accuracy or adaptive time-stepping algorithms for complex dynamical behaviors. The application of low-precision data types and computation to different integration schemes might be a valuable direction for future research.

Major Comment 4: ***I did not understand if parallelization was conducted over time at all. Paralleliza-tion over time should absolutely not be applied here, as dynamics in each time step depend on dynamics from previous time steps. More generally, how parallelization was performed here in comparison to state-of-the-art whole-brain modeling tools should be explained in much more detail. Modern implementations in languages such as Python already parallelize over regions, model realizations, etc, what is the difference here? Is it only that the memory is implemented in a distributed way instead of remotely accessed?***

Response: In our current implementation, we do not parallelize over time steps, as the evolution of brain dynamics is inherently causal, i.e., each time step depends on the results of the previous one. As you correctly mentioned, parallelization over the temporal dimension would violate this dependency and is therefore not applicable directly. However, temporal pipelining is a potential future direction for acceleration. Specifically, on brain-inspired computing chips, it is possible to divide the simulation across multiple cores, each responsible for a segment of the temporal evolution. This forms a temporal pipeline where the simulation is partitioned into sequential segments (e.g., time steps 1-100, 101-200, 201-300, etc.), with each core handling one segment. Within a single simulation run, adjacent segments cannot execute in parallel due to the causal dependency between consecutive time steps. For example, core 1 must complete processing time steps 1-100 before core 2 can begin processing time steps 101-200 using the results from core 1. However, this pipelining approach enables different simulation instances to overlap their execution: while core 1 processes time steps 1-100 of simulation instance A, core 2 can simultaneously process time steps 101-200 of a previously started simulation instance B, and core 3 can process time steps 201-300 of an even earlier simulation instance C. Despite introducing additional communication overhead for data transfer between cores, this temporal pipelining offers greater flexibility in resource mapping and has the potential to improve overall throughput on heterogeneous computing platforms by enabling concurrent processing of multiple simulation instances.

Existing whole-brain modeling tools, such as TVB [1], neurolib [2], and BrainPy [3], typically rely on high-performance computing (HPC) platforms and run on general-purpose CPUs or GPUs. They primarily exploit CPU multi-core parallelism (including vectorization) or invoke GPU acceleration through general-purpose libraries like JAX [4] and numba [5]. However, these tools do not finely control the hierarchical resource scheduling to fully utilize the hardware capabilities. Moreover, the parallel efficiency of the implementations is constrained by their underlying hardware architectures: CPUs possess limited parallel computing resources, while GPUs, despite having abundant computational units, may suffer from memory bandwidth bottlenecks. And both platforms employ considerably complex circuit designs to ensure generality, which introduce heavy overhead. In contrast, our approach encompasses two complementary aspects. (1) Like existing tools, our pipeline also utilizes GPU acceleration, but we provide detailed design and characterization of GPU resource mapping, explicitly examining the computational characteristics of whole-brain model inversion and designing tailored parallelization schemes. (2) More importantly, our approach mainly focuses on a brain-inspired computing chip with a non-von Neumann manycore architecture that supports near-memory or in-memory computing. This architecture tightly couples computation resources with memory, thereby reducing memory access overhead. As a specialized chip, it also enables more efficient processing through specialized low-precision circuit designs that minimize computational complexity.

Specifically, the parallelism in the brain-inspired computing chip operates at two levels: inter-core parallelism, where multiple candidate models are evaluated concurrently across different cores (population-level model inversion), and intra-core data parallelism, where differential equations for multiple nodes within a single model are solved in parallel inside each core. In practice, we map different model simulations to distinct cores to exploit the manycore distribution of the chip, while leveraging vectorization within each core to accelerate numerical updates of internal state variables. This hybrid strategy, combining model-level and node-level parallelism, allows for efficient resource utilization and significantly accelerating model inversion without violating the causal structure of temporal dynamics.

Revision: Added comparison to state-of-the-art whole-brain modeling tools and highlighted the advantages of the proposed methods in the revised *Discussion* section: "Existing whole-brain modeling tools, such as TVB [1], neurolib [2], and BrainPy [3], typically rely on high-performance computing (HPC) platforms and run on general-purpose CPUs or GPUs. They leverage thread- or process-level parallelism to accelerate the simulation of brain regions, connectivities, or model instances. However, they do not finely control the resource scheduling to fully utilize the hardware capabilities. And the parallel efficiency is constrained by their underlying hardware architectures: CPUs possess limited parallel computing resources, while GPUs, despite having abundant computational units, may suffer from memory bandwidth bottlenecks. And both platforms employ considerably complex circuit designs to ensure generality, which introduce heavy overhead. By contrast, our work focuses on detailed mapping between algorithmic components and architectural resources, and explores the application of the non-von Neumann brain-inspired computing architectures, which features tightly coupled computation and memory with specialized low-precision circuit designs. Specifically, we adopt a population-based metaheuristic inversion algorithm that is more amenable to parallelization and analyze multiple levels of parallelism in model inversion. We develop tailored mapping strategies that distribute multi-level parallel tasks across the computing and memory resources of both architectures. These strategies enable fine-grained and low-overhead scheduling, particularly for the brain-inspired computing architecture, providing more efficient parallelism."

Added discussion about temporal pipelining on the brain-inspired computing chip in the revised *Discussion* section: "Alternatively, we can divide T time steps of a model simulation across multiple cores, each responsible for a segment of the temporal evolution. This forms a temporal pipeline where the simulation is partitioned into sequential segments (e.g., time steps 1-100, 101-200, 201-300, etc.), with each core handling one segment. Within a single simulation run, adjacent segments cannot execute in parallel due to the causal dependency between consecutive time steps. For example, core 1 must complete processing time steps 1-100 before core 2 can begin processing time steps 101-200 using the results from core 1. However, this pipelining approach enables different simulation instances to overlap their execution: while core 1 processes time steps 1-100 of simulation instance A, core 2 can simultaneously process time steps 101-200 of a previously started simulation instance B, and core 3 can process time steps 201-300 of an even earlier simulation instance C. Despite introducing additional communication overhead for data transfer between cores, it provides considerable flexibility for resource mapping. This may help improve resource utilization and overall throughput. The complex spatiotemporal mapping problem is beyond the scope of this work and thus left for future work."

References:

[1] P. Sanz Leon, S. A. Knock, M. M. Woodman, L. Domide, J. Mersmann, A. R. McIntosh, and V. Jirsa, "The virtual brain: a simulator of primate brain network dynamics," *Frontiers in neuroinformatics*, vol. 7, p. 10, 2013.

[2] C. Cakan, N. Jajcay, and K. Obermayer, "neurolib: A simulation framework for whole-brain neural mass modeling," *Cognitive Computation*, vol. 15, no. 4, pp. 1132–1152, 2023.

[3] C. Wang, T. Zhang, X. Chen, S. He, S. Li, and S. Wu, "Brainpy, a flexible, integrative, efficient, and

extensible framework for general-purpose brain dynamics programming," *elife*, vol. 12, p. e86365, 2023.
[4] S. Schoenholz and E. D. Cubuk, "Jax md: a framework for differentiable physics," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11 428–11 441, 2020.
[5] S. K. Lam, A. Pitrou, and S. Seibert, "Numba: A llvm-based python jit compiler," in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, 2015, pp. 1–6.

Major Comment 5: ***How long are the rsfMRI time series that simulations are compared to? Empirical brain dynamics are highly non-stationary as mentioned throughout the manuscript, with parameters varying and fixed points shifting through time, and I worry if quantization of firing rates can robustly account for nonstationary brain activity, when compared to high-resolution models.***

Response: The simulation time series length in our study strictly corresponds to the time series length provided in the dataset. Specifically, we use the DMF model to simulate resting-state brain activities over 16.4 minutes, and employ the latter 14.4 minutes (1200 time points) of the time series for comparison with the rsfMRI data from each subject in the dataset. This time series length is consistent with previous works that have used HCP rs-fMRI data [1].

For the robustness of our quantization method in the context of non-stationary brain activities, we are glad to discuss in detail, as it is crucial for understanding the applicability of our method to whole-brain dynamics models.

First, to clarify, our quantization method is designed to operate in the established simulation procedure for whole-brain models. Most current macroscopic whole-brain models are used to fit resting-state data, which is also the primary scenario we consider. The standard simulation procedure typically begins from an arbitrarily specified initial state. After a warm-up period, when the model's signal activities reach a largely stable state, the signal characteristics are analyzed and compared with empirical resting-state data. When we mentioned "time variation" or "non-stationary" in the manuscript, we were mainly referring to the warm-up or transient regime of the simulation between the initial state and the stable state. Our quantization method builds upon the simulation procedure by adding a quantization parameter selection phase during the transient regime and implementing the stable-state simulation using low-precision data types and computation.

Second, we wish to define the applicability and constraints of semi-dynamic quantization more explicitly. The robustness of the low-precision quantization is determined by the **numerical range of variables** rather than the specific dynamic behaviors. Low-precision data types have inherently limited the representational range and precision. Therefore, as long as the ranges of variables and their derivatives of the model do not undergo substantial changes over time, low-precision quantization can remain effective. The DMF model converges to a fixed point with variables fluctuating within a bounded range, thus satisfying this requirement. However, this does not imply that only models converging to fluctuations around a fixed point can utilize the proposed semi-dynamic quantization method. For instance, in oscillator-based models such as the Kuramoto or Hopf model, state variables may oscillate over a wider range. Nevertheless, as long as their ranges remain relatively stable, semi-dynamic quantization can still be applicable. Previous literature empirically indicates that the resting-state FC does not exhibit extremely drastic transient changes (although some time-varying information might be extractable) [2]. Therefore, we believe that semi-dynamic quantization should be capable of meeting the computational demands of resting-state modeling of brain dynamics.

Looking toward future extension, we can envision scenarios where the ranges of variables change significantly during model execution. If such changes are predictable, semi-dynamic quantization could

still be employed. For instance, when simulating brain lesions [3], perturbations are typically introduced to a baseline resting-state brain model. These perturbations can cause the model to transfer from one stable state to another. Since the timing of these perturbations is known as a priori, high-precision floating-point numbers could be used for a new warm-up phase, while a new low-precision quantization scheme could be employed once another stable state is reached. By iteratively applying the warm-up and simulation stages, the semi-dynamic quantization method could be potentially extended to more complex scenarios. While these scenarios are theoretically plausible, they go beyond the scope of the current paper; therefore, we just discuss them in the revised Discussion section.

We also acknowledge that the semi-dynamic quantization method has inherent limitations. When a dynamical model exhibits frequent or unpredictable large-magnitude numerical fluctuations, this method would likely become unsuitable. To our knowledge, existing macroscopic brain dynamics models do not exhibit such characteristics under typical simulation conditions, but we have explicitly mentioned this limitation in the revised Discussion section.

Revision: Added the time series information in the revised part *Goodness-of-fit indicators and other quality metrics* of *Methods* section: "...we calculate the FC matrix for each subject based on the Pearson correlation coefficient between the BOLD time series. The simulated time series length strictly corresponds to the time series length provided in the dataset (i.e., 14.4 minutes). The FC matrices are then averaged..."

Modified the descriptions related to "time variation" or "non-stationary" to ensure a more rigorous and precise expression. In *Introduction*: "With the semi-dynamic quantization strategy, we can address the large temporal variations during the transient phase and achieve stable long-duration simulation of dynamic models using low-precision integers once the numerical ranges stabilize."

Provided detailed explanations regarding the prevalence of transient and stable regimes in the part *Dynamics-aware quantization for macroscopic brain models* of the *Results* section: "We introduce semi-dynamic quantization to address this problem. Most current macroscopic whole-brain models are developed based on the resting-state data. The simulation of the brain's resting state activity can be divided into two stages. In the warm-up stage, the variables may start from an arbitrary state and experience a transient period with significant changes before converging to a relatively stable state. In the simulation stage, the model typically runs in a relatively stable state where the variables fluctuate randomly or oscillate within a certain range. These stages represent a common practice in the model simulation scenario, regardless of which specific model is employed."

Clarified the scope of applicability of the proposed method in the part *Dynamics-aware quantization for macroscopic brain models* of the *Results* section: "The applicability of semi-dynamic quantization depends on the numerical range of variables. As long as the ranges of variables and their derivatives of the model do not undergo substantial changes over time, low-precision quantization can remain effective. Previous literature empirically indicates that the resting-state FC does not exhibit extremely drastic transient changes (although some time-varying information might be extractable) [2]. Therefore, we believe that semi-dynamic quantization should be capable of meeting the computational demands of resting-state modeling of brain dynamics across various model types. It is also worth..."

Added discussions regarding the broader applicability of semi-dynamic quantization in potential scenarios in the revised part *Semi-dynamic quantization for brain dynamics* of the *Methods* section: "...In the subsequent simulation stage, all variables are represented as integers. We can envision scenarios where the ranges of variables change significantly during model execution. If such changes are predictable, semi-dynamic quantization could still be employed. For instance, when simulating brain lesions [3], perturbations are typically introduced to a baseline resting-state brain model. These perturbations can

cause the model to transfer from one stable state to another. Since the timing of these perturbations is known as a priori, high-precision floating-point numbers could be used for the transition phase, while a new low-precision quantization scheme could be employed once another stable state is reached. By iteratively applying the warm-up and simulation stages, the semi-dynamic quantization method could be potentially extended to more complex scenarios."

Added discussions regarding the limitations of the quantization approaches in the revised *Discussion* section: "...allowing broader scientific computing tasks to benefit from advanced computing architectures. However, the proposed method also suffers inherent limitations: When a dynamical model exhibits frequent or unpredictable large-magnitude fluctuations in its numerical range, the semi-dynamic quantization would likely become unsuitable. There exists a fundamental trade-off between quantization precision and generality: higher quantization precision enables support for a broader variety of dynamical models with better simulation fidelity, but may result in reduced computational efficiency. Future work should explore adaptive quantization schemes that can dynamically balance this trade-off based on the specific characteristics of the target dynamic system and the available computational resources. Additionally, the computational operators supported by existing brain-inspired computing chips are limited and may not be capable of implementing arbitrary dynamic systems. These constraints highlight the need for future research in developing more flexible quantization strategies and expanding the computational capabilities of brain-inspired computing architectures to accommodate a broader range of scientific computing applications."

References:

[1] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium *et al.*, "The wu-minn human connectome project: an overview," *Neuroimage*, vol. 80, pp. 62–79, 2013.

[2] T. O. Laumann, A. Z. Snyder, A. Mitra, E. M. Gordon, C. Gratton, B. Adeyemo, A. W. Gilmore, S. M. Nelson, J. J. Berg, D. J. Greene *et al.*, "On the stability of bold fmri correlations," *Cerebral cortex*, vol. 27, no. 10, pp. 4719–4732, 2017.

[3] J. Wei, B. Wang, Y. Yang, Y. Niu, L. Yang, Y. Guo, and J. Xiang, "Effects of virtual lesions on temporal dynamics in cortical networks based on personalized dynamic models," *NeuroImage*, vol. 254, p. 119087, 2022.

Major Comment 6: ***Another important concern is scalability to larger numbers of parameters to infer. Recent work (Zhang et al, PNAS 2024) emphasizes the need for inferring more parameters, such as region-specific parameters, toward personalized whole-brain models. The manuscript should demonstrate scalability to more than 3 parameters to some extent, or acknowledge lack of scalability if this cannot be shown.***

Response: The scalability you mentioned is an important factor. In fact, our method supports scaling to larger numbers of parameters, which we will discuss from both low-precision quantization and hardware acceleration perspectives.

From the quantization perspective, the number of model parameters does not fundamentally change the basic quantization workflow. Regardless of how many tunable parameters exist in the model, the quantization process relies on determining quantization parameters based on floating-point simulation results during the warm-up stage and the QPS stage. To address your concern, we have explicitly performed model inversion using the proposed low-precision simulation to train a model with region-specific local excitatory recurrence parameters (w) on empirical data. We used the PSO algorithm to search for optimal parameters. Here the noise amplitude is also included as an optimizable parameter. We conducted 30 repeated simulations using both high-precision and low-precision models with the searched parameters. The

correlation between low-precision simulated and empirical FC results is, on average, 0.7299 ($p < 10^{-4}$), with an SD of 0.0115. The correlation between high-precision simulated and empirical FC results also reaches 0.7228 ($p < 10^{-4}$), with an SD of 0.0093. Meanwhile, the correlation between the FC results from the two precisions reaches 0.9603 on average ($p < 10^{-4}$). As a baseline, the correlation between the empirical SC and FC is 0.4545.

The results demonstrate that our quantization approach maintains high fidelity even with increased parameter complexity. In addition, we find that region-specific parameters lead to larger difference of the numerical ranges across brain regions. The proposed group-wise quantization strategy can effectively handle this heterogeneity.

From the hardware acceleration perspective, the primary impact of increasing the parameter number is on the population size and iteration count required by population-based optimization algorithms like PSO and CMA-ES, as higher-dimensional search spaces demand greater search efforts. However, these algorithmic changes do not affect the applicability of our proposed method. In fact, larger population sizes increase the parallelism potential of the optimization algorithm, allowing better utilization of more parallel hardware resources and potentially increasing the advantage of GPUs and brain-inspired computing chips over CPUs. Even when the parallel architectures cannot simultaneously evaluate the entire population, all computations can be completed through simple time-division multiplexing. The number of iterations clearly does not impact the hardware deployment feasibility.

Therefore, our approach demonstrates robust scalability to larger parameter spaces, with the quantization strategy adapting naturally to increased model complexity while the parallel hardware architectures providing enhanced benefits as computational demands grow.

Revision: Added a part *Scalability to larger numbers of parameters* in the *Methods* section: "Recent work emphasizes the need for inferring whole-brain models with larger numbers of parameters [1, 2], including region-specific parameters. The proposed pipeline can scale effectively to accommodate such increased parameter complexity. From the quantization perspective, the number of model parameters does not fundamentally change the basic quantization workflow. Regardless of how many tunable parameters exist in the model, the quantization process relies on determining quantization parameters based on floating-point results during the warm-up stage and the QPS stage. We explicitly perform model inversion using the proposed low-precision simulation to train a model with region-specific local excitatory recurrence parameters (w) on empirical data. We use the PSO algorithm to search for optimal parameters. Here the noise amplitude is also included as an optimizable parameter. We conducted 30 repeated simulations using both high-precision and low-precision models with the searched parameters. The correlation between low-precision simulated and empirical FC results is, on average, 0.7299 ($p < 10^{-4}$), with an SD of 0.0115. The correlation between high-precision simulated and empirical FC results also reaches 0.7228 ($p < 10^{-4}$), with an SD of 0.0093. Meanwhile, the correlation between the FC results from the two precisions reaches 0.9603 on average ($p < 10^{-4}$). As a baseline, the correlation between the empirical SC and FC is 0.4545. The results demonstrate that our quantization approach maintains high fidelity even with increased parameter complexity. In addition, we find that region-specific parameters lead to larger difference of the numerical ranges across brain regions. The proposed group-wise quantization strategy can effectively handle this heterogeneity.

From the hardware acceleration perspective, the primary impact of increasing the parameter number is on the population size and iteration count of the population-based optimization algorithms like PSO and CMA-ES, as higher-dimensional search spaces demand greater search efforts. However, these algorithmic changes do not affect the applicability of our proposed method. In fact, larger population sizes increase the parallelism potential of the optimization algorithm, allowing better utilization of more parallel hardware

resources and potentially increasing the advantage of GPUs and brain-inspired computing chips over CPUs. Even when the parallel architectures cannot simultaneously evaluate the entire population, all computations can be completed through simple time-division multiplexing. The number of iterations clearly does not impact the hardware deployment feasibility.

In summary, the proposed approaches can scale to larger parameter spaces, with the quantization strategy adapting naturally to increased model complexity while the parallel hardware architectures providing enhanced benefits as computational demands grow."

References:

[1] P. Wang, R. Kong, X. Kong, R. Liégeois, C. Orban, G. Deco, M. P. Van Den Heuvel, and B. Thomas Yeo, "Inversion of a large-scale circuit model reveals a cortical hierarchy in the dynamic resting human brain," *Science advances*, vol. 5, no. 1, p. eaat7854, 2019.

[2] S. Zhang, B. Larsen, V. J. Sydnor, T. Zeng, L. An, X. Yan, R. Kong, X. Kong, R. C. Gur, R. E. Gur *et al.*, "In vivo whole-cortex marker of excitation-inhibition ratio indexes cortical maturation and cognitive ability in youth," *Proceedings of the National Academy of Sciences*, vol. 121, no. 23, p. e2318641121, 2024.

Major Comment 7: ***The manuscript methods claim that a particle swarm algorithm was adapted to perform metaheuristic inferences. How does this compare in performance and speed to state-of-the-art methods in whole-brain model inference e.g. evolutionary algorithms (Zhang et al, PNAS 2024, Cakan et al, Cog Computation 2023) and simulation-based inference (Hashemi et al, Mach Learn Sci Technol 2024)? Are they comparable? Are other algorithms, such as those mentioned, able to be easily and readily integrated with the hardware and model presented here?***

Response: The specific optimization method is not the primary innovation of this work. We selected PSO as it is a representative population-based optimization algorithm, which aligns well with the hardware acceleration approach. Since the focus of this paper is on low-precision quantization and hardware resource mapping, a purely algorithmic comparison between PSO and other methods might go beyond our scope in this work and would be easily influenced by various hyper-parameters. However, we agree that it is necessary to more clearly articulate the potential and effectiveness of integrating other algorithms.

First, we believe most population-based optimization algorithms can be readily integrated into our proposed pipeline. This includes CMA-ES algorithms in [1] and grid search methods in [2]. These approaches all require evaluating large numbers of parameter sets (e.g., populations in evolutionary algorithms, particle swarms in PSO, and grid points in grid search) where the evaluation process can be fully parallelized. Therefore, they can take full advantage of the parallelism offered by advanced computing architectures.

In these algorithms, both high-precision and low-precision simulations serve as evaluation or cost functions. As long as the goodness-of-fit between low-precision simulation results and empirical data shows similar distribution patterns as high-precision simulation across the parameter space, these search algorithms should achieve valid results with low-precision simulation. This similarity in distribution is what we want to demonstrate in Fig. 3 and our response to your Major Comment 1.

To further address your concerns, we have tested our low-precision simulation approach with two different algorithms: PSO and CMA-ES, comparing all results against high-precision models. Unlike previous experiments, here we used the larger HCP dataset containing 1004 participants and employed a validation procedure similar to that described in [1]. We first performed algorithmic search on the training set, then selected optimal parameters using the validation set, and finally conducted repeated

simulations on the test set to calculate FC correlations. The results in Table R5 demonstrate that the FC correlations achieved by different algorithm-precision combinations are similar. No significant differences were observed between PSO and CMA-ES results, indicating that the choice of optimization algorithms does not substantially impact the effectiveness of our low-precision approach.

Table R5. Model inversion results using different algorithms with different precisions. Each row represents an algorithm-precision combination used during the parameter search process. We perform both high-precision and low-precision simulations using the searched parameter sets repeatedly and calculated the corresponding FC. Each column shows the correlation coefficients between pairs of high-precision simulated FC, low-precision simulated FC, and empirical FC. All correlation results are statistically significant with $p < 10^{-4}$. As a baseline, the correlation between the empirical SC and FC is 0.4759.

	FC corr. (high precis. and emp.)	FC corr. (low precis. and emp.)	FC corr. (high precis. and low precis.)
PSO (high precision)	0.5630 ($\pm$0.0085)	0.5612 ($\pm$0.0095)	0.9964 ($\pm$0.0005)
PSO (low precision)	0.5620 ($\pm$0.0107)	0.5713 ($\pm$0.0110)	0.9814 ($\pm$0.0018)
CMA-ES (high precision)	0.5618 ($\pm$0.0079)	0.5593 ($\pm$0.0069)	0.9952 ($\pm$0.0006)
CMA-ES (low precision)	0.5615 ($\pm$0.0149)	0.5660 ($\pm$0.0145)	0.9907 ($\pm$0.0012)

Interestingly, we observed a subtle cross-precision effect: parameters optimized using high-precision models show slightly reduced performance when evaluated with low-precision models, and vice versa. This suggests that low-precision simulation introduces minor alterations to the spatial distribution of fitting indicators in the parameter space compared to high-precision simulation. We believe this phenomenon is closely related to the observations in Fig. 3 and Supplementary Fig. S3 regarding the optimal parameter differences between models with different precisions. This issue will be discussed in detail in our response to your Minor Comment 7.

Regarding the simulation-based inference (SBI) method you mentioned [3], we agree it is an important and emerging approach. If we understand correctly, SBI generates large datasets through parameter sampling and simulation, then trains artificial neural networks to estimate the entire posterior distribution of unknown parameters. The large-scale sampling and simulation phase in SBI is indeed parallelizable and could benefit from our proposed acceleration methods. However, SBI differs fundamentally from the objective of our pipeline: it aims to estimate the entire posterior distribution of unknown parameters (i.e., uncertainty over a range of plausible parameter values), rather than finding a single-point estimate for data fitting. Therefore, while our methods could accelerate the simulation component of SBI, the overall framework addresses a different computational scenario.

We also would like to acknowledge the limitations of our proposed pipeline. First, algorithms with lower parallelism, such as some Bayesian estimation methods, may not fully utilize the parallel resources of advanced computing architectures, resulting in less pronounced acceleration benefits. Second, the goodness-of-fit obtained from low-precision simulations varies less smoothly with parameters compared to high-precision models (as shown in Fig. 3b), which may make our approach less suitable for some local optimization algorithms.

Revision: Added the model inversion results using different algorithm-precision combinations in Supplementary Table S7 (corresponding to Table R5).

Described the algorithms and results in the revised part *Metaheuristics for model inversion* of the *Methods* section: "In fact, most population-based optimization algorithms can be readily integrated into our proposed pipeline. This includes CMA-ES algorithms in [1] and grid search methods in [2]. These approaches all require evaluating large numbers of parameter sets (e.g., populations in evolutionary algorithms, particle swarms in PSO, and grid points in grid search) where the evaluation process can be fully parallelized. Therefore, they can take full advantage of the parallelism offered by advanced computing architectures. In these algorithms, both high-precision and low-precision simulations serve as evaluation or cost functions. As long as the goodness-of-fit between low-precision simulation results and empirical data shows similar distribution patterns as high-precision simulations across the parameter space, these search algorithms should achieve valid results with low-precision simulations. This similarity in distribution is what we want to demonstrate in Fig. 3.

We test our low-precision simulation approach with two different algorithms: PSO and CMA-ES, comparing all results against high-precision models. Unlike previous experiments, here we use the larger HCP dataset containing 1004 participants and employed a validation procedure similar to that described in [1]. We first performed algorithmic search on the training set, then selected optimal parameters using the validation set, and finally conducted repeated simulations on the test set to calculate FC correlations. The results in Supplementary Table S7 demonstrate that the FC correlations achieved by different algorithm-precision combinations are similar. No significant differences were observed between PSO and CMA-ES results, indicating that the choice of optimization algorithm does not substantially impact the effectiveness of our low-precision approach. The FC correlations are not as high as those reported in [1] because we employ a distinct model with fewer parameters."

Pointed out the limitations of the proposed pipeline in the revised part *Metaheuristics for model inversion* of the *Methods* section: "Beyond population-based metaheuristics, the proposed pipeline has certain limitations when integrating with other algorithms. First, algorithms with lower parallelism, such as some Bayesian estimation methods, may not fully utilize the parallel resources of advanced computing architectures, resulting in less pronounced acceleration benefits. Second, the goodness-of-fit obtained from low-precision simulations varies less smoothly with parameters compared to high-precision models (as shown in Fig. 3b), which may make the approaches less suitable for some local optimization algorithms."

References:

[1] S. Zhang, B. Larsen, V. J. Sydnor, T. Zeng, L. An, X. Yan, R. Kong, X. Kong, R. C. Gur, R. E. Gur *et al.*, "In vivo whole-cortex marker of excitation-inhibition ratio indexes cortical maturation and cognitive ability in youth," *Proceedings of the National Academy of Sciences*, vol. 121, no. 23, p. e2318641121, 2024.

[2] C. Cakan, N. Jajcay, and K. Obermayer, "neurolib: A simulation framework for whole-brain neural mass modeling," *Cognitive Computation*, vol. 15, no. 4, pp. 1132–1152, 2023.

[3] M. Hashemi, A. Ziaeemehr, M. M. Woodman, J. Fousek, S. Petkoski, and V. K. Jirsa, "Simulation-based inference on virtual brain models of disorders," *Machine Learning: Science and Technology*, vol. 5, no. 3, p. 035019, 2024.

Major Comment 8: ***Especially for wider audiences unfamiliar with the details of neuromorphic hardware, applicability and flexibility of the modeling and inference methodology should be discussed to address potential concerns. For instance, are there hardware changes needed to implement a changed structural connectome? If so, how long do they take? Same question to implement different firing rate differential equations.***

Response: This is an important consideration for the adoption of our methodology. Regarding the TianjicX

neuromorphic chip used in our work, it is a programmable platform that executes different operations based on configured instructions [1]. Although TianjicX is a prototype chip designed for academic research, it also has a toolchain that allows us to generate chip instructions using Python programming. This means that when model parameters, connectomes, or equations change, only code modifications are required without any hardware changes.

Specifically, the required modifications are summarized below.

(1) Structural connectome changes: When the numerical values in the structural connectome change, we only need to adjust the connection matrix data configured in the chip. This typically requires a few minutes.

(2) Scale changes: When the connectome scale (e.g., number of brain regions) changes, we only need to modify corresponding variables in the code to adjust instruction fields in most cases.

(3) Differential equation changes: When firing rate differential equations change, the code needs to be adjusted according to the new equations. This process is more complex and may require several hours for implementation.

Essentially, using neuromorphic chips like TianjicX is similar to using CPUs and GPUs, except that the toolchain support may not be very mature, and the specific coding process is usually more complex. However, these modifications are manageable for skilled programmers and do not require hardware changes.

Looking at the broader landscape of neuromorphic chips, not all neuromorphic chips share the same programmable characteristics as TianjicX we used. Some specialized neuromorphic chips may not support the functionalities required for various whole-brain models. However, programmability and configurability has emerged as a dominant trend in the neuromorphic computing field [2]. Many influential neuromorphic chips developed in recent years demonstrate excellent programmability, with several already providing comprehensive toolchains that facilitate easy deployment and adaptation [3, 4]. Therefore, we are highly optimistic about the applicability and flexibility of current and future neuromorphic platforms. As the field continues to improve and toolchains become increasingly sophisticated, we anticipate that deploying our methodology will become more straightforward across diverse neuromorphic architectures.

Revision: Added more descriptions regarding the usage of TianjicX in the revised part *Implementation of model operations on brain-inspired computing chips* of *Methods* section: "...We use a brain-inspired computing chip, TianjicX[1], as the validation platform. It is a programmable platform that executes different operations based on configured instructions as mentioned above. Although TianjicX is a prototype chip designed for academic research, it also has a toolchain that allows us to generate chip instructions using Python programming."

Added descriptions of the effort required for various changes in algorithms and models in the revised part *Implementation of model operations on brain-inspired computing chips* of *Methods* section: "When model parameters, connectomes, or equations change, code modifications are required as follows: When the numerical values in the structural connectome change, we only need to adjust the connection matrix data configured in the chip; when the connectome scale (e.g., number of brain regions) changes, we need to modify corresponding variables in the code to adjust instruction fields; when firing rate differential equations change, the code needs to be adjusted according to the new equations, which may require longer time for implementation."

Added discussions on the applicability and flexibility of broader neuromorphic hardware platforms

in the revised part *Implementation of model operations on brain-inspired computing chips* of *Methods* section: "Looking at the broader landscape of neuromorphic chips, not all neuromorphic chips share the same programmable characteristics, nor do all specialized neuromorphic chips support the full range of functionalities required for various whole-brain models. However, programmability and configurability has emerged as a dominant trend in the neuromorphic computing field [2]. Therefore, we are highly optimistic about the applicability and flexibility of current and future neuromorphic platforms."

References:

[1] S. Ma, J. Pei, W. Zhang, G. Wang, D. Feng, F. Yu, C. Song, H. Qu, C. Ma, M. Lu *et al.*, "Neuromorphic computing chip with spatiotemporal elasticity for multi-intelligent-tasking robots," *Science Robotics*, vol. 7, no. 67, p. eabk2948, 2022.

[2] D. Kudithipudi, C. Schuman, C. M. Vineyard, T. Pandit, C. Merkel, R. Kubendran, J. B. Aimone, G. Orchard, C. Mayr, R. Benosman *et al.*, "Neuromorphic computing at scale," *Nature*, vol. 637, no. 8047, pp. 801–812, 2025.

[3] W. Fang, Y. Chen, J. Ding, Z. Yu, T. Masquelier, D. Chen, L. Huang, H. Zhou, G. Li, and Y. Tian, "Spikingjelly: An open-source machine learning infrastructure platform for spike-based intelligence," *Science Advances*, vol. 9, no. 40, p. eadi1480, 2023.

[4] G. Orchard, E. P. Frady, D. B. D. Rubin, S. Sanborn, S. B. Shrestha, F. T. Sommer, and M. Davies, "Efficient neuromorphic signal processing with loihi 2," in *2021 IEEE Workshop on Signal Processing Systems (SiPS)*. IEEE, 2021, pp. 254–259.

Minor Comment 1: ***"can integrate macroscopic empirical data" - references are needed to back this point up***

Response: We have added appropriate references [1–5] (i.e., refs. [6-10] in the revised manuscript) to support this statement.

References:

[1] X. Kong, R. Kong, C. Orban, P. Wang, S. Zhang, K. Anderson, A. Holmes, J. D. Murray, G. Deco, M. van den Heuvel *et al.*, "Sensory-motor cortices shape functional connectivity dynamics in the human brain," *Nature communications*, vol. 12, no. 1, p. 6373, 2021.

[2] J. Vohryzek, J. Cabral, P. Vuust, G. Deco, and M. L. Kringelbach, "Understanding brain states across spacetime informed by whole-brain modelling," *Philosophical Transactions of the Royal Society A*, vol. 380, no. 2227, p. 20210247, 2022.

[3] M. Breakspear, "Dynamic models of large-scale brain activity," *Nature neuroscience*, vol. 20, no. 3, pp. 340–352, 2017.

[4] G. Deco and M. L. Kringelbach, "Great expectations: using whole-brain computational connectomics for understanding neuropsychiatric disorders," *Neuron*, vol. 84, no. 5, pp. 892–905, 2014.

[5] G. Deco, G. Tononi, M. Boly, and M. L. Kringelbach, "Rethinking segregation and integration: contributions of whole-brain modelling," *Nature reviews neuroscience*, vol. 16, no. 7, pp. 430–439, 2015.

Minor Comment 2: ***Ref 14-16 not about diagnosing brain disorders***

Response: You are correct that refs. 14-16 are not specifically about diagnosing brain disorders. Whole-brain dynamics modeling is usually used for investigating abnormal mechanisms of brain diseases and predicting therapeutic interventions, rather than direct clinical diagnosis. Considering your comment, we have revised the texts to more accurately reflect the actual applications of whole-brain modeling in

neuroscience research, replacing "diagnosing brain disorders" with "investigating abnormal mechanisms of brain diseases and predicting therapeutic effects".

Minor Comment 3: ***"brain-inspired computing chips are increasingly oriented to intelligent computing workloads rather than scientific computing ones" - what intelligent vs scientific means is unclear***

Response: For "intelligent computing workloads", we refer to applications primarily focus on artificial intelligence tasks such as pattern recognition, computer vision, and natural language processing. These workloads typically involve neural network computation, feature extraction, and decision-making process that mimic cognitive functions.

In contrast, "scientific computing workloads" encompass numerical simulation, mathematical modeling, high-performance computing, and computational science problems such as solving differential equations, brain simulation and molecular dynamics simulation. These applications typically require high-precision floating-point operations.

Most contemporary brain-inspired computing chips are indeed designed with artificial intelligence applications as their primary targets [1, 2], emphasizing energy efficiency for neural network operations rather than the high-precision numerical computation traditionally required in scientific computing domains. Our work demonstrates an innovative application of such hardware to scientific computing problems in computational neuroscience, bridging this gap through the proposed framework.

Revision: Added descriptions of the two phrases in the revised *Introduction* section: "First, brain-inspired computing chips are increasingly oriented to intelligent computing workloads [1, 2] (i.e., artificial intelligence tasks such as pattern recognition, computer vision, and natural language processing, typically involving neural network computation) rather than scientific computing ones (e.g., numerical simulation, mathematical modeling, requiring high-precision floating-point computation for solving differential equations), so they prioritize low-precision computing to reduce hardware resource costs and power consumption [3–6]."

References:

[1] G. Li, L. Deng, H. Tang, G. Pan, Y. Tian, K. Roy, and W. Maass, "Brain-inspired computing: A systematic survey and future trends," *Proceedings of the IEEE*, 2024.

[2] D. Kudithipudi, C. Schuman, C. M. Vineyard, T. Pandit, C. Merkel, R. Kubendran, J. B. Aimone, G. Orchard, C. Mayr, R. Benosman *et al.*, "Neuromorphic computing at scale," *Nature*, vol. 637, no. 8047, pp. 801–812, 2025.

[3] Y. Zhang, P. Qu, Y. Ji, W. Zhang, G. Gao, G. Wang, S. Song, G. Li, W. Chen, W. Zheng *et al.*, "A system hierarchy for brain-inspired computing," *Nature*, vol. 586, no. 7829, pp. 378–384, 2020.

[4] G. Orchard, E. P. Frady, D. B. D. Rubin, S. Sanborn, S. B. Shrestha, F. T. Sommer, and M. Davies, "Efficient neuromorphic signal processing with loihi 2," in *2021 IEEE Workshop on Signal Processing Systems (SiPS)*. IEEE, 2021, pp. 254–259.

[5] S. Höppner, Y. Yan, A. Dixius, S. Scholze, J. Partzsch, M. Stolba, F. Kelber, B. Vogginger, F. Neumärker, G. Ellguth *et al.*, "The spinnaker 2 processing element architecture for hybrid digital neuromorphic computing," *arXiv preprint arXiv:2103.08392*, 2021.

[6] Y. Zhong, Y. Kuang, K. Liu, Z. Wang, S. Feng, G. Chen, Y. Yang, X. Cui, Q. Wang, J. Cao *et al.*, "Paicore: a 1.9-million-neuron 5.181-tsops/w digital neuromorphic processor with unified snn-ann and on-chip learning paradigm," *IEEE Journal of Solid-State Circuits*, 2024.

Minor Comment 4: ***"they prioritize low-precision computing to reduce hardware resource costs and power consumption." - references are needed here.***

Response: We have added references [1–4] (i.e., refs. [26, 34, 47, 48] in the revised manuscript).

References:

[1] Y. Zhang, P. Qu, Y. Ji, W. Zhang, G. Gao, G. Wang, S. Song, G. Li, W. Chen, W. Zheng *et al.*, "A system hierarchy for brain-inspired computing," *Nature*, vol. 586, no. 7829, pp. 378–384, 2020.
[2] G. Orchard, E. P. Frady, D. B. D. Rubin, S. Sanborn, S. B. Shrestha, F. T. Sommer, and M. Davies, "Efficient neuromorphic signal processing with loihi 2," in *2021 IEEE Workshop on Signal Processing Systems (SiPS)*. IEEE, 2021, pp. 254–259.
[3] S. Höppner, Y. Yan, A. Dixius, S. Scholze, J. Partzsch, M. Stolba, F. Kelber, B. Vogginger, F. Neumärker, G. Ellguth *et al.*, "The spinnaker 2 processing element architecture for hybrid digital neuromorphic computing," *arXiv preprint arXiv:2103.08392*, 2021.
[4] Y. Zhong, Y. Kuang, K. Liu, Z. Wang, S. Feng, G. Chen, Y. Yang, X. Cui, Q. Wang, J. Cao *et al.*, "Paicore: a 1.9-million-neuron 5.181-tsops/w digital neuromorphic processor with unified snn-ann and on-chip learning paradigm," *IEEE Journal of Solid-State Circuits*, 2024.

Minor Comment 5: ***Fig 3b, 3d, and similar: all axes need to be labeled***

Response: We have checked all similar cases and added labels. The original Fig. 3d is now in Supplementary Fig. S9.

Revision: Added labels in figure axes.

Minor Comment 6: ***From the HCP dataset, how are subjects elected? The population appears smaller than the pool of subjects for whom data is available in this large-scale dataset.***

Response: As you pointed out, a larger sample size would be more appropriate if the goal is to explore universal characteristics of a population. However, the primary objective of using the data in our study is to propose a dynamics-aware quantization framewor and to verify that this framework can enable accurate low-precision simulation of brain dynamics models while maintaining their dynamic characteristics. In this context, our focus is on validating the effectiveness and accuracy of the framework, rather than conducting extensive statistical analyses of population features.

To address the concern regarding the sample size and generalization, we have presented the key inversion results using the larger HCP 1004 dataset in the response to your Major Comment 7. For your convenience, we copy the results here. The results in Table R6 demonstrate that our proposed pipeline can successfully identify solutions that significantly exceed the baseline (correlation between SC and FC). The FC correlations are not as high as those reported in [1] because we employ a distinct model with fewer parameters.

Revision: Added explanations in the revised part *Multimodal neuroimaging dataset and data processing* of the *Methods* section: "We utilize T1w MRI, dMRI, and rs-fMRI from 45 subjects from the Human Connectome Project (HCP) Retest data release [2]. The subject population appears to be relatively small as the primary objective of using the data in our study is to validate the effectiveness and accuracy of the framework, rather than conducting extensive statistical analyses of population features."

We also added the inversion results of the larger HCP 1004 dataset in Supplementary Table S7 (corresponding to Table R6).

References:

Table R6. Model inversion results using different algorithms with different precisions. Each row represents an algorithm-precision combination used during the parameter search process. We perform both high-precision and low-precision simulations using the searched parameter sets repeatedly and calculated the corresponding FC. Each column shows the correlation coefficients between pairs of high-precision simulated FC, low-precision simulated FC, and empirical FC. All correlation results are statistically significant with $p < 10^{-4}$. As a baseline, the correlation between the empirical SC and FC is 0.4759.

	FC corr. (high precis. and emp.)	FC corr. (low precis. and emp.)	FC corr. (high precis. and low precis.)
PSO (high precision)	0.5630 ($\pm$0.0085)	0.5612 ($\pm$0.0095)	0.9964 ($\pm$0.0005)
PSO (low precision)	0.5620 ($\pm$0.0107)	0.5713 ($\pm$0.0110)	0.9814 ($\pm$0.0018)
CMA-ES (high precision)	0.5618 ($\pm$0.0079)	0.5593 ($\pm$0.0069)	0.9952 ($\pm$0.0006)
CMA-ES (low precision)	0.5615 ($\pm$0.0149)	0.5660 ($\pm$0.0145)	0.9907 ($\pm$0.0012)

[1] S. Zhang, B. Larsen, V. J. Sydnor, T. Zeng, L. An, X. Yan, R. Kong, X. Kong, R. C. Gur, R. E. Gur *et al.*, "In vivo whole-cortex marker of excitation-inhibition ratio indexes cortical maturation and cognitive ability in youth," *Proceedings of the National Academy of Sciences*, vol. 121, no. 23, p. e2318641121, 2024.

[2] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium *et al.*, "The wu-minn human connectome project: an overview," *Neuroimage*, vol. 80, pp. 62–79, 2013.

Minor Comment 7: ***Fig S3: what is the interpretation behind higher inferred G in low-precision than high-precision models?***

Response: Thank you for your careful observation. Indeed, the inferred G corresponding to the maximum value of the curve from the low-precision model in Supplementary Fig. S3 is higher than that from the high-precision model. Your question prompted us to carefully examine this phenomenon. Through comparative experiments, we attempted to analyze the impact of quantization of different variables on the inferred G.

We believe that the deviation in inferred G values may be attributed to the following factors:

First, the quantization of the structural connectivity (SC) matrix is likely the primary cause of G deviation. Quantization may slightly alter the structure of SC. For instance, empirical SC matrices are typically not normally distributed but have many values concentrated around zero with a few very large values. In such cases, limited-range integers struggle to simultaneously represent both weak and strong connections, making it inevitable that the structure of quantized SC differs from that of high-precision SC. Since G can be viewed as a global scaling coefficient for SC, two slightly different SC matrices would likely require slightly different G to achieve optimal fitting performance. As shown in Fig. R6, when only SC is quantized (the yellow line), the inferred G is higher than that of the high-precision model, even slightly higher than the low-precision model with all variables quantized. This suggests that SC quantization is likely the primary cause of the G deviation observed in Fig S3. This bias could be addressed through several potential solutions: first, improving the quantization precision of SC as much as hardware permits; second, compensating for the inferred G based on changes in SC before and after quantization, for example, when the norm of SC decreases after quantization, the inferred G may tend to

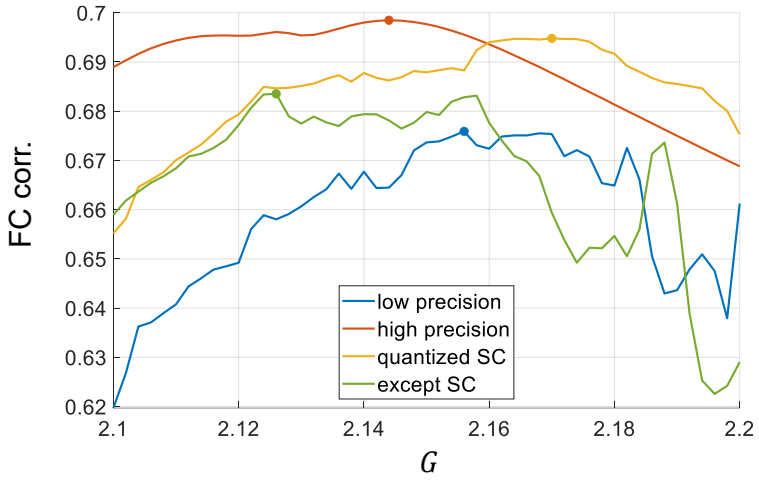

Fig. R6. Curves of FC correlation with respect to G from different quantization configurations. The values of w and I_0 correspond to those in Fig. S3. We reduce the step size of G and set the range around the optimal point to show the curves more clearly.

be larger.

Second, the quantization of state variables also impacts on the inferred G. When we quantize all state variables except SC (green curve in Fig. R6), the trend of FC correlation with respect to G is similar to that of high precision, but exhibits noticeable fluctuations. These fluctuations, compared to the curve shift caused by SC quantization, may introduce more random effects on the inferred G.

From the complete search range (e.g., G=1 to 5), the difference between the inferred G values from low-precision quantization and high-precision models is relatively small (approximately 0.02). If we aim to search for a parameter G that is applicable to the high-precision model, we can first use the low-precision model to rapidly search across the entire parameter space, then employ the high-precision model for local optimization around the search results.

This is a very subtle and interesting phenomenon caused by quantization. The analysis of this phenomenon provides us with clearer insights into the effects of low-precision quantization on model parameter inference. Thank you again for your meticulous observation.

Revision: Pointed out the difference in the inferred G in the caption of Supplementary Fig. S3: "The high-precision model and the low-precision model with group-wise quantization demonstrate nearly identical extreme positions (i.e., the G values corresponding to the maximum indicator values). While we observe slight deviations in the inferred optimal G values from models with different precisions, we provide detailed analyses and discussions of this phenomenon in Section S1.12."

Pointed out the difference in the inferred G in the revised part *Consistency of goodness-of-fit indicators between low- and high-precision models* of *Results* section: "Similar to what we observe in the three-dimensional distribution, the static and dynamic indicators obtained using models of different precision levels show the same trend as the parameter G changes, and the positions of the optimal points are almost consistent. It is worth noting that the inferred optimal G value from the low-precision model is slightly larger than that from the high-precision model. We provide detailed analyses and discussions of this phenomenon in the Supplementary Fig. S12."

Added more detailed analyses and discussions in S1.12 in Supplementary Information.

Added descriptions and discussions of the phenomenon in the revised *Discussion* section: In addition, we reveal the underlying mechanisms that determine how quantization affects the consistency of models with different precisions. First, under certain model parameters, variables associated with nonlinear transformations may exhibit significantly wider value ranges, affecting the selection of quantization parameters and increasing the rounding errors in low-precision models, thus degrading quantization performance. While this effect does not significantly impact the validity of the low-precision models in this work, it might limit low-precision implementation of other dynamic systems. An interesting topic in future work is to further explore this effect and seek possible solutions. Second, the quantization of variables, especially the SC, may introduce a slight bias in the inferred optimal parameters. This bias might be addressed by either improving the quantization precision of SC as much as hardware permits, or by compensating for the inferred parameters based on changes in SC before and after quantization.

Minor Comment 8: ***To my knowledge, no code has been made available yet. The manuscripts mentions that code will be made available, and a link is provided to an empty Github repository.***

Response: Thank you for your concern about code availability. We submitted the code (as a compressed archive) along with the manuscript during our initial submission. We have now incorporated the code for the new experiments included in this revision and uploaded the updated compressed archive. Please find the revised code in the supplementary materials.

Regarding the GitHub repository, since the paper has not yet been published, we are currently providing the code only to the editors and reviewers. Once the paper is accepted and published, we will make the code publicly available through GitHub for all readers.

Revision: Uploaded new codes.

3. Response to Reviewer 3

Overall Comment: ***This manuscript presents a hybrid hardware-software framework for accelerating the inversion of coarse-grained whole-brain models. By introducing a dynamics-aware quantization strategy and deploying simulations on neuromorphic (TianjicX) and GPU architectures, the authors demonstrate considerable speedups (up to ×400) while maintaining key features of the underlying dynamical system.***

The manuscript addresses a critical challenge in computational neuroscience — the acceleration of macroscopic brain model inversion. The proposed pipeline directly tackles this issue with a quantization framework, including semi-dynamic quantization, group-wise range adaptation, and multi-timescale integration. The implementation is both scalable and practical, leveraging TianjicX neuromorphic hardware and GPU parallelism to support large-scale parameter exploration through strong system-level integration. Notably, the work introduces a platform-enabling insight by demonstrating an unconventional yet efficient use of neuromorphic hardware (TianjicX) for non-spiking, coarse-grained dynamical modeling — thereby expanding the potential applications of such architectures beyond traditional neuron-level simulations. The manuscript could be further improved by addressing the following points.

Response: We sincerely appreciate you for the positive feedback on our motivation and performance benefits. For the concerns about the language, illustrations, and the interpretation of results, we have tried our best to address them in the revised manuscript. Details can be found in the following responses.

Comment 1: ***Generalization. The method is tested exclusively on the dynamic mean-field (DMF) model, which is smooth, low-dimensional, and relatively well-behaved. The authors do not demonstrate generalizability to models with richer or more complex dynamics (e.g., oscillatory or chaotic regimes). To broaden the impact, the authors should either include another model class or more explicitly discuss the limitations of applying the approach beyond DMF (see integration or oscillatory behavior, for example, in: 10.1201/9781003143499).***

Response: This concern is crucial for demonstrating the broader applicability of the proposed methods beyond the DMF model. Our answers are detailed below.

1. Rationale for DMF model selection: Our choice of the DMF model as the primary case study is based on several considerations. First, the DMF model represents one of the most widely adopted models in whole-brain dynamics research, with numerous influential studies utilizing DMF or its variants for whole-brain network modeling and empirical data fitting [1–3]. Second, the DMF model is theoretically grounded, being derived from microscopic spiking neuron models [4], which enables it to bridge dynamical processes across different scales within the brain. This theoretical foundation distinguishes it from more phenomenological models such as the Hopf or Kuramoto model.

It is important to note that despite the differences in the underlying equations, the simulation procedures for various models when applied to resting-state brain modeling are remarkably similar. Models typically begin simulation from randomly initialized states, followed by an initial transient regime. Analysis is then conducted on the time series with the initial transient period removed after the signal behavior reaches a relatively stable state. For example, in the Hopf model, during resting-state conditions, the oscillations are relatively stable with amplitude fluctuations that remain within bounded ranges. From this perspective, the DMF model is not a special case but rather representative of the common simulation workflow employed across different whole-brain dynamics models.

2. Demonstration of generalizability: To address your concern about generalizability, we have con-

ducted additional experiments using the Hopf model [5, 6], which exhibits different characteristics compared to the DMF model and addresses potential representativeness limitations from two key perspectives.

First, from a modeling foundation perspective, the Hopf model represents a phenomenological approach to whole-brain dynamics, in contrast to the DMF model which is a biophysically grounded model derived from microscopic spiking neuron networks. Second, from a dynamical behavior perspective, the Hopf model under certain parameter regimes exhibits oscillatory dynamics mentioned in [7], fundamentally different from the DMF model which converges to fluctuations around a fixed point attractor. This difference allows us to validate our approach across distinct dynamical regimes.

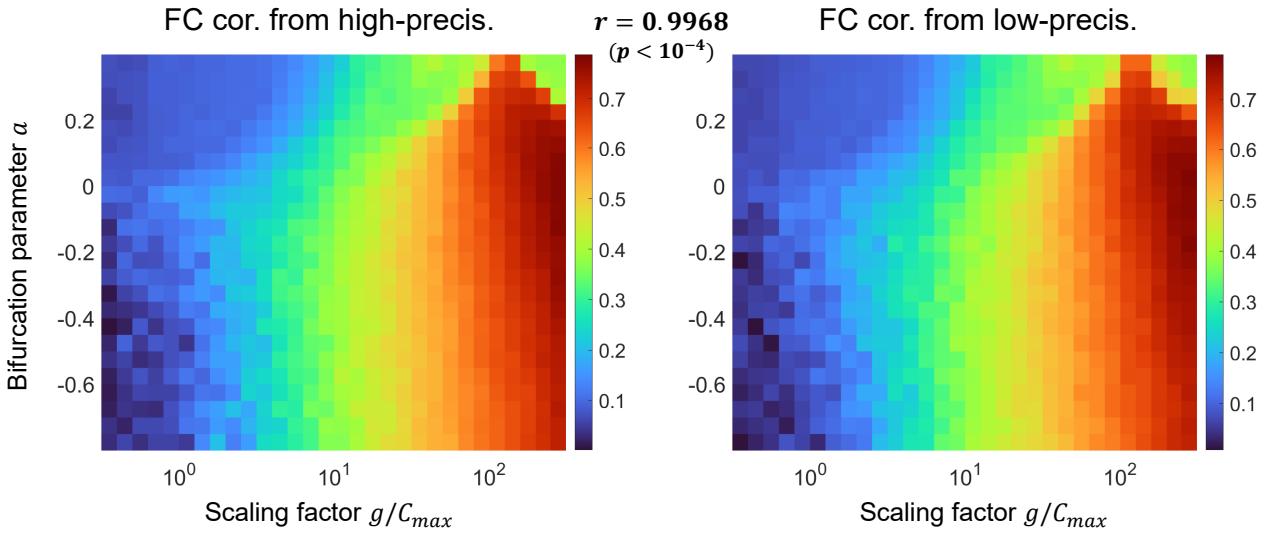

Fig. R7. Hopf model FC correlation across the parameter space. Correlation coefficients between empirical and simulated FC results from the high-precision model (left) and the low-precision model (right). Parameters a and g represent the bifurcation parameter and global connectivity scaling, respectively, while C_{max} denotes the maximum value of the SC matrix.

We validated the low-precision simulation of the Hopf model. Since the Hopf model we employed has only two parameters (the bifurcation parameter a and the global scaling g of the connectivity), we directly computed the correlation between simulated and empirical FC results across the two-dimensional parameter space, and compared the FC correlation distributions from high-precision and low-precision models. Figure R7 demonstrates that the quantization approach can maintain strong correspondence with high-precision simulations for oscillatory dynamics. Specifically, the correlation between high-precision and low-precision FC correlations across the parameter space can reach $r = 0.9968$ ($p < 10^{-4}$), indicating that the low-precision model can effectively capture the parameter-dependent FC patterns observed in the high-precision model.

Our experimental results also demonstrate the acceleration potential of the Hopf model simulation on advanced computing architectures. We conducted experiments on the Hopf model with 68, 148, and 512 nodes, and compared the performance of CPU, GPU and Brain-inspired computing chip implementations. The results are summarized in Table R7. Unlike the DMF model, the Hopf model requires fewer time steps for simulation, resulting in significantly longer simulation time on both CPU and GPU platforms, with simulation time accounting for a higher proportion of the total inversion time. Additionally, the warm-up phase in the brain-inspired computing chip pipeline also consumes more time. Despite these challenges, the brain-inspired computing chip pipeline still demonstrates significant acceleration compared to the CPU implementation, achieving speedup factors of up to 90.3× for simulation time and 30.4× for overall

inversion time in the 512-node case.

Table R7. Computational performance of Hopf modeling. $T_{overall\_inversion}$ is the total time consumed during the entire model inversion process, while $T_{simulation}$ represents the model simulation time. For the GPU and the brain-inspired computing chip, the $T_{simulation}$ results include both computation time and intra-device data transfer time. The remaining time in the entire process is used for data transfer between host and device, metric calculation, search algorithm iterations, memory management, and other workloads.

#Node	Platform	$T_{simulation}$	$T_{overall\_inversion}$
68	CPU	4185.04	4188.03
	Brain-inspired (ours)	**169.56 (24.7×)**	**361.39 (11.6×)**
148	CPU	13960.00	13973.20
	Brain-inspired (ours)	**359.53 (38.8×)**	**896.67 (15.6×)**
512†	CPU	217477.84	221446.53
	Brain-inspired (ours)	**2409.57 (90.3×)**	**7286.81 (30.4×)**

* All time units in the table are in seconds. Numbers in the parentheses represent speedup factors compared to CPU execution.
† Measured and estimated using synthetic data.

3. Scope and limitations of the proposed methods: The core principle underlying our quantization framework is that the effectiveness depends on the numerical range stability of variables rather than specific dynamical behaviors. As long as the ranges of state variables and their derivatives remain relatively stable over time, semi-dynamic quantization can be applied effectively. We believe that most macroscopic brain models under resting-state conditions can satisfy this requirement. However, we also acknowledge that particularly complex dynamical processes may be difficult to represent using low-precision data and computation (we also discuss this limitation in detail in the following response to Comment 3).

In addition, since most brain-inspired computing chips are specialized processors, the computational operators supported by existing neuromorphic chips (including the TianjicX chip we used) are inherently limited and may not be capable of implementing arbitrary brain dynamics computational process.

We have incorporated these important limitations into the revised Discussion section, and they represent substantial opportunities for future optimization and improvement in both hardware design and algorithm development.

Revision: Added descriptions of the DMF model's representativeness and validation of the method's generalization with the Hopf model in the revised part *Dynamics-aware quantization for macroscopic brain models* of the *Results* section: "...We specifically focus on the simulation of the DMF model and the hemodynamic model for the resting-state brain. The choice of the DMF model as the primary case study is based on its widespread use in whole-brain dynamics research [1–3] and its biophysically grounded foundation derived from microscopic spiking neuron models [4]. To further validate the generalization of our dynamics-aware quantization framework, we also conduct experiments over the Hopf model [5, 6], which represents a phenomenological approach with different oscillatory dynamics compared to fixed-point attractor behavior of the DMF model."

Added the low-precision simulation results of the Hopf model in Supplementary Fig. S10 (as shown in Fig. R7) and the acceleration results in Supplementary Table S6 (corresponding to Table R7). Added a new part *Generalization of the proposed pipeline to other brain dynamics models* in the *Methods* section, which includes the following contents: "Different whole-brain models employ distinct differential equations to describe brain regions, and the signals may exhibit different characteristics [8]. The DMF model we

primarily focus on evolves from random initial states to fixed points during simulation and exhibits random fluctuations around these fixed points driven by noise. In contrast, oscillator-based models such as Hopf and Kuramoto models demonstrate pronounced periodic oscillation over time, with variables showing regular temporal patterns rather than the stochastic fluctuations observed in the DMF model. Therefore, to demonstrate the generalization of our proposed pipeline, we conduct additional validation experiments over the Hopf model [5, 6].

The whole-brain dynamics of the Hopf model can be described by

$$\frac{dz_j}{dt} = (a + i\omega_j)z_j - |z_j|^2 z_j + g\sum_{k=1}^{N} C_{jk}(z_k - z_j) + \eta_j, \tag{R2}$$

where $z_j = x_j + iy_j$ is the complex state variable of the brain region j, $|z_j|$ is the module of z_j, the parameter a is the bifurcation parameter which we make constant across nodes, ω_j is the intrinsic angular frequency, g is a global scaling of the connectivity C, and η_j is uncorrelated white noise. From the modeling foundation perspective, the Hopf model represents a phenomenological approach to whole-brain dynamics, in contrast to the DMF model which is a biophysically grounded model derived from microscopic spiking neuron networks. From the dynamical behavior perspective, the Hopf model under certain parameter regimes exhibits oscillatory dynamics mentioned in [7], fundamentally different from the DMF model which converges to fluctuations around a fixed-point attractor. The difference allows us to validate our approach across distinct dynamical regimes.

We first validate the low-precision simulation of the Hopf model. Since the Hopf model we employ has only two parameters (the bifurcation parameter a and the global scaling parameter g), we directly computed the correlation between simulated and empirical FC results across the two-dimensional parameter space, and compared the correlation distributions of high-precision and low-precision models. Supplementary Fig. S10 demonstrates that the quantization approach can maintain strong correspondence with high-precision simulations for oscillatory dynamics. Specifically, the correlation between high-precision and low-precision FC results across the parameter space can reach $r = 0.9968$ ($p < 10^{-4}$), indicating that the low-precision model can effectively capture the parameter-dependent FC patterns observed in the high-precision model.

We also validate the acceleration potential of the Hopf model simulation on the advanced computing architectures. We conduct experiments over the Hopf model with 68, 148, and 512 nodes, and compare the performance of CPU and brain-inspired computing chip implementations. The results are summarized in Supplementary Table S6. Unlike the DMF model, the Hopf model requires smaller time steps for simulation, resulting in significantly longer simulation times on CPU, with simulation time accounting for a higher proportion of the total inversion time. Additionally, the warm-up phase in the brain-inspired computing chip pipeline also consumes more time. Despite these challenges, the brain-inspired computing chip pipeline still demonstrates significant acceleration compared to the CPU implementation, achieving speedup factors of up to 90.3× for simulation time and 30.4× for overall inversion time in the 512-node case."

Added discussions regarding the limitations of the quantization approaches in the revised *Discussion* section: "...allowing broader scientific computing tasks to benefit from advanced computing architectures. However, the proposed method also suffers inherent limitations: When a dynamical model exhibits frequent or unpredictable large-magnitude fluctuations in its numerical range, the semi-dynamic quantization would likely become unsuitable. There exists a fundamental trade-off between quantization precision and generality: higher quantization precision enables support for a broader variety of dynamical models with better simulation fidelity, but may result in reduced computational efficiency. Future work should explore adaptive quantization schemes that can dynamically balance this trade-off based on the specific characteristics of the target dynamic system and the available computational resources. Additionally, the computational operators

supported by existing brain-inspired computing chips are limited and may not be capable of implementing arbitrary dynamic systems. These constraints highlight the need for future research in developing more flexible quantization strategies and expanding the computational capabilities of brain-inspired computing architectures to accommodate a broader range of scientific computing applications."

References:

[1] G. Deco, J. Cruzat, and M. L. Kringelbach, "Brain songs framework used for discovering the relevant timescale of the human brain," *Nature communications*, vol. 10, no. 1, p. 583, 2019.

[2] M. Demirtaş, J. B. Burt, M. Helmer, J. L. Ji, B. D. Adkinson, M. F. Glasser, D. C. Van Essen, S. N. Sotiropoulos, A. Anticevic, and J. D. Murray, "Hierarchical heterogeneity across human cortex shapes large-scale neural dynamics," *Neuron*, vol. 101, no. 6, pp. 1181–1194, 2019.

[3] P. Wang, R. Kong, X. Kong, R. Liégeois, C. Orban, G. Deco, M. P. Van Den Heuvel, and B. Thomas Yeo, "Inversion of a large-scale circuit model reveals a cortical hierarchy in the dynamic resting human brain," *Science advances*, vol. 5, no. 1, p. eaat7854, 2019.

[4] K.-F. Wong and X.-J. Wang, "A recurrent network mechanism of time integration in perceptual decisions," *Journal of Neuroscience*, vol. 26, no. 4, pp. 1314–1328, 2006.

[5] G. Deco, M. L. Kringelbach, V. K. Jirsa, and P. Ritter, "The dynamics of resting fluctuations in the brain: metastability and its dynamical cortical core," *Scientific reports*, vol. 7, no. 1, p. 3095, 2017.

[6] A. Ponce-Alvarez and G. Deco, "The hopf whole-brain model and its linear approximation," *Scientific reports*, vol. 14, no. 1, p. 2615, 2024.

[7] E. E. Tsur, *Neuromorphic Engineering: The Scientist's, Algorithms Designer's and Computer Architect's Perspectives on Brain-Inspired Computing*. CRC Press, 2021.

[8] J. Cabral, M. L. Kringelbach, and G. Deco, "Functional connectivity dynamically evolves on multiple time-scales over a static structural connectome: Models and mechanisms," *NeuroImage*, vol. 160, pp. 84–96, 2017.

Comment 2: ***While the quantization pipeline is well-designed, the remaining components (e.g., simulation parallelism, GPU mapping) reflect well-established practices. The manuscript contributes more as a system integration effort than a source of novel algorithmic or neuroscientific insights. The authors should highlight this point.***

Response: Thank you for this insightful comment. Our work aims to explore the acceleration of macroscopic brain dynamics modeling using brain-inspired computing architectures alongside GPUs. Both low-precision quantization and simulation parallelism are indispensable components of our pipeline. While both components offer novel contributions, we acknowledge that they differ significantly in nature and complexity. The low-precision quantization, as a novel computational methodology, requires algorithmic exploration and innovation. We have conducted detailed experimental analyses of the phenomena and challenges arising from model quantization, revealing interesting insights and validating the effectiveness through extensive experiments.

In contrast, the parallel resource mapping and hardware deployment, as you correctly identified, is a more engineering-oriented problem that demands systematic design rather than scientific innovation. We have comprehensively considered the characteristics and requirements of models, algorithms, and hardware to design the mapping strategy for macroscopic brain dynamics models. We have optimized the entire pipeline across heterogeneous platforms and provided complete evaluation of computational overhead on different platforms. The contribution of this component lies not in algorithmic or neuroscientific advances, but in designing a novel hardware system. In addition, our work also provides new perspectives and guidance for future design of neuromorphic hardware. Following your suggestion, we have highlighted the distinction between these components in the revised Introduction and Discussion sections.

Revision: Highlighted the distinction of the two contributions in the revised *Introduction* section: " ...we propose a comprehensive pipeline enabling macroscopic brain dynamics models to benefit from modern computing architectures. Our pipeline addresses the challenges from two complementary perspectives. From the computational methodology perspective, we analyze...

In this way, we enable the majority of the model simulation process to be deployed on low-precision platforms. From the system engineering perspective, we propose... "

In the revised *Discussion* section: " We primarily address two key challenges. The first is implementing macroscopic brain dynamics on specialized computing architectures like brain-inspired computing chips, which offer high efficiency but low precision. This challenge needs computational algorithmic innovation.

...The second challenge is to fully utilize the highly parallel computing and memory resources of brain-inspired computing chips and GPUs to accelerate the costly model inversion process, which requires the development of new systems that leverage heterogeneous computing platforms along with considerable engineering optimization efforts. "

Comment 3: ***A critical limitation of the framework is the requirement for a warm-up phase using high-precision floating-point simulation (typically on a CPU) to determine the quantization parameters. This step breaks the standalone nature of the system and may reduce its deployability in embedded or edge applications. Moreover, it assumes convergence to a relatively stable state during warm-up — a condition that may not hold for more volatile brain dynamics. This trade-off between quantization precision and model generality deserves a clearer treatment in the manuscript.***

Response: We appreciate your insightful observation regarding the limitations of our semi-dynamic quantization framework. This comment touches on two fundamental aspects of our approach that deserve detailed clarification.

Regarding the warm-up phase requirement: Most current macroscopic whole-brain models are developed to fit resting-state MRI data. Even without employing low-precision quantization methods, the standard simulation procedure in prior studies typically begins from an arbitrarily specified initial state. After a warm-up period (e.g., a fixed duration of several seconds), when the model's signal activities reach a largely stable state, the signal characteristics are analyzed and compared with empirical resting-state data. Therefore, the warm-up phase is not specific to low-precision simulation. Our quantization method builds upon the established procedure by adding a quantization parameter selection phase during the warm-up phase and implementing stable simulation using low-precision computation.

Regarding the system deployability: First, we recognize that the integration of general-purpose processors within the system is currently unavoidable. The warm-up phase is essentially also a simulation process. If the accelerators can support high-precision computation, this step may not necessarily need to be completed on the host. Therefore, mixed precision computing could be a valuable consideration for future development of brain-inspired computing architectures. We had discussed this possibility in the Discussion section of the original manuscript: "...existing brain-inspired computing chips typically support only 8-bit or lower precision data types, with limited mixed-precision or dynamic quantization capabilities. This constrains the selection of quantization methods and degrades the effectiveness of low-precision models, while simultaneously increasing the workloads on the host and compromising the overall performance. The inversion time could be reduced by either adding high-precision computation paths within cores or incorporating some high-precision computing cores on the chip." However, aside from the warm-up phase, other essential pipeline components, including search algorithms and data I/O operations, are also currently all handled by CPUs. Contemporary advanced computing architectures, including both

GPUs and brain-inspired computing chips, are fundamentally positioned as specialized accelerators or co-processors designed to perform specific tasks. In most complex applications and deployment scenarios, these architectures cannot independently satisfy all functional requirements and cannot operate completely standalone without CPUs or FPGAs unless we design specialized circuits for each specific scenario. We believe this is a universal limitation of current computing systems.

The most promising approaches to improve the system integration under this constraint are System-on-Chip (SoC) or System-in-Package (SiP) designs, which can maintain the entire system within a single chip or package. These designs not only reduce the physical size but also potentially minimize the data transfer between the host and accelerators. Such integrated designs could significantly enhance the deployability of the proposed pipeline while preserving the computational advantages of specialized accelerators. We have discussed the above limitation and the possible solutions in the revised Discussion section.

Second, beyond the system integration approaches, we wish to further clarify the impact of CPU dependency on overall system performance. As mentioned above, the pipeline relies on CPU for several steps, including the warm-up phase. It accelerates the most computationally intensive simulation of the original modeling workflow using GPU or the brain-inspired computing chip. From the results, this partial acceleration approach has already achieved substantial performance improvements, demonstrating that the cost of the high-precision warm-up stage is acceptable. However, we acknowledge that the warm-up process represents a part of bottleneck in the system, particularly when dealing with large-scale algorithms or models (see Fig. 5a in the manuscript).

Regarding the assumption of convergence to stable states: We wish to define the applicability and constraints of semi-dynamic quantization more explicitly. The robustness of the low-precision quantization is determined by the **numerical range of variables** rather than the specific dynamics behaviors. Low-precision data types have inherently limited the representational range and precision. Therefore, as long as the ranges of the variables and their derivatives of the model do not undergo substantial changes over time, low-precision quantization can remain effective. The DMF model converges to a fixed point with variables fluctuating within a bounded range, thus satisfying this requirement. However, this does not imply that only models converging to fluctuations around a fixed point can utilize the proposed semi-dynamic quantization method. For instance, in oscillator-based models such as the Kuramoto or Hopf model, state variables may oscillate over a wider range. Nevertheless, as long as their ranges remain relatively stable, semi-dynamic quantization can still be applicable. Previous literature empirically indicates that the resting-state FC does not exhibit extremely drastic transient changes (although some time-varying information might be extractable) [1]. Therefore, we believe that semi-dynamic quantization should be capable of meeting the computational demands of resting-state modeling of brain dynamics.

Looking toward future extension, we can envision scenarios where the ranges of variables change significantly during model execution. If such changes are predictable, semi-dynamic quantization could still be employed. For instance, when simulating brain lesions [2], perturbations are typically introduced to a baseline resting-state brain model. These perturbations can cause the model to transition from one stable state to another. Since the timing of these perturbations is known as a priori, high-precision floating-point numbers could be used for a new warm-up phase, while a new low-precision quantization scheme could be employed once another stable state is reached. By iteratively applying the warm-up and simulation stages, the semi-dynamic quantization method could be potentially extended to more complex scenarios. While these scenarios are theoretically plausible, they go beyond the scope of the current paper; therefore, we just discuss them in the revised Discussion section.

We also acknowledge that the semi-dynamic quantization method has inherent limitations. When a dynamical model exhibits frequent or unpredictable large-magnitude numerical fluctuations, this method

would likely become unsuitable. To our knowledge, existing macroscopic brain dynamics models do not exhibit such characteristics under typical simulation conditions, but we have explicitly mentioned this limitation in the revised manuscript.

Revision: Explicitly pointed out the bottleneck effect of host time in the revised part *Scalability of model inversion on different architectures* of the *Results* section: "We can predict that the host time would become the dominant bottleneck in the brain-inspired computing chip pipeline as the model size increases. This highlights a significant opportunity for future optimization. Performance might be further improved through more efficient quantization algorithms, or by offloading portions of the warm-up process to GPU (like what we do in Supplementary Table S4) or other specific hardware accelerators."

Pointed out the bottleneck effect of host time in the revised *Discussion* section: "..., while simultaneously increasing the workloads on the host and compromising the overall performance. Our analysis indicates that the host-side warm-up process becomes a bottleneck for the entire pipeline as the model size increases. While using GPU for warm-up can effectively alleviate this bottleneck, it inevitably increases system complexity. We believe the inversion time could be more efficiently reduced by either adding..."

Added discussion about the deployability limitation and possible solutions in the revised *Discussion* section: "Third, aside from the warm-up phase, other essential pipeline components, including search algorithms and data I/O operations, are all handled by the host. In most complex applications and deployment scenarios, these advanced computing architectures cannot independently satisfy all functional requirements and cannot operate completely standalone without CPUs or FPGAs unless we design specialized circuits for each specific scenario. We believe this is a universal limitation of current computing systems. One of the most promising approaches to improve the system integration under this constraint are System-on-Chip (SoC) or System-in-Package (SiP) designs. These designs not only reduce the physical size but also potentially minimize the data transfer between the host and the accelerators, thus significantly enhancing the deployability of the proposed pipeline. Last, we deploy model inversion..."

Provided detailed explanations regarding the prevalence of transient and stable regimes in the part *Dynamics-aware quantization for macroscopic brain models* of the *Results* section: "We introduce semi-dynamic quantization to address this problem. Most current macroscopic whole-brain models are developed based on the resting-state data. The simulation of the brain's resting state activity can be divided into two stages. In the warm-up stage, the variables may start from an arbitrary state and experience a transient period with significant changes before converging to a relatively stable state. In the simulation stage, the model typically runs in a relatively stable state where the variables fluctuate randomly or oscillate within a certain range. These stages represent a common practice in the model simulation scenario, regardless of which specific model is employed."

Clarified the scope of applicability of the proposed method in the part *Dynamics-aware quantization for macroscopic brain models* of the *Results* section: "The applicability of semi-dynamic quantization depends on the numerical range of variables. As long as the ranges of variables and their derivatives of the model do not undergo substantial changes over time, low-precision quantization can remain effective. Previous literature empirically indicates that the resting-state FC does not exhibit extremely drastic transient changes (although some time-varying information might be extractable) [1]. Therefore, we believe that semi-dynamic quantization should be capable of meeting the computational demands of resting-state modeling of brain dynamics across various model types. It is also worth..."

Added discussions regarding the broader applicability of semi-dynamic quantization in potential scenarios in the revised part *Semi-dynamic quantization for brain dynamics* of the *Methods* section: "...In the subsequent simulation stage, all variables are represented as integers. We can envision scenarios where

the ranges of variables change significantly during model execution. If such changes are predictable, semi-dynamic quantization could still be employed. For instance, when simulating brain lesions [2], perturbations are typically introduced to a baseline resting-state brain model. These perturbations can cause the model to transfer from one stable state to another. Since the timing of these perturbations is known as a priori, high-precision floating-point numbers could be used for the transition phase, while a new low-precision quantization scheme could be employed once another stable state is reached. By iteratively applying the warm-up and simulation stages, the semi-dynamic quantization method could be potentially extended to more complex scenarios."

Added discussions regarding the limitations of the quantization approaches and the trade-off between quantization precision and generality in the revised *Discussion* section: "...allowing broader scientific computing tasks to benefit from advanced computing architectures. However, the proposed method also suffers inherent limitations: When a dynamical model exhibits frequent or unpredictable large-magnitude fluctuations in its numerical range, the semi-dynamic quantization would likely become unsuitable. There exists a fundamental trade-off between quantization precision and generality: higher quantization precision enables support for a broader variety of dynamical models with better simulation fidelity, but may result in reduced computational efficiency. Future work should explore adaptive quantization schemes that can dynamically balance this trade-off based on the specific characteristics of the target dynamic system and the available computational resources."

References:

[1] T. O. Laumann, A. Z. Snyder, A. Mitra, E. M. Gordon, C. Gratton, B. Adeyemo, A. W. Gilmore, S. M. Nelson, J. J. Berg, D. J. Greene *et al.*, "On the stability of bold fmri correlations," *Cerebral cortex*, vol. 27, no. 10, pp. 4719–4732, 2017.

[2] J. Wei, B. Wang, Y. Yang, Y. Niu, L. Yang, Y. Guo, and J. Xiang, "Effects of virtual lesions on temporal dynamics in cortical networks based on personalized dynamic models," *NeuroImage*, vol. 254, p. 119087, 2022.

Comment 4: ***The generalization of the proposed approach to other neuromorphic platforms, particularly, analog designs (e.g., 10.3390/app12094528). These architectures are primarily optimized for spiking neural network simulations and may not natively support the type of low-precision, non-spiking differential equation integration used in this work. Adapting the framework to these platforms would likely require substantial reengineering or conceptual reformulation. This should be discussed, particularly in light of the growing diversity in neuromorphic hardware design.***

Response: This is a critical consideration given the growing diversity in neuromorphic hardware architectures. Our detailed answers are as follows.

Common advantages of neuromorphic computing architectures: Our work preliminarily validates the feasibility of using brain-inspired computing chips for modeling macroscopic brain dynamics and demonstrates their performance advantages in this domain. These advantages likely stem from several key characteristics: (1) data locality enabled by tightly coupled memory and computation; (2) parallelism provided by highly distributed computing and storage resources; and (3) efficiency achieved through streamlined specialized circuits. Our proposed methods are designed to fully leverage these characteristics while satisfying the hardware constraints. Since these features represent common architectural advantages inspired by the brain [1], we believe neuromorphic computing architectures possess inherent advantages in this neuroscience application domain.

Functional and implementation differences across neuromorphic platforms: While architectural

commonalities exist, as you correctly note, significant differences also exist among neuromorphic platforms. Most neuromorphic chips adopt application-specific circuits. For instance, most existing analog neuromorphic circuits like [2], while efficiently supporting spiking neurons, may not directly support other types of dynamic models. Therefore, we acknowledge that not all existing neuromorphic circuits can be directly applied to macroscopic brain dynamics modeling. From the perspective of the Neural Engineering Framework's three fundamental principles [2–4], the primary difference between macroscopic brain models and neuronal models lies in "Representation": spiking signals of neurons differ from non-spiking signals of brain regions. Fortunately, the "Transformation" and "Dynamics" of the two kinds of models appear to have no fundamental differences, though specific implementations may require adjustments.

For other digital neuromorphic chips, the proposed methods can be readily adapted as long as their circuits support the non-spiking data types and computational operators required for whole-brain models. However, adapting our framework to analog circuits presents greater complexity, potentially requiring adjustments from both hardware and algorithmic perspectives.

At the hardware level, the proposed parallel mapping methods may not require substantial modifications due to architectural similarities, but the computational circuits need to make the following adjustments. First, circuits originally designed for implementing neurons may need to be adapted according to the differential equations of brain dynamics to support non-spike voltage/current representation, or alternatively, specialized spike encoding/decoding methods could be designed to represent the real-valued signals of brain regions using spike trains. Second, weighted synaptic connections used for implementing "Transformation" may need to support non-spike inputs or computation for specialized spike representation. Third, "Dynamics" require implementing more complex recurrent mechanisms than typical neural networks.

At the algorithmic level, the low-precision quantization needs to be adjusted according to specific circuit implementations: if continuous voltage/current values are implemented using analog circuits, the discrete integer constraints in the quantization method can be replaced with noise constraints; if spike trains are used to encode brain region signals, methods similar to the proposed integer quantization may be applicable.

In summary, applying the proposed framework to neuromorphic platforms designed for spiking neural networks presents challenges but can be achieved through algorithm-hardware co-design. The proposed quantization methods can be transferred to other digital platforms supporting non-spiking data, but may not be directly applicable to analog circuit platforms. We have added a discussion on these limitations in the revised manuscript, as we believe they represent important opportunities for future research and development.

Future trends and opportunities: The TianjicX chip [5] used in our work is a hybrid brain-inspired computing chip that can simultaneously support both spiking and non-spiking signals while providing high configurability. Many current neuromorphic platforms adopt similar designs, including Loihi 2 [6], SpiNNaker 2 [7], BrainScaleS 2 [8], and PAICORE [9]. Hybrid integration and configurability represent current trends in neuromorphic computing [1]. Future developments may see more mixed-signal neuromorphic chips, or configurable chips integrating specialized neuromorphic architectures with general-purpose processors. We believe that the methods can be relatively easily adapted to these heterogeneous platforms, so we are optimistic about the future generalization of the proposed framework.

Revision: Added a part *Generalization to diverse neuromorphic platforms* in the revised *Methods* section: "Brain-inspired computing chips are inherently advantageous for this neuroscience application due to common architectural features inspired by the brain like data locality, massive parallelism, and specialized circuit efficiency [1]. Our proposed methods are designed to leverage these characteristics, making

neuromorphic platforms a natural fit for modeling macroscopic brain dynamics.

However, significant differences between platforms pose adaptation challenges. Many neuromorphic platforms, especially analog designs, are highly optimized for the efficient simulation of spiking neurons [2]; however, the specialization makes them unable to directly support the non-spiking models we use. Drawing from the Neural Engineering Framework [2–4], the primary hurdle is "Representation", i.e., the whole-brain models use continuous signals, not spikes. The "Transformation" and "Dynamics" principles, conversely, are fundamentally similar and may require only implementation adjustments.

Adapting our framework to digital chips is straightforward once they support non-spiking data types. Analog circuits, in contrast, require deeper modifications at both the hardware and algorithmic levels. Hardware adjustments would involve re-purposing neuron circuits for non-spiking differential equations, or alternatively, designing spike-based encoding schemes. Synaptic circuits would also need to process these new signal types and support more complex recurrent dynamics. Algorithmically, quantization methods would need to be adapted for the target hardware, for instance by replacing discrete integer constraints with noise constraints for analog circuits.

Fortunately, the evolution of neuromorphic computing is likely to simplify this process. A trend towards hybrid, configurable architectures that support both spiking and non-spiking models is evident in platforms like the TianjicX chip [5], Loihi 2 [6], SpiNNaker 2 [7], BrainScaleS 2 [8], and PAICORE [9]. This shift toward mixed-signal and reconfigurable designs [1] will broaden the applicability of our pipeline, making future platforms increasingly suitable for macroscopic brain modeling."

References:

[1] D. Kudithipudi, C. Schuman, C. M. Vineyard, T. Pandit, C. Merkel, R. Kubendran, J. B. Aimone, G. Orchard, C. Mayr, R. Benosman *et al.*, "Neuromorphic computing at scale," *Nature*, vol. 637, no. 8047, pp. 801–812, 2025.

[2] A. Hazan and E. Ezra Tsur, "Neuromorphic neural engineering framework-inspired online continuous learning with analog circuitry," *Applied Sciences*, vol. 12, no. 9, p. 4528, 2022.

[3] C. Eliasmith and C. H. Anderson, *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. MIT press, 2003.

[4] E. E. Tsur, *Neuromorphic Engineering: The Scientist's, Algorithms Designer's and Computer Architect's Perspectives on Brain-Inspired Computing*. CRC Press, 2021.

[5] S. Ma, J. Pei, W. Zhang, G. Wang, D. Feng, F. Yu, C. Song, H. Qu, C. Ma, M. Lu *et al.*, "Neuromorphic computing chip with spatiotemporal elasticity for multi-intelligent-tasking robots," *Science Robotics*, vol. 7, no. 67, p. eabk2948, 2022.

[6] G. Orchard, E. P. Frady, D. B. D. Rubin, S. Sanborn, S. B. Shrestha, F. T. Sommer, and M. Davies, "Efficient neuromorphic signal processing with loihi 2," in *2021 IEEE Workshop on Signal Processing Systems (SiPS)*. IEEE, 2021, pp. 254–259.

[7] S. Höppner, Y. Yan, A. Dixius, S. Scholze, J. Partzsch, M. Stolba, F. Kelber, B. Vogginger, F. Neumärker, G. Ellguth *et al.*, "The spinnaker 2 processing element architecture for hybrid digital neuromorphic computing," *arXiv preprint arXiv:2103.08392*, 2021.

[8] C. Pehle, S. Billaudelle, B. Cramer, J. Kaiser, K. Schreiber, Y. Stradmann, J. Weis, A. Leibfried, E. Müller, and J. Schemmel, "The brainscales-2 accelerated neuromorphic system with hybrid plasticity," *Frontiers in Neuroscience*, vol. 16, p. 795876, 2022.

[9] Y. Zhong, Y. Kuang, K. Liu, Z. Wang, S. Feng, G. Chen, Y. Yang, X. Cui, Q. Wang, J. Cao *et al.*, "Paicore: a 1.9-million-neuron 5.181-tsops/w digital neuromorphic processor with unified snn-ann and on-chip learning paradigm," *IEEE Journal of Solid-State Circuits*, 2024.

Response Letter

Manuscript ID: NCOMMS-25-20318A
Title: Modeling Macroscopic Brain Dynamics with Brain-inspired Computing Architecture
Authors: Zhong Zheng, Jing Wei, Yaru Xu, Chenhua Li, Tianying Lu, Qili Guo, Xueyao Ji, Hao Guo, Gang Wang, Lei Deng

We would like to thank the reviewers for spending valuable time and raising insightful comments which we feel have substantially improved our manuscript. The point-by-point responses are provided as follows. All the revisions in the manuscript and Supplementary Information (SI) are marked in blue for your convenience.

1. Responses to Reviewer 1

Overall Comment: ***I have read the author responses to my comments and the changes made in the manuscript. All of my concerns were addressed satisfactorily. I'm satisfied with the generalisability of this method across models. The split of times is useful, and the speed-up from the GPU is promising. The additional correlation based quantitative results support the utility of the model.***

Response: We greatly appreciate your recognition of our contributions and insightful feedback. We are pleased that our previous responses and revisions can make you satisfied.

Comment 1: ***One thing the authors could optionally expand on is the analysis of the variation in the correlation with the scaling factor and bifurcation parameter in the Hopf model simulation. Does the values of scaling factors where the correlation is high also a useful regime?***

Response: Thank you for this insightful suggestion. The variation in the correlation with the scaling factor and bifurcation parameter is indeed an important issue that has been discussed in detail in previous work on Hopf models [1, 2]. In Fig. S10, our results are consistent with those previous findings, showing that the fitting performance is better (with higher correlation) when the scaling factor is large enough. Since our primary focus is on the similarity between results from high-precision and low-precision models rather than the specific indicator distributions, we did not provide detailed analyses in this aspect previously. Additionally, due to differences in the empirical data and the computational method for scaling factor normalization, the specific optimal regime we identify shows some variation from previous work.

Meanwhile, we did not distinguish between stable and unstable regions in the parameter space of the Hopf model to demonstrate the generalizability of our quantization method. Our results show that low-precision simulations maintain certain usability even with unstable parameters. However, such scenarios may have limited practical application value, which is why we did not elaborate on this aspect in the previous manuscript.

Revision: We have added the analyses of these results in the part *Generalization of the proposed pipeline to other brain dynamics models* of *Methods* Section: " ..., indicating that the low-precision model can effectively capture the parameter-dependent FC patterns observed in the high-precision model. Our results are also consistent with previous findings [1, 2], demonstrating that the fitting performance improves

(with higher correlation) when the scaling factor is sufficiently large. Additionally, due to differences in the empirical data and the computational method for scaling factor normalization, the specific optimal regime we identify shows some variation from previous work. Furthermore, we note that our analysis in Supplementary Fig. S10 does not distinguish between stable and unstable regions [2] within the parameter space. If we focus exclusively on the stable dynamical regions, the correlation between low-precision and high-precision simulation results would likely be even stronger. ”

References:

[1] Y. Sanz Perl, A. Escrichs, E. Tagliazucchi, M. L. Kringelbach, and G. Deco, “Strength-dependent perturbation of whole-brain model working in different regimes reveals the role of fluctuations in brain dynamics,” *PLOS Computational Biology*, vol. 18, no. 11, p. e1010662, 2022.

[2] A. Ponce-Alvarez and G. Deco, “The hopf whole-brain model and its linear approximation,” *Scientific reports*, vol. 14, no. 1, p. 2615, 2024.