nature portfolio

Peer Review File

Bridging known and unknown dynamics by machine learning inference from sparse observations

Corresponding Author: Professor Ying-Cheng Lai

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

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This manuscript uses the pre-training framework of transformer and also uses the reservoir computing to predict the long-term behavior s of complex dynamics where the data of the target system are incompletely collected. The idea of integrating the different frameworks of machine learning for rencontriucting dynamical systems only using limited information is interesting, which will certainly attracts attention from the community of complex dynamical systems and other related fields. The work deserves publication in NC; however, before the final acceptance, the following issues need to be substantially addressed.

1. This reviewer notices that the investigated dynamical system could be non-autonomous; however, such a configuration, though broadening the use of the current framework, brings many indeterministic issues. Since the pre-training procedures also need to include the non-autonomous systems as many as possible, it is unclear how to choose these non-autonomous systems as well as the time-variant input, and how to set the ratio of the data from the autonomous/non-autonomous systems. More justifications need to be done; or just focus the discussion merely on the typical dynamical systems, which are autonomous.

2. The long-term prediction of climate dynamics is used an illustrative example. However, such kind of discussion is also related crucially to the system which is either autonomous or non-autonomous. As we known, the climate system actually is not stationary but always time-variant. Then, how to provide the data for model pre-training is crucial to the prediction fidelity that one could achieve only using very limited observations. More discussion is anticipated.

3. In addition to the reservoir computing, is there any other neural networks that could be used in dynamics prediction. As this reviewer knows, at least the recurrent neural networks such as the neural ODEs, the KAN, and the diffusion models could be the potential candidates for this task. Thus, the authors are suggested to make some comparisons using different neural networks and highlight their respective advantages and weaknesses.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The manuscript introduces a hybrid machine-learning framework with transformers and reservoir computing to reconstruct and predict nonlinear dynamics from sparse observations. This issue occurs in scenarios where historical data from the target system is unavailable. To solve this problem, the developed hybrid architecture leverages transformers to infer missing dynamics from sparse inputs and reservoir computing to predict long-term attractors. By training on synthetic chaotic systems and testing on three distinct models, the chaotic food chain, Lorenz, and Lotka-Volterra systems, the authors demonstrate the effectiveness of the method under metrics of MSE and prediction stability. The reconstruction is accurate

under up to 80% sparsity and moderate noise levels.

Overall, the manuscript presents a potentially promising framework for dynamics reconstruction under extreme data limitations. However, although the present method can reconstruct the dynamics from a few observations in the given examples, to my understanding, there are fundamental limitations of the methodology that may prevent its practical use in more general and real examples. Considering that Nature Communications is highly selective and publishes novel ideas and general machine learning methods, this manuscript might be better suited to a more specialized journal such as Communications Physics.

1. According to my understanding, one possible restriction about the feasibility of the method is that, if the training data has fewer patterns of dynamics and the testing set has more, the method may not be effective anymore. That is, when the parameter k in Fig.1 is smaller than parameter m, the model does not have sufficient training patterns to learn, and it is questionable that the method can still construct the dynamics from a few measurements. As in Supplementary Fig.1, the training set has several times more dynamics than the testing dataset, and the test dataset only has three cases. What if the training data set has a similar size to the test data set, e.g., half of the system in Supplementary Fig.1 as the training dataset and the other half as the test data?

This is an important issue: If one could only reconstruct very few patterns from training on a lot of patterns, the method does not have the potential of generalization to more data or a new dataset encountered in nature. Currently, the method relies on synthetic chaotic systems for training. If the target system exhibits dynamics fundamentally dissimilar to the training set, performance may degrade. The paper assumes qualitatively similar dynamics, but it is challenging to rigoriously quantify this similarity.

2. Another limitation is about the resolution the method can learn. Since the goal is to reconstruct the dynamics from a few measurements, it can have an infinite number of dynamical systems to fit with the same set of observations, as noticed by the authors. In the examples shown, it is not convincing that the method can distinguish closely related dynamics. The current example shows the possibility of distinguishing the three-species chaotic food chain system, the chaotic Lorenz system, and the Lotka-Volterra system. However, what if the two systems to be tested are similar with a slightly different form of equation of parameters?

For example, if the two systems are both chaotic food chain systems, with a 2-fold change of a key parameter that leads to slightly different trajectories' behavior, can the method still be able to distinguish such two dynamical systems? In practice, the situation happens when some key parameters vary and can affect the system. So, if the current method cannot distinguish this scenario, it is not practically feasible to accurately learn the underlying equations. Indeed, the fact that the major results in this paper are based on predicting the three different dynamical systems indicates that the method may not be robust for more general cases as mentioned above.

3. Another related issue is that, although the manuscript considers the noisy observations, it is not very clear if the noisey effect can be similar to some perturbation on the underlying equation: If we have three systems of the chaotic food chain systems, with the second having a 2-fold change of a key parameter and the third having some noise, can method distinguish these three cases?

In the Supplementary Figs.10 and 11, the authors have shown the accuracy with respect to the noise strength, which is a good benchmark. Then, it will be nice to show the effect of the noise as well as the effect of perturbations on the underlying deterministic equation. Whether the method can distinguish these two sources of alternation on the dynamics? My suspicion is that the method may struggle to distinguish these when the strength of the parameter perturbation and noise are similar.

Besides the above limitations, I have some curious questions that might be further addressed:

1. There is a crucial hyperparameter in the training procedure, the smoothness penalty. According to my knowledge, it would significantly affect the training procedure because too smooth fitting would make the machine learn relatively coarse-graining dynamics, and less smooth fitting would make the fitted curve to be rough. Since this parameter may affect the reconstruction, the authors should show its effect in more detail.

2. Another question is about the role of the transformer. To my understanding, the transformer mainly encodes the training data and transforms the data to be an input of the reservoir computing. I wonder how necessary the transformer is and what unique advantages it has. What if we do not use the transformer and only have the reservoir computing, and would the performance be significantly worse? Besides, in Supplementary Fig.14, the authors have shown the case with using the LSTM, what if other advanced time sequence models, such as xLSTM, were used?

3. The training involves six NVIDIA GPUs, which may be prohibitive for some users. It would be helpful if more details of the computational efficiency, e.g., training time, were provided.

4. The manuscript mentions failure cases for systems "lacking dynamics". There can be a lot to be documented given my points 2,3 on the limitation, which might help reduce ambiguity about the performance of the method for more general cases.

About the code, I appreciate that the authors provide the code, as publicly available code and data enhance reproducibility. I have read the code on their Github website. They have organized the code well, with clear documentation in each of the code files. I suggest that the authors further provide a Jupiter notebook to help the reader use the code.

(Remarks on code availability)
I appreciate that the authors provide the code, as publicly available code and data enhance reproducibility. I have read the code on their Github website. They have organized the code well, with clear documentation in each of the code files. I suggest that the authors further provide a Jupiter notebook to help the reader use the code.

Reviewer #3

(Remarks to the Author)
The manuscript “Bridging known and unknown dynamics: machine learning inference from sparse observations” presents what is called a hybrid machine learning scheme for reconstructing the nonlinear dynamics when there is limited, sparse training data. They “hybrid” scheme is really two essentially independent techniques: First, a transformer is used to interpolate sparsely sampled data from a nonlinear system; Second, a reservoir computer is used to predict the future dynamical behavior of the system. This application of reservoir computing has been studied extensively in the literature (and nothing new in this regard is presented here); the novelty in this work is the use of a transformer to interpolate sparsely sampled data of a nonlinear dynamical system.

This is indeed an interesting application of a transformer, as interpolation of complex time series is an important and difficult task that is relevant for a variety of fields. However, there significant shortcomings in the presentation and in the analysis that prevent a complete evaluation of this technique. For that reason, I recommend the authors address the following before considering resubmission to Nature Communications.

1. Nothing new is done with or learned about reservoir computing in this work. My suggestion would be to minimize (or potentially eliminate) the portions that show that reservoir computing can reconstruct time series from evenly sampled systems: This is well-established in the literature. All the presence of a reservoir computer is serving here is to show that the interpolation performed by the transformer is sufficiently accurate for a reservoir computer to work; there are more quantitative metrics to quantify the accuracy of the interpolation.

2. Given the information presented in this work, it is not clear that the transformer's performance at interpolation is impressive. Addressing the following would help to clarify what exactly the transformer is doing and to evaluate the performance:

2a. The definition of sparsity used here is with respect to a sampling time rather than to the information content of the signal. Quantifying sparsity in this way is so dependent on the sampling time as to be completely useless. In an extreme case, one could imagine using a vanishingly small sampling time to sample a sine wave that has a frequency of 1 Hz. One could then obtain reconstruction with a sparsity of nearly 100%, but that would hardly be impressive. Please use a more appropriate definition of sparsity that depends on the information content of the signal. And please define sparsity much earlier in the text.

2b. This point is related to the previous point. The authors claim that the transformer can interpolate a signal sampled with up 80% sparsity, which sounds very impressive. However, for example, for the Lorenz system, the bandwidth is in the 1-2Hz range, implying that all that is required for an accurate reconstruction is a sampling rate of 2-4 Hz according to Nyquist. When sampling at 10 Hz as is done here, one would expect to be able to perform an accurate reconstruction with sparsity of at least 60% using basic trigonometric interpolation. However a comparison of the transformer interpolation technique with any simple interpolation techniques such as trigonometric or splines interpolation is lacking. Such a comparison should be done. I do not know the bandwidths of the other chaotic systems considered here, but they should be presented or the power spectra should be shown, so that the performance of the transformer can be compared with the Nyquist sampling criterion.

2c. In general, an information theoretic point of view is sorely lacking. The Nyquist criterion applies to random systems. It is expected that for deterministic chaotic systems, one can reconstruct the dynamics with less information than required by Nyquist. The quantity of information needed is quantified by the entropy rate of the chaotic system. A discussion of these considerations should be included in the text. Furthermore, what is the entropy rate of the chaotic system being interpolated? What is the entropy of the information being presented to the transformer? What are the fundamental (information theory) limitations of this technique?

2d. The authors state on p. 3 that the transformer can obtain “high reconstruction accuracy even when the available data is only 20% of that required to faithfully represent the dynamical behavior of the underlying system according to the Nyquist criterion.” I did not find any evidence for this claim in the text. In light of the comments above, I recommend that the authors provide some evidence for this claim in the text. And what is the Nyquist criterion for each system considered?

2e. The authors should also discuss the relation of their ideas to those of compressed sensing.

3. It is an important consideration (and limitation) that this technique seems to require that the number of variables provided to the transformer in the inference phase to be the same as those used for the testing phase. This limitation should be included in the main text, not hidden in the supplement.

4. The 28 different training systems all have associated power spectra and entropy rates. How do these power spectra compare with the test power spectra in terms of the total bandwidth? How do the entropy rates compare?

5. If you do not have many segments (which is the assumption), how do you know whether to believe either the transformer or the reservoir computer output? It is shown here that the transformer does not perform well for filtered noise. Is there some analysis that must be done to determine whether one should trust the transformer interpolation and/or the reservoir computer predictions?

6. The references in this manuscript are inappropriate. The fields of time series analysis, reservoir computing, and transformers are all enormous fields, yet twenty of the forty-four references in this manuscript are to papers written by the final author. For example, in referring to papers showing “that reservoir computing has the demonstrated ability to generate the long-term behavior of chaotic systems,” the authors neglect to cite the seminal paper by Pathak (which is cited elsewhere in this paper as Ref. 5) Instead, five papers are cited, all by the final author.

7. The authors write on p. 15: “Specifically, for the chaotic food chain, Lorenz and Lotka-Volterra systems, we set $\Delta s = 10dt = 0.1$, corresponding to approximately 1 over 40 $\sim$ 50 cycles of oscillation.” What does “1 over 40~50 cycles of oscillation” mean?

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have fully addressed my concerns in my review report. Also I have gone through all the responses that the authors made to the other comments in the last round of review. I think the present form could be accepted by NC.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors have made efforts to revise the manuscript in accordance with the reviewers' comments. They have also addressed most of my previous questions. I am now more convinced by the advancement of the method. If the authors can further address the following points, I would consider recommending it for publication. The comments below follow the order of the authors' responses.

1. "The following analogy from natural language processing (NLP), particularly the development of large-scale models such as ChatGPT that is also based on the transformer structure, may provide some insights."

It is not appropriate to draw an analogy between the current work and NLP, because the data in NLP are typically non-stationary. In contrast, the present problem primarily concerns stationary dynamical systems. Therefore, there is a conceptual gap between the two. If the authors insist on maintaining this analogy, it may lead to confusion or raise further questions. For instance, one may ask whether the proposed method can be extended to language-processing datasets. If the authors cannot address this discrepancy convincingly, they should remove such analogies both from the manuscript and their responses.

2. "In our work, the critical factor is not the number of testing cases relative to training, but rather ensuring that the training set captures a sufficiently diverse range of dynamical behaviors. Once the model has acquired this generalization ability, it can infer the governing dynamics of a variety of new systems from sparse observations."
"Based on this, we now show that once the transformer is well-trained under a sufficiently large number of synthetic systems, it can infer the governing dynamics of an arbitrary new system from sparse observations."

This is exactly the point I wanted to raise in my previous comments. The authors have conducted further analysis with a larger number of test datasets, and their main argument is that once the training set encompasses a sufficiently diverse range of dynamical behaviors, the model can generalize effectively. In this context, I would like to ask whether the model can be treated as a foundation model—for example, can it generalize across a sufficiently large and diverse space of dynamical systems? How to justify that the trained model can truly infer the governing dynamics of "an arbitrary new" dynamical system?

If the authors' argument holds, this should indeed be possible as an extrapolation. If not, what are the key factors limiting such generalization as a foundation model? What cases are the limitations of the current method in serving as a foundation model? The authors should clarify this point explicitly.

3. "There are few cases where the transformer struggles to produce accurate reconstruction due to the particularly complex dynamics and extremely sparse observation."

The authors should specify these cases clearly in the manuscript, such that readers can better understand the limitations of the method and the contexts in which it may fail.

(Remarks on code availability)
The code is well documented.

Reviewer #3

(Remarks to the Author)
The fundamental results appear to be that transformers can provide a useful technique to interpolate chaotic time series. This result is interesting and useful, but might be better suited for a more specialized journal or for Communication Physics.

I have one major additional comment on the manuscript: I believe that the definitions for sampling time and sparsity chosen result in substantial exaggeration of the performance of the transformer as an interpolator for chaotic time series. I think this is misleading, and I recommend re-evaluating the performance using better metrics. Please see attached pdf for detailed comments. Nevertheless, it is clear that the transformer is indeed an effective interpolator, and so this work will eventually be a valuable contribution to the nonlinear dynamics literature.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
I appreciate the authors' response to my comments and the revisions made to the manuscript. The authors have made commendable progress during the revision, including the addition of computational experiments on more nonlinear dynamical systems. The study presents an interesting approach; however, as demonstrated below, I remain uncertain about whether it reaches the conceptual depth expected for publication in Nature Communications.

In particular, the theoretical foundation for the generalization of the proposed method remains unclear, as acknowledged by the authors themselves. There is still no strong evidence that the method can serve as a foundation model or be reliably extrapolated to a broad class of previously unseen dynamical systems, even with the added numerical experiments. In other words, although the method demonstrates clear technical progress, it falls short of delivering a substantive conceptual breakthrough. This concern, also noticed by Reviewer 3, limits the perceived scope and general impact of the work.

Given these considerations, and should other reviewers express similar reservations, I would prefer to recommend that the manuscript might be better considered for publication in a more specialized journal.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors have addressed my concerns about the presentation of their results sufficiently. The claims presented in the manuscript are now supported by the data presented. I still think that the fundamental result is that transformers can provide a useful technique to interpolate chaotic time series. The authors appear to agree on this point. In my opinion, this result is interesting and potentially useful, but I still feel it might be better suited for a more specialized journal or for Communication Physics.

From a technical point of view, I can recommend publication of this manuscript. I hope the authors will consider the points raised in the attached pdf in a final revision, as I think they will further improve the manuscript.

Please see the attached PDF for full comments.

(Remarks on code availability)

Version 3:

Reviewer comments:

Reviewer #3

(Remarks to the Author)
The authors are playing with semantics to claim high sparsity when they are clearly oversampling, even by the textbook definition, which they misinterpret. They use this oversampling to claim incredible performance in terms of sparsity, when in reality the performance of the transformer is better than basic techniques only in a small region of sparsity, as discussed in previous rounds of review.

For the reasons above, I continue to recommend publication in a specialized journal. The most exaggerated claims have been removed from the manuscript at this point, and so from a technical point of view, the paper can be published.

(Remarks on code availability)

Summary of main changes

In response to the referees' comments, we have

1. investigated reconstructing the unseen dynamics of nonautonomous systems,
2. compared the performance of the transformer with recurrent neural networks such as xLSTM, traditional interpolation methods, and compressed sensing,
3. demonstrated that the performance of dynamics reconstruction on unseen systems scales with the number of training systems follows a power-law relationship,
4. provided extensive numerical results and arguments for the general applicability of our framework,
5. significantly revised and reorganized the structure of the manuscript to incorporate a broader range of perspectives and content,
6. redefined the sparsity metric S_r based on the Nyquist criterion,
7. calculated and analyzed the power spectral density and entropy rates of both training and target systems,
8. updated our codes to GitHub and data to Zenodo, and uploaded a Jupyter notebook to GitHub.

Point-by-point response to referee comments

Report of Referee $\#1$

General Comment: "*This manuscript uses the pre-training framework of transformer and also uses the reservoir computing to predict the long-term behaviors of complex dynamics where the data of the target system are incompletely collected. The idea of integrating the different frameworks of machine learning for rencontriucting dynamical systems only using limited information is interesting, which will certainly attracts attention from the community of complex dynamical systems and other related fields. The work deserves publication in NC; however, before the final acceptance, the following issues need to be substantially addressed.*"

Response: We appreciate the referee's positive assessment of our work, especially by stating "will certainly attract attention from the community of complex dynamical systems and other related fields" and "The work deserves publication in NC." The points raised by the referee have been fully addressed in the revised paper.

Comment 1: "*This reviewer notices that the investigated dynamical system could be non-autonomous; however, such a configuration, though broadening the use of the current framework, brings many indeterministic issues. Since the pre-training procedures also need to include the non-autonomous systems as many as possible, it is unclear how to choose these non-autonomous systems as well as the time-variant input, and how to set the ratio of the data from the autonomous/non-autonomous systems. More justifications need to be done; or just focus the discussion merely on the typical dynamical systems, which are autonomous.*"

Response: We agree with the referee that including nonautonomous dynamical systems can broaden the applicability of our framework. In our original work, both the training and target systems were autonomous. To ensure reliable performance on nonautonomous systems during deployment, the training data for the transformer should be from nonautonomous systems. For this purpose, we studied four nonautonomous dynamical systems: forced Rössler, forced Lorenz, mega extreme and memristive Laser systems. We used the first three as the training systems and the laser system as the unseen target. We studied two training strategies: (1) mixed training (with nonautonomous systems), including the three nonautonomous systems aforementioned and k-3 autonomous systems, for a total of k systems, and (2) purely autonomous training (without nonautonomous systems), using k autonomous systems. Both strategies are trained 50 times and the trained transformer is deployed to the nonautonomous laser system.

In our response to Comment 1 of Referee #2, we show that the performance of the transformer on a target system scales with the number of the training systems as a power law. We thus set $k = 10$ and $k = 20$ to obtain the reliable generalization. Since the total number of systems in our training pool is larger than k, we randomly select the required number of systems from the pool for each training. When the number of training systems is small ($k = 10$), the proportion of nonautonomous systems ($3/10$) is relatively high, and the model trained with the mixed strategy outperforms the one trained on autonomous systems only. However, for $k = 20$, the ratio of nonautonomous systems drops to $3/20$. In this case, the performance difference between the two strategies becomes negligible. Further improvement would require a larger and more diverse pool of nonautonomous systems, for which a systematic study of balancing the training data from autonomous versus nonautonomous dynamics would be helpful. In the revised manuscript, for clarity we focus on autonomous systems in the main text while reporting the new results on nonautonomous systems in Supplementary Materials.

We have included a new section in Supplementary Materials (on pages 22-24): Supplementary Note 9: Nonautonomous systems, with a new figure Fig. S20 (on page 23), which reads

- In our framework, both the training systems and the target unseen systems are autonomous, where the transformer is trained using data from a diverse set of dynamical systems and is then evaluated by recovering the underlying dynamics from sparse observations on the target system. When the target system is nonautonomous, training solely on autonomous systems may not be sufficient for reconstructing the target dynamics. To evaluate the capability of our framework in dealing with nonautonomous systems, we collect four such systems: the forced Rössler, forced Lorenz, mega extreme [26], and memristive laser [27] systems. Concretely, we use the first three systems as part of the training set and treat the laser system as an unseen target. The equations for the four nonautonomous systems are listed in Table S5.

  We compare two training strategies: (1) mixed training that includes the three nonautonomous systems and k-3 autonomous systems, for a total of k systems, and (2) purely autonomous training without any nonautonomous systems, using k autonomous systems. Both strategies are trained 50 times and evaluated on the nonautonomous laser system. The structure of the training with and without the nonautonomous systems and testing on the nonautonomous system is illustrated in Fig. S20(a). As the performance of transformer scales with the number of training systems in a power-law fashion, including a sufficient number of training systems can be crucial: we test this by setting $k = 10$ and $k = 20$ for comparison. Since the total number of systems in our training pool is larger than k, we randomly select the required number of systems from the pool for each training event. As shown in Fig. S20(b), when the number of training systems is small ($k = 10$), the fraction of nonautonomous systems ($3/10$) is relatively high, and the model trained with the mixed strategy outperforms the one trained solely on autonomous systems. However, for $k = 20$ where the ratio of nonautonomous systems is $3/20$, the performance difference between the two strategies becomes negligible.

We have also included the following sentence in the last paragraph in the Discussion section on page 13:

- It is worth noting that both the training and target dynamical systems in our experiments are autonomous. However, real-world systems can often be nonautonomous. To adapt the framework to target nonautonomous systems, we have developed a mixed training strategy that involves both autonomous and nonautonomous systems (Supplementary Note 9).

Comment 2: "*The long-term prediction of climate dynamics is used an illustrative example. However, such kind of discussion is also related crucially to the system which is either autonomous or non-autonomous. As we known, the climate system actually is not stationary but always time-variant. Then, how to provide the data for model pre-training is crucial to the prediction fidelity that one could achieve only using very limited observations. More discussion is anticipated.*"

Response: We agree with the referee that long-term prediction of climate dynamics can be nonautonomous, as the climate system is not stationary. We have added the following description to the last paragraph of Discussion on page 13:

- With regard to long-term prediction, climate dynamics are not stationary but often time-variant, i.e., nonautonomous. When providing the reservoir computer with high-fidelity outputs generated by the transformer from sparse observations, long-term climate prediction becomes feasible.

Comment 3: "*In addition to the reservoir computing, is there any other neural networks that could be used in dynamics prediction. As this reviewer knows, at least the recurrent neural networks such as the neural*

ODEs, the KAN, and the diffusion models could be the potential candidates for this task. Thus, the authors are suggested to make some comparisons using different neural networks and highlight their respective advantages and weaknesses. ”

Response: Reservoir computing has been demonstrated to be particularly suitable for predicting nonlinear dynamics (e.g., cited Refs. [3-22] in the main text). In the present work, we further validated its capability in performing long-term prediction of chaotic dynamics based on the transformer-generated output. Can other machine learning methods serve as alternatives to reservoir computing for this task? We have evaluated the performance of three alternative machine-learning schemes suggested by the referee: neural ODEs, the KAN, and diffusion models, and found that, unsurprisingly, reservoir computing consistently achieves the best reconstruction performance and the fastest single-realization convergence regardless of the sparsity level.

Report of Referee #2

Major Comment: "*The manuscript introduces a hybrid machine-learning framework with transformers and reservoir computing to reconstruct and predict nonlinear dynamics from sparse observations. This issue occurs in scenarios where historical data from the target system is unavailable. To solve this problem, the developed hybrid architecture leverages transformers to infer missing dynamics from sparse inputs and reservoir computing to predict long-term attractors. By training on synthetic chaotic systems and testing on three distinct models, the chaotic food chain, Lorenz, and Lotka-Volterra systems, the authors demonstrate the effectiveness of the method under metrics of MSE and prediction stability. The reconstruction is accurate under up to 80% sparsity and moderate noise levels.*

Overall, the manuscript presents a potentially promising framework for dynamics reconstruction under extreme data limitations. However, although the present method can reconstruct the dynamics from a few observations in the given examples, to my understanding, there are fundamental limitations of the methodology that may prevent its practical use in more general and real examples. Considering that Nature Communications is highly selective and publishes novel ideas and general machine learning methods, this manuscript might be better suited to a more specialized journal such as Communications Physics. "

Response: We thank the referee for stating that "the manuscript presents a potentially promising framework for dynamics reconstruction under extreme data limitations." The novelty of our hybrid machine-learning framework lies in: (1) the available observational data are random and sparse and (2) no training data from the target system are available. By training with "triple-randomness" across numerous training systems (varying training systems, input sequence length, and sparsity level), in the testing phase the transformer is able to reconstruct the dynamics of an unseen system with arbitrary data length and sparsity level. It is worth emphasizing that, once well-trained, the model can reconstruct any number of new target systems and distinguish between the systems with similar dynamics. We shall present supporting evidence in the Response.

Comment 1: "*According to my understanding, one possible restriction about the feasibility of the method is that, if the training data has fewer patterns of dynamics and the testing set has more, the method may not be effective anymore. That is, when the parameter k in Fig.1 is smaller than parameter m, the model does not have sufficient training patterns to learn, and it is questionable that the method can still construct the dynamics from a few measurements. As in Supplementary Fig.1, the training set has several times more dynamics than the testing dataset, and the test dataset only has three cases. What if the training data set has a similar size to the test data set, e.g., half of the system in Supplementary Fig.1 as the training dataset and the other half as the test data?*

This is an important issue: If one could only reconstruct very few patterns from training on a lot of patterns, the method does not have the potential of generalization to more data or a new dataset encountered in nature. Currently, the method relies on synthetic chaotic systems for training. If the target system exhibits dynamics fundamentally dissimilar to the training set, performance may degrade. The paper assumes qualitatively similar dynamics, but it is challenging to rigoriously quantify this similarity. "

Response: A key point is that the effectiveness of the model does not solely depend on the ratio between training and testing dataset sizes, but rather on whether the training set is sufficiently diverse and representative of the underlying dynamical patterns. The following analogy from natural language processing (NLP), particularly the development of large-scale models such as ChatGPT that is also based on the transformer structure,

may provide some insights. These models are trained on massive corpora - spanning billions of words - to learn a broad range of linguistic structures, reasoning patterns, and contextual understanding. Despite any single test input being minuscule compared to the training set, the model's learned generalization ability enables it to generate coherent and contextually appropriate responses across an infinite variety of queries. OpenAI's report (cited as Ref. [33] in the main text) showed that the model performance follows a power-law relationship with the size of the training dataset and the model size. This phenomenon suggests that once a model is trained on a sufficiently large and diverse dataset, it can generalize to unseen situations, provided they fall within the learned distribution of patterns.

In our work, the critical factor is not the number of testing cases relative to training, but rather ensuring that the training set captures a sufficiently diverse range of dynamical behaviors. Once the model has acquired this generalization ability, it can infer the governing dynamics of a variety of new systems from sparse observations - even if that specific system was not present in the training dataset. This is because the model has already learned a sufficiently expressive representation of dynamical behaviors, much like how an NLP model can generate well-structured sentences despite never having seen a particular phrase during training.

To provide more detailed support, we have simulated the transformer training on different number k of the training systems. Specifically, we range k from 1 to 28. For each value of k, we randomly sample a subset from the pool of 28 chaotic systems and calculated the average MSE over 50 iterations. The result presents a power-law behavior: the MSE on unseen target systems decreases with increasing k in a power-law fashion. This demonstrates that training our transformer-based framework on a diverse of chaotic systems with sufficient data is crucial. In the revised manuscript, we added a new figure Fig. 4 (on page 9 in Results 1.3), with panel (a) demonstrating the power-law relationship. The new content in this section on pages 8-9 reads as

- Transformer has the ability to reconstruct the time series of previously unknown dynamical systems, particularly under extreme sparsity. This capability stems from the generalizability of the transformer during its training on sufficiently diverse chaotic systems with large data. OpenAI reported a power-law relation between the transformer performance and model/data size [33]. Here we study how reconstruction performance depends on the number of training systems. Specifically, we train the transformer on k chaotic systems, where k ranges from 1 to 28. For each value of k, we randomly sample a subset from the pool of 28 chaotic systems and calculate the average MSE over 50 iterations. To ensure robustness, the MSE is averaged across the sparsity measure S_m whose value ranges from 0 to 1 at the interval of 0.05. As shown in Fig. 4(a), the MSE decreases with increasing k following a power-law trend with saturation, demonstrating that training our transformer-based framework on a diverse of chaotic systems with sufficient data is crucial for successful dynamics reconstruction.

In addition, we have added the following brief description in the third paragraph in Discussion (on page 13):

- In essence, the reconstruction performance on unseen target systems follows a power-law trend with respect to the number of synthetic systems used during the training phase.

In Fig. S1, we demonstrated that using only half of the system as the training dataset is insufficient. This is because transformers require a sufficiently diverse range of dynamical systems with abundant data to acquire robust generalization capability. Based on this, we now show that once the transformer is well-trained under a sufficiently large number of synthetic systems, it can infer the governing dynamics of an arbitrary new system from sparse observations. To validate this, we evaluated the performance of our previously trained transformer (the first version uploaded to GitHub) on 28 additional new chaotic systems, which were not used during training. The results demonstrate that the well-trained model is capable of reconstructing the dynamics of these

unseen systems, even with extremely high sparsity. It is worth noting that the number of training systems is $m = 28$ and the number of unseen target systems is $k = 31$ (three examples in the main text and the additional 28 systems). We have included the following explanation in Results 1.3 on page 9:

- Moreover, once the model has acquired this generalization ability, it can infer the governing dynamics of arbitrary new system from sparse observations. To demonstrate this, we have shown that our trained transformer performs well on 28 additional unseen target systems [34] (see Supplementary Note 12).

In addition, we have added a new section (Supplementary Note 12 entitled "Reconstructing dynamics of additional chaotic systems" in Supplementary Materials) with two new figures Figs. S25, S26 on pages 34-35 with examples of reconstructing 28 new target systems and their overall performance for extreme high sparsity. The description reads

- To further demonstrate that the transformer can reconstruct arbitrary systems, we evaluate the well-trained transformer (described in the main text) on 28 additional chaotic systems. Specifically, the transformer was previously trained on 28 systems and evaluated on three target systems: the food chain, Lorenz, and Lotka-Volterra systems. Now we introduce 28 new chaotic systems as targets for an extensive assessment. Representative reconstruction results are shown in Fig. S25. For most systems, the transformer can faithfully reconstruct the target dynamics for $S_r = 0.93$ and $L_s = 2000$. Figure S26 presents the performance of dynamics reconstruction of the 28 additional chaotic systems in terms of MSE (a) and reconstruction stability $R_s(\cdot)$ (b) for $S_r = 0.93$. The overall MSE is low and the reconstruction is highly stable across all 28 systems, demonstrating the power and generalizability of our transformer based reconstruction framework.

Furthermore, as the referee noted, rigorously quantifying the similarity between different dynamical systems remains a significant challenge. In many cases, even minor modifications to the governing equations can result in drastically different system behaviors. The transformer in our study does not memorize or store specific dynamical patterns encountered during training. Rather, it learns to organize the observed data and infer the underlying dynamics. This generalization capability allows the transformer to reconstruct the behavior of previously unseen systems during testing, even though such systems were not explicitly included in the training phase. Our results support this, demonstrating successful reconstruction across a diverse range of unseen dynamics (more than the number of training systems). There are few cases where the transformer struggles to produce accurate reconstruction due to the particularly complex dynamics and extremely sparse observation. In such a case, reducing the sparsity level is a viable solution to improve the performance.

Comment 2: "*Another limitation is about the resolution the method can learn. Since the goal is to reconstruct the dynamics from a few measurements, it can have an infinite number of dynamical systems to fit with the same set of observations, as noticed by the authors. In the examples shown, it is not convincing that the method can distinguish closely related dynamics. The current example shows the possibility of distinguishing the three-species chaotic food chain system, the chaotic Lorenz system, and the Lotka-Volterra system. However, what if the two systems to be tested are similar with a slightly different form of equation of parameters?*

For example, if the two systems are both chaotic food chain systems, with a 2-fold change of a key parameter that leads to slightly different trajectories' behavior, can the method still be able to distinguish such two dynamical systems? In practice, the situation happens when some key parameters vary and can affect the system. So, if the current method cannot distinguish this scenario, it is not practically feasible to accurately learn

the underlying equations. Indeed, the fact that the major results in this paper are based on predicting the three different dynamical systems indicates that the method may not be robust for more general cases as mentioned above."

Response: We are grateful to the referee for proposing the direction involving parameter variations within the same dynamical system. Our original results were based on three distinct dynamical systems, as they serve as the benchmark systems, and we conducted experiments from different perspectives to demonstrate the workings of our framework. Addressing the issue of parameter variations could further demonstrate generalizability of our framework.

Following the referee's suggestion, we consider the chaotic food chain system as a concrete example. As is common in ecological studies, the environmental carrying capacity K is an appropriate bifurcation parameter [25]. Often, a two-fold change in a key parameter can be too large for a chaotic system, as it may lead to a critical transition. For example, in the chaotic food chain system, decreasing K from 0.98 (our setting) to 0.97 would result in a less chaotic system. Moreover, further increasing K from 0.98 will lead to a crisis and collapse of the system. These considerations have led us to use the bifurcation values: $\mathrm{K} = 0.96$, 0.97, and 0.98 to evaluate the robustness of our framework.

We input extremely sparse observations of the food chain system with sparsity $S_r = 0.93$ from each of the three bifurcation parameters into the transformer model to reconstruct the time series. The reconstructed time series are then fed into a reservoir computer to predict the corresponding attractors. Each predicted attractor is compared with the ground truth at the three parameters. The results are conveniently quantified by the confusion matrix, which shows that each predicted attractor closely aligns with the ground truth attractor at the same parameter value, while remaining distinct from the attractors at the other parameter values. The results demonstrate that our framework is capable of successfully recovering the dynamics for the target system at different bifurcation parameter values. Nonetheless, as expected, if the parameter difference is too small, the framework would fail in recovering the similar dynamics, because the underlying dynamics are statistically indistinguishable. For instance, the dynamics from $\mathrm{K} = 0.98$ to $\mathrm{K} = 0.981$ are nearly identical. In such cases, even if the full dynamics were observable, it would practically be difficult for any machine-learning model to distinguish between them, especially with sparse data.

We have added the a new figure Fig. S16 on page 20 with the following content to Supplementary Note 7 (on pages 17-18):

- A pertinent question is, can the proposed framework recover correctly similar dynamics from nearby values of the bifurcation parameter? To address this question, we consider the chaotic food chain system with a varying bifurcation parameter as an example. Specifically, the environmental carrying capacity parameter K in Eq. (2) in the main text is a commonly used bifurcation parameter [25]. An example of the bifurcation diagram is shown in Fig. S16(a). We use three K values: 0.96, 0.97, 0.98, collect extremely sparse data with $S_r = 0.93$ from each of the parameter values, and reconstruct the time series through the transformer. The reconstructed time series for each parameter value are then fed into a reservoir computer to generate the corresponding attractors for comparison with the ground truth. The results are shown using the confusion matrix in Fig. S16(b), where the diagonal entries exhibit the smallest values in each row, indicating that each predicted attractor closely aligns with the ground-truth attractor at the same parameter, while the remaining entries characterize the distinction from the attractors at the other parameter values. The results demonstrate that our framework is capable of successfully recovering the dynamics from the target system at different values of the bifurcation parameter. Nonetheless, as expected, if the parameter difference is too small, the framework would fail in recovering the similar dynamics,

because the underlying dynamics are statistically indistinguishable. For instance, the dynamics from $\mathrm{K} = 0.98$ to $\mathrm{K} = 0.981$ are nearly identical. In such cases, even if the full dynamics were observable, it would be practically difficult for any machine-learning model to distinguish between them, especially with sparse data.

Comment 3: "*Another related issue is that, although the manuscript considers the noisy observations, it is not very clear if the noisey effect can be similar to some perturbation on the underlying equation: If we have three systems of the chaotic food chain systems, with the second having a 2-fold change of a key parameter and the third having some noise, can method distinguish these three cases?*

In the Supplementary Figs.10 and 11, the authors have shown the accuracy with respect to the noise strength, which is a good benchmark. Then, it will be nice to show the effect of the noise as well as the effect of perturbations on the underlying deterministic equation. Whether the method can distinguish these two sources of alternation on the dynamics? My suspicion is that the method may struggle to distinguish these when the strength of the parameter perturbation and noise are similar."

Response: It is interesting to compare "three chaotic food chain systems, with the second having a 2-fold change of a key parameter and the third having some noise". Following the suggestion by the referee, we consider three chaotic food chain systems: the first with $\mathrm{K} = 0.98$ (an example in the main text), the second also with $\mathrm{K} = 0.98$ but under noise of a moderate level σ, and the third with $\mathrm{K} = 0.97$. For each system, we apply our transformer reservoir-computing framework to reconstruct the attractor and to compare it with the ground truth. The performance is conveniently characterized by a 3 by 3 confusion matrix, where we compare each predicted attractor with the ground-truth attractor. The results show that the predicted attractors for $\mathrm{K} = 0.97$ and $\mathrm{K} = 0.98$ align most closely with the ground truth, consistent with our previous findings. However, when noise is added to the observations at $\mathrm{K} = 0.98$, the predicted attractor closely matches the ground truth attractor of the noiseless $\mathrm{K} = 0.98$ system, rather than the attractor at $\mathrm{K} = 0.97$ or the noisy attractor at $\mathrm{K} = 0.98$. This demonstrates a key distinction between parameter perturbation and observational noise: while noise perturbs the observed data, the framework is robust enough to reconstruct the underlying true dynamics. In contrast, a change in the bifurcation parameter alters the system's governing equations and thereby the dynamics themselves, which our framework correctly identifies as distinct.

We have added a new figure Fig. S17 on page 20 with the following content to Supplementary Note 7 (on pages 18-19):

- The observational noise considered in this Section does not alter the dynamics of the system. In contrast, variations in the bifurcation parameter can result in changes in the underlying dynamics. Is our framework robust enough to reconstruct the correct dynamics in these scenarios? To address this question, we consider three chaotic food chain systems: the first with $\mathrm{K} = 0.98$ (an example in main text), the second also with $\mathrm{K} = 0.98$ but under noise of a moderate level σ, and the third with $\mathrm{K} = 0.97$. For each system, we apply our transformer reservoir-computing framework to reconstruct the attractor and to compare it with the ground truth. The performance is conveniently characterized by the 3 by 3 confusion matrix, where we compare each predicted attractor with the ground-truth attractor. Figures S17(a) and S17(b) present examples of the reconstructed time series and the corresponding confusion matrix, respectively. The results show that the predicted attractors for $\mathrm{K} = 0.97$ and $\mathrm{K} = 0.98$ align most closely with their corresponding ground-truth attractors, consistent with our previous findings. However, when noise is added to the observations at $\mathrm{K} = 0.98$, the predicted attractor closely matches the ground-truth attractor

of the noiseless $\mathrm{K} = 0.98$ system, rather than the attractor at $\mathrm{K} = 0.97$ or the noisy attractor at $\mathrm{K} = 0.98$. The results demonstrate a key distinction between parameter perturbation and observational noise: while noise perturbs the observed data, the framework is robust enough to reconstruct the underlying true dynamics. In contrast, a change in the bifurcation parameter alters the system's governing equations and thereby the dynamics themselves, which our framework correctly identifies as distinct.

Comment 4: "*Besides the above limitations, I have some curious questions that might be further addressed:*

There is a crucial hyperparameter in the training procedure, the smoothness penalty. According to my knowledge, it would significantly affect the training procedure because too smooth fitting would make the machine learn relatively coarse-graining dynamics, and less smooth fitting would make the fitted curve to be rough. Since this parameter may affect the reconstruction, the authors should show its effect in more detail."

Response: The referee is insightful that smoothness penalty is a crucial hyperparameter in the training procedure and its effects should be studied. To this end, we selected four values of smoothness penalty: $\alpha_s = 0$ (without smoothness), $\alpha_s = 0.1$ (our setting), $\alpha_s = 0.5$, and a large penalty $\alpha_s = 1$. For each setting, we train the dynamics reconstruction framework 50 times and evaluate its performance on the target systems. The results show that excessive smoothness leads the model to learn overly coarse-grained dynamics, causing a large MSE. When no smoothness penalty is applied, while the MSE remains low, the reconstructed curves exhibit poor smoothness.

We have added the description in the last paragraph of Methods 3.2 (on page 18):

- It is worth noting that the smoothness penalty is a crucial hyperparameter that should be carefully selected. Excessive smoothness leads the model to learn overly coarse-grained dynamics, while absence of a smoothness penalty causes the reconstructed curves to exhibit poor smoothness (Supplementary Note 5).

In addition, we have added a new figure Fig. S7 with the following description in Supplementary Note 5 (on page 11):

- Smoothness penalty α_s is also a crucial hyperparameter in the training procedure. We study the effect of this penalty on reconstruction quality through two examples in Figs. S7(a) and S7(b), where training with smoothness yields smoother reconstructions. To evaluate the impact of α_s on the reconstruction accuracy, we test four values: $\alpha_s = 0$ (no smoothness), $\alpha_s = 0.1$ (our setting), $\alpha_s = 0.5$, and a large penalty $\alpha_s = 1$. In each case, the dynamics reconstruction framework is trained 50 times and evaluated on the target system. The results in Figs. S7(c) and S7(d) show that excessive smoothness leads the model to learn overly coarse-grained dynamics, resulting in a high MSE. Conversely, when the smoothness is small or absent, the MSE remains similarly low. We conclude that the smoothness penalty should be chosen appropriately to ensure both accurate and smooth reconstruction.

Comment 5: "*Another question is about the role of the transformer. To my understanding, the transformer mainly encodes the training data and transforms the data to be an input of the reservoir computing. I wonder how necessary the transformer is and what unique advantages it has. What if we do not use the transformer and only have the reservoir computing, and would the performance be significantly worse? Besides, in Supplementary Fig.14, the authors have shown the case with using the LSTM, what if other advanced time sequence models, such as xLSTM, were used?*"

Response: We have provided an introduction to transformer in Methods. It should be noted that transformer and reservoir computing serve entirely distinct purposes. The transformer is trained on a wide variety of synthetic dynamical systems, where each training iteration involves a randomly selected system, sparsity level, and sequence length. Once trained, the transformer can be applied to unseen target systems with arbitrary sparsity and sequence lengths for recovering the time series from sparse observations. Reservoir computing is not able to perform this task due to its one layer structure, less adjustable parameters, and more importantly, its recurrence structure. However, the transformer alone produces only short-term reconstructions. To capture the long-term behavior or the "dynamical climate" of the target system, reservoir computing is employed to generate arbitrarily long future dynamics using the short reconstructed time series from the transformer. This is where reservoir computing comes in and both models play complementary roles in the framework rather than performing the same task.

We demonstrated that RNNs, such as LSTM and xLSTM, capture dynamics differently from the transformer. RNNs store information in their hidden states and rely on the memory of past inputs to make predictions. As a result, they tend to overfit to the specific systems seen during training, particularly the most recent ones, and gives poor generalization. In contrast, the transformer does not rely on memory but instead learn to reconstruct dynamics by identifying the correlations among the observed points through its attention mechanism. This allows them to capture more fundamental and system-independent properties. As a result, LSTM-based RNNs are not appropriate for the reconstruction stage. In addition, we have shown the superiority of the transformer over compressed sensing and traditional interpolation methods. For further details, please refer to our Response to Comment 2 of Referee #3. We have also demonstrated the necessity of using reservoir computing over neural ODEs, KANs, and diffusion models, as addressed in our Response to Comment 3 of Referee #1.

We have revised Fig. S21 and modified the previous section title into Supplementary Note 10: Reconstructing dynamics with traditional methods, which includes the content about xLSTM (on page 24). Accordingly, we have added the following description (on page 25 and 26):

- In addition, the extended LSTM (xLSTM) [29] is a variant of the traditional LSTM architecture designed to enhance its capacity for modeling complex temporal dependencies in sequential data. In addition to inheriting the gating mechanisms of LSTM: input, forget, and output gates, which enable it to capture long-term dependencies, xLSTM introduces two critical extensions: (i) a learnable exponential decay parameter introduced from the last observed value for a missing data value and (ii) a residual skip connection between two stacked LSTM blocks of identical configuration. The structure of xLSTM is shown in Fig. S21(a).

- We investigate whether LSTM and xLSTM, as representative of RNNs, can be appropriate substitutes for the transformer for dynamics reconstruction from sparse data.

- We demonstrate that RNNs, such as LSTM and xLSTM, capture dynamics differently from the transformer. The former store information in their hidden states and rely on memory of past inputs to make predictions. As a result, they tend to overfit to the specific systems seen during training, particularly the most recent ones, exhibiting poor generalizability. In contrast, the transformer does not rely on the memory but instead learn to reconstruct dynamics by identifying the correlations among observed points through its attention mechanism. This allows them to capture more fundamental and system-independent properties.

Comment 6: "*The training involves six NVIDIA GPUs, which may be prohibitive for some users. It would be helpful if more details of the computational efficiency, e.g., training time, were provided.*"

Response: We use six NVIDIA GPUs for the all the simulations in our work, as some experiments require substantial computational time. For example, each hyperparameter setting involves 50 independent training runs. However, a single training run typically takes approximately 30 minutes using one NVIDIA RTX A6000 GPU.

We have revised the description in the last paragraph of Methods 3.3 on page 19:

- All simulations are run using Python on computers with six RTX A6000 NVIDIA GPUs. A single training run of our framework typically takes about 30 minutes using one of the GPUs.

Comment 7: "*The manuscript mentions failure cases for systems "lacking dynamics". There can be a lot to be documented given my points 2,3 on the limitation, which might help reduce ambiguity about the performance of the method for more general cases.*"

Response: We have thoroughly addressed the concerns raised by the referee, including major points 1-3, and have demonstrated that our model is applicable to general scenarios. Specifically, we have shown that our framework can be trained on a set of training systems and then successfully applied to an unlimited range of unseen target systems. Even when applied to the same target system with slightly different equation parameters, the framework is able to reconstruct the correct dynamics. Moreover, the proposed framework can effectively distinguish between perturbations on the equation parameters and observational noise, since the former alters the underlying dynamics while the latter only obscures them.

We acknowledge that there are a few failure cases in dynamics reconstruction, which may arise due to the high complexity of certain systems. We note that such limitations are common across predictive models, as seen in hallucinations in language generation or inaccuracies in weather forecasting. There is no model which performs without mistakes. For failure cases in our study, a simple and effective solution is to reduce the sparsity, i.e., to increase the amount of observed data so that more information from the system is available for reconstruction. Furthermore, we have shown that the performance of our framework on unseen target systems depends on the number of training systems, following a power-law relationship. Preparing a better-trained model by increasing the diversity and number of training systems is another viable solution.

Comment 8: "*About the code, I appreciate that the authors provide the code, as publicly available code and data enhance reproducibility. I have read the code on their Github website. They have organized the code well, with clear documentation in each of the code files. I suggest that the authors further provide a Jupiter notebook to help the reader use the code.*"

Response: We have provided a Jupyter notebook in our GitHub repository (under the 'examples' folder), which includes the process of generating synthetic and target systems, as well as reconstructing the dynamics of target systems using the transformer.

We have revised the description in the GitHub repository as follows:

- In addition to using our pre-trained model, we provide the script chaos_transformer_train.py for readers to train the transformer themselves. After training, please ensure that the 'save_file_name' variable in chaos_transformer_read_and_test.py is updated to match your saved model file name so that your own trained model is used during testing. A Jupyter notebook is provided in the 'examples' folder to help readers understand and reproduce the code workflow.

Furthermore, readers may generate more diverse synthetic systems for training, which can enhance the reconstruction performance on previously unseen target systems during testing. The performance degradation follows a power-law relationship with respect to the diversity of training systems.

In addition, we have updated our data on Zenodo and the codes on GitHub.

Report of Referee #3

Major Comment: "*The manuscript "Bridging known and unknown dynamics: machine learning inference from sparse observations" presents what is called a hybrid machine learning scheme for reconstructing the nonlinear dynamics when there is limited, sparse training data. They "hybrid" scheme is really two essentially independent techniques: First, a transformer is used to interpolate sparsely sampled data from a nonlinear system; Second, a reservoir computer is used to predict the future dynamical behavior of the system. This application of reservoir computing has been studied extensively in the literature (and nothing new in this regard is presented here); the novelty in this work is the use of a transformer to interpolate sparsely sampled data of a nonlinear dynamical system.*

This is indeed an interesting application of a transformer, as interpolation of complex time series is an important and difficult task that is relevant for a variety of fields. However, there significant shortcomings in the presentation and in the analysis that prevent a complete evaluation of this technique. For that reason, I recommend the authors address the following before considering resubmission to Nature Communications."

Response: We appreciate the referee's positive assessment of our work by commenting "This is indeed an interesting application of a transformer, as interpolation of complex time series is an important and difficult task that is relevant for a variety of fields." We thank the referee for the insightful comments and have made significant revisions to address them.

Comment 1: "*Nothing new is done with or learned about reservoir computing in this work. My suggestion would be to minimize (or potentially eliminate) the portions that show that reservoir computing can reconstruct time series from evenly sampled systems: This is well-established in the literature. All the presence of a reservoir computer is serving here is to show that the interpolation performed by the transformer is sufficiently accurate for a reservoir computer to work; there are more quantitative metrics to quantify the accuracy of the interpolation.*"

Response: We wish to clarify the role of reservoir computing (RC) in our hybrid machine-learning framework. The referee is absolutely right that RC has been studied extensively in the literature. In our hybrid machine learning framework, the roles of RC are: (1) validating the reliability of reconstructed time series by transformer, (2) refining the output dynamics of transformer, and (3) generating arbitrarily long time series of the target system. In brief, transformer is responsible for reconstructing the time series from sparse observations of the unseen dynamical system, and the reservoir computer further learns and refines the inferred dynamics by the transformer, generating more stable and accurate predictions.

When the sparsity level is low, it is reasonable that only a transformer would be sufficient to reconstruct the dynamics, in spite of the dynamics being new. However, as the observational data become more sparse, the transformer alone can fail to reconstruct the complete dynamical structure. In such cases, the reservoir computer is exploited to further refine the dynamics inferred by the transformer.

To clarify the role of reservoir computing in our hybrid machine learning framework, we illustrate the difference between the outputs of transformer and those of the hybrid transformer/reservoir computing under high sparsity by adding a new figure, Fig. S13 (on page 17) in Supplementary Note 6, Section "Performance of long-term climate reconstruction" with the following descriptions (on pages 14-16):

- In addition, the reservoir computer plays a crucial role in our hybrid machine learning framework, whose roles are: (1) validating the reliability of reconstructed time series by the transformer, (2) refining the

output dynamics of the transformer, and (3) generating arbitrarily long time series of the target system. This is particularly necessary when the observational data are highly sparse, where the transformer alone becomes insufficient to accurately reconstruct the full dynamics. In such cases, the reservoir computer learns from the transformer's output, refines and consolidates the reconstructed dynamics.

Figure S13 presents examples of dynamics reconstruction for three target systems: the chaotic food chain, Lorenz, and Lotka-Volterra systems, all under extreme sparsity conditions. The observed sparse data from these systems are processed by a transformer which is well trained on data from other systems. While the output of the transformer approaches the ground truth, there is a mismatch in the fine structural details, as shown in Fig. S13. To overcome this difficulty, we train an independent reservoir computer for each target system, using the transformer-reconstructed time series as the training data. The reservoir computer produces more accurate reconstructed attractors better capturing the underlying dynamics.

The referee is insightful that "more quantitative metrics to quantify the accuracy of the interpolation," which is quite important to the first part of our framework - transformer for sparsely sampled data of a new nonlinear dynamical system. We have revised significantly the structure of the manuscript to place greater emphasis on the transformer, highlighting its role and performance from multiple perspectives. Please see our Responses below for the detailed revisions.

Comment 2a: "*Given the information presented in this work, it is not clear that the transformer's performance at interpolation is impressive. Addressing the following would help to clarify what exactly the transformer is doing and to evaluate the performance:*

The definition of sparsity used here is with respect to a sampling time rather than to the information content of the signal. Quantifying sparsity in this way is so dependent on the sampling time as to be completely useless. In an extreme case, one could imagine using a vanishingly small sampling time to sample a sine wave that has a frequency of 1 Hz. One could then obtain reconstruction with a sparsity of nearly 100%, but that would hardly be impressive. Please use a more appropriate definition of sparsity that depends on the information content of the signal. And please define sparsity much earlier in the text."

Response: We agree that defining sparsity solely based on the proportion of missing time points without considering the information content of the signal is problematic and can be misleading. To address this concern, we have revised our definition of sparsity to incorporate the fundamental constraints from the Nyquist sampling theorem, thereby making it dependent on the information content of the signal rather than arbitrary sampling choices. Previously, the sparsity is defined as:

$$S_m = \frac{L_s - L_s^O}{L_s},$$

where L_s is the total number of samples in a complete dataset, and L_s^O is the number of observational points within the dataset. To distinguish the updated formulation, we define the revised, information-theoretic sparsity as:

$$S_r = \frac{L_s - L_s^{\mathrm{O}}}{L_s - L_s^N},$$

where $L_s^N = 2f_{\max}T$ represents the minimum number of samples required according to the Nyquist theorem, with $f_{\max}$ being the effective cutoff frequency of the signal and T the total time duration corresponding to L_s.

Under this information theoretic definition, sparsity has more meaningful bounds: $S_r = 0$ represents the fully observed case (at our predefined sampling rate), $S_r = 1$ represents the theoretical minimum sampling case (exactly at the Nyquist rate), and $S_r > 1$ indicates sub-Nyquist sampling, where perfect reconstruction becomes theoretically impossible without additional constraints.

As the natural frequency ranges of chaotic systems vary significantly, using a vanishingly small sampling time is not meaningful. Previously, we chose the sampling rate for each system to ensure that the attractor remains sufficiently smooth while limiting the number of sampled points. After a recalculation based on S_r, we find that the difference between the two definitions is not large in practice. For instance, in the chaotic food chain system, a previously defined sparsity of $S_m = 0.8$ corresponds to $S_r = 0.93$ using our new definition. To improve clarity, we now adopt S_r as the primary measure throughout the manuscript. However, for intuitive understanding, we continue to report the corresponding S_m values alongside S_r wherever appropriate.

We have clarified the definition of sparsity that depends on the information content of the signal and defined it earlier in our manuscript (the second paragraph on page 3), which reads:

- In practice, the natural sampling rate Δs is chosen to generate the complete dataset of a dynamical system. It should ensure that the attractor remains sufficiently smooth while limiting the number of sampled points. Let the total number of samples in the dataset be L_s and the actual number of randomly selected observational points be L_s^O. The sparsity measure of the dataset can be defined as $S_m = (L_s - L_s^O)/L_s$. However, this measure depends on the natural sampling rate of the dynamical system.

  To quantitatively describe the extent of sparsity in the observational data from an information-theoretic perspective, we introduce a metric that incorporates the constraints from the Nyquist sampling theorem:

$$S_r = \frac{L_s - L_s^O}{L_s - L_s^N},$$

  where $L_s^N = 2f_{\max} \cdot T$ represents the minimum number of samples required according to the Nyquist theorem with $f_{\max}$ being the effective cutoff frequency of the signal and T the total time duration corresponding to L_s. In this definition, $S_r = 0$ indicates fully observed data (at the natural sampling rate), $S_r = 1$ corresponds to the theoretical minimum sampling case (at the Nyquist rate), and $S_r > 1$ indicates sub-Nyquist sampling where perfect reconstruction becomes theoretically impossible without additional constraints. Our framework takes into account not only high sparsity but also the randomness in observations. More information about the determination of "cutoff" frequency of chaotic systems can be found in Supplementary Note 3.

It is worth noting that, unlike periodic signals, determining the bandwidth of chaotic systems poses unique challenges due to their broad spectra. Instead of identifying a single dominant frequency, we calculate an "effective bandwidth" by finding the frequency that captures 98% of the signal's power in the cumulative power spectral density, where the Welch's method is used to estimate the power spectral density, followed by determining the cutoff frequency where the cumulative power reaches the 98% threshold.

We have added one new section in the Supplementary Materials, Supplementary Note 3: Effective bandwidth of chaotic systems, to describe how we calculate the effective bandwidth, where Table S3 (on page 30) lists the dominant frequency, cutoff frequency, and sampling rate for each system. The added section on pages 4-5 reads :

- The Nyquist sampling theorem in signal processing establishes the conditions for reconstructing a continuous time band-limited signal from its discrete samples. Specifically, to perfectly reconstruct a band-limited signal, the sampling rate must be at least twice the highest frequency component present in

the signal; this minimum sampling rate is known as the Nyquist sampling rate. Mathematically, if a continuous-time signal $x(t)$ contains no frequency components above $f_{\max}$ Hz, it can be faithfully represented by discrete samples taken at a rate $f_s > 2f_{\max}$ Hz. It is worth noting that the theorem requires perfect uniformly sampling with no noise in the measurement with ideal reconstruction filters.

For periodic signals such as sinusoidal waves, the bandwidth determination is straightforward, as the highest harmonic component determines the Nyquist rate. Chaotic systems present unique challenges for bandwidth determination due to their complex and broad spectra. In particular, chaotic systems typically contain continuous, broadband power spectra rather than discrete frequency components, and the power spectral density decays with the frequency, theoretically extending to infinite frequencies with diminishing power. In addition, different chaotic systems exhibit distinct spectral distributions. Since chaotic signals can contain energy across an infinitely wide frequency range in principle, a strict application of the Nyquist theorem would require an infinitely high sampling rate. However, this is neither practical nor necessary for applications. To address this challenge, we define the "effective bandwidth" based on the cumulative power spectral density, which defines a practical frequency limit that captures the significant dynamical information of the system, while excluding negligible high-frequency components.

Specifically, we define a "cutoff" frequency $f_{\max}$ as the frequency below which 98% of the total signal power is contained. We calculate the power spectral density (PSD) by employing Welch's method [22], where the time series $x(t)$ is divided into overlapping segments and a window function is applied to each segment. The PSD is then computed and averaged as:

$$P_{xx}(f) = \frac{1}{K}\sum_{i=1}^{K}|\sum_{n=0}^{N_x-1} w(n)x_i(n)e^{-j2\pi fn}|^2,$$

where K is the number of segments, N_x is the segment length, $w(n)$ is the window function, and $x_i(n)$ is the i-th segment of the signal. The dominant frequency f_d is computed as the maximum frequency across all dimensions, where in each dimension, it is identified as the frequency corresponding to the maximum value in PSD. Afterward, we compute the cumulative power distribution by integrating the PSD from zero frequency to each frequency point:

$$C(f) = \frac{\int_0^f P_{xx}(\xi)d\xi}{\int_0^\infty P_{xx}(\xi)d\xi}.$$

We identify the frequency $f_{\max}$ at which the cumulative power reaches our predefined threshold, i.e., $C(f_{\max}) = 0.98$. Table S3 shows the dominant frequency f_d, effective bandwidth $f_{\max}$, and the sampling rate Δs for each chaotic system in our study. The quantity L_s^N defined in the main text, where $L_s^N = 2f_{\max} \cdot T$, can be calculated as $L_s^N = 2f_{\max} \cdot L_s \cdot \Delta s$.

Due to the revision of the sparsity measure from S_m to S_r, we have recalculated the corresponding sparsity values and updated the descriptions in the main text accordingly. In addition, we have revised Figs. 2, 3, 5, S5, S9, S11, S12, S14, S15, S19, and S21, wherever it is presented. Furthermore, for the new Figs. 4, S20, and S22 added in response to the referees' comments, we have also adopted the revised definition of sparsity S_r.

Calculating the Nyquist frequency for the chaotic systems made us recognize that the stochastic signal, which we have used as a counterexample, possessed a different Nyquist sampling rate than that in our original version. To ensure a more appropriate comparison, we adjusted the Gaussian filter from a standard deviation of $\sigma_g = 12$ to $\sigma_g = 9$, yielding a cutoff frequency that is more closely aligned with that of the target systems. We

reran the simulations for reconstructing the dynamics of the stochastic signal and obtained consistent results with our previous findings: successful reconstruction by the transformer requires that the sparse data be associated with some nonlinear dynamical process. We have revised the pertinent figures, Figs. S18 and S19 (on page 21 and 22, respectively) and descriptions in Supplementary Note 8 on the end of page 20 accordingly:

- A Gaussian filter with the standard deviation $\sigma_g = 9$ of the Gaussian kernel is applied to smooth out the generated noisy data.

Comment 2b: "*This point is related to the previous point. The authors claim that the transformer can interpolate a signal sampled with up 80% sparsity, which sounds very impressive. However, for example, for the Lorenz system, the bandwidth is in the 1-2Hz range, implying that all that is required for an accurate reconstruction is a sampling rate of 2-4 Hz according to Nyquist. When sampling at 10 Hz as is done here, one would expect to be able to perform an accurate reconstruction with sparsity of at least 60% using basic trigonometric interpolation. However a comparison of the transformer interpolation technique with any simple interpolation techniques such as trigonometric or splines interpolation is lacking. Such a comparison should be done. I do not know the bandwidths of the other chaotic systems considered here, but they should be presented or the power spectra should be shown, so that the performance of the transformer can be compared with the Nyquist sampling criterion.*"

Response: We fully agree with the observation that, under moderate sparsity conditions - approximately 40% of the data available (which corresponds to S_r about 0.7 and S_m value about 0.6), accurate reconstruction can be achieved using basic interpolation techniques. As the referee correctly pointed out, our transformer model is designed to be broadly adaptable to various sparsity levels. During training, the model is exposed to a wide range of chaotic systems with randomized sparsity. This training strategy allows the model to generalize well and handle new, unseen dynamical systems with arbitrary sequence lengths and sparsity levels during deployment. Once trained, the transformer can reconstruct time series across the full spectrum of observational sparsity, including both low- and high-sparsity regimes, which would not be possible with any existing interpolation methods.

We wish to stress that, if the available datasets have a low level of sparsity, traditional methods such as linear or spline interpolation can be effective due to their simplicity and computational efficiency. However, as the sparsity level increases, these classical methods by design face significant limitations, primarily due to their inability to capture the underlying dynamics of the systems. In such scenarios, the advantage of our transformer model becomes evident. A direct comparison between our approach and the traditional interpolation methods is meaningful and important. To this end, we have included a new figure, Fig. 4 (in the main text on page 9), to visually demonstrate this comparison. In panel (c), we present an example under high sparsity ($S_r = 1.0$). While both linear and spline interpolation fail to recover the underlying temporal structure, our transformer successfully reconstructs the correct dynamical behavior. We have also added a new section to the main text: key features of dynamics reconstruction on pages 8-10, to further show the characteristics of our transformer based dynamics reconstruction. To explain the details about the comparison of dynamics reconstruction methods, we have added the following content to Results 1.3 (the second and third paragraph on page 10):

- While our method is applicable across a wide range of sparsity, traditional techniques such as linear and spline interpolation can also achieve high accuracy when the sparsity level is low. These classical methods are simple and computationally efficient, and perform adequately in regimes with sufficient observational data. However, as the data sparsity level increases, the limitations of traditional interpolation become

evident. To explicitly demonstrate this, we take two examples of traditional interpolation methods as an example: linear and spline interpolation, where the former approximates missing values by connecting the nearest available data points with straight lines and the latter constructs piecewise polynomial functions to produce smooth transitions between observed points [37,38]. Both methods, despite their simplicity, by design lack the capacity to capture the intrinsic dynamics of complex systems. Figure 4(c) shows a representative example for sparsity $S_r = 1.0$ ($S_m = 0.86$), where the transformer successfully reconstructs the underlying dynamics, but the linear and spline interpolation methods fail to recover the correct temporal structure.

To quantify the performance, we calculate the MSE between the reconstructed time series and the ground truth. Figure 4(d) shows the reconstruction performance across a range of sparsity levels for a fixed sequence length, $L_s = 2,000$ (approximately 400 oscillation cycles). Results are averaged over 50 independent realizations, with shadowed areas indicating the standard deviation. When the sparsity measure S_r is low, all three methods - transformer, linear, and spline interpolation - perform comparably. However, once S_r exceeds approximately 0.7, the transformer begins to outperform the other methods, with its advantage becoming more pronounced as the available data become increasingly more sparse.

The bandwidth values of the other chaotic systems are listed in Table S3 in Supplementary Materials, as mentioned in the response to the previous comment. The S_r values have also been changed accordingly to include the Nyquist sampling information.

Comment 2c: "*In general, an information theoretic point of view is sorely lacking. The Nyquist criterion applies to random systems. It is expected that for deterministic chaotic systems, one can reconstruct the dynamics with less information than required by Nyquist. The quantity of information needed is quantified by the entropy rate of the chaotic system. A discussion of these considerations should be included in the text. Furthermore, what is the entropy rate of the chaotic system being interpolated? What is the entropy of the information being presented to the transformer? What are the fundamental (information theory) limitations of this technique?*"

Response: We agree that more information theoretic point of view is needed. We have significantly revised the structure of the paper, e.g., redefining the previous sparsity measure S_r, adding a new section in the main text on the key features of dynamics reconstruction, and quantifying the entropy rate of the input and output of the transformer (i.e., the entropy rate of the chaotic system being presented to the transformer and being interpolated).

For deterministic chaotic systems, if we have access to data from the target system, we can train a machine-learning framework to reconstruct the dynamics, with less information than that required by the Nyquist criterion, as shown in Fig. 1(a). However, if the data from the target system is completely unknown and only sparse observation is available, existing methods fail to reconstruct the dynamics. We have shown the failure of some typical interpolation methods in the response of the previous comment of the referee. In addition, we have thoroughly presented the failure of machine-learning methods such as LSTM and xLSTM for this task in Supplementary Information (please see our Response to Referee #2, comment 5 for details). Furthermore, a comparison between the transformer and compressed sensing methods is provided in our response to Comment 2e below. We wish to emphasize that the Nyquist sampling theorem assumes uniform sampling intervals and noise-free measurements. In contrast, our framework operates under conditions where the observations are both randomly sampled and contaminated by noise, which has been briefly stated in Discussion (on page 13):

- Moreover, we have demonstrated the superiority of our proposed hybrid machine-learning scheme to traditional interpolation methods, traditional recurrent neural networks, and compressed sensing (Supplementary Note 10).

Furthermore, for chaotic systems, the information content is quantified by the entropy rate. We used the Kolmogorov-Sinai entropy to quantify the entropy rate, which measures the average number of nats of information produced per unit time by the system and reflects its intrinsic unpredictability. In our study, since we assume that the governing equations of systems are unknown, we apply a recurrence-based approach that is both data-driven and robust to sparse observability. We have added a new subsection in Methods 3.7, on page 20, which reads

- We estimate the entropy rate of the dynamical system using the Kolmogorov-Sinai (KS) entropy, denoted as h_{KS}. This quantity measures the average number of nats of information produced per unit time by the system and reflects its intrinsic unpredictability. Since the governing equations of systems are assumed to be unknown, we apply a recurrence-based approach that is both data-driven and robust to sparse observability [36]. This method is based on the distribution of first recurrence times, i.e., the time it takes for the system to return close to a previous state. Given time series $\mathbf{X} \in \mathbb{R}^{L_s \times D}$, where $\mathbf{x}(t)$ is its state vector at time t, we compute the pairwise Euclidean distances. Specifically, for each time point t, we identify the smallest time lag τ such that:

$$||\mathbf{x}(t) - \mathbf{x}(t+\tau)|| < r,$$

where r is a fixed recurrence threshold. Through this process, a set of recurrence times $\tau(k)$ can be obtained, which defines the empirical distribution $\rho(\tau)$. The entropy rate is then estimated as:

$$h_{KS} \approx \frac{1}{\tau_{\min}} \sum_{\tau} \rho(\tau) \log(\frac{1}{\rho(\tau)}),$$

where $\tau_{\min}$ is the minimum observed recurrence time. In our work, we set the recurrence threshold $r = 0.5$ after normalizing the data to unit variance.

We have calculated the entropy rate for sparse observations, the interpolated time series by the transformer in comparison with the ground truth. We have added one panel to Fig. 4 on page 9, and added the following description to the last paragraph of pages 9-10 in Results 1.3:

- How much information is contained in the chaotic systems being presented to and interpolated by the transformer? We employ Kolmogorov-Sinai (KS) entropy [35, 36], denoted as h_{KS}, to estimate the entropy rate of the dynamical systems (Methods). For the transformer outputs, i.e., the reconstructed time series, and the ground truth data, we estimate h_{KS} directly using this method. However, a challenge is that the input to the transformer is sparse, which we meet by considering only those time steps in the input sequence where all dimensions are observed, while discarding time points with missing values in any variable. The processed input is then used to calculate the entropy rate. The estimated h_{KS} for both the input and the reconstructed time series generated by the transformer are shown in Fig. 4(b), where the black dashed line denotes the KS entropy calculated from the ground truth. While increased sparsity S_r corresponds to a gradual increase in entropy in the reconstructed time series, it remains closer to the ground truth than the entropy of the sparse observations, which exhibit substantial divergence. From an information theoretic point of view, the transformer receives sparse observations characterized by high uncertainty and disorder, yet it successfully reconstructs the dynamics to closely match the ground truth.

Our framework focuses on random and sparse observations, in contrast to the strict assumption of uniform sampling and noise-free measurements underlying the Nyquist theorem. Our transformer-based dynamics reconstruction framework relaxes these conditions, allowing for randomly sampled observations with the presence of observational noise. Under this setting, the number of observational points can be near the threshold required by the Nyquist theorem. However, we acknowledge that when the number of observations falls significantly below the Nyquist rate, the reconstruction performance degrades considerably. We have redefined sparsity relative to the Nyquist theorem and explicitly demonstrated how the performance changes as the sparsity level increases.

Comment 2d: "*The authors state on p. 3 that the transformer can obtain "high reconstruction accuracy even when the available data is only 20% of that required to faithfully represent the dynamical behavior of the underlying system according to the Nyquist criterion." I did not find any evidence for this claim in the text. In light of the comments above, I recommend that the authors provide some evidence for this claim in the text. And what is the Nyquist criterion for each system considered?*"

Response: We agree that the original statement lacked sufficient quantitative evidence. By the sentence "faithfully represent the dynamical behavior of the underlying system", we mean the empirical resolution threshold required to capture the dynamics, particularly the structure of the chaotic attractor. That is, below this sampling frequency, the sampled chaotic attractor is poorly resolved. In our study, we carefully selected the generation sampling rate Δs to ensure satisfactory resolution of the attractors for each dynamical system. For instance, in the Lorenz system, a sampling rate of $\Delta s = 0.02$ yields well-resolved attractors (as used in our study), whereas a coarser rate such as $\Delta s = 0.1$ introduces zigzag artifacts and fails to preserve the underlying dynamical structure, despite satisfying the Nyquist criterion. The statement "faithfully represent the dynamical behavior of the underlying system" is valid under full sampling. Compared to this, our framework requires only less than 20% of the samples. We have revised the sentence in the first paragraph on page 4 accordingly:

- We demonstrate that it can successfully reconstruct the dynamics of approximately three dozen prototypical nonlinear systems with high reconstruction accuracy even when the available data is only 20% of that required to faithfully represent the dynamical behavior of the underlying system.

Comment 2e: "*The authors should also discuss the relation of their ideas to those of compressed sensing.*"

Response: We thank the referee for pointing out the connection of our framework to compressed sensing (CS) that recovers a signal from a small number of observations under the assumption that the signal is sparse or compressible in a known basis (such as Fourier or wavelet). The term "sparse" in CS is referred to as the signal having only a few non-zero components after transformation into that basis. Compared with CS, our transformer-based framework learns generalizable reconstruction rules from a variety of synthetic systems, enabling it to handle unseen systems with non-sparse dynamics and severe observational sparsity, without requiring any prior sparse basis.

We have added the following description as the last paragraph on page 10 in Results 1.3:

- In addition to traditional interpolation methods, compressed sensing (CS) can also work as a signal reconstruction framework. CS assumes the the signal is sparse in a known basis and often employs optimization-based recovery [39,40]. It is important to note that the definition of the term "sparse" in CS is referred to as the signal having only a few non-zero components when expressed in an appropriate basis (e.g., Fourier or wavelet). However, the strict assumption can limit the applicability of CS. In contrast,

our method is model-free, data-driven, and capable of generalizing across unseen complex dynamical systems, regardless of the dynamics are sparse or not. Simulation results show that for a target system, when the observational sparsity is high, our framework outperforms CS significantly (Supplementary Note 10).

In Supplementary Materials, we have added a new figure Fig. S22 on page 25 to Supplementary Note 10, and a detailed description of the comparison on page 26:

- Compressed sensing (CS) is a signal reconstruction framework, which recovers signals from limited observations, provided that the signal is sparse or compressible in a known basis. It is important to note that the definition of "sparse" in CS differs from what is used in our work - it is referred to as the signal having only a few non-zero components when expressed in an appropriate basis (e.g., Fourier or wavelet). Let y_{cs} be the observed measurements, i.e., the compressed or sampled data. The goal of CS is to recover the full signal x_{cs}, assumed to be sparse in a known basis, by solving the optimization problem:

$$\min ||x_{cs}||_1 \quad \text{subject to} \quad y_{cs} = A_{cs} x_{cs}, \tag{1}$$

  where A_{cs} is a sensing matrix determined by the sampling scheme and basis. In our study, we apply CS in the discrete cosine transform (DCT) domain to reconstruct time series, from randomly sampled observations. The basic framework of CS is illustrated in Fig. S22(a).

  We evaluate the performance of CS in reconstructing dynamical systems from sparse and randomly sampled data, as tested for the transformer. Figure S22(b) demonstrates the MSE of CS and transformer reconstruction with varying sparsity S_r. It can be seen that CS can outperform the transformer under low sparsity conditions. However, as the sparsity level increases, the transformer yields more accurate reconstructions, demonstrating its stronger generalizability and recoverability under extreme observational limitations.

Comment 3: "*It is an important consideration (and limitation) that this technique seems to require that the number of variables provided to the transformer in the inference phase to be the same as those used for the testing phase. This limitation should be included in the main text, not hidden in the supplement.*"

Response: Yes, our method requires that the number of variables (i.e., system dimension) provided to the transformer model during inference match the number used during training. While the transformer architecture allows for flexible sequence lengths, it does not support varying input dimensionality across training and testing. This is a general constraint in machine-learning models and not specific to our approach. We agree that this limitation should be clearly stated in the main text, and we have added sentence to Results on pages 5-6:

- Altogether, 28 synthetic chaotic systems with same dimensions as the target systems are used to train the transformer, enabling it to learn to extract dynamic behaviors from sparse observations (see Supplementary Note 2 for a detailed description).

The following sentence has been added to the Discussion (in the first paragraph on page 13):

- It is worth noting that the dimension of the systems (i.e. the number of variables) provided to the transformer in the inference phase should match those in the testing phase.

Comment 4: "*The 28 different training systems all have associated power spectra and entropy rates. How do these power spectra compare with the test power spectra in terms of the total bandwidth? How do the entropy rates compare?*"

Response: We have calculated the PSD for the training and target systems, and added a new figure (Fig. S2). Because they all are nonlinear dynamical systems, the target systems share similar frequency components with the training systems, which may explain the ability of the transformer to generalize to the target systems. We have added the following description to Supplementary Note 3 in the last paragraph on page 5:

- To assess the frequency properties of the training and target systems more systematically, we analyze the power spectral density for each system. The calculated PSDs are shown in Fig. S2, with the black dashed vertical line in each panel indicating the dominant frequency f_d. The three colors in each panel represent the PSDs of the three state variables of the system. A common feature across all systems in both the training and testing phases is that, due to their nonlinear and chaotic nature, they exhibit broadband spectra with no dominant single frequency. We find that the target systems share similar frequency components with the training systems, providing an intuitive explanation of the ability of the transformer to generalize to the target systems.

In addition, we have calculated the entropy rates for the training and target systems, as listed in Table S3. The chaotic systems present a diverse KS entropy rates. Since each chaotic system is assigned a different sampling time Δs, chosen to ensure a smooth attractor while limiting the number of sample points, it is important to also compare the frequencies and entropy rates in a consistent manner. We normalize both the dominant frequency and the KS entropy rate by the sampling time to achieve this. Specifically, we compute the scaled dominant frequency as $f_d^s = f_d \cdot \Delta s$ and the scaled entropy rate as $h_{KS}^s = h_{KS} \cdot \Delta s$ for each system, ensuring fair comparison across systems. We have added a new figure, Fig. S3 to present the statistics of f_d^s and h_{KS}^s for both training and target systems. The results show a broad distribution across the systems, where the target systems fall within the wide range of the training systems in both measures, contributing to the strong generalizability observed.

We have added the following description in Supplementary Note 3 in the last paragraph on pages 5-6:

- In addition, since each chaotic system is assigned a different sampling time Δs, chosen to ensure a smooth attractor while limiting the number of sample points, it is important to also compare the experimental frequencies and entropy rates in a consistent manner. We normalize both the dominant frequency and the KS entropy rate by the sampling time to achieve this. Specifically, we compute the scaled dominant frequency as $f_d^s = f_d \cdot \Delta s$ and the scaled entropy rate as $h_{KS}^s = h_{KS} \cdot \Delta s$ for each system, ensuring fair comparison across systems. As depicted in Fig. S3, both f_d^s and h_{KS}^s show a broad distribution across systems. It can be seen that the target systems fall within the wide range of the training systems in both measures, contributing to the strong generalizability observed.

Comment 5: "*If you do not have many segments (which is the assumption), how do you know whether to believe either the transformer or the reservoir computer output? It is shown here that the transformer does not perform well for filtered noise. Is there some analysis that must be done to determine whether one should trust the transformer interpolation and/or the reservoir computer predictions?*"

Response: Our results have demonstrated that even under extremely sparse observations, the combination of transformer and reservoir computing can achieve satisfactory interpolation and long-term prediction. However, there are inherent constraints related to data segment length and sparsity. Specifically, if the available data is insufficient or the sparsity level exceeds a critical threshold, the framework becomes ineffective. We have reported these threshold conditions based on numerical experiments conducted on three representative target

systems. A natural and reasonable solution is to supply longer observation windows or more segments when possible. However, the main challenge lies in determining when the available data is sufficient. In our study, we systematically evaluate the performance of both the transformer and reservoir computing modules under varying sequence lengths and numbers of segments across multiple unseen target systems. The statistical results reveal consistent trends, suggesting that these insights are likely to generalize to new systems with similar sparsity levels and sequence lengths.

In fact, this challenge reflects a broader issue in time series analysis. For example, while weather forecasts can be extensively validated using historical data, it remains fundamentally uncertain whether predictions for the coming days will be accurate. Only statistical confidence can be offered.

We would also like to mention that the transformer does not perform well on stochastic signals (e.g., filtered noise), as described in Supplementary Materials. The conclusion is that the transformer is designed to extract and utilize the latent dynamics from the sparse observations and, in the absence of such dynamics, reconstruction fails. A similar requirement holds for reservoir computing: meaningful predictions are only possible when the time series itself is predictable.

Comment 6: "*The references in this manuscript are inappropriate. The fields of time series analysis, reservoir computing, and transformers are all enormous fields, yet twenty of the forty-four references in this manuscript are to papers written by the final author. For example, in referring to papers showing "that reservoir computing has the demonstrated ability to generate the long-term behavior of chaotic systems," the authors neglect to cite the seminal paper by Pathak (which is cited elsewhere in this paper as Ref. 5) Instead, five papers are cited, all by the final author.*"

Response: Our apologies! We have removed approximately half of the references by the senior author, particularly those that are only marginally related to this work. In addition, we have cited two papers by Pathak et al. to support the statement that "reservoir computing has the demonstrated ability to generate the long-term behavior of chaotic systems." However, we have retained three closely related references for this statement, due to their direct relevance. For example, Ref. [13] reported long-term prediction of chaotic systems using reservoir computing with "rare" updates from ground truth data and Ref. [21] explores the injection of noise as a hyperparameter to enhance long-term prediction performance in chaotic systems using reservoir computing.

Furthermore, we have expanded the reference list to include important works from other relevant fields such as time series analysis, sparse observations, transformers, entropy rates, compressed sensing, and related topics. Some examples are:

- On page 15: Transformers [46], originally developed for natural language processing, satisfy these requirements due to their basic attention structure. In particular, transformers has been widely applied and proven effective for time series analysis, such as prediction [47-49], anomaly detection [50], and classification [51].
- On page 9: We employ Kolmogorov-Sinai (KS) entropy [35,36], denoted as h_{KS}, to estimate the entropy rate of the dynamical systems (Methods).
- On page 10: CS assumes that the signal is sparse in a known basis and often employs optimization-based recovery [37,38].

Comment 7: *"The authors write on p. 15: "Specifically, for the chaotic food chain, Lorenz and Lotka-Volterra systems, we set $\Delta s = 10dt = 0.1$, corresponding to approximately 1 over 40 $\sim$ 50 cycles of oscillation." What does "1 over 40 50 cycles of oscillation" mean? "*

Response: The simulation time interval for each system is $dt = 0.01$. However, as the natural frequency ranges of these chaotic systems vary significantly, interpolating a system with an extremely low frequency using a small sampling interval is not particularly meaningful, as noted by the referee. Therefore, we select Δ_s as the effective sampling rate in generating each system, chosen empirically to ensure that the attractor remains sufficiently smooth while limiting the number of sampled points.

Since a common feature of chaotic systems is their broad power spectral density, it is generally not meaningful to define a precise period. Nevertheless, to provide readers with an intuitive understanding, we count the number of oscillation cycles within L_s data points and then estimate how many cycles occur per sampled point. In response to the comments of the referee, we have recalculated the frequencies and the number of oscillation cycles. We found that, for the three target systems, one point corresponds to approximately 30 to 50 cycles of oscillation. For the training systems, the range is broader, extending from approximately 20 to 130 cycles. The revised sentence in Methods 3.3 on page 18 is as follows:

- The training and testing data are obtained by sampling the time series at the interval Δ_s chosen to ensure an acceptable generation. Specifically, for the chaotic food chain, Lorenz and Lotka-Volterra systems, we set $\Delta_s = 1$, $\Delta_s = 0.02$, and $\Delta_s = 0.1$ respectively, corresponding to approximately 1 over $30 \sim 50$ cycles of oscillation. A similar procedure is also applied to other synthetic chaotic systems. (See Table S3 for Δ_s values for each system)

Summary of main changes

In response to the reviewers' new comments, we have made a number of major changes.

1. We have provided extensive new clarification and explanation of our hybrid machine learning framework for dynamics reconstruction.
2. We have elaborated on the potential of our framework as a foundation model.
3. We have demonstrated the rule for sampling rate selection and clarified the meaning of high sparsity.
4. We have revised and recalculated the entropy rate h_{KS}.

Point-by-point response to referee comments

Reviewer #1

General Comment: "*The authors have fully addressed my concerns in my review report. Also I have gone through all the responses that the authors made to the other comments in the last round of review. I think the present form could be accepted by NC.* "

Response: We thank the reviewer for recommending the paper for Nature Communications.

Reviewer #2

General Comment: "*The authors have made efforts to revise the manuscript in accordance with the reviewers' comments. They have also addressed most of my previous questions. I am now more convinced by the advancement of the method. If the authors can further address the following points, I would consider recommending it for publication. The comments below follow the order of the authors' responses.*"

Response: We are grateful to the reviewer for stating " I am now more convinced by the advancement of the method." We have addressed his/her comments and made the corresponding changes in the manuscript.

Comment 1: "*"The following analogy from natural language processing (NLP), particularly the development of large-scale models such as ChatGPT that is also based on the transformer structure, may provide some insights."*

It is not appropriate to draw an analogy between the current work and NLP, because the data in NLP are typically non-stationary. In contrast, the present problem primarily concerns stationary dynamical systems. Therefore, there is a conceptual gap between the two. If the authors insist on maintaining this analogy, it may lead to confusion or raise further questions. For instance, one may ask whether the proposed method can be extended to language-processing datasets. If the authors cannot address this discrepancy convincingly, they should remove such analogies both from the manuscript and their responses."

Response: We agree with the reviewer that drawing an analogy between our work and natural language processing (NLP) may not be appropriate due to the conceptual gap between time-series data from stationary dynamical systems and language data that are typically nonstationary. While transformers have indeed demonstrated great success in NLP tasks, they generally require substantially larger datasets and overwhelmingly more model parameters than are feasible with our available computational resources. Moreover, while our hybrid machine-learning framework shares the basic transformer architecture, its implementation details are tailored specifically to the task of dynamical systems reconstruction and differ considerably from those of large language models. In light of these, we have removed the statements on the language-processing analogy.

Comment 2: "*"In our work, the critical factor is not the number of testing cases relative to training, but rather ensuring that the training set captures a sufficiently diverse range of dynamical behaviors. Once the model has acquired this generalization ability, it can infer the governing dynamics of a variety of new systems*

from sparse observations." "Based on this, we now show that once the transformer is well-trained under a sufficiently large number of synthetic systems, it can infer the governing dynamics of an arbitrary new system from sparse observations."

This is exactly the point I wanted to raise in my previous comments. The authors have conducted further analysis with a larger number of test datasets, and their main argument is that once the training set encompasses a sufficiently diverse range of dynamical behaviors, the model can generalize effectively. In this context, I would like to ask whether the model can be treated as a foundation model—for example, can it generalize across a sufficiently large and diverse space of dynamical systems? How to justify that the trained model can truly infer the governing dynamics of "an arbitrary new" dynamical system?

If the authors' argument holds, this should indeed be possible as an extrapolation. If not, what are the key factors limiting such generalization as a foundation model? What cases are the limitations of the current method in serving as a foundation model? The authors should clarify this point explicitly."

Response: We appreciate the reviewer's great insights. In our work, the transformer is trained on a large and diverse set of chaotic dynamical systems. As shown in Fig. 4(a), we empirically observe a power-law decrease in the reconstruction error on new target systems as the number of training systems increases. This trend suggests that our hybrid machine-learning framework has the potential to acquire generalization capability and functions as a foundation model.

Previously, it was not clear whether our framework can function as a universal foundation model. To gain insights, we conducted new computational experiments on a wide variety of nonlinear dynamical systems as targets. The results revealed that our framework is able to reconstruct the dynamics of these systems when treated as unseen targets, thereby demonstrating the capability for zero-shot inference under sparse observations. This ability aligns with the expectations of a foundation model.

To further examine the limit of this generalization, we performed extensive evaluations, including tests under varying sparsity levels and sequence lengths, robustness tests, and information-theoretic analyses such as power spectral density and entropy rate. The results suggest that our framework is fully capable of implicit or behavioral inference on previously unseen systems. In fact, it utilizes global patterns from sparse observations and interpolate in a manner reflecting the true underlying dynamics. One indication of the success is the superior performance under high-sparsity conditions, where traditional interpolation methods, such as linear interpolation, spline interpolation, and compressed sensing, and standard machine learning models such as LSTM and xLSTM, fail. In addition, when applied to input data without any intrinsic deterministic dynamics (described in the counter-example in Supplementary Note 8), the framework fails to reconstruct the time series. We attribute this generalization ability to the "triple-randomness" training strategy introduced in this work, which prevents overfitting and encourages the machine-learning framework to extract patterns across diverse dynamical systems.

We acknowledge that, while our results point to the emergence of extrapolation capabilities, the theoretical foundation for this generalization remains unclear. Our findings are empirical and grounded in physical intuition, and formalizing the mechanisms responsible for the inner workings of our machine-learning framework warrants future research. We have added the following description in Discussion (on page 14):

- An open question is whether our proposed framework can function as a universal foundation model. The observed power-law relationship between the reconstruction error and training set diversity suggests its potential to function as a foundation model. In addition, extensive experiments on a broad set of target systems have demonstrated the emergence of extrapolation capabilities, where the model utilizes global

patterns in sparse observations to infer the underlying dynamics. However, establishing the theoretical base for this generalization remains challenging. Our current insights are empirical and based on physical intuition, warranting the development of a rigorous formalization of inner mechanisms of the proposed hybrid machine-learning framework.

Comment 3: "*"There are few cases where the transformer struggles to produce accurate reconstruction due to the particularly complex dynamics and extremely sparse observation." The authors should specify these cases clearly in the manuscript, such that readers can better understand the limitations of the method and the contexts in which it may fail.*"

Response: It is indeed important to specify these few cases in the manuscript to help readers better understand the limitations of the current method. We have added the following sentence at the end of the fourth paragraph in Discussion (on page 14):

- While the proposed machine-learning framework demonstrates promising performance on most of the target systems, we note that there are few cases where the transformer struggles to produce accurate reconstruction due to the particularly complex dynamics and extremely sparse observation (Supplementary Notes 11 and 12).

Remarks on code availability: "*The code is well documented.*"

Response: We are glad that the referee is happy with our code.

Reviewer $\#3$

General Comment: "*The fundamental results appear to be that transformers can provide a useful technique to interpolate chaotic time series. This result is interesting and useful, but might be better suited for a more specialized journal or for Communication Physics.*

I have one major additional comment on the manuscript: I believe that the definitions for sampling time and sparsity chosen result in substantial exaggeration of the performance of the transformer as an interpolator for chaotic time series. I think this is misleading, and I recommend re-evaluating the performance using better metrics. Please see attached pdf for detailed comments. Nevertheless, it is clear that the transformer is indeed an effective interpolator, and so this work will eventually be a valuable contribution to the nonlinear dynamics literature."

Response: We are happy that the referee stated "It is clear that the transformer is indeed an effective interpolator, and so this work will eventually be a valuable contribution to the nonlinear dynamics literature."

The definitions of sampling time and sparsity are universal, and a comparison between the performance of our transformer-based framework and that of traditional interpolation methods and other machine-learning approaches demonstrates the superiority of our framework. We provide detailed reasoning and justifications below.

Major comments: "*The authors have carefully considered the comments of the reviewers and have performed extensive revisions in light of them. The manuscript and the science therein has been improved as a result.*

However, in light of the revisions, it is clear that the results are either misunderstood or significantly overstated. In particular, the claims of performance under conditions of "extreme sparsity" are over vastly overstated (see comment 2a and 2b below). It appears that the chaotic systems are heavily oversampled and then downsampled to create artificially high sparsity.

This work is indeed interesting, but I cannot recommend publication until the claims presented in it are re-evaluated and presented accurately."

Response: A general observation was that, some excessively large sampling rate usually will result in jagged attractors unsatisfactorily representing the system dynamics, whereas an overly small sampling rate will lead to unnecessary oversampling. As such, it is necessary to select the sampling rate Δs for each system carefully, depending on its natural dominant frequency.

To clarify this point, we have produced a new figure, Fig. S2 in Supplementary Note 2, with examples of the following simulated target systems: chaotic food chain, Lorenz, and Lotka-Volterra, under varying sampling rates Δs. The orange-colored attractors represent the data ultimately used in our study. It can be seen that increasing the sampling rate leads to increasingly jagged time series and attractors, and selecting Δs that is too large contradicts our goal of reconstructing smooth dynamical systems: the interpolated or the ground-truth attractors do not resemble those in the right-most column of Fig. S2. The new content in Supplementary Note 2 (on page 4) to explain the figure reads

- When obtaining time series data from different dynamical systems, it is necessary to carefully choose the sampling rate. The goal of our work is to reconstruct the dynamics of new target systems that are smooth dynamical systems. To maintain the smoothness, selecting an appropriate sampling rate Δs for data generation is essential. An excessively large Δs can produce jagged attractors that inadequately represent the long-term dynamics, while some too-small sampling rate leads to oversampling. We determine the appropriate Δs values for each system by ensuring the attractors remain smooth while limiting the number of sampled points, as listed in Tab. S3. Figure S2 presents examples of the attractors of the simulated target systems under varying sampling rates, where the orange-colored attractors represent the data ultimately utilized in training our hybrid machine-learning framework.

In addition, we emphasized that the term "extreme sparsity" means not only sparse observational points but, more importantly, a setting in which the traditional interpolation methods (e.g., linear or cubic), engineering methods (e.g., compressed sensing), and other machine-learning methods (e.g., LSTM and xLSTM) fail to reconstruct the chaotic time series.

Comment 1: "*In response to Comment 1, the authors claim that "The reservoir computer produces more accurate reconstructed attractors [than the transformer] better capturing the underlying dynamics." This is not clear from the Figure S13. In Figure S13, neither the transformer alone nor the hybrid system deliver an especially good construction of the chaotic food chain or Lotka-Volterra system, and it appears that the transformer alone may deliver a better reconstruction of the Lorenz system (though again it is not totally clear from the figure). Can the authors try to quantify this statement with some sort of MSE or other metric? If not, it is still not clear whether the reservoir is providing any utility other than to show that the interpolation performed by the transformer is sufficiently accurate for a reservoir computer to work. It is already known that reservoir computers, given accurate training data, can replicate low dimensional chaotic attractors.*"

Response: In previous Fig. S13 (now Fig. S14 on page 18), the green attractors in the left column represent the reconstruction from the transformer alone, while the blue attractors in the right column are reconstructed from our hybrid machine-learning framework. For the chaotic food chain and Lorenz system, it is evident even visually that the attractors reconstructed by the transformer alone frequently deviate from the ground-truth attractors. Conversely, the attractors reconstructed using the hybrid method align closely with the ground truth attractors, demonstrating superior reconstruction accuracy compared to the case where only the transformer is used. For the Lotka-Volterra system, we previously clarified that the relatively poor reconstruction performance arises due to the partial usage of data - only the first three out of four dynamical variables are utilized. This is stated in Supplementary Note in the paragraph below Eq. (S3).

Quantifying performance differences via metrics such as the MSE is challenging due to the fundamentally distinct roles of the transformer and reservoir computing. Particularly, the transformer is primarily responsible for interpolation or missing data imputation, while the reservoir computer leverages the interpolated data by the transformer solely for long-term prediction. Using the MSE for long-term prediction is inappropriate because of the hallmark of chaos: sensitive dependence on initial conditions. In fact, even from the same initial condition, the machine-learning predicted evolution of the state variables will diverge from the true evolution after a few Lyapunov times.

It is worth emphasizing that the role of reservoir computing in our hybrid framework lies in illustrating that the interpolated data from the transformer is accurate and can be used for training a reservoir computer for subsequent long-term prediction of the attractor. More specifically, the workflow of our hybrid framework is as follows. Initially, sparse observations of an unseen system are interpolated by the transformer. The resulting interpolated time series, while accurate, remains limited in length. We then invoke reservoir computing to extend these accurate short series into arbitrarily long prediction, effectively converting sparse observations of limited length into long trajectories on the attractor.

We agree with the referee that the primary novelty and core contribution of our work lie in applying the transformer to interpolating unseen complex time series. The subsection on reservoir computing for long-term dynamics prediction was moved to the last subsection in Results, and the following details were moved to SI (on page 9) as the last paragraph in Supplementary Note 4:

- An assumption of our hybrid machine-learning framework is that a number of sparse data segments are available. The corresponding transformer-reconstructed segments are then used as the training data for the reservoir computer for it to find the relationship between the dynamical state at the current step and that in the immediate future. The trained reservoir computer can then predict the attractor or generate arbitrarily long time series of the target system.

Comment 2: "*In response to Comment 2a, the authors propose a new sparsity metric, S_r, that depends on the Nyquist criteria. This is an improved metric to S_m, which depends only on the sampling time Δs, which is chosen in an apparently arbitrary way. Still, S_r unnecessarily depends also on the arbitrary Δs, and I would recommend to remove this dependence by defining sparsity as*

$$S = \frac{L_s^N - L_s^O}{L_s^N}$$

More important is the fact that now that the Nyquist criterion is considered, it is clear that the transformer presented in this manuscript works only when there is not significant sparsity in an information theory sense. The evidence in the manuscript is as follows:

a) On p. 3, the authors state that "S_r = 1 corresponds to the theoretical minimum sampling case (at the Nyquist rate)" and that $S_r < 1$ corresponds to oversampling in the Nyquist sense.

b) Throughout the manuscript (e.g., on p. 8 of the main text and on p. 14 of the supplement), the authors state that the performance is poor when $S_r > 1.0$.

From these two statements, one can not conclude that the transformer works in the presence of sparsity, when sparsity is defined while considering information theory.

At issue is the selection of what is called in the manuscript as the "natural sampling rate" Δs. This quantity is apparently chosen arbitrarily so that "the attractor remains sufficiently smooth." The result of using this sampling rate rather than the Nyquist sampling rate (or something based on it) is that the systems are significantly oversampled, such that they can be downsampled by a large factor while still retaining the information content. This allows for claims of "high reconstruction accuracy even when the available data is only 20% of that required to faithfully represent the dynamical behavior of the underlying system" (p. 4) even though high performance is never obtained for $S_r > 1$.

Further evidence for this oversampling is provided, e.g, on p. 8 $L_s = 2000$ while $L_s^N = 280$. This implies that the Δs used oversamples the system by a factor of $2000/280 \approx 7$. That this system is oversampled before being downsampled is confirmed by the fact that on p. 8, S_r = 1 (implying that they are providing to the transformer exactly the amount of information required by Nyquist). Yet, the definition of sparsity using only Δs, $S_m = 0.86$, which makes it seem as if the transformer is performing with only 14% of the needed data.

I suggest that Δs not be called the "natural sampling rate" because there is nothing natural about it. It is arbitrarily (no systematic criteria is presented) chosen. The Nyquist sampling rate (or something related to it) would be more natural. Further, I suggest that any claims of high sparsity based on Δs be removed, as the apparent "sparsity" is likely due to initial oversampling. Indeed, throughout the manuscript, there are reference to conditions of "high sparsity" when $S_r < 1$. These are overstated and should be removed."

Response: We thank the reviewer for the insightful comment. Upon evaluation, we found this newly proposed metric S results in negative values that may not be appropriate for characterizing sparsity. For instance, in the chaotic food chain system, $S = -6.14$ corresponds to fully observed time series, $S = -2.57$ corresponds to 50% data sparsity ($S_m = 0.5$), while $S = -0.43$ corresponds to 80% sparsity ($S_m = 0.8$). This unbounded negative property might hinder an intuitive interpretation.

We agree with the referee that using the term "natural sampling rate" could be inappropriate. We have now simply used the term "sampling rate" instead. In general, the sampling rate needs to be selected carefully to ensure the smoothness of the attractors: the chosen Δs should be such that the reconstructed attractors are sufficiently smooth yet not over sampled, with limited data points. The new figure, Fig. S2, demonstrates this criterion - higher sampling intervals produce jagged attractors unsuitable for accurate interpolation.

Regarding the reviewer's comment on sparsity defined strictly by the Nyquist criteria, we stress that Nyquist sampling provides a theoretical baseline to ensure sufficient information. However, sampling beyond the Nyquist criterion does not inherently imply oversampling in practical interpolation contexts. Interpolation methods typically aim at reconstructing smooth trajectories rather than simply achieving the minimal Nyquist sampling. Examples can be found in the rightmost column of Fig. S2, which are not smooth and thus not accepted as the correctly interpolated attractors. To our knowledge, information theory does not explicitly define the Nyquist criterion as the definitive boundary between sparse and non-sparse sampling, notwithstanding the concept of sparsity in compressed sensing which is different. In fact, exceeding the Nyquist sampling rate does not mean oversampling, especially for reconstructing the attractors of smooth dynamical systems.

With respect to the term "high sparsity," our usage emphasizes the practical and empirical scenarios where

quite a few observational points are available and the traditional methods (e.g., linear and cubic interpolation, compressed sensing) and conventional machine-learning methods (e.g., LSTM, xLSTM) fail to reconstruct the time series. Nevertheless, we agree that our previous claim "high reconstruction accuracy even when available data is only 20% of that required to faithfully represent the dynamical behavior of the underlying system" is inappropriate. It has now been removed.

Comment 2b: "*The additional work in response to comment 2b is convincing and appreciated. However, in light of the above comment, it is not clear that this method helps with (information) sparsity. It is clear that it works with irregular sampling (which is useful in its own right). If the authors would like to claim that their method works under conditions of high sparsity, I suggest that they choose conditions that are actually sparse according to their own (new) metric S_r.*"

Response: We appreciate the referee's statement "the additional work in response to comment 2b is convincing and appreciated". We emphasize that our concept of high sparsity means not only quite a few available observations but also the conditions under which traditional reconstruction methods fail. We have added the following statement to Sec. 1.3 in Results (at the end of page 10):

- Hereafter, the term "high sparsity" is used to mean not only a limited number of available observations but also the failure to reconstruct the system accurately at this level of sparsity by the traditional methods such as linear and spline interpolation, compressed sensing, and conventional machine-learning techniques.

Comment 2c: "*I do not think that the entropy rate is being computed accurately here, as the results do not make sense. As the sparsity of information input to the transformer is increased (as S_r increases), the quantity of information per unit time (i.e., the entropy rate) input into the transformer necessarily decreases. However, the bars labeled "Observations" in Fig. 4b increase with S_r. This cannot be correct. I suggest the authors revisit this computation.*"

Response: We have revisited our entropy-rate calculations with two improvements. (1) Previously, we removed time steps containing missing values in any variable before calculating the KS entropy of the inputs. For a fair comparison, we have now omitted this preprocessing step for the inputs. (2) Initially, we defined the entropy rate h_{KS} as $(1/\tau_{\min}) \cdot \sum_\tau \rho(\tau) \log(1/\rho(\tau))$. We reexamined this definition and concluded that using the mean recurrence time $\langle\tau\rangle$ in the denominator provides a more representative measure than using the minimum recurrence time $\tau_{\min}$ so as to account for all the recurrences.

With the revised definition, we recalculated h_{KS} for the transformer inputs, outputs, and the ground truth across varying sparsity levels, and updated panel (b) of Fig. 4 (on page 9) accordingly. It is worth noting that, even with the improvement, the calculated input entropy rate still increases with S_r. Through a detailed analysis, we have identified the main reason as the denominator used in calculating h_{KS}. Regardless of whether τ_{min} or $\langle\tau\rangle$ is used, the sparse input yields a small denominator due to the prevalent zero values, making the "recurrence" occurrences artificially frequent and resulting in an increased entropy-rate value.

We have removed the descriptions of input preprocessing in Sec. 1.3 in Results and revised the description in Sec. 3.7 of Methods (in the first paragraph on page 21), which reads

- The entropy rate is then estimated by the following formula [35]:

$$h_{KS} \approx \frac{1}{\langle\tau\rangle} \sum_\tau \rho(\tau) \log(\frac{1}{\rho(\tau)}),$$

where $\langle\tau\rangle$ is the mean observed recurrence time.

In addition, we have recalculated the h_{KS} values for all the systems and updated them in Tab. S3. Panel (b) of Fig. S4 has also been updated. These results illustrate that revised calculations do not alter the conclusion.

Comment 2d: "*The first paragraph on p. 4 now reads: "We demonstrate that it can successfully reconstruct the dynamics of approximately three dozen prototypical nonlinear systems with high reconstruction accuracy even when the available data is only 20% of that required to faithfully represent the dynamical behavior of the underlying system."*

This 20% value depends on Δs which is apparently chosen arbitrarily (in the sense that no systematic definition of it that depends on the actual sampled information is presented). In other words, this seemingly incredible 20% value is a result of choosing Δs to be too small such that the system is oversampled. This statement is a significant exaggeration and should be removed entirely."

Response: We have removed the statement related to 20%, and now the sentence in the first paragraph on page 4 reads

- We demonstrate that it can successfully reconstruct the dynamics of approximately three dozen prototypical nonlinear systems with high reconstruction accuracy.

Comment 2e: "*The additional work to compare transformer with compressed sensing is interesting and adds to the work significantly. It shows that the story of this transformer has much more depth than was presented previously.*

However, this depth is completely hidden in the supplement. The main text simply states that "Simulation results show that for a target system, when the observational sparsity is high, our framework outperforms CS significantly (Supplementary Note 10)."

This is perhaps true, but certainly misleading. When the sparsity is low to moderate (< 0.75 or so), compressed sensing performs better than the transformer and is likely to be computationally less expensive to train and operate. When the sparsity is high (> 1), both the transformer and compressed sensing perform poorly, though the transformer performs better. Only in a narrow range of sparsities does the transformer work well when the compressed sensing does not. This should be reflected in the discussion of the compressed sensing results so as not to overstate the performance of the transformer."

Response: We thank the referee for stating "The additional work to compare transformer with compressed sensing is interesting and adds to the work significantly." We agree that the comparison deserves a more thorough discussion in the main text. The relevant statement in the last paragraph of Sec. 1.3 (on page 10) has been revised to

- Simulation results show that, for a target system, when the observational sparsity is low to moderate, compressed sensing performs better than the transformer. However, when the observational sparsity is high, our hybrid machine-learning framework outperforms compressed sensing significantly (Supplementary Note 10).

In addition, we would like to clarify that sparsity S_r does not vary linearly with the sampling rate. Instead, as Δs increases linearly, the resulting sparsity increases in a concave manner. That is, a small increase in Δs at low values leads to a large drop in the data availability. For example, in the chaotic food chain system, as shown

in Fig. S2 (a), increasing Δs from 1 to 2 reduces the data availability by 50% (corresponding to $S_r = 0.58$), while increasing Δs further to 7 results in $S_r = 1$ (only a 0.42 increment). This demonstrates that the the range between $S_r = 0.75$ and $S_r = 1$ gives significant sparsity. We have added the following description to Supplementary Note 2 (on pages 4-5):

- It is worth noting that the the third column of Fig. S2 corresponds to doubling the selected sampling rate, which reduces the data availability to approximately 50%. The last column shows that a sampling rate 7-8 times higher than the selected rate yields about 14% data availability, which is consistent with the Nyquist criterion ($S_r = 1$). This demonstrates a nonlinear (concave) relationship between the sampling rate and sparsity. For example, in the chaotic food chain system, increasing Δs from 1 to 2 reduces the data availability to 50% (corresponding to $S_r = 0.58$), whereas further increasing it to 7 results in $S_r = 1$ (only a 0.42 increment).

Comment 3: "*Comments 3, 4 and 6 have been addressed clearly, and the associated changes have improved the manuscript in my opinion. Comment 5 has essentially not been addressed but is a difficult question. I suggest that the authors make a comment in the discussion or conclusion about the difficulty of validating the predictions of their model and leave it to future work.*"

Response: We thank the referee for the positive evaluation of our previous Response. We have added the following description in the first paragraph of Discussion (on page 14):

- Moreover, there are constraints related to data segment length and sparsity. Specifically, performance deteriorates if the available data is insufficient or the sparsity exceeds a critical threshold. Determining these thresholds for a new target system is challenging. A reasonable solution is to supply longer observation windows or more segments when possible. In this work, we have provided such conditions through extensive experiments and the statistical results consistently reveal trends that are likely to generalize to other systems with similar levels of sparsity and sequence length. It is worth mentioning that this challenge in fact emerges commonly in time series analysis. For example, while weather forecasts can be extensively validated using historical data, it remains fundamentally uncertain whether predictions for the coming days will be accurate. Only statistical confidence can be offered.

Summary of main changes

In response to the reviewers' new comments, we have made a number of changes.

1. New clarification and explanation of our proposed dynamics reconstruction framework have been provided.
2. The generalization ability of our framework for dynamics reconstruction have been further discussed.
3. The meanings of smoothness and sparsity have been clarified and demonstrated.
4. The description of the entropy rate h_{KS} has been removed.

Point-by-point response to reviewer comments

Reviewer #2

General Comment: "*I appreciate the authors' response to my comments and the revisions made to the manuscript. The authors have made commendable progress during the revision, including the addition of computational experiments on more nonlinear dynamical systems. The study presents an interesting approach; however, as demonstrated below, I remain uncertain about whether it reaches the conceptual depth expected for publication in Nature Communications.*

In particular, the theoretical foundation for the generalization of the proposed method remains unclear, as acknowledged by the authors themselves. There is still no strong evidence that the method can serve as a foundation model or be reliably extrapolated to a broad class of previously unseen dynamical systems, even with the added numerical experiments. In other words, although the method demonstrates clear technical progress, it falls short of delivering a substantive conceptual breakthrough. This concern, also noticed by Reviewer 3, limits the perceived scope and general impact of the work.

Given these considerations, and should other reviewers express similar reservations, I would prefer to recommend that the manuscript might be better considered for publication in a more specialized journal."

Response: We thank the reviewer for his/her positive comments: "The authors have made commendable progress during the revision," and "the study presents an interesting approach."

Regarding the reviewer's concern about the conceptual depth, we fully agree that a rigorous theoretical framework would be a valuable asset. However, it is important to recognize that this limitation is not unique to our work. In fact, the machine-learning community has long acknowledged that a theoretical understanding of deep learning models, especially regarding generalization, remains underdeveloped. Classical statistical learning theory (e.g., VC dimension, bias-variance trade-off) has proven inadequate in explaining the empirical success of over parameterized deep neural networks [43,44].

This gap is particularly evident in the domain of foundation models and their derivatives. As stated in the Stanford CRFM report [45], the community currently has a quite limited theoretical understanding of foundation models, despite their transformative capabilities in applications such as natural language processing (e.g., GPT-3) and computer vision (e.g., DALL-E). These models exhibit remarkable generalization and adaptation across tasks, yet existing theoretical tools fall short of explaining their behavior. A simple Google Scholar search for "foundation model" combined with "deep learning" yields approximately 6,290,000 results; yet, according to the report we mentioned, none succeed in providing a theoretical explanation for their generalization capabilities. Nonetheless, the lack of a formal theory has not hindered the widespread success and adoption of these deep learning methods.

In our work, we demonstrate the generalization capabilities of our framework to form a foundation model through extensive computational experiments and physical insights. Specifically, we statistically evaluated the performance across a broad range of previously unseen target systems and compared our method against both classical interpolation techniques and other machine learning approaches. We also identified a power-law relationship between the target error and the number of training systems, which is a standard indicator of generalization potential in foundation models. In addition, we presented supporting physical intuition, including a counter-example, and conducted information-theoretic analyses such as power spectral density and dominant frequency characterizations. These elements collectively aim to either indicate or provide insight into the generalization behavior of our framework, as the reviewer noted "the method demonstrates clear technical progress."

Moreover, we note that reviewer 3 thinks that " transformers can provide a useful technique to interpolate chaotic time series" while not requiring a theoretical breakthrough.

We would like to emphasize that our contribution lies in proposing a generalizable foundation model for dynamical systems, bridging empirical capability and cross-system adaptability, rather than developing a theoretical framework from first principles, which remains a significant open challenge in the machine-learning field.

We have revised the following description in Discussion (the first paragraph on page 14):

- A question is whether our proposed framework can function as a universal. The observed power-law relationship between the reconstruction error and training set diversity suggests its potential to function as a foundation model. In addition, extensive experiments on a broad set of target systems have demonstrated the emergence of extrapolation capabilities, where the model utilizes global patterns in sparse observations to infer the underlying dynamics. However, establishing the theoretical base for this generalization remains challenging. Classical statistical learning theory (e.g., VC dimension, bias-variance trade-off) has proven inadequate in explaining the empirical success of over parameterized deep neural networks [43,44]. As stated in the Standford CRFM report [45], the community currently has a quite limited theoretical understanding of foundation models. Our current insights are empirical and based on physical intuition, warranting the development of a rigorous formalization of the inner mechanisms of the proposed hybrid machine-learning framework.

Reviewer $\#3$

General Comment: "*The authors have addressed my concerns about the presentation of their results sufficiently. The claims presented in the manuscript are now supported by the data presented. I still think that the fundamental result is that transformers can provide a useful technique to interpolate chaotic time series. The authors appear to agree on this point. In my opinion, this result is interesting and potentially useful, but I still feel it might be better suited for a more specialized journal or for Communication Physics.*

From a technical point of view, I can recommend publication of this manuscript. I hope the authors will consider the points raised in the attached pdf in a final revision, as I think they will further improve the manuscript."

Response: We thank the reviewer for recommending the publication of our manuscript in *Nature Communications* and for his/her encouraging comments "The authors have addressed my concerns about the presentation of their results sufficiently" and "this result is interesting and potentially useful." We are grateful to the reviewer for devoting tremendous time and effort in evaluating our work.

Comment 1: "*Authors: "A general observation was that, some excessively large sampling rate usually will result in jagged attractors unsatisfactorily representing the system dynamics, whereas an overly small sampling rate will lead to unnecessary oversampling. As such, it is necessary to select the sampling rate* Δs *for each system carefully, depending on its natural dominant frequency. To clarify this point, we have produced a new figure, Fig. S2 in Supplementary Note 2, with examples of the following simulated target systems: chaotic food chain, Lorenz, and Lotka-Volterra, under varying sampling rates* Δs*. The orange-colored attractors represent the data ultimately used in our study. It can be seen that increasing the sampling rate leads to increasingly jagged time series and attractors, and selecting* Δs *that is too large contradicts our goal of reconstructing smooth dynamical systems: the interpolated or the ground-truth attractors do not resemble those in the rightmost column of Fig. S2. "*

(in their response to Comment 2): "In general, the sampling rate needs to be selected carefully to ensure the smoothness of the attractors: the chosen Δs *should be such that the reconstructed attractors are sufficiently smooth yet not over sampled, with limited data points. The new figure, Fig. S2, demonstrates this criterion - higher sampling intervals produce jagged attractors unsuitable for accurate interpolation."*

Reviewer Comment: The authors seem to be concerned with an arbitrary concept of smoothness (here and also in their response to Comment 2), which they assert is somehow important but they do not explain why, nor do they quantify it. They state without evidence or citation that "An excessively large Δs *can produce jagged attractors that inadequately represent the long-term dynamics," and that "higher sampling intervals produce jagged attractors unsuitable for accurate interpolation." However, if one considers a sine wave with 5 samples per period, it will appear "jagged" to the eye, but it is sufficiently sampled to later on produce an accurate reconstruction with standard interpolation techniques. Indeed, I strongly suspect that the orange attractors shown in Fig. S2 are oversampled precisely because they appear very smooth to the eye.*

Of course there is nothing wrong with oversampling if one wants an attractive visualization. The issue I am raising is that using the oversampled sampling rate in the definition of sparsity allows one to claim incredible performance in terms of "sparsity" when the information itself is not actually sparse. See also the related example about Fourier interpolation in a comment below."

Response: By "smoothness," we mean not only the smooth visual appearance but also the continuity and bound of the derivative. Specifically, we ensure that the norm of the first derivative of the chaotic signals

remains below a predefined small threshold. We apologize for omitting this clarification in the previous version and have now added the following statement to the main text (on page 3):

- Specifically, the norm of the first derivative of the chaotic signals is ensured to stay below a predefined small threshold.

Regarding the reviewer's example of a sinusoidal signal, we wish to emphasize the difference between periodic and chaotic signals. In terms of the power spectrum, the maximum frequency of a periodic signal is finite, so the Nyquist frequency can naturally be defined. A chaotic system typically exhibits broad and continuous power spectral density, spanning a wide and potentially infinite range, rendering difficult to theoretically define the Nyquist frequency. Empirically, it is often possible to estimate an effective "cutoff" frequency to approximate the spectral support of the chaotic signal, but this is inherently a simplification.

We would like to clarify the role of the "orange" attractors in Fig. S2. These attractors represent the ground-truth trajectories, i.e., the target smooth signals we aim to recover. They are not the input data to the machine, i.,e., the machine-learning model does not receive these orange trajectories as input (unless the data is fully observed, i.e., with zero sparsity). Instead, the inputs typically consist of randomly and sparsely sampled (and visually jagged) trajectories, and our framework is designed to reconstruct a smooth ground trajectory from them. In this context, the example of a "jagged" sine wave reconstructed via interpolation is entirely consistent with our goal: a smooth sinusoidal wave is needed to compare with the interpolated one for performance evaluation. To clarify this, we have added the following description to Supplementary Note 2 (on pages 4-5):

- Note that that the orange-colored attractors represent the ground-truth trajectories, i.e., the target smooth signals we aim to recover. They are not the input data to the machine: the machine-learning model does not receive these orange trajectories as input (unless the data is fully observed, i.e., with zero sparsity). Instead, the inputs typically consist of randomly and sparsely sampled (and visually jagged) trajectories, and our framework is designed to reconstruct the smooth ground-truth trajectory from them.

Importantly, our work demonstrates that for chaotic systems, standard interpolation techniques often fail to produce accurate reconstructions under high sparsity, due to the irregularity and broadband nature of the input data. Our transformer-based method is specifically designed to handle these challenges with demonstrated success in recovering smooth, high-fidelity trajectories where classical and other machine-learning methods fail. We believe this addresses the question raised by the reviewer, as the "incredible performance" of our framework under high sparsity arises from both the limited observational data and the failures of previous approaches.

Comment 2: "*Authors: "In previous Fig. S13 (now Fig. S14 on page 18), the green attractors in the left column represent the reconstruction from the transformer alone, while the blue attractors in the right column are reconstructed from our hybrid machine-learning framework. For the chaotic food chain and Lorenz system, it is evident even visually that the attractors reconstructed by the transformer alone frequently deviate from the ground-truth attractors. Conversely, the attractors reconstructed using the hybrid method align closely with the ground truth attractors, demonstrating superior reconstruction accuracy compared to the case where only the transformer is used. For the Lotka-Volterra system, we previously clarified that the relatively poor reconstruction performance arises due to the partial usage of data - only the first three out of four dynamical variables are utilized. This is stated in Supplementary Note in the paragraph below Eq. (S3). Quantifying performance differences via metrics such as the MSE is challenging due to the fundamentally distinct roles of the transformer and reservoir computing. Particularly, the transformer is primarily responsible for interpolation or*

missing data imputation, while the reservoir computer leverages the interpolated data by the transformer solely for long-term prediction. Using the MSE for long-term prediction is inappropriate because of the hallmark of chaos: sensitive dependence on initial conditions. In fact, even from the same initial condition, the machine-learning predicted evolution of the state variables will diverge from the true evolution after a few Lyapunov times."

Comment: From the figures presented here, it still seems to me that In Figure S13, neither the transformer alone nor the hybrid system deliver an especially good construction of the chaotic food chain or Lotka-Volterra system, and it appears that the transformer alone may deliver a better reconstruction of the Lorenz system (though it is not totally clear from the figure). However, since the authors do not attempt to quantify agreement in any way, the reader is left no choice but to take their word for it."

Response: The suboptimal performance on the Lotka-Volterra system arises from discarding the information in the last dimension, as we previously clarified. For the food chain and Lorenz systems, the hybrid method performs better than the transformer alone. However, a different conclusion may be drawn based solely on visualization. To avoid confusion, we have removed the statement claiming that the reservoir computing produces more accurate or stable reconstructions. Instead, we emphasize its capacity to generate arbitrarily long dynamics based on the output of the transformer.

Comment 3: "*Authors: "It is worth emphasizing that the role of reservoir computing in our hybrid framework lies in illustrating that the interpolated data from the transformer is accurate and can be used for training a reservoir computer for subsequent long-term prediction of the attractor. More specifically, the workflow of our hybrid framework is as follows. Initially, sparse observations of an unseen system are interpolated by the transformer. The resulting interpolated time series, while accurate, remains limited in length. We then invoke reservoir computing to extend these accurate short series into arbitrarily long prediction, effectively converting sparse observations of limited length into long trajectories on the attractor. We agree with the reviewer that the primary novelty and core contribution of our work lie in applying the transformer to interpolating unseen complex time series. The subsection on reservoir computing for long-term dynamics prediction was moved to the last subsection in Results, and the following details were moved to SI (on page 9) as the last paragraph in Supplementary Note 4:"*

Comment: I think we are in agreement on this point now. Using a reservoir computer to perform attractor reconstruction from a time series is a well-established method, and its only purpose here is to qualitatively show that the transformer interpolation is 'good enough.'"

Response: We are happy that the reviewer agreed with us on this point.

Comment 4: "*Authors: "We thank the reviewer for the insightful comment. Upon evaluation, we found this newly proposed metric S results in negative values that may not be appropriate for characterizing sparsity. For instance, in the chaotic food chain system, $S = -6.14$ corresponds to fully observed time series, $S = -2.57$ corresponds to 50% data sparsity ($S_m = 0.5$), while $S = -0.43$ corresponds to 80% sparsity ($S_m = 0.8$). This unbounded negative property might hinder an intuitive interpretation.*"

Comment: This is a good point. An intuitive interpretation is also important."

Response: We thank the reviewer for the positive evaluation of our analysis.

Comment 5: "*Authors: "We agree with the referee that using the term "natural sampling rate" could be inappropriate. We have now simply used the term "sampling rate" instead. In general, the sampling rate needs to be selected carefully to ensure the smoothness of the attractors: the chosen Δs should be such that the reconstructed attractors are sufficiently smooth yet not over sampled, with limited data points. The new figure, Fig. S2, demonstrates this criterion - higher sampling intervals produce jagged attractors unsuitable for accurate interpolation."*

"Regarding the reviewer's comment on sparsity defined strictly by the Nyquist criteria, we stress that Nyquist sampling provides a theoretical baseline to ensure sufficient information. However, sampling beyond the Nyquist criterion does not inherently imply oversampling in practical interpolation contexts. Interpolation methods typically aim at reconstructing smooth trajectories rather than simply achieving the minimal Nyquist sampling. Examples can be found in the rightmost column of Fig. S2, which are not smooth and thus not accepted as the correctly interpolated attractors."

Comment: Indeed, interpolation methods aim at reconstructing smooth trajectories from "jagged" time series measurements. The "true trajectory" is smooth. If the time series measurement is already "smooth," then there is no need for interpolation.

Again, a sine wave that is sampled 5 times per period will not appear smooth to the eye, and yet a simple Fourier interpolation will be able to reconstruct the smooth sine wave from it. One may need to sample that sine wave 50 times per period for the time series to appear "smooth" – yet it would seem incorrect to claim that Fourier interpolation can achieve accurate reconstruction of the sine wave from measurements with 90% sparsity."

Response: We completely agree with the reviewer that "If the time series measurement is already "smooth," then there is no need for interpolation." As we stated earlier, the orange attractors in Fig. S3 is the smooth ground-truth trajectory. The model is provided with time series data of varying sparsity levels, including highly jagged observations, and the goal is to reconstruct the underlying smooth trajectory. In fact, a key feature of our proposed framework is its flexibility in handling a wide range of sparsity levels, including highly sparse, randomly sampled data from previously unseen systems. For a thorough evaluation, we tested the model under different sampling conditions, from fully observed (smooth) trajectories to gradually more sparse (more jagged) inputs, and observed superior performance, as detailed in the main text and illustrated in Fig. 4.

The reviewer is correct in noting that a sinusoidal wave sampled five times per period can be accurately reconstructed via Fourier interpolation. However, this principle does not extend directly to chaotic systems because of their broad (in principle infinite) frequency bands. As discussed in our response to Comment 1, chaotic time series exhibit broad and continuous power spectral densities. If one were to apply a sliding window and approximate the dominant frequency within each window, the instantaneous frequency content would vary significantly across time. In the high-frequency windows, sampling five points may be insufficient, and random sampling can make the situation worse. This variability, combined with the random sparsity patterns, make Fourier interpolation or other classical methods inappropriate, thereby justifying our proposed framework. Furthermore, we clarify that the notion of "sparsity" in our manuscript is defined in a relative sense. The term "high sparsity" is used to denote conditions under which baseline interpolation methods perform poorly, whereas our framework continues to produce accurate reconstruction.

We have added the following description to the Supplementary Note 3 (on pages 6-7) to address the referee's concern:

- In addition, we emphasize that, for chaotic systems that typically exhibit broad and continuous power spectral densities, if one were to apply a sliding window and approximate the dominant frequency within

each window, the instantaneous frequency content would vary significantly across time. In high-frequency windows, sampling several points may not be sufficient, and random sampling worsens the situation. This variability, in combination with the random sparsity patterns, make Fourier interpolation or other classical methods inappropriate, providing further justification for our proposed framework.

Comment 6: *"Authors: "To our knowledge, information theory does not explicitly define the Nyquist criterion as the definitive boundary between sparse and non-sparse sampling, notwithstanding the concept of sparsity in compressed sensing which is different. In fact, exceeding the Nyquist sampling rate does not mean oversampling, especially for reconstructing the attractors of smooth dynamical systems."*

Comment: This is exactly what the word "oversampling" means according to any standard digital signal processing textbook. For example, on p. 585 of the textbook Digital Signal Processing by Li Tan and Jean Jiang (2013, 2nd edition, ISBN: 978-0-12-415893-1), it is stated that: "Oversampling uses a sampling rate that is much higher than the Nyquist rate. We can define an oversampling ratio as $f_s/2f_{\max} \gg 1$ *" "*

Response: We thank the reviewer for pointing out the reference to the definition of oversampling. Having looked into it, we are more confident that the Nyquist criterion is not the definitive threshold for determining oversampling. The term "much higher" is qualitative and does not provide a strict quantitative threshold that clearly separates standard sampling from oversampling. This raises a natural question: how high must the number $f_s/2f_{\max}$ to be to qualify as "much" higher than one? The referenced textbook provides only two illustrative examples (Examples 12.7 and 12.8, on page 588 of the mentioned textbook). In the first example, the sampling rate f_s is 640 kHz and the maximum frequency $f_{\max}$ is 20 kHz, resulting in a ratio of $f_s/(2f_{\max}) = 16$. In the second example, the ratio is $f_s/(2f_{\max}) = 80\ \text{MHz}/(2 \cdot 4\ \text{kHz}) = 10,000$.

In our case, the estimated ratio of $f_s/2f_{\max}$ for the ground-truth trajectories (orange attractors in Fig. S3) is approximately 7, which is significantly below the oversampling levels used in examples in the textbook, which has now been cited as Ref. [27] in the main text.

Comment 7: *"Authors: "With respect to the term "high sparsity," our usage emphasizes the practical and empirical scenarios where quite a few observational points are available and the traditional methods (e.g., linear and cubic interpolation, compressed sensing) and conventional machine-learning methods (e.g., LSTM, xLSTM) fail to reconstruct the time series. Nevertheless, we agree that our previous claim "high reconstruction accuracy even when available data is only 20% of that required to faithfully represent the dynamical behavior of the underlying system" is inappropriate. It has now been removed."*

"We appreciate the referee's statement "the additional work in response to comment 2b is convincing and appreciated". We emphasize that our concept of high sparsity means not only quite a few available observations but also the conditions under which traditional reconstruction methods fail. We have added the following statement to Sec. 1.3 in Results (at the end of page 10): Hereafter, the term "high sparsity" is used to mean not only a limited number of available observations but also the failure to reconstruct the system accurately at this level of sparsity by the traditional methods such as linear and spline interpolation, compressed sensing, and conventional machine learning techniques. "

Comment: The paper has been improved by the clarification that by "high sparsity" the authors are not referring to sparsity of information but rather to the fact that their method is better at interpolation than other methods. In my opinion, sparsity is not the correct word for this, but it is sufficient that the authors have clarified what they mean."

Response: We thank the reviewer for the positive evaluation of our revision. We are glad that the reviewer stated "it is sufficient that the authors have clarified what they mean."

Comment 8: "*Authors: Response to Comment 2c: "I do not think that the entropy rate is being computed accurately here, as the results do not make sense. As the sparsity of information input to the transformer is increased (as S_r increases), the quantity of information per unit time (i.e., the entropy rate) input into the transformer necessarily decreases. However, the bars labeled "Observations" in Fig. 4b increase with S_r. This cannot be correct. I suggest the authors revisit this computation."*

Response: We have revisited our entropy-rate calculations with two improvements. (1) Previously, we removed time steps containing missing values in any variable before calculating the KS entropy of the inputs. For a fair comparison, we have now omitted this preprocessing step for the inputs. (2) Initially, we defined the entropy rate h_{KS} as $(1/\tau_{\min})\, P(\tau)\rho(\tau)\log(1/\rho(\tau))$. We reexamined this definition and concluded that using the mean recurrence time $\langle\tau\rangle$ in the denominator provides a more representative measure than using the minimum recurrence time $\tau_{\min}$ so as to account for all the recurrences. With the revised definition, we recalculated h_{KS} for the transformer inputs, outputs, and the ground truth across varying sparsity levels, and updated panel (b) of Fig. 4 (on page 9) accordingly. It is worth noting that, even with the improvement, the calculated input entropy rate still increases with S_r. Through a detailed analysis, we have identified the main reason as the denominator used in calculating h_{KS}. Regardless of whether $\tau_{\min}$ or $\langle\tau\rangle$ is used, the sparse input yields a small denominator due to the prevalent zero values, making the "recurrence" occurrences artificially frequent and resulting in an increased entropy-rate value. We have removed the descriptions of input preprocessing in Sec. 1.3 where $\langle\tau\rangle$ is the mean observed recurrence time. In addition, we have recalculated the h_{KS} values for all the systems and updated them in Tab. S3. Panel (b) of Fig. S4 has also been updated. These results illustrate that revised calculations do not alter the conclusion."

Comment: I still do not think that the entropy rate is being computed accurately here, as the results do not make sense. As the sparsity of information input to the transformer is increased (as S_r increases), the quantity of information per unit time (i.e., the entropy rate) input into the transformer necessarily decreases. However, the bars labeled "Observations" in Fig. 4b still increase with S_r. This cannot be correct. I think the manuscript would be significantly improved by an accurate estimate of these entropy rates-it would be the best way to quantify the sparsity of information. If for some reason this is not possible, I suggest the authors remove these computations-or in the interest of transparency leave them in the supplement but note that they cannot be correct-since presenting clearly incorrect entropy-rate estimates will only spark confusion and misunderstanding."

Response: We have carefully revisited our entropy-rate calculations and confirmed that the computations were implemented correctly. As we previously explained, the observed increase in the entropy rate with increasing sparsity S_r arises from the denominator in the definition, specifically, from the use of the mean recurrence time $\langle\tau\rangle$. With higher sparsity, the number of zero (missing) values increases, causing the recurrence detection to trigger more frequently, which leads to a smaller $\langle\tau\rangle$ value and therefore an over estimated entropy rate.

However, we agree with the reviewer's concern that this result might contradict intuitive expectations, where increased sparsity should correspond to decreased information content. We have removed the entropy-rate analysis from the main text and Supplementary Material. Panel (b) of Fig. 4 has been replaced by a panel on the analysis of scaled dominant frequencies across systems (originally shown in Fig. S4 in the previous version, now removed). We have also updated the corresponding description in the main text, as follows (the last paragraph on page 9):

- To assess the frequency properties of the training and target systems more systematically, we analyze the power spectral density (PSD) for each system. The calculated PSDs are shown in Supplementary Note 3. Since each chaotic system is assigned a different sampling time Δs, chosen to ensure a smooth attractor while limiting the number of sample points, it is important to also compare the experimental frequencies in a consistent manner. We normalize both the dominant frequency by the sampling time to achieve this. Specifically, we compute the scaled dominant frequency as $f_d^s = f_d \cdot \Delta s$ for each system, ensuring a fair comparison across the systems. As depicted in Fig. 4(b), f_d^s exhibits a broad distribution across systems. It can be seen that the target systems fall within the wide range of the training systems in both measures, contributing to the strong generalizability observed.

For the sake of transparency, we retained the entropy-rate computation scripts in the publicly available GitHub repository.

Comment 9: "*Authors: "We have removed the statement related to 20%, and now the sentence in the first paragraph on page 4 reads. We demonstrate that it can successfully reconstruct the dynamics of approximately three dozen prototypical nonlinear systems with high reconstruction accuracy. "*

Comment: I believe the clarity and accuracy of the paper has been improved by this change."

Response: We thank the reviewer for the previous suggestion, which improved the clarity and accuracy of our paper.

Comment 10: "*Authors: "We thank the referee for stating "The additional work to compare transformer with compressed sensing is interesting and adds to the work significantly." We agree that the comparison deserves a more thorough discussion in the main text. The relevant statement in the last paragraph of Sec. 1.3 (on page 10) has been revised to*

Simulation results show that, for a target system, when the observational sparsity is low to moderate, compressed sensing performs better than the transformer. However, when the observational sparsity is high, our hybrid machine-learning framework outperforms compressed sensing significantly (Supplementary Note 10).

In addition, we would like to clarify that sparsity Sr does not vary linearly with the sampling rate. Instead, as Δs increases linearly, the resulting sparsity increases in a concave manner. That is, a small increase in Δs at low values leads to a large drop in the data availability. For example, in the chaotic food chain system, as shown 9 in Fig. S2 (a), increasing Δs from 1 to 2 reduces the data availability by 50% (corresponding to $S_r = 0.58$), while increasing Δs further to 7 results in $S_r = 1$ (only a 0.42 increment). This demonstrates that the the range between $S_r = 0.75$ and $S_r = 1$ gives significant sparsity. We have added the following description to Supplementary Note 2 (on pages 4-5):

It is worth noting that the the third column of Fig. S2 corresponds to doubling the selected sampling rate, which reduces the data availability to approximately 50%. The last column shows that a sampling rate 7-8 times higher than the selected rate yields about 14% data availability, which is consistent with the Nyquist criterion ($S_r = 1$). This demonstrates a nonlinear (concave) relationship between the sampling rate and sparsity. For example, in the chaotic food chain system, increasing Δs from 1 to 2 reduces the data availability to 50% (corresponding to S_r = 0.58), whereas further increasing it to 7 results in $S_r = 1$ (only a 0.42 increment)."

Comment: These additions have improved the clarity of the manuscript"

Response: We appreciate the reviewer for the positive evaluation of our revision.

Point-by-point response to reviewer comments

Reviewer $\#3$

General Comment: "*The authors are playing with semantics to claim high sparsity when they are clearly oversampling, even by the textbook definition, which they misinterpret. They use this oversampling to claim incredible performance in terms of sparsity, when in reality the performance of the transformer is better than basic techniques only in a small region of sparsity, as discussed in previous rounds of review.*

For the reasons above, I continue to recommend publication in a specialized journal. The most exaggerated claims have been removed from the manuscript at this point, and so from a technical point of view, the paper can be published."

Response: We thank the reviewer for reading the revised manuscript again and for his/her recommendation. We have further tuned down the main claim in Abstract, Introduction, and Discussion, as follows.

In Abstract, the penultimate sentence has been revised to

- The proposed hybrid machine-learning framework is tested using various prototypical nonlinear systems, demonstrating that the dynamics can be faithfully reconstructed from reasonably sparse data.

The last sentence in the penultimate paragraph in Introduction has been changed to:

- We demonstrate that it can successfully reconstruct the dynamics of approximately three dozen prototypical nonlinear systems with high reconstruction accuracy, providing a viable framework for reconstructing complex and nonlinear dynamics in situations where training data from the target system do not exist and the observations or measurements are insufficient.

The first sentence in the last paragraph in Discussion has been rewritten as

- Overall, our hybrid transformer/reservoir-computing framework has been demonstrated to be effective for dynamics reconstruction and prediction of long-term behavior in situations where only sparse observations from a newly encountered system are available.

The authors have carefully considered the comments of the reviewers and have performed extensive revisions in light of them. The manuscript and the science therein has been improved as a result.

However, in light of the revisions, it is clear that the results are either misunderstood or significantly overstated. In particular, the claims of performance under conditions of "extreme sparsity" are over vastly overstated (see comment 2a and 2b below). It appears that the chaotic systems are heavily oversampled and then downsampled to create artificially high sparsity.

This work is indeed interesting, but I cannot recommend publication until the claims presented in it are re-evaluated and presented accurately.

Comment 1:
In response to Comment 1, the authors claim that "The reservoir computer produces more accurate reconstructed attractors [than the transformer] better capturing the underlying dynamics." This is not clear from the Figure S13. In Figure S13, neither the transformer alone nor the hybrid system deliver an especially good construction of the chaotic foodchain or Lotka-Volterra system, and it appears that the transformer alone may deliver a better reconstruction of the Lorenz system (though again it is not totally clear from the figure). Can the authors try to quantify this statement with some sort of MSE or other metric? If not, it is still not clear whether the reservoir is providing any utility other than to show that the interpolation performed by the transformer is sufficiently accurate for a reservoir computer to work. It is already known that reservoir computers, given accurate training data, can replicate low dimensional chaotic attractors.

Comment 2:
In response to Comment 2a, the authors propose a new sparsity metric, S_r, that depends on the Nyquist criteria. This is an improved metric to S_m, which depends only on the sampling time Δs, which is chosen in an apparently arbitrary way. Still, S_r unnecessarily depends also on the arbitrary Δs, and I would recommend to remove this dependence by defining sparsity as

$$S = \frac{L_s^N - L_s^O}{L_s^N}$$

More important is the fact that now that the Nyquist criterion is considered, it is clear that the transformer presented in this manuscript works only when there is not significant sparsity in an information theory sense. The evidence in the manuscript is as follows:
a) On p. 3, the authors state that "$S_r = 1$ corresponds to the theoretical minimum sampling case (at the Nyquist rate)" and that $S_r < 1$ corresponds to oversampling in the Nyquist sense.
b) Throughout the manuscript (e.g., on p. 8 of the main text and on p. 14 of the supplement), the authors state that the performance is poor when $S_r > 1.0$.

From these two statements, one can not conclude that the transformer works in the presence of sparsity, when sparsity is defined while considering information theory.

At issue is the selection of what is called in the manuscript as the "natural sampling rate" Δs. This quantity is apparently chosen arbitrarily so that "the attractor remains sufficiently smooth." The result of using this sampling rate rather than the Nyquist sampling rate (or something based on it) is that the systems are significantly oversampled, such that they can be downsampled by a large factor while still

retaining the information content. This allows for claims of "high reconstruction accuracy even when the available data is only 20% of that required to faithfully represent the dynamical behavior of the underlying system" (p. 4) even though high performance is never obtained for $S_r > 1$.

Further evidence for this oversampling is provided, e.g, on p. 8 L_s=2000 while L_s^N=280. This implies that the Δs used oversamples the system by a factor of 2000/280 ≈ 7. That this system is oversampled before being downsampled is confirmed by the fact that on p. 8, $S_r = 1$ (implying that they are providing to the transformer exactly the amount of information required by Nyquist). Yet, the definition of sparsity using only Δs, $S_m = 0.86$, which makes it seem as if the transformer is performing with only 14% of the needed data.

I suggest that Δs not be called the "natural sampling rate" because there is nothing natural about it. It is arbitrarily (no systematic criteria is presented) chosen. The Nyquist sampling rate (or something related to it) would be more natural. Further, I suggest that any claims of high sparsity based on Δs be removed, as the apparent "sparsity" is likely due to initial oversampling. Indeed, throughout the manuscript, there are reference to conditions of "high sparsity" when $S_r < 1$. These are overstated and should be removed.

Comment 2b:
The additional work in response to comment 2b is convincing and appreciated. However, in light of the above comment, it is not clear that this method helps with (information) sparsity. It is clear that it works with irregular sampling (which is useful in its own right). If the authors would like to claim that their method works under conditions of high sparsity, I suggest that they choose conditions that are actually sparse according to their own (new) metric S_r.

Comment 2c:
I do not think that the entropy rate is being computed accurately here, as the results do not make sense. As the sparsity of information input to the transformer is increased (as S_r increases), the quantity of information per unit time (i.e., the entropy rate) input into the transformer necessarily decreases. However, the bars labeled "Observations" in Fig. 4b increase with S_r. This cannot be correct. I suggest the authors revisit this computation.

Commend 2d:
The first paragraph on p. 4 now reads: "We demonstrate that it can successfully reconstruct the dynamics of approximately three dozen prototypical nonlinear systems with high reconstruction accuracy even when the available data is only 20% of that required to faithfully represent the dynamical behavior of the underlying system."

This 20% value depends on Δs which is apparently chosen arbitrarily (in the sense that no systematic definition of it that depends on the actual sampled information is presented). In other words, this seemingly incredible 20% value is a result of choosing Δs to be too small such that the system is oversampled. This statement is a significant exaggeration and should be removed entirely.

Comment 2e:
The additional work to compare transformer with compressed sensing is interesting and adds to the work significantly. It shows that the story of this transformer has much more depth than was presented previously.

However, this depth is completely hidden in the supplement. The main text simply states that

“Simulation results show that for a target system, when the observational sparsity is high, our framework outperforms CS significantly (Supplementary Note 10).”

This is perhaps true, but certainly misleading. When the sparsity is low to moderate (<0.75 or so), compressed sensing performs better than the transformer and is likely to be computationally less expensive to train and operate. When the sparsity is high (>1), both the transformer and compressed sensing perform poorly, though the transformer performs better. Only in a narrow range of sparsities does the transformer work well when the compressed sensing does not. This should be reflected in the discussion of the compressed sensing results so as not to overstate the performance of the transformer.

Comments 3, 4 and 6 have been addressed clearly, and the associated changes have improved the manuscript in my opinion. **Comment 5** has essentially not been addressed but is a difficult question. I suggest that the authors make a comment in the discussion or conclusion about the difficulty of validating the predictions of their model and leave it to future work.

I still think that the fundamental result is that transformers can provide a useful technique to interpolate chaotic time series. The authors appear to agree on this point. In my opinion, this result is interesting and potentially useful, but I still feel it might be better suited for a more specialized journal or for Communication Physics.

The authors have addressed my concerns about the presentation of their results sufficiently. The claims presented in the manuscript are now supported by the data presented. I recommend its publication. I hope they will consider the points raised below in a final revision, as I think they will further improve the manuscript.

I address each point from the authors' most recent response below.

Authors: "A general observation was that, some excessively large sampling rate usually will result in jagged attractors unsatisfactorily representing the system dynamics, whereas an overly small sampling rate will lead to unnecessary oversampling. As such, it is necessary to select the sampling rate Δs for each system carefully, depending on its natural dominant frequency. To clarify this point, we have produced a new figure, Fig. S2 in Supplementary Note 2, with examples of the following simulated target systems: chaotic food chain, Lorenz, and Lotka-Volterra, under varying sampling rates Δs. The orange-colored attractors represent the data ultimately used in our study. It can be seen that increasing the sampling rate leads to increasingly jagged time series and attractors, and selecting Δs that is too large contradicts our goal of reconstructing smooth dynamical systems: the interpolated or the ground-truth attractors do not resemble those in the right-most column of Fig. S2. "

***(in their response to Comment 2)*:** "In general, the sampling rate needs to be selected carefully to ensure the smoothness of the attractors: the chosen Δs should be such that the reconstructed attractors are sufficiently smooth yet not over sampled, with limited data points. The new figure, Fig. S2, demonstrates this criterion - higher sampling intervals produce jagged attractors unsuitable for accurate interpolation."

Reviewer Comment: The authors seem to be concerned with an arbitrary concept of smoothness (here and also in their response to Comment 2), which they assert is somehow important but they do not explain why, nor do they quantify it. They state without evidence or citation that "An excessively large Δs can produce jagged attractors that inadequately represent the long-term dynamics," and that "higher sampling intervals produce jagged attractors unsuitable for accurate interpolation." However, if one considers a sine wave with 5 samples per period, it will appear "jagged" to the eye, but it is sufficiently sampled to later on produce an accurate reconstruction with standard interpolation techniques. Indeed, I strongly suspect that the orange attractors shown in Fig. S2 are oversampled precisely because they appear very smooth to the eye.

Of course there is nothing wrong with oversampling if one wants an attractive visualization. The issue I am raising is that using the oversampled sampling rate in the definition of sparsity allows one to claim incredible performance in terms of "sparsity" when the information itself is not actually sparse. See also the related example about Fourier interpolation in a comment below.

Authors: "In previous Fig. S13 (now Fig. S14 on page 18), the green attractors in the left column represent the reconstruction from the transformer alone, while the blue attractors in the right column are reconstructed from our hybrid machine-learning framework. For the chaotic food chain and Lorenz system, it is evident even visually that the attractors reconstructed by the transformer alone frequently

deviate from the ground-truth attractors. Conversely, the attractors reconstructed using the hybrid method align closely with the ground truth attractors, demonstrating superior reconstruction accuracy compared to the case where only the transformer is used. For the Lotka-Volterra system, we previously clarified that the relatively poor reconstruction performance arises due to the partial usage of data - only the first three out of four dynamical variables are utilized. This is stated in Supplementary Note in the paragraph below Eq. (S3).

Quantifying performance differences via metrics such as the MSE is challenging due to the fundamentally distinct roles of the transformer and reservoir computing. Particularly, the transformer is primarily responsible for interpolation or missing data imputation, while the reservoir computer leverages the interpolated data by the transformer solely for long-term prediction. Using the MSE for long-term prediction is inappropriate because of the hallmark of chaos: sensitive dependence on initial conditions. In fact, even from the same initial condition, the machine-learning predicted evolution of the state variables will diverge from the true evolution after a few Lyapunov times."

Comment: From the figures presented here, it still seems to me that In Figure S13, neither the transformer alone nor the hybrid system deliver an especially good construction of the chaotic food chain or Lotka-Volterra system, and it appears that the transformer alone may deliver a better reconstruction of the Lorenz system (though it is not totally clear from the figure). However, since the authors do not attempt to quantify agreement in any way, the reader is left no choice but to take their word for it.

Authors: "It is worth emphasizing that the role of reservoir computing in our hybrid framework lies in illustrating that the interpolated data from the transformer is accurate and can be used for training a reservoir computer for subsequent long-term prediction of the attractor. More specifically, the workflow of our hybrid framework is as follows. Initially, sparse observations of an unseen system are interpolated by the transformer. The resulting interpolated time series, while accurate, remains limited in length. We then invoke reservoir computing to extend these accurate short series into arbitrarily long prediction, effectively converting sparse observations of limited length into long trajectories on the attractor. We agree with the referee that the primary novelty and core contribution of our work lie in applying the transformer to interpolating unseen complex time series. The subsection on reservoir computing for long-term dynamics prediction was moved to the last subsection in Results, and the following details were moved to SI (on page 9) as the last paragraph in Supplementary Note 4:"

Comment: I think we are in agreement on this point now. Using a reservoir computer to perform attractor reconstruction from a time series is a well-established method, and its only purpose here is to qualitatively show that the transformer interpolation is "good enough."

Authors: "We thank the reviewer for the insightful comment. Upon evaluation, we found this newly proposed metric S results in negative values that may not be appropriate for characterizing sparsity. For instance, in the chaotic food chain system, $S = -6.14$ corresponds to fully observed time series, $S = -2.57$ corresponds to 50% data sparsity ($Sm = 0.5$), while $S = -0.43$ corresponds to 80% sparsity ($Sm = 0.8$). This unbounded negative property might hinder an intuitive interpretation."

Comment: This is a good point. An intuitive interpretation is also important.

Authors: We agree with the referee that using the term "natural sampling rate" could be inappropriate. We have now simply used the term "sampling rate" instead. In general, the sampling rate needs to be selected carefully to ensure the smoothness of the attractors: the chosen Δs should be such that the reconstructed attractors are sufficiently smooth yet not over sampled, with limited data points. The new

figure, Fig. S2, demonstrates this criterion - higher sampling intervals produce jagged attractors unsuitable for accurate interpolation."

"Regarding the reviewer's comment on sparsity defined strictly by the Nyquist criteria, we stress that Nyquist sampling provides a theoretical baseline to ensure sufficient information. However, sampling beyond the Nyquist criterion does not inherently imply oversampling in practical interpolation contexts. Interpolation methods typically aim at reconstructing smooth trajectories rather than simply achieving the minimal Nyquist sampling. Examples can be found in the rightmost column of Fig. S2, which are not smooth and thus not accepted as the correctly interpolated attractors."

Comment: Indeed, interpolation methods aim at reconstructing smooth trajectories from "jagged" time series measurements. The "true trajectory" is smooth. If the time series measurement is already "smooth," then there is no need for interpolation.

Again, a sine wave that is sampled 5 times per period will not appear smooth to the eye, and yet a simple Fourier interpolation will be able to reconstruct the smooth sine wave from it. One may need to sample that sine wave 50 times per period for the time series to appear "smooth"--yet it would seem incorrect to claim that Fourier interpolation can achieve accurate reconstruction of the sine wave from measurements with 90% sparsity.

Authors: "To our knowledge, information theory does not explicitly define the Nyquist criterion as the definitive boundary between sparse and non-sparse sampling, notwithstanding the concept of sparsity in compressed sensing which is different. In fact, exceeding the Nyquist sampling rate does not mean oversampling, especially for reconstructing the attractors of smooth dynamical systems."

Comment: This is exactly what the word "oversampling" means according to any standard digital signal processing textbook. For example, on p. 585 of the textbook *Digital Signal Processing* by Li Tan and Jean Jiang (2013, 2nd edition, ISBN: 978-0-12-415893-1), it is stated that:
"Oversampling uses a sampling rate that is much higher than the Nyquist rate. We can define an oversampling ratio as $f_s / 2f_{max} \gg 1$ "

Authors: "With respect to the term "high sparsity," our usage emphasizes the practical and empirical scenarios where quite a few observational points are available and the traditional methods (e.g., linear and cubic interpolation, compressed sensing) and conventional machine-learning methods (e.g., LSTM, xLSTM) fail to reconstruct the time series. Nevertheless, we agree that our previous claim "high reconstruction accuracy even when available data is only 20% of that required to faithfully represent the dynamical behavior of the underlying system" is inappropriate. It has now been removed."

"We appreciate the referee's statement "the additional work in response to comment 2b is convincing and appreciated". We emphasize that our concept of high sparsity means not only quite a few available observations but also the conditions under which traditional reconstruction methods fail. We have added the following statement to Sec. 1.3 in Results (at the end of page 10):

Hereafter, the term "high sparsity" is used to mean not only a limited number of available observations but also the failure to reconstruct the system accurately at this level of sparsity by the traditional methods such as linear and spline interpolation, compressed sensing, and conventional machine-learning techniques. "

Comment: The paper has been improved by the clarification that by "high sparsity" the authors are not referring to sparsity of information but rather to the fact that their method is better at interpolation than other methods. In my opinion, sparsity is not the correct word for this, but it is sufficient that the authors have clarified what they mean.

Authors: Response to Comment 2c: "I do not think that the entropy rate is being computed accurately here, as the results do not make sense. As the sparsity of information input to the transformer is increased (as Sr increases), the quantity of information per unit time (i.e., the entropy rate) input into the transformer necessarily decreases. However, the bars labeled "Observations" in Fig. 4b increase with Sr. This cannot be correct. I suggest the authors revisit this computation."

Response: We have revisited our entropy-rate calculations with two improvements. (1) Previously, we removed time steps containing missing values in any variable before calculating the KS entropy of the inputs. For a fair comparison, we have now omitted this preprocessing step for the inputs. (2) Initially, we defined the entropy rate h_{KS} as $(1/\tau_{min}) \cdot P\, \tau\, \rho(\tau)\, \log(1/\rho(\tau))$. We reexamined this definition and concluded that using the mean recurrence time $\langle \tau \rangle$ in the denominator provides a more representative measure than using the minimum recurrence time τ_{min} so as to account for all the recurrences. With the revised definition, we recalculated h_{KS} for the transformer inputs, outputs, and the ground truth across varying sparsity levels, and updated panel (b) of Fig. 4 (on page 9) accordingly. It is worth noting that, even with the improvement, the calculated input entropy rate still increases with S_r. Through a detailed analysis, we have identified the main reason as the denominator used in calculating h_{KS}. Regardless of whether τ_{min} or $\langle \tau \rangle$ is used, the sparse input yields a small denominator due to the prevalent zero values, making the "recurrence" occurrences artificially frequent and resulting in an increased entropy-rate value. We have removed the descriptions of input preprocessing in Sec. 1.3 where $\langle \tau \rangle$ is the mean observed recurrence time. In addition, we have recalculated the hKS values for all the systems and updated them in Tab. S3. Panel (b) of Fig. S4 has also been updated. These results illustrate that revised calculations do not alter the conclusion."

Comment: I still do not think that the entropy rate is being computed accurately here, as the results do not make sense. As the sparsity of information input to the transformer is increased (as Sr increases), the quantity of information per unit time (i.e., the entropy rate) input into the transformer necessarily decreases. However, the bars labeled "Observations" in Fig. 4b still increase with Sr. This cannot be correct. I think the manuscript would be significantly improved by an accurate estimate of these entropy rates—it would be the best way to quantify the sparsity of information. If for some reason this is not possible, I suggest the authors remove these computations—or in the interest of transparency leave them in the supplement but note that they cannot be correct—since presenting clearly incorrect entropy-rate estimates will only spark confusion and misunderstanding.

Authors: "We have removed the statement related to 20%, and now the sentence in the first paragraph on page 4 reads. We demonstrate that it can successfully reconstruct the dynamics of approximately three dozen prototypical nonlinear systems with high reconstruction accuracy. "

Comment: I believe the clarity and accuracy of the paper has been improved by this change.

Authors: "We thank the referee for stating "The additional work to compare transformer with compressed sensing is interesting and adds to the work significantly." We agree that the comparison deserves a more thorough discussion in the main text. The relevant statement in the last paragraph of Sec. 1.3 (on page 10) has been revised to

Simulation results show that, for a target system, when the observational sparsity is low to moderate, compressed sensing performs better than the transformer. However, when the observational sparsity is high, our hybrid machine-learning framework outperforms compressed sensing significantly (Supplementary Note 10).

In addition, we would like to clarify that sparsity Sr does not vary linearly with the sampling rate. Instead, as Δs increases linearly, the resulting sparsity increases in a concave manner. That is, a small increase in Δs at low values leads to a large drop in the data availability. For example, in the chaotic food chain system, as shown 9 in Fig. S2 (a), increasing Δs from 1 to 2 reduces the data availability by 50% (corresponding to Sr = 0.58), while increasing Δs further to 7 results in Sr = 1 (only a 0.42 increment). This demonstrates that the the range between Sr = 0.75 and Sr = 1 gives significant sparsity. We have added the following description to Supplementary Note 2 (on pages 4-5):

It is worth noting that the the third column of Fig. S2 corresponds to doubling the selected sampling rate, which reduces the data availability to approximately 50%. The last column shows that a sampling rate 7-8 times higher than the selected rate yields about 14% data availability, which is consistent with the Nyquist criterion (Sr = 1). This demonstrates a nonlinear (concave) relationship between the sampling rate and sparsity. For example, in the chaotic food chain system, increasing Δs from 1 to 2 reduces the data availability to 50% (corresponding to Sr = 0.58), whereas further increasing it to 7 results in Sr = 1 (only a 0.42 increment)."

Comment: These additions have improved the clarity of the manuscripit.