Using weak predictors for training diffusion models

Marlin Lee

Abstract

Diffusion models have shown remarkable success across various domains, yet their performance can be limited by the complexity of the underlying data distribution. This paper introduces a novel approach to enhance diffusion model capacity through the integration of weak teachers - auxiliary losses that provide supplementary learning signals during training. We specifically focus on Sea Surface Temperature (SST) prediction, demonstrating how the existence of weak predictors can be used to better guide the training of diffusion models. Our methodology explores data-driven weak learners with a mathematically guided integration that only uses them when they will align with the correct distribution of answers. The experimental results show that our approach achieves a 39.0% improvement in prediction accuracy compared to the diffusion models of the baseline, while maintaining computational efficiency. Our findings suggest that weak teacher integration could be a promising direction for improving diffusion models in scientific applications.

1 Introduction

Recent advances in diffusion models have demonstrated their remarkable capability in generating high-quality samples across diverse domains, from image synthesis to scientific predictions. However these models trade this performance with requiring extra model capacity and slower inference time. This issue can prevent application in situations where these are limited. A fundamental challenge may lie in the commonly used loss function which is the simple mean squared error(MSE) between the original and predicted sample. The diffusion process of iteratively adding and removing noise means much of the predicted sample has little function, however the loss function still puts training pressure to optimize it. While previous works have attempted to address this through architectural modifications or specialized training schemes, these approaches often introduce additional complexity without fully resolving the core inefficiency. We seek to explore methods of resolving this tension.

Our work makes several key contributions:

- Introduce a new theoretical framework to integrate real weak teachers into the training process, aiming to only apply the teachers when their guidance is relevant to the student and the underlying task.
- Demonstrate practical benefits in Sea Surface Temperature prediction, a domain with many real weak teachers that can guide the model.

2 Related Work

2.1 Diffusion Models

Diffusion models have emerged as a powerful framework for generative modeling, with the foundational work of Ho et al. [2020] introducing the modern formulation that balances model capacity with training stability. This approach defines a forward process that gradually adds Gaussian noise to data and a reverse process that learns to denoise the data. Nichol and Dhariwal [2021] further

Preprint. Under review.

enhanced this framework by introducing several architectural improvements and training techniques that increased sample quality while maintaining computational efficiency. Recent work by Salimans and Ho [2022] has focused on addressing the computational overhead of diffusion models through progressive distillation, demonstrating that careful model compression can preserve performance while significantly reducing inference time.

2.2 Knowledge Distillation

The concept of knowledge distillation, first formalized by Hinton et al. [2015], demonstrates how a smaller model can learn from a larger teacher model's soft predictions rather than just hard labels. This approach has proven particularly effective when the teacher model can provide additional insights about the task structure. Sun et al. [2021] extended this framework to scenarios with weak teachers, showing that even imperfect guidance can improve student model performance when properly integrated into the training process. Our work builds upon these insights, adapting the teacher-student framework to the unique characteristics of diffusion models and developing a principled approach for determining when weak teacher signals become statistically relevant.

2.3 Sea Surface Temperature Prediction

Sea Surface Temperature prediction has seen significant advances through deep learning approaches. Choi et al. [2021] demonstrated the effectiveness of deep learning models for regional SST prediction near the Korean Peninsula, establishing baseline performance metrics for neural approaches to this task. Xu et al. [2023] extended this work to global predictions, highlighting the challenges of maintaining accuracy across diverse oceanic regions. The integration of physical constraints into neural models was explored by Yuan et al. [2023], who showed how partial differential equations could guide the learning process. Recent evaluation work by NOAA Technical Report [2023] has provided comprehensive benchmarks for assessing SST prediction accuracy, while the availability of high-quality observational data from National Centers for Environmental Prediction [2023] has enabled robust model validation. Our work builds upon these foundations by combining the strengths of diffusion models with domain-specific weak predictors derived from physical constraints and existing forecasting systems.

3 Theoretical Framework

Our analysis exploits the well-defined statistical properties of diffusion models to determine when two samples become indistinguishable. This approach relies on the consistent behavior of the diffusion model forward process, where transitioning to any timestep t is equivalent to sampling from a normal distribution with known parameters, as shown in Equation (1).

$$P(X_t|X_0) = N(\alpha_t X_t, \beta_t) \tag{1}$$

Given two samples X_0 and W_0, this formulation allows us to analyze their statistical distinguishability as they progress through timesteps. Because both conditional distributions are normal with identical variance, their difference follows a normal distribution characterized by Equation (2).

$$N(\alpha(X_0 - W_0), 2(1-\alpha)I) \tag{2}$$

This transforms our analysis into a hypothesis testing framework in which we seek to determine whether our sample is coming from a distribution with mean zero. We compute the Mahalanobis distance of the sampled distribution, shown in Equation (3).

$$\frac{\alpha||(X_0 - W_0)||_2}{\sqrt{2(1-\alpha)}} \tag{3}$$

Under the null hypothesis of zero mean, this distance follows a chi-squared distribution, allowing the calculation of the p-value. We define two samples as statistically different if we can reject the null hypothesis with confidence $\alpha = 0.05$.

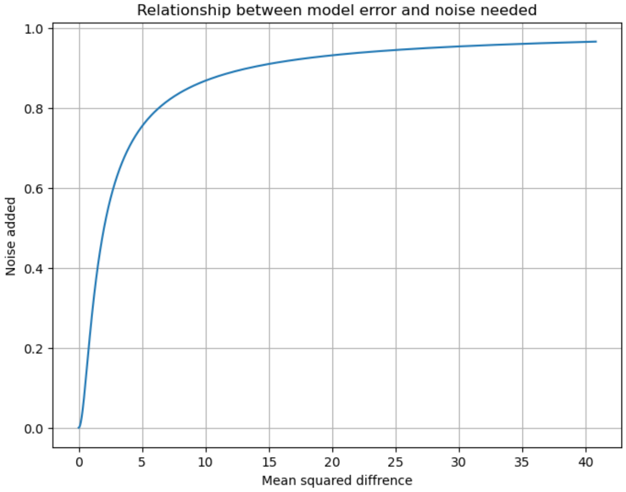

Figure 1: Relationship between Mean Squared Error (MSE) and diffusion timesteps at which they become indistinguishable.

A key insight is that this statistical threshold must exist at some timestep between t = 0 and t = T. At t = 0, the absence of variance ensures perfect discrimination between distributions. Conversely, at t = T, both conditional distributions converge to standard Gaussians, making discrimination impossible. Therefore, there must exist some critical timestep where statistical distinction becomes impossible while remaining possible at t-1.

This threshold depends on two parameters: the scaling factor α (a function of monotonic timestep t) and the distance between X_0 and W_0. For any pair of samples, we can deterministically compute their threshold timestep. Moreover, since the distance metric reduces to Mean Squared Error (MSE), we establish a direct mapping between MSE and timestep, as illustrated in Figure 1.

We extend this framework to analyze weak predictors that produce predictions W_0^i. These predictions serve as guiding loss terms specifically when the diffusion process reaches the regime where they represent statistically plausible predictions. For specialized cases like energy conservation where the predictor outputs a scalar rather than a full prediction, we adapt our framework using a single-dimensional normal distribution:

$$D \sim N(\alpha(\sum X_0 - \sum X_1), 2n(1-\alpha)) \tag{4}$$

where n represents the dimensionality of the original input space. This enables p-value computation through a standard Student's t-test.

For each weak predictor i, we determine its threshold timestep beyond which its predictions become statistically indistinguishable from ground truth. because this needs to be across the whole dataset we compute it for each example in the training step and find the average threshold value. The resulting loss function for each predictor is only applied when t exceeds this threshold, ensuring we only incorporate predictions when they become statistically relevant.

4 Methodology

To test our theory we devised a use case of predicting SST data. This task involves being shown a number of prior temperatures and then predicting what it is in the future. We chose this task because environmental science has a number of weak predictors making finding them easy. We also chose it because it is a relatively simpler task having only a single channel dimension per observation. This task can be important for a number of reasons especially for a number of downstream tasks that want future temperatures for the prediction of other more complicated features like weather.

4.1 Data Processing

The SST data was collected from the NOAA OI SST V2 High Resolution Dataset, which contains daily measurements from September 1981 to October 2024 with a spatial resolution of $0.25°$. Each diffusion input was a non-overlapping $16°$ by $16°$ chunk with latitudes ranging from $16.25°$ to $32.25°$ and longitudes ranging from $262°$ to $358°$. This region covers a substantial portion of the North Atlantic Ocean. An example input is shown in Figure **??**.

The dataset was split chronologically, with 75% used for training and the remainder for testing. We normalized the data using the training set statistics, scaling all temperatures to the range [0,1]. Land areas were masked with a value of -1 and excluded from both training and evaluation loss calculations. For the prediction task, each sample consisted of three consecutive daily measurements as input features, with the target being the temperature distribution seven days in the future. We also conditioned the model on the day of the year to allow the model to learn seasonal dynamics.

4.2 Model set up

We adopted an unconventional diffusion model architecture optimized for efficient integration with weak predictors. The model consisted of a compact 5.3M parameter U-Net that directly predicted X_0 across 1000 training steps. This smaller architecture, compared to typical diffusion models which often use hundreds of millions of parameters, was chosen due to the task selected to highlight the situation where model capacity is a constraint.

We employed a linear noise scheduler and conducted a statistical analysis of the distinguishability between the latest input day and the target prediction using the method described in our Theoretical Framework section. Our analysis revealed that these distributions became statistically indistinguishable (p > 0.05) at t=232, according to the Mahalanobis distance metric defined in Equation (3). To reduce redundant computation, we modified the standard diffusion training process to focus only on the final 250 timesteps where the distinction between input and target stop being ambiguous. During inference, we initialized the diffusion process at t=250 using a noisy version of the previous measurement, rather than starting from pure Gaussian noise at t=1000. This modification significantly reduced computational overhead while maintaining prediction quality, as the early timesteps of the diffusion process primarily serve to destroy information we already have in our input measurement.

4.3 Evaluation method

We employed two distinct metrics to evaluate our model's performance. The first metric was the Mean Squared Error (MSE) between the model's direct X_0 predictions and the ground truth future temperatures. While this metric directly measures the model's training objective, it has limited practical value as it does not reflect how diffusion models are deployed in real scenarios, where the full sampling process is used. Additionally, this metric is not directly comparable to other modeling approaches that may use different loss functions or prediction mechanisms.

Our primary evaluation metric was the Mean Absolute Error (MAE) between temperatures generated through the complete diffusion sampling process and the ground truth future temperatures. For each test sample, we initialized the diffusion process at t=250 with the noisy input measurement and performed the full sampling procedure to generate the prediction. This metric provides a more realistic assessment of model performance and enables direct comparison with other forecasting approaches in the literature. We compute MAE only over ocean points (excluding masked land areas) to ensure meaningful comparison across different spatial regions.

4.4 Weak predictors

For the task of SST prediction, there are several physics-guided metrics that can function as weak predictors. For this project, we selected three weak predictors to integrate into our model:

1) Persistence prediction assumes the temperature remains unchanged from the most recent measurement. While simplistic, this predictor was foundational to our approach, informing our diffusion process initialization at t=250 as described in Section 4.2.

2) Conservation of Energy enforces that the total energy in the system, proportional to the sum of temperatures across the spatial region, should remain approximately constant over short time

periods. This predictor became statistically indistinguishable from ground truth at t=173 according to Equation (4), reflecting the fact that this problem is not a closed system and therefore does not perfectly maintain energy.

3) NOAA's Climate Forecast System version 2 (CFSv2) provides operational climate predictions. These forecasts, while computationally expensive to generate, offer valuable guidance for our model. Statistical analysis showed these predictions became indistinguishable from ground truth at t=58, suggesting they provide the strongest signal among our weak predictors.

Having already incorporated the persistence predictor into our diffusion process initialization, our experiments focused on integrating the conservation of energy and CFSv2 predictors. The earlier threshold (t=58) for CFSv2 predictions compared to energy conservation (t=173) aligns with our expectation that operational forecasts provide more precise guidance than general physical constraints.

4.5 Experimental Design

To evaluate the effectiveness of our weak predictor integration approach, we conducted three sets of experiments:

4.5.1 Baseline

We first established a baseline by training our diffusion model without any extra guidance. This model was trained using only the standard diffusion loss on the final 250 timesteps as described in Section 4.2. This baseline represents the performance achievable through direct diffusion modeling alone.

4.5.2 Full Integration

In our second experiment, we integrated both weak predictors (conservation of energy and CFSv2) across all applicable timesteps. This approach ignores our mathematical theory and instead applies a constant weight at all timesteps. The loss weights for these predictors were set to 0.1 relative to the main diffusion loss to prevent overfitting to any single predictor. This serves as a second baseline to test the naive approach of assuming guidance is always beneficial.

4.5.3 Conditional Integration

In our third experiment, we integrated both predictors only for timesteps before there threshold. Specifically, for $t \geq 173$, we applied the energy conservation loss, and for $t \geq 58$ we incorporated the CFSv2 prediction loss. . This approach tests the hypothesis that consistent guidance from weak predictors can improve model performance.

For all experiments, we used the Adam optimizer with a learning rate of 1e-5 and trained for 6000 steps with batch size 128. Each configuration was evaluated using the MAE metric described in Section **??**, with results computed over the entire test set.

5 Results

Our experiments demonstrate that the conditional integration of weak predictors can meaningfully improve diffusion model performance for SST prediction. Table 5 summarizes the MAE scores across our three experimental configurations.

Configuration	MAE
Baseline	.0170
Full Integration	.0115
Conditional Integration	.0104

Table 1: Mean Absolute Error for different experimental configurations

The baseline model achieved an MAE of .0170, establishing our reference point for model performance. Full integration of weak predictors reduced this error to .0115, demonstrating that even naive predictor integration provides substantial benefits with a 32.4% error reduction. Our conditional integration approach achieved the best performance with an MAE of .0104, representing a 39.0% improvement over the baseline and a 9.8% improvement over full integration.

These results support our theoretical framework's prediction that weak predictors should only be applied after specific timesteps. The fact that conditional integration outperforms full integration suggests that applying weak predictors before their statistical relevance threshold can actually hinder model performance. This aligns with our hypothesis that weak predictors are most useful when they are plausible outputs of process.

6 Limitations and Future Work

A significant limitation of our study emerged from the choice of task domain. While SST prediction initially appeared ideal due to its abundance of physics-based weak predictors, it proved suboptimal for demonstrating our framework's full potential. The combination of temporal conditioning and the relatively low-frequency nature of daily temperature changes meant that models larger then 5.3M parameters could easily memorize the underlying patterns. Even with our small model the high baseline performance made it challenging to fully demonstrate the benefits of weak predictor integration, as the model was not sufficiently constrained by capacity limitations.

The most promising direction for future work emerged from a theoretical insight developed during this project. Our framework naturally extends beyond real weak predictors to accommodate synthetic ones, eliminating the requirement for task-specific guidance. This generalization could expand the approach for many other domains.

We plan to better validate the theory by applying the framework to a more traditional diffusion tasks, such as image generation, where the underlying dynamics are more complex and model capacity constraints are more significant. This would provide a clearer demonstration of the benefits in a better studied area.

7 Conclusion

This paper introduced a novel framework for integrating weak predictors into diffusion models based on statistical distinguishability thresholds. Our theoretical analysis provided a principled approach for determining when weak predictors become relevant during the diffusion process, leading to a 39.0% improvement in SST prediction accuracy compared to the baseline. While our initial validation on SST prediction demonstrated promising results, the limitations we encountered suggest that the framework's full potential may be better realized in domains with more complex underlying distributions. The natural extension of our approach to synthetic weak predictors opens up possibilities for improving diffusion models across a broader range of applications. As diffusion models continue to evolve, the integration of weak predictors—both real and synthetic—may offer a promising direction for enhancing model performance while maintaining computational efficiency.

References

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *Proceedings of the 38th International Conference on Machine Learning*, pages 8162–8171, 2021.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *International Conference on Learning Representations*, 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Wei Sun, Tianlong Chen, and Yu Wang. Teacher-student framework: A reinforcement learning approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):9231–9239, 2021.

Min-Kyu Choi, Young-Gyu Park, and Jin-Hee Lee. Deep-learning model for sea surface temperature prediction near the korean peninsula. *Ocean Science*, 17(5):1431–1442, 2021.

Jiahao Xu, Dongxiao Wang, Xiang Li, and Yang Yang. Short-term prediction of global sea surface temperature using deep learning networks. *Journal of Atmospheric and Oceanic Technology*, 40 (3):567–582, 2023.

Tianyu Yuan, Zhuoyi Yang, and Jinghua Chi. A space-time partial differential equation based physics-guided neural network for sea surface temperature prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

NOAA Technical Report. Evaluating climate model predictions of sea surface temperature. Technical report, National Oceanic and Atmospheric Administration, 2023.

National Centers for Environmental Prediction. Sea surface temperature data, 2023.