WHY DID THIS MODEL FORECAST THIS FUTURE? INFORMATION-THEORETIC TEMPORAL SALIENCY FOR COUNTERFACTUAL EXPLANATIONS OF PROBABILISTIC FORECASTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Probabilistic forecasting of multivariate time series is significant to several research domains where multiple futures exist for a single observed sequence. Identifying the observations on which a well-performing model bases its forecasts can enable domain experts to form data-driven hypotheses about the causal relationships between features. Consequently, we begin by revisiting the question: what constitutes a causal explanation? One hurdle in the landscape of explainable artificial intelligence is that what constitutes an explanation is not well-grounded. We build upon Miller's framework of explanations derived from research in multiple social science disciplines, and establish a conceptual link between counterfactual reasoning and saliency-based explanation techniques. However, the complication is a lack of a consistent and principled notion of saliency. Also, commonly derived saliency maps may be inconsistent with the data generation process and the underlying model. We therefore leverage a unifying definition of information-theoretic saliency grounded in preattentive human visual cognition and extend it to forecasting settings. In contrast to existing methods that require either explicit training of the saliency mechanism or access to the internal parameters of the underlying model, we obtain a closed-form solution for the resulting saliency map for commonly used density functions in probabilistic forecasting. To empirically evaluate our explainability framework in a principled manner, we construct a synthetic dataset of conversation dynamics and demonstrate that our method recovers the true salient timesteps for a forecast given a well-performing underlying model.

1 INTRODUCTION

The existence of multiple valid futures for a given observed sequence is a crucial attribute of several forecasting tasks, especially surrounding the dynamics of low-level human behavior. These tasks include the forecasting of trajectories of pedestrians (Huang et al., 2019; Mohamed et al., 2020; Rudenko et al., 2020; Salzmann et al., 2021; Zhang et al., 2019), vehicles (Carrasco et al., 2021; Gilles et al., 2022; Zeng et al., 2020; Zhao et al., 2020), and autonomous robots (Ivanovic et al., 2021; Vemula et al., 2017), or other more general nonverbal cues of humans (Adeli et al., 2020; Barquero et al., 2022; Nguyen & Celiktutan, 2022; Raman et al., 2021; Yao et al., 2018) and artificial virtual agents (Ahuja et al., 2019) in group conversation settings. Consequently, rather than making single (i.e. point) predictions, several machine learning methods in these settings have attempted to forecast a distribution over plausible futures (Mohamed et al., 2020; Raman et al., 2021). In this work, we introduce and address a novel research question towards gaining domain-relevant insights into such forecasts: given a reliable underlying model, how can we identify the observed timesteps that are salient for the model's *probabilistic* forecasts over a particular future window?

1.1 NOTIONS OF INTERPRETABILITY IN FORECASTING TASKS & DRAWBACKS

Recently, several works have proposed techniques for making interpretable *non-probabilistic* predictions for point forecasting tasks (Lim et al., 2020; Oreshkin et al., 2020; Pan et al., 2021). The

Table 1: Classes of Explanatory Question and the Reasoning Required to Answer. Reproduced from Miller (2019, Table 3)

Question	Reasoning	Description
What?	Associative	Reason about which unobserved events could have occurred given the observed events
How?	Interventionist	Simulate a change in the situation to see if the event still happens
Why?	Counterfactual	Simulating alternative causes to see whether the event still happens

general approach has been to train architectures to produce forecasts that are not only accurate but also interpretable. However, in what Lipton (2017) terms *The Mythos of Interpretability*, the notion of what renders these models interpretable is often not well-grounded and subject to presenting speculation in the guise of explanation Lipton & Steinhardt (2018). Examples of these operationalizations of interpretability[1] for forecasting tasks include: (i) injecting a suitable inductive bias into the model through a set of basis functions and identifying how they combine to produce an output (Oreshkin et al., 2020) (an approach that has recently been applied to the probabilistic setting as well (Rügamer et al., 2022)); (ii) employing a self-attention mechanism to learn temporal patterns while attending to a common set of features (Lim et al., 2020); and (iii) applying the notion of saliency maps from computer vision (Dabkowski & Gal, 2017) to time-series data as a measure of how much each feature contributes to the final forecast (Pan et al., 2021). A broader review of notions of interpretability across domains is in Appendix A.

Irrespective of the notion of interpretability, these methodologies are underpinned by two common attributes: (a) the interpretability mechanism needs explicit training as part of the model architecture, and (b) what constitutes a *good* explanation is subject to the biases, intuition, or the visual assessment of the human observer (Adebayo et al., 2018; Miller, 2019); a phenomenon we refer to as the *interpretation being in the eye of the beholder*. This is especially true for when saliency maps have been used as tools for post-hoc explanations: the computed map may not measure the intended saliency, and even be independent of both the model and data generating process (Adebayo et al., 2018; Atrey et al., 2020; Lapuschkin et al., 2019). Research on saliency-based methods is further confounded by the lack of a common notion of saliency. As Barredo Arrieta et al. (2020, Sec. 5.3) point out, "there is absolutely no consistency behind what is known as saliency maps, salient masks, heatmaps, neuron activations, attribution, and other approaches alike."

1.2 Saliency-Based Explanations for Counterfactual Reasoning & Drawbacks

Due to the inconsistencies mentioned above, we argue that the premise of what constitutes an explanation needs to be well-grounded within existing frameworks of how people define, generate, and present explanations. Miller (2019) recently turned to the vast body of research on the topic in philosophy, psychology, and cognitive science and highlighted the importance of causality in explanation. Specifically, in Table 1 we reproduce the categorization of explanatory questions he proposed based on Pearl and Mackenzie's *Ladder of Causation* (Pearl & Mackenzie, 2018). Using an abstract notion of an 'event' that needs explaining, Miller argues that the *what*-questions involve associative reasoning. *How* questions require interventionist reasoning to determine the set of causes that, if removed, would prevent the event from happening. The *why*-questions are the most challenging, as they require counterfactual reasoning to undo events and simulate other events. These also require associative and interventionist reasoning.

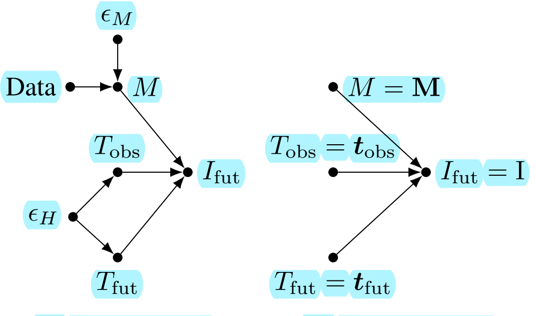
To apply Miller's (2019) framework to forecasting, consider a model $\mathbf{M}$ that predicts features over a future window $\boldsymbol{t}_{\text{fut}}$ by observing features over a window $\boldsymbol{t}_{\text{obs}}$. We argue that most of the existing interpretability approaches—including the two discussed categories involving the injection of inductive biases and attention-based mechanisms—are associative in nature. For a fixed $\boldsymbol{t}_{\text{fut}}$ and single $\boldsymbol{t}_{\text{obs}}$, they reason about the (unobserved) importance of features over $\boldsymbol{t}_{\text{obs}}$ using model parameters or attention coefficients based on one prediction from $\mathbf{M}$ (the 'event'). In contrast, we argue that the perturbation-based saliency methods have the potential to support counterfactual reasoning. The saliency masks are commonly learned by perturbing different parts of the input over

[1] Barredo Arrieta et al. (2020) argue for the importance of distinguishing between interpretability and explainability as different concepts, the latter denoting any active action or procedure taken by a model with the intent of clarifying or detailing its internal functions. From this perspective, what the cited works term interpretability is closer to the notion of explainability.

$\boldsymbol{t}_{\text{obs}}$—thereby simulating alternative 'causes' from $\mathbf{M}$'s perspective—and observing the effect on an error metric (the 'event'). However, while saliency-based approaches can, in theory, be promising in answering such why-questions, we identify several issues with how such methods are being applied for deriving explanations: (i) the feature-level manipulations used to produce saliency maps produce distortion interventions that may be inconsistent with the data generation process (Adebayo et al., 2018; Lapuschkin et al., 2019) and not lie on the real-world data manifold or preserve real-world semantics (Atrey et al., 2020); (ii) the notion of saliency in terms of minimizing an error metric is an arbitrary choice, requiring the ground-truth future that may not be available at test; and (iii) for forecasting applications, the saliency map is explicitly retrained for every single observed-future sequence (Pan et al., 2021) and does not capture salient structural patterns across samples.

1.3 Our Approach: Building upon a Unifying Framework of Bottom-Up Saliency

Our broad conceptual contribution is to link counterfactual reasoning with a principled notion of saliency. In Figure 1 we formally express counterfactual reasoning for explaining forecasts using causal graphs (Pearl, 2009). The generic graph expresses relationships between the random variables prior to training the forecasting model denoted by M. The exogenous variable ϵ_M captures the randomness in the training process and modeling choices including the distribution family for representing probabilistic forecasts. The exogenous variable ϵ_H captures the the randomness in the human observer's choice of observed and future windows to examine the model. Our central idea is to evaluate the information in the model's predicted distribution denoted by I_{fut}. Specifically, we propose posing the the following counterfactual question: *What information would* $\mathbf{M}$ *have about the future over a given* $\boldsymbol{t}_{\text{fut}}$ *if it observed the features over* $\boldsymbol{t}_{\text{obs}}$*?* The modified graph for evaluating this question is depicted in Figure 1b. Once the model $\mathbf{M}$ has been trained and the windows $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$ have been chosen, the effect of the exogenous variables on the variable I_{fut} disappears. This yields a deterministic system where the change in the information about the future can be evaluated for different values of $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$. Given that the focus of this work is the explanation phase, we assume the modified graph is already available. Note that while the procedure starting from training the forecasting model in the generic graph implicitly follows Pearl's *abduct-action-prediction* process (Pearl, 2009, p.207), estimating the distribution over the exogenous variables from the *abduction* step is conceptually not applicable in this setting.


ϵ_M
Data
M
T_{obs}
I_{fut}
ϵ_H
T_{fut}
$M = \mathbf{M}$
$T_{\text{obs}} = \boldsymbol{t}_{\text{obs}}$
$I_{\text{fut}} = \mathrm{I}$
$T_{\text{fut}} = \boldsymbol{t}_{\text{fut}}$

(a) Generic Graph **(b)** Modified Graph

Figure 1: Causal Graphs for Explaining Forecasts

To relate such counterfactual reasoning to saliency, we propose the following implication:

$$\text{observing the features at a timestep } t \in \boldsymbol{t}_{\text{obs}} \text{ results in a change in the model's information about the future } I_{\text{fut}} \text{ over the given window } \boldsymbol{t}_{\text{fut}} \implies t \text{ is salient.} \tag{1}$$

This notion of saliency related to surprisal or information associated with observations is grounded in preattentive human cognition (Loog, 2011; vanderHeijden, 1996). Concretely, the statement entails that an observation that changes the certainty of the future is surprising, and thereby salient. Note that the antecedent (on the left of the implication) is a counterfactual statement. The implication entails that for the antecedent to be true t must be salient. However, knowing the antecedent is false is not sufficient to conclude that t is not salient, i.e. there can be other notions of saliency that make t salient. In this work, rather than defining saliency in a top-down manner as a function of some task-specific error metric that cannot be computed at test, as other works have done (Pan et al., 2021), we build our computational framework upon a general expression of bottom-up, or task-agnostic, preattentive saliency (Loog, 2011). Specifically, we derive a closed-form solution for identifying salient timesteps using the differential entropy in $\mathbf{M}$'s forecast over $\boldsymbol{t}_{\text{fut}}$. We then empirically validate the framework using the counterfactual reasoning illustrated in Figure 1 for a forecasting task where human-verifiable ground-truth saliency exists.

Preattentive saliency captures what is perceived to be subconsciously informative before conscious, or attentive, processing by the brain. Here, a surprising or unexpected observation is salient. When applied to forecasting tasks, this idea of surprisal (or unexpectedness or informativeness) needs to be *tied to the future outcome*, which guides our proposal of evaluating the information in the predicted future density. Our approach is counterfactual because we examine changes in this future density in response to *different* realizations of real-world sequences $\boldsymbol{t}_{\text{obs}}$. In contrast, the only existing saliency-based method for time-series forecasting follows associative reasoning within Miller's (2019) categorization: Pan et al. (2021) retrain a saliency map for every *single* $\boldsymbol{t}_{\text{obs}}$ from scratch, one at a time, using a single (point) prediction over the $\boldsymbol{t}_{\text{fut}}$ instead of a distribution.

Our framework addresses all the aforementioned concerns we identified with saliency-based approaches in forecasting: (i) the counterfactuals in our framework pertain to windows of real observed features rather than random perturbations of the input, and therefore preserve real-world semantics; (ii) our information-theoretic perspective of saliency is principled, and given the underlying model, can be used to compute saliency for unseen test data where the ground-truth future is unavailable; (iii) we utilize a distribution over the futures that the model believes could have occurred for a single input: this distribution captures structural predictive relationships between features across multiple samples. Moreover, our framework involves no training for computing the saliency, and can be applied to any underlying model that outputs a distribution over futures.

2 Background: Information Theoretic Preattentive Saliency

Loog (2011) developed a general closed-form expression for saliency using a surprisal-based operational definition of bottom-up attention from the field of computational visual perception. Let $L : \mathbb{R}^n \to \mathbb{R}^d$ be a general n-dimensional d-valued image. A continuously differentiable feature mapping $\phi : \mathbb{R}^n \to \mathbb{R}^N$ relates every image location $x \in \mathbb{R}^n$ from L to N features. Further, let p_X denote the probability density function over image locations x and p_Φ denote the probability density function over all feature vectors $\phi(x)$. The distribution p_X captures any prior knowledge that would make one location more salient than another. In the absence of such prior knowledge, p_X is typically chosen to be uniform.

The saliency $S(x)$ of a location x is then defined in terms of the amount of information—or surprise—of its associated feature vector $\phi(x)$ relative to the other feature vectors extracted from the same image. The intuition is that the larger the information content of a certain combination of features, $-\log p_\Phi(\phi(x))$, the more salient the location x:

$$S(x) > S(x') \iff -\log p_\Phi(\phi(x)) > -\log p_\Phi(\phi(x')). \tag{2}$$

This general definition unifies seemingly different definitions of saliency encountered in the literature. It relates to salient observations being considered unexpected, rare, or surprising (Garcia et al., 2001; Itti & Koch, 2001; Torralba, 2003; Walker et al., 1998). It also incorporates the notion of saliency considered in decision-theoretic settings (Lingyun et al., 2007; Torralba, 2003), where bottom-up saliency increases with an increase in the information associated with feature vectors (Gao & Vasconcelos, 2007; Rosenholtz, 1999).

Contrary to approaches that determine saliency maps through an explicit data-driven density estimation (Bruce, 2005; Dabkowski & Gal, 2017; Fong & Vedaldi, 2017; Gao & Vasconcelos, 2007; Huang & Gao, 2020; Itti & Baldi, 2005; Jiang et al., 2019; Li & Yu, 2016; Li et al., 2013; Simonyan et al., 2013; Torralba, 2003; Zhou et al., 2016), a closed form expression for saliency can be given once the feature mapping ϕ is fixed. When ϕ is continuously differentiable, the information content $-\log p_\Phi$ on $\phi(\mathbb{R}^n) \subset \mathbb{R}^N$ over all image features can be obtained from $\log p_X$ through a simple change of variables (Boothby, 1975) from x to $\phi(x)$. The amount of information for the feature vector $\phi(x)$ at every location x, and thereby the saliency $S(x)$, is then given by the expression:

$$-\log p_\Phi(\phi(x)) = -\log \frac{p_X(x)}{\sqrt{\det(J_\phi^t(x) J_\phi(x))}} \tag{3}$$

$$= -\log p_X(x) + \frac{1}{2} \log \det(J_\phi^t(x) J_\phi(x)), \tag{4}$$

where $J_\phi : \mathbb{R}^n \to \mathbb{R}^{N \times n}$ denotes the Jacobian matrix of ϕ, and $\_^t$ indicates matrix transposition.

A crucial implication of this result is that the computation of saliency can be performed based on purely *local* measurements and without the need to refer to previously observed data or any explicit

Algorithm 1 Temporal Saliency in Probabilistic Forecasting

```
Require: The probability density function $p_{Y|X}$
Input: A fixed $\boldsymbol{t}_{\text{fut}}$ of interest, a sequence of $m$ preceding observed windows $O = [\boldsymbol{t}^1_{\text{obs}}, \ldots, \boldsymbol{t}^m_{\text{obs}}]$, and the
    behavioral features $\boldsymbol{X}^j$ for every $\boldsymbol{t}^j_{\text{obs}}$
Output: The saliency map $S(O)$ over the observed windows

for each $\boldsymbol{t}^j_{\text{obs}} \in O$ do
    Compute the feature mapping $\phi(\boldsymbol{t}^j_{\text{obs}}) \leftarrow h(Y|X = \boldsymbol{X}^j)$
end for
Compute the saliency map $S(\boldsymbol{t}_{\text{obs}}) \leftarrow \det(J^t_\phi(\boldsymbol{t}_{\text{obs}}) J_\phi(\boldsymbol{t}_{\text{obs}}))$
```

density estimate of these. Following Equation 4, Loog simplifies the definition of the saliency map to

$$S(x) \coloneqq \det(J^t_\phi(x) J_\phi(x)), \tag{5}$$

given that a monotonic transformation of the saliency map does not essentially alter the map.

3 Closed-Form Temporal Saliency for Probabilistic Forecasting

3.1 Setup and Notation

Given the relevance to forecasting tasks involving human behavior, we borrow the notation from the task formulation of Social Cue Forecasting (SCF) (Raman et al., 2021). Let $\boldsymbol{t}_{\text{obs}} \coloneqq [o1, o2, ..., oT]$ denote a window of consecutively increasing observed timesteps, and $\boldsymbol{t}_{\text{fut}} \coloneqq [f1, f2, ..., fT]$ denote an unobserved future time window, with $f1 > oT$. Given a set of n interacting participants, let us denote their features over $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$ respectively as $\boldsymbol{X} \coloneqq [\boldsymbol{b}^i_t; t \in \boldsymbol{t}_{\text{obs}}]^n_{i=1}$ and $\boldsymbol{Y} \coloneqq [\boldsymbol{b}^i_t; t \in \boldsymbol{t}_{\text{fut}}]^n_{i=1}$. The vector $\boldsymbol{b}^i_t$ encapsulates the multimodal cues of interest from participant i at time t, such as head and body pose, facial expressions, gestures, etc. The task is to forecast the density $p_{Y|X}$, denoting a distribution over possible future features given an observed sequence $\boldsymbol{X}$ of the *same features*. Here Y and X denote the multivariate random variables associated with the future and observed sequences respectively. While the original formulation of SCF dealt with conversation settings, it is general enough to apply to other tasks such as pedestrian trajectory forecasting, where the features are simply the locations of individuals. In the rest of this work, for simplicity we denote the feature array at an individual timestep t in $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$ as $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ respectively, even though they correspond to the same feature types. Given $p_{Y|X}$, the focus of this work is to compute the saliency $S(\boldsymbol{t}_{\text{obs}})$ of an observed $\boldsymbol{t}_{\text{obs}}$ towards the future occurring over a fixed choice of $\boldsymbol{t}_{\text{fut}}$.

3.2 Defining $\boldsymbol{\phi}$ in Terms of the Uncertainty over the Future Window $\boldsymbol{t}_{\text{fut}}$

The closed-form expression for information-based spatial saliency in Equation 4 makes it explicit that the choice of the feature mapping ϕ determines the actual form of the saliency map. To extend the saliency framework to forecasting settings, we formalize the logical implication in Equation 1 by mapping the window $\boldsymbol{t}_{\text{obs}}$ to the differential entropy of the model's predicted future distribution over $\boldsymbol{t}_{\text{fut}}$ conditioned on the observed features $\boldsymbol{X}$. That is, we define $\phi : \boldsymbol{t}_{\text{obs}} \mapsto h(Y|X = \boldsymbol{X})$, where the conditional differential entropy of Y given $\{X = \boldsymbol{X}\}$ is defined as

$$h(Y|X = \boldsymbol{X}) \coloneqq - \int p_{Y|X}(\boldsymbol{Y}|\boldsymbol{X}) \log p_{Y|X}(\boldsymbol{Y}|\boldsymbol{X}) d\boldsymbol{Y}. \tag{6}$$

Following Equation 6, our framework for computing the saliency map is summarized in Algorithm 1. Consider that a domain expert selects a specific $\boldsymbol{t}_{\text{fut}}$ corresponding to a high-order semantic behavior they wish to analyse. Examples include a speaking-turn change (Keitel et al., 2015; Levinson & Torreira, 2015) an interaction termination (Bohus & Horvitz, 2014; van Doorn, 2018), or a synchronous behavior event (Bilakhia et al., 2013). Given an underlying forecasting model and looking back in time before $\boldsymbol{t}_{\text{fut}}$, we compute $h(Y|X = \boldsymbol{X})$ for different *observed multivariate features* $\boldsymbol{X}$ corresponding to different locations of a sliding $\boldsymbol{t}_{\text{obs}}$. The computed differential entropy values are then inserted into Equation 5 to obtain the saliency of *different* $\boldsymbol{t}_{\text{obs}}$ *locations* towards the future over the chosen $\boldsymbol{t}_{\text{fut}}$.

Differential entropy possesses favorable properties that make it a suitable choice as ϕ for computing the saliency map. First, the scale of the forecast density does not affect the resulting saliency map (see Cover & Thomas, Theorem 8.6.4):

$$h(aY) = h(Y) + \log |a|, \text{for } a \neq 0, \text{and} \tag{7}$$

$$h(\boldsymbol{A}Y) = h(Y) + \log |\det(\boldsymbol{A})|, \text{when } \boldsymbol{A} \text{ is a square matrix.} \tag{8}$$

That is, scaling the distribution changes the differential entropy by only a constant factor. So the saliency map resulting from inserting the entropy into Equation 5 remains unaffected since the Jacobian term only depends on the relative change in entropy across different choices of $\boldsymbol{t}_{\text{obs}}$. Similarly, translating the predicted density leaves the saliency map unaffected (see Cover & Thomas, Theorem 8.6.3):

$$h(Y + c) = h(Y). \tag{9}$$

3.3 COMPUTING $h(Y|X = \boldsymbol{X})$

To compute the differential entropy of the future distribution, we need the density function $p_{Y|X}$. Typically, this density is modeled as a multivariate Gaussian distribution (Raman et al., 2021; Rangapuram et al., 2018; Salinas et al., 2019a;b). When the decoding of the future is non-autoregressive, the parameters of the distributions for all $t \in \boldsymbol{t}_{\text{fut}}$ are estimated in parallel. In these cases, the differential entropy has a closed-form expression. Assuming a d-dimensional predicted Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{K}$, the expression for the entropy of a multivariate Gaussian distribution is given by (see Cover & Thomas, Theorem 8.4.1)

$$h(Y|X = \boldsymbol{X}) = h(\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{K})) = \frac{1}{2} \log[(2\pi e)^d \det(\boldsymbol{K})]. \tag{10}$$

When $\boldsymbol{K}$ is diagonal, so that the predicted distribution is factorized over participants and features, we can simply sum the $\log$ of the individual variances to obtain the feature mapping ϕ. Note that from Equation 10, for a multivariate Gaussian distribution, the differential entropy only depends on the covariance, or the *spread* of the distribution, aligning with the notion of differential entropy as a measure of total uncertainty. (See (Cover & Thomas, Tab. 17.1; Lazo & Rathie, 1978) for closed-form expressions for a large number of commonly employed probability density functions.)

Another common approach for inferring the future density function is the use of probabilistic autoregressive decoders (Ha & Eck, 2017; Raman et al., 2021; Salinas et al., 2019b; Salzmann et al., 2021). Here, one possible decoding approach (Ha & Eck, 2017; Salzmann et al., 2021) involves taking a specific sample $\widehat{\boldsymbol{Y}}_t$ from the density predicted at each $t \in \boldsymbol{t}_{\text{fut}}$, and passing it back as input to the decoder for estimating the density at timestep $t + 1$. Therefore, the density at $t + 1$ depends on the randomness introduced in sampling $\widehat{\boldsymbol{Y}}_t$. Figure 2 illustrates the concept for two timesteps. Here, a single forecast would only output the shaded red distribution for Y_2. This precludes the direct computation of $h(Y_1, Y_2)$ that requires the full joint distribution p_{Y_1,Y_2}. In such cases, we have two broad options: using a simplifying assumption to retain computational simplicity, or approximating the differential entropy by sampling.

The simpler option is to redefine our feature-mapping as $\phi : \boldsymbol{t}_{\text{obs}} \mapsto \sum_{t \in \boldsymbol{t}_{\text{fut}}} h(Y_t|\widehat{\boldsymbol{Y}}_{<t}, \boldsymbol{X})$, i.e. we sum the differential entropy of the individual densities estimated at each timestep as an approximation of the total uncertainty over the predicted sequence. Note that following the chain rule for differential entropy (see Cover & Thomas, Eq. 8.62), the joint entropy can indeed be written as the sum of individual conditionals. However, in general,

$$h(Y|X = \boldsymbol{X}) = \sum_{t \in \boldsymbol{t}_{\text{fut}}} h(Y_t|Y_{<t}, \boldsymbol{X}) \neq \sum_{t \in \boldsymbol{t}_{\text{fut}}} h(Y_t|\widehat{\boldsymbol{Y}}_{<t}, \boldsymbol{X}). \tag{11}$$

Employing this approximation (of summing the entropies across all timesteps) relies on the observation that for autoregressive decoding, the parameters of the predicted distribution for Y_t is computed as a deterministic function of the decoder hidden state. That is, Y_t is conditionally independent of $Y_{<t}$ given the hidden state of the decoder $\boldsymbol{s}_t$ at timestep t. The underlying assumption is that for a well-trained decoder, $\boldsymbol{s}_t$ encodes all relevant information from other timesteps to infer the distribution of Y_t. So at inference, despite being a function of the single sample $\widehat{\boldsymbol{Y}}_{t-1}$, the predicted distribution conditioned on $\boldsymbol{s}_t$ provides a reasonable estimate of the the uncertainty in Y_t.

The benefit of employing this assumption is that when each Y_t is modeled using a density function that has a closed-form expression for differential entropy (Cover & Thomas, Tab. 17.1; Lazo & Rathie, 1978), every item in the sum can be computed anlaytically, and we obtain a closed-form expression for the saliency map. Apart from a multivariate Gaussian, the other common choice for modeling Y_t is using a Gaussian mixture (Ha & Eck, 2017; Salzmann et al., 2021). While a closed-form expression for the differential entropy of a Gaussian mixture is not known, approximations that approach the true differential entropy can be obtained efficiently (Huber et al., 2008; Michalowicz et al., 2008; Zhang & Luo, 2017) to directly compute the feature mapping ϕ.

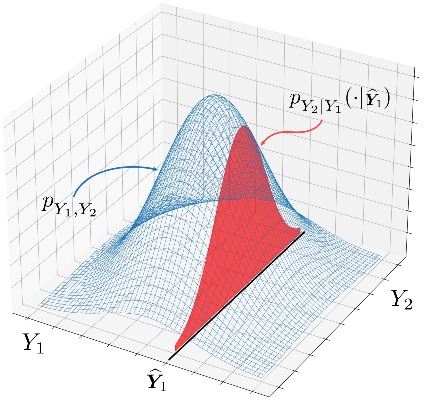


Figure 2: Illustrating predicted densities under the greedy autoregressive decoding approach for two timesteps. For simplicity, we depict a joint Gaussian distribution and omit the conditioning on $\boldsymbol{X}$ everywhere.

What if the differential entropy for Y_t does not have an analytical expression or a computationally efficient approximation? In such cases, $h(Y|X = \boldsymbol{X})$ can be estimated using sampling or other non-parametric approaches (Ajgl & Šimandl, 2011; Ariel & Louzoun, 2020; Beirlant et al., 1997; Brewer, 2017). Such sampling-based methods can provide approximations that converges to the true entropy.

4 Empirical Validation of the Proposed Framework

How do we empirically validate that our framework retrieves verifiable salient timesteps towards a future outcome? Crucially, to avoid issues arising from qualitative visual assessment of saliency maps (Adebayo et al., 2018; Barredo Arrieta et al., 2020), a human-verifiable salient relationship between a $\boldsymbol{t}_{\text{obs}}$ and $\boldsymbol{t}_{\text{fut}}$ is needed to serve as ground truth. However, this is non-trivial for real-world data; ground-truthing such saliency in behavior datasets is especially hard because there often exists both supporting and opposing evidence in social psychology for predictive relationships between behaviors (Kalma, 1992; Levinson & Torreira, 2015). So we have a *chicken-egg* problem surrounding evaluation on real-world data: if we had a way of computing such ground-truth predictive saliency, we would not need this research to begin with. Furthermore, even if this is addressed, using real-world data would also entail training a forecasting model on the data. This model is bound to be arbitrarily imperfect, resulting in a correspondingly imperfect saliency map from our framework (garbage in, garbage out). So, when using an imperfect model, do we attribute any flaws in the saliency map to the explainability method or to the imperfect underlying forecasts? This also makes it non-trivial to compare against other gradient and attention-based XAI techniques that require training a specific model to access its parameters. As such, for a fair empirical comparison, any candidate XAI method would need to i. compute saliency by changing observed timesteps one-at-a-time (counterfactual reasoning); and ii. evaluate change in the information in the probablistic forecast to measure the same notion of saliency. No such directly comparable saliency-based explanation frameworks exist for probabilistic forecasts. The closest saliency-based method (Pan et al., 2021) computes saliency maps using random perturbations for non-probabilistic forecasts and requires access to model parameters.

Consequently, our experimental setup is as follows. First, we construct a synthetic dataset of conversation dynamics where a high-order semantic predictive relationship exists between behaviors. Next we assume a perfect underlying forecasting model trained over this dataset: the best a well-trained model can do is to predict the true variance over plausible futures given an observation. Note that the model does not need to have any conception of saliency or high-order semantic concepts. The task is to predict the distribution $p_{Y|X}$ over low-level head pose features denoted by real-valued quaternions, matching the real-world scenario (Barquero et al., 2022; Raman et al., 2021). Our goal is to evaluate if our framework empirically retrieves the timesteps that are salient by construction.

4.1 The Dataset

We choose to synthesize turn-taking dynamics in multi-party conversations, a setting that has received extensive domain interest over the last decade (De Kok & Heylen, 2009; Ishii et al., 2013; 2017; Keitel et al., 2015; Malik et al., 2020; Petukhova & Bunt, 2009; Rochet-Capellan & Fuchs, 2014).

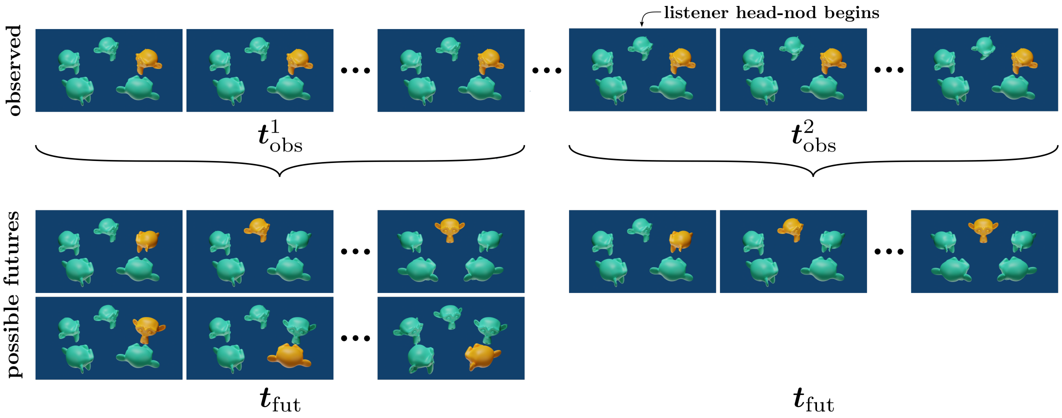


Figure 3: Illustrating the synthetic conversation dynamics dataset. Speakers are denoted in orange and listeners in green. For a fixed $\boldsymbol{t}_{\text{fut}}$ we depict two preceding $\boldsymbol{t}_{\text{obs}}$ windows. By construction, when observing a stable speaking turn over $\boldsymbol{t}^1_{\text{obs}}$, two valid futures are possible over $\boldsymbol{t}_{\text{fut}}$. These correspond to a turn handover to the immediate left or right of the current speaker. Over $\boldsymbol{t}^2_{\text{obs}}$, when a listener nods to indicate the desire to take the floor, the future over $\boldsymbol{t}_{\text{fut}}$ becomes certain, corresponding to the listener successfully taking over the speaking turn. Here $\boldsymbol{t}^2_{\text{obs}}$ is consequently more salient than $\boldsymbol{t}^1_{\text{obs}}$ towards forecasting the turn change over $\boldsymbol{t}_{\text{fut}}$.

Specifically, we synthesize a group conversation following established patterns of social behavior. First, the visual focus of attention of listeners is usually the speaker, while the speaker might look at different listeners over the speaking turn (Hung et al., 2008). Second, head gestures and gaze patterns are predictive of the next speaker (De Kok & Heylen, 2009; Ishii et al., 2017; Malik et al., 2020; Petukhova & Bunt, 2009). We implement a simplified version of these dynamics as follows. The speaker rotates towards the geometric center of the formation after acquiring a speaking turn, while the listeners orient towards the speaker. A listener nods their head to indicate a desire to acquire the floor, following which the current speaker rotates to look at the listener and hands over the speaking turn. We represent 3D head poses with quaternions given their common use in representing human motion and pose (Pavllo et al., 2018; Raman et al., 2021). Concretely following the notation in subsection 3.1, $\boldsymbol{b} = [q_w, q_x, q_y, q_z]$. We simulate the turn changes to occur once clockwise and once anticlockwise around the group so that each participant yields the floor once each to the participant on their immediate left and right. We provide code, the dataset consisting of the head-pose and speaking status features, and an animated visualization in the Supplementary material.

4.2 Computing Temporal Saliency

By construction, for a human observer it is impossible to guess the future with certainty by simply observing the current speaker speaking: there exist two valid examples of the future head behavior in the dataset for such a sequence. Only the commencement of a head nod by a listener makes the future turn handover certain. Figure 3 illustrates this mechanism. Therefore, the timestep corresponding to a head nod is salient towards forecasting the features over a future window involving a turn change in this dataset. Can our framework identify this timestep as salient? A perfect model would predict the true variance over the two valid future quaternion trajectories. This variance vanishes when the future becomes certain as soon as one of the listeners starts nodding. We model the future distribution using a Gaussian function for simplicity (setting std. to 10^{-10} for the single future), but a more complex distribution that predicts the appropriate change in variance would also work in practice.

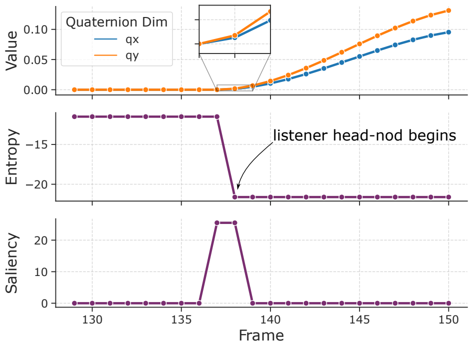


Figure 4: Computing Saliency. We plot the quaternion dimensions qx and qy for the listener that nods over $\boldsymbol{t}^2_{\text{obs}}$ in Figure 3 (top). The observation of the head-nod beginning in frame 138 makes the future over frames $183 - 228$ ($\boldsymbol{t}_{\text{fut}}$) certain. This is reflected in the reduction in the mean entropy over future values of all participants (middle). The saliency map obtained using our framework (bottom) correctly identifies the beginning of the head-nod in frame 138 as the salient timestep towards forecasting the future over $\boldsymbol{t}_{\text{fut}}$.

We now implement Algorithm 1 as follows. We identify a window where a turn change occurs in the data (frames 183-228) and denote this 45 frame window as the $\boldsymbol{t}_{\text{fut}}$ of interest. While we manually identify an interesting event for illustration, such a window could also correspond to an interesting prediction by a model. We generate a set of candidate $\boldsymbol{t}_{\text{obs}}$ by sliding a 30 frame window over a horizon of 100 frames prior to $\boldsymbol{t}_{\text{fut}}$, with a stride of 1 frame. For every observed $\boldsymbol{t}_{\text{obs}}$, we fit a Gaussian density to the quaternion and speaking status features of all participants over the futures that can occur during $\boldsymbol{t}_{\text{fut}}$. We then set the entropy of this Gaussian density as the feature ϕ for that $\boldsymbol{t}_{\text{obs}}$. We obtain the saliency map using Equation 5, and plot the listener features, entropy, and saliency across timesteps in Figure 4. The ground truth salient timesteps here are frames 138 and 139 where the head nod begins. Our framework retrieves these timesteps as salient with 100% accuracy, empirically validating our framework. Note that once the nod is already in motion, the saliency drops as expected since the future is then already certain given the data. So observing a head nod in motion does not provide any additional information about the future.

5 Discussion and Conclusion: Saliency & XAI with Domain Experts in the Loop

We have proposed a computational framework obtaining counterfactual explanations of model forecasts and established its link to a principled notion of bottom-up task-agnostic saliency. Specifically, we obtain a convenient closed-form expression of the saliency of observed timesteps towards a future outcome for commonly used probability density functions to represent forecasts (Ha & Eck, 2017; Raman et al., 2021; Rudenko et al., 2020; Salzmann et al., 2021). We empirically validate our framework using a synthetic experiment that allows for quantitative validation, avoiding established issues arising from observer biases in visual assessment of saliency maps (Adebayo et al., 2018; Barredo Arrieta et al., 2020). Our goal with this work is to aid domain experts in forming new hypothses from real-world complex data. Given a model that has learned structural predictive relationships from data, our framework identifies timesteps salient towards the model's forecasts through counterfactual reasoning. We thereby envision a human-in-the-loop XAI methodology where the domain expert can form hypotheses surrounding the *features* occurring at these salient timesteps. These hypotheses can then be verified through subsequent controlled experiments.

Loog's (2011) unifying framework subsumes all forms of saliency, although identifying the appropriate ϕ for a specific domain is non-trivial. In this work we have established both theoretically and empirically how expressing ϕ in terms of the information about the future enables principled counterfactual reasoning in forecasting settings. Nevertheless, we reiterate that the salient timesteps retrieved by our framework ought to be treated as *candidate* causes until subsequently examined along with a domain expert. Our stance on human-in-the-loop XAI also aligns with research on saliency-based and general XAI in other domains (Atrey et al., 2020; Joshi et al., 2021).

In principle, when it is possible to have access to the true $p_{Y|X}$, the salient timesteps identified by our framework reflect the *true* predictive structural relationships captured by the underlying model across the entire data. However, estimating this density analytically entails identifying the multiple futures in the data corresponding to every occurrence of the same observed features. In practice, subtle variations in behaviors and sensor measurement errors make it infeasible to estimate $p_{Y|X}$ analytically, so a model is trained to capture generalized patterns from the given data. In these cases, our framework identifies the sequences that *the model considers salient* for its forecasts *given the data*. Consequently, subsequent causal analysis of the features over the salient timesteps is crucial, especially in the healthcare and human behavior domains to avoid potential prejudices against certain behaviors, or worse, misdiagnoses of affective conditions.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. Socially and Contextually Aware Human Motion and Pose Forecasting. *arXiv:2007.06843 [cs]*, July 2020.

Chaitanya Ahuja, Shugao Ma, Louis-Philippe Morency, and Yaser Sheikh. To React or not to React: End-to-End Visual Pose Forecasting for Personalized Avatar during Dyadic Conversations. *arXiv:1910.02181 [cs]*, October 2019.

Jiří Ajgl and Miroslav Šimandl. Differential entropy estimation by particles. *IFAC Proceedings Volumes*, 44(1): 11991–11996, January 2011. ISSN 1474-6670. doi: 10.3182/20110828-6-IT-1002.01404.

Ahmed M Alaa and Mihaela van der Schaar. Attentive state-space modeling of disease progression. *Advances in neural information processing systems*, 32, 2019.

Gil Ariel and Yoram Louzoun. Estimating differential entropy using recursive copula splitting. *Entropy. An International and Interdisciplinary Journal of Entropy and Information Studies*, 22(2):236, 2020.

Akanksha Atrey, Kaleigh Clary, and David Jensen. Exploratory Not Explanatory: Counterfactual Analysis of Saliency Maps for Deep Reinforcement Learning, February 2020.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, and Katja Hansen. How to Explain Individual Classification Decisions. pp. 29.

Sarah Adel Bargal, Andrea Zunino, Donghyun Kim, Jianming Zhang, Vittorio Murino, and Stan Sclaroff. Excitation Backprop for RNNs. *arXiv:1711.06778 [cs]*, March 2018.

German Barquero, Johnny Núñez, Sergio Escalera, Zhen Xu, Wei-Wei Tu, Isabelle Guyon, and Cristina Palmero. Didn't see that coming: A survey on non-verbal social human behavior forecasting. *arXiv:2203.02480 [cs]*, March 2022.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020. ISSN 1566-2535. doi: 10.1016/j.inffus.2019.12.012.

Jan Beirlant, Edward J Dudewicz, László Györfi, Edward C Van der Meulen, et al. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39, 1997.

Sanjay Bilakhia, Stavros Petridis, and Maja Pantic. Audiovisual Detection of Behavioural Mimicry. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 123–128, Geneva, Switzerland, September 2013. IEEE. ISBN 978-0-7695-5048-0. doi: 10.1109/ACII.2013.27.

Dan Bohus and Eric Horvitz. Managing Human-Robot Engagement with Forecasts and… um… Hesitations. *Proceedings of the 16th International Conference on Multimodal Interaction*, pp. 8, 2014.

WM Boothby. An introduction to differential geometry and riemannian manifolds, 1975.

Brendon J Brewer. Computing entropies with nested sampling. *Entropy. An International and Interdisciplinary Journal of Entropy and Information Studies*, 19(8):422, 2017.

Neil DB Bruce. Features that draw visual attention: An information theoretic perspective. *Neurocomputing*, 65: 125–133, 2005.

Sandra Carrasco, David Fernández Llorca, and Miguel Ángel Sotelo. SCOUT: Socially-COnsistent and UndersTandable Graph Attention Network for Trajectory Prediction of Vehicles and VRUs. *arXiv:2102.06361 [cs]*, May 2021.

Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. *arXiv:1608.05745 [cs]*, August 2016.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*.

Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. *Advances in neural information processing systems*, 30, 2017.

Iwan De Kok and Dirk Heylen. Multimodal end-of-turn prediction in multi-party meetings. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, pp. 91–98, 2009.

Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3429–3437, 2017.

Dashan Gao and Nuno Vasconcelos. Bottom-up saliency is a discriminant process. In *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–6. IEEE, 2007.

Jose A. Garcia, Joaquín Fdez-Valdivia, Xose R Fdez-Vidal, and Rosa Rodriguez-Sanchez. Information theoretic measure for visual target distinctness. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23 (4):362–383, 2001.

Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. THOMAS: Trajectory Heatmap Output with learned Multi-Agent Sampling. *arXiv:2110.06607 [cs]*, January 2022.

David Ha and Douglas Eck. A Neural Representation of Sketch Drawings. *arXiv:1704.03477 [cs, stat]*, May 2017.

Kun Huang and Shenghua Gao. Image saliency detection via multi-scale iterative CNN. *The Visual Computer*, 36(7):1355–1367, 2020.

Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. STGAT: Modeling Spatial-Temporal Interactions for Human Trajectory Prediction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6271–6280, Seoul, Korea (South), October 2019. IEEE. ISBN 978-1-72814-803-8. doi: 10.1109/ICCV.2019.00637.

Marco F. Huber, Tim Bailey, Hugh Durrant-Whyte, and Uwe D. Hanebeck. On entropy approximation for Gaussian mixture random vectors. In *2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 181–188, Seoul, August 2008. IEEE. ISBN 978-1-4244-2143-5. doi: 10.1109/MFI.2008.4648062.

Hayley Hung, Dinesh Babu Jayagopi, Sileye Ba, Jean-Marc Odobez, and Daniel Gatica-Perez. Investigating automatic dominance estimation in groups from visual attention and speaking activity. pp. 233. ACM Press, 2008. ISBN 978-1-60558-198-9. doi: 10.1145/1452392.1452441.

Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, Masafumi Matsuda, and Junji Yamato. Predicting next speaker and timing from gaze transition patterns in multi-party meetings. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, pp. 79–86, 2013.

Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. Prediction of Next-Utterance Timing using Head Movement in Multi-Party Meetings. In *Proceedings of the 5th International Conference on Human Agent Interaction*, HAI '17, pp. 181–187, New York, NY, USA, October 2017. Association for Computing Machinery. ISBN 978-1-4503-5113-3. doi: 10.1145/3125739.3125765.

Aya Abdelsalam Ismail, Mohamed Gunady, Héctor Corrada Bravo, and Soheil Feizi. Benchmarking Deep Learning Interpretability in Time Series Predictions. *arXiv:2010.13924 [cs, stat]*, October 2020.

Laurent Itti and Pierre Baldi. A principled approach to detecting surprising events in video. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pp. 631–637. IEEE, 2005.

Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3): 194–203, 2001.

Boris Ivanovic, Yifeng Lin, Shubham Shrivastava, Punarjay Chakravarty, and Marco Pavone. Propagating state uncertainty through trajectory forecasting. *arXiv preprint arXiv:2110.03267*, 2021.

Lai Jiang, Zhe Wang, Mai Xu, and Zulin Wang. Image saliency prediction in transformed domain: A deep complex neural network method. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8521–8528, 2019.

Shalmali Joshi, Sonali Parbhoo, and Finale Doshi-Velez. Pre-emptive learning-to-defer for sequential medical decision-making under uncertainty, September 2021.

Akko Kalma. Gazing in triads: A powerful signal in floor apportionment. *British Journal of Social Psychology*, 31(1):21–39, March 1992.

Anne Keitel, Moritz M Daum, et al. The use of intonation for turn anticipation in observed conversations without visual signals as source of information. *Frontiers in psychology*, 6(108):265–273, 2015.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. *Nature Communications*, 10(1):1096, December 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08987-4.

A Verdugo Lazo and P Rathie. On the entropy of continuous probability distributions (corresp.). *IEEE Transactions on Information Theory*, 24(1):120–122, 1978.

Stephen C. Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6, June 2015. ISSN 1664-1078. doi: 10.3389/fpsyg.2015.00731.

Guanbin Li and Yizhou Yu. Visual saliency detection based on multiscale deep CNN features. *IEEE transactions on image processing*, 25(11):5012–5024, 2016.

Longyuan Li, Junchi Yan, Xiaokang Yang, and Yaohui Jin. Learning Interpretable Deep State Space Model for Probabilistic Time Series Forecasting. *arXiv:2102.00397 [cs, stat]*, January 2021.

Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense and sparse reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2976–2983, 2013.

Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting. *arXiv:1912.09363 [cs, stat]*, September 2020.

Zhang Lingyun, Matthew H Tong, and Garrison W Cottrell. Information attracts attention: A probabilistic account of the cross-race advantage in visual search. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29, 2007.

Zachary C. Lipton. The Mythos of Model Interpretability, March 2017.

Zachary C. Lipton and Jacob Steinhardt. Troubling Trends in Machine Learning Scholarship, July 2018.

Marco Loog. Information theoretic preattentive saliency: A closed-form solution. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 1418–1424, Barcelona, Spain, November 2011. IEEE. ISBN 978-1-4673-0063-6 978-1-4673-0062-9 978-1-4673-0061-2. doi: 10.1109/ICCVW.2011.6130417.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Usman Malik, Julien Saunier, Kotaro Funakoshi, and Alexandre Pauchet. Who speaks next? Turn change and next speaker prediction in multimodal multiparty interaction. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 349–354. IEEE, 2020.

Joseph V. Michalowicz, Jonathan M. Nichols, and Frank Bucholtz. Calculation of Differential Entropy for a Mixed Gaussian Distribution. *Entropy*, 10(3):200–206, September 2008. ISSN 1099-4300. doi: 10.3390/entropy-e10030200.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267: 1–38, February 2019. ISSN 0004-3702. doi: 10.1016/j.artint.2018.07.007.

Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction. *arXiv:2002.11927 [cs]*, February 2020.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

Felix Mujkanovic, Vanja Doskoč, Martin Schirneck, Patrick Schäfer, and Tobias Friedrich. timeXplain – A Framework for Explaining the Predictions of Time Series Classifiers. *arXiv:2007.07606 [cs, stat]*, July 2020.

Tan Viet Tuyen Nguyen and Oya Celiktutan. Context-Aware Human Behaviour Forecasting in Dyadic Interactions. *Understanding Social Behavior in Dyadic and Small Group Interactions*, pp. 1–19, 2022.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *arXiv:1905.10437 [cs, stat]*, February 2020.

Qingyi Pan, Wenbo Hu, and Ning Chen. Two Birds with One Stone: Series Saliency for Accurate and Interpretable Multivariate Time Series Forecasting. In *Twenty-Ninth International Joint Conference on Artificial Intelligence*, volume 3, pp. 2884–2891, August 2021. doi: 10.24963/ijcai.2021/397.

Madhurima Panja, Uttam Kumar, and Tanujit Chakraborty. An Interpretable Probabilistic Autoregressive Neural Network Model for Time Series Forecasting. *arXiv:2204.09640 [cs, stat]*, April 2022.

Dario Pavllo, David Grangier, and Michael Auli. QuaterNet: A Quaternion-based Recurrent Model for Human Motion. *arXiv:1805.06485 [cs]*, July 2018.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic books, 2018.

Volha Petukhova and Harry Bunt. Who's next? Speaker-selection mechanisms in multiparty dialogue. In *Workshop on the Semantics and Pragmatics of Dialogue*, 2009.

Chirag Raman, Hayley Hung, and Marco Loog. Social Processes: Self-Supervised Meta-Learning over Conversational Groups for Forecasting Nonverbal Social Cues. *arXiv:2107.13576 [cs]*, July 2021.

Syama Sundar Rangapuram, Matthias W Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. Deep state space models for time series forecasting. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Amélie Rochet-Capellan and Susanne Fuchs. Take a breath and take the turn: How breathing meets turns in spontaneous dialogue. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1658), December 2014. ISSN 0962-8436. doi: 10.1098/rstb.2013.0399.

Ruth Rosenholtz. A simple saliency model predicts a number of motion popout phenomena. *Vision research*, 39 (19):3157–3163, 1999.

Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. Human motion trajectory prediction: A survey. *The International Journal of Robotics Research*, 39(8):895–935, 2020.

David Rügamer, Philipp F. M. Baumann, Thomas Kneib, and Torsten Hothorn. Probabilistic Time Series Forecasts with Autoregressive Transformation Models. *arXiv:2110.08248 [cs]*, February 2022.

David Salinas, Michael Bohlke-Schneider, Laurent Callot, Roberto Medico, and Jan Gasthaus. High-dimensional multivariate forecasting with low-rank Gaussian Copula Processes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a.

David Salinas, Valentin Flunkert, and Jan Gasthaus. DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks. *arXiv:1704.04110 [cs, stat]*, February 2019b.

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-Feasible Trajectory Forecasting With Heterogeneous Data. *arXiv:2001.03093 [cs]*, January 2021.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMLR, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding using deep networks. *arXiv preprint arXiv:1705.08498*, 2017.

Antonio Torralba. Modeling global scene factors in attention. *JOSA A*, 20(7):1407–1418, 2003.

Felix van Doorn. Rituals of Leaving: Predictive Modelling of Leaving Behaviour in Conversation. *Master of Science Thesis, Delft University of Technology*, 2018.

AHC vanderHeijden. Perception for selection, selection for action, and action for perception, 1996.

Anirudh Vemula, Katharina Muelling, and Jean Oh. Modeling cooperative navigation in dense human crowds. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1685–1692. IEEE, 2017.

Kevin N Walker, Timothy F Cootes, and Christopher J Taylor. Locating salient object features. In *BMVC*, volume 98, pp. 557–566. Citeseer, 1998.

Taiping Yao, Minsi Wang, Bingbing Ni, Huawei Wei, and Xiaokang Yang. Multiple Granularity Group Interaction Prediction. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2246–2254, Salt Lake City, UT, June 2018. IEEE. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00239.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833. Springer, 2014.

Wenyuan Zeng, Shenlong Wang, Renjie Liao, Yun Chen, Bin Yang, and Raquel Urtasun. DSDNet: Deep Structured self-Driving Network. *arXiv:2008.06041 [cs]*, August 2020.

Chenchen Zhang and Yuan Luo. Approximating the differential entropy of gaussian mixtures. In *GLOBECOM 2017-2017 IEEE Global Communications Conference*, pp. 1–6. IEEE, 2017.

Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. SR-LSTM: State Refinement for LSTM towards Pedestrian Trajectory Prediction. *arXiv:1903.02793 [cs]*, March 2019.

Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Benjamin Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, Congcong Li, and Dragomir Anguelov. TNT: Target-driveN Trajectory Prediction. *arXiv:2008.08294 [cs]*, August 2020.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.

A Broader Related Work: Explainable Methods for Time-Series Tasks Across Domains

The larger focus of explainability techniques involving time-series data has been on the task of **classifying** time-series. The goal has been to estimate the relevance of each input feature at a given timestep towards each output class. Here, saliency approaches often overlap with techniques developed for image data and can be categorized into:

Gradient-Based Techniques. The broad approach involves evaluating the gradient of the output class with respect to the input (Baehrens et al.). Several variants have been proposed (Bargal et al., 2018; Lundberg & Lee, 2017; Shrikumar et al., 2017; Smilkov et al., 2017; Sundararajan et al., 2017).

Perturbation-Based Techniques. The idea is to examine how the output changes in response to some perturbation of the input. Perturbations are implemented by either occluding contiguous regions of the input (Mujkanovic et al., 2020; Zeiler & Fergus, 2014); performing an ablation of the features (Suresh et al., 2017); or randomly permuting features (Molnar, 2020). Ismail et al. (2020) provide a benchmark of a subset of these techniques.

Attention-Based Techniques. These incorporate an attention mechanism into the model that is trained to attribute importance to previous parts of the input towards an output at each timestep. Such techniques have been specially employed for healthcare data, with early methods applying a reverse-time attention (Choi et al., 2016), and later ones applying the attention to probabilistic state-space representations (Alaa & van der Schaar, 2019).

Some of these broad ideas have been applied to the **regression** setting to make interpretable forecasts of future time-series features. Lim et al. (2020) leveraged self-attention layers for capturing long-term dependencies. Pan et al. (2021) recently proposed computing saliency as a mixup strategy between series images and their perturbed version with a learnable mask for each sample. They view saliency in terms of minimizing the mean squared error between the predictions and ground-truths for a particular instance. Focusing on the univariate point-forecasting problem, Oreshkin et al. (2020) proposed injecting inductive biases by computing the forecast as a combination of a trend and seasonality model. They argue that this decomposition make the outputs more interpretable.

Developing explainable techniques for the probabilistic forecasting setting remains largely unexplored and subject to non-overlapping notions of explainability. Rügamer et al. (2022) transform the forecast using predefined basis functions such as Bernstein polynomials. They relate interpretability to the coefficients of these basis functions (a notion similar to that of Oreshkin et al. (2020)). Panja et al. (2022) embed the classical linear ARIMA model into a non-linear autoregressive neural network for univariate probabilistic forecasting. As before, the explainability here also stems from the 'white-box' nature of the linear ARIMA component. Li et al. (2021) propose an automatic relevance determination network to identify useful exogenous variables (i.e. variables that can affect the forecast without being a part of the time-series data). To the best of our knowledge, saliency-based methods have not yet been considered within this setting.