Distribution-aware Evaluation of Multimodal Trajectory Predictions with Energy Score[⋆]

Novin Shahroudi[1][0000−0002−3121−7603] and Meelis Kull[1][0000−0001−9257−595X]

University of Tartu, Tartu 51009, Estonia
{novin.shahroudi, meelis.kull}@ut.ee
https://ml.cs.ut.ee

Abstract. Trajectory Prediction has received much attention in recent years due to the deployment of autonomous vehicles in real-world scenarios. With it, many works have addressed challenges in producing more reliable trajectory predictions. While modeling and representation of trajectory prediction have evolved to address its spatiotemporal complexities, its evaluation has primarily remained primitive. The most current methods provide rich probabilistic spatiotemporal outputs. However, the evaluation metrics used to assess such rich outputs evaluate limited aspects of the predictive distribution, making evaluation and comparison of models uninformative and sometimes even misleading. We propose using the Energy Score as a distribution-aware alternative to frequently used metrics in the trajectory prediction literature, such as minFDE and minADE. Energy Score is a strictly proper scoring rule more commonly known and used in the forecasting community. We present the formulation of the energy score for spatiotemporal data and showcase its capabilities through preliminary empirical results supporting our proposal. By introducing the energy score as an alternative evaluation metric for trajectory predictions, we aim to enhance the assessment of trajectory prediction models and foster more informative and reliable comparisons among different approaches.

Keywords: Trajectory Prediction · Energy Score · Distribution-aware Evaluation.

1 Introduction

Trajectory prediction has gained significant attention in recent years due to the increasing deployment of autonomous vehicles in real-world scenarios. As a result, numerous studies have focused on addressing the challenges associated with generating more reliable trajectory predictions [1,6,10–12,14,16,18]. While the modeling and representation of trajectory prediction have evolved to handle the spatiotemporal complexities involved, the evaluation of these predictions has remained relatively primitive. Current trajectory prediction methods often provide

[⋆] Supported by University of Tartu.

rich probabilistic spatiotemporal outputs. However, the evaluation metrics commonly used to assess such outputs, such as minimum Final Displacement Error (minFDE) and minimum Average Displacement Error (minADE), evaluate only limited aspects of the predictive distribution. [15] provides a summary showing that most of the works rely on these metrics, and the trend has continued to this date. As an example, the leaderboard of ETH/UCY[1], one of the most famous datasets, ranks the state-of-the-art (SOTA) methods based on these metrics. A common choice of evaluation seems to have been stuck with the literature from the initial work that introduced the dataset [4]. Limitations of these metrics can lead to uninformative and sometimes misleading evaluation and comparison of different trajectory prediction models as pointed out in [8].

Inspired by advancements in the forecasting literature, particularly scenario forecasting, we propose using Energy Score as a distribution-aware alternative to conventional evaluation metrics in trajectory prediction. Energy Score has a relatively long history in the forecasting community as a multivariate scoring rule, and one of the latest studies shows its capabilities compared to other multivariate measures [19]. It offers a more comprehensive evaluation of the predictive distribution by capturing its probabilistic characteristics. By incorporating it as an evaluation metric for trajectory predictions, we aim to enhance the assessment of multimodal trajectory prediction models to enable a more informative and reliable comparison among different approaches. In this paper, we present different formulations of the Energy Score specifically tailored for spatiotemporal data. Our main contributions can be summarized as follows:

1. Proposing three different variations of the Energy Score for evaluating multimodal trajectory predictions
2. Designing an experimental setup showcasing the desired behavior of the Energy Score in evaluating multimodal trajectory predictions, demonstrating its potential as a more powerful alternative evaluation metric.

The code to our experiments is available at https://github.com/novinsh/trajectory_prediction_evaluation_via_energy_score.

2 Background

Trajectory prediction is relevant in many contexts and applications, such as in robotics and autonomous vehicles. For example, an autonomous vehicle has to interact with the environment where other agents and objects could be present, and predicting their trajectories is crucial to operating in an environment. Cognition and planning tasks such as collision avoidance rely on trajectory prediction. This is even more important in safety-critical environments as the cost of inaccurate trajectory prediction could be detrimental to the relevant downstream tasks/decisions, or similarly, the benefit of an accurate one could be highly rewarding. Here we provide a formal definition of the terms central to our work.

[1] https://paperswithcode.com/sota/trajectory-prediction-on-ethucy

Trajectory A *trajectory* consists of a sequence of spatial variables with the dimensionality of S, encoding the coordinates of an agent throughout time. The sequence can be divided into past and future: $\{Y_i^{\text{past}}, Y_i^{\tau}\}$, where $Y_i^{\text{past}} = \{Y_i^t | t \in [1, T_{obs}]\}$ is the observed trajectory with T_{obs} time steps, and $Y_i^{\tau} = \{Y_i^t | t \in \tau\}$ is the future path with T_{pred} time steps, $\tau = (T_{obs}, T_{obs} + T_{pred}]$, and i is the index of the trajectories from N unique or non-unique agents. In our work, $Y_i^t \in \mathbb{R}^2$ because we consider $S = 2$. A trajectory can be seen as a multivariate time series, which in our case, consists of two time series coupled based on a spatial dependency.

Trajectory Prediction A *trajectory prediction* aims to predict K future trajectories $X_i^{\tau} = \{X_{i,k}^t | k = 1, \ldots, K\}$ from the observed information in the past or available cues as covariates in τ. $\mathbf{X}_i$ is a spatiotemporal random variable, and in this work, we consider the K trajectories to be independent samples of the predictive joint distribution $\mathbf{F}_{\mathbf{X_i}}(x)$ over the future trajectories [2]. $\mathbf{F}_{\mathbf{X_i}}(x)$ is a $T_{pred} \times S$ multivariate distribution, and hence to be able to capture its modalities, K must be sufficiently large. $K = 1$ corresponds to a deterministic forecast as opposed to $K > 1$ as a multimodal (probablistic) forecast.

Trajectory Evaluation A *trajectory evaluation* is a process to assess how well trajectory predictions X_i^{τ} follow the ground truth Y_i^{τ}. A distribution-aware evaluation aims at minimizing $|\mathbf{F}_{\mathbf{X_i}}(x) - \mathbf{G}_{\mathbf{Y_i}}(y)|$, which is some measure of distance between the two distributions that would inform the quality of the predicted trajectory distribution.

Common Evaluation Metrics In reality, we observe one sample $\mathbf{y}$ per each agent for its future ground truth that is $T_{pred} \times S$ dimensional. Final Displacement Error (FDE) and Average Displacement Error (ADE) are two of the most common measures of trajectory prediction's quality that are not distribution-aware but can be applied to the samples obtained from the predictive distribution $\mathbf{F}_{\mathbf{X}}$ are as follows:

$$\begin{aligned} DE(\mathbf{F}_{\mathbf{X}}, \mathbf{y}) =& \|X - y\|_2 \qquad (1) \\ ADE(\mathbf{F}_{\mathbf{X}}, \mathbf{y}) =& \mathbb{E}_{i,k,t \in \tau} DE(X_{i,t}^t, y_i^t) \qquad (2) \\ FDE(\mathbf{F}_{\mathbf{X}}, \mathbf{y}) =& \mathbb{E}_{i,k,t=T} DE(X_{i,k}^t, y_i^t) \qquad (3) \end{aligned}$$

$$\begin{aligned} minADE(\mathbf{F}_{\mathbf{X}}, \mathbf{y}) =& \mathbb{E}_{i,t \in \tau} \min_{k<K} DE(X_{i,k}^t, y_i^t) \qquad (4) \\ minFDE(\mathbf{F}_{\mathbf{X}}, \mathbf{y}) =& \mathbb{E}_{i,t=T} \min_{k<K} DE(X_{i,k}^t, y_i^t) \qquad (5) \end{aligned}$$

[2] Broadly speaking the predictive distribution can be represented explicitly or implicitly with different computational and approximation implications depending on the modeling technique. In either case, the K i.i.d. samples of the predictive distribution allow for a universal and non-parametric evaluation, however, the sampling itself could be a costly process, but that is out of the scope of our work.

The minimum versions in Eq. 4 and 5 are referred to as Minimum of N (MoN) and introduced by [6], and they are categorized as *lower-bound metrics* by [8]. They are also referred to as topK and topK%, where in the latter, instead of a specific number, a percentage of the K trajectories are chosen for the error calculation. The final displacement error variations only consider the last timestep of the trajectory $T = T_{obs} + T_{pred}$.

Proper Scoring Rule A *scoring rule* provides a summary measure for evaluating probabilistic predictions. At the same time, a scoring rule could be considered a cost or loss function that evaluates a prediction and outputs a cost or score. The score could be used to compare different predictions or prediction models with one another. The minimum score is realized when the true set of probabilities is reported by the prediction model. A scoring rule is proper if it satisfies this property as defined in Definition 1. We provide this definition based on [5].

Definition 1. *A negatively oriented strictly proper scoring rule* $\mathbf{S}$ *maps a probability distribution* F *and a realization* y *to a real number, i.e.,* $\mathbf{S}(F, y)$*. The expected value of* $\mathbf{S}(F, .)$ *under* Q*, is written as* $\mathbf{S}(F, G)$ *where* $y \sim G$*. A scoring rule is strictly proper if* $\mathbf{S}(F, G) \geq \mathbf{S}(G, G)$ *with the equality if and only if* $F = Q$*, and proper if* $\mathbf{S}(F, G) \geq \mathbf{S}(G, G)$ *for all* F *and* G*.*

Because of this property, a strictly proper scoring rule is not only useful for evaluation but also for learning and optimization of a probabilistic model as it encourages the model towards optimal prediction.

3 Energy Score as a Distribution-aware Metric

Energy Score was introduced by [5] and they showed that it is a multivariate generalization of Continuous Ranked Probability Score (CRPS) [7,13] and it is a strictly proper scoring rule. Even though CRPS is suitable for probabilistic evaluation, it is a univariate scoring rule and hence not appropriate for evaluating multivariate time series. Energy Score is rooted in Energy Distance formulated originally by [17], where they provide a unifying theory of energy statistics and discuss different properties and use cases of the energy distance in depth. They adopted the name energy inspired by the notion of Newton's gravitational potential energy, which is a function of the distance between two bodies. Similarly, energy statistics are functions of the distance between two statistical observations, with the statistical potential energy being zero if and only if the underlying null hypothesis is true.

3.1 Energy Score for Trajectory Evaluation

Since trajectories are essentially a multivariate time series, we can adopt energy score as a multivariate scoring rule to evaluate trajectories. In the time series forecasting literature, the energy score is usually applied to a single time series, reporting on the marginal and temporal interdependence of the time series. It is deemed a multivariate evaluation since the temporal variables are evaluated jointly. In the case of 2D trajectories, the energy score is to be applied to two spatially coupled time series. Therefore, the time series is multivariate in the temporal and spatial sense. So, the energy score has to evaluate all the temporal and spatial variables jointly. As far as we know, no other work has discussed such an application of energy score. To this end, we introduce three variations of Energy Score for different use cases of evaluation in the context of trajectory prediction, which also applies to broader use cases in multivariate time series evaluation. The energy score is applied to a probabilistic trajectory prediction with a predictive distribution $\boldsymbol{F_X}$ and assessed against a ground truth observation $\boldsymbol{y}$ as in Eq. 6. The overall performance is then the average performance of all the agents.

$$
\begin{aligned}
ES(\boldsymbol{F_X}, \boldsymbol{y}) =& \overbrace{\mathbb{E}\Big(\|\boldsymbol{X} - \boldsymbol{y}\|_p^\beta\Big)}^{ED} - \frac{1}{2}\overbrace{\mathbb{E}\Big(\|\boldsymbol{X} - \tilde{\boldsymbol{X}}\|_p^\beta\Big)}^{EI} && (6) \\
\overline{ES}(\boldsymbol{F_{X_i}}, \boldsymbol{y_i}) =& \mathbb{E}_i ES(\boldsymbol{F_{X_i}}, \boldsymbol{y_i}) && (7)
\end{aligned}
$$

The energy score operates on K uniform samples (trajectories) sampled from the predictive joint distribution, and the expectations are calculated with respect to the K samples. Since $\boldsymbol{X} \in \mathbb{R}^{K \times T_{pred} \times S}$ and $\boldsymbol{y} \in \mathbb{R}^{1 \times T_{pred} \times S}$ there are many ways that distance between them can be calculated. $\|.\|_p$ is a Euclidean norm, and for a rotation-invariant distance calculation, the Lp norm with $p = T_{pred} \times S$ should be used. For strictly proper evaluation, $\beta \in (0, p)$ and with a default value being $\beta = 1$. Small values of β should be used for heavy tail data to ensure corresponding moments exist. Refer to [17] for further information. We provide three different estimations of the energy score from Eq. 6. For clarity we provide the estimations for each term separately. The first term is similar to the displacement error that measures the distance of the prediction with the observation (ED). The second term is an intra-distance (EI) which can be thought of as a form of entropy. EI will be zero when the predictions are all the same or for a deterministic prediction.

$$
X_k^{t,s} = \begin{pmatrix} x_k^{T_{obs}1} & \dots & x_k^{T_{obs}S} \\ \vdots & \ddots & \vdots \\ x_k^{T1} & \dots & x_k^{TS} \end{pmatrix} \qquad y^{t,s} = \begin{pmatrix} y^{T_{obs}1} & \dots & y^{T_{obs}S} \\ \vdots & \ddots & \vdots \\ y^{T1} & \dots & y^{TS} \end{pmatrix}
$$

Entry-wise variation The distance between the matrices will be an entry-wise matrix norm. With $p = 2$ it would be equal to Frobenius distance.

$$ED(\boldsymbol{X}, \boldsymbol{y}) = \frac{1}{K} \sum_{k=1}^{K} \Big(\sum_{t=T_{obs}}^{T} \sum_{s=1}^{S} |X_k^{t,s} - y^{t,s}|^p \Big)^{\beta/p} \tag{8}$$

$$EI(\boldsymbol{X}) = \frac{1}{K^2} \sum_{k=1}^{K} \sum_{l=1}^{K} \Big(\sum_{t=T_{obs}}^{T} \sum_{s=1}^{S} |x_k^{t,s} - \tilde{x}_l^{t,s}|^p \Big)^{\beta/p} \tag{9}$$

Column-wise variation In this case, the distance is marginalized over the spatial dimension, and the emphasis is on the temporal sequence by finding the distance between temporal sequences. This variation ought to be more sensitive to temporal discrepancies in $\boldsymbol{X}$.

$$ED(\boldsymbol{X}, \boldsymbol{y}) = \frac{1}{K} \sum_{k=1}^{K} \Bigg(\frac{1}{S} \sum_{s=1}^{S} \Big(\sum_{t=T_{obs}}^{T} |x_k^{t,s} - y^{t,s}|^p \Big)^{\beta/p} \Bigg) \tag{10}$$

$$EI(\boldsymbol{X}) = \frac{1}{K^2} \sum_{k=1}^{K} \sum_{l=1}^{K} \Bigg(\frac{1}{S} \sum_{s=1}^{S} \Big(\sum_{t=T_{obs}}^{T} |x_k^{t,s} - \tilde{x}_l^{t,s}|^p \Big)^{\beta/p} \Bigg) \tag{11}$$

Row-wise variation Similar to the column-wise but it marginalizes over the temporal dimension and hence only reports on the spatial dependency. It does not capture cross-interaction between each spatial sequence. For a deterministic forecast, EI is zero, which makes this estimation identical to the ADE.

$$ED(\boldsymbol{X}, \boldsymbol{y}) = \frac{1}{K} \sum_{k=1}^{K} \Bigg(\frac{1}{T_{pred}} \sum_{t=T_{obs}}^{T} \Big(\sum_{s=1}^{S} |x_k^{t,s} - y^{t,s}|^p \Big)^{\beta/p} \Bigg) \tag{12}$$

$$EI(\boldsymbol{X}) = \frac{1}{K^2} \sum_{k=1}^{K} \sum_{l=1}^{K} \Bigg(\frac{1}{T_{pred}} \sum_{t=T_{obs}}^{T} \Big(\sum_{s=1}^{D} |x_k^{t,s} - \tilde{x}_l^{t,s}|^p \Big)^{\beta/p} \Bigg) \tag{13}$$

4 Experiments and results

We conduct a series of experiments with a synthetic setup demonstrating Energy Score's ability to evaluate trajectory predictions and compare them to the lower-bound metrics. We consider the following variations of the Energy Score we introduced earlier:

- Energy Score (**ES**): entry-wise variation
- Energy Score Temporal (**EST**): column-wise variation
- Energy Score Spatial (**ESS**): row-wise variation
- Final Energy Score (**FES**): ES applied only to the final step of the trajectory.

In all of the variations, we set the parameters to be $p = 2$ and $\beta = 1$. For the lower bound metrics, we consider the commonly used variations of top1 and top10% of FDE and ADE. We also include ADE and FDE that is calculated over all K of the trajectories. We do not consider CRPS because it would be equivalent to the average of energy scores obtained by applying energy scores to each temporal and spatial variable independently, meaning all the information concerning the spatiotemporal aspects is excluded. In such case, when energy score is applied to each temporal and spatial variable independently, there is only one way to estimate Eq. 6.

4.1 Synthetic data

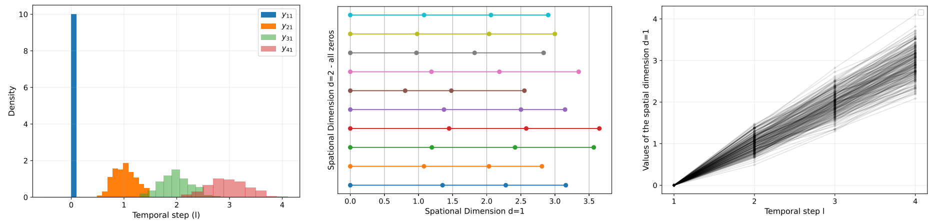
We define a synthetic trajectory-generating process as a bi-variate normal distribution representing 2D spatial data ($S = 2$), coupled temporally through an autoregressive process with coefficients $c^{(t)}$ that controls autocorrelation between each timestep, $a^{(t)}$ that controls deviation from the mean and $b^{(t)}$ that controls deviation from the covariance of each step.

$$\begin{aligned}\boldsymbol{y}^{(t)} =& c^{(t)} \cdot \boldsymbol{y}^{(t-1)} + \mathcal{N}(\boldsymbol{\mu}^{(t)} + a^{(t)}, \boldsymbol{\Sigma}^{(t)} + b^{(t)}), \quad y^0 = (0,0)\\ \boldsymbol{\mu}^{(t)} =& (\mu^1, \mu^2) \quad \boldsymbol{\Sigma}^{(t)} = \begin{pmatrix}\sigma^{11} & \sigma^{12}\\ \sigma^{21} & \sigma^{22}\end{pmatrix}\end{aligned}$$

For our experiments, we generate trajectories with length $T_{pred} = 4$ and consider the second spatial dimensional to be all zeros, and so effectively, the generating process is one-dimensional and can be summarized as follows:

$$\boldsymbol{y}^{(t)} = c^{(t)} \cdot \boldsymbol{y}^{(t-1)} + \mathcal{N}(\mu^{(t)} + a^{(t)}, (\sigma^{(t)} + b^{(t)})^2), \quad t \in \{1,2,3\}, \quad y^0 = 0$$

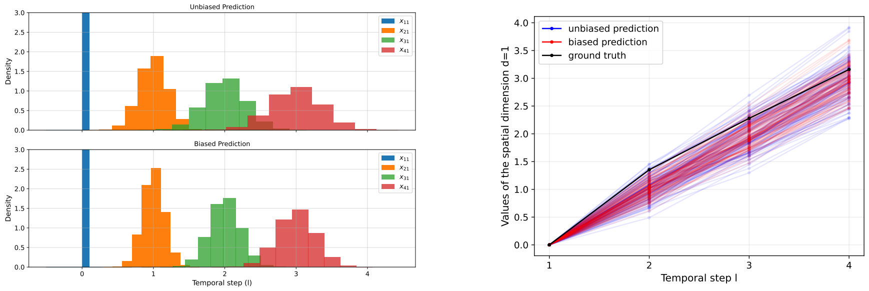
$\boldsymbol{a} = (a^{(1)}, a^{(2)}, a^{(3)})$, $\boldsymbol{b} = (b^{(1)}, b^{(2)}, b^{(3)})$ and $\boldsymbol{c} = (c^{(1)}, c^{(2)}, c^{(3)})$ allow creating different deviations for each step of the generating process from its main parameter values and hence allows to create so-called synthetic predictions with different deviations to study robustness and sensitivity of different metrics with respect to such deviations. We create N different observations $\{\boldsymbol{y_i}\}_{i=1}^N$ by simulating the above process, where $\boldsymbol{y}_i = \{y^{(0)}, \dots, y^{(3)}\}$. For each i^{th} observation, we also create K samples of the same process for the trajectory prediction $\{X_i\}_{i=1}^N$, where $X_i = \{X_{i,k}^{(0)}, \dots, X_{i,k}^{(3)}\}_{k=1}^K$, but with different parameters $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ to emulate the discrepancies in the prediction model. An unbiased prediction has the same parameter set as the $\boldsymbol{y_i}$. So observations is a $N \times T_{pred} \times S$ dimensional vector, and predictions is a $N \times K \times T_{pred} \times S$ dimensional vector where for all of the experiments $T_{pred} = 4$ and $S = 2$. As can be seen from the definition of our synthetic setup, the second spatial dimension $s = 2$ is all zeros, and effectively, the trajectory is one-dimensional, but the evaluations and calculations still operate in a spatial dimensional of two. This was intentional to make our synthetic setup simpler and the interpretation of the results easier a rationale that if a metric does not exhibit desired behavior on lower spatial dimensionality, it is certain that in higher dimensionality it would not be any more desirable.



(a) (b) (c)

Fig. 1: Example of a synthetically generated observation (ground truth). (a) the distribution per each time step for the first spatial dimension, (b) 10 trajectories in 2D where the second spatial dimension of all trajectories are zero and hence they are plotted with a y-offset for visualization purpose, (c) value of trajectories along the first spatial dimension.

Figure 1 shows N trajectories that constitute the observation samples $\boldsymbol{y}$. As shown in Fig. 1b these trajectories are one-dimensional and they can be deemed as trajectories that vary in their velocity. In Fig. 2 one observation is shown against K trajectories that constitute the biased and unbiased predictions. The biased prediction corresponds to *case 4* (small variance) of our sensitivity study Appendix A plotted together with case 1 (unbiased prediction).



(a) (b)

Fig. 2: Example of a synthetically generated prediction. (a) the distribution of an unbiased vs. biased prediction per each time step for the first spatial dimension, (b) One out of N observations depicted against K predicted trajectories.

4.2 Robustness to the size of trajectories

Similar to the experiment in Appendix F of [2], we demonstrate how each metric is sensitive to the number of trajectories as shown in Table 1. Assuming that each trajectory is a uniform sample of the predictive distribution, the more we sample, the more it occupies the space; hence, a metric that can converge to a stable score with a minimum number of trajectories is more desired. Authors in [2] previously showed that topK% scores are more desirable over topK types since they are more consistent. We show this in our experiment as well and also show that Energy Score is consistent across different trajectory sizes. The bias prediction in Table 1 is the prediction with a small variance as in our sensitivity study in Appendix A. In the next section, we also show why topK% metrics do not behave ideally, and they still suffer from similar issues as top1.

		$t =$	0	1	2	3
metric	pred	K				
FDE top1	unb	50	0.0	1.1	1.7	2.0
		100	0.0	0.6	1.0	1.1
		300	0.0	0.2	0.4	0.4
	b	50	0.0	1.5	2.4	2.9
		100	0.0	1.0	1.5	1.8
		300	0.0	0.5	0.6	0.8
FDE top10%	unb	50	0.0	23.1	32.6	40.6
		100	0.0	23.1	32.6	40.7
		300	0.0	23.0	32.3	40.4
	b	50	0.0	20.6	28.9	36.3
		100	0.0	20.6	28.9	36.3
		300	0.0	20.5	28.7	36.1
FES	unb	50	0.0	6.0	8.4	10.7
		100	0.0	6.0	8.4	10.7
		300	0.0	5.9	8.2	10.5
	b	50	0.0	6.1	8.5	11.0
		100	0.0	6.1	8.6	10.9
		300	0.0	6.0	8.4	10.7

		$t =$	0	1	2	3
metric	pred	K				
ADE top1	unb	50	0.0	0.5	2.6	4.9
		100	0.0	0.3	2.0	3.9
		300	0.0	0.1	1.1	2.7
	b	50	0.0	0.8	3.0	5.3
		100	0.0	0.5	2.1	4.2
		300	0.0	0.3	1.4	3.0
ADE top10%	unb	50	0.0	1.4	4.7	7.3
		100	0.0	1.4	4.5	7.2
		300	0.0	1.3	4.4	7.1
	b	50	0.0	1.8	4.9	7.6
		100	0.0	1.8	4.8	7.4
		300	0.0	1.7	4.7	7.3
ES	unb	50	0.0	6.0	10.9	16.1
		100	0.0	6.0	10.9	16.0
		300	0.0	5.9	10.7	15.7
	b	50	0.0	6.1	11.1	16.4
		100	0.0	6.1	11.1	16.4
		300	0.0	6.0	11.0	16.1

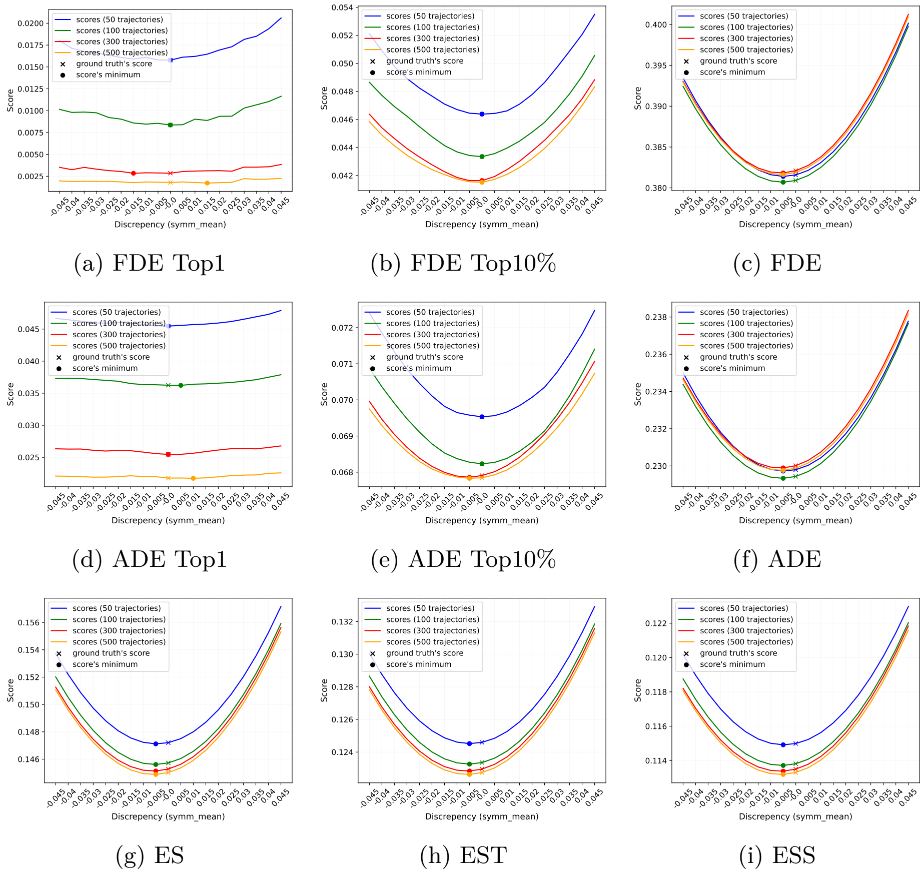
Table 1: The error for each step is calculated based on a moving window over the trajectories horizon (temporal step). topK metrics are prone to the size of trajectories (their error decreases with the number of trajectories), while topK% and Energy Score variations are stable across the size of trajectories. All metrics report a larger error on the bias prediction b vs. unbiased prediction *unb*. Also, they report larger errors on later steps as expected due to the cumulative nature of the error into the farther future. All the scores obtained from the metrics in the table were multiplied by 100.

4.3 Propriety Study

In this study, we empirically aim to demonstrate and examine the properness (propriety) of the variations of Energy Scores we introduced against different variations of ADE/FDE. We use the parameters $\mu^{(t)} = 0$ and $\sigma^{(t)} = 0.2$ for $t = 1, 2, 3$ with $N = 1000$ and $K = 500$ to generate ground truth observations and predictions. We consider two categories of predictions, namely the mean deviated and variance deviated predictions, by choosing 19 equidistant values in the range $[0.045, 0.045]$ for $\boldsymbol{a}$ and $\boldsymbol{b}$ respectively for each of the categories. Errors of the mean deviated and variance deviated predictions could be seen in respectively in Fig. 3 and 4. We expect a metric to assign its lowest score or error to the unbiased prediction which is closest to the ground truth. On the other hand, as the deviations in the parameters of the generating process that generated the ground truth become larger and hence farther from the truth, we expect the score to reflect that. In these figures, we depict these two important anchor points: the score of the unbiased prediction and the lowest score assigned by the metric via a cross and a dot, respectively. What we observe is all metrics except for the top1 types seem to be able to distinguish the mean deviated predictions from the truth and uniquely report the optimal prediction, as we can see that the dotted and crossed points are nearly identical or very close in Fig. 3. The reason for them not quite matching all the time could be attributed to randomness and approximation errors in the metrics. On the other hand, for the deviations in the variance, it is clear ADE/FDE, and their lower-bound variations fail altogether, while Energy Score variations still uniquely report the optimal. This empirical evidence suggests that ADE/FDE and its variations are not proper, and their usage is limited to certain situations, e.g., detecting mean bias, while Energy Score, as a strictly proper scoring rule, provides a more robust evaluation.

As we can see in Fig. (4c, 4f), FDE and ADE are not proper as they assign a lower error to a prediction that has a lower variance (dots) compared to the unbiased prediction (crosses). Also, the difference between trajectories with different sizes is not visible on the plots as the displacement errors are averaged over the size of the trajectories, and the average displacement error (y-axis) has a hugely different scale than the discrepancy (x-axis). FDE Top1 exhibits a similar behavior as depicted in Fig. 4a while ADE Top1 may not seem to exhibit this behavior or be less severe as depicted in Fig. 4b, but since it only considers one trajectory it could be merely due to chance and that more clear as more trajectories are considered in their Top 10% counterparts as depicted in Fig. 4b, 4e, respectively. In the meantime, Energy Score variations are consistent and robustly identify the optimal. Another important observation is that as the number of trajectories considered for calculation of the scores, Energy Scores are able to identify the optimal better, while that is not the case for the ADE and FDE metrics, and actually, quite the opposite happens.

A final observation from our propriety study is that perhaps there is an optimal value to be chosen for parameter N% that for Top N% metrics minimizes their undesired non-propriety behavior; however, finding this optimal value does



(a) FDE Top1

(b) FDE Top10%

(c) FDE

(d) ADE Top1

(e) ADE Top10%

(f) ADE

(g) ES

(h) EST

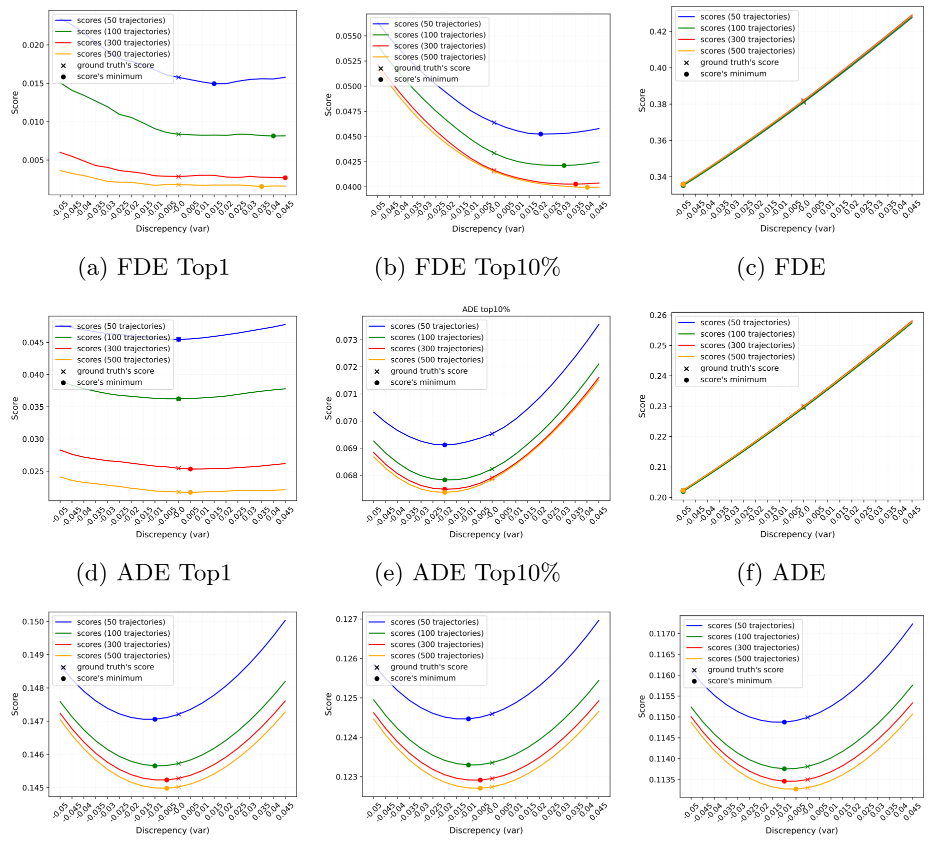
(i) ESS

Fig. 3: Plot of different metrics across different predictions that deviate on the mean parameter. The lowest reported score versus the unbiased prediction is depicted in the circle and cross, respectively.

not guarantee desired behavior on other possible discrepancies and should it be considered to be found as a hyperparameter for the evaluation to be more robust, would be deemed more like a hack than a remedy to resolve this problem and the conclusion is to abandon the usage of these scores as they provide very limited view for evaluation the least if not misleading.

5 Discussion

Our empirical studies demonstrate desired behavior of the Energy Scores we proposed for the evaluation of the multimodal trajectory predictions. Our studies are limited to a simple case of trajectory prediction derived from simple discrepancies that can materials in one dimension with the rationale that if an



(a) FDE Top1 (b) FDE Top10% (c) FDE

(d) ADE Top1 (e) ADE Top10% (f) ADE

(g) ES (h) EST (i) ESS

Fig. 4: Plot of different metrics across different predictions that deviate on the variance parameter. The lowest reported score versus the unbiased prediction is depicted in the circle and cross, respectively.

evaluation measure is not exhibiting desired behaviors, i.e., uniquely identifying the optimal prediction in the simple cases, then it will still be undesirable in a higher dimension where there are more plausible discrepancies. One of the consequences of our study is that it shows that conventional ADE and FDE metrics, together with their lower-bound variations, do not faithfully report the quality of trajectory predictions as promised.

A premise of lower-bound metrics is that they offer a better evaluation of the multimodal trajectories. While that is true when compared to Top1 and full ADE/FDE calculation, objectively, that is not true, and they could be misleading as, by definition, they exclude a good chunk of the trajectories that could be used for evaluation and effectively missing information on certain parts of the predictive distribution.

One of the limitations of Energy Score is the execution time $\mathcal{O}(n^2)$, which can be reduced to $\mathcal{O}(2kn)$ with k-band variation at the cost of approximation accuracy introduced in [19]. An implication of employing a distribution-aware measure is that it uniquely reports on a predictive distribution that information theoretically is optimal. That makes any downstream task that uses a trajectory prediction better positioned in decisions related to choosing a trajectory prediction model over the other. Moreover, the energy score can be used as an objective function during training as long as the function generating the K samples[3] is differentiable. We are aware of two works in the time series forecasting applications that employ energy score as a loss [3, 9], respectively, for multivariate probabilistic forecasting of electricity price scenarios and multivariate ensemble post-processing. Their work could be an inspiration for developing methods for spatiotemporal modeling, such as in the case of multimodal trajectory prediction. Nonetheless, as noted by these works, the multivariate evaluation and employing a differentiable generator function are present challenges subject to future research.

6 Conclusion and Future work

We proposed three different variations of the Energy Score for distribution-aware evaluation of multimodal trajectory predictions. This addresses the gap in the literature for a thorough, robust, and reliable assessment of multimodal trajectories. Even though our experiments are not exhaustive, they showcase simple cases under which lower-bound metrics such as minADE and minFDE fail, while Energy Score offers a better alternative. At the same time, the variations of the energy score we presented may not be ideal. Still, our preliminary results suggest it is a better alternative to existing metrics commonly employed in the trajectory prediction literature. A main advantage of ES is that it has a capacity for different formulations to adapt to the requirements of the task and address aspects that could only be seen objectively in the view of a downstream task. In our work, we rely on the forecasting literature to argue and propose the usage of Energy Score for trajectory prediction evaluation as its properties, such as propriety, have been studied and proved to provide superiority over others. To this end, we list the following potential future directions to adopt ES for the evaluation of trajectory predictions:

1. Evaluation with real data and a more realistic synthetic setup to better understand the difference between each variation of the Energy Score.
2. Evaluating state-of-the-art (SOTA) methods, especially those focusing on uncertainty estimation using Energy Score. It is likely to obtain different rankings among SOTA, which would underline the importance of energy score and distribution-aware evaluation.
3. How energy score compares with other metrics such as likelihood-based and distribution-aware evaluation of trajectory predictions.

[3] Also referred to as scenarios in time series literature.

4. Incorporate other forms of distance (kernel design for energy score) that would better capture cross-interactions in spatiotemporal dimensions.
5. How Energy Score can facilitate usage and evaluation of trajectory predictions in the downstream tasks.
6. Employing Energy Score as a loss function for training models for multimodal trajectory prediction.

Acknowledgements This work was supported by the Estonian Research Council grant PRG1604 and the European Social Fund via IT Academy program. We would also like to express our sincere gratitude to Karl Riis for his invaluable discussions.

References

1. Bae, I., Park, J.H., Jeon, H.G.: Non-Probability Sampling Network for Stochastic Human Trajectory Prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6477–6487 (2022), https://openaccess.thecvf.com/content/CVPR2022/html/Bae_Non-Probability_Sampling_Network_for_Stochastic_Human_Trajectory_Prediction_CVPR_2022_paper.html
2. Bhattacharyya, A., Hanselmann, M., Fritz, M., Schiele, B., Straehle, C.N.: Conditional Flow Variational Autoencoders for Structured Sequence Prediction (Aug 2020), http://arxiv.org/abs/1908.09008, arXiv:1908.09008 [cs, stat]
3. Chen, J., Janke, T., Steinke, F., Lerch, S.: Generative machine learning methods for multivariate ensemble post-processing (Sep 2022). https://doi.org/10.48550/arXiv.2211.01345, http://arxiv.org/abs/2211.01345, arXiv:2211.01345 [physics, stat]
4. Ess, A., Leibe, B., Van Gool, L.: Depth and Appearance for Mobile Scene Analysis. In: 2007 IEEE 11th International Conference on Computer Vision. pp. 1–8 (Oct 2007). https://doi.org/10.1109/ICCV.2007.4409092, iSSN: 2380-7504
5. Gneiting, T., Raftery, A.E.: Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association **102**(477), 359–378 (Mar 2007). https://doi.org/10.1198/016214506000001437, https://doi.org/10.1198/016214506000001437, publisher: Taylor & Francis _eprint: https://doi.org/10.1198/016214506000001437
6. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social GAN: Socially Acceptable Trajectories With Generative Adversarial Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2255–2264 (2018), https://openaccess.thecvf.com/content_cvpr_2018/html/Gupta_Social_GAN_Socially_CVPR_2018_paper.html
7. Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. Weather and Forecasting **15**(5), 559–570 (Oct 2000). https://doi.org/10.1175/1520-0434(2000)015¡0559:DOTCRP¿2.0.CO;2, https://journals.ametsoc.org/view/journals/wefo/15/5/1520-0434_2000_015_0559_dotcrp_2_0_co_2.xml, publisher: American Meteorological Society Section: Weather and Forecasting
8. Huang, R., Xue, H., Pagnucco, M., Salim, F., Song, Y.: Multimodal Trajectory Prediction: A Survey (Feb 2023), http://arxiv.org/abs/2302.10463, arXiv:2302.10463 [cs]

9. Janke, T., Steinke, F.: Probabilistic multivariate electricity price forecasting using implicit generative ensemble post-processing. In: 2020 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS). pp. 1–6 (Aug 2020). https://doi.org/10.1109/PMAPS47429.2020.9183687, iSSN: 2642-6757

10. Ma, Y.J., Inala, J.P., Jayaraman, D., Bastani, O.: Likelihood-Based Diverse Sampling for Trajectory Forecasting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13279–13288 (2021), https://openaccess.thecvf.com/content/ICCV2021/html/Jason_Likelihood-Based_Diverse_Sampling_for_Trajectory_Forecasting_ICCV_2021_paper.html

11. Mangalam, K., An, Y., Girase, H., Malik, J.: From Goals, Waypoints & Paths To Long Term Human Trajectory Forecasting. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15213–15222. IEEE, Montreal, QC, Canada (Oct 2021). https://doi.org/10.1109/ICCV48922.2021.01495, https://ieeexplore.ieee.org/document/9709992/

12. Mangalam, K., Girase, H., Agarwal, S., Lee, K.H., Adeli, E., Malik, J., Gaidon, A.: It Is Not the Journey But the Destination: Endpoint Conditioned Trajectory Prediction. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 759–776. Lecture Notes in Computer Science, Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-58536-5_45

13. Matheson, J.E., Winkler, R.L.: Scoring Rules for Continuous Probability Distributions. Management Science **22**(10), 1087–1096 (1976), https://www.jstor.org/stable/2629907, publisher: INFORMS

14. Mohamed, A., Qian, K., Elhoseiny, M., Claudel, C.: Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14412–14420. IEEE, Seattle, WA, USA (Jun 2020). https://doi.org/10.1109/CVPR42600.2020.01443, https://ieeexplore.ieee.org/document/9156583/

15. Rudenko, A., Palmieri, L., Herman, M., Kitani, K.M., Gavrila, D.M., Arras, K.O.: Human motion trajectory prediction: a survey. The International Journal of Robotics Research **39**(8), 895–935 (Jul 2020). https://doi.org/10.1177/0278364920917446, https://doi.org/10.1177/0278364920917446, publisher: SAGE Publications Ltd STM

16. Salzmann, T., Ivanovic, B., Chakravarty, P., Pavone, M.: Trajectron++: Dynamically-Feasible Trajectory Forecasting with Heterogeneous Data. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 683–700. Lecture Notes in Computer Science, Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-58523-5_40

17. Székely, G.J., Rizzo, M.L.: Energy statistics: A class of statistics based on distances. Journal of Statistical Planning and Inference **143**(8), 1249–1272 (Aug 2013). https://doi.org/10.1016/j.jspi.2013.03.018, https://www.sciencedirect.com/science/article/pii/S0378375813000633

18. Yue, J., Manocha, D., Wang, H.: Human Trajectory Prediction via Neural Social Physics. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 376–394. Lecture Notes in Computer Science, Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-19830-4_22

19. Ziel, F., Berk, K.: Multivariate Forecasting Evaluation: On Sensitive and Strictly Proper Scoring Rules. Papers (Oct 2019), https://ideas.repec.org//p/arx/papers/1910.07325.html, number: 1910.07325 Publisher: arXiv.org

Supplementary Material

Appendix A Sensitivity Study

In this study, we will conduct a similar study as the last but with more prediction discrepancies and show the Energy Score's power. The data-generating process that generates observations (ground truths) is the same with $\mu^{(t)} = 0$ and $\sigma^{(t)} = 0.2$ for $t = 1, 2, 3$. To assess the sensitivity of different metrics, we consider the following prediction models:

1. Unbiased: $\boldsymbol{a} = (0, 0, 0), \boldsymbol{b} = (0, 0, 0), \boldsymbol{c} = (1, 1, 1)$
2. Symmetric Mean: $\boldsymbol{a} = (0.025, 0.025, 0.025), \boldsymbol{b} = (0, 0, 0), \boldsymbol{c} = (1, 1, 1)$
3. Asymmetric Mean: $\boldsymbol{a} = (0.025, -0.05, 0.05), \boldsymbol{b} = (0, 0, 0), \boldsymbol{c} = (1, 1, 1)$
4. Large Variance: $\boldsymbol{a} = (0, 0, 0), \boldsymbol{b} = (0.05, 0.05, 0.05), \boldsymbol{c} = (1, 1, 1)$
5. Small Variance: $\boldsymbol{a} = (0, 0, 0), \boldsymbol{b} = (-0.05, -0.05, -0.05), \boldsymbol{c} = (1, 1, 1)$
6. Symmetric Mean bias and Large Variance $a = (0.025, 0.025, 0.025), b = (0.05, 0.05, 0.05), \boldsymbol{c} = (1, 1, 1)$

Instead of looking at the scores produced by each of the metrics, we look at the p-values obtained from a Diebold-Mariano test (DM-test) which reports whether the scores of a biased prediction are statistically different from the unbiased prediction or not. To obtain the p-value following steps are taken: 1) obtain N scores corresponding to unbiased and biased predictions respectively, 2) calculate the difference between the two and obtain N differences, 3) normalize the differences according to the DM-test, 4) calculate the two-tailed p-value.

$$\begin{aligned}
\mathit{\Delta}_i =& SC_i^{unbiased} - SC_i^{biased} \\
\bar{\mathit{\Delta}} =& \frac{1}{N}\sum_{i=1}^{N} \mathit{\Delta}_i \\
z =& \frac{\mathit{\Delta}}{\sqrt{\frac{\sigma(\bar{\mathit{\Delta}})}{N}}} \rightarrow \mathcal{N}(0, 1) \\
\text{p-value} =& 2(1 - \mathit{\Phi}(z)) * 100
\end{aligned}$$

$\mathit{\Phi}$ is the cumulative distribution function of a Normal distribution and the p-value ranges from 0% to 100% with lower values indicating that a metric is better capable of distinguishing between an unbiased and a biased prediction. We consider p-values under 10% to be statistically significant (equivalent to 5% in a one-sided test). The parameters for the biased prediction model are manually found such that at least one of the metrics fall under the 10% threshold and the rest not to be too close to either of 0% or 100% bounds.

$t =$	1	2	3
metric			
FDE top1	57.91	94.96	75.07
FDE top10%	**5.20**	**9.38**	6.54
ADE top1	57.91	81.05	94.06
ADE top10%	**5.20**	62.24	**3.29**

$t =$	1	2	3
metric			
ES	94.07	14.50	2.10
FES	94.07	**7.80**	**0.75**
EST	94.07	11.99	1.47
ESS	94.07	20.57	3.41

Table 2: p-values of symmetric mean biased prediction (case 2).

In Table 2 we can see that Top10% metrics are more successful compared to Top1 metrics in distinguishing the biased prediction from the unbiased prediction for most of the time steps. However, in the ADE Top10%, we can observe that although the mean bias increases over time, the discrimination is not monotonic as one would expect, whereas, for all of the Energy Score variations, the discrimination is monotonic. All energy score variations consider the first step of the biased prediction to be nearly indifferent from the biased prediction but for later steps gradually they report more difference, and that is desired as the symmetric mean bias causes an accumulation of bias to the later steps by definition. Overall, since we are interested in a multivariate scoring that considers all trajectories and their full length when looking at $t = 3$ we can conclude that all energy score variations correctly show strong discrimination for the symmetric mean bias (case 2) and our proposed energy score variations discriminate better. It is also, noteworthy that after FES, the EST variation shows strong discrimination which is expected due to the fact that the FES only looks at the final step. Also, this shows that EST variation behaves the way we expect, which is to be sensitive to temporal deviations, and since the mean deviations gradually accumulate the EST also gradually responds. Note that FDE and ADE top1/top10% yield the same value for $t = 1$ because by definition they are the same for the first step. Also the same is true for all energy score variations for $t = 1$.

$t =$	1	2	3
metric			
FDE top1	57.91	28.53	33.13
FDE top10%	**5.20**	73.09	48.62
ADE top1	57.91	**4.86**	42.54
ADE top10%	5.20	67.97	**2.75**

$t =$	1	2	3
metric			
ES	94.07	94.74	1.28
FES	94.07	92.25	8.14
EST	94.07	91.53	**0.47**
ESS	94.07	97.45	8.86

Table 3: p-values of asymmetric mean biased prediction (case 3).

Table 3 shows the results for the asymmetric case (case 3). Since the first step shares the same mean bias as in case 2 the scores for all metrics are the same. All metrics behave similarly for all time-steps as in the previous case except for both FDE top10% and ADE top1 on steps 3 and 4. For ADEs it could be partly because of the fact that the assymetric mean bias affects the temporal correlation between the steps and ADE is perhaps responding to that change but since it only considers 1 trajectory, it could also be a random effect. Since FDE top10% is not sensitive to temporal correlations, the possibles explanations coule be some effects of randomness or propagated effects from earlier steps. On the other hand all the energy score variations behave as expected and similarly as before. Interestingly the p-values for the second step are comparable with the first step and that could be explained by the fact that the assymmetric bias introduced in the first step is counterbalanced (by having a negative bias for the second step) and that is clearly reflected in the energy scores' behavior. The final step performances for energy scores are comparable and again as expected EST is the best since this task is essentially about the temporal discrimination.

$t =$	1	2	3
metric			
FDE top1	30.65	4.78	3.39
FDE top10%	**0.29**	**2.92**	3.09
ADE top1	30.65	8.35	86.80
ADE top10%	**0.29**	37.74	**0.00**

$t =$	1	2	3
metric			
ES	19.16	3.32	6.37
FES	19.16	7.28	30.05
EST	19.16	**2.61**	**5.40**
ESS	19.16	5.94	8.83

Table 4: p-values of large variance biased prediction (case 4).

Table 4 shows that topK and in particular topK% metrics are more sensitive on large variance (case 4) while energy score variations are less sensitive. At the same time, they show less consistency across time-steps compared to energy score variations, however, since energy score also exhibits some inconsistency across steps, it is more likely to be of a systematic nature such as the accumulative effect propagation from earlier steps to the later steps. Larger variance means that the samples from earlier steps could affect the later steps and visually that means that the later steps can have more overlaps with the ground truth hence increasing the chance of the metric to consider the prediction to be close to the ground truth. The fact that the FES on step 4 has high p-value is aligned with this line of reasoning. Again, among energy score variations EST performes the best.

$t =$	1	2	3
metric			
FDE top1	0.6	0.08	0.36
FDE top10%	**0.0**	**0.00**	**0.00**
ADE top1	0.6	14.96	2.49
ADE top10%	**0.0**	0.22	6.99

$t =$	1	2	3
metric			
ES	0.15	**0.08**	**0.01**
FES	0.15	1.08	0.10
EST	0.15	**0.08**	**0.01**
ESS	0.15	0.09	**0.01**

Table 5: p-values of small variance biased prediction (case 5).

As shown in Table 5 all metrics show comparable discrimination in the small variance case (case 5), as opposed to case 4. Both categories of lower-bound and energy score metrics show more sensitivity to lower variance. That is more accentuated for the energy score metrics. It is hard to explain this behavior however it seems to be aligned with the results from [19]. But more importantly, we can see that lower-bound scores tend to exhibit an overreaction to lower variance by having p-values that are zero. This can be explained by the fact that they only take a limited number of trajectories into account.

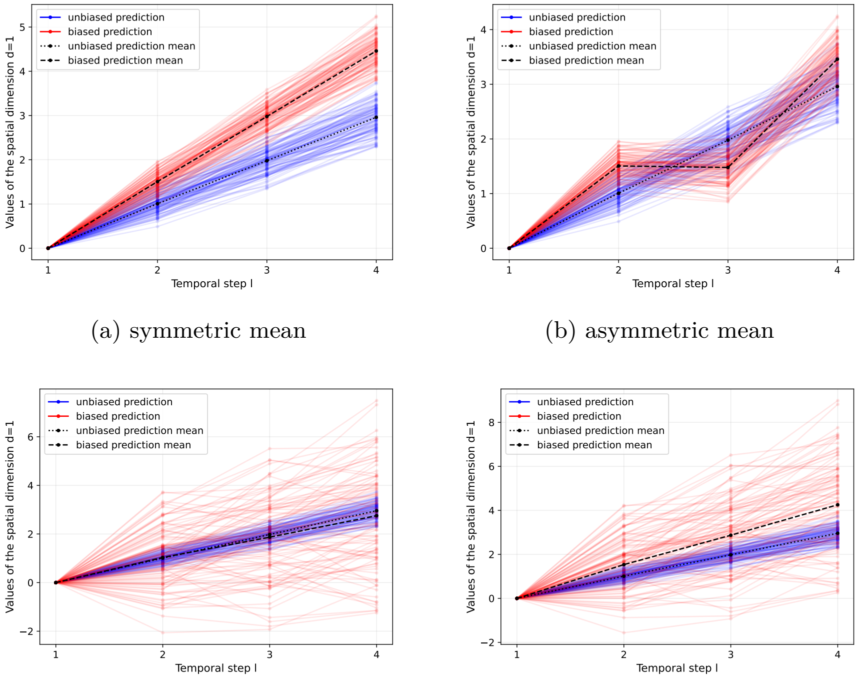
$t =$	1	2	3
metric			
FDE top1	73.42	**8.14**	41.69
FDE top10%	**9.66**	10.59	49.67
ADE top1	73.42	35.81	12.35
ADE top10%	**9.66**	19.24	**0.00**

$t =$	1	2	3
metric			
ES	37.41	2.79	0.37
FES	37.41	**1.87**	0.26
EST	37.41	2.02	**0.20**
ESS	37.41	5.11	0.79

Table 6: p-values of mean biased and large variance (case 6).

Finally, in Table 6 we see the results for the case where we have symmetric mean and large variance discrepency in the prediction and it summarizes the power of each metric. Energy score variations come out more successful with the results being coherent with the previous cases. And again, EST demonstrating a more powerful discrimination.

In summary, we can see that the energy scores are more consistent in their discrimination across different steps. For example, whenever the deviations are expected to accumulate over the time steps, their discrimination responds accordingly i.e. p-value decreases monotonically, as opposed to the topK% variations where the decrease is not monotonic which for the topK% metrics it could be attributed to the fact that at i different modes of the predictive distribution is evaluated which inevitably leads to such inconsistencies that ultimately causes difficulty in interpretation of the results especially if one is interested



(a) symmetric mean

(b) asymmetric mean

(c) large variance

(d) Symmetric mean and large variance

Fig. 5: Visualization of the biased predictions against the unbiased (case 1). Respectively the biased predictions depicted in red in (a), (b), (c), and (d) correspond to cases 2, 3, 4, and 6 with exaggerated parameters for visualization purposes. Visualization of Case 5 corresponding to small variance can be found in earlier figures.

in a temporal assessment. The discrimination power of topK% metrics could also be affected by the percentage parameter which is not obvious how should be tuned and probably is a function of the underlying complexity of the data and some characteristics of the predictive distribution which make them a less appealing and inconvenient choice for robust evaluation. Moreover, only taking into account a limited number of trajectories for evaluation can lead to under- or over-estimation of the true error.