EuroSun2026

TOWARDS FOUNDATION MODELS FOR SOLAR THERMAL SYSTEMS

Florian Ebmeier, Nicole Ludwig, Jannik Thuemmel, Georg Martius, Volker H. Franz

University of Tübingen

florian.ebmeier@uni-tuebingen.de

Summary

This work assesses the feasibility of state-of-the-art machine learning models for monitoring solar thermal systems (STS), with the goal of a single model that performs different tasks without task-specific training. Based on the publicly available PaSTS dataset, we employ a two-stage masked autoencoding framework: a convolutional tokenizer (a 1D-CNN-based VQVAE) compresses each sensor signal into discrete tokens, and a Vision Transformer (ViT) learns to predict tokens that were masked out. Any task that can be expressed as a masking pattern at inference time can then be addressed by the same model.

Two tasks are evaluated. (a) Estimation of missing sensor data reaches a mean absolute error of 8.67 K averaged over five temperature sensors, substantially outperforming linear regression (10.06 K) and a per-timestep mean profile (12.33 K). (b) Anomaly detection reaches an F1 score of 0.76 ± <0.01 and a system-wise F1 of 0.40 ± 0.01, matching all but the strongest of the dedicated baseline models. Since all of these baselines are single-purpose detectors, the claim here is generality rather than per-task accuracy. Central to our findings is that a better tokenizer does not translate into better downstream performance: a larger codebook clearly degrades anomaly detection while leaving sensor estimation unchanged, so that the smallest codebook acts as a regularizer. The masking pattern used during training mildly trades the two tasks against each other. Overall, the model generalizes across system realizations, handles missing data by construction, and is a first step toward foundation models for solar thermal and heating systems. We discuss how to close the gap toward specialized models and real-world applicability.

Keywords: Solar thermal systems, deep learning, foundation models, machine learning

Field of application for the findings: Sustainable Heating and Cooling Systems for Buildings

Addressed audience: Researchers in energy system modeling and machine learning practitioners in the renewable energy sector.

---

1. Motivation and Problem Definition

Deep learning is increasingly defined by foundation models (FMs) (Bommasani et al., 2021), large-scale architectures trained on diverse data that are applicable to numerous downstream tasks. These FMs have revolutionized language processing and vision, and the same recipe has since been carried into scientific domains such as climate, weather and earth system modeling (Bodnar et al., 2025; Brown et al., 2025; Price et al., 2025) and structural biology (Jumper et al., 2021). However, the trend of FMs has only begun to reach energy systems, and small-scale thermal systems in particular remain unaddressed.

A data-driven approach for small-scale thermal systems must generalize across system realizations, handle incomplete data, and be adaptable to diverse downstream tasks. Solar Thermal Systems (STS) present the typical challenges of small-scale energy systems. Real-world installations are diverse and rarely match standardized guidelines, frequently missing sensors, experiencing intermittent dropouts, and operating under unique boundary conditions. Thus, the classical approach of developing individual simulations or models for every small-scale system is economically unfeasible.

The masked autoencoding paradigm is a natural fit for these requirements. Missing sensors and intermittent dropouts are masked inputs by construction and are easily handled by a masked autoencoder. Additionally, any task that can be articulated as a specific mask at inference time can be attempted by the same model, without being specifically trained for it.

In this work, we address these challenges by applying a scalable masked autoencoding framework to the public PaSTS dataset (Ebmeier et al., 2024), following the evaluation protocol of Ebmeier et al. (2025). Our contributions are as follows. We show that a single model, trained once on nominal data, can serve both sensor estimation and anomaly detection zero-shot by expressing each task as an inference-time mask. We then quantify the remaining gap to task-specific models, focusing on generality rather than per-task accuracy. In doing so, we identify a decoupling between tokenizer quality and downstream performance that has to be resolved before the framework can be scaled.

2. Related Work

2.1 Machine Learning in Solar Thermal Systems

Recent work in machine learning for solar thermal systems and related small-scale energy systems has focused on specific applications like fault detection (Feierl et al., 2023; Ebmeier et al., 2025), short-term forecasting (Heidari and Khovalyg, 2020; Rana et al., 2022), as well as control and load management (Lissa et al., 2021). These approaches often rely on simulation data or individual case studies, struggling to generalize across diverse system realizations. To date, only Ebmeier et al. (2025) has utilized the multi-system PaSTS dataset (Ebmeier et al., 2024) to address generalization. However, most of these works employ standard convolutional neural network (CNN) and long short-term memory (LSTM) architectures rather than scalable foundation-style models, and are trained for one task at a time.

2.2 Foundation Models and Masked Autoencoding

Masked autoencoding partitions the input into a hidden and a visible part and trains a model to reconstruct the hidden part from the visible one (Devlin et al., 2019; He et al., 2022). While autoregressive models dominate language processing, it has emerged as a more recent alternative in foundation models (Chang et al., 2022; Shi et al., 2024; Google DeepMind, 2025). These models are particularly suited for engineering applications, as they inherently handle missing sensor data and variable input modalities by treating them as masked during inference. The state-of-the-art in generative modeling has shifted toward Transformer-based architectures, in particular the Vision Transformer (ViT) (Dosovitskiy et al., 2021), and discrete tokenization via vector quantization, introduced with the Vector Quantized Variational Autoencoder (VQVAE) (van den Oord et al., 2017) and extended in later work (Esser et al., 2021).

3. Methods

Following the framework of MaskGIT (Chang et al., 2022) and MD4 (masked discrete diffusion) (Shi et al., 2024), our model utilizes a two-stage approach trained exclusively on nominal data from the PaSTS dataset. The overall structure can be seen in Fig. 1. A day of measurements from one system is a matrix $x \in \mathbb{R}^{V \times T}$ with V sensors and T time steps.

3.1 Stage 1: Univariate Tokenization

A VQVAE with 1D-ConvNext blocks and strided convolutions tokenizes T time steps into L discrete tokens. An encoder E maps a univariate series to L latent vectors, each of which is replaced by its nearest entry in a learned codebook $C = \{e_1, \ldots, e_{N_c}\}$ with N_c entries, and a decoder D maps the quantized latents back to the time domain. To ensure the model learns local dynamics before cross-correlations, each of the V sensors is processed independently as a univariate series. Once trained via standard reconstruction and vector quantization loss, Stage 1 weights are frozen, and a day becomes a token grid of size $V \times L$.

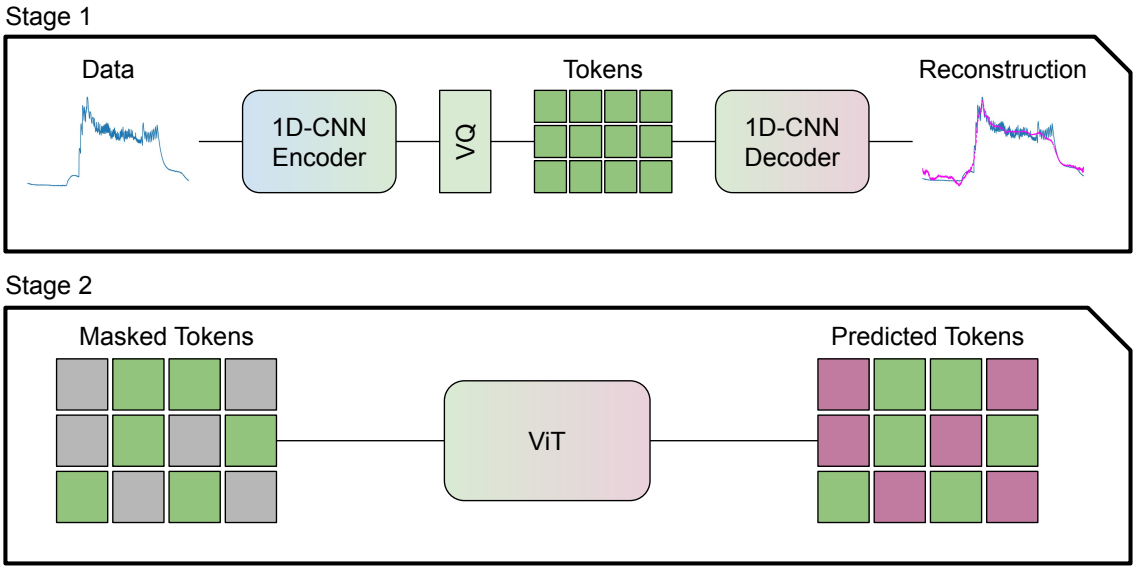

Fig. 1: Schematic of the two-stage model. In Stage 1 a standard CNN-based VQVAE is trained. Stage 2 is a ViT-based masked autoencoder.

3.2 Stage 2: Masked Latent Modeling

Masking. During training, a fraction γ_t of the $N = V \cdot L$ tokens is replaced by a mask token, with γ_t following a cosine masking schedule, $\gamma_t = 1 - \cos(\pi t/2)$ for a schedule time $t \sim \mathcal{U}(0, 1)$, so that γ_t increases from 0 to 1. Which tokens are masked is decided by sampling. Each token receives a random score

$$s_{v,l} = \varepsilon_{v,l} + \alpha\, \pi_{v,l}, \tag{1}$$

where $\varepsilon_{v,l}$ is per-token Gumbel noise, $\pi_{v,l}$ is a structural prior on the sampling distribution over the time l and sensor v index. The factor α weights the two contributions. In general, setting $\alpha = 0$ recovers the standard uniform masking without any prior. For this work, we assumed a block prior, where blocks of b consecutive tokens along the time axis of sensor v get assigned the same prior value, again sampled from per-block Gumbel noise. Here a block size of 1 also recovers uniform masking. The round$(\gamma_t N)$ highest-scoring tokens are then masked. A block size of $b = L$ spans a sensor's entire time axis, mirroring the sensor estimation task. The resulting masked token grid is modeled by a ViT architecture.

Training objective. For uniform masking ($\alpha = 0$ or $b = 1$), training can be motivated by the variational lower bound of Shi et al. (2024),

$$\mathcal{L}_{\text{vlb}}(x) = -\int_0^1 \frac{\dot{\gamma}_t}{\gamma_t}\, \mathbb{E}_{\mathcal{M}_t}\Big[\sum_{i \in \mathcal{M}_t} \log p_\theta\big(x^i \mid x_t\big)\Big]\, \mathrm{d}t, \tag{2}$$

where $\mathcal{M}_t$ is the set of masked positions at schedule time t, x_t the correspondingly masked grid, and $\dot{\gamma}_t$ the schedule derivative. The expectation is taken over the random masked set $\mathcal{M}_t$, i.e. over the forward masking process $x_t \sim q(x_t \mid x)$ of Shi et al. (2024). In practice the model is optimized via the cross-entropy (CE) loss, rescaled by the same factor,

$$\mathcal{L}_{\text{Stage 2}}(x) = \mathbb{E}_{t \sim \mathcal{U}(0,1),\, \mathcal{M}_t}\left[\frac{\dot{\gamma}_t}{\gamma_t} \sum_{i \in \mathcal{M}_t} \mathrm{CE}\big(x^i,\, p_\theta(\cdot \mid x_t)\big)\right]. \tag{3}$$

Structured priors. For uniform masking, Eq. (3) coincides with this bound, up to the discretization of the schedule by the fixed number of masked tokens. With the block prior ($\alpha > 0$ and $b > 1$), the masks are no longer sampled independently per token, and Eq. (3) becomes a heuristic reweighting of the cross-entropy that is not guaranteed to bound the likelihood. We nevertheless use it as the training objective and leave a principled treatment of structured masking priors to future work.

3.3 Downstream Tasks as Inference-Time Masks

Tasks are framed as inference-time masking problems. A masked autoencoder partitions the data into sets of hidden and visible tokens. Several downstream tasks follow directly from this partition. If we imagine a classical forecasting task, we can, at inference, treat the future as the hidden partition and the past as the visible partition. More generally, the flexibility of these general partitions lets us apply the model to any task that can be defined as such a partition or a composite of partitions. Here we focus on two different tasks: sensor estimation, as an example of a simple partition task, and anomaly detection as an example of a more complicated composite task.

- **Sensor Estimation:** Defined by applying a mask to the entirety of a single variable. The sensor is reconstructed purely from the remaining variables. The ViT predicts the missing tokens in a single forward pass.
- **Anomaly Detection:** Treated as an out-of-distribution task. We form a Monte Carlo estimate of the objective in Eq. (3) by sampling the cross-entropy across the entire masking schedule, using the same masking prior as during training. We draw one sample at each of 50 points along the schedule, covering masking rates from roughly 8% to 99%. For uniform masking this estimates $\mathcal{L}_{\text{vlb}}(x)$, a bound on the negative log-likelihood. With the block prior ($\alpha > 0$ and $b > 1$) it is a likelihood-inspired score. Days yielding high values relative to a threshold are considered out-of-distribution and thus classified as faulty.

4. Experimental Setup

Dataset. While STS sensor data is increasingly logged, public datasets remain scarce. We use the PaSTS dataset (Ebmeier et al., 2024), which contains 83 real-world domestic STS installations of varying configuration, age and system runtime, recorded at one-minute resolution for a total of 39 878 days. We adopt the reannotated fault labels, the dataset splits, as well as the evaluation procedure of Ebmeier et al. (2025), so that our anomaly detection results can be compared directly against the baselines reported there. The data is split along systems, such that each system is fully contained in exactly one set and evaluation is always performed on unseen systems. The training set comprises 30 systems with 13 229 days, the validation set 9 systems with 1 153 days, and the test set 44 systems with 25 496 days. Training and validation systems are mostly faultless, whereas the test systems contain both nominal and faulty behavior. Training uses only the nominal days of the training systems. Every data sample is a single day of a system with $V = 8$ sensors and $T = 1440$ one-minute time steps. There are five temperature sensors — the collector 'TSA1', the collector inlet 'TSE', the tank 'TW', the collector outlet 'TSV' and the ambient temperature 'TAM' — the volumetric flow sensor 'VF', the pump control signal 'pwm' and the frost-protection control signal 'ctr'. We additionally use the same preprocessing as defined in Ebmeier et al. (2025).

Metrics. Sensor estimation is evaluated on the nominal days of the test set. This test set carries a large number of unlabeled faults, which adds noise to the reported errors. We report the mean absolute error (MAE) over the masked sensor time-series in kelvin (K) for the five temperature sensors as well as averaged over all temperature sensors. For the anomaly detection we report the overall F1 score for an optimal global threshold, the system-wise F1 score as defined in Ebmeier et al. (2025) and the area under the precision-recall curve (AUC-PR). Unless stated otherwise, numbers are seed-means over ten seeds and all deviations are standard deviations across seeds.

Configurations. All configurations use a patch size of 8 and two stride-2 convolution stages, compressing 1440 time steps into $L = 45$ tokens per sensor. We observed that the codebook size N_c of the tokenizer affects the two downstream tasks differently, so we report $N_c = 128$, $N_c = 512$ and $N_c = 4096$, all with a masking block size of $b = 5$, a prior weight of $\alpha = 2$ and 10k Stage 2 steps. The ViT has a channel dimension of 256 with 6 layers unless stated otherwise. The featured configuration uses the smallest codebook of $N_c = 128$ entries. For the anomaly detection we additionally report a $N_c = 128$ variant trained with uniform masking ($b = 1$). Both stages are optimized with schedule-free AdamW (Defazio et al., 2024) at a learning rate of 10^{-4}, weight decay 10^{-5}, batch size 512 and 16-bit mixed precision, with 5k steps for Stage 1.

Baselines. We compare the sensor estimation against two simple baselines. The first is **linear regression**. For each temperature sensor a model predicts the temperature sensor from the other seven sensors at the same time step. The second is a **mean profile**, which returns the per-timestep training mean of each sensor. Both baselines are fitted on the nominal training set. Additionally, we report the **autoencoder floor**, which is the

reconstruction error obtained by encoding and decoding the ground-truth tokens without any masking. This represents the best result possible for the Stage 2 model under the tokenizer.

5. Results

5.1 Sensor Estimation

Tab. 1: Results of the sensor estimation task. Mean absolute error in K per temperature sensor (seed-mean ± standard deviation over ten seeds), obtained by masking one entire sensor and reconstructing it from the other seven. Errors are temperature differences and therefore given in K. All FM rows use a masking block size of 5, so they differ only in the codebook size N_c. The featured configuration is marked. The autoencoder floor is the reconstruction round-trip without masking and bounds the performance of the FM. The two baselines are independent of the tokenizer.

Method	TSA1	TSE	TW	TSV	TAM	Mean
Mean profile (per-timestep)	14.49	9.43	17.36	13.04	7.35	12.33
Linear regression (7→1)	11.59	7.67	17.51	8.27	5.28	10.06
FM (ours), $N_c = 128$ (featured)	7.06 ± 0.15	6.96 ± 0.25	17.77 ± 0.41	7.26 ± 0.10	4.31 ± 0.13	8.67 ± 0.11
Autoencoder floor, $N_c = 128$	3.51	1.22	1.60	2.86	0.98	2.04
FM (ours), $N_c = 512$	6.75 ± 0.12	6.72 ± 0.23	18.33 ± 0.74	7.44 ± 0.15	4.04 ± 0.08	8.66 ± 0.19
Autoencoder floor, $N_c = 512$	2.47	0.97	0.79	2.06	0.47	1.35
FM (ours), $N_c = 4096$	6.66 ± 0.15	6.81 ± 0.20	17.93 ± 0.73	7.50 ± 0.15	3.81 ± 0.14	8.54 ± 0.17
Autoencoder floor, $N_c = 4096$	1.66	0.79	0.70	1.62	0.33	1.02

Overall error. Table 1 reports the results for the sensor estimation task. The featured model reaches 8.67 K averaged over the five sensors, against 10.06 K for the per-timestep linear regression and 12.33 K for the mean profile. While the autoencoder floor improves monotonically with the codebook size (from 2.04 K to 1.02 K), the sensor estimation MAE varies by at most 0.13 K between the codebooks. This difference is comparable to the seed variability and an order of magnitude less than the change of the floor. We further investigate this in section 5.4.

Per-sensor breakdown. The per-sensor breakdown gives more insight into the sensor estimation. The tank temperature 'TW' is not meaningfully estimated by any of the models. We observe comparable errors between 17.36 K and 18.33 K. The tank temperature is driven by the auxiliary heater and by hot-water consumption, neither of which is observed by the seven remaining sensors, so no model can recover it from these inputs. Excluding TW, the featured model reaches 6.40 K against 8.20 K for linear regression and 11.08 K for the mean profile. We specifically discuss this in section 6. The difference between the baseline methods and the FM comes from the other sensors, with the sensor experiencing the most dynamics (TSA1) showing the largest improvements over the baselines.

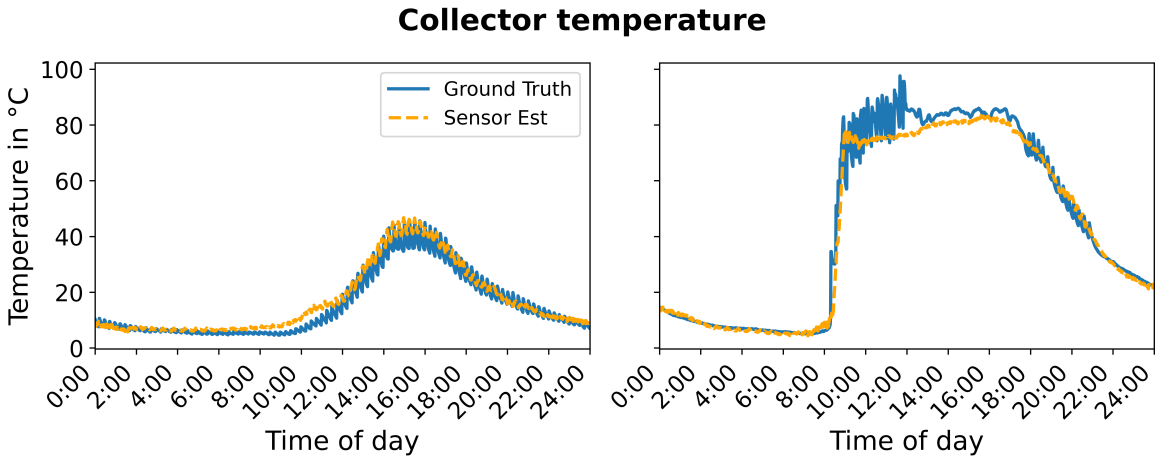

Fig. 2: Two examples of the sensor estimation task on the collector temperature ($N_c = 512$, $b = 45$ model). The mean behavior is captured well, but the reconstruction is visibly smoothed relative to the ground truth.

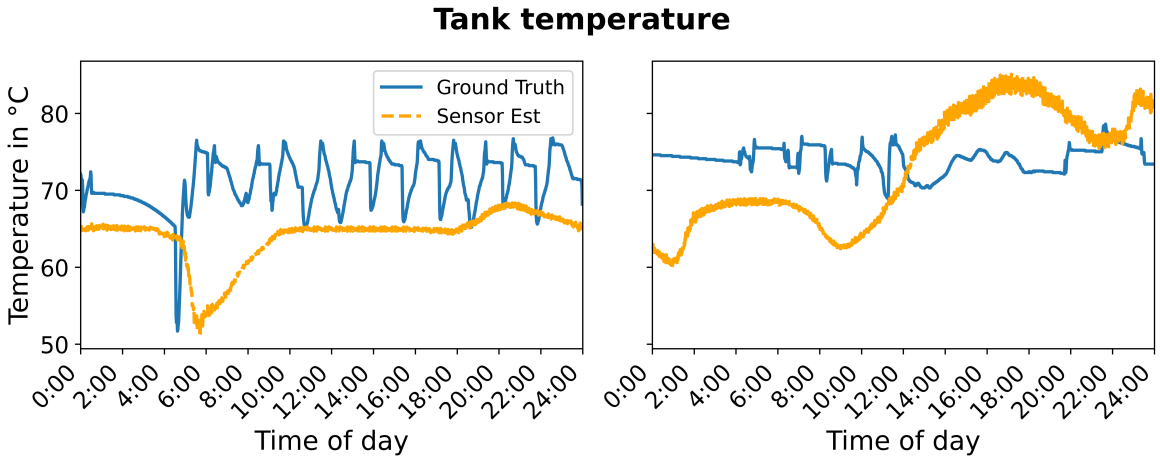

Fig. 3: Two examples of the sensor estimation task on the tank temperature for the same days and the same model as Fig. 2. The tank temperature follows a generic temperature curve, not matching the system's temperature. The periodic sawtooth pattern in the measured tank temperature (left) is consistent with the cycling of an auxiliary gas boiler.

Example days. Visual inspection (Fig. 2) of the collector temperature estimation for two different days shows good agreement in behavior both for a day with low dynamics (left) and high dynamics (right). There is a clear smoothing effect, which is expected and common in machine learning models. We also observe a mean bias, where the estimated temperature of the low dynamics day is slightly above the curve, and the estimated temperature of the high dynamics day is slightly below the curve. A different behavior is shown by the tank temperature of the same days (Fig. 3). Here the estimates are generic temperature curves, largely independent of the observed tank temperatures.

5.2 Anomaly Detection

Tab. 2: Zero-shot anomaly detection on the reannotated PaSTS labels. Baseline numbers are taken from Ebmeier et al. (2025). All reported deviations are standard deviations over ten seeds; deterministic methods carry no standard deviations. The FM rows vary the codebook size N_c at a fixed masking block size of $b = 5$; the last row uses uniform masking ($b = 1$). Bold marks the best value in each column (ties are all marked).

Method	System-wise F1	Overall F1	AUC-PR
Unscaled PCA-R	0.30	0.69	0.74
Rescaled PCA-R	0.40	0.76	**0.82**
MSE-βVAE	0.32 ± 0.02	0.72 ± 0.01	0.75 ± 0.01
MVE-βVAE	**0.46 ± 0.02**	**0.77 ± 0.01**	0.79 ± 0.01
FM (ours), $N_c = 128$, $b = 5$ (featured)	0.40 ± 0.01	0.76 ± <0.01	**0.82 ± <0.01**
FM (ours), $N_c = 512$, $b = 5$	0.37 ± 0.02	0.73 ± 0.01	0.81 ± <0.01
FM (ours), $N_c = 4096$, $b = 5$	0.34 ± 0.02	0.72 ± 0.01	0.79 ± 0.01
FM (ours), $N_c = 128$, $b = 1$	0.42 ± 0.02	0.76 ± <0.01	**0.82 ± <0.01**

Table 2 shows both the results of our anomaly detection and the models implemented by Ebmeier et al. (2025). The featured $N_c = 128$ model reaches an overall F1 of 0.76 ± <0.01 and a system-wise F1 of 0.40 ± 0.01, and the detection performance degrades monotonically with increasing codebook size. We note that even though we have a deterministic Stage 1 model, the Stage 2 model, trained with the MD4 rescaled cross-entropy loss (Shi et al., 2024), is an uncertainty-aware model. Our model clearly improves upon the baseline methods without uncertainty estimation (MSE-βVAE and Unscaled PCA-R) and is on par with Rescaled PCA-R, which the uniform-masking variant exceeds on the system-wise F1 (0.42 vs. 0.40). Only the MVE-βVAE remains clearly ahead, while our AUC-PR of 0.82 matches the best baseline value. Notably, the uniform-masking variant, whose anomaly score is a valid variational bound on the likelihood, is also the best detector.

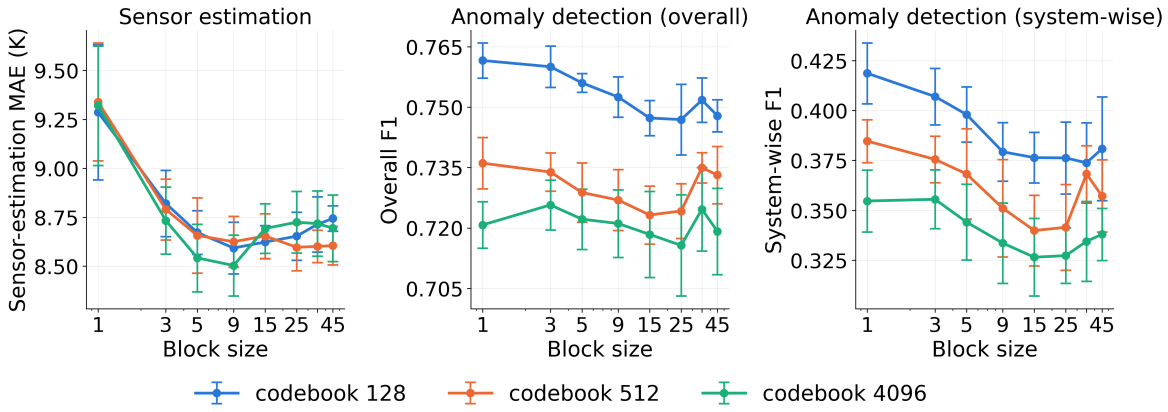

Fig. 4: Block size sweep for the three codebook sizes. Points are seed-means over ten seeds. Error bars are one standard deviation. The block-size axis is logarithmic. A block size of 1 corresponds to uniformly random masking, and 45 to masking whole sensors. The three panels show the sensor estimation MAE, the overall F1 score and the system-wise F1 score against the block size.

5.3 Effect of the Masking Prior

To analyze the impact of the masking prior on the downstream task performance, we swept the masking block size b for all three codebook sizes. For every setting we used ten different seeds and the results can be found in Fig. 4. Uniform masking ($b = 1$) is clearly the worst setting for sensor estimation, at approximately 9.3 K for all codebook sizes, while any block size $b \geq 5$ reaches a plateau of 8.5–8.8 K with a possible optimum around $b = 5$ to 9. We observed a reversed ordering for the anomaly detection task. The best performance is observed for $b = 1$ for every codebook. The performance decreases up to $b = 15$–25, and partially recovers for the largest block sizes. The masking prior therefore mildly trades the two tasks against each other, with blocky masking benefiting sensor estimation at a moderate cost in detection performance.

5.4 Hyperparameter sensitivity

During development we varied hyperparameters such as the channel dimensions or depth of the ViT and the tokenizer. However, with sensible hyperparameter choices the performance of both downstream tasks was largely independent of the hyperparameters. The only exception to this is the codebook size N_c of the tokenizer. Here we observe a trade-off between reconstruction quality and anomaly detection performance. Increasing N_c from 128 to 4096 lowered the reconstruction floor from 2.04 K to 1.02 K, but monotonically degraded the anomaly detection, with the overall F1 dropping from 0.76 to 0.72 and the system-wise F1 from 0.40 to 0.34. The sensor estimation performance, in contrast, remained essentially unchanged.

A plausible mechanism for the trade-off would be that a larger codebook allows anomalous subsequences to have corresponding codebook entries. A sufficiently good Stage 2 model is then able to select these codebook entries and more accurately reconstruct the anomalies as well as the nominal data, weakening the out-of-distribution signal and thus worsening anomaly detection performance. This inherent trade-off needs to be resolved for better training performance to lead to better downstream task performance. Here we feature the $N_c = 128$ result, as the smaller codebook acts as a regularizer that improves anomaly detection at no measurable cost in sensor estimation.

6. Discussion

Sensor estimation gap A substantial gap remains between the autoencoder floor and the sensor estimation task. A possible cause is the reconstruction–generation trade-off. Under this hypothesis, a larger codebook encodes more informative codes that are harder to predict, so that the task difficulty grows with the codebook size. However, increasing N_c from 512 to 4096 grows the number of codes used by a factor of 7.1 (as codebook utilization drops from 0.96 to 0.85), but the predictive perplexity of the masked tokens only by a factor of 3.0 (from 6.8 to 20.6), so the additional codes are not disproportionately harder to predict. Moreover, the $N_c = 128$ tokenizer has a substantially worse floor on the dynamic sensors (3.51 K on TSA1, against 1.66 K

for $N_c = 4096$), yet reaches the same sensor estimation error, showing that the floor is not the limiting factor. Consequently, we suspect that a large portion of the gap is due to the difficulty of estimating the data itself. The clearest example is the tank temperature, as it is estimated poorly by the FM and both baselines. TW is largely driven by the heating demand and the auxiliary heating source. There is neither metadata nor time-series data about this in the dataset. Adding data from the auxiliary heater (e.g. burner on/off) or the hot-water draw would improve this reconstruction more than any model change. More generally, even though some of the gap between the floor and the task performance might be closed by improving the model itself, task performance would improve more from additional data sources.

Anomaly detection gap. The relevant claim in this comparison is generality rather than detection accuracy. All baselines in Table 2 are single-purpose anomaly detectors. Our model reaches its performance without being designed specifically for anomaly detection, and the same model also performs sensor estimation. Our results exceed MSE-βVAE, are on par with Rescaled PCA-R, even exceeding it on the system-wise metric with uniform masking, and sit below only MVE-βVAE. The remaining gap to MVE-βVAE is not due to a lack of uncertainty estimation in the anomaly scores. Stage 2 predicts a full categorical distribution over codebook entries, is trained with a proper scoring rule, and the anomaly score is likelihood-based. It is a valid bound for uniform masking and a likelihood-inspired reweighting for the block prior. A more likely explanation is the regularization of the latent space or the computation of the anomaly score itself. The resolution of the tokens may contribute as well. As tokens cover 32 time steps, the model cannot express uncertainty at a finer resolution. The strongest baseline, MVE-βVAE, instead estimates uncertainty directly in the signal domain, at full temporal resolution.

Decoupling of stage 1 and task performance. The decoupling of Stage 1 quality and downstream performance is one of the more consequential findings of the paper. The premise of scaling a framework like ours is that better training performance translates into better downstream task performance. We observe the opposite: a larger codebook, which clearly improves reconstruction, leaves sensor estimation unchanged and degrades anomaly detection. Anomaly detection relies on the model being better at reconstructing nominal rather than anomalous data. A more expressive codebook improves reconstruction for both and thus weakens this signal, while the smaller codebook acts as a regularizer. Such a strong impact of the regularization on anomaly detection performance was also found in Ebmeier et al. (2025). For a large-scale foundation model, into which we would want to invest more capacity and data, this conflict has to be resolved.

Smoothing. Due to the compression of 1440 steps into 45 tokens in the tokenizer and training with the standard mean-squared error loss, we observe smoothing even in the autoencoder floor, without any masking. Stage 2 then decodes deterministically by taking the most likely token. This further pushes reconstruction toward the conditional mean, directly visible as smoothing. For many downstream applications this is not a problem. However, if control decisions, such as pump switching in a solar thermal system, rely on sharp increases and decreases of a sensor, this effect has to be considered.

7. Conclusion and Outlook

A two-stage masked autoencoder, trained once on nominal data from multiple real solar thermal installations, can serve several downstream tasks zero-shot by expressing each as an inference-time mask. It estimates missing sensors better than a linear cross-sensor baseline and detects anomalies on par with all but the strongest dedicated reconstruction baseline, without having been trained for either task. Missing sensors and dropouts, which are endemic in this domain, are handled by construction rather than by preprocessing.

Downstream task performance is nonetheless modest. Anomaly detection performance lags behind the state-of-the-art dedicated system and sensor estimation error is still above what would be reasonable to use in operational setups. Neither the capacity of the Stage 2 model nor masking priors closed the gap between the tokenizer floor and the sensor estimation error. Most fundamentally, better Stage 1 training currently does not translate into better downstream performance: it leaves sensor estimation unchanged and degrades anomaly detection. This decoupling has to be resolved before additional scaling can pay off.

There are three clear directions following from this. First, we may use different masking distributions at inference and during training. For example, we could vary the masking distribution during training and use only the valid bound ($\alpha = 0$) at inference, returning anomaly detection to its roots as out-of-distribution detection. Second, the uncertainty internally estimated by the model should be carried through the decoder into

data space. This may counteract smoothing and might be necessary as a step toward operational setups. Third, the tokenization step has to be re-evaluated. More diverse data sources will need specific modality-dependent tokenization steps and it has to be critically evaluated whether vector quantization is the right choice for the time-series modality.

Beyond the model, the natural extension is data. The model could be further extended with a multi-domain dataset encompassing different types of small- to medium-scale heating systems and incorporating more diverse data sources, if such a dataset existed. Building one is, in our view, a more important prerequisite for a foundation model in this domain than any further architectural refinement. Despite the current lack of such a dataset, this architecture provides a viable foundation for general-purpose energy system modeling.

Acknowledgments

This project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645. The authors thank the international Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Florian Ebmeier.

References

Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A Weyn, Haiyu Dong, et al. A foundation model for the Earth system. *Nature*, 641(8065):1180–1187, 2025. doi: 10.1038/s41586-025-09005-y.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. doi: 10.48550/arXiv.2108.07258.

Christopher F Brown, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, et al. AlphaEarth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025. doi: 10.48550/arXiv.2507.22291.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11305–11315, June 2022. doi: 10.1109/CVPR52688.2022.01103.

Aaron Defazio, Xingyu Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The road less scheduled. *arXiv preprint arXiv:2405.15682*, 2024. doi: 10.48550/arXiv.2405.15682.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Florian Ebmeier, Nicole Ludwig, Georg Martius, and Volker H. Franz. PaSTS: An operational dataset for domestic solar thermal systems. In *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems*, e-Energy '24, page 529–534, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704802. doi: 10.1145/3632775.3662159.

Florian Ebmeier, Nicole Ludwig, Jannik Thuemmel, Georg Martius, and Volker H Franz. Fault detection in solar thermal systems using probabilistic reconstructions. *arXiv preprint arXiv:2511.10296*, 2025. doi: 10.48550/arXiv.2511.10296.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12868–12878, 2021. doi: 10.1109/CVPR46437.2021.01268.

Lukas Feierl, Viktor Unterberger, Claudio Rossi, Bernhard Gerardts, and Manuel Gaetani. Fault detective: Automatic fault-detection for solar thermal systems based on artificial intelligence. *Solar Energy Advances*, 3, January 2023. ISSN 2667-1131. doi: 10.1016/j.seja.2023.100033.

Google DeepMind. Gemini Diffusion: Scalable non-autoregressive language modeling, 2025. URL `https://deepmind.google/models/gemini-diffusion/`. Accessed: 2026-02-19.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. doi: 10.1109/CVPR52688.2022.01553.

Amirreza Heidari and Dolaana Khovalyg. Short-term energy use prediction of solar-assisted water heating system: Application case of combined attention-based LSTM and time-series decomposition. *Solar Energy*, 207:626–639, 2020. ISSN 0038-092X. doi: 10.1016/j.solener.2020.07.008.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Paulo Lissa, Conor Deane, Michael Schukat, Federico Seri, Marcus Keane, and Enda Barrett. Deep reinforcement learning for home energy management system control. *Energy and AI*, 3:100043, 2021. ISSN 2666-5468. doi: 10.1016/j.egyai.2020.100043.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025. doi: 10.1038/s41586-024-08252-9.

Mashud Rana, Subbu Sethuvenkatraman, Rahmat Heidari, and Stuart Hands. Solar thermal generation forecast via deep learning and application to buildings cooling system control. *Renewable Energy*, 196:694–706, 2022. ISSN 0960-1481. doi: 10.1016/j.renene.2022.07.005.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 103131–103167. Curran Associates, Inc., 2024. doi: 10.52202/079017-3277.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html`.