Interpretability-by-Design with Accurate Locally Additive Models and Conditional Feature Effects

Vasilis Gkolemis[1,2] Loukas Kavouras[2] Dimitrios Kyriakopoulos[2] Konstantinos Tsopelas[2] Dimitrios Rontogiannis[4] Giuseppe Casalicchio[5] Theodore Dalamagas[2] Christos Diou[1,3]

[1]Department of Informatics and Telematics, Harokopio University of Athens [2]IMSI, ATHENA RC [3]Archimedes AI, ATHENA RC [4]Max Planck ISS [5]LMU Munich

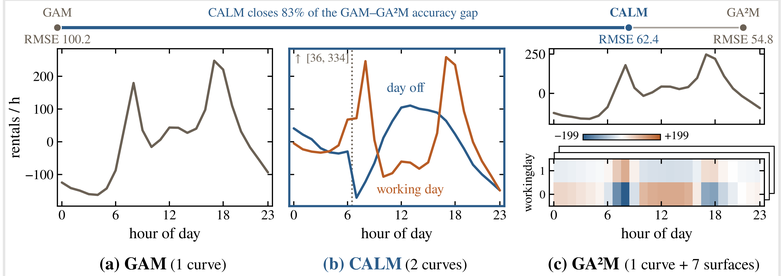
CALM: In a nutshell

Interpretable-by-design, without a (high) accuracy tax.

CALM sits *between* GAMs and GA^{2}Ms:
close to GAMs in **interpretability** and close to GA^{2}Ms in **accuracy**.

The GAM / GA^{2}M dilemma

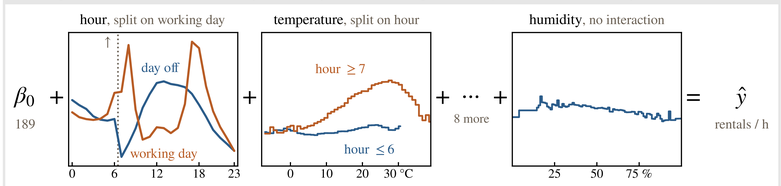
- **GAMs** ($y = \beta_0 + \sum_i f_i(x_i)$) are fully auditable, one 1D curve per feature, but they **underfit whenever features interact**.
- **GA^{2}Ms** add pairwise terms $f_{ij}(x_i, x_j)$ and gain accuracy, but **lose two interpretability properties**: (i) a prediction no longer splits into one contribution per feature, and (ii) no single plot shows how a feature affects the prediction.
- **CALM** keeps univariate shape functions, but allows *several* per feature, each active in a different region. Effects stay locally additive, yet vary across regions, so interactions are captured without ever leaving 1D plots.



The effect of the hour on bike rentals (Bike Sharing). The GAM averages the interaction with the working day away, CALM ($d_{\max} = 1$) keeps it as two curves, and the GA^{2}M spreads it over a curve and seven surfaces.

CALM: Model overview

On Bike Sharing, the hour splits on the working day: on working days rentals peak at the morning and evening commute, whereas on days off they rise to a single broad peak around midday.



The prediction sums one plot per feature, each holding one curve per region. The temperature splits on the hour; the humidity interacts with no feature and keeps one curve.

How to use CALM

```
$ pip install calm-additive
from calm_additive import CALMRegressor   # or CALMClassifier
calm = CALMRegressor(max_partition_depth=2,  # up to 4 curves per feature
                     teacher="xgb")          # or your own black box
calm.fit(X_train, y_train)
calm.predict(X_test)
calm.print_summary()        # the model on one screen
calm.plot_curves("hour")    # one curve per region
```

Reading a CALM plot

In the CALM panel on the left, each curve is the contribution of the hour in one region: at 8 am it is $+246$ on a working day and -122 on a day off.
Dotted lines mark *hidden jumps*: the temperature and the year split on the hour at 6.5, so moving the hour across 6.5 adds a jump in $[36, 334]$ to the change of its own curve. Arrows summarize it: ↑ means $\Delta y > \Delta f_i$, ↓ means $\Delta y < \Delta f_i$, ↕ means it depends.

Interpretability property	GAM	GA^{2}M	CALM
P1 Local feature contribution	✓	✗	✓
P2 Regional feature sensitivity	✓	✗	✓*
P3 Global feature property (e.g. monotonicity)	✓	✗	✓
Captures interactions	✗	✓	✓

*exact unless the change crosses a dotted line; the arrows still answer qualitative questions. In a GA^{2}M the contribution is not even unique.

Humans actually read CALM plots better

In a within-subjects user study ($N = 25$), participants answered questions on the effect of the hour in a Bike Sharing model, once from CALM's curves and once from a GA^{2}M's curve and heatmaps. They answered **54 %** correctly with CALM and **34 %** with the GA^{2}M ($p < 0.05$), and the gap is widest on how long the hour takes to lower the prediction (**40 %** vs. **12 %**). The advantage holds among the 13 advanced or expert participants (65.4 % vs. 30.8 %).

CALM: Conditionally Additive Local Models

$$g\left(\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]\right) \;=\; \beta_0 + \sum_{i=1}^{d} f_i^{(r_i(\mathbf{x}_{-i}))}(x_i)$$

- **Conditional feature effects.** Each feature i owns a *set* of univariate shape functions $\{f_i^{(r)}\}_{r=1}^{R_i}$.
- **Region selection.** $r_i(\mathbf{x}_{-i})$ picks which curve is active, based only on the other features.
- **Readable regions.** Each region is a conjunction of at most $d_{\max}$ threshold rules, read off a tree T_i over $\mathbf{x}_{-i}$: $\mathcal{R}_i^{(r)} = \{\, \mathbf{x}_{-i} \mid \wedge_k (x_{j_k} \,\mathrm{op}_k\, \tau_k) \,\}$. So CALM captures interactions of up to $d_{\max} + 1$ features.
- **Cheap expressivity.** Each feature has at most $2^{d_{\max}}$ curves, yet every instance picks one region per feature, so the curves combine into up to $2^{d \cdot d_{\max}}$ distinct additive models.
- **Two knobs.** They bound what a reader must inspect: $d_{\max}$ (2 by default) caps how many curves each feature has and K how many interactions the model captures in total.

Training: a 3-step distillation pipeline

Step 1: Reference model.
Fit a black box $f_{\texttt{ref}}$ (XGBoost by default). It is used *only* to detect interactions, then discarded.

Step 2: Find near-additive regions.
Interactions show up as *heterogeneity* of the local effects h of x_i around their average $\mu_i^{\mathcal{R}}$:

$$H_i^{\mathcal{R}} \;=\; \sqrt{\mathbb{E}\Big[\big(h(X_i, \mathbf{X}_{-i}) - \mu_i^{\mathcal{R}}(X_i)\big)^2\Big]}, \qquad \mathbf{X}_{-i} \in \mathcal{R}.$$

A CART-style tree T_i per feature splits on the x_j that most reduces $H_i^{\mathcal{R}}$. Then a greedy selection keeps only the trees that improve an additive surrogate of $f_{\texttt{ref}}$, so a redundant tree is discarded and CALM stays small. Under specific assumptions, CALM's error in approximating $f_{\texttt{ref}}$ is at most d times the sum of the squared heterogeneities, so partitioning tightens this bound.

Step 3: Region-aware backfitting.
Shape functions are fitted by gradient boosting, each update restricted to its own region. Regions differ across features, so the updates are *coupled*: a single additive predictor with region-gated components, whose idealized version converges to an empirical risk minimizer over the fixed-tree CALM class.

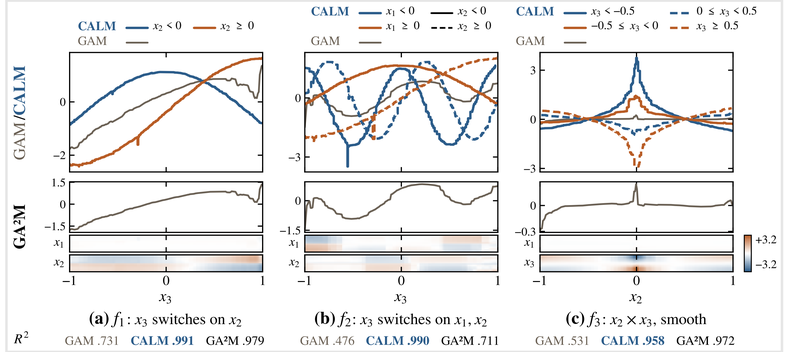
Funded by the European Union under Grant Agreements No. 101289094 (AIRIS) and No. 101135826 (AIDAPT). Views and opinions expressed are however those of the authors and do not necessarily reflect those of the European Union or the granting authority (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

CALM: Paper & Code

- **Paper:** arxiv.org/abs/2602.16503
- **Page and docs:** givasile.github.io/calm-additive
- **Package:** github.com/givasile/calm-additive
- **Experiments:** github.com/givasile/calm-paper

Synthetic tasks: near-GA^{2}M accuracy, simpler explanation

Three independent features $x_i \sim \mathcal{U}(-1, 1)$, one interaction per case: in Case 1, x_3 switches shape on the sign of x_2; in Case 2, on the signs of x_1 and x_2, which a GA^{2}M cannot express; in Case 3, x_2 and x_3 interact as a smooth product, which CALM approximates by regions.



Top: CALM's curves, one per region, over the GAM's single curve (gray). Bottom: the GA^{2}M's main curve and its two heatmaps, three views to combine per feature.

The GAM misses every interaction. CALM matches the GA^{2}M in Case 1, is far more accurate in Case 2 and slightly below it in Case 3.

Real datasets: 24 tabular tasks

We compare CALM with the black box XGBoost, two GAMs and two GA^{2}Ms over 5-fold CV, with no tuning per dataset. A win or a loss counts only when two means differ by more than the larger of their standard deviations.

CALM wins/draws/losses	BB	GAM		GA^{2}M	
	XGB	NAM	EBM	EB^{2}M	NODE
Classification (10)	1/5/4	5/5/0	3/7/0	0/7/3	1/7/2
Regression (14)	2/1/11	10/4/0	8/6/0	1/5/8	4/6/4

- **Never behind a GAM:** CALM closes a median 37 % of the GAM to XGBoost gap in regression and 31 % in classification.
- **Competitive with GA^{2}Ms:** it matches or beats NODE-GA^{2}M on most datasets; EB^{2}M is more accurate on 8 of the 14 regression datasets, the cost of 1D curves instead of 2D surfaces.
- **Sparser & faster:** CALM adds 8.2 curves to the GAM in regression and 7.9 in classification, against 14.3 and 15.2 EB^{2}M surfaces; it fits 19 of the 24 datasets in under a minute (median 21 s).
- **Stable:** on 77 % of the 376 features, all 5 fold refits agree on whether to split the feature (90 % on at least 4).

Takeaway

A new model class between GAMs and GA^{2}Ms:
it never loses to a GAM, and it is read more accurately than a GA^{2}M.