从变系数模型到个体化响应函数学习

一条分散但连续的方法谱系

From Varying Coefficients to Unit-Level Response Function Learning:
A Narrative Review

Anonymous Author(s)

Affiliation withheld for draft review

July 2026

摘要

中文摘要。经典监督学习通常把不同样本视为同一条总体 response law 在不同输入处的观测。与此不同，统计学、纵向数据分析、空间统计、因果推断和现代元学习中已经形成了一组允许不同 unit 拥有不同 response function 的方法。本文将这棵分散的方法树分成四条主线：由 observed modifier 索引响应曲线的 varying-coefficient 与 functional-coefficient models；借助重复测量推断 subject-specific 曲线的 random-effects 与 functional mixed-effects models；在识别假设下估计异质性处理效应或连续剂量反应的 causal machine learning；以及从 context 或 evidence 生成整条函数的 Hypernetwork、Neural Process 与相关元学习方法。本文区分真实 factual prediction、半合成 response benchmark 与纯合成机制恢复，审计代表性实验，并总结适用数据、失败边界和未形成统一机器学习范式的原因。核心结论是："每个 unit 有自己的 response function"并非空白；仍未被统一解决的是，如何把 world-side token outcome generation 与 learner-side prediction 分开。Model-level U 选择 token, $U = u^\star$ 后实际 outcome 为 $Y_{u^\star}(x) = f_\theta(x, E; u^\star)$。Learner 看到 predictors $X = x$ 后,先把它当 factual evidence 做 point、Gaussian 或 Cauchy abduction，再把它当 factual query 输入共享 generator。Distributional prediction 为 $p_{\phi,\theta}(y \mid x^F, x^Q) = \int p_\theta(y \mid x^Q, u) q_\phi(u \mid x^F) du$。Same-token counterfactual 固定 factual abduction，只改变 query。Cauchy 有 location/scale 但无 finite mean/variance；它表达 globally open candidate uncertainty，不保证任何 outcome 都能由任何 unit 生成。

Abstract. This narrative review studies a family of methods that assign or infer a distinct response function for each observational unit, subject, location, state, or task. We connect four literatures: varying- and functional-coefficient models, random and functional mixed-effects models, heterogeneous treatment and continuous dose-response estimation, and modern conditional function generators such as Hypernetworks and Neural Processes. We separate real factual prediction from semi-synthetic counterfactual benchmarks. For the current method, model-level U selects a token, $U = u^\star$ is its actual embedding, and $Y_{u^\star}(x) = f_\theta(x, E; u^\star)$. Predictors first provide factual evidence for point, Gaussian, or Cauchy unit abduction and then provide the factual generator query. Distributional prediction composes candidates as $p_{\phi,\theta}(y \mid x^F, x^Q) = \int p_\theta(y \mid x^Q, u) q_\phi(u \mid x^F) du$. A same-token counterfactual fixes factual abduction and changes only the query. Cauchy abduction has location and scale but no finite mean or variance; its heavy tails do not imply unrestricted outcome support.

关键词： varying-coefficient model；functional coefficient；random effects；functional mixed effects；individualized treatment effect；dose-response；Hypernetwork；Neural Process；unit-level response function；one-shot function learning

Keywords: varying coefficients; functional mixed effects; individualized response; dose-response; conditional function generation; unit-level response function; one-shot learning

1 问题、范围与综述方法

1.1 研究问题

设 unit i 的 response law 为

$$R_i : \mathcal{X} \to \mathcal{P}(\mathcal{Y}).$$

本文关心的不是某一种估计器，而是下面这个更基础的问题：

> 学术界以哪些形式允许不同 unit 拥有不同 response function? 这些方法依靠什么共享结构获得可学习性，在哪类数据上表现良好，又为什么没有以一个统一名称成为通用机器学习范式?

这里的 "unit" 可以是病人、消费者、地理区域、设备、时间序列所处的局部状态，也可以是元学习中的 task。不同文献对 unit 的定义并不相同，因此 "看起来都在学习一条个体曲线" 并不意味着它们具有相同 estimand 或 identification semantics。

1.2 三层纳入范围

本文采用叙述性而非穷尽式检索，把文献分成三层：

1. **直接谱系**：论文明确使用 varying-coefficient、functional-coefficient、random-coefficient 或 subject-specific functional effects；
2. **实质同构**：方法从 unit/task evidence 生成参数、函数或函数分布，即使不使用 VCM 名称；
3. **因果特化**：方法学习 CATE、ITE 或 continuous dose-response，但必须单独核对其 ignorability、overlap、cross-world 或实验设计假设。

泛泛的 "个性化预测" 不自动纳入。只有当模型显式表达 unit-conditioned relation、curve、mechanism 或 function distribution 时，才属于本文讨论的核心对象。

1.3 证据等级

本文把实验主张分成三档：

- **真实 factual prediction**：输入和 outcome 均来自真实观测，可做 held-out prediction；
- **半合成 response/counterfactual benchmark**：covariates 来自真实数据，但 treatment、outcome 或完整 response curve 由研究者生成；
- **纯合成机制恢复**：unit function 与所有 query outcome 均由 simulator 给出。

这一区分很重要。半合成数据可以比较方法是否恢复了已知曲线，却不能直接证明真实病人的 counterfactual response 已被恢复。

2 一个统一形式与三个必要区分

2.1 从共享函数到 unit-conditioned function

经典监督学习通常估计一条共享 law：

$$Y_i \sim R(\cdot \mid X_i).$$

最简单的 unit heterogeneity 是让 observed descriptor Z_i 修改这条 law：

$$R_i(\cdot \mid x) = G_\theta(Z_i)(\cdot \mid x).$$

进一步，可以把 unit function 本身视为 latent random object：

$$F_i \sim P_F, \qquad Y_{it} \sim F_i(\cdot \mid X_{it}), \qquad p(F_i \mid \mathcal{O}_i)$$

由 unit 的证据集合 $\mathcal{O}_i$ 更新。前一种写法给出一个确定的 context-indexed function；后一种写法保留了 learner 对 unit function 的后验不确定性。

2.2 observed-indexed 不等于不可约的 unit-specific

经典 VCM 可以写成

$$\mathbb{E}(Y_i \mid X_i, Z_i) = \beta_0(Z_i) + X_i^\top \beta(Z_i).$$

因此每个 unit 确实被赋予一条

$$G_i(x) = \beta_0(Z_i) + x^\top \beta(Z_i),$$

但两个具有相同 Z_i 的 unit 得到相同函数。VCM 学习的是**观测 modifier 到函数的映射**，不是从该 unit 的独有历史中识别一条任意 latent function。

2.3 稳定 unit type、动态 state 与 event noise

在纵向或动态问题中，更稳妥的表示是

$$Y_{it} \sim R_\theta(\cdot \mid A_{it}, U_i, S_{it}, E_{it}),$$

其中：

- U_i 是跨 repeated events 保持的 response-relevant unit type；
- S_{it} 是随时间变化的内部或外部状态；
- E_{it} 是事件级随机性；
- A_{it} 是当前 query、exposure、action 或 treatment。

如果把 S_{it} 与 U_i 混在一起，“同一个 unit 的 response function” 就会随每次记录任意改变；如果把 E_{it} 吸收到 U_i，模型又会把偶然 realization 误当成稳定个性。

2.4 predictive heterogeneity 不等于 causal response

下面两个对象在观测数据中可能数值相似，但语义不同：

$$m(a, x) = \mathbb{E}(Y \mid A = a, X = x),$$

$$\mu(a, x) = \mathbb{E}\{Y(a) \mid X = x\}.$$

只有在 treatment assignment、confounding、consistency 和 support 条件成立时，前者才能识别后者。更进一步，即便识别了 $\mu(a, x)$，它通常仍是特征为 x 的人群条件平均，而不是某个不可约个体的完整 $Y_i(a)$。

3 方法谱系

3.1 前史：random coefficients 与 mixed effects

个体参数并不是从 VCM 才开始出现。Laird 与 Ware 的纵向 random-effects model 把总体固定效应与 subject-specific random effects 分开：

$$Y_{it} = X_{it}^{\top}\beta + Z_{it}^{\top}b_i + \varepsilon_{it}, \qquad b_i \sim P_b.$$

重复测量使研究者能够对 b_i 做 shrinkage estimation；新 subject 的估计则借助 population distribution。(Laird and Ware, 1982) 奠定了这一纵向建模传统。

Guo 后来把 random effect 从有限维向量扩展为随机函数：

$$G_i(t) = G_0(t) + B_i(t),$$

并在统一平滑空间中同时估计 population-average 与 subject-specific curves。(Guo, 2002) 因而是经典统计学中最接近 “每个 subject 有一条自己的 response curve” 的文献之一。

这条路线的代价是：它通常需要同一 subject 的多次观测。只有一个 outcome 时，个体随机斜率或随机函数主要由 population prior 与协变量决定，unit-specific evidence 极弱。

3.2 Varying-coefficient models：从一个参数向量到一族局部关系

局部回归在 Cleveland、Grosse 与 Shyu 的工作中已经提供了 “参数随局部位置变化” 的计算基础。(Cleveland et al., 1992) Hastie 与 Tibshirani 将这类思想系统化为 varying-coefficient models：

$$g\{\mathbb{E}(Y \mid X, R)\} = \sum_{j=0}^{p} X_j \beta_j(R_j).$$

他们给出平滑估计与 backfitting 算法，并把框架扩展到 Gaussian regression、广义线性模型和 Cox proportional hazards，从而把 GAM、dynamic GLM 与局部 effect modification 连接起来。(Hastie and Tibshirani, 1993)

VCM 的关键贡献不只是 “多加了 interaction”。它把一个 unit 的 descriptor R_i 映射成一组局部 coefficient functions，从而间接产生该 unit 的 response function。与此同时，它用 “对 X 保持简单、只对低维 modifier 做非参数平滑” 降低完全非参数回归的维数灾难。

3.2.1 1993 原论文究竟展示了什么？

原论文的第一个定量例子不是 Burns 或 Aorta，而是一个单缸乙醇发动机的 NOx 排放实验，共 $n = 88$ 个观测。令 E 表示 air–fuel equivalence ratio，C 表示 compression ratio，作者拟合

$$\mathbb{E}(\mathrm{NOx} \mid E, C) = \beta_0(E) + \beta_1(E)C.$$

这里 E 是 effect modifier，C 是保持局部线性的 predictor。最终 varying-slope 模型的 residual sum of squares 为 2.65（72 degrees of freedom）；表中三个更简单的结构分别为 5.19、6.33 和 3.20。论文据 approximate F-tests 判断：E 与 C 的 interaction 不能省略，且 C 的 coefficient 需要随 E 非线性变化。这是有意义的结构比较与 in-sample fit，但不是现代意义上的 held-out benchmark。(Hastie and Tibshirani, 1993)

第二个主要应用把 VCM 放进 logistic regression，研究回顾性 myocardial-infarction 数据中 blood pressure、cholesterol ratio 与 antihypertensive treatment 的交互。作者明确提醒：风险因子是

在病例发生 MI 之后测量的，因此该例更多是在展示 flexible interaction 以及 naive observational interpretation 的陷阱，不能当作个体化 treatment effect 的证据。(Hastie and Tibshirani, 1993)

这个历史事实也澄清了本文的术语：VCM 中的 modifier 可以是年龄、时间、位置、剂量或系统状态，它为具有相同 modifier 的 observations 分配同一条条件响应函数；它不自动等于某个 subject 的独有 evidence，更不自动识别不可约的 individual mechanism。

Fan 与 Zhang 随后建立了 local linear、two-step optimal estimation 等统计理论；当不同 coefficient functions 具有不同 smoothness 时，两步估计可以优于朴素 one-step procedure。(Fan and Zhang, 1999)；其后续综述总结了 VCM 在经济、金融、流行病学、医学、生态、纵向和 survival data 中的发展。(Fan and Zhang, 2008)

3.3 直接后代：时间、空间、高维和树结构

3.3.1 纵向与 functional data

Hoover 等研究了不规则重复测量中的 time-varying coefficients，允许同一 subject 内相关，并给出 smoothing spline、local polynomial、cross-validation 和收敛性质。(Hoover et al., 1998)

Huang、Wu 与 Zhou 进一步用 basis functions 做全局平滑，提供 subject bootstrap、置信区域与假设检验，并处理 observation times 稀疏且不规则的 repeated measurements。(Huang et al., 2002)

3.3.2 非线性时间序列

Chen 与 Tsay 的 functional-coefficient autoregressive model 让 AR coefficients 随滞后状态变化，包含或近似多种 threshold 和 smooth-transition dynamics。(Chen and Tsay, 1993)

Cai、Fan 与 Yao 发展了局部线性估计、模型检验和 forecasting 方法。他们强调：只平滑一个低维状态 U_t，同时保持对 lag vector 的线性结构，可以比完全非参数时间序列更节省样本。(Cai et al., 2000)

3.3.3 空间异质性

Geographically Weighted Regression 在每个地理位置拟合局部回归关系，并用空间 kernel 决定邻近观测的权重。它本质上是以 location 为 modifier 的变系数模型，适合 housing、环境、流行病学和区域经济中的 spatial nonstationarity。(Brunsdon et al., 1996)

3.3.4 学习 modifier 与高维选择

传统 VCM 预先指定哪个变量是 modifier。Adaptive VCM 进一步学习一个低维 index，使 coefficients 成为未知线性组合的函数，从而在灵活性和维数约简之间折中。(Fan et al., 2003)

在 large-p, small-n 情形中，Xue 与 Qu 通过非凸正则化选择非零 coefficient functions，证明 oracle property 与最优收敛率；其模拟中真实模型选择频率高于 LASSO、adaptive LASSO 和 SCAD。(Xue and Qu, 2012)

3.3.5 从平滑变化到分段变化

Model-based recursive partitioning 先拟合一个参数模型，再检验 parameter instability，并沿最不稳定的 partitioning variable 递归切分，因此得到 piecewise-constant 的局部响应模型。(Zeileis et al., 2008)

Decision-tree-boosted VCM 则直接用 boosted trees 表示 effect modifiers，减少对 modifier space 的平滑结构假设，并保留 coefficient-level explanation。(Zhou and Hooker, 2022)

截至 2026 年, 这条统计学路线仍在继续发展。Fabbrico、Pedone 与 Stingo 的 INVENT 模型把 covariates 同时作为 predictors 与 effect modifiers, 用正交 spline basis 和稀疏 Bayesian prior 选择线性、非线性 main effects 与 interactions, 并显式处理这种双重角色造成的 likelihood identifiability 问题。它产生 subject-specific coefficients，但这些 coefficients 仍是 observed covariates 的平滑函数，而不是从独有 repeated history 推断的 latent unit function。(Fabbrico et al., 2026)

3.4 因果机器学习：从异质效应到整条 dose-response curve

二元 treatment 下，causal forest 估计

$$\tau(x) = \mathbb{E}\{Y(1) - Y(0) \mid X = x\},$$

将 random forest 的局部分组能力用于 heterogeneous treatment effects 与统计推断。(Wager and Athey, 2018)

Bayesian multi-task Gaussian Process 把 factual 与 counterfactual outcomes 当作 vector-valued function，并给出个体 treatment-effect estimate 的 pointwise credible intervals。(Alaa and van der Schaar, 2017)

连续 treatment 把一个 contrast 扩展成整条曲线。DRNet 使用 treatment-specific layers 与 dose strata heads 学习多个 treatment 的 dose-response。(Schwab et al., 2020)

SCIGAN 用生成器补全 counterfactual response curves，再训练 inference network 为新样本预测曲线。(Bica et al., 2020)

VCNet 是这条因果谱系与经典 VCM 的最直接连接。它让 neural prediction head 的权重成为 treatment t 的 spline functions：

$$\widehat{\mu}(t, x) = f_{\theta(t)}\{z(x)\}.$$

论文同时估计 generalized propensity density，并用 functional targeted regularization 估计整条 average dose-response curve。(Nie et al., 2021)

然而，这些论文中的 “individualized” 通常指

$$\mu(t, x) = \mathbb{E}\{Y(t) \mid X = x\},$$

即 observed covariates 所索引的条件平均 potential-outcome curve。它们没有从一个人的一次 factual treatment 中识别该人的不可约 latent $Y_i(t)$ trajectory。

3.5 现代 function generators：同一思想的深度学习重述

Hypernetwork 用一个网络生成另一个网络的参数：

$$\theta_i = h_\phi(O_i), \qquad G_i(x) = f_{\theta_i}(x).$$

原始 HyperNetworks 在 character-level language modeling、handwriting generation 和 neural machine translation 上达到接近当时 SOTA 的结果，并在 CNN 上以更少参数取得有竞争力的图像识别表现。(Ha et al., 2017)

Conditional Neural Process 从少量 context input-output pairs 推断一条 task-specific function，把 Gaussian Process 的函数先验思想与可扩展神经网络结合，并用于回归、分类和 image completion。(Garnelo et al., 2018)

这类方法证明了 “从 evidence 生成一条函数” 已经进入主流 ML。它们与 VCM 的区别在于 modifier 不再只是一个低维标量，生成的也可能是整张网络；但它们通常不自动拥有 causal 或 physical mechanism semantics。

3.6 Cauchy random effect：组件级邻居而非同一设定

RECaST 用 random calibration effect $\beta_i = g(\theta_T, x_i)/f(\theta_S, x_i)$ 把 source model 的 prediction 校准到 target population。在 source/target 都取线性结构、feature $x_i \sim \mathcal{N}_p(0, I_p)$ 时，两个线性投影之比 exact 服从 Cauchy；其 heavy tails 允许较大的 source–target discrepancy，并服务于 posterior predictive uncertainty 与 prediction-set coverage。(Hickey et al., 2024)

这是当前 Cauchy 选择的重要**组件级邻居**，但不是相同 learning object。RECaST 的 β_i 校准 source-to-target mapping，论文也明确采用 Bayesian parent-parameter posterior；它不从单个 unit evidence abduce 对 fixed actual $U = u^\star$ 的 candidate embedding law，也不学习 $u \mapsto \{x \mapsto p_\theta(\cdot \mid x, u)\}$ 的 shared generator family。因此，可以借用它对 Cauchy ratio、heavy-tail robustness 与 coverage 的论证，不能把它写成已经提出了本文的四层 ontology。

4 四类方法到底学习了什么?

方法族	典型形式	unit evidence	primary object	最强可解释含义
VCM / functional coefficient	$G_i(x) = \beta_0(Z_i) + x^\top \beta(Z_i)$	observed modifier Z_i	确定的局部 coefficient functions	相同 Z 的 units 共享同一 response function
Random/functional mixed effects	$G_i = G_0 + B_i$	group identity + repeated outcomes	random coefficient/function posterior	在层级先验下的 subject-specific curve
CATE / dose-response	$\mu(a, x) = E\{Y(a) \mid X = x\}$	pretreatment covariates	conditional potential-outcome mean/contrast	在识别假设下的条件平均因果响应
Hypernetwork / Neural Process	$O_i \mapsto p(G_i \mid O_i)$ 或 $\theta_i = h(O_i)$	task context、descriptor、history	task/unit-conditioned function	数据驱动的函数生成；不自动是因果机制

对照 guard。表中的 random coefficient/function posterior 与 $p(G_i \mid \mathcal{O}_i)$ 是 mixed-effects、Gaussian Process 或 Neural Process 等相邻文献的真实学习对象, 不能为了贴合当前方法而改名或抹去。反过来，这些文献中的 prior/likelihood/posterior 语义也不能自动移植给当前的 $q_\phi(u \mid O)$：除非另行给出 generative model 与 Bayesian update，q_ϕ 只是 learner-side candidate uncertainty，不预设为 population prior 更新得到的 Bayesian posterior，也不包含 ϕ, θ 的参数不确定性。

这个表说明，文献中的 “每个 unit 一条函数” 至少有三种不同强度：

1. 根据已观测特征给 unit **分配**一条函数；
2. 根据重复证据对 unit 的随机函数做 **posterior shrinkage**；
3. 从一次或少量证据中 **amortize** 一个函数或函数分布。

第三种是最相邻的 function-generator 谱系，也最容易遭遇不可识别性；当前契约与它的关键区别是，直接 amortize 的是关于 fixed actual $U = u^\star$ 的 candidate law $q_\phi(u \mid O)$，再与 shared $f_\theta(x, E; u)$ composition；普通预测输出是 outcome distribution $p_{\phi,\theta}(y \mid O, x)$，而不是 sampled unit 或 function posterior。

5 哪些问题和数据最适合?

5.1 低维且有语义的 effect modifier

VCM 最理想的 modifier 通常是一到三个有明确科学含义的变量，例如：

- 年龄、calendar time、disease stage；
- treatment dose、exposure intensity；

- 经纬度或空间位置；
- population density、lagged system state；
- 一个通过 domain knowledge 或 single-index 学到的低维 score。

此时研究问题本身就是 “关系如何随这个轴变化”，coefficient curves 比黑箱 feature interaction 更容易解释。

5.2 重复测量与 longitudinal trajectories

当每个 subject 在多个时间点被观察时，random slopes、functional mixed effects 和 longitudinal VCM 可以同时利用 population sharing 与 within-subject evidence。典型场景包括生长曲线、生物标志物、CD4 trajectory、wearables、设备退化和个体行为序列。

这里必须区分 unit 与 observation：当 modifier $U = t$ 是 calendar time 时，$\beta(t)$ 首先表示 population relationship 随时间改变；它并不因为每个时间点不同，就自动成为 “每个人一条函数”。只有再加入 subject identity、random effects 或足够的 within-subject history，才能把 time-varying population effect 与 subject-specific trajectory 分开。

5.3 平滑非线性动力系统

如果系统动力学随当前状态连续变化，functional-coefficient AR 可以比固定 AR 更灵活，又比完全非参数高阶 autoregression 更省样本。若真实机制存在硬 threshold，tree/TAR 往往更合适。

5.4 空间非平稳性

当 “同一协变量在不同地点产生不同关系” 是核心科学问题时，GWR 与 spatial VCM 很自然。但必须防止把遗漏的空间 confounder 或不稳定局部样本误解成真实局部机制差异。

5.5 个体化 treatment 与 dose selection

连续 dose-response 在精准医疗、policy intensity、pricing 和 resource allocation 中尤其重要。这类方法需要：足够样本量、充分 overlap、可信的 pretreatment covariates，以及 treatment assignment mechanism 的合理建模。

5.6 多 task 快速适应

当训练数据包含许多相关 tasks，每个新 task 只给少量 context pairs 时，Neural Process、Hypernetwork 和 meta-learning 可以学习跨函数的生成规律。这里的关键不是 “每个 task 单独训练”，而是用大量旧 tasks 学到一个可快速条件化的 function prior。

6 代表性实验结果：必须区分证据类型

文献与数据	数据性质	代表性结果	可以支持什么	不能支持什么
Hastie–Tibshirani ethanol engine, $n = 88$	真实受控实验；主要是 in-sample fit	varying-slope VCM RSS 2.65；三个更简单结构为 5.19、6.33、3.20	approximate F-tests 支持 E–C interaction 与 nonlinear coefficient variation	没有 held-out test；不能把 RSS 直接包装成现代 predictive performance

文献与数据	数据性质	代表性结果	可以支持什么	不能支持什么
Hastie–Tibshirani myocardial-infarction example	真实回顾性 observational data	logistic VCM 展示 blood pressure/cholesterol curves 与 treatment 的 flexible interaction	VCM 可扩展到 binary outcome，并把 treatment-specific curves 变成可检查对象	covariates 在 MI 后测量；原文明确警告不能作 naive causal interpretation
Cai–Fan–Yao Canadian lynx，前 102 点训练、后 12 点预测	真实时间序列 held-out forecasting	一步预测 AAPE 相比 TAR 再降低约 25%；相对 linear AR(2)，nonlinearity test $p < 0.001$	平滑 state-dependent dynamics 可改善短期预测	作者指出 absolute error 只在小数第二位改善，实际幅度不应夸大
Hoover et al. HIV-exposed children growth	真实纵向流行病学	展示 gender、HIV status、maternal vitamin A effects 随时间变化	irregular repeated measurements 下可估 time-varying effects	原论文不是现代 predictive leaderboard
Huang–Wu–Zhou AIDS，283 名 HIV-positive males、每人 1–14 次测量	真实不规则纵向数据	展示 baseline、smoking、age、pre-CD4 effects 的时间变化，并发展 bootstrap inference	稀疏 observation times 下可做曲线估计与检验	不提供真实个体 counterfactual curve ground truth
VCNet：IHDP 与 News	真实 covariates；continuous treatment 与 outcome 半合成	testing AMSE：IHDP 0.117 vs DRNet 0.230；News 0.024 vs DRNet 0.114	VCM-style neural head 对平滑 continuous-treatment benchmark 有效	News 上 GPS 为 0.022，略优于 VCNet；结果不是现实因果真值验证
SCIGAN：TCGA、News、MIMIC	真实 covariates；response curves 半合成	$\sqrt{\text{MISE}}$：1.89、3.71、2.09；DRNet 为 3.64、4.98、4.45	生成式曲线补全在已知 simulator truth 上可优于若干 baseline	counterfactual outcomes 仍由研究者生成；论文明确说真实数据无法直接完整评估
HyperNetworks 的 sequence/image experiments	真实通用 ML benchmark	sequence modeling 接近当时 SOTA，CNN 以较少参数保持有竞争力	conditional parameter generation 可以扩展到高维任务	不是 unit-level causal response 的证据

这张表揭示了一个结构性事实：**经典 VCM 的真实数据证据多强调拟合、解释、推断或 forecasting；现代 individualized causal response 的完整曲线指标则主要来自半合成数据。**两者不能用同一证据标准混在一起。

7 这类方法为什么有效?

7.1 它在完全 pooling 与完全分离之间建立了正确折中

完全 pooling 假设所有 units 共用一条函数；完全分离则试图为每个 unit 独立拟合函数，通常不可学习。变系数、mixed effects 与 function generator 都采用中间道路：

$$\text{共享函数生成规律} \quad + \quad \text{unit-specific coordinate/evidence.}$$

7.2 它是一种结构化维数约简

完全非参数 $E(Y \mid X, Z)$ 需要在 (X, Z) 的联合空间平滑。VCM 保留对 X 的简单结构，只让 coefficient 随低维 Z 变化，因此能在灵活性与样本复杂度之间取得较好平衡。

7.3 它把异质性变成可研究对象

树或深网也能拟合 interactions，但一条 $\beta_j(t)$ 曲线直接回答："变量 X_j 的作用如何随年龄、时间、地点或剂量变化？"这对科学解释、机制假设生成和 subgroup discovery 很有价值。

7.4 它自然支持层级 shrinkage 与 uncertainty

Mixed effects、GP 与 Neural Process 不必把每条 unit function 当成一个确定点。它们可以保留 $p(G_i \mid \mathcal{O}_i)$，在 evidence 稀少时向 population structure 收缩，并给出函数或 contrast 的 uncertainty。

8 根本缺点与失败边界

8.1 一个 factual point 不能识别一条任意函数

若每个 unit 只观察

$$Y_i = G_i(X_i),$$

则对任意 $x^\star \neq X_i$，都有无穷多条函数在 X_i 处一致、在 $x^\star$ 处分叉。所以：

No cross-unit structure, no unit-level function learning.

所有可行方法都必须引入共享 family、smoothness、低秩、random-effects distribution、task prior、结构方程、intervention data 或其他 restriction。"每个 unit 有完全自由的一条函数"不是一个可学习模型。

8.2 高维 modifier 会重新遭遇维数灾难

VCM 的优势依赖 modifier 维度较低。Cai 等明确指出，多维 modifier 的局部平滑在实践中很快失效。Adaptive index、sparsity、trees 与 deep encoders 缓解了问题，但同时弱化了 coefficient curve 的直接解释。

8.3 条件线性既是优势，也是表达边界

说 VCM"本质上仍是线性模型"并不完全准确：$\beta_j(U)$ 与 link function 可以产生高度非线性的 joint response surface 和连续 interaction。更准确的限制是：**给定 modifier U 后，经典 VCM 通常对预先指定的 X 保持线性或加性结构。**这正是它节省样本并保持 coefficient interpretation 的来源，也使它不适合直接承担图像、文本等需要自动学习高阶 representation 的任务。

可以先用 basis、tree 或 deep encoder 得到 $h(X)$，再学习 varying coefficients；但此时 coefficient 对原始变量的直接含义会减弱，模型也逐步离开经典 VCM 的"纯粹"形态。

8.4 smoothness 可能是假结构

Kernel、spline 和 GP 通常假定相邻 modifier values 具有相似 response。若真实变化是 hard threshold、稀有 subgroup 或结构突变，平滑模型会产生 bias；此时 tree、mixture 或 change-point 模型更合适。

8.5 边界、tail 与 extrapolation 不稳定

局部模型在 modifier support 的边缘只有单侧、少量邻居。原始 VCM 的 survival example 已展示 tail coefficient 对少数 observations 很敏感。对于 dose-response，这一问题转化为 positivity：若某类人几乎不接受某个 dose，就没有足够数据支撑该处的个体曲线。

8.6 预测异质性不自动具有因果含义

观测到不同 A 与 Y 的关系随 X 变化，不代表 treatment effect 随 X 变化。Unmeasured confounding、selection、post-treatment modifiers 和 measurement error 都可能制造伪异质性。连续 treatment 还要求每个 relevant (x, a) 附近有足够 support。

8.7 coefficient 本身依赖坐标与模型设定

改变 feature scale、basis、link function 或加入相关 covariates 都可能改变局部 coefficients。因此 “unit 的 coefficient vector” 并不天然等于 “unit 的物理机制”。Coefficient curve 是模型内解释对象，而不是无条件可识别的世界属性。

8.8 新 unit 的 cold start

如果新 unit 没有 repeated outcomes，也没有信息充分的 descriptors，subject-specific curve 只能退回 population average。Neural function generator 并没有消除这个信息边界，只是把跨 unit prior 学得更灵活。

8.9 function-level evaluation 很困难

普通 prediction 只需比较 $\hat{Y}_i$ 与 Y_i。完整 response function 评估需要在多个 query points 上观察同一 unit，但现实中通常做不到；因果问题还受到 fundamental problem of causal inference 的限制。因此很多论文只能评估 factual loss、semi-synthetic integrated error、policy value 或 uncertainty coverage 的某个投影。

8.10 深度版本的样本量与校准代价

SCIGAN 明确指出其 GAN 至少需要数千个训练样本。深度 individualized-response 方法还容易出现优化不稳定、overconfident extrapolation 与 function uncertainty 未校准的问题。

9 为什么没有以 VCM 的名字成为主流机器学习?

首先，问题的前提需要修正：**核心思想已经成为主流，只是术语没有。**

VCM 中的思想	在其他社区中的名字
coefficient 随 context 变化	smooth interaction、GAM、conditional parameterization
group/subject-specific coefficients	random slopes、mixed effects、hierarchical Bayes
局部关系随空间变化	local models、GWR、spatially varying coefficients
不同区域使用不同参数	model-based/linear trees、mixture of experts；普通 RF/GBDT 是更宽松的分区类比
treatment effect 随 unit features 变化	CATE、causal forest、HTE
参数由 context 生成	Hypernetwork、conditional computation
从少量 evidence 推断一条函数	meta-learning、Neural Process

VCM 中的思想	在其他社区中的名字

树模型确实解释了 VCM 思想为何被主流 ML 吸收，但不能简单说 “随机森林就是变系数模型”。普通 regression tree 通常给每个 leaf 一个常数预测, 标准 RF/GBDT 也不显式区分 predictor X 与 modifier U，更没有可直接阅读的 $\beta_j(U)$。真正结构上接近 VCM 的是 leaf 内拟合参数模型的 model-based/linear trees，以及直接用 boosted trees 表示 coefficient functions 的 tree-boosted VCM。树方法的优势在于自动选择 split 和高维交互，而代价是平滑 coefficient curve 与显式科学语义变弱。

VCM 术语本身没有统治通用 ML，主要有六个原因。

1. **它最自然的场景是结构化科学数据。**图像和文本任务通常没有一个先验明确、低维且可平滑的 coefficient modifier。
2. **通用 ML 优化 aggregate predictive risk。** VCM 更重视 structured heterogeneity、局部效应与可解释曲线，这些不一定提升标准 leaderboard 指标。
3. **树和神经网络减少了人工结构选择。**它们自动发现分区或 representation，研究者不必预先声明哪个变量修改哪个 coefficient；但标准模型通常也不再输出显式 coefficient functions。
4. **完整 unit curve 没有可直接观察的标签。**难以形成 ImageNet 式统一 benchmark。
5. **灵活性与可解释性存在张力。**一旦用高维 encoder 或整网参数生成器替代低维 spline，方法虽更强，却不再像经典 VCM 那样容易解释。
6. **软件与社区分散。** GAM、mixed-model、spatial、causal forest、meta-learning 各自拥有工具链，没有一个统一的 “unit response function” API。

因此，VCM 没有失败；它更像一组被不同领域分别吸收的设计原则。

10 Unit-Level Response Function Learning 的更精确位置

10.1 VCM 是一个严格特例

可以把 VCM 写成一个退化的 unit-selection law：

$$q(du \mid O_i) = \delta_{\beta(Z_i)}(du),$$

$$R_i(\cdot \mid x) = R_\theta\{\cdot \mid x, U_i = \beta(Z_i)\}.$$

它具有三个限制：selection law 退化成 point mass；unit state 完全由 observed modifier 决定；response 对 explicit input 保持线性或广义线性。

10.2 Functional mixed effects 是 repeated-measure 特例

当 $\mathcal{O}_i$ 包含同一 subject 的多个 (X_{it}, Y_{it}) 时，可以推断

$$p(G_i \mid \mathcal{O}_i)$$

并通过 population prior 做 shrinkage。这比 VCM 更接近 literal subject-specific curve，但它的主要信息来源是 repeated outcomes，不是 one-shot evidence。

10.3 Causal individualized response 是带识别条件的特化

当 explicit input 是 treatment A，还必须声明：

- consistency / well-defined intervention;
- no unmeasured confounding 或随机化设计;
- positivity / overlap;
- structural invariance 与需要的 cross-world coupling;
- target 是 CATE、conditional mean curve、distributional response，还是 literal unit effect。

不满足这些条件时，学到的是 treatment-indexed predictor，而不是 causal response function。

10.4 更有辨识度的研究对象

一个比 "每个 unit 有自己的函数" 更强、也更可检验的 operational contract，首先要把 world-side outcome computation 与 learner-side prediction 分开。

第一层：world-side token selection 与 outcome generation。 Model-level U 选择 token；在声明的 context 与 time scale 内，一次 realization $U = u^\star$ 是该 token 的 fixed embedding。面对 predictors/query x 与 event noise $E = e$，实际 outcome 是

$$Y_{u^\star}(x) = f_\theta(x, e; u^\star).$$

样本编号不参与这次计算。Token–embedding injectivity 可以作为 identity-encoding assumption，但 one-shot observation 不识别 $u^\star$ 的唯一坐标，也不排除坐标重参数化。

第二层：predictors 作为 factual evidence 做 unit abduction。 Learner 看到 $X = x$ 时，首先把它记作 x^F，用来推断这个实际 token 可能对应什么 embedding。Abduction 有三种基本形式：

$$\widehat{u} = a_\phi(x^F) \qquad \text{(point)},$$

$$q_\phi(u \mid x^F) = \mathcal{N}\big(u; \mu_\phi(x^F), \Sigma_\phi(x^F)\big) \qquad \text{(Gaussian)},$$

$$q_\phi(u \mid x^F) = \prod_j \mathrm{Cauchy}\big(u_j; m_{\phi,j}(x^F), \gamma_{\phi,j}(x^F)\big) \qquad \text{(Cauchy)}.$$

Point mode 做最强的 recognition commitment；Gaussian 把 ambiguity 表达成围绕中心的 local、finite-variance uncertainty；Cauchy 有 location 与 scale，但没有 finite mean 或 variance，因而用 polynomial tails 保留远离 "典型个体" 的 candidate mechanisms。三者都描述 learner 对同一 actual $u^\star$ 的 epistemic abduction result，不是每次 sample 都创造一个新的 physical unit。Cauchy 的 full support 也不保证任何 outcome 都能由任何 unit 生成；outcome support 仍由 f_θ 与 E 决定。

第三层：candidate-conditioned generation。 对每个 candidate u，learner 调用同一个共享 generator $f_\theta(x^Q, E; u)$。为读者友好，正文把 event noise 诱导的预测写作

$$p_\theta(\cdot \mid x^Q, u) := \mathcal{L}_E\{f_\theta(x^Q, E; u)\}.$$

这不是另一个模型；p_θ 只是 f_θ 与 event noise 诱导的 outcome distribution。不同 embeddings 仍可能在全部目标 queries 上产生同一 predictive map，因此 unit identity 的可区分性与 response equivalence 是两件事。严格的 measure-theoretic kernel 记号可以在理论附录需要时恢复，但不应遮住 generator ontology。

第四层：看到 $X = x$ 后的 ordinary prediction。 Factual prediction 中，同一个数值 x 第二次以 query 身份出现，即 $x^Q = x$。Distributional abduction 的最终预测是

$$p_{\phi,\theta}(y \mid x^F, x^Q) = \int_{\mathcal{U}} p_\theta(y \mid x^Q, u) q_\phi(u \mid x^F)\, du.$$

等价的 sampling procedure 是

$$\widetilde{U} \sim q_\phi(\cdot \mid x^F), \qquad E \sim p_E, \qquad \widetilde{Y} = f_\theta(x^Q, E; \widetilde{U}).$$

Tilde 表示 computational candidate，不是 world-side identity redraw。Point mode 则直接使用 $f_\theta(x^Q, E; a_\phi(x^F))$。普通 prediction 返回 outcome distribution 或其 point summary，不返回一个 physical unit，也不必返回一条 sampled function。

第五层：same-token counterfactual linkage。若 factual predictors 是 x，查询同一 token 在新输入 x' 下的 outcome，必须固定从 $x^F = x$ 得到的 point 或 distributional abduction result：

$$p_{\phi,\theta}(y \mid x, x') = \int p_\theta(y \mid x', u) q_\phi(u \mid x)\, du.$$

不能改成 $q_\phi(u \mid x')$，因为这会用 alternative query 重新识别 token，可能把 unit 一起换掉。Fixed abduction + changed query 给出 Layer 3 的 same-token linkage；它本身仍不提供 intervention identification 或 strict singular causal attribution。

当前 tractable default 让 neural abduction operator 输出 Cauchy location/scale，并让 generator 对 x 保持 linear 或 bilinear。这是可替换的 inductive bias，不是上述 ontology 的必要条件。

在这一表述下，可以把研究贡献界定为：

> **We separate world-side token outcome generation from learner-side unit recognition, allow point, Gaussian, or Cauchy abduction from factual predictors, and preserve same-token counterfactuals by fixing factual abduction while changing only the mechanism query.**

这一定义应坚持八条边界：

1. model-level U、actual realization $u^\star$ 与 learner candidates 分开；
2. 样本/序列编号只做 bookkeeping，不充当 unit ontology；
3. token–embedding injectivity 与 one-shot coordinate identification 分开；
4. factual evidence x^F 与 mechanism query x^Q 的角色分开，即使 factual 时数值相同；
5. Cauchy location/scale 与不存在的 mean/variance 分开；
6. heavy-tailed candidate uncertainty 与 unrestricted outcome support 分开；
7. prediction adequacy 与 physical/causal mechanism identification 分开；
8. one-shot factual supervision 与 evaluator-only function truth 分开。

11 仍然开放的研究问题

11.1 One-shot 条件下究竟识别了什么?

未来理论不应把 token–embedding injectivity 的建模假设误写成 actual $u^\star$ 的数据 identification，也不应默认恢复唯一的真实 unit mechanism。更诚实的目标是描述 candidate representations、response-equivalence classes、partial identification region，或明确给出在何种 shared-family assumptions 下可以恢复 target contrasts。即使 actual unit representations 不同，它们也可能在目标 query domain 上满足 $u \sim_R v$。

11.2 如何构造可信的 function-level benchmark?

至少需要三类 benchmark:

- simulator 提供完整 unit functions，用于机制恢复与校准;
- 有 repeated interventions 或 crossover design 的真实数据，用于有限 query 验证;
- 只有 factual outcome 的大规模现实数据,用于 support-aware prediction 与 policy evaluation。

这三类证据必须分开报告，不能用 semi-synthetic curve recovery 代替真实因果验证。

11.3 如何评价 induced function-space law，而不只评价一个点?

候选指标包括 integrated log score、integrated CRPS、query-wise calibration、simultaneous coverage、finite-contrast error、policy regret，以及在 support 边界外拒绝预测的质量。

11.4 如何学习共享结构而不抹平真正异质性?

过强 pooling 会退化为 population regression; 过弱 pooling 会造成每个 unit 的函数不可学习。需要研究 hierarchical priors、prototype mixtures、low-rank function manifolds、neural stochastic processes 与 sparse mechanism libraries 之间的可识别性和样本复杂度。

11.5 如何处理动态 unit?

现实 unit 会随时间改变。未来模型需要同时推断稳定 U_i 与动态 S_{it}，并回答: 何时 trajectory 的改变来自干预，何时来自自然 state transition，何时只是 event noise。

11.6 如何把因果识别和 function generation 接起来?

Hypernetwork 可以生成函数,但不保证反事实正确;causal estimators 可以识别某些 contrasts,却未必生成完整 outcome distribution。一个成熟框架需要把 design/identification、unit abduction、mechanism evaluation、uncertainty reduction 和 falsifiable evaluation 连成一条证据链。

12 结论

围绕 unit-specific response function，学术界并非没有前人，而是已经形成一棵跨越统计学与机器学习的大树: random-effects 提供个体参数的层级推断; VCM 把 observed modifier 映射为局部 coefficient functions; functional mixed effects 把 subject-specific deviation 扩展为随机函数; GWR、functional-coefficient time series 与 model-based trees 处理空间、状态与分段异质性; causal forest、DRNet、SCIGAN 与 VCNet 学习 heterogeneous treatment/dose response; Hypernetwork 和 Neural Process 则把 function generation 推向通用高维任务。

这棵树最重要的共同原则是:

> 不要求 units 共享同一条 response function; 要求它们共享可学习的函数生成规律。

它没有形成统一范式，主要不是因为思想无效，而是因为各社区拥有不同的 unit、evidence、query、estimand 与验证方式。真正值得继续提出的统一问题，是如何在有限 unit evidence 下，把 world-side $U = u^\star$ 的 token outcome generation、由 factual predictors 得到的 point/Gaussian/Cauchy abduction、candidate-conditioned generator evaluation 与最终 outcome prediction 分开; same-token query 固定 factual abduction, 只改变 mechanism query。同时必须诚

实说明：token–embedding injectivity 是建模假设而非 one-shot identification，不同 embeddings 可以产生相同 predictive behavior，Cauchy heavy tails 不等于 unrestricted outcome support；只有额外的 design 与 identification 条件才能进一步获得 causal 或 mechanism interpretation。

参考文献

Ahmed M. Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/6a508a60aa3bf9510ea6acb021c94b48-Abstract.html.

Ioana Bica, James Jordon, and Mihaela van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/bea5955b308361a1b07bc55042e25e54-Abstract.html.

Chris Brunsdon, A. Stewart Fotheringham, and Martin E. Charlton. Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4):281–298, 1996. doi: 10.1111/j.1538-4632.1996.tb00936.x.

Zongwu Cai, Jianqing Fan, and Qiwei Yao. Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, 95(451):941–956, 2000. doi: 10.1080/01621459.2000.10474284.

Rong Chen and Ruey S. Tsay. Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, 88(421):298–308, 1993. doi: 10.1080/01621459.1993.10594322.

William S. Cleveland, Eric Grosse, and William M. Shyu. Local regression models. In John M. Chambers and Trevor J. Hastie, editors, *Statistical Models in S*, pages 309–376. Wadsworth and Brooks/Cole, 1992.

Davide Fabbrico, Matteo Pedone, and Francesco Claudio Stingo. An interpretable varying coefficients approach to non-linear regression. *Statistics and Computing*, 36(4):152, 2026. doi: 10.1007/s11222-026-10903-y.

Jianqing Fan and Wenyang Zhang. Statistical estimation in varying coefficient models. *The Annals of Statistics*, 27(5):1491–1518, 1999. doi: 10.1214/aos/1017939139.

Jianqing Fan and Wenyang Zhang. Statistical methods with varying coefficient models. *Statistics and Its Interface*, 1(1):179–195, 2008. doi: 10.4310/SII.2008.v1.n1.a15.

Jianqing Fan, Qiwei Yao, and Zongwu Cai. Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):57–80, 2003. doi: 10.1111/1467-9868.00372.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S. M. Ali Eslami. Conditional neural processes. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1704–1713, 2018. URL https://proceedings.mlr.press/v80/garnelo18a.html.

Wensheng Guo. Functional mixed effects models. *Biometrics*, 58(1):121–128, 2002. doi: 10.1111/j.0006-341X.2002.00121.x.

David Ha, Andrew Dai, and Quoc V. Le. Hypernetworks. In *International Conference on Learning Representations*, 2017. URL https://research.google/pubs/hypernetworks-2/.

Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779, 1993. doi: 10.1111/j.2517-6161.1993.tb01939.x.

Jimmy Hickey, Jonathan P. Williams, and Emily C. Hector. Transfer learning with uncertainty quantification: Random effect calibration of source to target (RECaST). *Journal of Machine Learning Research*, 25(338):1–40, 2024. URL https://www.jmlr.org/papers/v25/22-1369.html.

Donald R. Hoover, John A. Rice, Colin O. Wu, and Li-Ping Yang. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822, 1998. doi: 10.1093/biomet/85.4.809.

Jianhua Z. Huang, Colin O. Wu, and Lan Zhou. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89(1):111–128, 2002. doi: 10.1093/biomet/89.1.111.

Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38 (4):963–974, 1982. doi: 10.2307/2529876.

Lizhen Nie, Mao Ye, Qiang Liu, and Dan Nicolae. VCNet and functional targeted regularization for learning causal effects of continuous treatments. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2103.07861.

Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M. Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5612–5619, 2020. doi: 10.1609/aaai.v34i04.6014.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839.

Lan Xue and Annie Qu. Variable selection in high-dimensional varying-coefficient models with global optimality. *Journal of Machine Learning Research*, 13:1973–1998, 2012. URL https://www.jmlr.org/papers/v13/xue12a.html.

Achim Zeileis, Torsten Hothorn, and Kurt Hornik. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2):492–514, 2008. doi: 10.1198/106186008X319331.

Yichen Zhou and Giles Hooker. Decision tree boosted varying coefficient models. *Data Mining and Knowledge Discovery*, 36(6):2237–2271, 2022. doi: 10.1007/s10618-022-00863-y.