LLM-GUIDED VINE COPULA SYNTHESIS FOR TRANSPORTING TREATMENT EFFECTS FROM SMALL-SCALE STUDIES UNDER COVARIATE SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating target-population effects from small source studies is challenging because limited or imbalanced treatment arms produce high variance, while covariate shift creates extrapolation error. We propose **VinePrior**, a structured causal-augmentation framework that combines regular-vine modeling with large-language-model knowledge. To the best of our knowledge, VinePrior is the first framework to introduce LLM-derived semantic priors into pair-copula family selection for causal data augmentation, combining pretrained knowledge with statistical dependence summaries to stabilize selection in small samples. The framework has two complementary components: (i) source-domain augmentation generates covariates, treatment states, and outcomes in causal order and iteratively expands the training data until the effect-estimation confidence interval stabilizes; and (ii) **CondVine** incorporates observed target values of effect-relevant covariates and conditionally generates the remaining variables, enabling target-aligned augmentation under weaker overlap. We establish an error bound connecting generator error, outcome-learning error, and source–target overlap to target-ATE error. Across benchmark datasets, VinePrior achieves lower ATE error than ten baselines, with particularly strong gains under small samples, high dimensionality, treatment imbalance, and covariate shift.

1 INTRODUCTION

Small-scale source studies are common in disease research, pilot clinical trials, and other settings where recruitment is difficultDemirel et al. (2024); Stuart et al. (2018). In randomized studies, treatment assignment supports identification of causal effects within the enrolled sample, yet scientific and policy decisions often concern a target population with a different covariate distributionStuart et al. (2018); Ackerman et al. (2019); Ek & Zachariah (2024). A limited source sample yields high-variance estimates of treatment-specific outcome functions, while source–target covariate shift creates extrapolation error in regions that are prevalent in the target population but poorly represented in the sourceMelnychuk et al. (2026); Demirel et al. (2024). We address the problem of reliably estimating the target-population average treatment effect from such small source studies.

Under conditional mean transportability, existing approaches typically adjust for source–target distribution shift through sampling weights, calibration weighting, or outcome standardizationCole & Stuart (2010); Stuart et al. (2011); Dahabreh et al. (2019; 2020); Josey et al. (2021). When overlap is weak, sampling weights may concentrate on a small number of source units, further reducing the effective sample sizeHwang et al. (2024); outcome regression must instead predict treatment responses in target-relevant regions where source data are sparse. Both difficulties become more severe when the source sample is small. Data augmentation offers a complementary route. Rather than reallocating weight among existing observations, it draws new training samples and covariate configurations from a fitted model, thereby providing a denser training design for downstream outcome learners. The reliability of these samples, however, depends on both the original data and the fitted generator. For high-capacity deep generative models with many trainable parameters, tens or hundreds of source observations may be insufficient for reliable training and hyperparameter tuningMargeloiu et al. (2024); Hollmann et al. (2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Vine copulas provide another class of generative models. A regular vine decomposes complex multivariate dependence into a sequence of low-dimensional pair-copulas and permits different variable pairs to use different copula families, thereby representing nonlinear, asymmetric, and tail dependenceBedford & Cooke (2002); Aas et al. (2009); Czado & Nagler (2022). Compared with end-to-end deep generators, this decomposition has a more explicit statistical structure and is well suited to studying generation in small samples. However, when several copula families fit a limited sample similarly, selection based on AIC, BIC, or likelihood can be sensitive to sample perturbationsDissmann et al. (2013); Gruber & Czado (2015). This observation motivates using variable semantics as probabilistic prior information while retaining data-based parameter estimation and model comparison. We therefore propose VinePrior, a structured causal-augmentation method for small source studies. A regular vine represents the source covariate distribution, while a large language model (LLM) converts variable descriptions and dependence summaries into semantic priors over candidate pair-copula families. VinePrior thus uses task-specific language-model knowledge without allowing the LLM to replace statistical fitting.

VinePrior does not treat covariates, treatment, and outcome as an unstructured table. Instead, it follows a staged generation strategy: it first generates covariates, then constructs treatment states through a controlled allocation module, and finally generates outcomes conditional on the covariates and treatment. It uses an iterative augmentation mechanism that continues until the effect-estimation confidence interval stabilizes. For a randomized source study, the synthetic allocation need not exactly reproduce the treatment proportion and may assign more samples to a sparsely represented arm. The augmented data provide a training interface for a range of downstream causal estimators.

Our source-domain procedure is designed for settings in which the source and target covariate distributions differ, while target-relevant regions remain represented within the support of the source distribution. We find that, under conditional mean transportability, source-domain augmentation stabilizes treatment-specific outcome learning and thereby improves generalization to the target population without directly using target data to fit the generator. When full-covariate overlap deteriorates, we introduce **CondVine**, a target-aligned extension that incorporates the observed target values of effect-relevant covariates into the generation process and requires overlap only on this lower-dimensional covariate set.

Contributions. Our contributions are fourfold. **Structured causal augmentation.** VinePrior integrates regular-vine covariate generation, controlled treatment allocation, and treatment-conditional outcome generation into a modular pipeline that follows the causal data-generating order. **Semantic-prior selection.** it converts variable semantics and dependence summaries into probabilistic priors over pair-copula families, stabilizing family selection in small samples. **End-to-end error characterization.** we connect generator error and downstream outcome-learning error to target-population ATE error, while quantifying the role of source–target overlap. **Transport under weaker overlap.** CondVine anchors effect-relevant covariates at observed target values and conditionally generates the remaining covariates from the source, replacing full-covariate overlap with overlap on a lower-dimensional effect-relevant set.

2 Related Work

VinePrior operates at the intersection of causal effect generalization and data augmentation. Classical research established methods such as outcome standardization, inverse probability weighting, and doubly robust estimation Horvitz & Thompson (1952); Robins (1986); Rosenbaum & Rubin (1983); Robins et al. (2000); Bang & Robins (2005). To address the "trial-to-population" generalization problem, relevant methods have adapted these tools through techniques such as outcome modeling, selection or calibration weighting, and augmented estimation Hainmueller (2012); Cole & Stuart (2010); Stuart et al. (2011); Dahabreh et al. (2019; 2020); Josey et al. (2021); Demirel et al. (2024); Ek & Zachariah (2024); Boughdiri et al. (2026); recent studies have also addressed effect-related covariates that are not jointly observed across different environments Li et al. (2025). While these methods leverage observed source samples to identify target effects, they struggle to directly stabilize the learning of treatment-specific outcomes in small-scale studies. VinePrior complements these methods through structural augmentation, whereas CondVine—in scenarios with low covariate overlap—utilizes effect-related target covariates to guide data generation.

Generative augmentation offers an alternative pathway for learning from limited tabular data. Tabular data synthesis techniques have evolved from adversarial and variational models to language models and diffusion models Xu et al. (2019); Borisov et al. (2022); Kotelnikov et al. (2023); Zhang et al. (2024); Stoian et al. (2024); Margeloiu et al. (2024); Kim et al. (2024); Lin et al. (2025). Despite their flexibility, reliably training these models on new datasets typically requires a larger number of observations than small source studies can provide. Foundation models shift this data requirement to the pre-training phase: by training on a vast array of real or synthetic tasks, they become capable of supporting predictions based on small datasets as well as general-purpose tabular reasoning Hollmann et al. (2022; 2025); Ye et al. (2026); Zhang et al. (2025); Robertson et al. (2026).

In the realm of causal inference, existing research has explored outcome-invariant augmentation and the sequential generation of data from randomized trials Akbar et al. (2026); Petrakos et al. (2025); Melnychuk & Feuerriegel (2026). VinePrior adheres to this causal generation sequence by fitting its dependency model directly to small-scale source studies and evaluating the augmentation efficacy based on downstream target effect estimation—rather than relying solely on distributional fidelity. In the field of structured generation, Vine Copulas construct multivariate dependencies using pair-copulas, allowing different edges to capture nonlinear, asymmetric, and upper- or lower-tail dependency characteristics Bedford & Cooke (2002); Aas et al. (2009); Czado & Nagler (2022). This flexibility introduces a significant model selection challenge, which has been addressed through sequential and Bayesian methods for selecting vine structures and copula families Dissmann et al. (2013); Gruber & Czado (2015). Subsequently, Vine Copulas have been applied to data synthesis—spanning learning-based, expert-guided, and privacy-aware approaches Sun et al. (2019); Sexton et al. (2022); Griesbauer et al. (2025)—with open-source implementations available for practical use Claassen et al. (2024). VinePrior integrates these research directions: it utilizes probability distributions over copula families—derived from Large Language Models (LLMs)—as priors, while retaining data-driven parameter estimation and full model calibration. We evaluate whether this combination improves copula family selection and downstream target effect estimation performance in small-sample settings.

3 PRELIMINARIES

Notation. Let $S = 1$ denote the source study and $S = 0$ the target population. We observe an independent source sample $\mathcal{D}_S = \{(X_i, T_i, Y_i)\}_{i=1}^{n}$, i.i.d. from P_S, and a target covariate sample $\mathcal{D}_T = \{X_j^{\text{tar}}\}_{j=1}^{N}$, i.i.d. from P_T^X. Here, $X \in \mathcal{X} \subseteq \mathbb{R}^d$ denotes pretreatment covariates, $T \in \{0, 1\}$ a binary treatment, and $Y \in \mathcal{Y} \subseteq \mathbb{R}$ the observed outcome. Let $P_S^X = \mathcal{L}(X \mid S = 1)$ and $P_T^X = \mathcal{L}(X \mid S = 0)$, with expectations $\mathbb{E}_S$ and $\mathbb{E}_T$. For measurable f, define $\|f\|_{L_2(P)} = \{\mathbb{E}_{X\sim P}[f(X)^2]\}^{1/2}$.

Problem setup. We estimate the target-population ATE from a small source study and a covariate-only target sample. Let $Y(t)$ be the potential outcome under $t \in \{0, 1\}$, with target estimand $\tau_T = \mathbb{E}[Y(1) - Y(0) \mid S = 0]$. The source may be randomized or observational and need not have balanced treatment allocation. Define $e_S(x) = \mathbb{P}(T = 1 \mid X = x, S = 1)$ and $\mu_t^S(x) = \mathbb{E}[Y \mid X = x, T = t, S = 1]$. Small or imbalanced source samples make μ_0^S and μ_1^S unstable, while covariate shift requires them to generalize to target-relevant regions with limited source support. We use structured causal augmentation to stabilize this outcome learning for target-effect estimation.

Assumption 3.1 (Source causal identification)**.** *For each $t \in \{0, 1\}$, we assume consistency, $Y = Y(t)$ if $T = t$; conditional mean exchangeability, $\mathbb{E}[Y(t) \mid X, S = 1, T = t] = \mathbb{E}[Y(t) \mid X, S = 1]$; and treatment positivity, $\mathbb{P}(T = t \mid X = x, S = 1) > 0$ for every x with positive density under P_S^X Dahabreh et al. (2019); Hwang et al. (2024); Robins (1986); Rosenbaum & Rubin (1983).*

Under Assumption 3.1, the source conditional potential-outcome means are identified as $\mu_t^S(x) = \mathbb{E}[Y(t) \mid X = x, S = 1] = \mathbb{E}[Y \mid X = x, T = t, S = 1]$; see Section B.2 and proposition B.1 for the derivation and observational-source extension.

Source-supported transport. We next state the standard conditions connecting the source response functions to the target population.

Assumption 3.2 (Conditional mean transportability and overlap)**.** *For each $t \in \{0, 1\}$, the conditional potential-outcome mean is transportable, $\mathbb{E}[Y(t) \mid X = x, S = 1] = \mathbb{E}[Y(t) \mid X =$*

$x, S = 0]$*, and source participation is positive for every* x *with positive density in the target population***?***Dahabreh et al. (2019); Cole & Stuart (2010); Demirel et al. (2024); Boughdiri et al. (2026). Equivalently,* $P_T^X \ll P_S^X$.

Under Assumptions 3.1 and 3.2, the target ATE is identified as $\tau_T = \mathbb{E}_T[\mu_1^S(X) - \mu_0^S(X)]$. Under these identification conditions, the conditional potential-outcome means are invariant across populations, so differences between the source and target covariate distributions affect the target ATE only through the distribution over X. Consequently, P_S^X and P_T^X may differ provided that $P_T^X \ll P_S^X$. When it exists, let $w_X(x) = \mathrm{d}P_T^X / \mathrm{d}P_S^X(x)$ denote the corresponding density ratio.

Source augmentation and target estimation. For estimated source response functions $\hat{\mu}_t$, target-effect error satisfies

$$|\hat{\tau}_T - \tau_T| \le \mathbb{E}_T\big|\hat{\mu}_1(X) - \mu_1^S(X)\big| + \mathbb{E}_T\big|\hat{\mu}_0(X) - \mu_0^S(X)\big| .$$

Hence source-only augmentation need not approximate P_T^X to improve target estimation. A small source sample may sparsely represent low-probability source regions that carry substantial target mass. Recovering and resampling such source-supported patterns can reduce outcome-model error in these regions and thereby tighten the target-weighted bound above. VinePrior targets this finite-sample limitation without assuming that augmentation removes population shift.

Reduced-overlap transport. Full-X overlap may be unnecessarily restrictive when population differences in treatment effects are governed by a lower-dimensional set of transport variablesLi et al. (2025); Melnychuk et al. (2026). Partition $X = (Z, W)$, where Z contains the effect-relevant transport variables and W the remaining covariates, and define $\theta_S(z) = \mathbb{E}[Y(1) - Y(0) \mid Z = z, S = 1]$.

Assumption 3.3 (Effect transportability over Z)**.** *The conditional treatment effect is transportable over* Z*,* $\mathbb{E}[Y(1) - Y(0) \mid Z = z, S = 1] = \mathbb{E}[Y(1) - Y(0) \mid Z = z, S = 0]$*, and* $P_T^Z \ll P_S^Z$.

Under Assumptions 3.1 and 3.3, $\tau_T = \mathbb{E}_{Z \sim P_T^Z}[\theta_S(Z)]$. Thus, if Z is sufficient for transporting treatment-effect heterogeneity, overlap is required only over Z, rather than the full covariate vector. This motivates CondVine, which anchors observed target values of Z and conditionally generates the remaining covariates W.

Regular vine copulas. Let F_X be the joint distribution of $X = (X_1, \ldots, X_d)$ with marginals $F_1, \ldots, F_d$. By Sklar's theorem, $F_X(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d))$, where C is a copula. A regular vine represents C through a sequence of trees $\mathcal{V} = (\mathcal{T}_1, \ldots, \mathcal{T}_{d-1})$, with edge set E_m in tree $\mathcal{T}_m$. For an edge $e = (j_e, k_e; D_e)$, let $u_{e,1} = F_{j_e|D_e}(x_{j_e} \mid x_{D_e})$ and $u_{e,2} = F_{k_e|D_e}(x_{k_e} \mid x_{D_e})$, and let $c_e(\cdot, \cdot; \theta_e)$ denote its conditional pair-copula density. Under the simplifying assumption, the joint density factorizes as

$$f_X(x) = \prod_{j=1}^{d} f_j(x_j) \prod_{m=1}^{d-1} \prod_{e \in E_m} c_e(u_{e,1}, u_{e,2}; \theta_e) . \tag{1}$$

Equation (1) decomposes high-dimensional dependence into structured bivariate components, mitigating the curse of dimensionality of direct multivariate density estimation in small samplesAas et al. (2009); Czado & Nagler (2022); Sun et al. (2019). However, reliable pair-copula family selection remains challenging when only limited observations are available. We address this selection problem in Section 4.

4 VinePrior: Semantic-Guided Causal Augmentation

4.1 Vine-based causal generation.

As summarized in Figure 1, VinePrior models the source covariate law through Equation (1) and embeds it in the causal order $X \to T \to Y$. For a fitted vine g, the augmented law is $q_g^{\mathrm{aug}}(x, t, y) = q_g^X(x) a(t \mid x) q_t^Y(y \mid x)$, where q_g^X denotes the vine-based covariate distribution, $a(t \mid x)$ the controlled treatment-allocation rule, and $q_t^Y(y \mid x)$ the treatment-conditional outcome model. This factorization preserves the distinct functions of covariate synthesis, treatment assignment, and outcome generation, allowing each component to be modeled or controlled according to its causal role.

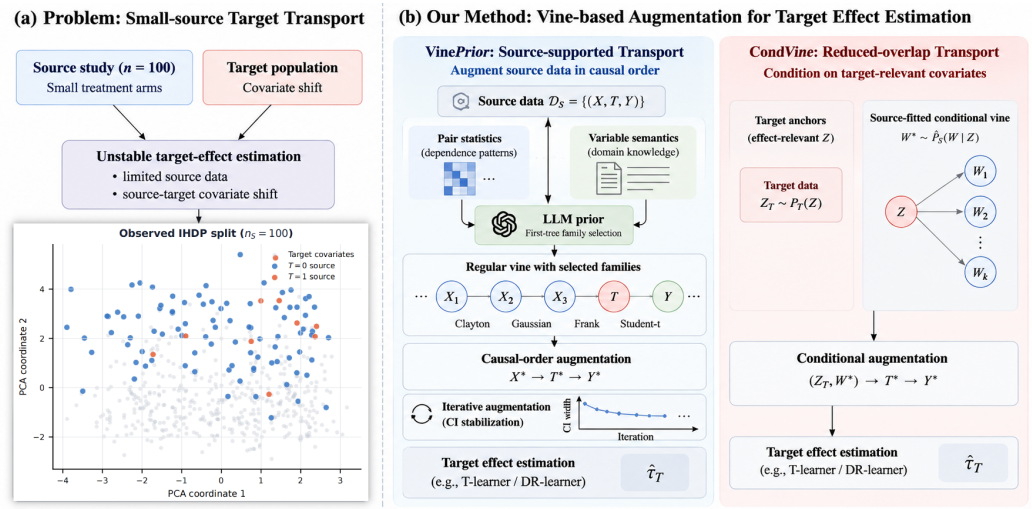


Figure 1: **Overview of the problem and our framework.** (a) Small source studies under covariate shift make target-effect estimation challenging, illustrated by the IHDP source–target split. (b) VinePrior performs semantic-guided, causal-order augmentation under source-supported shift, while CondVine uses target-aligned conditional generation under reduced overlap.

4.2 Semantic-prior-guided first-tree family selection.

VinePrior uses semantic guidance specifically for pair-copula family selection in the first vine treeDissmann et al. (2013); Czado & Nagler (2022). Continuous covariates are mapped to the copula scale through fitted marginals. For categorical covariates, empirical category probabilities define intervals on a continuous latent scale; observations are embedded within these intervals before the probability-integral transform, and generated values are mapped back through the same thresholds showed by Lemma A.1.

The first tree captures unconditional pairwise dependencies underlying the subsequent conditional trees. For each edge $e \in \mathcal{E}_1$, VinePrior constructs a prompt summary s_e containing variable descriptions and types together with dataset-specific dependence statistics, including rank correlation, Kendall's τ, and empirical upper- and lower-tail dependence. Conditioned on s_e, the LLM returns a normalized semantic prior π_e over the candidate pair-copula families $\mathcal{F}$, and VinePrior selects $\widehat{f}_e^{\mathrm{LLM}} = \arg\max_{f \in \mathcal{F}} \pi_e(f)$, where $\pi_e(f)$ is the LLM-assigned probability of family f for edge e.

The LLM affects only first-tree family selection; source observations determine the marginal distributions, copula parameters, and higher-tree components. VinePrior therefore introduces semantic and pretrained knowledge into the discrete family-selection step while retaining data-driven estimation for the remaining vine model. Beyond marginal fidelity, we evaluate dependence preservationJiang et al. (2026) using known-family recovery, held-out likelihood, and rank- and tail-dependence errors. An likelihood-gated calibration restricts selection used only as an evaluation diagnostic. Further prompt, calibration, evaluation, and fitting details are provided in the appendix.

4.3 Causal augmentation in causal order.

After fitting the covariate vine, VinePrior generates observations sequentially according to $X \to T \to Y$: $\widetilde{X}_i \sim Q_{\widehat{g}}^X$, $\widetilde{T}_i \sim a(\cdot \mid \widetilde{X}_i)$, and $\widetilde{Y}_i = \widehat{\mu}_{\widetilde{T}_i}(\widetilde{X}_i) + \widetilde{\varepsilon}_{i,\widetilde{T}_i}$. Here, $\widehat{\mu}_t$ is fitted from source observations under treatment t, and $\widetilde{\varepsilon}_{i,t}$ is drawn from the corresponding fitted residual distribution. Across covariate generators, the treatment-allocation rule, outcome and residual models, and augmentation budget are fixed so that performance differences reflect covariate generation.

Iterative stabilization. Synthetic observations are added in batches that satisfy prespecified distribution-fidelity and effect-stability criteria. Iteration stops when the effect-estimation interval width changes by less than ζ or no batch is accepted. An outer bootstrap repeats vine fitting, aug-

mentation, and outcome learning; the acceptance gates and uncertainty procedure are specified in Section A.11.

Target standardization. Let $\widehat{\mu}_t^{\mathrm{aug}}$ denote the response function fitted on the augmented source sample. VinePrior estimates $\widehat{\tau}_T^{\mathrm{VP}} = N^{-1}\sum_{j=1}^{N}\{\widehat{\mu}_1^{\mathrm{aug}}(X_j^{\mathrm{tar}}) - \widehat{\mu}_0^{\mathrm{aug}}(X_j^{\mathrm{tar}})\}$. Augmentation is used to improve source response learning, while observed target covariates determine the standardization population. Theorem 5.1 formalizes how generator and response-function errors propagate to target-ATE error.

4.4 CondVine: generative transport under reduced overlap.

CondVine extends source augmentation to target-aligned generation under Assumption 3.3. Write $X = (Z, W)$, where Z contains the effect-relevant transport variables. For each observed Z_j^{tar}, CondVine draws $\widetilde{W}_{j\ell} \sim \widehat{Q}_S(W \mid Z = Z_j^{\mathrm{tar}})$, $\ell = 1, \ldots, L$, from the conditional distribution implied by the source vine. This defines the target-aligned covariate law $\widehat{Q}_C^X(Z, W) = \widehat{P}_T^Z(Z)\widehat{Q}_S(W \mid Z)$. Target observations therefore anchor the effect-relevant coordinates, while the conditional vine generates the remaining covariates. With $\widetilde{X}_{j\ell} = (Z_j^{\mathrm{tar}}, \widetilde{W}_{j\ell})$, CondVine estimates $\widehat{\tau}_T^{\mathrm{CV}} = (NL)^{-1}\sum_{j=1}^{N}\sum_{\ell=1}^{L}\{\widehat{\mu}_1^{\mathrm{aug}}(\widetilde{X}_{j\ell}) - \widehat{\mu}_0^{\mathrm{aug}}(\widetilde{X}_{j\ell})\}$. Proposition 5.3 establishes target-ATE identification under Z-level effect transportability and overlap, rather than full-X overlap. Target covariates are used only for standardization or as CondVine anchors and do not enter the fitting of the source VinePrior generator. Validation and cross-fitting protocols are provided in the appendix. Complete pseudocode is provided in Algorithm 1.

5 Theoretical Guarantees for VinePrior and CondVine

We characterize how generator error propagates to target-effect estimation and how CondVine replaces full-covariate overlap with lower-dimensional overlap. Proofs and technical conditions are provided in the appendix.

Statistical conditions. Source and target observations are independent i.i.d. samples, and outcomes and fitted predictions are bounded in absolute value by M. For finite-sample augmentation, the generator is fitted independently of an honest outcome-learning fold, with cross-fitting used to rotate these rolesChernozhukov et al. (2018). The response learner satisfies a uniform-deviation bound for bounded squared loss.

5.1 Transport amplification

Let $\widetilde{\tau}_T = \mathbb{E}_T[\widehat{\mu}_1(X) - \widehat{\mu}_0(X)]$ and $w_X = \mathrm{d}P_T^X/\mathrm{d}P_S^X$; the following result quantifies how source response error is amplified under target shift.

Proposition 5.1 (Transport amplification)**.** *Under Assumptions 3.1 and 3.2, if* $\mathbb{E}_S[w_X(X)^2] < \infty$*, then*

$$|\widetilde{\tau}_T - \tau_T| \le \sum_{t\in\{0,1\}} \mathbb{E}_T|\widehat{\mu}_t(X) - \mu_t^S(X)| \le \sqrt{\mathbb{E}_S[w_X(X)^2]} \sum_{t\in\{0,1\}} \|\widehat{\mu}_t - \mu_t^S\|_{L_2(P_S^X)}. \quad (2)$$

Thus, Equation (2) shows that source response-learning error transfers to target-ATE error, amplified by the source–target shift factor $\mathbb{E}_S[w_X^2]$.

5.2 Generator error and augmented learning

For the following finite-sample result, consider a randomized source study, so both treatment arms share P_S^X. Let g_{LLM} be the direct selector in Section 4.2, completed by data-based parameter and higher-tree fitting. Define $\kappa_{\mathrm{LLM}} = D_{\mathrm{KL}}(P_S^X \| Q_{g_{\mathrm{LLM}}}^X)$ and, for arm t, $\xi_t = \mathbb{E}_{P_S^X} D_{\mathrm{KL}}\{P_S(Y \mid X, T = t) \| Q_{g_{\mathrm{LLM}}}(Y \mid X, T = t)\}$; the KL chain rule and Pinsker's inequality bound the arm-specific discrepancy by $\sqrt{(\kappa_{\mathrm{LLM}} + \xi_t)/2}$. Combining n_t^{out} real observations with m_t generated observations, define $\alpha_t = n_t^{\mathrm{out}}/(n_t^{\mathrm{out}} + m_t)$, $a_t = \inf_{h\in\mathcal{H}_t} \|h - \mu_t^S\|_{L_2(P_S^X)}^2$, and the bounded-loss deviation term $\varepsilon_t(n, \delta)$.

Proposition 5.2 (Augmented outcome-learning bound)**.** *Conditional on the fitted generator, simultaneously for $t \in \{0,1\}$, with probability at least $1-\delta_H$,*

$$\|\widehat{\mu}_t^{\mathrm{aug}} - \mu_t^S\|_{L_2(P_S^X)}^2 \le a_t + 8M^2(1-\alpha_t)\sqrt{\frac{\kappa_{\mathrm{LLM}}+\xi_t}{2}} + 2\varepsilon_t(n_t^{\mathrm{out}} + m_t, \delta_H/2). \tag{3}$$

The bound in Equation (3) exposes the augmentation trade-off: additional samples can reduce the finite-sample learning term, whereas generator discrepancy incurs a fidelity penalty weighted by $1-\alpha_t$. Known-family recovery, held-out likelihood, and dependence diagnostics provide empirical proxies for generator fidelity.

Theorem 5.1 (End-to-end target-ATE bound)**.** *Let $r_N(\delta_T) = 4M\sqrt{\log(2/\delta_T)/(2N)}$. Under Assumptions 3.1 and 3.2 and the conditions of Proposition 5.2, with probability at least $1-\delta_H-\delta_T$,*

$$|\widehat{\tau}_T^{\mathrm{VP}} - \tau_T| \le \sqrt{\mathbb{E}_S[w_X(X)^2]}\sum_{t=0}^{1}\left[\begin{array}{c} a_t + 8M^2(1-\alpha_t)\sqrt{\frac{\kappa_{\mathrm{LLM}}+\xi_t}{2}} \\ + 2\varepsilon_t(n_t^{\mathrm{out}} + m_t, \delta_H/2)\end{array}\right]^{1/2} + r_N(\delta_T). \tag{4}$$

Theorem 5.1 and Equation (4) link generator fidelity, augmented response learning, and source–target shift in a single target-ATE bound.

Semantic-prior calibration. For a finite complete-vine shortlist $\mathcal{G}$, let $R_X(g)$ and $\widehat{R}_X(g)$ be the population and independent validation risks under a loss bounded in $[L_c, U_c]$, with $B_c = U_c - L_c$. Given a fixed safeguarded semantic prior Π_ρ, let $\lambda_g = \log(1/\Pi_\rho(g)) + \log(2/\delta_G)$, $c_G(g;\delta_G) = B_c\sqrt{\lambda_g/(2n_{\mathrm{val}})}$, and $\widehat{g}_{\mathrm{cal}} = \arg\min_g\{\widehat{R}_X(g) + c_G(g;\delta_G)\}$.

Theorem 5.2 (Semantic-prior oracle inequality)**.** *Conditional on the fitting fold, with probability at least $1-\delta_G$ over the independent validation fold,*

$$R_X(\widehat{g}_{\mathrm{cal}}) \le \inf_{g\in\mathcal{G}}\left\{R_X(g) + 2c_G(g;\delta_G)\right\}. \tag{5}$$

The prior-dependent penalty in Equation (5) favors semantically supported candidates while preserving a finite-class safeguard. This theorem applies to the optional calibrated variant; the direct VinePrior selector remains unchanged.

5.3 CondVine under reduced overlap

Write $\theta_S(z) = \mathbb{E}_{W\sim P_S(W|z)}[\mu_1^S(z,W) - \mu_0^S(z,W)]$; the following proposition gives the reduced-overlap identification and contraction results.

Proposition 5.3 (Reduced-overlap transport)**.** *Under Assumptions 3.1 and 3.3, $\tau_T = \mathbb{E}_{P_T^Z}[\theta_S(Z)]$. For $d = \widehat{\theta} - \theta_S$ and $w_Z = \mathrm{d}P_T^Z/\mathrm{d}P_S^Z$, if $\mathbb{E}_{P_S^Z}[w_Z^2] < \infty$, then $|\mathbb{E}_{P_T^Z} d(Z)|$ is bounded by $\{\mathbb{E}_{P_S^Z}[w_Z^2]\}^{1/2}\|d\|_{L_2(P_S^Z)}$. Whenever w_X also exists, $\chi^2(P_T^Z\|P_S^Z) \le \chi^2(P_T^X\|P_S^X)$.*

Hence CondVine identifies the target ATE under Z-level effect transportability with an overlap factor no larger than its full-X counterpart. The appendix further separates CondVine error into response-learning and conditional-generation components. With one conditional draw per target anchor, empirical target standardization contributes the usual $N^{-1/2}$ sampling term.

6 Experiments

Baselines. We compare the unaugmented source estimator (No aug.), causal transport baselines (IPSW and calibration weighting)Cole & Stuart (2010); Josey et al. (2021), statistical generators (SMOTE, KDE, and Gaussian copula)Chawla et al. (2002); Botev et al. (2010); Liu et al. (2009), and deep tabular generators (CTGAN, TVAE, and TabDDPM)Xu et al. (2019); Kotelnikov et al. (2023). Vine+AIC replaces semantic family selection with AIC while retaining the same vine-generation pipelineDissmann et al. (2013), providing the primary structural ablation. Target-X standardizes source-fitted response functions over observed target covariates without synthetic augmentation. All generators share the same treatment, outcome, augmentation, and downstream-estimation settings.

Table 1: **Target-domain ATE estimation error.** (a) IHDP: $n = 100$, $d = 25$, $s = 2.0$, four estimators and four fixed seeds. (b) Controlled $d = 20$ simulations: VinePrior under source-supported shift (Simulation, $s = 2.0$) and CondVine under reduced overlap (Sim-Shift, $s = 3.0$), using T-learner RF and ten paired seeds. Values are mean ± standard deviation of absolute target-ATE error. Lower is better; bold values indicate the best result in each column, while bold method names denote our methods. Gau-Copula denotes Gaussian copula.

(a) IHDP

Method	S-Lin	S-RF	T-RF	DR-RF
No aug.	0.335±0.124	0.386±0.073	0.363±0.162	0.442±0.153
IPSW	0.381±0.148	0.511±0.382	0.511±0.369	0.457±0.355
CW	1.059±1.831	1.482±1.586	1.530±1.615	1.245±1.463
SMOTE	0.370±0.103	0.392±0.169	0.394±0.184	0.485±0.223
KDE	1.537±0.246	0.742±0.300	0.911±0.213	0.638±0.525
Gau-Copula	1.554±0.158	1.812±0.338	1.759±0.313	1.634±0.407
TVAE	2.003±2.072	2.099±1.917	1.693±1.535	1.397±0.943
CTGAN	3.314±0.319	3.136±0.323	2.056±0.397	1.397±0.687
TabDDPM	99.4±84.7	0.766±0.260	184.3±223.4	82.9±108.2
Target-X	0.473±0.610	0.390±0.628	0.385±0.633	0.468±0.551
VinePrior	**0.190±0.112**	**0.066±0.056**	**0.059±0.051**	**0.169±0.188**

(b) Controlled simulations

Method	Sim.	Sim-Shift
No aug.	1.028±0.560	0.955±0.400
IPSW	1.328±0.900	1.387±1.042
CW	1.819±1.170	1.122±0.633
SMOTE	1.013±0.592	0.972±0.439
KDE	1.031±0.575	0.960±0.412
Gau-Copula	1.008±0.580	0.963±0.411
CTGAN	1.032±0.579	0.891±0.374
TVAE	0.979±0.513	0.933±0.382
TabDDPM	0.914±0.673	0.987±0.481
Target-X	1.027±0.558	0.954±0.399
VinePrior	**0.217±0.180**	0.928±0.429
CondVine	0.757±0.504	**0.633±0.293**

Table 2: **Iterative CI stabilization.**

Seed	Iters	CI	Red.
42	3	12.10→1.50	87.6%
1024	3	6.96→0.91	87.0%
456	3	6.22→0.68	89.1%
789	2	4.36→1.88	56.9%
Mean	**2.75**	**7.41→1.24**	**80.1%**

Table 3: **Fidelity diagnostics versus target-ATE error.**

Method	X-SMD	Corr. MSE	ATE Error
Gau-Copula	**0.037**	**0.006**	1.812
KDE	0.049	0.017	0.742
SMOTE	0.193	0.014	0.392
IPSW	0.230	0.015	0.511
CTGAN	0.198	0.029	3.136
VinePrior	0.131	0.018	**0.066**

Table 4: **VinePrior versus random.**

Seed	VinePrior	Random
42	**0.030**	0.334±0.013
123	**0.007**	0.472±0.014
456	0.122	**0.062±0.011**
789	**0.079**	0.470±0.014
Mean	**0.059**	0.335

Table 5: **Semantic prior versus AIC.**

Selection	S-RF	T-RF	DR-RF
Vine+AIC	0.369±0.225	0.371±0.265	0.420±0.163
VinePrior	**0.066±0.056**	**0.059±0.051**	**0.169±0.188**
Reduction	**82.1%**	**84.1%**	**59.8%**

Table 6: **Sensitivity to prompt design.**

Prompt	S-RF	T-RF	DR-RF
5-family	0.240±0.212	0.249±0.215	0.338±0.278
15-family	**0.209±0.241**	**0.216±0.229**	**0.298±0.195**
Reduction	**12.9%**	**13.3%**	**11.8%**

Experimental setup and datasets. We study target-ATE estimation from a source sample of $n = 100$. IHDP provides a semi-synthetic causal benchmark with $d = 25$ covariates, substantial treatment imbalance, and known ground truth. We construct source and target populations to study target-ATE estimation under covariate shift. We additionally construct two controlled $d = 20$ simulated randomized controlled trial (RCT) settings with balanced treatment assignment that instantiate source-supported and reduced-overlap regimes: Simulation retains source support in target-relevant regions, whereas Sim-Shift weakens full-covariate overlap while preserving the prespecified three-variable effect set. The primary metric is absolute target-ATE error. All methods within a simulation column use the same ten paired seeds, a T-learner random forest with 300 trees, and the same augmentation budget. We also evaluate S-Lin, S-RF, and DR-RF. We report the mean and standard deviation of absolute target-ATE error. Full data-generating processes (DGPs), hyperparameters, and per-seed results are provided in Appendix Section C.

6.1 Main Results

Target-effect accuracy across overlap regimes. Table 1 evaluates structured augmentation across overlap regimes. VinePrior achieves the lowest error across all four IHDP estimators and performs best in the source-supported simulation. Under reduced overlap, target alignment provides the larger improvement, while under stronger shift conditional generation becomes the best-performing alignment rule (Table 17). Thus, VinePrior and CondVine address complementary sources of target-effect error.

Iterative confidence-interval stabilization. Across four IHDP seeds, iterative augmentation reduces bootstrap CI width by 80.1% on average and stabilizes within 2–3 accepted rounds. Most uncertainty reduction occurs early, with diminishing gains thereafter. Together with the source-

fidelity gate, this supports CI stabilization as a data-dependent stopping rule rather than unrestricted synthetic expansion.

6.2 Ablation Studies

LLM prior versus AIC. Replacing AIC with semantic first-tree selection while holding the remaining pipeline fixed consistently improves IHDP performance across RF-based estimators. The gain is largest in the small-sample regime and diminishes as the source sample grows, indicating that semantic priors are most useful when likelihood-based family selection is data limited. The gain also persists across multiple LLM backbones; see Table 21.

LLM prior versus random families. Semantic selection outperforms repeated random family selection on three of four paired IHDP seeds and avoids the severe failures observed across random runs. Since choosing the best random realization would require the unknown target effect, the semantic prior provides a single-shot, leakage-free alternative rather than benefiting from favorable random selection.

6.3 Sensitivity Analysis

Source sample size. VinePrior improves over Vine+AIC at $n = 50$ and 100, whereas the advantage disappears and can reverse for $n \geq 200$ (Table 22). This localizes the benefit of semantic guidance to the intended small-source regime, where likelihood-based family selection is less stable.

Shift strength. VinePrior's advantage over Vine+AIC is largest at moderate shift ($s = 2.0$); gains are smaller under weak shift and less stable under severe shift (Table 23). This identifies an intermediate regime where source augmentation is most useful; as source support deteriorates, target alignment through CondVine becomes more appropriate.

Marginal versus joint fidelity. Low-order fidelity does not necessarily predict causal utility. Table 3 shows that Gaussian copula achieves the lowest SMD (0.037) and correlation MSE (0.006) but an S-RF target-ATE error of 1.812, whereas VinePrior obtains 0.066 despite less favorable low-order diagnostics. Thus, matching marginal and pairwise statistics alone is insufficient; dependence relevant to treatment and outcome learning is more consequential for target-effect estimation.

Prompt design. A richer 15-family semantic configuration reduces 10-seed S-RF, T-RF, and DR-RF errors by 12.9%, 13.3%, and 11.8%, respectively, relative to the compact five-family design (Table 6). These consistent but moderate gains suggest that the main benefit comes from semantic guidance rather than a narrowly tuned prompt.

LLM backbone robustness. Across Kimi-2.6, Qwen-3.5, Qwen Plus, and DeepSeek V3, semantic selection reduces DR-RF target-ATE error by 78.3%–95.7% relative to the corresponding same-run Vine+AIC baseline (Table 21). The improvement across all four models indicates that the semantic-selection gain is not specific to a single LLM backbone.

7 Conclusion

We introduced VinePrior and CondVine for target-effect estimation from small source studies under covariate shift. VinePrior uses semantic priors to guide vine-copula family selection and augments the source data in causal order, while CondVine extends this idea to reduced-overlap settings through target-aligned conditional generation. Our theoretical and empirical results show that the two methods address complementary regimes: VinePrior improves target-effect estimation when target-relevant regions remain supported by the source, whereas CondVine remains effective as full-covariate overlap weakens. Together, these results demonstrate the value of structured, dependence-aware augmentation for causal generalization from limited source data. Our current framework assumes sufficient measured covariates for source causal identification and focuses on population-average treatment effects. CondVine also requires the effect-relevant transport variables to be specified in advance. Future work could relax these requirements through data-driven selection of transport variables and augmentation strategies, and extend semantic guidance beyond pair-copula family selection to richer dependence models and heterogeneous treatment effects.

AI USE STATEMENT

Generative AI tools were used to assist with literature discovery, research brainstorming, code development, and language editing. AI-assisted research ideas were checked against the relevant literature, generated code was reviewed and tested by the authors, and all AI-assisted text was manually reviewed and revised. Separately, LLMs are an explicit component of VinePrior for semantic pair-copula family selection, as described in the main text. The authors take full responsibility for the final content, claims, code, and experimental results of this work.

ETHICS STATEMENT

This work introduces a methodological framework for causal effect estimation and does not involve the collection of new human-subject data. Experiments use established benchmark data and controlled simulations. We are not aware of additional ethical concerns beyond those associated with the responsible use of causal inference and generative AI methods.

REPRODUCIBILITY STATEMENT

We provide the assumptions, theoretical results, and methodological details of VinePrior and CondVine in the main text and appendix. The appendix further provides complete data-generating processes, implementation and hyperparameter details, prompt configurations, evaluation protocols, additional ablations and sensitivity analyses, per-seed experimental results, and proofs of the theoretical claims. These materials are intended to support independent reproduction of the reported results.

REFERENCES

Kjersti Aas, Claudia Czado, Arnoldo Frigessi, and Henrik Bakken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics*, 44(2):182–198, 2009.

Benjamin Ackerman, Ian Schmid, Kara E Rudolph, Marissa J Seamans, Ryoko Susukida, Ramin Mojtabai, and Elizabeth A Stuart. Implementing statistical methods for generalizing randomized trial findings to a target population. *Addictive behaviors*, 94:124–132, 2019.

Uzair Akbar, Niki Kilbertus, Hao Shen, Krikamol Muandet, and Bo Dai. An analysis of causal effect estimation using outcome invariant data augmentation. *Advances in Neural Information Processing Systems*, 38:98456–98497, 2026.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Tim Bedford and Roger M Cooke. Vines–a new graphical model for dependent random variables. *The Annals of statistics*, 30(4):1031–1068, 2002.

Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*, 2022.

Zdravko I Botev, Joseph F Grotowski, and Dirk P Kroese. Kernel density estimation via diffusion. 2010.

Ahmed Boughdiri, Clément Berenfeld, Julie Josse, and Erwan Scornet. A unified framework for the transportability of population-level causal measures. *Advances in Neural Information Processing Systems*, 38:84220–84264, 2026.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Judith N Claassen, Elco E Koks, Marleen C de Ruiter, Philip J Ward, and Wiebke S Jäger. Vinecopulas: an open-source python package for vine copula modelling. *Journal of Open Source Software*, 9(101):6728, 2024.

Stephen R Cole and Elizabeth A Stuart. Generalizing evidence from randomized clinical trials to target populations: the actg 320 trial. *American journal of epidemiology*, 172(1):107–115, 2010.

Claudia Czado and Thomas Nagler. Vine copula based modeling. *Annual Review of Statistics and Its Application*, 9:453–477, 2022.

Issa J Dahabreh, Sarah E Robertson, Eric J Tchetgen, Elizabeth A Stuart, and Miguel A Hernán. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694, 2019.

Issa J Dahabreh, Sarah E Robertson, Jon A Steingrimsson, Elizabeth A Stuart, and Miguel A Hernan. Extending inferences from a randomized trial to a new target population. *Statistics in medicine*, 39(14):1999–2014, 2020.

Ilker Demirel, Ahmed Alaa, Anthony Philippakis, and David Sontag. Prediction-powered generalization of causal inferences. *arXiv preprint arXiv:2406.02873*, 2024.

Jeffrey Dissmann, Eike C Brechmann, Claudia Czado, and Dorota Kurowicka. Selecting and estimating regular vine copulae and application to financial returns. *Computational statistics & data analysis*, 59:52–69, 2013.

Sofia Ek and Dave Zachariah. Externally valid policy evaluation from randomized trials using additional observational data. *Advances in Neural Information Processing Systems*, 37:69787–69813, 2024.

Elisabeth Griesbauer, Claudia Czado, Arnoldo Frigessi, and Ingrid Hobæk Haff. Tvinesynth: A truncated c-vine copula generator of synthetic tabular data to balance privacy and utility. *arXiv preprint arXiv:2503.15972*, 2025.

Lutz Gruber and Claudia Czado. Sequential bayesian model selection of regular vine copulas. 2015.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 20(1):25–46, 2012.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Inwoo Hwang, Yesong Choe, Yeahoon Kwon, and Sanghack Lee. On positivity condition for causal inference. In *Forty-first International Conference on Machine Learning*, 2024.

Xiangjian Jiang, Nikola Simidjievski, and Mateja Jamnik. Tabstruct: Measuring structural fidelity of tabular data. In *International Conference on Learning Representations*, volume 2026, pp. 41147–41209, 2026.

Kevin P Josey, Seth A Berkowitz, Debashis Ghosh, and Sridharan Raghavan. Transporting experimental results with entropy balancing. *Statistics in medicine*, 40(19):4310–4326, 2021.

Jinhee Kim, Taesung Kim, and Jaegul Choo. Epic: Effective prompting for imbalanced-class data synthesis in tabular data classification via large language models. *Advances in Neural Information Processing Systems*, 37:31504–31542, 2024.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International conference on machine learning*, pp. 17564–17579. PMLR, 2023.

Baohong Li, Yingrong Wang, Anpeng Wu, Ming Ma, Ruoxuan Xiong, and Kun Kuang. Generalizing causal effects from randomized controlled trials to target populations across diverse environments. In *Forty-second International Conference on Machine Learning*, 2025.

Xiaofeng Lin, Chenheng Xu, Matthew Yang, and Guang Cheng. Ctsyn: A foundation model for cross tabular data generation. In *International Conference on Learning Representations*, volume 2025, pp. 47938–47959, 2025.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(10), 2009.

Andrei Margeloiu, Xiangjian Jiang, Nikola Simidjievski, and Mateja Jamnik. Tabebm: A tabular data augmentation method with distinct class-specific energy-based models. *Advances in Neural Information Processing Systems*, 37:72094–72144, 2024.

Valentyn Melnychuk and Stefan Feuerriegel. Gdr-learners: Orthogonal learning of generative models for potential outcomes. In *International Conference on Learning Representations*, volume 2026, pp. 105267–105301, 2026.

Valentyn Melnychuk, Dennis Frauen, Jonas Schweisthal, and Stefan Feuerriegel. Overlap-adaptive regularization for conditional average treatment effect estimation. In *International Conference on Learning Representations*, volume 2026, pp. 53748–53788, 2026.

Niki Z Petrakos, Erica EM Moodie, and Nicolas Savy. A framework for generating realistic synthetic tabular data in a randomized controlled trial setting. *Statistics in Medicine*, 44(18-19):e70227, 2025.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-pfn: In-context learning for causal effect estimation. *Advances in Neural Information Processing Systems*, 38:174811–174848, 2026.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7 (9-12):1393–1512, 1986.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Jean-Thomas Sexton, Michael Morin, and Jonathan Gaudreault. Vine copula based data generation for machine learning with an application to industrial processes. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*, 2022.

Mihaela CÄ Stoian, Salijona Dyrmishi, Maxime Cordy, Thomas Lukasiewicz, and Eleonora Giunchiglia. How realistic is your synthetic data? constraining deep generative models for tabular data. *arXiv preprint arXiv:2402.04823*, 2024.

Elizabeth A Stuart, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 174(2):369–386, 2011.

Elizabeth A Stuart, Benjamin Ackerman, and Daniel Westreich. Generalizability of randomized trial results to target populations: design and analysis possibilities. *Research on social work practice*, 28(5):532–537, 2018.

Yi Sun, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Learning vine copula models for synthetic data generation. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pp. 5049–5057, 2019.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

Han-Jia Ye, Si-Yang Liu, and Wei-Lun Harry Chao. A closer look at tabpfn v2: Understanding its strengths and extending its capabilities. *Advances in Neural Information Processing Systems*, 38: 135605–135637, 2026.

Hengrui Zhang, Jiani Zhang, Zhengyuan Shen, Balasubramaniam Srinivasan, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space. In *International Conference on Learning Representations*, volume 2024, pp. 52829–52857, 2024.

Xingxuan Zhang, Gang Ren, Han Yu, Hao Yuan, Hui Wang, Jiansheng Li, Jiayun Wu, Lang Mo, Li Mao, Mingchao Hao, et al. Limix: Unleashing structured-data modeling capability for generalist intelligence. *arXiv preprint arXiv:2509.03505*, 2025.

A Supplementary Method Details

A.1 Complete VinePrior and CondVine procedure

Algorithm 1 VinePrior and CondVine

Require: Source data $\mathcal{D}_S$, target covariates $\mathcal{D}_T$, variable descriptions, candidate families $\mathcal{F}$, allocation rule, and augmentation budget
Ensure: Target-ATE estimate $\widehat{\tau}_T$
1: Fit marginal transforms, vine structure, and dependence summaries
2: Use LLM priors to select first-tree families; fit all copula parameters
3: Fit treatment-specific outcome and residual models from $\mathcal{D}_S$
4: **repeat**
5: Generate covariates using VinePrior or target-anchored CondVine
6: Generate treatment and outcome; retain the batch if acceptance criteria hold
7: **until** the interval stabilizes or the augmentation budget is reached
8: Fit the downstream learner, standardize over the target population, and **return** $\widehat{\tau}_T$

A.2 Likelihood-Gated Semantic Calibration

The primary VinePrior rule selects the first-tree family with the largest semantic-prior probability. We additionally consider a likelihood-gated variant that allows the data to exclude clearly inferior candidates before the semantic prior resolves the remaining ambiguity. For each first-tree edge e, candidate-family parameters are estimated on a fitting fold, and $\ell_e(f)$ denotes the average log-likelihood of family $f \in \mathcal{F}$ on a disjoint validation fold. Given a tolerance $\delta \geq 0$, define

$$\mathcal{F}_e(\delta) = \left\{ f \in \mathcal{F} : \ell_e(f) \geq \max_{f' \in \mathcal{F}} \ell_e(f') - \delta \right\}. \tag{6}$$

The calibrated family is then

$$\widehat{f}_e^{\,\mathrm{cal}} = \arg\max_{f \in \mathcal{F}_e(\delta)} \pi_e(f). \tag{7}$$

Thus, the held-out likelihood determines the statistically competitive set, whereas the semantic prior selects among candidates that fit the data similarly. At $\delta = 0$, the rule reduces to held-out likelihood selection up to ties; increasing δ gives the semantic prior greater scope to resolve family ambiguity. Dependence summaries used in the prompt are computed only from the fitting fold. After family selection, the chosen model is refitted on the complete source sample for generation.

Table 7: Candidate pair-copula families supplied to the semantic selector. The rotation is part of the family label and is selected before numerical parameter fitting.

Family group	Included labels	Dependence represented	Tail behavior
Elliptical	Gaussian, Student	symmetric positive or negative	none; both tails
Frank	Frank	symmetric positive or negative	none
Gumbel	Gumbel0/90/180/270	asymmetric, rotation-specific	upper/lower or mirrored
Clayton	Clayton0/90/180/270	asymmetric, rotation-specific	lower/upper or mirrored
Joe	Joe0/90/180/270	strong asymmetric dependence	upper/lower or mirrored
Null	Independence	negligible rank dependence	none

A.3 Validation and Refitting Protocol

In known-family simulations, the validation sample is drawn independently from the source data-generating process while the fitting sample remains at the designated small size. In semi-synthetic and real-data analyses, repeated cross-fitting rotates the fitting and validation roles and reports out-of-fold likelihood. Results from this calibrated variant are reported separately from the direct VinePrior selector.

A.4 Prompt and Candidate-Family Construction

For each first-tree edge, the prompt contains the two variable names, descriptions and types; fitted marginal families; Pearson, Spearman, and Kendall dependence; and marginal skewness and kurtosis as tail-shape summaries. It contains no row-level observations. The rich candidate set comprises Gaussian, Frank, Student- t, the four rotations of Clayton, Gumbel, and Joe, and independence. The response schema requires a nonnegative probability for every candidate and a short rationale. Family aliases are normalized, the probabilities are renormalized to sum to one, and the highest-probability valid family is passed to source-data parameter fitting. Independence is represented by the Gaussian copula at zero correlation. Prompts, parsed probability vectors, and final edge specifications are cached for exact reruns.

A.5 Exact Semantic-Prior Prompt

The following is the substantive prompt template used by the rich selector; bracketed fields are filled separately for every first-tree edge. Line wrapping is added only for presentation. No target outcomes, target treatment labels, or row-level source observations are supplied.

```
SYSTEM ROLE
You are a copula-family selection expert for the first tree of a
regular vine under a small-sample setting (n <= 200).

TASK
Assign probabilities to the allowed pair-copula families for the
dependence between two variables. Use both variable semantics and the
reported source statistics. Concentrate probability on one to three
plausible families and return valid JSON only.

ALLOWED FAMILIES
Gaussian, Student, Frank,
Gumbel0, Gumbel90, Gumbel180, Gumbel270,
Clayton0, Clayton90, Clayton180, Clayton270,
Joe0, Joe90, Joe180, Joe270, Independence.

EDGE CONTEXT
dataset: [dataset description]
var1: [name], [type], [description]
var2: [name], [type], [description]
marginals: [fitted marginal 1], [fitted marginal 2]

SOURCE SUMMARIES
```

```
Kendall tau: [value]     Spearman rho: [value]
Pearson r: [value]       skewness: [value 1], [value 2]
kurtosis: [value 1], [value 2]

SELECTION GUIDANCE
Use Independence only for negligible rank dependence and no semantic
reason for association. Use Student for symmetric heavy-tailed
dependence; Frank or Gaussian for symmetric light-tailed dependence;
and rotated Gumbel, Clayton, or Joe families when sign and asymmetric
tail behavior support them. Respect variable types and the sign of tau.

OUTPUT JSON
{"family_probs": {"Gaussian": p_1, ..., "Joe270": p_15,
 "Student": p_16, "Independence": p_17},
 "top_family": "family with largest probability",
 "reason": "one sentence tied to semantics or a supplied statistic"}
```

The direct selector normalizes nonnegative returned masses and chooses the largest-probability family. In the single-family schema used by some backbone replications, the same prompt returns `copula_name` and `reason`; all remaining fitting steps are unchanged. The compact five-family prompt used in Table 6 restricts the catalog but preserves the same edge context and JSON-only response rule.

A.6 Response Validation, Caching, and Failure Handling

The parser maps capitalization, hyphenation, and common aliases to the canonical labels in Table 7. Invalid or negative probabilities are removed, omitted valid families receive zero mass, and the remaining vector is renormalized. A zero-mass or unparsable response invokes a prespecified source-only fallback rather than using target information. The selected family, rationale, normalized probability vector, model name, prompt version, and edge identifiers are cached. Cached responses make repeated downstream-estimator comparisons deterministic and prevent API variation from becoming an uncontrolled experimental factor. The matched Qwen Plus and DeepSeek V3 runs in Table 21 returned valid responses for all first-tree edges and therefore used no fallback selections.

Information boundary. For an edge e, the model receives s_e, which contains variable descriptions, types, fitted marginal names, and aggregate source summaries. It does not receive individual records, outcome labels, the true ATE, or any evaluation metric. Once $\widehat{f}_e$ is chosen, its parameter is estimated only from source pseudo-observations. Target covariates enter after source generator fitting, either in final standardization or as the observed Z-anchors of CondVine.

A.7 Vine Structure and Parameter Fitting

Marginal models are fitted coordinate-wise and converted into pseudo-observations. The first tree is constructed from empirical absolute Kendall dependence subject to the regular-vine proximity condition. Semantic selection changes the family attached to each first-tree edge, not the edge set. Given a selected family, its numerical parameter is estimated from the source pseudo-observations. Conditional pseudo-observations for subsequent trees are obtained recursively through fitted pair-copula h-functions; higher-tree families and parameters are then fitted by the common statistical procedure used by Vine+AIC and VinePrior. This controlled difference makes Table 5 a family-selection ablation rather than a comparison of unrelated generators.

The simplified-vine assumption treats the parameter of $c_{j_e,k_e|D_e}$ as constant over the realized value of X_{D_e}. This yields Equation (1) and tractable recursive simulation. Parameter admissibility is enforced by the family-specific fitting routine. Independence is represented by a zero-correlation Gaussian component so it can use the common h-function interface.

A.8 Categorical Latent Representation

For a categorical variable with ordered categories $c_1, \ldots, c_K$, let $\widehat{p}_k$ be its empirical source probability and define thresholds $b_k = \Phi^{-1}(\sum_{r \le k} \widehat{p}_r)$, $k = 1, \ldots, K-1$. An observation in category c_k is embedded in the corresponding latent interval $(b_{k-1}, b_k]$, with finite tail truncation used only to generate a numerical representative. The latent value then enters the same probability-integral transform and vine fit as a continuous coordinate. Generated latent values are mapped back by the stored thresholds, preserving the original categorical support. Cumulative probabilities are clipped away from zero and one before applying Φ^{-1}; a constant category is stored as a degenerate coordinate and mapped back without threshold estimation.

Lemma A.1 (Recovery of fitted category probabilities)**.** *Let $a_k = \sum_{r \le k} \widehat{p}_r$, with $a_0 = 0$ and $a_K = 1$, and decode a latent draw U^* as c_k when $U^* \in (a_{k-1}, a_k]$. If $U^* \sim \mathrm{Unif}(0,1)$, then $\Pr\{C^* = c_k\} = \widehat{p}_k$ for every k. Equivalently, if $Z^* \sim N(0,1)$ and the thresholds are $b_k = \Phi^{-1}(a_k)$, threshold decoding recovers the fitted categorical distribution exactly.*

Proof. The decoding interval for category c_k has uniform probability $a_k - a_{k-1} = \widehat{p}_k$. Since $\Phi(Z^*)$ is uniform, the latent-normal statement follows by the probability-integral transform. More generally, if the generated latent CDF is G, then the category-probability error is at most $2 \sup_z |G(z) - \Phi(z)|$. Thus any departure from the fitted category frequencies is controlled directly by marginal latent-scale error rather than by the vine dependence model.

A.9 Treatment and Outcome Generation

After drawing $\widetilde{X}$, treatment is sampled from the fixed allocation rule $a(t \mid \widetilde{X})$. In the controlled randomized experiments this rule is Bernoulli with probability one half. For observational sources, the corresponding source treatment mechanism is estimated using the same source fold available to all covariate generators. Treatment-specific response functions $\widehat{\mu}_0$ and $\widehat{\mu}_1$ are trained on source observations, and residuals are stored separately by arm. A synthetic response is generated as $\widetilde{Y} = \widehat{\mu}_{\widetilde{T}}(\widetilde{X}) + \widetilde{\varepsilon}_{\widetilde{T}}$, with the residual drawn from the matching source-arm residual distribution. This preserves the causal order $X \to T \to Y$ and prevents the covariate generator from conditioning on post-treatment information.

All comparisons in Table 1 hold a, the response learner, the residual procedure, and the synthetic budget fixed. Hence Vine+AIC and VinePrior differ only in first-tree family selection, while the statistical and deep generators differ only in the law used to draw $\widetilde{X}$. Generated rows are concatenated with the observed source sample before fitting the downstream S-, T-, or DR-learner.

A.10 Target Standardization and Cross-Fitting

For VinePrior, the learned contrast is standardized over the observed target covariates, $N^{-1} \sum_j \{\widehat{\mu}_1^{\mathrm{aug}}(X_j^{\mathrm{tar}}) - \widehat{\mu}_0^{\mathrm{aug}}(X_j^{\mathrm{tar}})\}$. For CondVine, X_j^{tar} is replaced by the target-aligned completion $(Z_j^{\mathrm{tar}}, \widetilde{W}_j)$. The target outcome is never used for generator fitting, model selection, batch acceptance, or early stopping.

The theoretical sample split separates generator fitting from downstream outcome learning. In a K-fold implementation, fold k fits the generator and nuisance functions without the units whose augmented risk is evaluated in that fold; the roles rotate and fold-specific target contrasts are averaged. The final descriptive generator may then be refitted on the complete source sample after all model choices are fixed. This is the cross-fitted analogue of the conditional-on-the-generator analysis in Proposition 5.2.

A.11 Adaptive Stopping and Outer-Bootstrap Calibration

At iteration r, five candidate batches are generated from the current training sample. A batch is admissible when its mean source-covariate SMD is below 0.2, its bootstrap ATE interval is narrower than at iteration $r-1$, and its ATE estimate remains within the prespecified stability tolerance of the original-source estimate. Among admissible batches, the one with the narrowest interval is added. Iteration stops when no batch is admissible, after five iterations, or when the relative interval-

width reduction is below 0.02. The first iteration uses semantic family selection and subsequent iterations use AIC after the training design has expanded. Target outcomes and the true target ATE are excluded from every acceptance decision. For final uncertainty, the outer bootstrap resamples original source units and repeats vine fitting, batch generation, acceptance, and downstream outcome learning.

Formally, let $\mathcal{F}_r$ contain the original sample and every accepted batch through iteration r. For candidate batch b, let $A_{r,b} \in \{0, 1\}$ indicate that all fidelity, interval-reduction, and effect-stability gates are satisfied, and select the admissible candidate with the smallest interval width. With maximum iteration $R_{\max}$, the stopping time is

$$R = \min\{r : \max_b A_{r,b} = 0 \text{ or the relative width change is at most } \zeta\} \wedge R_{\max}.$$

The accepted design $\mathcal{D}^{(R)}$ and final estimator $\widehat{\tau}_T = \Psi(\mathcal{D}^{(R)}, \mathcal{D}_T)$ are therefore outputs of one adaptive map, not of a fixed final regression alone.

For outer-bootstrap replicate b, source units are resampled and the entire map is rerun: marginal fitting, semantic family selection, vine parameters, synthetic batches, acceptance decisions, stopping time, response learning, and target standardization. If $\widehat{\tau}_T^{*(b)}$ denotes the resulting estimate, interval endpoints are obtained from the empirical bootstrap law of $\widehat{\tau}_T^* - \widehat{\tau}_T$. The calibration target is coverage of the full adaptive estimator, $\Pr\{\tau_T \in \mathrm{CI}_{1-\alpha}^*\}$, rather than coverage conditional on the selected batches. Under the usual bootstrap regularity conditions and a gate-margin condition that keeps the population criteria away from their decision thresholds, this procedure propagates both generator-estimation and adaptive-stopping uncertainty. The reported experiments repeat this complete procedure rather than bootstrapping only the terminal response model.

A.12 CondVine Conditional Sampling

The controlled CondVine experiment fixes the DGP-defined effect-transport set before estimation. Source-fitted marginal transforms map Z and W to the copula scale, and the fitted vine is reordered so the observed target coordinates Z^{tar} form the conditioning block. Recursive h-functions then draw W from $\widehat{Q}_S(W \mid Z = Z^{\mathrm{tar}})$; inverse marginal and categorical-threshold maps return the draws to their original support. Treatment and outcomes are subsequently generated using the same allocation and response modules as VinePrior. Diagnostics report fidelity of the conditional draws, target alignment of the anchored coordinates, and ATE error over the paired shift grid. The main experiment uses one draw per target anchor; repeated draws are averaged in the general estimator stated in Section 4.4.

A.13 Prespecifying and Auditing the Effect-Transport Set

CondVine treats Z as an effect-transport set rather than as a generic dimension-reduction variable. In controlled simulations, Z is fixed from the known data-generating process before any method is evaluated. In an application, we recommend freezing Z before target-effect estimation using the following source-outcome and target-covariate information boundary.

First, candidate coordinates must be pretreatment variables with a substantive role as plausible effect modifiers. Second, source-only heterogeneity analyses can assess whether treatment contrasts vary with each candidate, using cross-fitted interaction or treatment-specific response models. Third, source–target support is evaluated for each candidate set using only covariates. Useful summaries include classifier-based density ratios, the effective sample size $\mathrm{ESS}_Z = (\sum_i \widehat{w}_Z(Z_i))^2 / \sum_i \widehat{w}_Z(Z_i)^2$, upper quantiles of $\widehat{w}_Z$, and nearest-source distances for target observations. The same summaries computed on full X show the overlap gained by moving to Z.

Two complementary diagnostics assess whether a proposed set is sufficiently rich. Residual effect heterogeneity examines variation of $\widehat{\mu}_1^S(Z, W) - \widehat{\mu}_0^S(Z, W)$ over W after conditioning on Z; conditional-shift diagnostics compare $P_S(W \mid Z)$ with $P_T(W \mid Z)$. A variable is most consequential when it contributes to both residual effect heterogeneity and conditional population shift. We therefore report sensitivity to adding or removing such variables whenever Z is not fixed by the scientific design. All rankings and overlap diagnostics exclude target treatment and outcome information. After Z is frozen, the source vine and response functions are refitted using the designated training sample.

Table 8: Logical dependency map for the theoretical results.

Main-paper statement	Required ingredients	Appendix location
Source-supported identification	source identification + mean transport	Section B.2
Transport amplification	density ratio + Cauchy–Schwarz	Section B.4
Augmented outcome bound	mixture risk + TV/KL control	Section B.5
End-to-end target-ATE bound	preceding two + target concentration	Section B.6
Likelihood-gate guarantee	finite-family uniform concentration	Section B.10
Semantic-prior oracle inequality	prior-weighted union bound	Section B.11
Reduced-overlap identification	effect transport over Z	Section B.14
CondVine error decomposition	response + conditional-law error	Section B.15

This protocol does not replace Assumption 3.3; it makes the effect-set choice auditable and connects observable diagnostics to the misspecification term B_Z in Proposition B.3. In particular, small residual heterogeneity or small conditional shift makes a large omitted-modifier contribution less plausible, while the exact decomposition continues to hold for any prespecified Z.

A.14 Computational Cost and Reproducibility

A d-dimensional regular vine contains $d(d-1)/2$ pair-copula edges, but VinePrior queries the LLM only for the $d-1$ unconditional edges in the first tree. All edge prompts can be issued independently and are cached. Marginal fitting, parameter estimation, higher-tree fitting, synthetic sampling, and downstream learning are local numerical operations. The backbone study reuses the same data split and augmentation settings for its AIC and semantic runs; the Qwen Plus and DeepSeek V3 calls each completed in under one minute on the reported environment. Exact runtime depends on the remote model service and is not used as an evaluation criterion.

Every experiment records the dataset identifier, source/target indices, seed, candidate families, selected edge families, fitted parameters, prompt version, synthetic budget, downstream learner, and evaluation output. The controlled simulation seeds are $\{42, 123, 456, 789, 1024, 1337, 1701, 2099, 2718, 3141\}$; IHDP four-seed comparisons use $\{42, 123, 456, 789\}$, except for the prespecified iterative-stabilization audit, which uses $\{42, 1024, 456, 789\}$. This distinction is reported explicitly in Table 10.

B Supplementary Theory

B.1 Proof Roadmap

Table 8 records how the assumptions and results in Sections 3 and 5 connect. The chain is deliberately modular: causal identification supplies the response functions, the generator bound controls their augmented estimation error, the density ratio transports that error, and target-sample concentration converts the population contrast to the reported estimator. CondVine replaces the final full-X transport step with its Z-level analogue.

B.2 Source Identification and Target Standardization

Under consistency, $Y = Y(t)$ whenever $T = t$. Conditional exchangeability then gives $\mathbb{E}[Y(t) \mid X = x, S = 1] = \mathbb{E}[Y(t) \mid X = x, T = t, S = 1]$, which equals $\mathbb{E}[Y \mid X = x, T = t, S = 1] = \mu_t^S(x)$. Source positivity ensures that both observed conditional means are defined on the source support. Under Assumption 3.2, the same conditional potential- outcome mean applies in the target population. Iterated expectation therefore yields $\mathbb{E}[Y(t) \mid S = 0] = \mathbb{E}_T \mu_t^S(X)$, and subtraction proves $\tau_T = \mathbb{E}_T\{\mu_1^S(X) - \mu_0^S(X)\}$. Absolute continuity ensures that the response functions are evaluated only on regions represented by the source law. This establishes the identification claim used in Proposition 5.1.

B.3 Notation and Regularity Conditions

For treatment arm $t \in \{0, 1\}$, let P_t denote the source law of $(X, Y) \mid T = t, S = 1$. The finite-sample augmentation result considers a randomized source study, so the X-marginal of each P_t is P_S^X. Let

$$Q_t = Q^X_{g_{\mathrm{LLM}}} Q_{g_{\mathrm{LLM}}}(Y \mid X, T = t)$$

be the corresponding generated law and let $M_t = \alpha_t P_t + (1-\alpha_t)Q_t$, where $\alpha_t = n_t^{\mathrm{out}}/(n_t^{\mathrm{out}} + m_t)$. For $h \in \mathcal{H}_t$, define the squared-loss risks $R_{P_t}(h) = \mathbb{E}_{P_t}(Y - h(X))^2$, $R_{Q_t}(h) = \mathbb{E}_{Q_t}(Y - h(X))^2$, and $R_{M_t}(h) = \alpha_t R_{P_t}(h) + (1 - \alpha_t) R_{Q_t}(h)$.

The generator is fitted without the outcome-learning fold. Conditional on the fitted generator, the real outcome-learning observations and generated observations are independent, with their counts treated as fixed. We assume that, with probability at least $1 - \delta$,

$$\sup_{h \in \mathcal{H}_t} \left| \widehat{R}_t(h) - R_{M_t}(h) \right| \le \varepsilon_t(n_t^{\mathrm{out}} + m_t, \delta),$$

where $\widehat{R}_t$ is the empirical risk formed from the n_t^{out} real and m_t generated observations. This condition accommodates the independent, stratified two-distribution sample; standard bounded-loss complexity bounds give explicit instances for particular response classes. Outcomes and fitted predictions are bounded by M, hence $0 \le (Y - h(X))^2 \le 4M^2$. We use $\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)| = \frac{1}{2}\|P - Q\|_1$.

The source-identification and transport results permit either randomized or observational source studies under Assumption 3.1. Randomization is required only for the finite-sample augmentation bound below, where it supplies the common arm-wise covariate marginal P_S^X. Cross-fitting can rotate the generator-fitting and outcome-learning roles while preserving the conditional independence used in each fold.

B.4 Proof of Transport Amplification

Under Assumptions 3.1 and 3.2, $\tau_T = \mathbb{E}_T[\mu_1^S(X) - \mu_0^S(X)]$. Therefore, the triangle inequality gives

$$|\widetilde{\tau}_T - \tau_T| \le \sum_{t \in \{0,1\}} \mathbb{E}_T \left| \widehat{\mu}_t(X) - \mu_t^S(X) \right|.$$

For $d_t(x) = \widehat{\mu}_t(x) - \mu_t^S(x)$, change of measure and Cauchy–Schwarz yield

$$\mathbb{E}_T|d_t(X)| = \mathbb{E}_S[w_X(X)|d_t(X)|] \le \sqrt{\mathbb{E}_S[w_X(X)^2]}\|d_t\|_{L_2(P_S^X)}.$$

Summing over t proves Proposition 5.1.

B.5 Proof of the Augmented Outcome-Learning Bound

For squared loss, the conditional-mean identity gives

$$R_{P_t}(h) - R_{P_t}(\mu_t^S) = \|h - \mu_t^S\|^2_{L_2(P_S^X)}.$$

Fix $\eta > 0$ and choose $h_{t,\eta} \in \mathcal{H}_t$ such that $\|h_{t,\eta} - \mu_t^S\|^2_{L_2(P_S^X)} \le a_t + \eta$. Because the loss takes values in $[0, 4M^2]$, the variational characterization of total variation implies, for every $h \in \mathcal{H}_t$,

$$|R_{P_t}(h) - R_{Q_t}(h)| \le 4M^2 \,\mathrm{TV}(P_t, Q_t).$$

Consequently,

$$|R_{P_t}(h) - R_{M_t}(h)| \le 4M^2(1 - \alpha_t)\,\mathrm{TV}(P_t, Q_t).$$

On the uniform-deviation event, empirical risk minimization gives

$$R_{M_t}(\widehat{\mu}_t^{\mathrm{aug}}) \le R_{M_t}(h_{t,\eta}) + 2\varepsilon_t(n_t^{\mathrm{out}} + m_t, \delta_H/2).$$

Applying the preceding distribution-shift inequality once to $\widehat{\mu}_t^{\mathrm{aug}}$ and once to $h_{t,\eta}$ yields

$$\begin{aligned} R_{P_t}(\widehat{\mu}_t^{\mathrm{aug}}) \le R_{P_t}(h_{t,\eta}) &+ 8M^2(1-\alpha_t)\,\mathrm{TV}(P_t, Q_t) \\ &+ 2\varepsilon_t(n_t^{\mathrm{out}} + m_t, \delta_H/2). \end{aligned}$$

The KL chain rule gives

$$\begin{aligned} D_{\mathrm{KL}}(P_t \| Q_t) &= D_{\mathrm{KL}}(P_S^X \| Q_{g_{\mathrm{LLM}}}^X) \\ &\quad + \mathbb{E}_{P_S^X} D_{\mathrm{KL}}\{P_S(Y \mid X, T=t) \| Q_{g_{\mathrm{LLM}}}(Y \mid X, T=t)\} \\ &= \kappa_{\mathrm{LLM}} + \xi_t. \end{aligned}$$

Pinsker's inequality therefore gives $\mathrm{TV}(P_t, Q_t) \le \sqrt{(\kappa_{\mathrm{LLM}} + \xi_t)/2}$. Subtracting $R_{P_t}(\mu_t^S)$, using the conditional-mean identity, and sending $\eta \downarrow 0$ proves the stated arm-wise bound. A union bound over $t \in \{0, 1\}$ gives probability at least $1 - \delta_H$.

The coefficient $8M^2$ thus comes from two changes of measure between the augmented mixture and the real arm distribution, each contributing at most $4M^2(1-\alpha_t)\,\mathrm{TV}(P_t, Q_t)$. The result also displays the synthetic-budget trade-off: m_t can improve the uniform-deviation term, while increasing $1 - \alpha_t$ places more weight on generator fidelity.

B.6 Proof of the End-to-End Target-ATE Bound

Define $\widetilde{\tau}_T = \mathbb{E}_T[\widehat{\mu}_1^{\mathrm{aug}}(X) - \widehat{\mu}_0^{\mathrm{aug}}(X)]$. Then

$$|\widehat{\tau}_T^{\mathrm{VP}} - \tau_T| \le |\widehat{\tau}_T^{\mathrm{VP}} - \widetilde{\tau}_T| + |\widetilde{\tau}_T - \tau_T|.$$

Conditional on the learned response functions, the target-sample contrast lies in $[-2M, 2M]$. Hoeffding's inequality therefore implies, with probability at least $1 - \delta_T$,

$$|\widehat{\tau}_T^{\mathrm{VP}} - \widetilde{\tau}_T| \le 4M\sqrt{\frac{\log(2/\delta_T)}{2N}} = r_N(\delta_T).$$

Apply Proposition 5.1 to the second term and substitute the square roots of the two bounds in Proposition 5.2. A union bound over the outcome-learning and target-averaging events proves Theorem 5.1.

B.7 Cross-Fitted Version of the Augmentation Bound

The conditioning argument used above extends fold by fold. Partition source indices into K folds. For fold k, fit the generator on the complement and treat the resulting $Q_t^{(-k)}$ as fixed while fitting or evaluating the fold-specific response function. Conditional on the complement, the same mixture-risk proof applies with $\kappa_{\mathrm{LLM}}^{(-k)}$, $\xi_t^{(-k)}$, $\alpha_t^{(k)}$, and $\varepsilon_t^{(k)}$. A union bound with failure probability δ_H/K gives simultaneous validity across folds. Jensen's inequality then bounds the error of the averaged response function by the average of the fold-specific squared errors. Thus cross-fitting uses all source observations across rotations while maintaining the independence needed by Proposition 5.2.

The theorem conditions on the fitted generator rather than treating the synthetic rows as newly randomized subjects. Generator-estimation uncertainty is propagated empirically by the outer bootstrap in Section A.11: each outer resample refits marginals, families, copula parameters, response functions, accepted batches, and the final target contrast.

B.8 Extension to Observational Source Studies

The identification results already permit observational sources under Assumption 3.1. Here we extend the finite-sample augmentation argument by targeting the common source-covariate risk with propensity-adjusted arm losses. Let $p_1(x) = e_S(x)$, $p_0(x) = 1 - e_S(x)$, and define $R_{S,t}(h) = \mathbb{E}_S[\mathbb{E}\{(Y - h(X))^2 \mid X, T = t, S = 1\}]$. Then $R_{S,t}(h) = \mathbb{E}_S[\mathbf{1}(T = t)(Y - h(X))^2/p_t(X)]$, whose minimizer is μ_t^S. In practice, p_t is replaced by a cross-fitted estimate $\widehat{p}_t$, clipped to $[\eta, 1-\eta]$, so that the observation used in a weighted loss does not train its own propensity model.

Proposition B.1 (Observational-source augmentation)**.** *Assume $|Y|, |h(X)| \le M$, source positivity with constant η, and $\|\widehat{e} - e_S\|_\infty \le r_e$, with both the true and fitted propensities bounded in $[\eta, 1-\eta]$. Let $\varepsilon_t^{\mathrm{obs}}$ be a uniform-deviation bound for the cross-fitted propensity-weighted real loss combined with the generated loss. Then the conclusion of Proposition 5.2 becomes*

$$\|\widehat{\mu}_t^{\mathrm{aug}} - \mu_t^S\|^2_{L_2(P_S^X)} \le a_t + 8M^2(1-\alpha_t)\sqrt{(\kappa_{\mathrm{LLM}} + \xi_t)/2} + 2\varepsilon_t^{\mathrm{obs}} + 8M^2\alpha_t r_e/\eta^2.$$

Justification. For any bounded squared loss, reciprocal-weight perturbation satisfies $|1/\widehat{p}_t(X) - 1/p_t(X)| \le r_e/\eta^2$. Hence replacing the ideal arm weight by its cross-fitted estimate perturbs the real-data risk uniformly by at most $4M^2 r_e/\eta^2$; its mixture contribution is multiplied by α_t, and the standard empirical-risk comparison doubles this term. The generated component is controlled exactly as in Section B.5. The sampling part of $\varepsilon_t^{\mathrm{obs}}$ also inherits the familiar positivity factor because the weights are bounded by $1/\eta$. Thus observational assignment adds an explicit nuisance-estimation and positivity cost while preserving the same generator-fidelity and augmented-sample terms. Fold rotation and the outer bootstrap propagate propensity, generator, and response-model uncertainty without using target outcomes.

B.9 Interpretation of the Synthetic-Budget Trade-off

Write the right-hand side of Equation (3) as $a_t + G_t(m_t) + E_t(m_t)$, where $G_t(m_t) = 8M^2\{m_t/(n_t^{\mathrm{out}} + m_t)\}\sqrt{(\kappa_{\mathrm{LLM}} + \xi_t)/2}$ is the generator-fidelity term and $E_t(m_t) = 2\varepsilon_t(n_t^{\mathrm{out}} + m_t, \delta_H/2)$ is the finite-sample learning term. For a fixed generator, G_t is nondecreasing and approaches the full discrepancy penalty, whereas standard complexity bounds make E_t decrease with the augmented design size. Consequently the bound does not advocate unlimited augmentation: the accepted-batch rule seeks the region in which the learning reduction remains empirically useful while source-fidelity and effect-stability criteria control the discrepancy term. This is the theoretical motivation for the stopping behavior reported in Tables 2 and 10.

B.10 Finite-Sample Guarantee for the Likelihood Gate

For one first-tree edge, let $L_e(f)$ denote the population validation log-likelihood and $\ell_e(f)$ its empirical counterpart. Let $f_e^\star \in \arg\max_{f \in \mathcal{F}} L_e(f)$. On the event $\sup_{f \in \mathcal{F}} |\ell_e(f) - L_e(f)| \le \eta_e$, every $\widehat{f}_e^{\mathrm{cal}} \in \mathcal{F}_e(\delta)$ satisfies

$$L_e(f_e^\star) - L_e(\widehat{f}_e^{\mathrm{cal}}) \le \delta + 2\eta_e.$$

Indeed, membership in the likelihood gate gives $\ell_e(\widehat{f}_e^{\mathrm{cal}}) \ge \max_f \ell_e(f) - \delta \ge \ell_e(f_e^\star) - \delta$; applying the uniform deviation bound to both candidates proves the claim. If each validation log-likelihood lies in an interval of width B_e, Hoeffding's inequality and a union bound over $\mathcal{F}$ give

$$\eta_e = B_e\sqrt{\frac{\log(2|\mathcal{F}|/\delta_e)}{2n_{\mathrm{val}}}}$$

with probability at least $1 - \delta_e$. The gate therefore excludes candidates with a population log-likelihood gap exceeding $\delta + 2\eta_e$, while the semantic prior determines the choice within the retained set.

B.11 Proof of the Semantic-Prior Oracle Inequality

Let $\mathcal{G}$ be finite and let the safeguarded prior be $\Pi_\rho(g) = (1-\rho)\Pi_{\mathrm{LLM}}(g) + \rho/|\mathcal{G}|$, with $0 < \rho \leq 1$. Conditional on the fitting fold, Π_ρ and all fitted candidates are fixed before the independent validation sample is observed. For a loss in $[L_c, U_c]$, Hoeffding's inequality implies

$$\Pr\Big(|\widehat{R}_X(g) - R_X(g)| > c_G(g;\delta_G)\Big) \leq \delta_G \Pi_\rho(g),$$

where $c_G(g;\delta_G) = B_c\sqrt{\{\log(1/\Pi_\rho(g)) + \log(2/\delta_G)\}/(2n_{\mathrm{val}})}$. Summing over $g \in \mathcal{G}$ shows that, with probability at least $1-\delta_G$, the deviations are simultaneously bounded by their respective penalties. On this event, for every $g \in \mathcal{G}$,

$$\begin{aligned} R_X(\widehat{g}_{\mathrm{cal}}) &\leq \widehat{R}_X(\widehat{g}_{\mathrm{cal}}) + c_G(\widehat{g}_{\mathrm{cal}};\delta_G) \\ &\leq \widehat{R}_X(g) + c_G(g;\delta_G) \\ &\leq R_X(g) + 2c_G(g;\delta_G). \end{aligned}$$

Taking the infimum proves Theorem 5.2.

For continuous densities, a directly applicable bounded loss is the clipped negative log-likelihood $\ell_g(x) = \mathrm{clip}\{-\log q_g(x), L_c, U_c\}$. Ordinary negative log-likelihood is covered when it is uniformly bounded on the evaluation support. In the latter case, population cross-entropy differs from $D_{\mathrm{KL}}(P_S^X \| Q_g^X)$ only by the entropy of P_S^X, so risk comparison and KL comparison agree.

B.12 Direct and Calibrated Selectors

The direct experimental selector and the calibrated theoretical selector have different roles. The direct rule $\widehat{f}_e^{\mathrm{LLM}} = \arg\max_f \pi_e(f)$ is fitted on the complete designated source sample and is the method evaluated in the primary ablations. The likelihood-gated rule first constructs $\mathcal{F}_e(\delta)$ on an honest split and then applies the same semantic ordering inside this statistically competitive set. It provides a safeguard when validation data are available; it is not used to replace or retroactively alter the direct decisions in Table 1.

At $\delta = 0$, the gate contains empirical likelihood maximizers and the semantic prior only resolves ties. As δ grows, more statistically near-optimal families enter, and semantic information receives greater scope. The finite-sample gate result controls the worst population likelihood gap by $\delta + 2\eta_e$. The oracle inequality instead scores complete candidate vines with a prior-dependent complexity penalty. The two results are complementary: the first is edge-level and set-valued, while the second is a complete-model risk comparison.

B.13 Robustness to an Imperfect Semantic Prior

The optional calibrated selector retains finite-sample protection when the semantic prior ranks an incorrect family highly. At the edge level, Section B.10 implies that every retained choice has population log-likelihood gap at most $\delta + 2\eta_e$, regardless of the ordering induced by π_e. The semantic prior therefore operates only within a statistically competitive set.

At the complete-vine level, the safeguarded prior $\Pi_\rho(g) = (1-\rho)\Pi_{\mathrm{LLM}}(g) + \rho/|\mathcal{G}|$ satisfies $\Pi_\rho(g) \geq \rho/|\mathcal{G}|$ for every candidate. Consequently, the penalty in Theorem 5.2 is always bounded by

$$c_G(g;\delta_G) \leq B_c\sqrt{\frac{\log(|\mathcal{G}|/\rho) + \log(2/\delta_G)}{2n_{\mathrm{val}}}}.$$

Combining this bound with Equation (5) shows that even an imperfect prior retains a finite-class oracle guarantee; when the semantic prior assigns greater mass to a low-risk candidate, its candidate-specific penalty becomes smaller. The uniform component $\rho/|\mathcal{G}|$ prevents any family from receiving zero support. This clarifies the roles of the two selectors: the direct selector uses the full small source sample and is the primary experimental method, whereas the likelihood-gated or safeguarded selector is appropriate when an honest validation fold is available and a formal protection against semantic-prior misspecification is desired.

B.14 Proof of CondVine Identification and Overlap Contraction

Source causal identification gives

$$\theta_S(z) = \mathbb{E}[Y(1) - Y(0) \mid Z = z, S = 1].$$

By Assumption 3.3, this equals $\mathbb{E}[Y(1) - Y(0) \mid Z = z, S = 0]$ on the target support. Iterated expectation over P_T^Z therefore yields

$$\tau_T = \mathbb{E}_{Z \sim P_T^Z}[\theta_S(Z)],$$

which proves the identification statement in Proposition 5.3. For $d(z) = \widehat{\theta}(z) - \theta_S(z)$, change of measure and Cauchy–Schwarz give

$$|\mathbb{E}_{P_T^Z} d(Z)| = |\mathbb{E}_{P_S^Z}[w_Z(Z)d(Z)]| \le \sqrt{\mathbb{E}_{P_S^Z}[w_Z(Z)^2]}\|d\|_{L_2(P_S^Z)}.$$

When the full density ratio exists, marginalization gives

$$w_Z(Z) = \mathbb{E}_{P_S^X}[w_X(X) \mid Z].$$

Conditional Jensen's inequality then implies

$$\mathbb{E}_{P_S^Z}[w_Z(Z)^2] \le \mathbb{E}_{P_S^X}[w_X(X)^2].$$

Since $\mathbb{E}_{P_S^Z} w_Z = \mathbb{E}_{P_S^X} w_X = 1$, subtracting one from both sides gives the overlap-contraction statement in Proposition 5.3. This is the chi-square-divergence contraction induced by projecting $X = (Z, W)$ onto Z.

B.15 CondVine Conditional-Generation Error

Let $d_S(z, w) = \mu_1^S(z, w) - \mu_0^S(z, w)$ and $\widehat{d}(z, w) = \widehat{\mu}_1^{\mathrm{aug}}(z, w) - \widehat{\mu}_0^{\mathrm{aug}}(z, w)$. For a fixed z,

$$\begin{aligned}|\widehat{\theta}_{\mathrm{CV}}(z) - \theta_S(z)| \le\ & \mathbb{E}_{\widehat{Q}_S(W|z)}|\widehat{\mu}_1^{\mathrm{aug}}(z, W) - \mu_1^S(z, W)| \\ & + \mathbb{E}_{\widehat{Q}_S(W|z)}|\widehat{\mu}_0^{\mathrm{aug}}(z, W) - \mu_0^S(z, W)| \\ & + 4M\,\mathrm{TV}\{\widehat{Q}_S(W \mid z), P_S(W \mid z)\}.\end{aligned}$$

The last constant follows because $\|d_S(z, \cdot)\|_\infty \le 2M$ and, under our total-variation convention, $|\mathbb{E}_P f - \mathbb{E}_Q f| \le 2\|f\|_\infty \mathrm{TV}(P, Q)$. Thus CondVine error separates into response-learning error and conditional-generation error. Our implementation uses one conditional draw per target anchor, yielding an order-$N^{-1/2}$ empirical-averaging term. More generally, L conditionally independent draws per anchor give an additional order-$(NL)^{-1/2}$ conditional Monte Carlo term. These terms complement the Z-level transport factor in Proposition 5.3.

Proposition B.2 (CondVine finite-sample decomposition)**.** *Let* $\Delta_t(z) = \mathbb{E}_{\widehat{Q}_S(W|z)}|\widehat{\mu}_t^{\mathrm{aug}}(z, W) - \mu_t^S(z, W)|$ *and* $b(z) = \mathrm{TV}\{\widehat{Q}_S(W \mid z), P_S(W \mid z)\}$. *Conditional on the learned response functions and conditional generator, the population target-aligned contrast obeys*

$$|\tau_{\mathrm{CV}} - \tau_T| \le \mathbb{E}_{P_T^Z}[\Delta_0(Z) + \Delta_1(Z) + 4Mb(Z)]\,.$$

If $|\mu_t^S|$ *and* $|\widehat{\mu}_t^{\mathrm{aug}}|$ *are bounded by* M*, then empirical averaging over* N *target anchors and* L *independent conditional completions per anchor adds, with probability at least* $1 - \delta$*, terms of order* $M\sqrt{\log(1/\delta)/N}$ *and* $M\sqrt{\log(1/\delta)/(NL)}$.

Proof. For each anchor, add and subtract the contrast integrated under the true conditional law $P_S(W \mid z)$. The two response-learning terms follow from the triangle inequality, while the difference between the generated and true conditional expectations is at most $4Mb(z)$. Averaging over

Table 9: Assumptions, observable diagnostics, and their roles in the analysis.

Condition	Reported diagnostic	Role in the analysis
Source treatment positivity	propensity range, arm counts, weighted ESS, and clipping frequency	controls stability of source response learning and the $1/\eta$ factor in Proposition B.1
Source–target support on X	density-ratio quantiles, target-to-source distances, and full-X ESS	diagnoses the shift factor in Proposition 5.1
Reduced overlap on Z	Z-level ESS and density-ratio summaries compared with their full-X counterparts	quantifies the overlap contraction established by Proposition 5.3
Effect transport over Z	source residual-effect heterogeneity, conditional $W \mid Z$ shift, and add/drop sensitivity	audits the variables contributing to B_Z in Proposition B.3
Vine generator fidelity	held-out likelihood, family recovery, rank and tail error, SMD, and correlation error	probes κ_{LLM}, ξ_t, and conditional-generation error
Semantic family prior	response-schema validity, prior entropy, selected-family stability, and likelihood-gate membership	distinguishes direct semantic selection from the safeguarded analysis in Section B.13
Adaptive augmentation	accepted batches, stopping iteration, interval-width trace, and bootstrap stability	verifies that uncertainty repeats the full adaptive map in Section A.11

P_T^Z gives the displayed population inequality. For the empirical estimator, add and subtract first the target-anchor expectation and then the conditional completion expectation. Conditional contrasts lie in $[-2M, 2M]$, so Hoeffding bounds the anchor-average and within-anchor completion deviations. A union bound completes the result.

For the reported CondVine experiments $L = 1$. We therefore evaluate one target-aligned completion for every observed anchor and report the resulting paired-seed variability; no separate large-L Monte Carlo extrapolation is used in the main results.

B.16 CondVine under an Approximate Effect-Transport Set

Exact reduced-overlap transport sets $\theta_S(z) = \theta_T(z)$, where $\theta_T(z) = \mathbb{E}[Y(1) - Y(0) \mid Z = z, S = 0]$. For a prespecified but possibly incomplete set Z, define its transport misspecification by $B_Z = |\mathbb{E}_{P_T^Z}\{\theta_S(Z) - \theta_T(Z)\}|$. The following result separates this term from estimable response and generation errors.

Proposition B.3 (Approximate effect-set decomposition)**.** *Under source identification and $P_T^Z \ll P_S^Z$, but without requiring exact conditional-effect transport over Z, the population CondVine contrast satisfies*

$$|\tau_{\mathrm{CV}} - \tau_T| \le \mathbb{E}_{P_T^Z}[\Delta_0(Z) + \Delta_1(Z) + 4Mb(Z)] + B_Z,$$

where Δ_t and b are defined in Proposition B.2. Empirical target standardization and conditional completion add the same sampling terms as in that proposition.

Proof. Add and subtract $\mathbb{E}_{P_T^Z}\theta_S(Z)$. The first difference is bounded by the response-learning and conditional-generation argument in Proposition B.2; the second is exactly B_Z. If the full-X conditional effect $\tau(x)$ is stable across populations but Z omits effect modifiers in W, then $\theta_S(z) - \theta_T(z) = \int \tau(z, w)\,\{\mathrm{d}P_S(w \mid z) - \mathrm{d}P_T(w \mid z)\}$. Therefore $|\theta_S(z) - \theta_T(z)| \le 4M\,\mathrm{TV}\{P_S(W \mid z), P_T(W \mid z)\}$, giving an interpretable diagnostic bound. When Z is a valid effect-transport set, $B_Z = 0$ and the result reduces to Proposition B.2.

B.17 Assumption-to-Diagnostic Map

Identification assumptions are scientific conditions, whereas diagnostics measure observable implications and finite-sample stress. To keep these roles distinct, Table 9 records what is assumed, what is reported, and which part of the analysis is affected. No diagnostic is used as a substitute for an identification condition or is tuned using the target effect.

For randomized sources, propensity diagnostics reduce to the known allocation probability and realized arm counts. For observational sources, they are computed out of fold. Support and generator diagnostics use no outcomes from the target population. This separation mirrors the information-use audit in Section C.4 and provides a compact checklist for reproducing each theoretical regime.

C Supplementary Experiments

C.1 Iterative Augmentation Traces

Table 10: IHDP iterative stabilization over four paired source samples.

Seed	Initial CI	Final CI	Reduction	Initial error	Final error	Iterations
42	12.10	1.50	87.6%	1.302	0.479	3
1024	6.96	0.91	87.0%	0.615	0.688	3
456	6.22	0.68	89.1%	0.412	0.170	3
789	4.36	1.88	56.9%	0.564	0.367	2
Mean	7.41	1.24	80.1%	0.723	0.426	2.75

The confirmatory uncertainty analysis resamples the original source units and repeats vine fitting, family selection, augmentation, batch acceptance, and outcome learning within each outer-bootstrap replicate. This treats the accepted synthetic sample as an output of the source study rather than as an independent source of subjects.

Table 11: Batch-acceptance and stopping criteria used by iterative augmentation. All quantities are computed without target outcomes.

Criterion	Operational rule	Purpose
Source fidelity	mean covariate SMD < 0.2	retain source-supported design
Interval improvement	candidate CI narrower than current CI	reduce sampling variability
Effect stability	source estimate within tolerance	control drift
Convergence	relative width change < 0.02	stop marginal iterations
Budget	at most five rounds	bound computation

The four reported splits are a prespecified audit set. In the hybrid strategy, semantic selection is used for the first small-sample round and AIC thereafter, when the accepted design is larger. A separate all-LLM audit produces a mean CI reduction of 79.1%, compared with 80.1% for the hybrid strategy, supporting the interpretation that semantic guidance is most useful at the small-sample cold start. This secondary comparison is descriptive and does not alter the accepted batches used in Table 10.

Table 12: Hybrid versus all-LLM iterative selection on the same four IHDP splits. Values are mean $\pm$ standard deviation; lower error and CI width are better.

Strategy	Rounds	Final CI	CI reduction	S-RF	T-RF	DR-RF
Hybrid LLM–AIC	2.75 ± 0.50	1.24 ± 0.57	$80.1\pm15.5\%$	0.307 ± 0.190	0.301 ± 0.195	0.509 ± 0.335
All-LLM	2.25 ± 0.50	1.32 ± 0.32	$79.1\pm11.2\%$	0.255 ± 0.133	0.284 ± 0.129	0.222 ± 0.169

The two strategies produce nearly identical interval compression. All-LLM uses fewer accepted rounds on average and has lower downstream DR-RF error, whereas the hybrid rule obtains a slightly narrower final interval while avoiding repeated LLM calls after the small-sample first round. This supports the hybrid rule as a computationally economical default rather than implying that later semantic calls are required for identification.

C.2 Dependence and Generator Diagnostics

The semantic prompt and the post-fit diagnostics serve different purposes. Prompt summaries are computed before family selection and contain only low-dimensional source statistics. Post-fit diagnostics evaluate the complete generator and do not feed back into the direct selector. For a pair of pseudo-observations (U, V), Kendall's τ and Spearman's ρ measure rank dependence, while Pearson correlation captures linear association. Marginal skewness and excess kurtosis summarize asymmetry and tail weight. Empirical upper- and lower-tail co-occurrence at threshold q are computed as

Table 13: Information available to each stage of VinePrior and CondVine. Target outcomes and the true target effect are used exclusively for post-hoc evaluation and never for model fitting, semantic selection, augmentation, or stopping decisions.

Stage	Source X	Source (T, Y)	Target X	Target Y	True τ_T
Marginal/vine fitting	✓	–	–	–	–
Semantic edge selection	aggregate only	–	–	–	–
Treatment/outcome generator	✓	✓	–	–	–
Batch acceptance	✓	✓	standardization only	–	–
CondVine anchoring	source fit	source fit	Z only	–	–
Final target standardization	learned model	learned model	✓	–	–
Reported error	–	–	–	evaluation	✓

$\widehat{\lambda}_U(q) = \Pr(U > q, V > q)/(1-q)$ and $\widehat{\lambda}_L(q) = \Pr(U \leq q, V \leq q)/q$, respectively. These finite-threshold quantities are used as summaries rather than asserted as asymptotic tail-dependence coefficients.

Complete-generator evaluation includes held-out negative log likelihood, rank-correlation error, tail-co-occurrence error, source mean SMD, and correlation-matrix MSE. For known-family simulations, recovery is scored only when the fitted edge corresponds to the known edge in the data-generating vine; rotated families are treated as distinct labels. Parameter error is reported conditional on correct family recovery, while density accuracy is reported unconditionally so an apparently correct label cannot mask poor numerical fitting.

C.3 Known-Family and Held-Out Validation Protocol

The known-family validation separates the small fitting sample from an independent evaluation sample generated by the same source DGP. Marginals, edge summaries, LLM decisions, and copula parameters use only the designated small fitting sample. The evaluation sample is used to compute family recovery and joint-density metrics after the decision has been frozen. For semi- synthetic data, where the true family is unavailable, repeated cross-fitting rotates fitting and validation folds. Each fold recomputes prompt summaries without its evaluation units, applies the unchanged selection rule, and scores the fitted density out of fold. The final generator is refitted on the full source sample only after the protocol and family-selection rule are fixed.

The direct selector evaluated in the main paper does not use held-out scores to revise its family decision. The optional likelihood-gated selector in Section A.2 is reported as a separate construction whenever it is evaluated. This separation prevents validation performance from silently becoming a second selection stage in the direct VinePrior results.

C.4 Information-Use and Leakage Audit

Variable descriptions are treated as external metadata and are fixed before observing target performance. Neither prompt rationales nor cached decisions are edited using target-ATE error. Random-family ablations are averaged over all prespecified assignments; the best random realization is not selected, because doing so would require the unavailable target effect. Similarly, the iterative acceptance rule uses source fidelity, interval width, and effect-stability criteria, not the final target error.

C.5 Reproducibility Checklist

For each reported aggregate, the experiment record contains (i) the source, target, and evaluation indices; (ii) the random seed and treatment-stratified source counts; (iii) marginal families and categorical thresholds; (iv) the vine edge set, candidate family set, semantic response, selected family, and fitted parameters; (v) the treatment and outcome generator settings; (vi) the number of generated and accepted observations; and (vii) predictions, the true target effect used for evaluation, and the resulting absolute error. Tables are aggregated from these machine-readable records rather than from manually transcribed intermediate logs.

The main simulation table uses the same ten seeds for every displayed method. The IHDP main table and semantic-versus-AIC ablation use four paired seeds. The matched-backbone table is explicitly

a fixed-split replication and is not presented as a multi-seed uncertainty estimate. Deep-generator failures remain in the aggregate. All model-facing prompts omit row-level data, and cached responses are stored together with the canonical parsed family so the selection step can be inspected independently of downstream results.

C.6 Scope of the Supplementary Experiments

The main claims are tied to the regimes tested by each experiment. Table 1 compares estimation error under the primary small- source settings. Table 10 evaluates adaptive stabilization of source augmentation. Table 19 and Table 21 isolate the discrete semantic-selection step. Table 17 evaluates reduced-overlap transport when the effect-relevant set is known from the DGP. Finally, Tables 22, 24 and 25 locate sensitivity to source size, covariate dimension, and prompt design. These studies are reported separately so that results from different seed sets, selectors, or overlap regimes are not pooled into a single aggregate.

C.7 Dataset Construction and Target Shift

IHDP. The semi-synthetic IHDP benchmark contains 747 units, 25 pretreatment covariates, a binary treatment with an empirical treated fraction near 0.19, and simulated potential outcomes. For every seed, target and source indices are formed before any generator is fitted. Let $r_i = (X_{i1} - \overline{X}_1)/\widehat{\sigma}_1$ denote the standardized first covariate. Target units are sampled without replacement with probabilities proportional to $\mathrm{expit}(sr_i)$, where s is the reported shift strength. The source sample is drawn from the remaining units with treatment-stratified sampling so its small-sample treatment fraction tracks the available source pool. Target treatment and outcomes are withheld from all fitting and selection stages. Known potential outcomes are used only to compute the final target-ATE error.

Controlled simulations. Both simulations use $d = 20$, $n = 100$ source units, a balanced randomized treatment, 500 target covariate observations, and a separate target-law test population of 2,000 observations. The effect-transport set is fixed before estimation as $Z = (X_1, X_2, X_3)$. Simulation uses a Gaussian conditional law with adequate target-relevant source support. Sim-Shift uses a nonlinear, mixed-type conditional law and reduced dispersion to create weak full-X overlap while retaining the known Z-level construction. The ten seeds are listed in Section A.14.

Table 14: Main experimental settings. A dash indicates that the quantity is determined by the available semi-synthetic benchmark split.

Dataset	d	source n	target N	treated rate	regime
IHDP	25	100	–	≈ 0.19	covariate shift
Simulation	20	100	500	0.50	source-supported
Sim-Shift	20	100	500	0.50	reduced overlap

C.8 Controlled Simulation Data-Generating Processes

For the source-supported Simulation, source and target transport coordinates follow $Z_S \sim \mathcal{N}(-s/2, I_3)$ and $Z_T \sim \mathcal{N}(s/2, I_3)$. For $j = 1, \ldots, 17$, the remaining coordinates satisfy $W_j = 0.5 \sin(j+1) Z_1 + \epsilon_j$, with independent standard Gaussian noise. Treatment is $T \sim \mathrm{Bernoulli}(0.5)$. The heterogeneous effect is $\tau(X) = 2 + 0.3Z_1$, and outcomes follow $Y(0) = 3Z_1 + 1.5W_1 + 0.5W_2 + \epsilon_Y$ and $Y(1) = Y(0) + \tau(X)$, where $\epsilon_Y \sim \mathcal{N}(0, 0.5^2)$. The reported setting uses $s = 2$.

Sim-Shift retains the same effect and outcome equations but uses $Z_S \sim \mathcal{N}(-s/2, 0.8^2 I_3)$ and $Z_T \sim \mathcal{N}(s/2, 0.8^2 I_3)$. Its conditional covariates include $W_1 \sim \mathrm{Bernoulli}\{\mathrm{expit}(Z_1^2 - 2)\}$, a zero-inflated W_2 that equals zero with probability one half and otherwise follows $\mathcal{N}(Z_1^2, 1)$, and heavy-tailed coordinates $W_j = 0.5 \sin(j+1) Z_1^2 + 0.8\epsilon_j$ for $j \geq 3$, with $\epsilon_j \sim t_3$. This construction tests conditional generation under nonlinear and tail-dependent $P_S(W \mid Z)$. The main Sim-Shift result uses $s = 3$; Table 17 also reports $s = 2$ and $s = 4$. The true target ATE is the empirical mean of $\tau(X)$ in the independent target-law test population.

Table 15: Shared implementation settings for the main comparisons.

Component	Setting	Component	Setting
Source size	100	IHDP shift	$s = 2$
Simulation seeds	10 paired	IHDP main seeds	4 paired
S-RF	300 trees, leaf 5	T-RF	200 trees/arm, leaf 5
DR-RF	200-tree nuisances	propensity clipping	$[0.05, 0.95]$
CTGAN/TVAE	300 epochs	TabDDPM	1,000 epochs, batch 64
Semantic scope	first tree only	target outcomes	evaluation only

C.9 Baseline Implementations and Fairness Controls

No augmentation fits the downstream estimator to the source sample alone. IPSW estimates a source-membership density ratio and weights source units; calibration weighting balances target covariate moments. SMOTE interpolates within treatment strata with at most five neighbors. KDE standardizes covariates and uses Scott bandwidth. Gaussian copula fits empirical marginals and a Gaussian dependence matrix. Vine+AIC uses the same marginals, vine edge set, causal-order generator, and synthetic budget as VinePrior, replacing only the first-tree semantic selector by AIC. CTGAN and TVAE train for 300 epochs; TabDDPM uses 1,000 epochs and batch size 64. Target-X copies observed target covariates into the common treatment/outcome generator and therefore serves as a response-learning reference rather than a feasible target-outcome oracle.

For each seed, all covariate generators receive the same source sample and produce the same number of synthetic covariate rows. Treatment allocation, arm-specific response models, residual resampling, downstream estimator, and target standardization are shared. Hyperparameters are fixed before inspecting target-ATE errors. Failed or numerically extreme deep-generator runs remain in the reported mean and standard deviation; they are not replaced by favorable seeds.

C.10 Downstream Estimators, Metrics, and Hyperparameters

S-Lin fits one linear outcome regression to (X, T). S-RF uses a pooled random forest with 300 trees and minimum leaf size five. T-RF fits separate 200-tree forests in the two treatment arms, again with minimum leaf size five. DR-RF fits the same arm-specific forests, clips estimated propensities to $[0.05, 0.95]$, constructs the doubly robust pseudo-outcome, and regresses it on X with a 200-tree forest. Every estimator is standardized over the same target covariates or CondVine completions.

The primary metric is $|\widehat{\tau}_T - \tau_T|$. Source mean standardized differences are $d^{-1} \sum_k |\overline{X}_{S,k} - \overline{X}_{G,k}|/s_{S,k}$, with zero-variance coordinates handled by the common preprocessing rule. Correlation MSE averages squared differences between the upper-triangular entries of the source and generated correlation matrices. CI width is the difference between the upper and lower endpoints of the bootstrap interval. Every aggregate reports the arithmetic mean and sample standard deviation over the stated paired seeds.

C.11 Complete Main-Result Protocol and Tables

All methods within a dataset use identical source–target splits, paired seeds, augmentation budgets, treatment and outcome generators, and downstream estimators. The controlled simulations use the ten seeds $\{42, 123, 456, 789, 1024, 1337, 1701, 2099, 2718, 3141\}$. In the reduced-overlap simulation, the effect-relevant set $Z = (X_1, X_2, X_3)$ is fixed by the DGP before fitting any method. We report per-seed results, full multi-estimator results, and the completed deep-generator baselines. An entry is transferred to the main table only after the paired protocol is complete; results from different seed sets or downstream estimators are not pooled.

The estimator comparison changes only the downstream learner: source–target splits, generated covariates, treatment and outcome assignments, and synthetic budgets remain fixed. The paired results show that the conclusions in the main T-RF comparison are not produced by a favorable re-run of the generators; complete per-seed records are included with the experiment outputs.

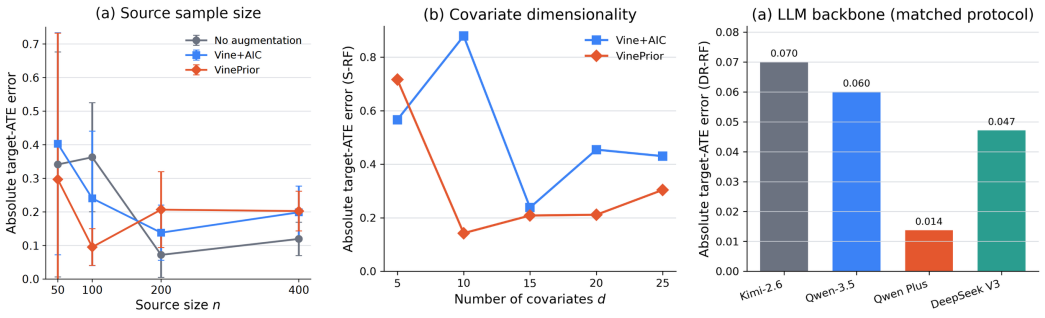


Figure 2: **Sensitivity analysis of VinePrior.** Effects of (a) source-sample size on IHDP, (b) covariate dimensionality, and (c) LLM backbone under matched protocols. Error bars denote one standard deviation when multiple paired seeds are available.

Table 16: **Estimator robustness of the controlled-simulation baselines.** Absolute target-ATE error (mean ± standard deviation) over the same ten paired seeds. Each generator is fitted once per seed, and all four downstream estimators use the resulting common training sample. The main text reports T-RF; lower is better.

(a) Simulation: $s = 2.0$

Method	S-Lin	S-RF	T-RF	DR-RF
No augmentation	0.614±0.114	1.901±0.481	1.028±0.560	0.765±0.288
IPSW	0.569±0.199	1.950±0.521	1.328±0.900	1.277±0.699
CW	2.015±0.281	2.308±0.052	1.819±1.170	1.578±1.421
SMOTE	0.566±0.265	1.015±0.591	1.013±0.592	1.017±0.556
KDE	0.668±0.293	1.031±0.573	1.031±0.575	1.002±0.542
Gaussian copula	0.598±0.272	1.007±0.576	1.008±0.580	0.957±0.496
CTGAN	0.770±0.447	1.033±0.579	1.032±0.579	1.012±0.553
TVAE	0.537±0.253	0.979±0.512	0.979±0.513	0.944±0.458
TabDDPM	0.790±0.339	1.056±0.712	0.914±0.673	1.049±0.713

(b) Sim-Shift: $s = 3.0$

Method	S-Lin	S-RF	T-RF	DR-RF
No augmentation	0.908±0.113	1.443±0.506	0.955±0.400	0.930±0.353
IPSW	0.925±0.098	1.397±0.836	1.387±1.042	1.864±1.254
CW	2.176±0.179	2.421±0.049	1.122±0.633	1.185±0.483
SMOTE	0.877±0.167	0.968±0.436	0.972±0.439	0.995±0.389
KDE	0.954±0.216	0.960±0.411	0.960±0.412	0.960±0.391
Gaussian copula	0.905±0.200	0.966±0.409	0.963±0.411	0.931±0.345
CTGAN	0.882±0.399	0.890±0.375	0.891±0.374	0.931±0.364
TVAE	0.810±0.251	0.937±0.384	0.933±0.382	0.891±0.371
TabDDPM	1.047±0.441	0.984±0.483	0.987±0.481	1.069±0.411

Table 17: Target-ATE error across reduced-overlap regimes. The effect-relevant set $Z = (X_1, X_2, X_3)$ is fixed by the simulation DGP. All entries use the same ten paired seeds and T-learner RF; lower is better.

Method	$s = 2$	$s = 3$	$s = 4$
No augmentation	1.028±0.560	0.955±0.400	1.271±0.357
Vine+AIC	0.971±0.570	0.917±0.415	1.240±0.413
VinePrior	0.990±0.578	0.928±0.429	1.218±0.392
Target-anchor ablation	0.835±0.496	0.900±0.197	1.213±0.261
CondVine	**0.757±0.504**	**0.633±0.293**	**1.123±0.593**
Target-X	1.027±0.558	0.954±0.399	1.270±0.357

At moderate shift, direct target anchoring is the strongest target-alignment ablation. As full-covariate overlap deteriorates, conditional generation becomes more valuable: CondVine attains the lowest mean error at $s = 3$ and $s = 4$. This transition supports the reduced-overlap mechanism rather than selecting a single alignment rule for every regime.

Table 18: Per-seed CondVine target-ATE error for the overlap experiment.

Seed	$s = 2$	$s = 3$	$s = 4$
42	1.063	1.150	1.179
123	0.025	0.742	1.114
456	1.404	1.662	2.073
789	0.187	0.067	0.459
1024	0.516	0.324	1.216
1337	1.453	1.540	2.077
1701	0.122	0.908	1.086
2099	0.915	0.695	0.076
2718	0.695	0.303	1.210
3141	1.194	0.959	0.743
Mean	0.757	0.835	1.123

C.12 Complete First-Tree Selector Ablations

The AIC, random-family, and direct semantic-prior selectors share the same first-tree edges, candidate families, marginal fits, synthetic budget, and outcome learner. Table 19 compares one deterministic semantic selection with ten random assignments on each of four paired source splits. Semantic selection wins on three splits and reduces the across-split mean from 0.335 to 0.059. In the expanded ten-seed audit, random selection exceeds error 0.7 in $10/100$ runs, compared with $0/10$ semantic selections. The likelihood-gated variant in Section A.2 is analyzed separately and is not pooled with these direct-selector results.

Table 19: Direct semantic-prior versus random first-tree selection on IHDP using T-RF. Random entries summarize ten assignments per source split.

Seed	Semantic prior	Random family
42	$\mathbf{0.030}$	0.334 ± 0.013
123	$\mathbf{0.007}$	0.472 ± 0.014
456	0.122	$\mathbf{0.062 \pm 0.011}$
789	$\mathbf{0.079}$	0.470 ± 0.014
Mean	$\mathbf{0.059}$	0.335

C.13 Fidelity and Target References

Table 20 compares low-order source fidelity with downstream causal utility. Gaussian copula has the smallest marginal SMD and correlation error but the largest S-RF target-ATE error, whereas VinePrior has the smallest ATE error. The Target-X comparison uses the same source-fitted response learner and therefore isolates the contribution of outcome learning.

Table 20: Source-distribution diagnostics and IHDP S-RF target-ATE error. Lower is better; matching low-order moments does not by itself ensure causal utility.

Method	Mean SMD	Correlation MSE	ATE error
Gaussian copula	$\mathbf{0.037}$	$\mathbf{0.006}$	1.812
KDE	0.049	0.017	0.742
SMOTE	0.193	0.014	0.392
IPSW	0.230	0.015	0.511
CTGAN	0.198	0.029	3.136
VinePrior	0.131	0.018	$\mathbf{0.066}$

C.14 LLM-Backbone Replication

We compare four backbones under the same protocol: a fixed IHDP split, $n_{\mathrm{rct}} = 100$, shift $= 2$, the direct five-family selector, a common augmentation budget, and the DR-RF estimator. Existing Qwen-3.5 and Kimi-2.6 results are complemented by matched Qwen Plus and DeepSeek V3 runs. All first-tree edges receive valid model responses without fallback selection.

Table 21: LLM-backbone sensitivity on the matched IHDP protocol. Absolute target-ATE error for DR-RF on the fixed seed-42 split; lower is better. Each AIC value is obtained from the same run as its corresponding LLM selector.

Backbone	Vine+AIC	VinePrior	Reduction
Kimi-2.6	0.388	0.070	82.0%
Qwen-3.5	0.277	0.060	78.3%
Qwen Plus	0.316	$\mathbf{0.014}$	95.7%
DeepSeek V3	0.316	0.047	85.1%

Table 22: Source-sample-size sensitivity on IHDP.

n	No augmentation	Vine+AIC	VinePrior
50	0.341 ± 0.335	0.403 ± 0.331	$\mathbf{0.297 \pm 0.435}$
100	0.363 ± 0.162	0.240 ± 0.200	$\mathbf{0.095 \pm 0.055}$
200	$\mathbf{0.072 \pm 0.067}$	0.138 ± 0.082	0.207 ± 0.113
400	$\mathbf{0.120 \pm 0.050}$	0.199 ± 0.078	0.202 ± 0.059

C.15 Completed Sensitivity Sweeps

Tables 21 to 24 provide the numerical values and protocols underlying the sensitivity panels. The source-size and shift sweeps use T-learner RF and four fixed IHDP seeds; estimator- and backbone-specific settings are stated in their respective captions.

Table 23: Covariate-shift sensitivity on IHDP.

Shift	No augmentation	Vine+AIC	VinePrior	Target-X
0.5	$\mathbf{0.202 \pm 0.226}$	0.945 ± 0.651	1.143 ± 0.852	0.831 ± 0.393
1.0	0.192 ± 0.126	0.212 ± 0.151	$\mathbf{0.158 \pm 0.050}$	0.453 ± 0.098
2.0	0.363 ± 0.162	0.240 ± 0.200	$\mathbf{0.095 \pm 0.055}$	0.385 ± 0.633
3.0	$\mathbf{0.208 \pm 0.206}$	0.371 ± 0.346	0.522 ± 0.383	0.972 ± 0.136

Table 24: Covariate-dimensionality sensitivity on IHDP using S-RF. Each row uses the same four paired seeds, source size, shift strength, and outcome pipeline; lower is better.

d	No augmentation	Vine+AIC	VinePrior
5	0.466 ± 0.162	$\mathbf{0.566 \pm 0.449}$	0.716 ± 0.585
10	0.389 ± 0.057	0.879 ± 0.704	$\mathbf{0.142 \pm 0.138}$
15	0.389 ± 0.065	0.238 ± 0.248	$\mathbf{0.208 \pm 0.412}$
20	0.385 ± 0.074	0.454 ± 0.495	$\mathbf{0.211 \pm 0.133}$
25	0.386 ± 0.073	0.430 ± 0.154	$\mathbf{0.304 \pm 0.349}$

These sweeps locate, rather than redefine, the operating regime. Semantic guidance improves Vine+AIC at $n \leq 100$, where family-selection uncertainty is largest; with larger source samples the unaugmented learner is already stable. In the source-only IHDP study the gain over Vine+AIC peaks at the moderate shift $s = 2$; the separate CondVine overlap experiment tests stronger shifts while retaining overlap on the effect-relevant coordinates. Across the dimension sweep, VinePrior improves over Vine+AIC for four of five values of d; the low-dimensional $d = 5$ case is the exception, where the discrete family-selection problem is comparatively small and both augmented methods are less stable than no augmentation.

C.16 Prompt and candidate-family sensitivity

We compare a minimal design containing variable types and rank dependence over five families with a richer design that adds variable descriptions, marginal shape, Kendall's τ, and empirical tail dependence over fifteen families. The semantic-prior backbone is held fixed, and both designs use identical source splits, paired seeds, augmentation budgets, and downstream estimators.

Table 25: Prompt and candidate-family sensitivity on IHDP. Absolute target-ATE error (mean $\pm$ standard deviation) over ten paired seeds; lower is better.

Design	Families	S-RF	T-RF	DR-RF
Minimal	5	0.240 ± 0.212	0.249 ± 0.215	0.338 ± 0.278
Rich	15	$\mathbf{0.209 \pm 0.241}$	$\mathbf{0.216 \pm 0.229}$	$\mathbf{0.298 \pm 0.195}$