STRUCTURAL REPLACEMENT CERTIFICATION FOR HIGH-DIMENSIONAL DISCRETE POSTERIORS FINITE-SAMPLE CERTIFICATES VIA LOCAL COMPRESSION AND UNIFIED TV COUPLINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Observation-wise product posteriors can become dependent after observations are marginalized out. Under an access model in which full posterior probabilities are available, the representation is frozen, and scoring is independent, we certify whether a given mixture-of-products candidate outperforms the entire product family in normalized Hamming Wasserstein distance. Under an exact fractional cover, the approximation loss of the local compression lower bound is at most twice the compression distortion plus a structural remainder, without requiring independence across groups. On the candidate side, we use direct component transport as a baseline, combine prefix aggregation and state lifting into a TV hierarchy with nonincreasing costs, and enlarge the set of valid couplings through a stopping tree that retains parent nodes. Once the reference, candidate, and metric are fixed, a common empirical event converts valid coupling costs into finite-sample upper endpoints, yielding the structural replacement criterion $U < L$. Validity holds for any fixed finite number of components. In synthetic experiments, 60 of 80 audits using codebooks frozen after learning pass numerically; a further 180 perturbation comparisons and 400 reference-structure comparisons verify reductions in coupling costs and upper endpoints. The additional certification gains from combining the new couplings with the codebook lower bound, applicability to trained RSSMs, and certification with rigorous rounding remain unverified.

Keywords: RSSM; structural replacement certification; local compression; feasible coupling; finite-sample certification

1 INTRODUCTION

PlaNet's recurrent state-space model (RSSM) combines deterministic memory with stochastic states (Hafner et al., 2019), while DreamerV2 uses a categorical prior factorized across blocks conditional on memory (Hafner et al., 2021). A future observation can simultaneously change multiple block probabilities even with factorized observation-wise posteriors; marginalizing observations can thus induce dependencies that a product prior cannot represent.

With the encoder, deterministic recurrent model, runtime state, and evaluation data distribution frozen, we compare the full product family with a given mixture-of-products candidate using the conditional aggregate posterior as reference. Observation-wise posterior probabilities yield a lower bound L on the original family's optimal error and an upper bound U on the candidate's full-space error. Thus, $U < L$ certifies improvement over every member of the original family without assuming optimal training of the original model. The main results are:

Approximation guarantee for the local lower bound. Under an exact fractional cover, local compression loses at most $2e + \tau$, even with arbitrary dependencies across groups.

Extension of valid couplings. State lifting and tail-prefix aggregation form a TV hierarchy with nonincreasing costs. Stopping trees enlarge the feasible set through exact splitting, parent-node

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.

retention, and optimization of stopping masses, and can be incorporated into the hierarchy, always preserving the full marginals.

Finite-sample certification. A fixed population transport potential gives a common one-sided scoring event for selecting feasible plans, lifting levels, and valid tree representations within the same empirical problem. Calibration and conditional detection analyses appear in the appendices.

We build on independence distances, mixture transport, and concentration tools; auxiliary statistical endpoints and comparisons at the same budget appear in the appendices. Empirical coupling costs require deterministic evaluation or deterministic safe upper bounds, with numerical errors accounted for separately. Experiments verify three-block coupling gains in costs and upper endpoints under perturbations and different reference structures. Incompletely reproduced earlier summaries are not core evidence; applicability to trained RSSMs and task gains remain unverified.

2 Related Work

2.1 Comparison Target and Access Model

PlaNet and DreamerV2 provide background on recurrent latent dynamics and discrete states, respectively (Hafner et al., 2019; 2021). Çelik et al. study Wasserstein distance from discrete distributions to independence models (Çelik et al., 2021), while Wasserstein conditional independence tests have finite-sample analyses (Warren, 2024; Neykov et al., 2024). We retain this objective, control the distance to the full family through local compression, and require the candidate's error upper bound to lie strictly below the original family's error lower bound.

Park, Balakrishnan, and Wasserman have proposed a framework for the relative fit of candidates and model families (Park et al., 2026). The IPM comparison in Proposition 6 of the published version includes Wasserstein-1: when the expectations exist and an exact optimal witness is attainable, a fixed model pair, or a model pair conditioned on independent fitting, satisfies a population expectation inequality with a factor of 3; finite-sample guarantees still require test calibration. Under the stronger access model of full observation-wise product probabilities, we construct a strict additive certificate for a specified Hamming metric. Sample-size comparisons must use the same access model.

2.2 Component Transport and Extensions of Feasible Couplings

Dusson et al. define finite-mixture transport and restricted couplings through distances between atoms; Definition 3.3 and Eq. (3.1) of preprint v1 use $p > 1$ and specified finite decompositions (Dusson et al., 2026). We use $p = 1$, and the source decomposition law need not be discrete, so Section 5.1 separately establishes the full-space upper bound. Delon–Desolneux use traces of Gaussian component covariances to control the relaxation of restricted second-order mixture transport relative to the full distance (Delon and Desolneux, 2020, Proposition 6). Our direct component transport likewise depends on the specified decompositions and generally differs from the distance between the full aggregate distributions.

Using this construction as a baseline, we enlarge the set of valid couplings through prefix aggregation, state lifting, and stopping trees. Stopping trees use exact single-block channels of the candidate, retain parent nodes, and optimize stopping masses. The specific guarantees are preservation of the full marginals, nested feasible sets, and integration with the unified TV hierarchy and common scoring event.

2.3 Concentration Tools and One-Sided Errors

Convex comparisons for bounded variables, binomial absolute deviations, and empirical Bernstein bounds are treated in (Bentkus, 2004; Hoeffding, 1956), (Berend and Kontorovich, 2013), and (Maurer and Pontil, 2009), respectively. Uniform CDF control is given in (Reeve, 2024), and the DKW–CVaR envelope of Thomas–Learned-Miller is given in (Thomas and Learned-Miller, 2019); the threshold optimization representation of CVaR originates in (Rockafellar and Uryasev, 2000), while one-sided CVaR concentration for unbounded light-tailed variables is separately studied in (Kolla et al., 2019). We incorporate the CVaR envelope into transport upper bounds to obtain locally adaptive endpoints. Guigues, Juditsky, and Nemirovski analyze asymmetric errors of optimal values in

stochastic optimization (Guigues et al., 2017). We combine label-frequency deviations, the opposite bias direction of the transport semidual, and the two-component bias constant to obtain a conditional detection bound without a logarithmic grid factor, and use a fixed population-optimal potential to jointly calibrate empirically feasible couplings.

3 Problem Definition and Structural Comparison Objective

3.1 Conditional Reference under a Frozen Representation

Fix the prediction condition c, encoder, deterministic recurrent model, and conditional law of future observations $\mu = \mu_c$, and access the full posterior probabilities for each observation. The latent state X has G blocks, with K_g categories in block g:

$$\mathcal{S}_g = [K_g] := \{1, \dots, K_g\}, \qquad \mathcal{S} = \prod_{g=1}^{G} \mathcal{S}_g, \qquad Q_o = \bigotimes_{g=1}^{G} q_{og}. \tag{1}$$

The common reference is $R = R_c = \mathbb{E}_{O\sim\mu_c} Q_O$, which may exhibit dependence after observations are marginalized out. We suppress the fixed condition c below and restore it explicitly when considering multiple conditions. Let $\mathrm{TV}(p, q) = \frac{1}{2}\sum_x |p(x) - q(x)|$ and $[x]_+ = \max\{x, 0\}$; the RSSM conditioning and probability conventions are given in Appendix B.6. We use the normalized Hamming distance, taking values in $[0, 1]$, and its associated first-order Wasserstein distance:

$$d_G(x, y) = \frac{1}{G}\sum_{g=1}^{G} \mathbf{1}\{x_g \neq y_g\}, \qquad W = W_{1,d_G}. \tag{2}$$

For product distributions $Q = \bigotimes_g q_g$ and $V = \bigotimes_g v_g$,

$$W(Q, V) = \kappa(Q, V) = \frac{1}{G}\sum_{g=1}^{G} \mathrm{TV}(q_g, v_g) = \frac{1}{2G}\sum_{g=1}^{G} \|q_g - v_g\|_1. \tag{3}$$

The proof is given in Appendix B.1. The original family and the candidate are, respectively,

$$\mathcal{F} = \left\{ \bigotimes_{g=1}^{G} u_g \right\}, \qquad P = \sum_{h=1}^{C} \pi_h V_h, \qquad V_h = \bigotimes_{g} v_{hg}. \tag{4}$$

Here u_g, v_{hg} are categorical probability vectors, and π gives the probability weights of the candidate components. A factorized neural prior at a given condition belongs to $\mathcal{F}$, so the family-level lower bound also constrains the actual prior.

3.2 Structural Replacement Criterion

The optimal error of the original family is $\Delta(R) = \inf_{Q\in\mathcal{F}} W(R, Q)$. On the joint event $L \leq \Delta(R)$ and $W(R, P) \leq U$, the inequality $U < L$ proves that the candidate outperforms every $Q \in \mathcal{F}$, and $L - U$ is a valid lower bound on the true improvement. True improvement, the population certificate margin, and the approximation-loss decomposition are given in Equations (5)–(6) of Appendix B.9; the comparison with the distance to the product of marginals is given in Appendix B.7.

3.3 Certification Routes

The lower bound compresses posterior probabilities using local codebooks to control the optimal error of the original product family. The upper bound takes direct component transport as its baseline: prefix aggregation and state lifting form a unified TV hierarchy; a stopping tree provides a separate branch that can be further combined with the hierarchy. Each construction preserves the same full

reference and candidate marginals. The common event in Section 6 yields endpoints that certify replacement when $U < L$. Throughout, G counts categorical blocks. Tables of dimensions and notation are given in Appendices B.5 and B.8, and auxiliary statistical endpoints in Appendix G.

4 Local Posterior Compression and Lower Bounds for the Independent Family

4.1 Local Codebooks That Preserve Category Information

For a nonempty coordinate set S, W_S uses the Hamming distance normalized by $|S|$, $\mathcal{F}_S$ is the within-group product family, and κ_S is the within-group version of (3). Write $\Delta_S(F) = \inf_{Q\in\mathcal{F}_S} W_S(F,Q)$. The complete conventions are given in Appendix B.6.

Choose a family $\mathcal{B}$ of nonempty, possibly overlapping groups, write $b_B = |B|$, and fix weights $w_B \geq 0$ satisfying

$$\sum_{B\ni g} w_B = 1 \quad (g = 1,\dots,G), \qquad \alpha_B = \frac{w_B b_B}{G}, \qquad \sum_B \alpha_B = 1. \tag{10}$$

This exact fractional cover counts each coordinate's cost exactly once. Each group has C_B product centers U_{Bz} on $\mathcal{S}_B$, with labels $A_B(O) \in [C_B]$ assigned from posterior probabilities by a fixed rule. The groups, weights, centers, and assignments are frozen before scoring, and the within-group distance uses all category probabilities. Write $r_{Bz} = \Pr(A_B(O) = z)$ and $\boldsymbol{r}_B = (r_{Bz})_{z\in[C_B]}$, and define

$$\begin{aligned} e_B &= \mathbb{E}\kappa_B(Q_O^B, U_{B,A_B(O)}), \\ D_B &= \inf_{u_B\in\mathcal{F}_B} \sum_z r_{Bz}\kappa_B(U_{Bz}, u_B), \\ e &= \sum_B \alpha_B e_B, \qquad D = \sum_B \alpha_B D_B, \\ \tau &= \sum_B \alpha_B \sqrt{\frac{\log C_B}{2b_B}}. \end{aligned} \tag{11}$$

Here e_B is the average compression distortion, and D_B is the average distance from the codebook to a single product distribution within the group. The value D_B can be computed coordinate by coordinate through weighted total-variation 1-median problems on the categorical simplices, without enumerating joint states. The two-center closed form and degenerate cases are given in Appendix E.1.

4.2 Approximation Guarantees Without an Intergroup Dependence Penalty

Theorem 1. Local compression lower bound and its approximation loss. Under the exact cover in (10), define $\ell = [D - e - \tau]_+$. Then

$$0 \leq \Delta(R) - \ell \leq 2e + \tau. \tag{14}$$

This result permits arbitrary intergroup dependence and allows the centers of overlapping groups to assign different probabilities on shared coordinates.

The proof is given in Appendix E.4. If only the coordinatewise coverage budget $\sum_{B\ni g} w_B \leq 1$ holds, ℓ remains valid, with an additional uncovered contribution to the approximation loss; the complete statement is given in Appendix E.8.

Two-center codebooks can use the binomial sharpening in Appendix E.6. This improves the constant in the same codebook lower bound, and the subsequent experiments use this version.

5 Full-Space Candidate Upper Bounds and Feasible TV Couplings

5.1 Direct Transport Satisfying the Weight Constraints

For each posterior index o and candidate component h, compute $c_h(o) = \kappa(Q_o, V_h)$. Define direct transport by

$$\mathcal{T} = \inf_{a^{\mathrm{tr}}} \mathbb{E} \sum_{h=1}^{C} a_h^{\mathrm{tr}}(O) c_h(O), \quad a_h^{\mathrm{tr}} \geq 0, \quad \sum_h a_h^{\mathrm{tr}} = 1, \quad \mathbb{E} a_h^{\mathrm{tr}} = \pi_h. \tag{22}$$

Direct transport satisfies $W(R, P) \leq \mathcal{T}$; see Appendix F.3 for the proof. Equation (22) uses the full G-dimensional cost and satisfies the candidate weights exactly.

$\mathcal{T}$ is a basic feasible upper bound under the specified posterior decomposition. For the finite source representation F in Section 5.3, its corresponding notation is $\mathcal{T}_{\varnothing}(F, P)$, satisfying $W(F, P) \leq \mathcal{T}_{\varnothing}(F, P)$. The subsequent constructions tighten the upper bound by enlarging the class of feasible couplings; certification relies on the validity of the resulting upper bound.

5.2 Prefix Coupling and Mixture Cancellation

Prefix coupling preserves the conditional dependence of the source mixture and exploits cancellation between probability deviations in opposite directions. It serves as the tail module of the unified hierarchy in Section 5.3. Fix a coordinate order. For any distribution F and product template $V = \bigotimes_g v_g$, define

$$\mathsf{Pref}(F, V) = \frac{1}{G} \sum_g \mathbb{E}_{X_{<g} \sim F} \, \mathrm{TV}(F_g(\cdot \mid X_{<g}), v_g). \tag{23}$$

Zero-probability prefixes contribute zero, so their conditional distributions may be defined arbitrarily. For a finite product mixture $F = \sum_{i=1}^{N_{\mathrm{mix}}} \alpha_i Q_i$, $Q_i = \bigotimes_g q_{ig}$, with $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$, we have

$$W(F, V) \leq \mathsf{Pref}(F, V) \leq \sum_i \alpha_i \kappa(Q_i, V). \tag{26}$$

Thus, cancellation within the conditional mixture can tighten the componentwise coupling cost. This cost depends on the coordinate order and generally differs from the exact Wasserstein distance. The computational form, prefix responsibilities, and proofs are given in Appendix F.4, and the exact cancellation quantity in F.6. Prefix enumeration may be exponentially expensive; safe computation conventions required for certification are given in F.7.

5.3 A Unified TV Hierarchy with Partial State Lifting and Tail Aggregation

Fix a finite product-mixture representation $F = \sum_{i=1}^{N} a_i Q_i$, with $a_i \geq 0$ and $\sum_i a_i = 1$, and the original candidate $P = \sum_h \pi_h V_h$. For scoring, take $N = n, a_i = 1/n, Q_i = Q_{O_i}$, so that $F = \widehat{R}_n$; this does not require the true observation law to have known finite support. The target label h below belongs to the transport plan, and its marginal mass always satisfies the original candidate weights. All costs remain normalized by the original total number of blocks G.

Common marginals.

For a set A of original coordinates, define the lifted source and target masses by

$$\alpha_A(i, x_A) = a_i Q_i^A(x_A), \qquad \beta_A(h, y_A) = \pi_h V_h^A(y_A). \tag{28}$$

Below, $\mathrm{Cpl}(\alpha_A, \beta_A)$ denotes the nonnegative plans with these two marginals. Appendix F.1 defines the comparison construction $\mathcal{T}_A$ using sourcewise tail coupling, whose empty-set level is direct

transport. It is used to establish the hierarchy's comparison relations and is no longer treated as a parallel main method.

Unified cost.

Fix a permutation of the original coordinates and relabel them as $1, \dots, G$. Let $A_s = \{1, \dots, s\}$ and $D_s = \{s+1, \dots, G\}$. Here, $s \in \{0, \dots, G\}$ is the expansion level. The empty coordinate space $\mathcal{S}_\varnothing$ is taken to be a singleton, the empty marginal has mass 1, and the empty-tail cost is 0; no normalized metric involving $1/|\varnothing|$ is defined. For any feasible $\Xi \in \mathrm{Cpl}(\alpha_{A_s}, \beta_{A_s})$, the mass of a target prefix $t = (h, y_{A_s})$ and its conditional source-tail distribution are, respectively,

$$\omega_t^{\mathrm{tar}} = \pi_h V_h^{A_s}(y_{A_s}), \qquad F_t^{\Xi} = \frac{\sum_{i, x_{A_s}} \Xi(i, x_{A_s}, h, y_{A_s}) Q_i^{D_s}}{\omega_t^{\mathrm{tar}}}. \tag{29}$$

Only positive-mass t are defined. F_t^{Ξ} is the actual tail mixture induced by the current plan and is generally not a product distribution. Define the tail prefix cost with the original normalization retained by

$$\mathsf{Pref}_{>s}^{(G)}(F, V_h) = \frac{1}{G} \sum_{g=s+1}^{G} \mathbb{E}_F \, \mathrm{TV}\left(F_g(\cdot \mid X_{s+1:g-1}), v_{hg}\right), \tag{30}$$

with value zero when $s = G$. Combining this with Section 5.2, the unified feasible cost and its optimum are

$$\begin{aligned} \mathcal{C}_s(\Xi) &= \sum_{i, x_{A_s}, h, y_{A_s}} \Xi(i, x_{A_s}, h, y_{A_s}) \frac{1}{G} \sum_{g \le s} \mathbf{1}\{x_g \neq y_g\} \\ &\quad + \sum_t \omega_t^{\mathrm{tar}} \mathsf{Pref}_{>s}^{(G)}(F_t^{\Xi}, V_h), \\ \mathcal{H}_s(F, P) &= \min_{\Xi \in \mathrm{Cpl}(\alpha_{A_s}, \beta_{A_s})} \mathcal{C}_s(\Xi). \end{aligned} \tag{31}$$

Theorem 2 (Unified TV coupling hierarchy). Under the finite product representation above,

$$\boxed{\begin{gathered} W(F, P) = \mathcal{H}_G(F, P) \le \mathcal{H}_{G-1}(F, P) \le \cdots \le \mathcal{H}_0(F, P) \le \mathcal{T}_\varnothing(F, P), \\ \mathcal{H}_s(F, P) \le \mathcal{T}_{A_s}(F, P). \end{gathered}} \tag{32}$$

The proof is given in Appendix F.2.

When $s = 0$, $\Xi = \lambda = (\lambda_{ih})$ need only have row sums a_i and column sums π_h. Thus, $\mathcal{H}_0$ optimizes $\sum_{h:\pi_h>0} \pi_h \mathsf{Pref}(\sum_i \lambda_{ih} Q_i/\pi_h, V_h)$ under precisely these constraints (zero-weight columns contribute zero). The original direct transport and prefix coupling therefore become basic constructions within the same main approach. Except at the full level, $\mathcal{T}_A$ and $\mathcal{H}_s$ still depend on the given source representation; the notation F, P does not assert decomposition invariance.

On finite spaces, the objective can be formulated as an LP. The practical implementation retains verified safe costs; limitations of approximate optimization are discussed in Appendix F.7.

5.4 Stopping-Tree Transport Retaining Parent Nodes

The stopping tree improves the direct-transport upper bound by exactly splitting the frozen candidate's product components and jointly selecting stopping masses and a transport plan. Construct a finite tree for each root V_h, denoting its root by ρ_h, node set by $\mathcal{N}_{\mathscr{T}}$, child set by $\mathrm{ch}(v)$, and internal-node set by $\mathcal{N}_{\mathscr{T}}^{\mathrm{int}}$. The product templates satisfy $T_{\rho_h} = V_h$, and every internal node decomposes exactly as $T_v = \sum_{w \in \mathrm{ch}(v)} \beta_{w|v} T_w$, with nonnegative child weights summing to 1; the single-block channel construction is given in Appendix F.5.

Denote the incoming, stopping, and continuing masses by f_v, σ_v, τ_v, respectively, and require

$$\begin{aligned} f_{\rho_h} &= \pi_h && (h \in [C]), \\ f_v &= \sigma_v + \tau_v, \quad \sigma_v, \tau_v \geq 0 && (v \in \mathcal{N}_{\mathscr{T}}), \\ f_w &= \beta_{w|v}\tau_v && (v \in \mathcal{N}_{\mathscr{T}}^{\mathrm{int}},\ w \in \mathrm{ch}(v)). \end{aligned} \tag{35}$$

For leaves, set $\tau_v = 0$. Let $\Sigma_{\mathscr{T}}$ be the set of stopping weights allowed by these constraints. The candidate then remains unchanged:

$$\sum_v \sigma_v T_v = P. \tag{36}$$

For a finite product mixture $F = \sum_{i=1}^N a_i Q_i$ ($a_i \geq 0,\ \sum_i a_i = 1$), define

$$\mathcal{T}_{\mathscr{T}}(F, P) = \min_{\substack{\gamma_{iv} \geq 0,\ \sum_v \gamma_{iv} = a_i \\ \sigma_v = \sum_i \gamma_{iv},\ \sigma \in \Sigma_{\mathscr{T}}}} \sum_{i,v} \gamma_{iv} \kappa(Q_i, T_v). \tag{38}$$

Proposition 2 (Validity and nestedness of stopping transport). With node templates and splitting proportions fixed,

$$W(F, P) \leq \mathcal{T}_{\mathscr{T}}(F, P) \leq \mathcal{T}_{\varnothing}(F, P). \tag{39}$$

Here, $\mathcal{T}_{\varnothing}(F, P)$ is the direct-transport cost under the same source representation. If $\mathscr{T}'$ is obtained by adding children while retaining all old nodes, templates, and channels, then $\mathcal{T}_{\mathscr{T}'} \leq \mathcal{T}_{\mathscr{T}}$. The proof is given in Appendix F.5, and the LP implementation in Appendix F.8.

Combining with the hierarchy. Stopping-tree transport and the prefix value $\mathcal{H}_0$ admit no universal ordering (Appendix I.5). Replace the target representation of the hierarchy by a valid stopping representation (σ_v, T_v), denote the resulting cost by $\mathcal{H}_s(F; \sigma, T)$, and jointly optimize to obtain

$$\mathcal{H}_s^{\mathscr{T}}(F, P) = \min_{\sigma \in \Sigma_{\mathscr{T}}} \mathcal{H}_s(F; \sigma, T). \tag{40}$$

Then

$$\begin{gathered} W(F, P) = \mathcal{H}_G^{\mathscr{T}} \leq \cdots \leq \mathcal{H}_0^{\mathscr{T}} \leq \mathcal{T}_{\mathscr{T}}, \qquad \mathcal{H}_s^{\mathscr{T}} \leq \mathcal{H}_s, \\ \mathcal{H}_s^{\mathscr{T}'} \leq \mathcal{H}_s^{\mathscr{T}} \quad \text{when } \mathscr{T}' \text{ retains all old nodes, templates, and channels,} \\ 0 \leq \mathcal{H}_s^{\mathscr{T}}(F, P) - W(F, P) \leq \frac{G - s}{G}. \end{gathered} \tag{41}$$

The proof is given in Appendix F.5. The full level is exact, but expansion may be exponentially expensive; enlarging only the target tree does not guarantee completeness. The implementation of the combined approach is given in Appendix F.8.

6 Independent Scoring and Structural Replacement Certification

The posterior, finite-component candidate, metric, conditional law, and codebook witnesses are determined from independent fitting data and frozen; conditional on the fitting results, the scoring rows are independent and identically distributed, with $n \geq 2$. For the same empirical reference $\widehat{R}_n = n^{-1} \sum_i Q_{O_i}$, let $C_n \geq W(\widehat{R}_n, P)$ be the minimum of the valid feasible costs already computed for the baseline, prefix couplings, hierarchies, stopping trees, and their combinations. Let $r_T^W = \sqrt{\log(2/\delta_T)/(2n)}$, and define

$$L = [\widehat{D} - \widehat{e} - r_e - \tau_* - q_n]_+, \qquad U_{\mathrm{coupling}} = \min\{1, C_n + r_T^W\}.$$

Here $\widehat{D}, \widehat{e}$ are the empirical codebook cost and mean distortion; the radii and τ_* are given in Appendix G.9. Costs must be evaluated deterministically or replaced by deterministic safe upper bounds, and the shared event covers coupling selection (Appendix G.10).

Theorem 4 (Joint reliability). Let U be the minimum of this upper endpoint and calibrated optional endpoints. Fix the budgets before scoring, write $\delta_L = \delta_H + \delta_e + \delta_r$, and denote the additional event budget by δ_{opt} (counting shared events only once). If $\delta_L + \delta_T + \delta_{\mathrm{opt}} \le \delta$, then

$$\Pr\{L \le \Delta(R),\ W(R,P) \le U\} \ge 1 - \delta. \tag{48}$$

Replacement is certified when $U < L$; otherwise, the output is insufficient evidence. The proofs, complete budgets, and detection conditions are given in Appendices G.5 and G.7–G.12.

7 Implementation and Scope of Applicability

Appendix H.3 constructs a two-component candidate and provides a positive certificate gap and finite-sample detection conditions; Appendix H.4 gives a capacity limit under equally sized independent modules. These results establish that the framework is implementable under explicit structural conditions, without guaranteeing certification for arbitrary RSSM posteriors.

8 Experiments

All current results use synthetic posteriors. Within each configuration, the methods share the reference, candidate, and scoring stream, retaining the original statistical budgets. Results use float64, and complete certification declares a numerical pass when $L - U > 10^{-9}$. The full protocols and scope of the numerical results are provided in Appendix J.4.

8.1 Complete Audits of Objects Learned and Then Frozen

For noisy module posteriors with $K = 4$, 512 independent fitting observations are used to learn the groups, two-center codebooks, and a two-component candidate, followed by freezing and independent scoring; the budget per run is $\delta = 0.05$. Table 1 selects successful strong-signal settings, failures with more modules, successful soft-posterior settings, and an independent negative control. Across all eight configurations, 60 of 80 repetitions achieved a numerical pass. The generative specification, learning rules, and complete results are provided in Appendix J.5.

Table 1. Complete synthetic audits of objects learned and then frozen: representative configurations, with 10 independent repetitions per row.

Condition	M	G	n	Numerical passes	Mean $L - U$
Strong	8	2048	4096	10/10	0.057847
Strong	32	4096	4096	0/10	-0.013715
Soft	8	2048	16384	10/10	0.003724
Independent	8	2048	4096	0/10	-0.048010

This experiment verifies the complete codebook audit of objects that are learned and then frozen, using the direct component baseline and strong statistical endpoints for the upper bound. Prefix couplings and stopping trees have not yet been run on these high-dimensional objects; their additional gains in complete certification remain to be verified.

8.2 Cost and Endpoint Gains from Feasible Couplings

We compare direct transport, full-prefix couplings, and stopping trees in three-block synthetic configurations, taking the minimum of the three costs as the actual combination. The baseline and new endpoints both retain the same strong statistical endpoints and original budgets. Table 2 fixes $n = 65536$ and summarizes all three perturbations and the correlated-reference control. The complete settings, other sample sizes, and endpoint definitions are provided in Appendices J.6–J.7.

Table 2. Three-block synthetic comparisons at $n = 65536$; all entries are means. Each of the first three rows has 20 repetitions, and the last row has 100.

Configuration	Direct cost	Prefix cost	Tree cost	U_{base}	U_{new}
Perturbation $\epsilon = 0$	0.050369	0.000573	0.000560	0.052774	0.007117
Perturbation $\epsilon = 0.03$	0.055346	0.010366	0.010558	0.057750	0.016923
Perturbation $\epsilon = 0.09$	0.065468	0.030429	0.030581	0.067827	0.036986
Correlated reference, product candidate	0.490000	0.320737	0.253889	0.490000	0.260446

The prefix is slightly tighter under nonzero perturbations, whereas the stopping tree is tighter for the correlated reference. All entries for the independent-reference control are 0.133333, with no tightening. These experiments verify only improvements in costs and upper endpoints: the product candidate in the correlated-reference row cannot outperform the entire product family, and the other comparisons were not accompanied by recomputation of the codebook lower bound. The actual combination does not solve the joint tree-augmented hierarchy. Dimension diagnostics and hierarchy consistency checks are provided in Appendices J.8 and I; results for trained RSSMs remain to be added.

9 Discussion

The certificate requires a consistent representation, reference conditional law, candidate, and metric, full posterior probability vectors for each observation, and independent scoring data. Adjacent trajectory frames or repeated generation under different conditions are not automatically independent.

Enlarging the feasible coupling set can lower costs without guaranteeing tighter statistical endpoints or certification. Codebook distortion, group-width penalties, scoring error, and high-dimensional prefix enumeration and optimization remain limiting factors. Evidence is restricted to synthetic audits and finite-space checks: certification gains from combining new couplings with the codebook lower bound, applicability to trained RSSMs, and rigorous numerical certification remain unverified. Distributional improvements do not directly imply policy gains.

10 Conclusion

We certify structural replacement using a codebook lower bound, prefix couplings, the unified TV hierarchy, stopping trees, and their combinations. After fitting and freezing the objects, independent scoring yields endpoints; $U < L$ certifies improvement over the entire product family, with valid combinations calibrated by a common statistical event. The appendices provide proofs, auxiliary endpoints, and conditional detection and capacity analyses.

Synthetic codebook audits support end-to-end feasibility; separate coupling comparisons verify cost and upper-endpoint reductions, subject to the limitations above.

AI Use Statement

Generative AI tools were used to assist with literature discovery and research brainstorming; to provide feedback on theoretical formulations, mathematical derivations, and proof development; to assist with experimental design and code development; and to draft, revise, and polish portions of the manuscript. AI-generated outputs were not treated as authoritative sources. AI-assisted mathematical arguments, experimental outputs, and manuscript text were reviewed by the author before inclusion. The author takes full responsibility for all claims, proofs, experimental results, citations, and other content in the final submission.

REFERENCES

Bentkus V. On Hoeffding's Inequalities. The Annals of Probability, 32(2): 1650–1673, 2004. DOI; Full text.

Berend D, Kontorovich A. A Sharp Estimate of the Binomial Mean Absolute Deviation with Applications. Statistics & Probability Letters, 83(4): 1254–1259, 2013. DOI.

Çelik T Ö, Jamneshan A, Montúfar G, Sturmfels B, Venturello L. Wasserstein Distance to Independence Models. Journal of Symbolic Computation, 104: 855–873, 2021. DOI; Preprint.

Delon J, Desolneux A. A Wasserstein-Type Distance in the Space of Gaussian Mixture Models. SIAM Journal on Imaging Sciences, 13(2): 936–970, 2020. Full text; DOI.

Dusson G, Ehrlacher V, Nouaime N. A Wasserstein-type Metric for Generic Mixture Models, Including Location-scatter and Group Invariant Measures. ESAIM: Control, Optimisation and Calculus of Variations, 32, article 19, 2026. DOI; Preprint v1. The definitions and equation numbers cited in Section 2.2 of this paper follow preprint v1.

Guigues V, Juditsky A, Nemirovski A. Non-asymptotic Confidence Bounds for the Optimal Value of a Stochastic Program. Optimization Methods and Software, 32(5): 1033–1058, 2017. DOI; Author-hosted published version.

Hafner D, Lillicrap T, Fischer I, et al. Learning Latent Dynamics for Planning from Pixels. Proceedings of the 36th International Conference on Machine Learning, PMLR 97: 2555–2565, 2019. Paper.

Hafner D, Lillicrap T, Norouzi M, Ba J. Mastering Atari with Discrete World Models. International Conference on Learning Representations, 2021. Paper.

Hoeffding W. On the Distribution of the Number of Successes in Independent Trials. The Annals of Mathematical Statistics, 27(3): 713–721, 1956. DOI.

Hoeffding W. Probability Inequalities for Sums of Bounded Random Variables. Journal of the American Statistical Association, 58(301): 13–30, 1963. DOI; Original text.

Kolla R K, Prashanth L A, Bhat S P, Jagannathan K. Concentration Bounds for Empirical Conditional Value-at-Risk: The Unbounded Case. Operations Research Letters, 47(1): 16–20, 2019. Paper.

Maurer A, Pontil M. Empirical Bernstein Bounds and Sample Variance Penalization. Proceedings of the 22nd Annual Conference on Learning Theory (COLT), 2009. Conference full text; Preprint.

McDiarmid C. On the Method of Bounded Differences. In: Siemons J (ed.). Surveys in Combinatorics, 1989. London Mathematical Society Lecture Note Series 141: 148–188. Cambridge University Press, 1989. Original source.

Neykov M, Wasserman L, Kim I, Balakrishnan S. Nearly Minimax Optimal Wasserstein Conditional Independence Testing. Information and Inference: A Journal of the IMA, 13(4): iaae033, 2024. DOI; Preprint.

Park B, Balakrishnan S, Wasserman L. Robust Universal Inference for Misspecified Models. Biometrika, 113(2), asaf070, 2026; online publication 2025. Published version; Preprint v4; DOI. Section 2.1 of this paper uses Proposition 6 in the published version; the corresponding result is Proposition 7 in preprint v4.

Peyré G, Cuturi M. Computational Optimal Transport: With Applications to Data Science. Foundations and Trends in Machine Learning, 11(5–6): 355–607, 2019. Paper.

Reeve H W J. A Short Proof of the Dvoretzky–Kiefer–Wolfowitz–Massart Inequality. arXiv:2403.16651, 2024. Paper.

Rockafellar R T, Uryasev S. Optimization of Conditional Value-at-Risk. Journal of Risk, 2(3): 21–41, 2000. DOI; Author-hosted full text.

Thomas P, Learned-Miller E. Concentration Inequalities for Conditional Value at Risk. Proceedings of the 36th International Conference on Machine Learning, PMLR 97: 6225–6233, 2019. Paper.

Warren A. Wasserstein Conditional Independence Testing. arXiv:2107.14184v2, 2024; first posted 2021. Preprint v2.

A Probability Tools and Supplementary Proofs

A.1 Bounded Differences and Transport–Entropy Control

Let μ be a product of m independent modules, and define $d_m(x, y) = m^{-1} \sum_{j=1}^{m} \mathbf{1}\{x_j \neq y_j\}$ on this module space, counting each module replacement as one coordinate change. For any function that is 1-Lipschitz with respect to d_m, changing one module changes its value by at most $1/m$. The bounded-differences moment-generating-function bound (McDiarmid, 1989) gives

$$\log \mathbb{E}_\mu e^{\lambda(f - \mathbb{E}_\mu f)} \leq \frac{\lambda^2}{8m}. \tag{A1}$$

For $\nu \ll \mu$, the KL variational inequality and Eq. (A1) give, for every $\lambda > 0$,

$$\mathbb{E}_\nu f - \mathbb{E}_\mu f \leq \frac{\mathrm{KL}(\nu\|\mu) + \log \mathbb{E}_\mu e^{\lambda(f - \mathbb{E}_\mu f)}}{\lambda} \leq \frac{\mathrm{KL}(\nu\|\mu)}{\lambda} + \frac{\lambda}{8m}.$$

When $0 < \mathrm{KL}(\nu\|\mu) < \infty$, take $\lambda = \sqrt{8m\,\mathrm{KL}(\nu\|\mu)}$. When KL is zero, let $\lambda \downarrow 0$; when KL is infinite, the following bound holds trivially. Taking the Wasserstein dual supremum over 1-Lipschitz functions then yields

$$W_{1,d_m}(\nu, \mu) \leq \sqrt{\frac{\mathrm{KL}(\nu\|\mu)}{2m}}. \tag{A2}$$

When the modules are the original individual blocks, take $m = G$. For any joint distribution F and product distribution V, this gives

$$W(F, V) \leq \sqrt{\mathrm{KL}(F\|V)/(2G)}. \tag{A3}$$

The KL variational inequality and transport–entropy control here are used in theoretical proofs; we do not introduce separate KL scoring or certification endpoints. The equally sized modules in Theorem 6 satisfy $d_G \leq d_M$, so an upper bound on the module distance also controls the original full-space distance. Lemma 1 applies the expected-maximum bound $\log C_0/\lambda + \lambda/(8b)$ to C_0 centered directions; optimization gives Eq. (13).

A.2 Binomial Constants

Two-step convex comparison. Let $b \geq 1$, let $Z_1, \ldots, Z_b$ be mutually independent random variables taking values in $[0, 1]$, and let $\varphi : [0, b] \to \mathbb{R}$ be convex. Write $p_g = \mathbb{E}Z_g$ and $\bar{p} = b^{-1} \sum_{g=1}^{b} p_g$. We first replace each variable with a Bernoulli variable of the same mean, and then average the success probabilities.

Step 1: endpoint replacement preserving the mean. Fix a coordinate g and let $S_0 = \sum_{h \neq g} Z_h$, so that S_0 and Z_g are independent. For each fixed value of S_0, the graph of a convex function on $[S_0, S_0 + 1]$ lies below the chord joining its endpoints. Thus,

$$\varphi(S_0 + z) \leq (1 - z)\varphi(S_0) + z\varphi(S_0 + 1), \qquad z \in [0, 1].$$

Set $z = Z_g$ and take expectations. Independence gives $\mathbb{E}[Z_g \mid S_0] = p_g$, hence

$$\begin{aligned} \mathbb{E}\varphi(S_0 + Z_g) &\leq \mathbb{E}\big[(1 - Z_g)\varphi(S_0) + Z_g\varphi(S_0 + 1)\big] \\ &= (1 - p_g)\mathbb{E}\varphi(S_0) + p_g\mathbb{E}\varphi(S_0 + 1). \end{aligned}$$

Take $B_g \sim \mathrm{Bern}(p_g)$ independently of all other variables. Then $\mathbb{E}B_g = p_g = \mathbb{E}Z_g$ and

$$\mathbb{E}\varphi(S_0 + B_g) = (1 - p_g)\mathbb{E}\varphi(S_0) + p_g\mathbb{E}\varphi(S_0 + 1).$$

Therefore, $\mathbb{E}\varphi(S_0 + Z_g) \leq \mathbb{E}\varphi(S_0 + B_g)$. We may choose $B_1, \ldots, B_b$ in advance to be mutually independent and independent of the entire Z vector. During coordinate-wise replacement, the Bernoulli variables already substituted and the variables not yet replaced remain mutually independent. The argument can therefore be repeated to obtain the first comparison:(Bentkus, 2004, Lemma 4.3)

$$\mathbb{E}\varphi\left(\sum_{g=1}^{b} Z_g\right) \leq \mathbb{E}\varphi\left(\sum_{g=1}^{b} B_g\right).$$

Step 2: averaging success probabilities while preserving the total mean. For $b \geq 2$, choose two Bernoulli variables $B_p \sim \mathrm{Bern}(p)$ and $B_q \sim \mathrm{Bern}(q)$. In this step, let S_0 denote the sum of the remaining $b-2$ variables. The two variables are independent of each other and of S_0. Expanding over the four possible combinations of their values and rearranging gives

$$\begin{aligned}\mathbb{E}\varphi(S_0 + B_p + B_q) = \mathbb{E}\varphi(S_0) + (p+q)\mathbb{E}[\varphi(S_0+1) - \varphi(S_0)] \\ + pq\,\mathbb{E}[\varphi(S_0+2) - 2\varphi(S_0+1) + \varphi(S_0)].\end{aligned}$$

The first two terms depend only on $p+q$. Convexity gives $\varphi(S_0+1) \leq [\varphi(S_0) + \varphi(S_0+2)]/2$, so the second difference in the last term is nonnegative. Moreover,

$$\left(\frac{p+q}{2}\right)^2 - pq = \frac{(p-q)^2}{4} \geq 0.$$

Replacing these two variables with Bernoulli variables that are independent of each other and of the remaining variables, both with success probability $(p+q)/2$, therefore preserves the total mean and does not decrease the expectation above.

To show that repeated averaging indeed converges to a common success probability, denote the probabilities after the mth averaging step by $p_g^{(m)}$, initialized at $p_g^{(0)} = p_g$. At each step, choose the largest and smallest success probabilities and replace both with their average. Write

$$A_m = \sum_{g=1}^{b} \left(p_g^{(m)} - \bar{p}\right)^2.$$

The average remains $\bar{p}$, and direct calculation gives

$$A_{m+1} = A_m - \frac{1}{2}\left(\max_g p_g^{(m)} - \min_g p_g^{(m)}\right)^2.$$

The nonnegative sequence A_m converges monotonically, so successive differences tend to zero, as does the range of the success probabilities. Since $\bar{p}$ always lies between the smallest and largest probabilities, each $p_g^{(m)} \to \bar{p}$. The expectation of φ of a finite Bernoulli sum is a polynomial in the success probabilities and is therefore continuous. Each averaging step also makes this expectation nondecreasing, so the limit gives the second comparison.(Bentkus, 2004, Eq. (3.3)) When $b = 1$, the second step is simply an equality. Letting $N \sim \mathrm{Bin}(b, \bar{p})$, the two steps together yield

$$\mathbb{E}\varphi\left(\sum_{g=1}^{b} Z_g\right) \leq \mathbb{E}\varphi\left(\sum_{g=1}^{b} B_g\right) \leq \mathbb{E}\varphi(N).$$

Take $\varphi(x) = |x - b\bar{p}|$. Since the total mean is unchanged throughout the comparisons,

$$\mathbb{E}\left|\sum_{g=1}^{b}(Z_g - p_g)\right| \leq \mathbb{E}|N - b\bar{p}|,$$

which is the binomial reduction required by Lemma 2. Only convexity is needed here, not strict convexity. The closed form of the binomial mean absolute deviation and its interval-wise maximizers are derived directly from the probability mass function in Appendix E.3.

Central maximizing term. Substituting $j = k + 1$ in Eq. (20), the maximization objective is proportional to $a(j)a(b + 1 - j)$, where $a(j) = j^j/(j - 1)!$. The successive ratio $a(j + 1)/a(j) = (1 + 1/j)^{j+1}$ decreases, so the maximum is attained at the center. Stirling's formula gives $\rho_b \sim 1/\sqrt{2\pi b}$. Finite-sample endpoints use the finite-b value of the central term, rounded upward.

Fair-binomial constant for soft modules. The quantity A_M in Appendix H.3 is the normalized absolute deviation in Eq. (55) at the fixed success probability $1/2$. It differs from ρ_b in Eq. (19), which takes the supremum over success probabilities. Let $N \sim \mathrm{Bin}(M, 1/2)$ and $k_0 = \lfloor M/2 \rfloor$. By symmetry and the binomial-coefficient identity

$$(k - M/2)\binom{M}{k} = \frac{M}{2}\left[\binom{M-1}{k-1} - \binom{M-1}{k}\right],$$

telescoping over $k = k_0 + 1, \ldots, M$, with the convention $\binom{M-1}{M} = 0$, yields

$$A_M = \frac{2}{M2^M}\sum_{k=k_0+1}^{M}(k - M/2)\binom{M}{k} = 2^{-M}\binom{M-1}{k_0} = \frac{1}{2}\frac{\binom{2\lfloor M/2 \rfloor}{\lfloor M/2 \rfloor}}{4^{\lfloor M/2 \rfloor}}.$$

This agrees with the known closed form of the binomial absolute deviation (Berend and Kontorovich, 2013, Eq. (1)). In particular, $A_1 = 1/2$. For an integer $m \geq 1$, write $c_m = \binom{2m}{m}/4^m$. Then

$$c_1 = \frac{1}{2}, \qquad \frac{c_{m+1}}{c_m} = \frac{2m+1}{2m+2} \geq \sqrt{\frac{m}{m+1}}.$$

After squaring, the last step is equivalent to $(2m + 1)^2 \geq 4m(m + 1)$. Induction gives $c_m \geq 1/(2\sqrt{m})$. Setting $m = \lfloor M/2 \rfloor$, for $M \geq 2$ we obtain $A_M \geq 1/(4\sqrt{m}) \geq 1/(2\sqrt{2M})$; for $M = 1$, the same bound holds directly. Together with $\rho_b \leq 1/(2\sqrt{b})$ from Appendix E.6,

$$b \geq M/a^2 \quad \Longrightarrow \quad 2aA_M - \rho_b \geq \frac{a}{\sqrt{2M}} - \frac{1}{2\sqrt{b}} \geq \left(\frac{1}{\sqrt{2}} - \frac{1}{2}\right)\frac{a}{\sqrt{M}} > 0.$$

Applying Stirling's formula to the central binomial coefficient above also gives $A_M \sim (2\pi M)^{-1/2}$, completing the conclusions on finite group width and asymptotic order in Appendix H.3.

A.4 Atoms at Quantiles

This section uses the notation of Appendices G.2–G.3: $p_{\mathrm{mix}} = \pi_1 \in (0, 1)$, and F is the distribution function of the random cost difference $D_c = c_1(O) - c_2(O)$. We use the generalized quantile $F^{-1}(q) = \inf\{t : F(t) \geq q\}$, $q \in (0, 1)$. The optimality condition for the concave transport dual objective is $F(t-) \leq p_{\mathrm{mix}} \leq F(t)$, allowing mass at an atom of the cost difference to be split according to Eq. (G5). DKW simultaneously controls the distribution function and its left limits, so the integrated envelope applies to both discrete and continuous costs.

B Proofs of Distance Identities and the Structural Objective

B.1 Transport Identity for Product Distributions (3)

The space defined in Eq. (2) has diameter $G^{-1}\#\{g : K_g \geq 2\} \leq 1$, equal to 1 when every $K_g \geq 2$. Throughout the paper, we use the diameter upper bound 1.

The coordinate-wise lower bound does not require independence of the coordinates at either endpoint: for arbitrary joint distributions F, F', we have $G^{-1}\sum_g \mathrm{TV}(F_g, F'_g) \le W(F, F')$. Product structure at both endpoints ensures that independently combined coordinate couplings have the required full marginals. General dependent distributions need not attain equality in this lower bound.

Proof of Eq. (3). We first establish the coordinate-wise cost that any coupling must incur. Take an arbitrary coupling (X, Y) with marginals Q, V. For each $g \in [G]$ and $k \in \mathcal{S}_g$,

$$\Pr(X_g = Y_g = k) \le \min\{q_g(k), v_g(k)\}.$$

Consequently,

$$\begin{aligned}\Pr(X_g \neq Y_g) &\ge 1 - \sum_{k\in\mathcal{S}_g} \min\{q_g(k), v_g(k)\} \\ &= \frac{1}{2}\sum_{k\in\mathcal{S}_g} |q_g(k) - v_g(k)| = \mathrm{TV}(q_g, v_g).\end{aligned}$$

The middle equality follows from $\min\{a, b\} = (a + b - |a - b|)/2$ and the fact that each marginal has mass 1. Summing according to Eq. (2) and taking the infimum over full couplings gives

$$W(Q, V) \ge \frac{1}{G}\sum_{g=1}^{G} \mathrm{TV}(q_g, v_g).$$

We now construct a full coupling attaining this lower bound. For each coordinate, write $m_g(k) = \min\{q_g(k), v_g(k)\}$ and $\varepsilon_g = 1 - \sum_k m_g(k) = \mathrm{TV}(q_g, v_g)$. First place mass $m_g(k)$ at each diagonal position (k, k). If $\varepsilon_g > 0$, couple the remaining masses at (k, ℓ), for $k, \ell \in \mathcal{S}_g$, as $[q_g(k) - m_g(k)][v_g(\ell) - m_g(\ell)]/\varepsilon_g$. If $\varepsilon_g = 0$, retain only the diagonal masses. This defines a coordinate coupling M_g. The remaining mass on each side sums to ε_g, so the marginals of M_g are exactly q_g, v_g. For any one category, at least one side has zero remaining mass, so all remaining mass is placed off the diagonal. Hence,

$$\Pr_{M_g}(X_g \neq Y_g) = \varepsilon_g = \mathrm{TV}(q_g, v_g).$$

Combine these coordinate pairs independently, defining

$$\Gamma(x, y) = \prod_{g=1}^{G} M_g(x_g, y_g).$$

Since both endpoints are product distributions, the full marginals satisfy

$$\begin{aligned}\sum_y \Gamma(x, y) &= \prod_{g=1}^{G}\sum_{y_g} M_g(x_g, y_g) = \prod_{g=1}^{G} q_g(x_g) = Q(x), \\ \sum_x \Gamma(x, y) &= \prod_{g=1}^{G}\sum_{x_g} M_g(x_g, y_g) = \prod_{g=1}^{G} v_g(y_g) = V(y).\end{aligned}$$

Thus, Γ is a valid full coupling, and

$$\mathbb{E}_\Gamma d_G(X, Y) = \frac{1}{G}\sum_{g=1}^{G} \Pr_{M_g}(X_g \neq Y_g) = \frac{1}{G}\sum_{g=1}^{G} \mathrm{TV}(q_g, v_g).$$

This gives the reverse inequality. Combined with the lower bound above, it proves Eq. (3); the last equality is the definition of TV. This completes the proof.

B.2 Productization and the Family-Distance Bound (7)

Proof of Eq. (7). We first show that productization does not increase Wasserstein distance under the metric in Eq. (2). For arbitrary joint distributions F, F', Eq. (3) applies to their productizations, while the coordinate-wise lower bound in Appendix B.1 applies to F, F' themselves. Thus,

$$\begin{aligned} W(\mathrm{Prod}(F), \mathrm{Prod}(F')) &= \frac{1}{G}\sum_{g=1}^{G} \mathrm{TV}(F_g, F'_g) \\ &\le W(F, F'). \end{aligned}$$

Here, both F, F' may have arbitrary coordinate dependencies; in particular, we may take $F = R, F' = P$.

Since $\mathrm{Prod}(R) \in \mathcal{F}$, the definition in Eq. (5) directly gives the right-hand inequality

$$\Delta(R) = \inf_{Q \in \mathcal{F}} W(R, Q) \le W(R, \mathrm{Prod}(R)).$$

To prove the left-hand inequality, take any $Q \in \mathcal{F}$. A product distribution satisfies $\mathrm{Prod}(Q) = Q$. By the contraction property just proved and symmetry of the distance,

$$W(Q, \mathrm{Prod}(R)) = W(\mathrm{Prod}(Q), \mathrm{Prod}(R)) \le W(Q, R) = W(R, Q).$$

The triangle inequality then gives

$$\begin{aligned} W(R, \mathrm{Prod}(R)) &\le W(R, Q) + W(Q, \mathrm{Prod}(R)) \\ &\le 2W(R, Q). \end{aligned}$$

This holds for every $Q \in \mathcal{F}$. Taking the infimum over Q and dividing by 2 proves the left-hand side of Eq. (7). The factor $1/2$ arises because each of the two distances in the triangle inequality is bounded by $W(R, Q)$. This completes the proof.

B.3 One-Hot Normalization Identity (8)

Proof of Eq. (8). Let $\mathrm{oh}_g(x_g) \in \{0, 1\}^{K_g}$ be the one-hot encoding of block g, equal to 1 only at position x_g and to 0 elsewhere. If $x_g = y_g$, the encodings coincide and their L^1 distance is zero. If $x_g \neq y_g$, the encodings differ at exactly two positions, each contributing 1. Therefore,

$$\| \mathrm{oh}_g(x_g) - \mathrm{oh}_g(y_g) \|_1 = 2\mathbf{1}\{x_g \neq y_g\}.$$

The full encoding $\mathrm{oh}(x)$ concatenates the block encodings, and the L^1 norm adds over blocks. Hence,

$$\begin{aligned} \| \mathrm{oh}(x) - \mathrm{oh}(y) \|_1 &= \sum_{g=1}^{G} \| \mathrm{oh}_g(x_g) - \mathrm{oh}_g(y_g) \|_1 \\ &= 2\sum_{g=1}^{G} \mathbf{1}\{x_g \neq y_g\} = 2G d_G(x, y). \end{aligned}$$

Dividing both sides by $2G$ yields Eq. (8). The factor 2 accounts for the two one-hot positions changed by a category change, and G normalizes by the original number of blocks; neither depends on the category counts K_g. This completes the proof.

B.4 Proposition 1: Irrelevant-Dimension Expansion

Proof. Projecting any coupling onto the original coordinates shows that the full cost is at least G/G' times the original cost. Conversely, use the original optimal coupling and couple the additional variables identically to attain equality. The same projection of any full product distribution gives the lower bound on the family distance; choosing Q_{ext} for the additional part gives the upper bound. Each source–candidate component cost is also multiplied by exactly G/G', so direct transport and the true improvement scale in the same way.

B.5 Stochastic Block Counts, Category Counts, and Dimension Calibration

The dimension used in concentration counts the conditionally independent categorical blocks in Eq. (1).

Symbol	Meaning	Where it appears
G	Total number of categorical stochastic blocks	Normalized Hamming cost and conditional-product concentration
K_g	Number of categories in block g	Simplex dimension, probability access, and within-group TV computation
$D_{\mathrm{flat}} = \sum_g K_g$	Length of the flattened probability or one-hot vector	Storage and computational cost; does not automatically provide independence
b_B	Number of stochastic blocks in group B	Local structural penalty $\sqrt{\log C_B / b_B}$ or ρ_{b_B}
M	Number of independent modules in the aggregate reference in a particular construction	Expressivity limitations of small candidates; distinct from G
n	Number of independent scoring observation rows	Statistical confidence radii; distinct from the coordinate count
C_B, C	Local codebook size and number of candidate components	Representation and optimization complexity; these quantities are distinct

For example, DreamerV2's 32 stochastic blocks with 32 categories each correspond to $G = 32$ and $D_{\mathrm{flat}} = 1024$.(Hafner et al., 2021)

If $\mathrm{oh}(x)$ denotes the blockwise one-hot encoding, then

$$d_G(x, y) = \frac{\|\mathrm{oh}(x) - \mathrm{oh}(y)\|_1}{2G}. \tag{8}$$

See Appendix B.3 for the proof.

Normalizing one-hot bit differences by D_{flat} multiplies the distance by $2G/D_{\mathrm{flat}}$. For uniform category counts $K_g \equiv K$, this factor is $2/K$. The one-hot bits within a categorical block are mutually exclusive, not K independent Bernoulli variables. Therefore, even when the cost units are adjusted accordingly, the concentration denominator cannot simply be changed from G to GK.

Proposition 1. Scaling under irrelevant-dimension expansion. Append the same r-block product distribution Q_{ext} to the reference, every posterior, and every candidate component. Write $G' = G + r$, $R' = R \otimes Q_{\mathrm{ext}}$, and $P' = P \otimes Q_{\mathrm{ext}}$. The new space is still normalized by all G' blocks, $W' = W_{1,d_{G'}}$, and $\Delta_{[G']}$ uses the product family on that space. Let $\mathcal{T}'$ be direct component transport for the expanded specified decompositions, normalized by G', and let $\eta' := \Delta_{[G']}(R') - W'(R', P')$. Then

$$W'(R', P') = \frac{G}{G'} W(R, P), \qquad \Delta_{[G']}(R') = \frac{G}{G'} \Delta(R), \qquad \mathcal{T}' = \frac{G}{G'} \mathcal{T}, \qquad \eta' = \frac{G}{G'} \eta. \quad (9)$$

See Appendix B.4 for the proof.

Duplicating sampled coordinates does not satisfy conditional independence; adding independent coordinates changes the normalized objective according to Eq. (9).

The following analysis uses the uniform block cost in Eq. (2).

B.6 Probabilities, Local Distances, and Notational Conventions

Conditioning and access model. The condition available at prediction time is $c = ((h_t, s_t), a_t)$, where t is the time index and h_t, s_t, a_t denote deterministic memory, the stochastic latent state, and the action, respectively. The deterministic recurrence gives $h' = H_{\theta_{\mathrm{dyn}}}(c)$. Fix the encoder, recurrent model, and conditional law of the future observation, $\mu_c = \mathcal{L}(O' \mid c)$. Once c is fixed, abbreviate O' as O and consistently denote the latent state by X. Their joint law is $\mu_c(do)Q_o(dx)$, and $R_c = \mathbb{E}_{O\sim\mu_c} Q_O$. The main text suppresses the fixed condition and writes $\mu = \mu_c$, $R = R_c$; formulas involving multiple conditions explicitly restore this input.

To generate a candidate, first sample a shared label according to π, then generate the blocks independently conditional on that label. To implement candidates across multiple conditions with a single network, that network must jointly output their components and weights. Throughout the paper, $\log$ denotes the natural logarithm, the entropy of a finite discrete variable is denoted by H, and we adopt the conventions $0 \log 0 = 0$ and $[x]_+ = \max\{x, 0\}$.

Integer intervals are consistently denoted by $[n] = \{1, \ldots, n\}$. On a finite space, $\mathrm{TV}(p, q) = \frac{1}{2}\sum_x |p(x) - q(x)|$ and $\mathrm{KL}(p\|q) = \sum_x p(x) \log[p(x)/q(x)]$. A term is zero when $p(x) = 0$, and is $+\infty$ when $p(x) > 0, q(x) = 0$. The states of $\mathrm{Bern}(p)$ are encoded as $\{0, 1\}$. Binary examples are obtained by relabeling $\{1, 2\}$, which does not change the distance.

For any nonempty coordinate set $S \subseteq [G]$, we consistently use

$$\mathcal{S}_S := \prod_{g\in S} \mathcal{S}_g, \qquad d_S(x_S, y_S) := \frac{1}{|S|} \sum_{g\in S} \mathbf{1}\{x_g \neq y_g\}, \qquad W_S := W_{1,d_S},$$

$$\mathcal{F}_S := \left\{ \bigotimes_{g\in S} u_g : \ u_g \text{ is a probability vector on } \mathcal{S}_g \right\},$$

$$\kappa_S\left(\bigotimes_{g\in S} u_g, \bigotimes_{g\in S} v_g \right) := \frac{1}{|S|} \sum_{g\in S} \mathrm{TV}(u_g, v_g), \qquad \Delta_S(F) := \inf_{Q\in\mathcal{F}_S} W_S(F, Q).$$

Here, F is any distribution on $\mathcal{S}_S$; κ_S is applied directly only to two product distributions. The coordinate marginal of a distribution F is denoted by F^S, or by F_S when the subscript has no other meaning, so $F^S = F_S$. The posterior notations $q_g(o) = q_{og}$ and $q_g(O) = q_{Og}$ are equivalent. In an empirical scoring representation, the index i in $Q_i := Q_{O_i}$ indexes observation rows; in a general finite-mixture representation, i indexes components. Neither is a coordinate index. On the full space, $W = W_{[G]}$ and $\Delta = \Delta_{[G]}$. When local coordinates are relabeled as $[b]$, d_b, W_{1,d_b}, and $\Delta_{[b]}$ are all normalized by b, not by G. The standard two-bit space uses $d_2(x, y) = \frac{1}{2}\sum_{g=1}^{2} \mathbf{1}\{x_g \neq y_g\}$, and its first-order transport distance is written as W_{1,d_2}.

B.7 Distance Comparison with the Product of Marginals

For any joint distribution F, define its productization as $\mathrm{Prod}(F) = \bigotimes_g F_g$. The optimal error of the original family satisfies

$$\frac{1}{2} W(R, \mathrm{Prod}(R)) \leq \Delta(R) \leq W(R, \mathrm{Prod}(R)). \quad (7)$$

See Appendix B.2 for the proof.

Equation (7) connects dependence to family-level error, but still involves the full joint distribution. Section 4 constructs lower bounds from local probability computations.

B.8 Complete Method and Notation Guide

Section 3.3 outlines the main-text methods: the codebook lower bound targets the original-family distance $\Delta(R)$, and all candidate upper bounds target the same full-space distance $W(R, P)$. Direct component transport provides the baseline; $\mathcal{H}_s$ denotes the unified hierarchy, $\mathcal{T}_{\mathscr{T}}$ denotes stopping-tree transport, and $\mathcal{H}_s^{\mathscr{T}}$ combines the two. There is no universal ordering between standalone $\mathcal{T}_{\mathscr{T}}$ and the prefix value $\mathcal{H}_0$; the combination retains both types of feasible scheme. The fully lifted level equals full-state transport. Once the node template and splitting proportions are fixed, the transport plan and stopping masses can be jointly optimized by linear programming. Small-range SAA, anchor, and CVaR are separately calibrated statistical endpoints; their applicability and budgets are given in Appendix G.11.

The notation m_{cert} denotes the certificate gap, M denotes the number of structural modules, and $\kappa(Q, V)$ is the transport cost in Eq. (3). The codebook costs D, D_B and cover weights w_B in Section 4 retain their original meanings; the reference quantity A_{ref} is defined in Appendix E.8. For TV, $\mathsf{Pref}(F, V)$ is the prefix coupling functional, Ξ is a lifted plan, $\mathcal{H}_s$ is the level-s cost for a specified finite source representation, and ω_t^{tar} is the target prefix mass.

Main objects and their scopes. The similarly formed approximation quantities below are not interchangeable. Hats indicate empirical computation, while $+/-$ indicate the endpoints or numerical safety directions specified individually in the text; these symbols do not themselves imply a coverage probability.

Symbol	Type, fixed meaning, and scope
R_c	Conditional aggregate reference; the cost range is denoted separately by B_{cost}
$D, D_B, D_{\mathrm{code}}$	Transport costs for a specified codebook decomposition
A_{ref}	Reference quantity in distance units
Prod; Cpl	Productization map; set of couplings with two specified marginals
a_h^{tr}	Transport assignment satisfying the candidate-weight constraints
$\mathcal{B}$	Family of coordinate groups
θ_{dyn}	RSSM deterministic recurrence parameters
$\mathcal{T}, \widehat{\mathcal{T}}, C_n$	Population/empirical component-transport optimal values for specified decompositions, and a valid feasible upper bound on the full empirical distance; the latter is not an alias for the former
τ, τ_2; τ_v	Structural remainders of local codebook compression; continuation mass at a stopping-tree node
$\mathscr{T}, T_v, \beta_{w\|v}$	Stopping tree, node product distribution, and child-node splitting proportion; see Section 5.4
$f_v, \sigma_v, \Sigma_{\mathscr{T}}$	Node arrival mass, stopping mass, and set of valid stopping weights

For finite probability vectors a, b, the coupling set $\mathrm{Cpl}(a, b)$ consists of nonnegative mass tables with row and column marginals a, b. Functions, local proof variables, and dummy summation indices are used within their declared scopes. When conditioning is restored or the space is changed, the corresponding probability law and normalization are restored as well.

B.9 True Improvement, Population Gap, and Approximation Loss

Define the optimal error of the original family, the candidate's true improvement, and the population certificate gap as

$$\Delta(R) = \inf_{Q \in \mathcal{F}} W(R, Q), \qquad \eta = \Delta(R) - W(R, P), \qquad m_{\mathrm{cert}} = \ell - u, \tag{5}$$

where $\ell \le \Delta(R)$ and $u \ge W(R, P)$. The identity

$$m_{\mathrm{cert}} = \eta - \big(\Delta(R) - \ell\big) - \big(u - W(R, P)\big) \tag{6}$$

decomposes the true improvement into the certificate gap and two approximation losses. Finite-sample endpoints L, U must additionally account for scoring error. On the joint event $L \le \Delta(R)$ and $W(R, P) \le U$, the inequality $L > U$ implies $W(R, P) < W(R, Q)$ for every $Q \in \mathcal{F}$, and $L - U$ is a valid lower bound on the candidate's true improvement.

C Current Reproduction Materials and Scope of Claims

The current configurations and random streams follow the v10 description. They constitute a new batch, not an exact reproduction of the earlier V0 random streams. Appendix J.5 runs complete codebook certification, whereas Appendices J.6–J.7 compare only valid coupling costs and upper endpoints for the population distance. Each comparison holds the frozen reference, candidate, scoring stream, and upper-endpoint budget fixed.

`learned_trials.csv` records all 80 complete codebook audits, including both successes and failures. `coupling_trials.csv` and `reference_trials.csv` record the costs and upper endpoints for 180 perturbation comparisons and 400 reference-structure controls. The three-block records do not include pass decisions for structural replacement. The `frozen/` directory stores learned groupings, codebooks, candidates, and frozen-object hashes. The `scores/` directory stores per-row costs, compression distortions, and assignments for the high-dimensional setting, or observation-row counts and tree-flow plans for the three-block setting. Full probability rows can be regenerated from the configurations, independent random seeds, code, and environment versions.

The current code, `learned_trials.csv`, and `summary.json` all use the label-frequency bound in Eq. (G1) for a_n. All high-dimensional configurations with groups use two centers, so this term is $B_0/(2\sqrt{n})$, and is zero when no valid group is present. This synchronization reuses the existing frozen objects and scoring data and updates only the bias diagnostics, without changing L, U, gaps, or pass decisions. The quantity S_n in Eq. (53) must use the synchronized a_n.

The high-dimensional audits allocate a total budget of 0.05 across eight event types. The three-block comparisons retain the numerical budget 0.05/7 for each type of upper-endpoint event. The five budgets sum to 0.05 x 5/7, with the remainder left unused. Budgets were not reallocated based on the results to reduce the radii. Different repetitions do not constitute a claim of simultaneous coverage over the entire batch.

The three-block experiments actually use the direct cost, the prefix level $\mathcal{H}_0$, the stopping tree, and the minimum of these three feasible costs; they do not solve the joint tree-augmented hierarchy $\mathcal{H}_s^{\mathcal{T}}$. Cost reductions and final upper-endpoint reductions are recorded separately, while small-amplitude SAA, anchor, and CVaR are identical across all comparison columns. The current manuscript does not recompute three-block structural replacement scores using the codebook lower bound, so these upper-endpoint comparisons provide no experimental evidence of additional $U < L$ certifications.

This round uses float64 and HiGHS. Marginal residuals, independent full-state OT comparisons, and endpoint-gain identities are used to check numerical consistency; they do not replace rigorous rounding-error envelopes. Positive gaps in Appendix J.5 are termed numerical passes. The paper's probability theorems apply to their mathematical definitions and to implementations with valid numerical envelopes.

Earlier V0 summaries and records of old endpoints from withdrawn interfaces are retained only in a separate working archive and are not part of the current certification evidence. Old finite-space

mechanism scripts and source materials are retained separately; rerunning the same scripts does not increase the number of independent cases.

D Empirical Support Obstruction

The common event does not control upward bias. Let the true $R = P$ be the product of G fair bits, with point-mass observation-wise posteriors. The empirical source F is supported on at most n states. For any coupling (X, Y), let $D = \mathbb{E}d_G(X, Y)$. The mismatch-bit vector together with X determines Y, so

$$G \log 2 = H(Y) \le H(X) + \sum_g h_{\mathrm{bin}}(\Pr(X_g \neq Y_g)) \le \log n + G h_{\mathrm{bin}}(D). \tag{D1}$$

Here, $h_{\mathrm{bin}}(p) = -p \log p - (1-p)\log(1-p)$. The full optimal cost is at most $1/2$, since coupling with an independent fair target attains $1/2$. When $\log n < G \log 2$,

$$W(F, P) \ge h^{-1}_{\mathrm{bin},[0,1/2]}\left(\log 2 - \frac{\log n}{G}\right). \tag{D2}$$

Even when the true error is zero, empirical transport can remain large. After adding independent flips at rate $\epsilon < 1/2$ to each point posterior, the true aggregate remains the fair product, and the empirical distribution is at distance at most ϵ from the original empirical distribution. The lower bound therefore decreases by only ϵ. Full support alone cannot eliminate this limitation.

E Supplementary Proofs for Local Codebooks

E.1 Two-Center Optimum

The local codebook in Section 4.1 satisfies, when $C_B = 2$,

$$D_B = \gamma_B \min\{r_{B1}, r_{B2}\}, \qquad \gamma_B = \kappa_B(U_{B1}, U_{B2}). \tag{12}$$

Proof of Equation (12). Fix a group B and let $C_B = 2$, so $r_{B1}, r_{B2} \ge 0$ and $r_{B1} + r_{B2} = 1$. Both centers U_{B1}, U_{B2} belong to $\mathcal{F}_B$. For any $u_B = \bigotimes_{g \in B} u_g \in \mathcal{F}_B$, applying the triangle inequality for total variation coordinatewise gives

$$\begin{aligned}
\gamma_B &= \frac{1}{b_B} \sum_{g \in B} \mathrm{TV}\big((U_{B1})_g, (U_{B2})_g\big) \\
&\le \frac{1}{b_B} \sum_{g \in B} \big[\mathrm{TV}\big((U_{B1})_g, u_g\big) + \mathrm{TV}\big(u_g, (U_{B2})_g\big)\big] \\
&= \kappa_B(U_{B1}, u_B) + \kappa_B(U_{B2}, u_B).
\end{aligned}$$

Both distances are nonnegative, and each weight is at least $\min\{r_{B1}, r_{B2}\}$. Therefore,

$$\begin{aligned}
&r_{B1}\kappa_B(U_{B1}, u_B) + r_{B2}\kappa_B(U_{B2}, u_B) \\
&\quad \ge \min\{r_{B1}, r_{B2}\}\,[\kappa_B(U_{B1}, u_B) + \kappa_B(U_{B2}, u_B)] \\
&\quad \ge \gamma_B \min\{r_{B1}, r_{B2}\}.
\end{aligned}$$

Taking the infimum over $u_B \in \mathcal{F}_B$ yields $D_B \ge \gamma_B \min\{r_{B1}, r_{B2}\}$.

For the reverse inequality, choose the center with the larger weight as u_B. When $r_{B1} \ge r_{B2}$, the choice $u_B = U_{B1}$ is a feasible product distribution with objective value

$$r_{B1}\kappa_B(U_{B1}, U_{B1}) + r_{B2}\kappa_B(U_{B2}, U_{B1}) = r_{B2}\gamma_B = \gamma_B \min\{r_{B1}, r_{B2}\}.$$

When $r_{B2} > r_{B1}$, choosing $u_B = U_{B2}$ similarly gives the objective value $r_{B1}\gamma_B$. Hence $D_B \le \gamma_B \min\{r_{B1}, r_{B2}\}$. Together with the preceding lower bound, this proves Equation (12), and the infimum is indeed attained at a codebook center.

When $r_{B1} = r_{B2} = 1/2$, both centers are optimal, without implying uniqueness of the optimum. If either weight is zero or the two centers coincide, the same argument gives $D_B = 0$. The proof uses only the triangle inequality for the within-group distance and the feasibility of the two centers; it does not require point-mass centers, binary categories, or independence between groups. This completes the proof.

E.2 Local Mixture Error Lemma

Proof. Coupling each component according to its label gives $W_S(F_{\text{code}}, Q) \le \sum_z r_z \kappa_S(U_z, Q)$; taking the infimum over Q proves the right-hand inequality. For the left-hand inequality, fix any product distribution $Q = \bigotimes_{g\in S} q_g$ and lift an optimal coupling of F_{code} and Q to (Z, X, Y) such that $X \mid Z = z \sim U_z$. For every center and coordinate, choose $f_{zg} : \mathcal{S}_g \to [0,1]$ satisfying $((U_z)_g - q_g) \cdot f_{zg} = \mathrm{TV}((U_z)_g, q_g)$, and write

$$J_z(Y) = \frac{1}{b} \sum_{g\in S} \left(f_{zg}(Y_g) - \mathbb{E}_{q_g} f_{zg} \right).$$

Since $f_{zg}(X_g) - f_{zg}(Y_g) \le \mathbf{1}\{X_g \ne Y_g\}$, the difference between the component cost and the coupling cost is at most a centered term at the X endpoint plus $J_Z(Y)$. The former averages to zero under $X \mid Z$. Thus,

$$\sum_z r_z \kappa_S(U_z, Q) - W_S(F_{\text{code}}, Q) \le \mathbb{E} \max_{z \le C_0} J_z(Y) \le \sqrt{\frac{\log C_0}{2b}}.$$

The final step uses only coordinate independence under $Y \sim Q$. For fixed z, each coordinate term has range width at most $1/b$, and Hoeffding's moment-generating-function bound for bounded variables gives $\log \mathbb{E}_Q e^{\lambda J_z} \le \lambda^2/(8b)$.(Hoeffding, 1963, Eq. (4.16)) Therefore,

$$\mathbb{E}_Q \max_{z \le C_0} J_z \le \inf_{\lambda > 0} \left\{ \frac{\log C_0}{\lambda} + \frac{\lambda}{8b} \right\} = \sqrt{\frac{\log C_0}{2b}}.$$

For $C_0 = 1$, take $\lambda \downarrow 0$. Independence between different J_z is unnecessary, and the lifted conditional law $Y \mid Z$ is not assumed to remain a product distribution. Taking the infimum over Q then proves the left-hand inequality.

E.3 Binomial Sharpening

Proof. Both J_z have mean zero. Let $Z_g = (f_{1g}(Y_g) - f_{2g}(Y_g) + 1)/2$. Then the $Z_g \in [0,1]$ are independent, and

$$\mathbb{E} \max(J_1, J_2) = \frac{1}{2}\mathbb{E}|J_1 - J_2| = \frac{1}{b}\mathbb{E}\left| \sum_{g=1}^{b} (Z_g - \mathbb{E}Z_g) \right|.$$

Write $\bar{p} = b^{-1} \sum_{g=1}^{b} \mathbb{E}Z_g$. With the means fixed, convex comparison with endpoint distributions replaces each Z_g by a Bernoulli variable of the same mean. With the total mean fixed, pairwise averaging of the success probabilities of independent Bernoulli variables does not decrease the expected absolute deviation above. General convex comparisons for these two steps are given in (Bentkus, 2004, Lemma 4.3 and Eq. (3.3)), respectively; a strictly convex version of the comparison at fixed total mean is given in (Hoeffding, 1956, Theorem 3). A direct convexity proof is provided in Appendix A.2. Consequently, $\mathbb{E} \max(J_1, J_2) \le b^{-1}\mathbb{E}|\operatorname{Bin}(b, \bar{p}) - b\bar{p}|$.

We use the classical identity for the binomial mean absolute deviation (e.g., (Berend and Kontorovich, 2013, Eq. (1))), deriving it directly from the probability mass function below. Let $N \sim \mathrm{Bin}(b, p)$ and $M \sim \mathrm{Bin}(b-1, p)$. First take $0 < p < 1$ and write $k = \lfloor bp \rfloor$. Using $j\binom{b}{j} = b\binom{b-1}{j-1}$ and the recursion for binomial tail probabilities gives

$$\begin{aligned}\mathbb{E}|N - bp| &= 2\mathbb{E}[(N - bp)\mathbf{1}\{N \geq k+1\}] \\ &= 2bp\{\Pr(M \geq k) - \Pr(N \geq k+1)\} \\ &= 2bp(1-p)\Pr(M = k).\end{aligned}$$

Thus, for $p \in [k/b, (k+1)/b]$ and $k \in \{0, \ldots, b-1\}$, with the endpoints handled by continuity,

$$\frac{1}{b}\mathbb{E}|N - bp| = 2\binom{b-1}{k}p^{k+1}(1-p)^{b-k}.$$

The maximum on this interval occurs at $p = (k+1)/(b+1)$; substitution yields Equation (20). At $p = 0, 1$, the absolute deviation is zero.

The maximum term in Equation (20) occurs at a central position, and $\rho_b \sim (2\pi b)^{-1/2}$; see Appendix A.2 for the derivation. For finite b, the worst-case parameter need not be $1/2$; for example, $\rho_2 = 8/27$. Take independent $Y_g \sim \mathrm{Bern}(p_*)$ and complementary directions $f_{1g}(y) = y$ and $f_{2g}(y) = 1 - y$, where $p_* = (k_* + 1)/(b+1)$ and k_* attains the maximum in Equation (20). This attains the stated value, making it a sharp bound for the specified class of two-direction maxima.

E.4 Proof of Theorem 1 and the Partial-Cover Extension

Proof. Let $S_B = \sum_z r_{Bz}U_{Bz}$. Under this specified decomposition, $D_{\mathrm{code}}(B; \{r_{Bz}, U_{Bz}\}) = D_B$. Coupling each posterior according to its assignment gives $W_B(R_B, S_B) \leq e_B$, and the distance to a fixed family of distributions is 1-Lipschitz. Lemma 1 therefore yields $\Delta_B(R_B) \geq D_B - e_B - \sqrt{\log C_B/(2b_B)}$. For any full product distribution Q and any coupling (X, Y), weighting the projected costs of the groups by α_B and applying Equation (10) gives

$$\sum_B \alpha_B W_B(R_B, Q_B) \leq \mathbb{E}d_G(X, Y). \tag{15}$$

Taking the infimum first over full couplings and then over Q yields $\ell \leq \Delta(R)$.

To control the loss of the lower bound, use the common-reference quantity A_{ref} defined in Equation (16) of Appendix E.8.

Form a full product distribution from the coordinatewise optimal u_g. Coupling each posterior to this distribution gives $\Delta(R) \leq A_{\mathrm{ref}}$. On the other hand, the triangle inequality and coordinatewise minimization give

$$\left| \frac{1}{b_B}\sum_{g \in B} A_{\mathrm{ref},g} - D_B \right| \leq e_B. \tag{17}$$

The exact cover implies $A_{\mathrm{ref}} \leq D + e$. If $D - e - \tau \geq 0$, then $\Delta(R) - \ell \leq 2e + \tau$; otherwise, $D < e + \tau$, and we likewise have $\Delta(R) \leq D + e < 2e + \tau$.

Partial cover. If $s_g \leq 1$, the left-hand side of Equation (15) is still no greater than the full cost, so the same ℓ remains valid. Equation (17) gives $A_{\mathrm{ref}} \leq D + e + r_{\mathrm{miss}}$; applying the same case analysis for the positive part yields the approximation-loss bound in Appendix E.8. The deterministic upper bound in Equation (18) follows by choosing the uniform probability vector at each coordinate.

E.5 Proof of Corollary 1 and the Partial-Cover Extension

Proof. For two-center groups, apply Lemma 2 within Lemma 1. A single-center group has only one centered direction, whose expectation is exactly zero, and therefore incurs no two-direction

penalty. The result then follows by the argument of Theorem 1. The binomial variance bound gives $\rho_b \le 1/(2\sqrt{b}) \le \sqrt{\log 2/(2b)}$, and hence $\tau_2 \le \tau$; with the same codebooks, assignments, and cover weights, $\ell_2 \ge \ell$. When every group has exactly two centers, $\tau_2 = \sum_B \alpha_B \rho_{b_B}$.

For a partial cover, treating the uncovered contribution as in E.4 gives the conclusion in Appendix E.6.

E.6 Binomial Sharpening for Two-Center Structures

When each local codebook has only two centers, the generic entropy bound for a finite maximum can be sharpened further. Define

$$\rho_b = \sup_{p\in[0,1]} \frac{1}{b}\mathbb{E}|\operatorname{Bin}(b,p) - bp|. \tag{19}$$

Lemma 2. Bound on the maximum of two directions. Let $Y_1, \dots, Y_b$ be independent, let f_{zg} map the space of Y_g to $[0,1]$, for $z = 1, 2$, and let $J_z = b^{-1}\sum_g (f_{zg}(Y_g) - \mathbb{E}f_{zg}(Y_g))$. Then

$$\mathbb{E}\max(J_1, J_2) \le \rho_b = \max_{k\in\{0,\dots,b-1\}} \frac{2\binom{b-1}{k}(k+1)^{k+1}(b-k)^{b-k}}{(b+1)^{b+1}}. \tag{20}$$

The proof is given in Appendix E.3, and the convex comparison for the binomial constant in Appendix A.2.

Corollary 1. Two-center local lower bound. Under the exact fractional cover in Equation (10) and the setup of Theorem 1, if every local codebook satisfies $1 \le C_B \le 2$, define

$$\tau_2 := \sum_B \alpha_B \begin{cases} 0, & C_B = 1, \\ \rho_{b_B}, & C_B = 2. \end{cases}$$

Here, C_B is the codebook size frozen before scoring, not the number of labels actually observed during scoring. Then

$$\ell_2 = [D - e - \tau_2]_+, \qquad 0 \le \Delta(R) - \ell_2 \le 2e + \tau_2. \tag{21}$$

The proof is given in Appendix E.5. With the same codebooks, assignments, and cover weights, $\tau_2 \le \tau$ and $\ell_2 \ge \ell$; when every group has exactly two centers, $\tau_2 = \sum_B \alpha_B \rho_{b_B}$.

Partial cover. If only $s_g := \sum_{B\ni g} w_B \le 1$ holds, the above ℓ_2 remains a valid lower bound, but the approximation loss becomes

$$0 \le \Delta(R) - \ell_2 \le 2e + \tau_2 + r_{\mathrm{miss}},$$

where r_{miss} is as defined in Equation (18). The version of Equation (21) without the uncovered contribution can be used only under an exact cover or when $r_{\mathrm{miss}} = 0$ is established separately.

E.7 Auxiliary Lemma for Local Mixtures and the Product Family

Lemma 1. Error from a local mixture to the product family. For a nonempty coordinate set S, let $b = |S|$ and $F_{\mathrm{code}} = \sum_{z=1}^{C_0} r_z U_z$, where $U_z \in \mathcal{F}_S$, $r_z \ge 0$, and $\sum_z r_z = 1$. Define the cost associated with this specified decomposition by $D_{\mathrm{code}}(S; \{r_z, U_z\}) := \inf_{Q\in\mathcal{F}_S} \sum_z r_z \kappa_S(U_z, Q)$, abbreviated as D_{code} within this lemma. Then

$$D_{\mathrm{code}} - \sqrt{\frac{\log C_0}{2b}} \le \Delta_S(F_{\mathrm{code}}) \le D_{\mathrm{code}}. \tag{13}$$

The proof is given in Appendix E.2; only coordinate independence of the product target is required.

E.8 Uncovered Contribution and Approximation Guarantee for Partial Covers

To state the error for partial covers, define the common-reference quantities

$$A_{\mathrm{ref},g} = \inf_{u_g} \mathbb{E}\,\mathrm{TV}(q_{Og}, u_g), \qquad A_{\mathrm{ref}} = \frac{1}{G}\sum_g A_{\mathrm{ref},g}. \tag{16}$$

If only $s_g = \sum_{B \ni g} w_B \le 1$ holds, continue to define D, e, τ using $\alpha_B = w_B b_B / G$, in which case $\sum_B \alpha_B \le 1$. The same lower bound $\ell = [D - e - \tau]_+$ remains valid, but its approximation loss must include the uncovered contribution

$$r_{\mathrm{miss}} = \frac{1}{G}\sum_g (1 - s_g) A_{\mathrm{ref},g} \le \frac{1}{G}\sum_g (1 - s_g)(1 - 1/K_g). \tag{18}$$

The approximation loss then satisfies

$$0 \le \Delta(R) - \ell \le 2e + \tau + r_{\mathrm{miss}}.$$

The uncovered contribution enters only the approximation loss; it is not an additional subtraction needed to make the lower bound valid. The proof is given in Appendix E.4.

F Supplementary Constructions and Proofs for TV Couplings

F.1 Comparison Construction for Partial State Lifting

Use the finite product-mixture representations $F = \sum_i a_i Q_i$ and $P = \sum_h \pi_h V_h$ from Section 5.3, where $Q_i = \bigotimes_g q_{ig}$ and $V_h = \bigotimes_g v_{hg}$. For a subset $A \subseteq [G]$ of the original coordinates, let Q_i^A, V_h^A be the true marginals, and use the lifted masses at the two endpoints from Equation (28) of the main text. Write $A^c = [G] \setminus A$. The empty coordinate space is a singleton; the empty marginal mass and empty product are both 1, and the empty sum is 0.

Let $\mathrm{Cpl}(\alpha_A, \beta_A)$ denote the set of nonnegative plans with these two marginals. For any feasible $\Xi \in \mathrm{Cpl}(\alpha_A, \beta_A)$, the inner product $\langle c_A, \Xi \rangle$ denotes the finite sum $\sum c_A(i, x_A, h, y_A)\Xi(i, x_A, h, y_A)$ over all (i, x_A, h, y_A). Define

$$\begin{aligned} c_A(i, x_A, h, y_A) &= \frac{1}{G}\left[\sum_{g \in A} \mathbf{1}\{x_g \ne y_g\} + \sum_{g \notin A} \mathrm{TV}(q_{ig}, v_{hg})\right], \\ \mathcal{T}_A(F, P) &= \min_{\Xi \in \mathrm{Cpl}(\alpha_A, \beta_A)} \langle c_A, \Xi \rangle. \end{aligned} \tag{F1}$$

The empty-set level $\mathcal{T}_\varnothing$ is component transport under the specified source decomposition, and equals $\widehat{\mathcal{T}}$ when $F = \widehat{R}_n$. We now prove

$$W(F, P) = \mathcal{T}_{[G]}(F, P) \le \mathcal{T}_{A'}(F, P) \le \mathcal{T}_A(F, P) \quad (A \subseteq A'). \tag{F2}$$

Proof. Step 1: Every valid lifted plan generates valid transport on the full space. Take any $\Xi \in \mathrm{Cpl}(\alpha_A, \beta_A)$, whose marginal constraints are

$$\sum_{h, y_A} \Xi(i, x_A, h, y_A) = a_i Q_i^A(x_A), \qquad \sum_{i, x_A} \Xi(i, x_A, h, y_A) = \pi_h V_h^A(y_A).$$

For each label pair (i, h) and each $g \notin A$, choose an optimal single-block coupling $M_g^{ih} \in \mathrm{Cpl}(q_{ig}, v_{hg})$ such that

$$\sum_v M_g^{ih}(u,v) = q_{ig}(u), \qquad \sum_u M_g^{ih}(u,v) = v_{hg}(v),$$
$$\sum_{u,v} M_g^{ih}(u,v)\mathbf{1}\{u \neq v\} = \mathrm{TV}(q_{ig}, v_{hg}).$$

Define the full labeled joint distribution by

$$\Gamma_\Xi(i,x,h,y) = \Xi(i,x_A,h,y_A)\prod_{g\notin A} M_g^{ih}(x_g,y_g).$$

This corresponds to first sampling the labels and expanded states according to Ξ, then independently generating each unexpanded coordinate pair while keeping the same label pair (i,h). Summing directly over the target side gives the full labeled source marginal:

$$\begin{aligned}\sum_{h,y}\Gamma_\Xi(i,x,h,y) &= \sum_{h,y_A}\Xi(i,x_A,h,y_A)\prod_{g\notin A}\left[\sum_{y_g} M_g^{ih}(x_g,y_g)\right]\\ &= \left[\prod_{g\notin A} q_{ig}(x_g)\right]\sum_{h,y_A}\Xi(i,x_A,h,y_A)\\ &= Q_i^{A^c}(x_{A^c})\, a_i Q_i^A(x_A) = a_i Q_i(x).\end{aligned}$$

The last equality uses the product structure of component Q_i. Similarly, on the target side, the product structure of V_h gives

$$\begin{aligned}\sum_{i,x}\Gamma_\Xi(i,x,h,y) &= \left[\prod_{g\notin A} v_{hg}(y_g)\right]\sum_{i,x_A}\Xi(i,x_A,h,y_A)\\ &= V_h^{A^c}(y_{A^c})\,\pi_h V_h^A(y_A) = \pi_h V_h(y).\end{aligned}$$

Summing over the labels i,h, respectively, then gives $\Pr_{\Gamma_\Xi}(X=x) = F(x)$ and $\Pr_{\Gamma_\Xi}(Y=y) = P(y)$. Thus, the (X,Y) marginal of Γ_Ξ is a valid coupling on the full space. These equalities hold for each full state and therefore preserve all cross-coordinate dependence within F, P. For example, when $F = \frac{1}{2}\delta_{(0,0)} + \frac{1}{2}\delta_{(1,1)}$, expanding the first coordinate still uses the same source label i to generate the second coordinate, so the source state is always 00 or 11.

Computing the cost of this full coupling coordinate by coordinate yields

$$\begin{aligned}\mathbb{E}_{\Gamma_\Xi} d_G(X,Y) &= \sum_{i,x_A,h,y_A}\frac{\Xi(i,x_A,h,y_A)}{G}\left[\sum_{g\in A}\mathbf{1}\{x_g\neq y_g\} + \sum_{g\notin A}\sum_{u,v} M_g^{ih}(u,v)\mathbf{1}\{u\neq v\}\right]\\ &= \langle c_A, \Xi\rangle.\end{aligned}$$

Thus, $W(F,P) \le \langle c_A, \Xi\rangle$ for every valid Ξ. Minimizing over Ξ gives

$$W(F,P) \le \mathcal{T}_A(F,P).$$

Step 2: Enlarging the expanded set does not increase the optimal cost. Let $A \subseteq A' \subseteq [G]$, and write $B = A' \setminus A$. For any old plan $\Xi \in \mathrm{Cpl}(\alpha_A, \beta_A)$, use the optimal single-block couplings from Step 1 and define

$$\Xi'(i,x_{A'},h,y_{A'}) = \Xi(i,x_A,h,y_A)\prod_{g\in B} M_g^{ih}(x_g,y_g).$$

As in Step 1, first sum over the other endpoint of the added coordinates, then use the marginal constraints of the old plan to obtain

$$\sum_{h,y_{A'}} \Xi'(i,x_{A'},h,y_{A'}) = \left[\prod_{g\in B} q_{ig}(x_g)\right] \sum_{h,y_A} \Xi(i,x_A,h,y_A) = a_i Q_i^{A'}(x_{A'}),$$
$$\sum_{i,x_{A'}} \Xi'(i,x_{A'},h,y_{A'}) = \left[\prod_{g\in B} v_{hg}(y_g)\right] \sum_{i,x_A} \Xi(i,x_A,h,y_A) = \pi_h V_h^{A'}(y_{A'}).$$

Hence $\Xi' \in \mathrm{Cpl}(\alpha_{A'},\beta_{A'})$. Each added coordinate previously incurred $\mathrm{TV}(q_{ig},v_{hg})$ in c_A, and now incurs $\mathbf{1}\{x_g \neq y_g\}$ in $c_{A'}$; its expectation under M_g^{ih} is exactly the original TV cost. Keeping the same normalization factor $1/G$, we have

$$\begin{aligned}\langle c_{A'},\Xi'\rangle &= \sum_{i,x_A,h,y_A} \frac{\Xi(i,x_A,h,y_A)}{G}\left[\sum_{g\in A}\mathbf{1}\{x_g\neq y_g\} + \sum_{g\in B}\mathrm{TV}(q_{ig},v_{hg}) + \sum_{g\notin A'}\mathrm{TV}(q_{ig},v_{hg})\right]\\ &= \langle c_A,\Xi\rangle.\end{aligned}$$

Every old plan therefore corresponds to a valid plan at the new level with the same cost, so $\mathcal{T}_{A'} \leq \langle c_A,\Xi\rangle$. Minimizing over the old plan Ξ then gives

$$\mathcal{T}_{A'}(F,P) \leq \mathcal{T}_A(F,P).$$

Step 3: The fully expanded level equals the full Wasserstein distance. Step 1 has already established $W(F,P) \leq \mathcal{T}_{[G]}(F,P)$. Conversely, take any full-state coupling $\zeta \in \mathrm{Cpl}(F,P)$, augment each state pair with the component labels at both endpoints, and define

$$\Xi_{\mathrm{full}}(i,x,h,y) = \begin{cases}\zeta(x,y)\dfrac{a_i Q_i(x)}{F(x)}\dfrac{\pi_h V_h(y)}{P(y)}, & F(x)>0,\ P(y)>0,\\ 0, & F(x)P(y)=0.\end{cases}$$

The two ratios are the component responsibilities conditional on the full source state and the full target state, respectively. If either denominator is zero, the corresponding row or column of ζ is also zero, so this definition does not require evaluating $0/0$. For states with $F(x)>0$,

$$\begin{aligned}\sum_{h,y}\Xi_{\mathrm{full}}(i,x,h,y) &= \frac{a_i Q_i(x)}{F(x)}\sum_{y:P(y)>0}\zeta(x,y)\left[\sum_h \frac{\pi_h V_h(y)}{P(y)}\right]\\ &= \frac{a_i Q_i(x)}{F(x)}\sum_y \zeta(x,y) = a_i Q_i(x).\end{aligned}$$

When $F(x)=0$, every $a_iQ_i(x)=0$, so the same marginal equality still holds. For states with $P(y)>0$, we similarly have

$$\begin{aligned}\sum_{i,x}\Xi_{\mathrm{full}}(i,x,h,y) &= \frac{\pi_h V_h(y)}{P(y)}\sum_{x:F(x)>0}\zeta(x,y)\left[\sum_i \frac{a_i Q_i(x)}{F(x)}\right]\\ &= \frac{\pi_h V_h(y)}{P(y)}\sum_x \zeta(x,y) = \pi_h V_h(y),\end{aligned}$$

and both sides are zero when $P(y)=0$. Therefore, $\Xi_{\mathrm{full}} \in \mathrm{Cpl}(\alpha_{[G]},\beta_{[G]})$. Moreover, $\sum_{i,h}\Xi_{\mathrm{full}}(i,x,h,y) = \zeta(x,y)$ for every state pair, so adding labels preserves the state-pair masses and cost:

$$\begin{aligned}\langle c_{[G]}, \Xi_{\text{full}} \rangle &= \sum_{x,y} d_G(x,y) \sum_{i,h} \Xi_{\text{full}}(i,x,h,y) \\ &= \sum_{x,y} d_G(x,y)\zeta(x,y).\end{aligned}$$

Hence $\mathcal{T}_{[G]}(F,P) \le \sum_{x,y} d_G(x,y)\zeta(x,y)$. Minimizing over ζ gives $\mathcal{T}_{[G]}(F,P) \le W(F,P)$. Combining this with Steps 1 and 2 proves Equation (F2). The empty set and zero-weight components are already covered by the conventions above. This completes the proof.

Optimizing the dependence between the target label and source state here changes only the joint coupling, not the candidate generative distribution.

Projecting the lifted plan back to (i,h) and comparing the costs on the expanded coordinates also gives

$$0 \le \mathcal{T}_{\varnothing} - \mathcal{T}_A \le |A|/G. \tag{F3}$$

Proof of Equation (F3). Monotonicity in Equation (F2), applied to $\varnothing \subseteq A$, already gives $\mathcal{T}_A \le \mathcal{T}_{\varnothing}$, so the left-hand side is nonnegative. We now prove the upper bound on the right.

Step 1: Aggregate the expanded plan into a valid component-transport table. The finite coupling set is nonempty and compact, and the objective is linear, so an optimal plan Ξ^* attaining $\mathcal{T}_A$ exists. Define

$$\lambda_{ih} := \sum_{x_A, y_A} \Xi^*(i, x_A, h, y_A).$$

This aggregates all plan entries with the same source label i and target label h. The marginal constraints of Ξ^* give

$$\begin{aligned}\sum_h \lambda_{ih} &= \sum_{x_A} \sum_{h, y_A} \Xi^*(i, x_A, h, y_A) = a_i \sum_{x_A} Q_i^A(x_A) = a_i, \\ \sum_i \lambda_{ih} &= \sum_{y_A} \sum_{i, x_A} \Xi^*(i, x_A, h, y_A) = \pi_h \sum_{y_A} V_h^A(y_A) = \pi_h.\end{aligned}$$

Thus, λ is a valid component-transport table at the empty-set level. Only its feasibility is needed; it need not be optimal at that level.

Step 2: Compare costs entry by entry before and after expansion. The per-entry cost at the empty-set level is

$$c_{\varnothing}(i,h) = \frac{1}{G} \sum_{g=1}^{G} \mathrm{TV}(q_{ig}, v_{hg}).$$

Subtracting c_A in Equation (F1) cancels all terms for the unexpanded coordinates, giving

$$\begin{aligned}c_{\varnothing}(i,h) - c_A(i, x_A, h, y_A) &= \frac{1}{G} \sum_{g \in A} \left[\mathrm{TV}(q_{ig}, v_{hg}) - \mathbf{1}\{x_g \ne y_g\}\right] \\ &\le \frac{|A|}{G}.\end{aligned}$$

This uses only $\mathrm{TV}(q_{ig}, v_{hg}) \le 1$ and nonnegativity of the indicator; the difference inside an individual pair of brackets may be negative. Thus, $c_{\varnothing}(i,h) \le c_A(i, x_A, h, y_A) + |A|/G$ holds entry by entry.

Step 3: Weight by the masses of the optimal plan. Feasibility of the aggregated plan λ and the entrywise comparison above give

$$\begin{aligned}\mathcal{T}_{\varnothing} &\le \sum_{i,h} \lambda_{ih} c_{\varnothing}(i,h) \\ &= \sum_{i,x_A,h,y_A} \Xi^*(i,x_A,h,y_A) c_{\varnothing}(i,h) \\ &\le \sum_{i,x_A,h,y_A} \Xi^*(i,x_A,h,y_A) \left[c_A(i,x_A,h,y_A) + \frac{|A|}{G} \right] \\ &= \mathcal{T}_A + \frac{|A|}{G}.\end{aligned}$$

The last equality uses both $\langle c_A, \Xi^* \rangle = \mathcal{T}_A$ and $\sum_{i,x_A,h,y_A} \Xi^*(i,x_A,h,y_A) = 1$. Rearranging and combining with the nonnegativity established in the first paragraph proves Equation (F3). When $A = \varnothing$, the difference is zero. Plan entries corresponding to zero weights or zero-probability marginals have zero mass, and the same argument applies. This completes the proof.

Thus, expanding only a few coordinates does not guarantee elimination of an arbitrary constant slack. If $K_A = \prod_{g \in A} K_g$, a dense transport plan contains approximately NCK_A^2 variables. The original categorical states must be retained; binary-event projections cannot directly replace the full block costs.

F.2 Complete Proof of the Unified Hierarchy

Proof. Use $A_s = \{1, \dots, s\}$ and $D_s = \{s+1, \dots, G\}$ from Section 5.3, with all costs normalized by the original number of blocks G. Conditional distributions such as F_t^{Ξ} are defined only for target prefixes $t = (h, y_{A_s})$ satisfying $\omega_t^{\mathrm{tar}} > 0$. If $\omega_t^{\mathrm{tar}} = 0$, nonnegativity and the column marginal constraint imply that all Ξ entries in that column are zero, and the corresponding constructions below are also set to zero. The empty tail space is a singleton, with tail cost zero.

Step 1: Equation (31) is the cost of a valid full coupling. Fix a feasible plan Ξ at level s. First sample (i, x_{A_s}, h, y_{A_s}) according to Ξ, then generate $z = x_{D_s}$ from the source component's product tail distribution $Q_i^{D_s}$. For a fixed positive-mass target prefix t, Equation (29) gives

$$\sum_{i,x_{A_s}} \Xi(i, x_{A_s}, h, y_{A_s}) Q_i^{D_s}(z) = \omega_t^{\mathrm{tar}} F_t^{\Xi}(z).$$

Thus, conditional on target prefix t, the generated source tail has exactly distribution F_t^{Ξ}. Take the tail prefix coupling $J_t(z,w)$ from Section 5.2, where $w = y_{D_s}$, such that

$$\sum_w J_t(z,w) = F_t^{\Xi}(z), \qquad \sum_z J_t(z,w) = V_h^{D_s}(w),$$

and its tail cost, computed with the original denominator G, satisfies

$$\sum_{z,w} J_t(z,w) \frac{1}{G} \sum_{j \in D_s} \mathbf{1}\{z_j \neq w_j\} = \mathsf{Pref}_{>s}^{(G)}(F_t^{\Xi}, V_h).$$

Where $F_t^{\Xi}(z) > 0$, define the conditional kernel

$$K_t(w \mid z) := \frac{J_t(z,w)}{F_t^{\Xi}(z)}.$$

This normalizes row z of the coupling table J_t into a probability vector; choose any probability vector for a zero-mass row. Keep the already generated source tail z unchanged, and generate only the target tail according to $K_t(\cdot \mid z)$. The corresponding full labeled plan is

$$\Gamma_{s,\Xi}(i, x_{A_s}, z, h, y_{A_s}, w) = \Xi(i, x_{A_s}, h, y_{A_s}) Q_i^{D_s}(z) K_t(w \mid z).$$

When $F_t^{\Xi}(z) = 0$, every $\Xi(i, x_{A_s}, h, y_{A_s}) Q_i^{D_s}(z)$ in the nonnegative sum above is zero, so arbitrarily completing the kernel on that row does not affect the plan. Summing over the target side gives the full labeled source marginal

$$\begin{aligned}\sum_{h, y_{A_s}, w} \Gamma_{s,\Xi}(i, x_{A_s}, z, h, y_{A_s}, w) &= Q_i^{D_s}(z) \sum_{h, y_{A_s}} \Xi(i, x_{A_s}, h, y_{A_s}) \\ &= a_i Q_i^{A_s}(x_{A_s}) Q_i^{D_s}(z) = a_i Q_i(x).\end{aligned}$$

Summing over the source side gives the full labeled target marginal

$$\begin{aligned}\sum_{i, x_{A_s}, z} \Gamma_{s,\Xi}(i, x_{A_s}, z, h, y_{A_s}, w) &= \omega_t^{\mathrm{tar}} \sum_z F_t^{\Xi}(z) K_t(w \mid z) \\ &= \omega_t^{\mathrm{tar}} \sum_z J_t(z, w) \\ &= \pi_h V_h^{A_s}(y_{A_s}) V_h^{D_s}(w) = \pi_h V_h(y).\end{aligned}$$

Both sides are zero at zero-mass target prefixes. Summing over labels then recovers F, P, while the relationship between the source tail and source label is preserved throughout. By the cost identity for J_t,

$$\begin{aligned}\mathbb{E}_{\Gamma_{s,\Xi}} d_G(X, Y) = \frac{1}{G} &\sum_{i, x_{A_s}, h, y_{A_s}} \Xi(i, x_{A_s}, h, y_{A_s}) \sum_{j \le s} \mathbf{1}\{x_j \neq y_j\} \\ &+ \sum_t \omega_t^{\mathrm{tar}} \mathsf{Pref}_{>s}^{(G)}(F_t^{\Xi}, V_h) = \mathcal{C}_s(\Xi).\end{aligned}$$

Therefore, $W(F, P) \le \mathcal{C}_s(\Xi)$. Minimizing over feasible Ξ gives $W(F, P) \le \mathcal{H}_s(F, P)$.

Step 2: Construct the next-level plan and verify the marginals in Equation (F4). Let $s < G$, and denote the added coordinate by $g = s + 1$, its source state by $a \in \mathcal{S}_g$, and its target state by $b \in \mathcal{S}_g$. Here, a is a state value, distinct from the mixture weight a_i. For each positive-mass t, let

$$f_t(a) := (F_t^{\Xi})_g(a) = \frac{\sum_{i, x_{A_s}} \Xi(i, x_{A_s}, h, y_{A_s}) q_{ig}(a)}{\omega_t^{\mathrm{tar}}}.$$

q_{ig} is the coordinate distribution of a single source component, whereas f_t is the coordinate distribution after mixing the source masses throughout the target column. To preserve the cost of the first term in the old tail, take a maximal coupling M_t of f_t and v_{hg} satisfying

$$\sum_b M_t(a, b) = f_t(a), \qquad \sum_a M_t(a, b) = v_{hg}(b), \qquad \sum_{a,b} M_t(a, b) \mathbf{1}\{a \neq b\} = \mathrm{TV}(f_t, v_{hg}).$$

When $\omega_t^{\mathrm{tar}} > 0$ and $f_t(a) > 0$, define

$$\Xi'(i, x_{A_s}, a, h, y_{A_s}, b) = \Xi(i, x_{A_s}, h, y_{A_s}) q_{ig}(a) \frac{M_t(a, b)}{f_t(a)}. \tag{F4}$$

If $\omega_t^{\mathrm{tar}} = 0$ or $f_t(a) = 0$, set the corresponding Ξ' entries to zero. For positive-mass t, $f_t(a) = 0$ implies $\sum_{i, x_{A_s}} \Xi(i, x_{A_s}, h, y_{A_s}) q_{ig}(a) = 0$, so each summand is zero individually, consistently with the summation identities below.

The three factors in Equation (F4) are, in order, the old plan mass, the probability that the source component takes state a, and the conditional probability that the target takes state b given a. The final factor is normalized by the row sum of M_t: $\sum_b M_t(a,b)/f_t(a) = 1$. Consequently,

$$\sum_b \Xi'(i, x_{A_s}, a, h, y_{A_s}, b) = \Xi(i, x_{A_s}, h, y_{A_s}) q_{ig}(a).$$

Summing over target prefixes then gives the source marginal at the new level:

$$\begin{aligned}\sum_{h,y_{A_s},b} \Xi'(i, x_{A_s}, a, h, y_{A_s}, b) &= q_{ig}(a) \sum_{h,y_{A_s}} \Xi(i, x_{A_s}, h, y_{A_s}) \\ &= a_i Q_i^{A_s}(x_{A_s}) q_{ig}(a) = a_i Q_i^{A_{s+1}}(x_{A_s}, a).\end{aligned}$$

The target marginal follows from the definition of f_t. For $f_t(a) > 0$,

$$\begin{aligned}\sum_{i,x_{A_s}} \Xi'(i, x_{A_s}, a, h, y_{A_s}, b) &= \frac{M_t(a,b)}{f_t(a)} \sum_{i,x_{A_s}} \Xi(i, x_{A_s}, h, y_{A_s}) q_{ig}(a) \\ &= \omega_t^{\mathrm{tar}} M_t(a,b).\end{aligned}$$

When $f_t(a) = 0$, the corresponding row of M_t is also zero, so this equality still holds. Summing over a then gives

$$\sum_{i,x_{A_s},a} \Xi'(i, x_{A_s}, a, h, y_{A_s}, b) = \omega_t^{\mathrm{tar}} v_{hg}(b) = \pi_h V_h^{A_{s+1}}(y_{A_s}, b).$$

Therefore, $\Xi' \in \mathrm{Cpl}(\alpha_{A_{s+1}}, \beta_{A_{s+1}})$. In addition, for every old plan entry,

$$\sum_{a,b} \Xi'(i, x_{A_s}, a, h, y_{A_s}, b) = \Xi(i, x_{A_s}, h, y_{A_s}),$$

This shows that the old explicit cost on the first s coordinates is unchanged.

Step 3: Derive the remaining-tail mixture in Equation (F5). Fix a positive-mass t. For $f_t(a) > 0$, define the conditional distribution of the remaining source tail by

$$\begin{aligned}R_{t,a} &:= F_t^{\Xi}(X_{D_{s+1}} \in \cdot \mid X_g = a) \\ &= \frac{\sum_{i,x_{A_s}} \Xi(i, x_{A_s}, h, y_{A_s}) q_{ig}(a) Q_i^{D_{s+1}}}{\omega_t^{\mathrm{tar}} f_t(a)}.\end{aligned}$$

This uses the product structure of each source component. For conditioning events with $f_t(a) = 0$, choose any tail distribution as a version; because $M_t(a,b) = 0$, this choice does not affect the mixture below.

When $v_{hg}(b) > 0$, the new target prefix (t, b) has mass $\omega_t^{\mathrm{tar}} v_{hg}(b)$. Substituting Equation (F4) into the definition in Equation (29) at the new level gives

$$\begin{aligned}F_{t,b}^{\Xi'} &= \frac{\sum_{i,x_{A_s},a} \Xi'(i, x_{A_s}, a, h, y_{A_s}, b) Q_i^{D_{s+1}}}{\omega_t^{\mathrm{tar}} v_{hg}(b)} \\ &= \frac{1}{\omega_t^{\mathrm{tar}} v_{hg}(b)} \sum_{a: f_t(a)>0} \frac{M_t(a,b)}{f_t(a)} \sum_{i,x_{A_s}} \Xi(i, x_{A_s}, h, y_{A_s}) q_{ig}(a) Q_i^{D_{s+1}} \\ &= \sum_a \frac{M_t(a,b)}{v_{hg}(b)} R_{t,a}.\end{aligned}$$

That is,

$$F^{\Xi'}_{t,b} = \sum_a \frac{M_t(a,b)}{v_{hg}(b)} F^{\Xi}_t(X_{D_{s+1}} \in \cdot \mid X_g = a). \tag{F5}$$

These weights are nonnegative and sum to 1 because $\sum_a M_t(a,b) = v_{hg}(b)$. The denominator $f_t(a)$ in Equation (F4) is a row sum of the coupling table, used to generate the target conditional on the source state. The denominator $v_{hg}(b)$ in Equation (F5) is a column sum, used to aggregate the source tails conditional on the target state. When $v_{hg}(b) = 0$, the new target column has zero mass, and its conditional distribution is not used.

Step 4: Prove Equation (F6) by convexity, then compare the three parts of the cost. Fix the target label h corresponding to t, and temporarily write

$$\Phi_h(R) := \mathsf{Pref}^{(G)}_{>s+1}(R, V_h).$$

The version of Equation (24) for the remaining tail is

$$\Phi_h(R) = \frac{1}{2G} \sum_{j=s+2}^{G} \sum_{u,k} |R(u,k) - R(u) v_{hj}(k)|,$$

where u is a prefix of the remaining coordinates $s+2, \ldots, j-1$, and $R(u)$ and $R(u,k)$ are the corresponding true marginals; the empty prefix has mass 1. The expression inside each absolute value is linear in R, so Φ_h is convex. If the remaining tail is empty, the entire sum is zero, and the conclusion still holds.

Apply convexity to Equation (F5), multiply by $v_{hg}(b)$, and sum over positive-mass b to obtain

$$\begin{aligned}
\sum_b v_{hg}(b) \Phi_h(F^{\Xi'}_{t,b}) &\le \sum_{a,b} M_t(a,b) \Phi_h(R_{t,a}) \\
&= \sum_a \left[\sum_b M_t(a,b) \right] \Phi_h(R_{t,a}) \\
&= \sum_a f_t(a) \Phi_h(R_{t,a}).
\end{aligned}$$

Zero-mass rows and columns contribute zero. Restoring the notation gives

$$\sum_b v_{hg}(b) \mathsf{Pref}^{(G)}_{>s+1}(F^{\Xi'}_{t,b}, V_h) \le \sum_a f_t(a) \mathsf{Pref}^{(G)}_{>s+1}(F^{\Xi}_t(X_{D_{s+1}} \in \cdot \mid X_g = a), V_h). \tag{F6}$$

To relate this to the total cost, first separate out the initial term for the added coordinate $g = s+1$ in Equation (30), then condition the subsequent terms on $X_g = a$, obtaining the exact decomposition

$$\mathsf{Pref}^{(G)}_{>s}(F^{\Xi}_t, V_h) = \frac{1}{G} \operatorname{TV}(f_t, v_{hg}) + \sum_a f_t(a) \Phi_h(R_{t,a}).$$

Thus, the right-hand side of Equation (F6) is indeed the old tail cost with its first term removed. The total cost can be divided into the following three parts.

First, the old explicit cost on the first s coordinates is unchanged because summing over the added states (a,b) recovers the original table Ξ. Second, the explicit cost of the added coordinate is

$$\frac{1}{G} \sum_t \omega^{\mathrm{tar}}_t \sum_{a,b} M_t(a,b) \mathbf{1}\{a \neq b\} = \frac{1}{G} \sum_t \omega^{\mathrm{tar}}_t \operatorname{TV}(f_t, v_{hg}),$$

which is exactly the cost of the first term in the old tail. Finally, the new remaining-tail cost is $\sum_t \omega_t^{\mathrm{tar}} \sum_b v_{hg}(b)\Phi_h(F_{t,b}^{\Xi'})$. Multiplying Equation (F6) by ω_t^{tar} and summing over t shows that it does not exceed the old remaining-tail cost. Combining the three parts gives

$$\mathcal{C}_{s+1}(\Xi') \le \mathcal{C}_s(\Xi).$$

This is a constructive comparison for every old feasible plan. The finite coupling polytope is nonempty and compact, and Equation (24) ensures continuity of the cost. We may therefore take an optimal plan Ξ at level s, giving

$$\mathcal{H}_{s+1}(F,P) \le \mathcal{C}_{s+1}(\Xi') \le \mathcal{C}_s(\Xi) = \mathcal{H}_s(F,P).$$

Step 5: Exactness of the full level and comparison with component transport at the same level. When $s = G$, the tail is empty, and both the feasible set and objective in Equation (31) coincide with those of $\mathcal{T}_{[G]}$. Full-state responsibility lifting in Appendix F.1 gives

$$\mathcal{H}_G(F,P) = \mathcal{T}_{[G]}(F,P) = W(F,P).$$

Here, adding the component labels at both endpoints to any state coupling leaves both the state-pair masses and Hamming cost unchanged; zero-mass states are set to zero as in Appendix F.1.

Next, fix any level $s < G$ and the same feasible plan Ξ. For every positive-mass t, F_t^{Ξ} is a mixture of product tails with probability weights $\Xi(i, x_{A_s}, h, y_{A_s})/\omega_t^{\mathrm{tar}}$. Applying Equation (26) to these tails and multiplying by $(G-s)/G$ to preserve the original normalization gives

$$\omega_t^{\mathrm{tar}} \mathsf{Pref}_{>s}^{(G)}(F_t^{\Xi}, V_h) \le \sum_{i, x_{A_s}} \Xi(i, x_{A_s}, h, y_{A_s}) \frac{1}{G} \sum_{g>s} \mathrm{TV}(q_{ig}, v_{hg}).$$

Summing over t and adding the same explicit prefix cost yields

$$\mathcal{C}_s(\Xi) \le \langle c_{A_s}, \Xi \rangle.$$

Minimizing both sides over the same set of feasible plans gives $\mathcal{H}_s \le \mathcal{T}_{A_s}$. When $s = G$, both tails are empty, and this comparison is an equality. In particular, $s = 0$ gives $\mathcal{H}_0 \le \mathcal{T}_{\varnothing}$. Together with the monotonicity above and equality at the full level, this proves all conclusions in Equation (32) of the main text:

$$\begin{aligned} W(F,P) &= \mathcal{H}_G(F,P) \\ &\le \mathcal{H}_{G-1}(F,P) \le \cdots \le \mathcal{H}_0(F,P) \\ &\le \mathcal{T}_{\varnothing}(F,P), \\ &\mathcal{H}_s(F,P) \le \mathcal{T}_{A_s}(F,P). \end{aligned}$$

This completes the proof.

F.3 Proof of the Direct-Transport Upper Bound

For any feasible assignment a^{tr} in Equation (22), first sample an observation and candidate label according to $\mu(do)a_h^{\mathrm{tr}}(o)$, then generate a full state pair using the coordinatewise optimal coupling in Equation (3). Its source marginal is $\int Q_o \, \mu(do) = R$, its target marginal is $\sum_h \mathbb{E} a_h^{\mathrm{tr}}(O) V_h = P$, and its cost is $\mathbb{E} \sum_h a_h^{\mathrm{tr}}(O)\kappa(Q_O, V_h)$. Taking the infimum over feasible assignments gives $W(R,P) \le \mathcal{T}$. Maximal couplings on finite category sets can be chosen using the explicit formula in Appendix B.1, so the construction also applies to nondiscrete observation laws. For a finite source representation, setting $\lambda_{ih} = a_i a_h^{\mathrm{tr}}(i)$ yields the empty-set level $\mathcal{T}_{\varnothing}$.

F.4 Computational Form and Proofs for Prefix Coupling

Given the full past $(X_{<g}, Y_{<g})$, sequentially construct a maximal coupling of $F_g(\cdot \mid X_{<g})$ and v_g. The first marginal is F, while the conditional distribution of the second marginal at every step is v_g, so the second marginal is V. Thus, prefix coupling is a valid coupling, and $W(F, V) \leq \mathsf{Pref}(F, V)$. Zero-probability prefixes contribute zero, so their conditional distributions may be defined arbitrarily.

Computational form and basic properties. Equation (23) is equivalent to

$$\mathsf{Pref}(F, V) = \frac{1}{2G} \sum_{g, x_{<g}, k} |F(x_{<g}, k) - F(x_{<g}) v_g(k)|. \tag{24}$$

Here, $F(x_{<g})$ and $F(x_{<g}, k)$ denote the prefix marginal $F_{<g}$ and extended-prefix marginal $F_{\leq g}$, respectively; for $g = 1$, the empty prefix has mass 1. By Equation (24), $\mathsf{Pref}(\cdot, V)$ is convex, continuous, takes values in $[0, 1]$, and vanishes if and only if $F = V$. It depends on the coordinate order and generally differs from the exact Wasserstein distance.

Derivation of Equation (24) from Equation (23). Fix a coordinate g. For a prefix with $F(x_{<g}) > 0$, we have $F_g(k \mid x_{<g}) = F(x_{<g}, k)/F(x_{<g})$. Using $\mathrm{TV}(a, b) = \frac{1}{2} \sum_k |a(k) - b(k)|$ and weighting by the prefix mass yields

$$\begin{aligned}
&\mathbb{E}_{X_{<g} \sim F} \, \mathrm{TV}(F_g(\cdot \mid X_{<g}), v_g) \\
&= \frac{1}{2} \sum_{x_{<g}: F(x_{<g}) > 0} F(x_{<g}) \sum_k \left| \frac{F(x_{<g}, k)}{F(x_{<g})} - v_g(k) \right| \\
&= \frac{1}{2} \sum_{x_{<g}: F(x_{<g}) > 0} \sum_k |F(x_{<g}, k) - F(x_{<g}) v_g(k)| \\
&= \frac{1}{2} \sum_{x_{<g}, k} |F(x_{<g}, k) - F(x_{<g}) v_g(k)| \, .
\end{aligned}$$

The last equality follows because, when $F(x_{<g}) = 0$, the identity $F(x_{<g}) = \sum_k F(x_{<g}, k)$ and nonnegativity imply that every $F(x_{<g}, k) = 0$. These prefixes therefore contribute zero to both sides, without requiring any definition of $0/0$. Summing over g and dividing by G gives Equation (24) from Equation (23). This completes the proof.

Prefix responsibilities for finite mixtures. For a specified finite product mixture $F = \sum_{i=1}^{N_{\mathrm{mix}}} \alpha_i Q_i$, $Q_i = \bigotimes_g q_{ig}$, with nonnegative weights summing to 1, define the following on positive-mass prefixes:

$$w_i(x_{<g}) = \frac{\alpha_i \prod_{j<g} q_{ij}(x_j)}{\sum_\ell \alpha_\ell \prod_{j<g} q_{\ell j}(x_j)}, \quad F_g(\cdot \mid x_{<g}) = \sum_i w_i(x_{<g}) q_{ig}. \tag{25}$$

At zero-mass prefixes, any probability responsibility vector may be chosen, with the conditional distribution defined by the right-hand side of Equation (25); this does not affect the integral. Here, $Q_i(x_{<g}) = \prod_{j<g} q_{ij}(x_j)$. Prefix responsibilities retain the full mixture dependence and yield Equation (26).

Proof of the right-hand inequality in Equation (26). Fix g. The product structure of each Q_i gives $F(x_{<g}, k) = \sum_i \alpha_i Q_i(x_{<g}) q_{ig}(k)$, which yields the conditional mixture representation in Equation (25) on positive-mass prefixes. Since $w_i(x_{<g}) \geq 0$ and $\sum_i w_i(x_{<g}) = 1$, we also have $v_g = \sum_i w_i(x_{<g}) v_g$. Applying the triangle inequality for absolute values category by category therefore gives

$$\begin{aligned}\mathrm{TV}(F_g(\cdot \mid x_{<g}), v_g) &= \frac{1}{2}\sum_k \left|\sum_i w_i(x_{<g})\big(q_{ig}(k) - v_g(k)\big)\right| \\ &\le \sum_i w_i(x_{<g})\frac{1}{2}\sum_k |q_{ig}(k) - v_g(k)| \\ &= \sum_i w_i(x_{<g})\,\mathrm{TV}(q_{ig}, v_g).\end{aligned}$$

This is the convexity of TV used here. Next, take expectation over the prefixes under F. For each i, the mean responsibility satisfies

$$\begin{aligned}\mathbb{E}_{X_{<g}\sim F} w_i(X_{<g}) &= \sum_{x_{<g}:F(x_{<g})>0} F(x_{<g})\frac{\alpha_i Q_i(x_{<g})}{F(x_{<g})} \\ &= \sum_{x_{<g}:F(x_{<g})>0} \alpha_i Q_i(x_{<g}) \\ &= \alpha_i \sum_{x_{<g}} Q_i(x_{<g}) = \alpha_i.\end{aligned}$$

Reintroducing the zero-mass prefixes is valid: if $F(x_{<g}) = \sum_j \alpha_j Q_j(x_{<g}) = 0$, nonnegativity of all summands implies that every $\alpha_i Q_i(x_{<g}) = 0$. Thus, the identity also covers the case $\alpha_i = 0$. For $g = 1$, all empty-prefix masses are 1, giving $w_i = \alpha_i$ directly. Combining this with the convexity of TV yields

$$\mathbb{E}_{X_{<g}\sim F}\,\mathrm{TV}(F_g(\cdot \mid X_{<g}), v_g) \le \sum_i \alpha_i\,\mathrm{TV}(q_{ig}, v_g).$$

Finally, summing over g and dividing by G gives

$$\begin{aligned}\mathsf{Pref}(F, V) &\le \frac{1}{G}\sum_g\sum_i \alpha_i\,\mathrm{TV}(q_{ig}, v_g) \\ &= \sum_i \alpha_i\left[\frac{1}{G}\sum_g \mathrm{TV}(q_{ig}, v_g)\right] = \sum_i \alpha_i \kappa(Q_i, V).\end{aligned}$$

The last equality uses Equation (3), with both Q_i and V being product distributions. The left-hand inequality in Equation (26) was established by the sequential maximal coupling above. This completes the proof.

Equation (27) and the cancellation quantity. Taking the expectation of the definition of c_F in Appendix F.6 and using the identity $\mathbb{E}_F w_i = \alpha_i$ just proved gives Equation (27).

For every $(g, x_{<g}, k)$, let $A_+ = \sum_i \alpha_i Q_i(x_{<g})[q_{ig}(k) - v_g(k)]_+$ and $A_- = \sum_i \alpha_i Q_i(x_{<g})[v_g(k) - q_{ig}(k)]_+$. The total subtraction is exactly $G^{-1}\sum_{g,x_{<g},k}\min(A_+, A_-)$, quantifying cancellation between deviations in opposite directions among components that have not yet been identified by the prefix.

F.5 Stopping-Tree Construction, Proposition 2, and Proofs for the Tree-Augmented Hierarchy

Starting from a product node $T = \bigotimes_g t_g$, select only one block g to split. For a normalized nonnegative channel $C(b \mid k)$ on a finite output alphabet, with $\sum_b C(b \mid k) = 1$, define

$$\beta_b = \sum_k t_g(k)C(b \mid k), \qquad t_g^{(b)}(k) = \frac{t_g(k)C(b \mid k)}{\beta_b}, \qquad T^{(b)} = t_g^{(b)} \otimes \bigotimes_{\ell \ne g} t_\ell \quad (\beta_b > 0). \tag{33}$$

Zero-weight branches are removed; the remaining branches satisfy $\sum_b \beta_b T^{(b)} = T$ exactly. Recursive splitting preserves the product structure of the nodes and leaves the candidate unchanged.

Construct a finite tree for each root V_h, and denote its root node by ρ_h. Let $\mathcal{N}_{\mathscr{T}}$ be the set of all nodes, $\mathrm{ch}(v)$ the children of node v, and $\mathcal{N}_{\mathscr{T}}^{\mathrm{int}} := \{v \in \mathcal{N}_{\mathscr{T}} : \mathrm{ch}(v) \neq \varnothing\}$ the set of internal nodes. The node templates T_v are all product distributions, and

$$\begin{aligned} T_{\rho_h} &= V_h && (h \in [C]), \\ T_v &= \sum_{w \in \mathrm{ch}(v)} \beta_{w|v} T_w && (v \in \mathcal{N}_{\mathscr{T}}^{\mathrm{int}}), \\ \beta_{w|v} &\geq 0 \quad (w \in \mathrm{ch}(v)), \qquad \sum_{w \in \mathrm{ch}(v)} \beta_{w|v} = 1 && (v \in \mathcal{N}_{\mathscr{T}}^{\mathrm{int}}). \end{aligned} \tag{34}$$

Equation (33) provides an exact blockwise construction satisfying Equation (34).

Exact splitting. By Equation (33), for each category k, we have $\sum_{b:\beta_b>0} \beta_b t_g^{(b)}(k) = t_g(k) \sum_b C(b \mid k) = t_g(k)$; all terms for zero-weight branches are zero. The other blocks remain unchanged, so $\sum_b \beta_b T^{(b)} = T$. Applying this recursively yields Equation (34).

Conservation of the stopping representation. Indeed, the nonroot incoming masses satisfy $\sum_{v \notin \{\rho_1,\ldots,\rho_C\}} f_v T_v = \sum_{v \text{ internal}} \tau_v T_v$. Using $f_v = \sigma_v + \tau_v$ and the root masses, termwise cancellation yields Equation (36). The stopping weights are masses in the transport representation, rather than a refitting of the candidate weights.

Proof of Proposition 2. For each (i, v), use the optimal product coupling from Equation (3), then mix according to γ_{iv}. The row constraints preserve the source marginal F, and Equation (36) preserves the target marginal P. When all mass stops at the roots, any old component plan can be reused, giving the second inequality. After extending an old leaf, its continuing mass can still be set to zero, so every old solution remains feasible. This completes the proof.

Equation (41). The first two lines follow from Theorem 2 for a fixed stopping representation, stopping all mass at the roots, and inclusion of the feasible sets of old stopping representations, respectively. For the last line, directly apply responsibility lifting to an optimal full-space coupling of the original distributions and retain only A_s: the cost of the expanded coordinates is unchanged, while the feasible product-coupling cost on the remaining coordinates is at most $(G - s)/G$. Then use $\mathcal{H}_s^{\mathscr{T}} \leq \mathcal{H}_s \leq \mathcal{T}_{A_s}$. The full level is therefore exact, but state expansion may be exponentially expensive; enlarging only the target tree does not automatically yield this completeness property.

F.6 Exact Cancellation Quantity for Prefix Coupling

Using the mixture representation from Section 5.2 and the prefix responsibilities from Appendix F.4, define

$$c_F(x) = \frac{1}{G} \sum_g \left[\sum_i w_i(x_{<g}) \, \mathrm{TV}(q_{ig}, v_g) - \mathrm{TV}\left(\sum_i w_i(x_{<g}) q_{ig}, v_g \right) \right] \in [0, 1].$$

The notation c_F also implicitly depends on the fixed source decomposition $(\alpha_i, Q_i)_{i=1}^{N_{\mathrm{mix}}}$ and target V; it is not a decomposition-invariant quantity determined solely by the aggregate source F. The following exact subtraction identity holds:

$$\boxed{\mathsf{Pref}(F, V) = \sum_i \alpha_i \kappa(Q_i, V) - \mathbb{E}_F c_F(X).} \tag{27}$$

Equation (27) quantifies cancellation between probability deviations in opposite directions; its derivation is given in Appendix F.4.

F.7 Computational Conventions for Prefix Coupling and the Hierarchy

Computation and limitations. For a finite source representation, compute the prefix mixture probabilities using Equation (25), then sum over nonzero prefixes using Equation (24). Log-sum-exp can be used to handle extremely small prefix probabilities. The number of prefixes may grow exponentially with the number of blocks. This manuscript uses only deterministically evaluated empirical coupling costs or costs with deterministic safe upper bounds for certification; see Appendices G.10-G.11.

The absolute-value linear representation in Equation (24) makes $\mathcal{C}_s$ convex in Ξ; after multiplication by ω_t^{tar}, division by an unknown mass is unnecessary. On finite spaces, this can be formulated as an LP, but both state expansion and enumeration of long prefixes may be exponentially expensive. The implementation initializes from the old component plan, accepts only improvements whose marginals have been repaired and whose costs have safe upper estimates, and retains the old feasible cost. Monotonicity across levels of the ideal optimum does not automatically guarantee monotone outputs from an arbitrary approximate solver.

F.8 Implementation Details for the Stopping Tree and Combined Approach

The single-block split in Equation (33) of Appendix F.5 removes zero-weight branches and satisfies $\sum_{b:\beta_b>0} \beta_b T^{(b)} = T$. After recursive splitting, every node remains a product distribution, and leaf weights are products of the weights along their paths. The proof is given in Appendix F.5. TV costs allow zero probabilities. Prespecified channels directly determine node probabilities; separately clipping node probabilities and weights must be avoided because it would invalidate the mixture identity. This construction neither assumes that the true reference follows a tree model nor changes the posterior or candidate.

Equation (38) is an LP: take γ_{iv}, τ_v as variables and substitute $\sigma_v = \sum_i \gamma_{iv}$ into Equation (35) to obtain linear constraints. There are $N|\mathcal{N}_{\mathscr{T}}|$ transport variables and an additional $O(|\mathcal{N}_{\mathscr{T}}|)$ flow variables; this is not a guarantee on solution time. The construction retains costs over all coordinates without enumerating latent states. Forcing all mass to the leaves may remove the original root plan and therefore does not provide the same monotonicity.

Combining with the existing TV hierarchy. The standalone $\mathcal{T}_{\mathscr{T}}$ and the prefix-aggregation value $\mathcal{H}_0$ admit no universal ordering: in the first example of Appendix I.5, with only roots retained, the former is strictly larger; for the tree constructed in the second example, it is strictly smaller. For each valid stopping representation $P = \sum_v \sigma_v T_v$, apply the lifting and tail-coupling definitions in Section 5.3 and Appendix F.1, and then jointly optimize the stopping masses. Let $\mathcal{H}_s(F; \sigma, T)$ denote the same construction with the original target components (π_h, V_h) replaced by (σ_v, T_v); its joint optimum is given in Equation (40).

When the stopping masses vary, the target-prefix mass is $\sigma_v T_v^{A_s}(y_{A_s})$, which still gives a linear constraint. The perspective form of Equation (24) ensures that the joint prefix objective can be formulated as a finite LP. In practice, one may compute only the tree direct-transport plan or a few prefix improvements and take the minimum with previously obtained feasible hierarchy costs, without solving the full joint optimization.

Parent-node retention ensures that old solutions can be reused; it does not guarantee a strict cost reduction at every refinement, nor does it establish that arbitrary high-dimensional posteriors can be tightly certified using small trees. Prespecifying exact channels and solving Equation (38) suffice to apply the results of this section. If old channels are replaced in a way that changes old templates, the nestedness relation cannot be invoked directly; the valid plan and safe cost from before the update must be retained for the same batch of Q_i, row masses, candidate P, and metric. The noninferiority relations above compare computational schemes only for the same posterior, row masses, candidate, and metric. After model retraining, the costs must be recomputed and recertified; their old numerical values cannot be reused directly.

G Statistical Endpoints, Strong Baselines, and Implementation Calibration

G.1 Label Frequencies and Scoring of the Minimum

Conditional on all independent fitting results, use the independent and identically distributed scoring rows from Appendix G.9. Codebook sizes and diameters are taken from the complete codebooks frozen before scoring, rather than from the labels that happen to appear during scoring. Define

$$\gamma_B = \max_{z,z'} \kappa_B(U_{Bz}, U_{Bz'}), \qquad B_0 = \sum_B \alpha_B \gamma_B, \qquad a_n = \sum_B \alpha_B \frac{\gamma_B \sqrt{C_B - 1}}{2\sqrt{n}}. \tag{G1}$$

For single-center groups, both γ_B and the corresponding bias term are zero. This bound depends explicitly on the frozen codebook sizes C_B; when the group family, coverage weights, and codebooks are fixed, $a_n = O(n^{-1/2})$.

Proposition G1 (Label-frequency scoring bound for the local TV optimum). For D in Eq. (11) and its empirical version $\widehat{D}$,

$$\begin{aligned} |\widehat{D} - D| &\le \sum_B \alpha_B \gamma_B \, \mathrm{TV}(\widehat{\boldsymbol{r}}_B, \boldsymbol{r}_B), \\ \mathbb{E}|\widehat{D} - D| \le a_n, \qquad & 0 \le D - \mathbb{E}\widehat{D} \le a_n. \end{aligned} \tag{G2}$$

Replacing any one row changes $\widehat{D}$ by at most B_0/n; labels from different groups may be dependent, and independence across groups is not required.

Proof. Fix B and any feasible center $u_B \in \mathcal{F}_B$. The reverse triangle inequality gives $|\kappa_B(U_{Bz}, u_B) - \kappa_B(U_{Bz'}, u_B)| \le \gamma_B$. Construct a maximal coupling of the population and empirical labels: the objective difference is zero when the labels coincide, and its absolute value is at most γ_B when they differ. Hence,

$$\left| \sum_z (\widehat{r}_{Bz} - r_{Bz}) \kappa_B(U_{Bz}, u_B) \right| \le \gamma_B \, \mathrm{TV}(\widehat{\boldsymbol{r}}_B, \boldsymbol{r}_B).$$

The right-hand side does not depend on u_B. Applying $|\inf f - \inf g| \le \sup |f - g|$ and summing with weights α_B gives the pointwise bound in Eq. (G2). Since the scoring rows are independent and identically distributed, each $\widehat{r}_{Bz}$ has variance $r_{Bz}(1 - r_{Bz})/n$; by the Cauchy-Schwarz inequality and $\sum_z r_{Bz} = 1$,

$$\mathbb{E}\,\mathrm{TV}(\widehat{\boldsymbol{r}}_B, \boldsymbol{r}_B) \le \frac{1}{2\sqrt{n}} \sum_z \sqrt{r_{Bz}(1 - r_{Bz})} \le \frac{\sqrt{C_B - 1}}{2\sqrt{n}}.$$

Taking expectations and summing with the weights gives $\mathbb{E}|\widehat{D} - D| \le a_n$. Let $f_u(o) = \sum_B \alpha_B \kappa_B(U_{B,A_B(o)}, u_B)$, with each $u_B \in \mathcal{F}_B$ optimized separately. Then $D = \inf_u \mathbb{E} f_u$ and $\widehat{D} = \inf_u n^{-1} \sum_i f_u(O_i)$. Interchanging expectation and infimum gives $\mathbb{E}\widehat{D} \le D$; combining this with the absolute-deviation bound yields the last relation in Eq. (G2).

For any two rows o, o', the reverse triangle inequality gives $|f_u(o) - f_u(o')| \le B_0$. Replacing one scoring row changes every empirical objective by at most B_0/n, and this bound is preserved after taking the infimum. This completes the proof.

Proposition G2 (Asymmetric confidence interval for the minimization endpoint). Let

$$q_n = B_0 \sqrt{\frac{\log(2/\delta_r)}{2n}}.$$

Then, with probability at least $1 - \delta_r$,

$$\boxed{\widehat{D} - q_n \le D \le \widehat{D} + a_n + q_n.} \tag{G3}$$

Proof. For $D = \inf_u \mathbb{E} f_u$ and $\widehat{D} = \inf_u \widehat{\mathbb{E}} f_u$, interchanging expectation and infimum gives $\mathbb{E}\widehat{D} \le D$; Proposition G1 additionally gives $0 \le D - \mathbb{E}\widehat{D} \le a_n$. Applying the two-sided bounded-differences inequality (McDiarmid, 1989) to $\widehat{D} - \mathbb{E}\widehat{D}$ proves the result. This is a specialization to the present objective of the classical asymmetry in the two sides of the deviation of a stochastic optimization optimum. (Guigues et al., 2017)

When only a valid lower endpoint is required, q_n may be replaced by the one-sided radius $B_0\sqrt{\log(1/\delta_r)/(2n)}$; below, Eq. (G3) is used to handle the detection analysis uniformly. When $B_0 = 0$, the radius is zero.

For $n \ge 2$ independent scalar samples $X_1, \dots, X_n$, all taking values in the same known interval of length a_0, write $\bar{X} = n^{-1}\sum_i X_i$ and $\mu_X^{\mathrm{av}} = n^{-1}\sum_i \mathbb{E}X_i$, and define the one-sided empirical Bernstein radius (Maurer and Pontil, 2009, Theorems 4 and 11)

$$\mathrm{EB}(X, a_0, \delta) = \sqrt{\frac{2s_X^2 \log(2/\delta)}{n}} + \frac{7a_0 \log(2/\delta)}{3(n-1)}, \tag{G4}$$

where $s_X^2 = (n-1)^{-1}\sum_i (X_i - \bar{X})^2$. This radius satisfies

$$\Pr\{\mu_X^{\mathrm{av}} \le \bar{X} + \mathrm{EB}(X, a_0, \delta)\} \ge 1 - \delta.$$

For independent and identically distributed samples, $\mu_X^{\mathrm{av}} = \mathbb{E}X_1$, corresponding to (Maurer and Pontil, 2009, Theorem 4); for independent but not necessarily identically distributed samples, the target remains the average of the row expectations, corresponding to (Maurer and Pontil, 2009, Theorem 11). In the latter case, s_X^2 is not interpreted as an unbiased estimator of a common population variance. This extension of the general operator does not remove the requirements concerning the target conditional law and independent and identically distributed observations in the individual audit subsections.

G.2 Computation and Asymmetric Intervals for Component Transport

First consider two-component weights $p_{\mathrm{mix}} := \pi_1 \in (0,1)$ and $\pi_2 = 1 - p_{\mathrm{mix}}$, so that $P = p_{\mathrm{mix}}V_1 + (1 - p_{\mathrm{mix}})V_2$, and let $\gamma = \kappa(V_1, V_2)$. Below, write the random costs as $c_h = c_h(O)$, with scoring samples $c_{ih} = c_h(O_i) = \kappa(Q_{O_i}, V_h)$; $\mathrm{EB}(c_h, \cdot, \cdot)$ takes the sample column $(c_{ih})_{i=1}^n$ as input. Let $d_i = c_{i1} - c_{i2}$, let $\widehat{F}(t) = n^{-1}\sum_{i=1}^n \mathbf{1}\{d_i \le t\}$ be the empirical distribution function of the cost differences, and sort the d_i. Let $k = \lfloor np_{\mathrm{mix}} \rfloor$ and $\omega_{\mathrm{frac}} = np_{\mathrm{mix}} - k \in [0,1)$, so that $np_{\mathrm{mix}} = k + \omega_{\mathrm{frac}}$. Then

$$\widehat{\mathcal{T}} = \frac{1}{n}\sum_i c_{i2} + \frac{1}{n}\left(\sum_{j=1}^{k} d_{(j)} + \omega_{\mathrm{frac}} d_{(k+1)}\right). \tag{G5}$$

Sorting and splitting mass at the boundary enforce the candidate weights exactly, allowing atoms, ties, and noninteger np_{mix}. When $p_{\mathrm{mix}} \in \{0,1\}$, simply use the corresponding mean cost. The two-component population semi-dual is

$$\mathcal{T} = \sup_{t \in [-\gamma, \gamma]} g(t), \qquad g(t) = p_{\mathrm{mix}} t + \mathbb{E}\min(c_1 - t, c_2). \tag{G6}$$

More generally, for any finite number C of candidate components, restore the full probability vector $\pi = (\pi_h)_{h \in [C]}$ below, with $\pi \cdot v = \sum_h \pi_h v_h$. The standard semi-discrete transport duality is given

in (Peyré and Cuturi, 2019, Sections 5.1-5.2, Eqs. (5.4)-(5.7)); for the costs of this section, it takes the form

$$\mathcal{T} = \sup_{v \in \mathbb{R}^C} \{\pi \cdot v + \mathbb{E} f_v(O)\}, \qquad f_v(o) = \min_h \{c_h(o) - v_h\}. \tag{G7}$$

After removing zero-weight candidate components and fixing $\max_h v_h = 0$, any $v_h < -1$ can be truncated to -1 without decreasing the objective, so a population optimum exists. All costs lie in $[0, 1]$. If a smaller uniform bound on the variation across rows is known before scoring,

$$\sup_{o,o'} \max_h |c_h(o) - c_h(o')| \le B_{\text{cost}} \le 1, \tag{G8}$$

then each f_v has range at most B_{cost}. The default is $B_{\text{cost}} = 1$; a smaller value must follow from fixed support or model constraints and cannot be replaced by the empirical range of the scoring samples.

Theorem G1 (Asymmetric transport interval for the maximization endpoint). For any finite number C of candidate components,

$$\Pr\left\{\mathcal{T} \le \widehat{\mathcal{T}} + B_{\text{cost}} \sqrt{\frac{\log(1/\delta_T)}{2n}}\right\} \ge 1 - \delta_T. \tag{G9}$$

For two components, further let

$$\beta_n = \gamma \rho_n, \qquad q_T = B_{\text{cost}} \sqrt{\frac{\log(2/\delta_T)}{2n}}.$$

Then, on the same event of probability at least $1 - \delta_T$,

$$\boxed{\widehat{\mathcal{T}} - \beta_n - q_T \le \mathcal{T} \le \widehat{\mathcal{T}} + q_T, \qquad U_{\text{SAA}} = \min\{1, \widehat{\mathcal{T}} + q_T\}.} \tag{G10}$$

Here β_n is a uniform upper bound on the upward bias of two-component empirical transport; it is attained over the full class that permits candidate weights to be chosen in advance, without claiming that every fixed p_{mix} attains the same worst-case constant. SAA stands for sample average approximation.

Proof. Fix a population optimizer v^* of Eq. (G7). The empirical supremum is at least the empirical value at v^*, and therefore

$$\mathcal{T} - \widehat{\mathcal{T}} \le \mathbb{E} f_{v^*} - \widehat{\mathbb{E}} f_{v^*}.$$

Applying the one-sided Hoeffding inequality to this fixed function gives Eq. (G9).

For two components, continue to use the scalar $p_{\text{mix}} = \pi_1$; the cost difference $D_c = c_1 - c_2$ is supported on $[-\gamma, \gamma]$. If its distribution function is F, define the generalized quantile as $F^{-1}(u) := \inf\{t : F(t) \ge u\}$ for $u \in (0, 1)$, with endpoint values having no effect on the integral. Then

$$\mathcal{T} = \mathbb{E} c_2 + \Psi_{p_{\text{mix}}}(F), \qquad \Psi_{p_{\text{mix}}}(F) = \int_0^{p_{\text{mix}}} F^{-1}(u)\, du = p_{\text{mix}} \gamma - \int_{-\gamma}^{\gamma} \min\{F(t), p_{\text{mix}}\}\, dt. \tag{G11}$$

This identity applies to distributions with atoms. The relation $\mathbb{E}\widehat{\mathcal{T}} \ge \mathcal{T}$ follows from the relationship between expectation and supremum. Using $\min(x, p_{\text{mix}}) = (x + p_{\text{mix}} - |x - p_{\text{mix}}|)/2$ and $\mathbb{E}\widehat{F} = F$ further gives

$$0 \le \mathbb{E}\widehat{\mathcal{T}} - \mathcal{T} = \frac{1}{2}\int_{-\gamma}^{\gamma} \Big[\mathbb{E}|\widehat{F}(t) - p_{\mathrm{mix}}| - |F(t) - p_{\mathrm{mix}}|\Big]\, dt \\ \le \frac{1}{2}\int_{-\gamma}^{\gamma} \mathbb{E}|\widehat{F}(t) - F(t)|\, dt \le \gamma\rho_n = \beta_n. \tag{G12}$$

Replacing one row changes every empirical dual objective by at most B_{cost}/n, and this remains true after taking the supremum. Applying the two-sided bounded-differences inequality (McDiarmid, 1989) to $\widehat{\mathcal{T}} - \mathbb{E}\widehat{\mathcal{T}}$ and combining it with Eq. (G12) yields Eq. (G10).

To show that the bias constant is attainable, let the posterior equal V_1 with probability p and V_2 with probability $1-p$, and let the candidate use the same weights. Population transport is zero, whereas empirical transport is exactly $\gamma|\widehat{p} - p|$. Choosing p that attains ρ_n makes its expectation equal to $\gamma\rho_n$. This bias does not enter the valid upper endpoint, but it still enters two-sided approximation and detection analysis. This completes the proof.

The one-sided reliability in Eq. (G9) applies to any fixed finite number of components of a single candidate P; the bias bound in Eq. (G12) is used only in the two-component case. These intervals continue to be used for the component transport baseline and diagnostics. For the TV upper endpoint in the main text, the common event in Appendices G.10-G.11 replaces the default SAA branch; its statistical target is the full W. New coupling costs are not substituted into Eq. (G10) as if they provided a confidence upper bound on $\mathcal{T}$.

G.3 Strong Anchor and CVaR Endpoints

Continue to use $p_{\mathrm{mix}} = \pi_1 \in (0,1)$, let $D_c = c_1 - c_2$, and let F be its distribution function. The lower-tail integral of two-component transport can be controlled by CDF envelopes; see (Thomas and Learned-Miller, 2019, Lemma 1 and Theorems 3-4) for the underlying tools. Following that paper's convention of parameterizing by upper-tail mass, if F_{-D_c} is the distribution function of $-D_c$, then

$$\mathrm{CVaR}_{p_{\mathrm{mix}}}(-D_c) := \frac{1}{p_{\mathrm{mix}}}\int_{1-p_{\mathrm{mix}}}^{1} F^{-1}_{-D_c}(u)\, du, \qquad \Psi_{p_{\mathrm{mix}}}(F) = -p_{\mathrm{mix}}\,\mathrm{CVaR}_{p_{\mathrm{mix}}}(-D_c).$$

Thus, this section uses the lower-tail integral in Eq. (G11), without division by the tail mass, rather than upper-tail CVaR itself; generalized quantile integrals are still used when atoms are present. When $p_{\mathrm{mix}} \in \{0,1\}$, the empirical optimum reduces to the corresponding component's mean cost as in Appendix G.2; a valid radius for mean estimation is still added to the reliable endpoint, and no CVaR involving division by zero is computed. Fix an independently fitted threshold $t_0 \in [-\gamma, \gamma]$. The one-sided derivatives of the semi-dual are

$$g'_+(t) = p_{\mathrm{mix}} - F(t), \qquad g'_-(t) = p_{\mathrm{mix}} - F(t-). \tag{G13}$$

Take $\varepsilon_n = \sqrt{\log(2/\delta_F)/(2n)}$, $F_- = [\widehat{F} - \varepsilon_n]_+$, and $F_+ = \min\{1, \widehat{F} + \varepsilon_n\}$. We next make explicit the correspondence between the two-sided DKW-Massart event used here and reference (Reeve, 2024).

To distinguish it from the CDF error radius in this section, denote the parameter in (Reeve, 2024) by $\varepsilon_{\mathrm{ref}}(n,\delta) = \log(1/\delta)/n$. The one-sided threshold in [9, Corollary 1] is $\varphi(I, \varepsilon_{\mathrm{ref}})\sqrt{\varepsilon_{\mathrm{ref}}/2}$, where $\varphi(I, \varepsilon_{\mathrm{ref}}) \le 1$. For the independent and identically distributed cost-difference samples, take $I = \mathbb{R}$ and relax this threshold to $\sqrt{\varepsilon_{\mathrm{ref}}/2}$, obtaining

$$\Pr\left\{\sup_{t\in\mathbb{R}}[\widehat{F}(t) - F(t)] > \sqrt{\frac{\log(1/\delta)}{2n}}\right\} \le \delta.$$

For the reverse direction, apply the same one-sided bound to the reflected variable $-D_c$. When atoms are present, its population and empirical distribution functions respectively satisfy

$$F_{-D_c}(u) = 1 - F((-u)-), \qquad \widehat{F}_{-D_c}(u) = 1 - \widehat{F}((-u)-).$$

Using $F(t) = \lim_{r \downarrow t} F(r-)$ and the same property of the empirical CDF then gives the same one-sided bound for $\sup_t[F(t) - \widehat{F}(t)]$. Taking $\delta = \delta_F/2$ in each direction and applying the union bound gives

$$\Pr\left\{\sup_{t\in\mathbb{R}} |\widehat{F}(t) - F(t)| \le \varepsilon_n\right\} \ge 1 - \delta_F, \qquad \varepsilon_n = \sqrt{\frac{\varepsilon_{\mathrm{ref}}(n, \delta_F/2)}{2}} = \sqrt{\frac{\log(2/\delta_F)}{2n}}.$$

On this event, $F_-(t) \le F(t) \le F_+(t)$ holds simultaneously for all t; taking left limits as $r \uparrow t$ also gives $F_-(t-) \le F(t-) \le F_+(t-)$, without consuming any additional failure-probability budget. The event therefore simultaneously controls both one-sided derivatives in Eq. (G13), allowing atoms in the cost distribution. Reference (Reeve, 2024) provides uniform CDF control here; the endpoints below are then constructed by combining it with the semi-dual and integral representations. Define

$$B_{\mathrm{int}} = \max\left\{\int_{t_0}^{\gamma} [p_{\mathrm{mix}} - F_-(t)]_+ dt, \int_{-\gamma}^{t_0} [F_+(t) - p_{\mathrm{mix}}]_+ dt\right\}. \tag{G14}$$

Theorem G2 (Integral rebalancing upper bound). On the DKW event, integration using Eq. (G13) gives $0 \le \mathcal{T} - g(t_0) \le B_{\mathrm{int}}$. Let $Y_i = \min(c_{i1} - t_0, c_{i2})$, whose range is at most B_{cost}. Then

$$U_{\mathrm{anchor}} = \min\{1, p_{\mathrm{mix}} t_0 + \overline{Y} + \mathrm{EB}(Y, B_{\mathrm{cost}}, \delta_Y) + B_{\mathrm{int}}\} \tag{G15}$$

bounds $W(R, P)$ on an event of probability at least $1 - \delta_F - \delta_Y$. Write $\overline{c}_2 = n^{-1}\sum_{i=1}^n c_{i2}$. The comparator that separately estimates the mean cost and lower-tail integral is

$$U_{\mathrm{CVaR}} = \min\left\{1, \overline{c}_2 + \mathrm{EB}(c_2, B_{\mathrm{cost}}, \delta_2) + p_{\mathrm{mix}}\gamma - \int_{-\gamma}^{\gamma} \min\{F_-(t), p_{\mathrm{mix}}\}\, dt\right\}. \tag{G16}$$

Its budget is $\delta_F + \delta_2$. There is no universal ordering between the two endpoints: U_{anchor} preserves the correlation between paired costs, whereas U_{CVaR} controls the two terms separately. Empirical CDF integrals can be computed by a linear scan after sorting. Equation (49) in Appendix G.11 replaces the default SAA branch while retaining the same strong endpoints and budgets; the small-range branch uses U_{small} from Appendix G.11 and its dedicated budget.

G.4 Numerical Feasible Costs and Range Boundaries

Numerical-feasible-cost convention. Fix the same empirical source, row masses, candidate, and metric. Let C^+_{old} and C^+_{new} be valid safe upper bounds on $W(\widehat{R}_n, P)$ from the old and new schemes, respectively; when the old scheme uses the original component plan, $C^+_{\mathrm{old}} = \widehat{\mathcal{T}}^+$. With the same radius r_T^W, define

$$\begin{aligned} C^+_{\mathrm{keep}} &= \min\{C^+_{\mathrm{old}}, C^+_{\mathrm{new}}\}, \\ U_{\mathrm{keep}} &= \min\{1, C^+_{\mathrm{keep}} + r_T^W\}, \qquad U^+_{\mathrm{old}} = \min\{1, C^+_{\mathrm{old}} + r_T^W\}. \end{aligned}$$

Then $U_{\mathrm{keep}} \le U^+_{\mathrm{old}}$; retaining the same anchor and CVaR endpoints also makes the combination formed by taking the minimum no worse. The comparison here is with the old feasible endpoint under the same numerical convention, rather than with exact SAA. To retain the proof of comparison with exact SAA in Eqs. (50)-(51), one still needs $C^+_{\mathrm{keep}} \le \widehat{\mathcal{T}}$; if only $C^+_{\mathrm{old}} \le \widehat{\mathcal{T}} + \epsilon_T$ is available, retain ϵ_T as in Appendix G.12. For example, when the two source rows and two candidate roots are both the equally weighted $\mathrm{Bern}(0.1), \mathrm{Bern}(0.9)$, the diagonal plan has exact optimal cost 0, whereas the crossed feasible plan has cost 0.8; retaining the latter does not imply that the new cost is at most 0.

Taking the minimum of these deterministic safe costs is covered by the same event in Eq. (44). Both the old and new costs must satisfy the same full empirical marginal constraints and safe-upper-bound conditions, and the comparison uses the same outer radius; one cannot take the minimum of unverified numerical point estimates and then claim dominance at the same budget.

Cost ranges cannot be transferred directly. In general, the old cost range $B_{\mathrm{cost}} < 1$ cannot be transferred to Eq. (44). Take one binary block, with $Q_0 = \mathrm{Bern}(0.1)$, $Q_1 = \mathrm{Bern}(0.9)$, candidate $P = V = \mathrm{Bern}(0.5)$, and $\Pr(O = 0) = 3/4$, $\Pr(O = 1) = 1/4$. Every row has cost 0.4, so $B_{\mathrm{cost}} = 0$ is valid; however, $R = \mathrm{Bern}(0.3)$ and $W(R, V) = 1/5$. Let $\widehat{p} = n^{-1} \sum_i \mathbf{1}\{O_i = 0\}$. Then

$$W(\widehat{R}_n, V) = 0.8|\widehat{p} - 1/2|.$$

When $n = 2$, the probability of observing one row of each posterior is $2(3/4)(1/4) = 3/8$, in which case $\widehat{R}_2 = V$ and the empirical distance is zero. Thus, incorrectly using $B_{\mathrm{cost}} = 0$ as the radius coefficient in Eq. (44) would underestimate the true distance with probability $3/8 > 0.05$, violating a claim with $\delta_W = 0.05$. This counterexample refutes the incorrect transfer of the radius coefficient, rather than Eq. (44).

Consequently, small-range SAA, anchor, and CVaR remain endpoints with separate statistical advantages; if small-range SAA is used together with the new common event, a joint budget must be allocated for its validity. The full two-sided interval for $\mathcal{T}$ in the original Eq. (G10) is not automatically preserved by the new event either.

G.5 Proof of Validity of the Codebook Lower Endpoint

Validity of the lower endpoint. By Eqs. (11) and (42), $\mathbb{E}Z_i = e$. On the intersection of the codebook event, the two-sided Hoeffding event, and the one-sided empirical Bernstein event,

$$\widehat{D} - q_n \le D, \qquad e \le \widehat{e} + r_e^{\mathrm{H}}, \qquad e \le \widehat{e} + \mathrm{EB}(Z, s_{\mathrm{cov}}, \delta_e).$$

Hence $e \le \widehat{e} + r_e$, and therefore

$$L = [\widehat{D} - q_n - (\widehat{e} + r_e) - \tau_*]_+ \le [D - e - \tau_*]_+ \le \Delta(R),$$

where the last step follows from Section 4.2 and Appendix E.6. The total failure probability of the three events is at most $\delta_H + \delta_e + \delta_r$; the events need not be mutually independent.

G.6 Proof of Conditional Detection and Numerical Errors

For two-component candidates, use $\beta_n = \kappa(V_1, V_2)\rho_n$; for a single component, take $\beta_n = 0$, in which case empirical component transport is a mean cost and its upward deviation is controlled by the one-sided Hoeffding inequality. A single component $P \in \mathcal{F}$ satisfies $\ell_* \le \Delta(R) \le W(R, P) \le \mathcal{T}$, so a strictly positive population certificate gap is impossible.

Proof. By Eq. (G3) and $\widehat{e} - e \le r_e^{\mathrm{H}}$,

$$L \ge [D - e - \tau_* - (a_n + 2q_n + r_e^{\mathrm{H}} + r_e)]_+ \ge \ell_* - (a_n + 2q_n + r_e^{\mathrm{H}} + r_e).$$

The first inequality follows from Eq. (43); the second uses $[x - a]_+ \ge [x]_+ - a$ for $a \ge 0$, without requiring $D - e - \tau_*$ to be positive. Since $U \le U_{\mathrm{coupling}} \le \widehat{\mathcal{T}} + r_T^W \le \mathcal{T} + \beta_n + 2r_T^W$, subtracting the two bounds gives Eq. (53).

Conditional on all independent fitting results, denote the extended common scoring event above by $\mathcal{E}$, so that $\Pr(\mathcal{E}) \ge 1 - \delta$. On $\mathcal{E}$, $\ell_* - \mathcal{T} > S_n$ implies $L > U$, and thus $\{\ell_* - \mathcal{T} > S_n,\ L \le U\} \subseteq \mathcal{E}^c$, yielding the random-threshold probability statement in Appendix G.12. Moreover, since $S_n \le \overline{S}_n$ for every dataset, if the frozen population gap satisfies $\ell_* - \mathcal{T} > \overline{S}_n$, then throughout $\mathcal{E}$,

$$L - U \geq (\ell_* - \mathcal{T}) - S_n \geq (\ell_* - \mathcal{T}) - \overline{S}_n > 0.$$

Consequently, $\Pr\{L > U\} \geq \Pr(\mathcal{E}) \geq 1 - \delta$. No additional statistical event or budget is required. This completes the proof.

When computing only the original lower bound and component transport baseline, the scoring cost for nonoverlapping groups, two centers, and two components is $O(n \sum_g K_g + n \log n)$. Overlapping groups are charged according to the total number of coordinate occurrences; general codebooks additionally require codebook-diameter computation and simplex-constrained optimization. The solution and prefix-summation costs of the new couplings are reported separately according to Sections 5.2-5.4.

In the formulas above, $\widehat{D}$ and $\widehat{\mathcal{T}}$ are exact optima. In practice, the objective value of a feasible solution to the local minimization problem is an upper bound on $\widehat{D}$ and cannot be substituted directly into the certification lower endpoint. One should use a verified dual lower bound $\widehat{D}^-$ or subtract a guaranteed optimization error from the feasible objective. Candidate transport has the opposite direction: the upper endpoint should use the primal feasible cost after repairing marginal mass residuals and its numerically safe upper estimate $\widehat{\mathcal{T}}^+$.

If $0 \leq \widehat{D} - \widehat{D}^- \leq \epsilon_D$ and $0 \leq \widehat{\mathcal{T}}^+ - \widehat{\mathcal{T}} \leq \epsilon_T$, and the old plan is retained so that the actual safe cost does not exceed $\widehat{\mathcal{T}}^+$, then this cost is at most $\widehat{\mathcal{T}} + \epsilon_T$. With the same statistical radii, reliability is preserved, while the right-hand side of Eq. (53) for the actual endpoints additionally subtracts $\epsilon_D + \epsilon_T$. Without an envelope for the optimality error, a feasible cost can still provide a valid upper endpoint, but the detection guarantee that omits optimization error cannot be directly retained; nor does retaining the old plan alone imply dominance over exact SAA. In this case, the comparison is with the old feasible endpoint under the same numerical convention in Appendices G.11 and G.4.

The sharpened binomial constant should be rounded upward; if the rounding direction cannot be guaranteed, the conservative alternative $\rho_n \leq 1/(2\sqrt{n})$ may be used.

Common scoring event. Take the minimum of valid upper bounds and the maximum of valid lower bounds only for the same frozen reference, candidate, and metric. The claimed scope over all models, conditions, seeds, and scoring sample sizes must be accounted for in the total budget; shared statistical events are counted only once. In Appendices G.10-G.11, deterministic selection among feasible couplings for the same empirical reference and fixed candidate is covered by one common event; selection involving new population local quantities or candidates still requires separate calibration. Events may share scoring rows and need not be mutually independent. The L and U values of different models cannot be combined.

Let $L_{\mathrm{num}}, U_{\mathrm{num}}$ denote computed endpoints that still require numerical safety corrections, and let L, U be the corresponding statistical endpoints computed exactly. Verified deterministic envelopes $\varepsilon_L, \varepsilon_U \geq 0$ should satisfy

$$L_{\mathrm{num}} - \varepsilon_L \leq L, \qquad U \leq U_{\mathrm{num}} + \varepsilon_U.$$

The envelopes must cover optimization, marginal-repair, integration, and rounding errors not already accounted for; the same errors already included through safe quantities such as $\widehat{D}^-$ and $\widehat{\mathcal{T}}^+$ are not subtracted or added again. The quantities ϵ_D, ϵ_T are the optimality-error bounds above, used in the detection remainder, and cannot automatically be identified with $\varepsilon_L, \varepsilon_U$. When these envelopes are available, test $L_{\mathrm{num}} - \varepsilon_L > U_{\mathrm{num}} + \varepsilon_U$; when only float64 results are available without such envelopes, label the outcome a numerical positive case.

G.7 Proof of the Common Transport Event in Theorem 3

Proof. On the finite state space, take a fixed optimal 1-Lipschitz potential $f^\star$ for the population transport problem, and shift it so that $0 \leq f^\star \leq 1$. Let $z(o) = \mathbb{E}_{Q_o} f^\star$. Kantorovich–Rubinstein duality (Peyré and Cuturi, 2019, §6.1) gives

$$W(R,P) = \mathbb{E}z - \mathbb{E}_P f^\star, \qquad W(\widehat{R}_n,P) \geq \frac{1}{n}\sum_i z(O_i) - \mathbb{E}_P f^\star. \tag{45}$$

Subtract the two relations and apply the one-sided Hoeffding inequality (Hoeffding, 1963) to the fixed $[0,1]$-valued variable $z(O)$ to obtain the result. Computing $f^\star$ is unnecessary. This completes the proof.

G.8 Proofs of Joint Validity and the Comparison Under the Same Budget

Proof of Theorem 4. Appendix G.9 provides the lower-endpoint event. Theorem 3 simultaneously covers all valid empirical coupling costs with failure probability $\delta_T/2$; combining this with the events for the endpoints actually enabled gives the result. Integrating over the independent fitting results preserves the conclusion. For general C, the two-component bias formula is not invoked, and the other reserved half, $\delta_T/2$, is not used to claim a general detection guarantee. This completes the proof.

Small-amplitude endpoint. We use the radius in (G10). Replacing δ_T there by δ_{small} and using $W(R,P) \leq \mathcal{T} \leq 1$ yields $\Pr\{W(R,P) \leq U_{\mathrm{small}}\} \geq 1 - \delta_{\mathrm{small}}$. When this branch is disabled, set $U_{\mathrm{small}} = 1$ and allocate no budget to it.

Equations (51)–(52). Under the exact-optimum convention in (50), the common radius equals the default SAA radius, and $C_n \leq \widehat{\mathcal{T}}$, so $U_{\mathrm{coupling}} \leq U_{\mathrm{SAA}}$. Taking the minimum with the same optional endpoints gives (51). Let $B = \min\{1, U_{\mathrm{small}}, U_{\mathrm{anchor}}, U_{\mathrm{CVaR}}\}$ and $A = \widehat{\mathcal{T}} + r_T^W$. Then $U_{\mathrm{TV,base}} = \min\{B, A\}$ and $U_{\mathrm{TV}} = \min\{B, A - d_n\}$. Because $d_n \geq 0$, their difference is $[d_n - (A - U_{\mathrm{TV,base}})]_+$, which is (52). On the common event in (48), $L > U$ implies $\eta > 0$. Thus, the false-certification event is contained in the complement of that event, with failure probability at most δ.

G.9 Complete Definition of Lower-Endpoint Scoring

The fitting data used to determine the model to be certified, the candidate, and the lower-bound witnesses are independent of the final scoring data. Appendices G.10–G.11 additionally permit adaptive optimization of empirical feasible couplings after the objects to be certified have been fixed. Conditional on all independent fitting results, $O_1, \ldots, O_n$ are independent and identically distributed according to the target conditional law, with $n \geq 2$. Each observation provides a complete posterior probability tensor; the coordinates, category probabilities, and assignments for different groups within the same row may have arbitrary dependence. Define

$$Z_i = \sum_B \alpha_B \kappa_B(Q^B_{O_i}, U_{B,A_B(O_i)}), \qquad 0 \leq Z_i \leq s_{\mathrm{cov}} := \sum_B \alpha_B \leq 1, \qquad \widehat{e} = \frac{1}{n}\sum_i Z_i. \tag{42}$$

Define $\widehat{r}_{Bz} = n^{-1}\sum_{i=1}^n \mathbf{1}\{A_B(O_i) = z\}$ and $\widehat{\boldsymbol{r}}_B = (\widehat{r}_{Bz})_{z\in[C_B]}$. Substituting the scoring frequency vector $\widehat{\boldsymbol{r}}_B$ into (11) gives the empirical optimum $\widehat{D}$; the optimization domain for each u_g remains the categorical simplex.

First compute a valid codebook lower endpoint. Let γ_B be the within-group diameter of the frozen complete codebook, let $B_0 = \sum_B \alpha_B \gamma_B$, and take

$$q_n = B_0\sqrt{\frac{\log(2/\delta_r)}{2n}}.$$

The label-frequency bias bound a_n and its proof are given in Appendix G.1. Proposition G2 gives $\widehat{D} - q_n \leq D \leq \widehat{D} + a_n + q_n$, with both sides using the same δ_r event. The operator EB in (43) uses (G4) in Appendix G.1, retaining the two-sided Hoeffding and one-sided empirical Bernstein budgets from v10.

Also take the two-sided distortion radius $r_e^{\mathrm{H}} = s_{\mathrm{cov}}\sqrt{\log(2/\delta_H)/(2n)}$, and use the following on the intersection of its event and the empirical Bernstein event:

$$\boxed{r_e = \min\{r_e^{\mathrm{H}}, \mathrm{EB}(Z, s_{\mathrm{cov}}, \delta_e)\}, \qquad L = [\widehat{D} - \widehat{e} - r_e - \tau_* - q_n]_+.} \tag{43}$$

For general codebooks, take $\tau_* = \tau$; when all $C_B \le 2$, take $\tau_* = \tau_2$ as defined in Appendix E.6. Single-center groups incur no two-direction penalty.

This lower endpoint satisfies $\Pr\{L \le \Delta(R)\} \ge 1 - (\delta_H + \delta_e + \delta_r)$; see Appendix G.5 for the proof. Independence of the three events is not required.

The coordinatewise coverage budget $s_g \le 1$ is sufficient for validity of this lower endpoint. Under partial coverage, r_{miss} characterizes the approximation loss of the population lower bound and is not subtracted again in (43); the detection condition in Appendix G.12 directly uses the corresponding population local lower bound. Likewise, the empirical minimization bias a_n enters only the detection analysis and is not subtracted again from the valid lower endpoint.

G.10 Scope of Selection and Monotonicity of the Common Upper Endpoint

Theorem 3 (Common certification of empirical feasible transport). Conditional on all independent fitting results, fix the posterior mapping, candidate P, conditional observation law, and metric. Let $O_1, \dots, O_n$ be independent and identically distributed, and let $\widehat{R}_n = n^{-1}\sum_i Q_{O_i}$. For $\delta_W \in (0,1)$,

$$q_n^W(\delta_W) = \sqrt{\frac{\log(1/\delta_W)}{2n}}, \qquad \Pr\{W(R,P) \le W(\widehat{R}_n, P) + q_n^W(\delta_W)\} \ge 1 - \delta_W. \tag{44}$$

See Appendix G.7 for the proof.

On the same event, all valid empirical upper bounds satisfying $W(\widehat{R}_n, P) \le C_n$ simultaneously satisfy

$$W(R,P) \le \min\{1, C_n + q_n^W(\delta_W)\}. \tag{46}$$

Thus, the scoring posteriors may be used to select feasible plans, the order of the original coordinates, lifting levels, computational groups, and the tree structures and stopping masses in Section 5.4 that satisfy the exact mixture constraints, without adding a separate selection budget for each choice. The costs used must come from feasible couplings that preserve the same full source and target marginals. The candidate, posterior mapping, metric, and target conditional law must still remain fixed, and all scoring rows and their masses must be retained. Lower-bound witnesses must still be frozen before scoring; the permission for in-sample selection in this section applies only to valid empirical coupling upper bounds. We require C_n to be evaluated deterministically given the scoring rows, or replaced by a deterministic safe upper bound; a point estimate based on finitely many latent-state samples cannot replace a valid cost. This restriction does not change the independent random scoring of the observation rows O_i in (44).

If $C_n = \mathcal{H}_s(\widehat{R}_n, P)$ and the costs are exact, the endpoints under the common radius, $U_s = \min\{1, \mathcal{H}_s(\widehat{R}_n, P) + q_n^W(\delta_W)\}$, satisfy

$$W(R,P) \le U_s \quad \text{simultaneously for all } s, \qquad U_G \le \cdots \le U_0 \le \min\{1, \widehat{\mathcal{T}} + q_n^W(\delta_W)\}. \tag{47}$$

Equation (47) concerns the nested construction under a common ordering of the original coordinates and its exact optimal values. It states that the endpoints are nonincreasing for each dataset, allowing equality between levels. If different levels use different orderings or are solved approximately and independently, the resulting valid costs are still covered by the common event, but the outputs are not automatically monotone. Retaining previously verified safe costs for the same empirical problem and taking their minimum ensures that the actual endpoints are nonincreasing. Using the same q_n^W

for the tree-augmented hierarchy $\mathcal{H}_s^{\mathscr{T}}$ in Section 5.4 yields the same non-worsening relations across levels and under valid tree extensions; direct feasible tree costs can also be substituted into (46). These constructions share the same outer transport event. The common event does not control the upward bias of empirical transport; see Appendix D for the high-dimensional empirical-support obstruction.

The common event covers valid empirical couplings under the same frozen reference, candidate, and metric. It does not permit changing the candidate or lower-bound witnesses after scoring, nor does it control the upward bias of empirical transport. The small amplitude B_{cost} of the original component costs cannot be used directly as a coefficient in the common radius; see Appendix G.4 for a counterexample.

G.11 Auxiliary Endpoints, Joint Budget, and Comparison Under the Same Budget

Complete budget and applicability conditions for Theorem 4. Conditional on the independent fitting results, freeze the posterior mapping, target observation law, metric, a candidate P with any finite number of components C, and the lower-bound witnesses used; the scoring rows are independent and identically distributed. Each failure-probability budget is fixed before scoring and may depend on the independent fitting results; conditional on these results, the budgets are independent of the scoring rows. The local codebooks satisfy the coordinatewise coverage budget, and the lower endpoint L is the codebook endpoint in Appendix G.9. Let $\delta_L = \delta_H + \delta_e + \delta_r$, and retain

$$\delta_W = \delta_T/2, \qquad r_T^W = q_n^W(\delta_T/2).$$

For any valid empirical cost $W(\widehat{R}_n, P) \le C_n$, take $U_{\mathrm{coupling}} = \min\{1, C_n + r_T^W\}$. One may additionally take the minimum with calibrated optional upper endpoints. Write δ_{opt} for the sum of the budgets for these additional events, counting shared events only once. If $\delta_L + \delta_T + \delta_{\mathrm{opt}} \le \delta$, the joint validity statement in (48) of the main text holds.

Here, $C_n \le \widehat{\mathcal{T}}$ is not required. That additional condition is used only for the comparison with exact SAA below and the detection guarantee in Appendix G.12.

See Appendix G.8 for the proof.

When only the two-component anchor and CVaR endpoints are enabled, they share the DKW event, so $\delta_{\mathrm{opt}} = \delta_F + \delta_Y + \delta_2$. If small-amplitude SAA is also enabled, its dedicated budget δ_{small} is included as described below. Multiple events may use the same scoring rows and need not be independent. The results reported in this paper retain the original numerical budgets for all retained events. Allocations to removed branches that do not participate in these TV endpoints are left unused and are not reassigned; the total budget need only be no larger than the stated δ.

Strong statistical baseline. For two-component candidates, U_{SAA} in this section refers specifically to the default endpoint in (G10) with $B_{\mathrm{cost}} = 1$. If a valid small-amplitude bound is additionally available before scoring, define, with a dedicated budget $\delta_{\mathrm{small}} > 0$,

$$r_{\mathrm{small}} = B_{\mathrm{cost}} \sqrt{\frac{\log(2/\delta_{\mathrm{small}})}{2n}}, \qquad U_{\mathrm{small}} = \min\{1, \widehat{\mathcal{T}} + r_{\mathrm{small}}\}.$$

This endpoint covers $W(R, P)$ with probability at least $1 - \delta_{\mathrm{small}}$; see Appendix G.8 for the proof. When it is disabled, set $U_{\mathrm{small}} = 1$ and allocate no budget to it. The comparison baseline is $U_{\mathrm{TV,base}} = \min\{U_{\mathrm{SAA}}, U_{\mathrm{small}}, U_{\mathrm{anchor}}, U_{\mathrm{CVaR}}\}$. The definitions, exact sorting-based computations, and budgets for the default SAA, anchor, and CVaR endpoints are given in Appendices G.2–G.3. These are statistical baselines and optional endpoints, rather than the coupling constructions of Section 5. The new combination retains the same optional endpoints:

$$U = U_{\mathrm{TV}} = \min\{U_{\mathrm{coupling}}, U_{\mathrm{small}}, U_{\mathrm{anchor}}, U_{\mathrm{CVaR}}\}. \tag{49}$$

When small-amplitude SAA is enabled, $\delta_{\mathrm{opt}} = \delta_F + \delta_Y + \delta_2 + \delta_{\mathrm{small}}$; the new and original combinations share the same small-amplitude event, whose budget is counted only once. Below, $U_{\mathrm{TV,base}}, U_{\mathrm{TV}}$ refer to the combinations above with the endpoints that are actually enabled.

Exact-optimum convention. First consider implementations that guarantee $W(\widehat{R}_n, P) \le C_n \le \widehat{\mathcal{T}}$, where $\widehat{\mathcal{T}}$ is the exact optimum in (G5) or of the corresponding transport problem. Let

$$r_T^W = q_n^W(\delta_T/2) = \sqrt{\frac{\log(2/\delta_T)}{2n}}, \qquad U_{\mathrm{coupling}} = \min\{1, C_n + r_T^W\}, \quad W(\widehat{R}_n, P) \le C_n \le \widehat{\mathcal{T}}. \tag{50}$$

The cost C_n may be an exact unified hierarchy value, a stopping-tree transport value, a tree-augmented hierarchy value, or a safe cost separately verified not to exceed $\widehat{\mathcal{T}}$. Merely retaining a general original primal feasible cost $\widehat{\mathcal{T}}^+$ does not automatically satisfy the latter requirement, because $\widehat{\mathcal{T}}^+ \ge \widehat{\mathcal{T}}$; such implementations use the numerical feasible-cost convention below.

As long as the posterior tensors, row masses, candidate, and metric define the same empirical problem, previously verified valid feasible costs may be retained under the same event. An old cost value from a different training checkpoint cannot be used for this fallback. In the two-component case, the other reserved half of the budget, $\delta_T/2$, controls $\widehat{\mathcal{T}} \le \mathcal{T} + \beta_n + r_T^W$, so the detection corollary in Appendix G.12 still has a common event under the same budget. If only upper-bound validity is claimed, the entire δ_T may be used in (44), but the original detection event cannot then also be retained.

When the original (G10) uses $B_{\mathrm{cost}} = 1$, we have $r_T^W = q_T$. Keeping the other endpoints and their budgets unchanged and using the actual combinations $U_{\mathrm{TV,base}}$ above and (49), the replacement satisfies

$$\begin{gathered} U_{\mathrm{coupling}} \le U_{\mathrm{SAA}}, \qquad U_{\mathrm{TV}} \le U_{\mathrm{TV,base}} \\ \text{for every dataset,} \\ \text{without increasing the total failure budget.} \end{gathered} \tag{51}$$

This guarantees that the final combination is no worse, but does not guarantee that the new branch alone is below the anchor or CVaR endpoint. Let $d_n = \widehat{\mathcal{T}} - C_n$. Including the effect of capping, the exact final gain is

$$U_{\mathrm{TV,base}} - U_{\mathrm{TV}} = \left[d_n - \left(\widehat{\mathcal{T}} + r_T^W - U_{\mathrm{TV,base}} \right) \right]_+ . \tag{52}$$

Obtaining an additional certification also requires this gain to exceed the previous gap $U_{\mathrm{TV,base}} - L \ge 0$.

Numerical and range conditions. Practical solvers must retain the previous valid feasible cost, repair the full marginals, and apply a safe upward correction. When feasibility is available without an envelope for the optimality error, the guarantee is non-worsening relative to the original feasible endpoint under the same numerical convention; dominance over exact SAA cannot be claimed automatically. See Appendix G.4 for details. Small-range SAA may have a separate statistical advantage and therefore cannot be removed on the basis of the common event.

Equation (48) also gives

$$\Pr\{L > U \text{ and } \eta \le 0\} \le \delta.$$

This is a frequentist false-certification guarantee, not a posterior probability of correctness after certification. When $L \le U$, the output is insufficient evidence. If multiple full candidate distributions P, conditions, or sample sizes are selected on the basis of the scoring results, their simultaneous validity requires separate calibration; selection among feasible couplings for the same candidate is already covered by the common transport theorem.

G.12 Conditional Detection Guarantee for Two-Component Candidates

Corollary 2. Finite-sample detection condition based on the population certificate gap. Under the conditions of Theorem 4, further restrict the candidate to at most two components, use the codebook lower endpoint and the exact-optimum convention, and require $C_n \le \widehat{\mathcal{T}}$. Use the other half of δ_T for the event $\widehat{\mathcal{T}} \le \mathcal{T} + \beta_n + r_T^W$; together with the validity event, the total budget remains δ_T. On this extended common event, write the population local lower bound corresponding to (43) and the total scoring-error term as

$$\ell_* = [D - e - \tau_*]_+, \qquad S_n = a_n + \beta_n + 2q_n + 2r_T^W + r_e^{\mathrm{H}} + r_e.$$

Here, τ_* follows the choice in Appendix G.9. Under either exact or partial coverage, $\ell_* \le \Delta(R)$, and

$$\boxed{L - U \ge (\ell_* - \mathcal{T}) - S_n.} \tag{53}$$

Thus, on this common event, $\ell_* - \mathcal{T} > S_n$ is sufficient for $U < L$. Because r_e depends on the scoring sample, S_n is a random threshold; conditional on all independent fitting results, its probabilistic interpretation is

$$\Pr\{\ell_* - \mathcal{T} > S_n,\ L \le U\} \le \delta.$$

Using $r_e \le r_e^{\mathrm{H}}$, which holds for every dataset, define the threshold that can be determined before scoring:

$$\overline{S}_n = a_n + \beta_n + 2q_n + 2r_T^W + 2r_e^{\mathrm{H}}, \qquad S_n \le \overline{S}_n.$$

Conditional on all independent fitting results, $\ell_*, \mathcal{T}, \overline{S}_n$ are fixed, yielding the ex ante detection guarantee

$$\boxed{\ell_* - \mathcal{T} > \overline{S}_n \quad \Longrightarrow \quad \Pr\{L > U\} \ge 1 - \delta.}$$

The probability here is only over the independent scoring rows. This deterministic condition is more conservative than the within-event condition using S_n, but it does not change the actual endpoints. Both conditions use the population certificate gap formed by the current lower bound and the direct transport baseline; tightening through other valid upper endpoints may still allow the actual certificate to pass.

See Appendix G.6 for the proof. For two-component candidates, take $\beta_n = \kappa(V_1, V_2)\rho_n$; for single-component candidates, take $\beta_n = 0$. The definition of a_n is given in Appendix G.1.

A single-component candidate cannot satisfy the strictly positive gap condition above; $\beta_n = 0$ retains the degenerate case of the statistical formulas, and validity continues to hold. See Appendix G.6 for the proof.

This result does not require control of $\mathcal{T} - W(R, P)$, nor does it imply guaranteed detection from an arbitrary true improvement $\eta > 0$. The effect of partial coverage on the population lower bound is already reflected in ℓ_*, so no additional uncovered contribution is subtracted in (53). The quantities a_n and β_n are the empirical biases of minimization and maximization, respectively, and both affect detection. When the frozen group family, cover weights, codebook sizes of all groups, and failure probabilities are fixed, the pure scoring terms are $O(n^{-1/2})$.

This section retains the radii and budget split from v10; it does not switch to smaller, purely one-sided radii. Validity for general finite C does not depend on this corollary, and this corollary does not fully characterize the detection capability of the new hierarchy. See Appendices G.4 and G.6 for the additional errors introduced on the right-hand side by approximate optimization and numerical computation.

H Proofs for Certifiable Structures and Capacity Bounds

H.1 Theorem 5, Positive Margins, and the Finite-Sample Connection

Proof. Let the number of positive-sign modules be $N_+ \sim \mathrm{Bin}(M, 1/2)$. By Equation (3), the costs to the negative and positive candidate components are $2aN_+/M$ and $2a(M-N_+)/M$, respectively. For every observation, the cost of any valid assignment is at least the smaller of these two costs. Therefore,

$$\mathcal{T} \geq \mathbb{E} \min \left\{ \frac{2aN_+}{M}, \frac{2a(M-N_+)}{M} \right\} = a - \frac{2a}{M}\mathbb{E}|N_+ - M/2| = a - 2aA_M.$$

Choose the negative component when $N_+ < M/2$ and the positive component when $N_+ > M/2$, splitting the mass equally at ties. Since N_+ and $M - N_+$ have the same distribution, the expected mass assigned to each component is exactly $1/2$, satisfying the weight constraints in Equation (22) and attaining the lower bound above. Thus, the first relation holds with equality. The second relation follows from $\ell_2 = [a - \rho_b]_+ \geq a - \rho_b$. This completes the proof.

Positive margin and group width. Here $0 < A_M \leq 1/2$. If $a \leq \rho_b$, the margin is $-\mathcal{T} \leq 0$; otherwise, it equals $2aA_M - \rho_b$. The inequality in Equation (56) therefore cannot be replaced by an equality for all parameter values. Since $\rho_b \leq 1/(2\sqrt{b})$, the condition $b > 1/(16a^2A_M^2)$ is sufficient for a positive margin. Moreover, Appendix A.2 proves that $A_M \geq 1/(2\sqrt{2M})$ for every $M \geq 1$, giving the explicit finite-dimensional guarantee

$$2aA_M - \rho_b \geq \frac{a}{\sqrt{2M}} - \frac{1}{2\sqrt{b}} \geq \left(\frac{1}{\sqrt{2}} - \frac{1}{2}\right)\frac{a}{\sqrt{M}} > 0 \qquad \left(b \geq \frac{M}{a^2}\right).$$

Together with $A_M \sim (2\pi M)^{-1/2}$ and $\ell_2 \leq a$, this implies that the population margin is of order $a/\sqrt{M}$ under this group-width condition. Under the conditions of Appendix G.12, take $\ell_* = \ell_2$ and substitute the margin lower bound from Theorem 5 into Equation (53) to obtain $L - U \geq 2aA_M - \rho_b - S_n$ in Appendix H.3. Using $S_n \leq \overline{S}_n$ then gives an a priori high-probability detection condition. For approximate optimization, subtract $\epsilon_D + \epsilon_T$ as in G.6.

Within-module correlation. In this binary construction, when $b \geq 2$, distinct state coordinates in the same module under the aggregate reference R satisfy $\mathrm{Cov}(X_g, X_{g'}) = \mathrm{Var}(1/2 + aZ_j) = a^2$ and $\mathrm{Var}(X_g) = 1/4$, so their correlation coefficient is $4a^2$. The observation-wise posterior remains independent conditional on the module signs.

H.2 Complete Proof of Theorem 6

Proof. Ignore zero-weight components and take an optimal state coupling ζ^* of R, P. Whenever $P(y) > 0$, augment the coupling with the candidate component label H according to the conditional probability

$$\Pr(H = h \mid X = x, Y = y) = \frac{\pi_h V_h(y)}{P(y)}$$

States with $P(y) = 0$ carry no mass. This lifting preserves the joint law of (X, Y) and its transport cost, and satisfies $\Pr(H = h) = \pi_h$ and $\mathcal{L}(Y \mid H = h) = V_h$. Denote the conditional source distribution and conditional coupling under this lifted coupling by

$$S_h = \mathcal{L}(X \mid H = h), \qquad \Gamma_h = \mathcal{L}(X, Y \mid H = h).$$

Then $\sum_h \pi_h S_h = R$ and $\Gamma_h \in \mathrm{Cpl}(S_h, V_h)$, so

$$W(R, P) = \sum_h \pi_h \mathbb{E}_{\Gamma_h} d_G(X, Y) \geq \sum_h \pi_h W(S_h, V_h).$$

Conversely, for each h, take an optimal coupling $\widetilde{\Gamma}_h$ of S_h, V_h. The mixture plan $\sum_h \pi_h \widetilde{\Gamma}_h$ has source and target marginals R, P, respectively, so its cost gives the reverse inequality

$$W(R,P) \le \sum_h \pi_h W(S_h, V_h).$$

Thus, the relation above is in fact an equality; the S_h here are determined by the globally optimal coupling. Moreover, since every V_h belongs to the product family over all coordinates, the triangle inequality gives

$$\Delta(R) \le W(R, V_h) \le W(R, S_h) + W(S_h, V_h).$$

Rearranging and taking the π_h-weighted sum yields

$$W(R,P) = \sum_h \pi_h W(S_h, V_h) \ge \Delta(R) - \sum_h \pi_h W(S_h, R). \tag{58}$$

For each retained positive-weight component, $R(x) = \sum_j \pi_j S_j(x) \ge \pi_h S_h(x)$ implies $S_h \ll R$, and $\mathrm{KL}(S_h \| R) \le \log(1/\pi_h) < \infty$. The transport-entropy bound in Appendix A.1 therefore applies without requiring full support of R or module independence of S_h. For any 1-Lipschitz function under d_G, changing one entire module changes its value by at most $(G/M)/G = 1/M$. Under the module-product reference R, the bounded-differences moment-generating-function bound (McDiarmid, 1989) thus gives $\log \mathbb{E}_R e^{\lambda(f - \mathbb{E}_R f)} \le \lambda^2/(8M)$. Combining this with the KL variational formula and Wasserstein duality in Appendix A.1 yields

$$W(S_h, R) \le \sqrt{\frac{\mathrm{KL}(S_h \| R)}{2M}}.$$

The lifted joint law is $\Pr(X = x, H = h) = \pi_h S_h(x)$, with marginals R and π. Therefore,

$$\sum_h \pi_h \,\mathrm{KL}(S_h \| R) = \sum_{h,x} \pi_h S_h(x) \log \frac{S_h(x)}{R(x)} = I(X;H).$$

Finally, Equation (58), concavity of the square root, and the entropy upper bound give

$$\begin{aligned} \eta &\le \sum_h \pi_h W(S_h, R) \\ &\le \sqrt{\frac{\sum_h \pi_h \,\mathrm{KL}(S_h \| R)}{2M}} = \sqrt{\frac{I(X;H)}{2M}} \\ &\le \sqrt{\frac{H(\pi)}{2M}} \le \sqrt{\frac{\log C}{2M}}. \end{aligned}$$

This proves Equation (57), where $H(\pi)$ is the entropy of the candidate component weights. This completes the proof.

H.3 Independent Soft-Module Construction and Explicit Detection Conditions

Partition the $G = Mb$ binary coordinates as $[G] = B_1 \sqcup \cdots \sqcup B_M$, where $|B_j| = b$ and $M, b \ge 1$. Let $Z_1, \ldots, Z_M \in \{-1, +1\}$ be independent fair signs, and set $q_g(1) = 1/2 + aZ_j$ within module j, where $0 < a < 1/2$. The observation-wise posterior remains a product over all coordinates. After marginalizing these independent signs, the aggregate reference is

$$R = \bigotimes_{j=1}^{M} \left[\tfrac{1}{2} \operatorname{Bern}(1/2 - a)^{\otimes b} + \tfrac{1}{2} \operatorname{Bern}(1/2 + a)^{\otimes b} \right].$$

Choose a candidate consisting of two global product distributions:

$$P_2 = \tfrac{1}{2} \operatorname{Bern}(1/2 - a)^{\otimes G} + \tfrac{1}{2} \operatorname{Bern}(1/2 + a)^{\otimes G}. \tag{54}$$

Take precisely these modules as the lower-bound groups, so $\mathcal{B} = \{B_1, \ldots, B_M\}$, $w_{B_j} = 1$, and $\alpha_{B_j} = 1/M$. The two centers in group j are $U_{j,\pm} = \operatorname{Bern}(1/2 \pm a)^{\otimes b}$, assigned according to Z_j. Both label frequencies are $1/2$, and $\gamma_{B_j} = \kappa_{B_j}(U_{j,-}, U_{j,+}) = 2a$. By Equations (12) and (21),

$$e_{B_j} = 0, \quad D_{B_j} = a, \qquad e = 0, \quad D = a, \quad \tau_2 = \rho_b, \quad \ell_2 = [a - \rho_b]_+.$$

Write

$$A_M = \frac{1}{M} \mathbb{E} |\operatorname{Bin}(M, 1/2) - M/2|. \tag{55}$$

Theorem 5. Certifiability of soft modules. For the candidate in Equation (54),

$$\mathcal{T} = a - 2aA_M, \qquad \ell_2 - \mathcal{T} \geq 2aA_M - \rho_b. \tag{56}$$

The proof is given in Appendix H.1.

Positive population margin and finite group width. The actual population certificate margin satisfies

$$\ell_2 - \mathcal{T} = [a - \rho_b]_+ - a + 2aA_M, \qquad \ell_2 - \mathcal{T} > 0 \iff \rho_b < 2aA_M.$$

A sufficient condition for a positive margin is $b > 1/(16a^2 A_M^2)$. A simpler conservative condition is

$$b \geq \frac{M}{a^2} \quad \Longrightarrow \quad \ell_2 - \mathcal{T} \geq \left(\frac{1}{\sqrt{2}} - \frac{1}{2} \right) \frac{a}{\sqrt{M}} > 0.$$

Under this group-width condition, the population margin is of order $a/\sqrt{M}$. The condition and scaling are derived in Appendix H.1.

Use the scoring procedure, budgets, and exact-optimum conditions in Appendix G.12, and take $\tau_* = \tau_2 = \rho_b$, so that $\ell_* = \ell_2$. On the extended common event in that subsection, Equation (53) gives

$$L - U \geq 2aA_M - \rho_b - S_n.$$

Therefore, $S_n < 2aA_M - \rho_b$ is a sufficient condition for detection on this event. Using $\overline{S}_n$, which can be determined before scoring as described in that subsection, gives the a priori guarantee

$$2aA_M - \rho_b > \overline{S}_n \quad \Longrightarrow \quad \Pr\{L > U\} \geq 1 - \delta.$$

The probability is conditional on all independent fitting results and concerns only the scoring rows. For approximate optimization, $\epsilon_D + \epsilon_T$ must additionally be subtracted from the margin lower bound, as specified in Appendix G.6.

When $b \geq 2$, the correlation coefficient between distinct state coordinates in the same module under the aggregate reference is $4a^2$. The observation-wise posterior remains independent conditional on

the module signs; see Appendix H.1 for the calculation. At the boundaries, $M = 1$ gives $P_2 = R$ and $\mathcal{T} = 0$, while $b = 1$ gives $R = \mathrm{Bern}(1/2)^{\otimes M} \in \mathcal{F}$, so no positive true improvement exists, consistently with the margin condition above.

H.4 Candidate Capacity Limit for Equally Sized Independent Modules

Theorem 6. Upper bound on improvement by global mixtures of products. Suppose there is a coordinate partition $[G] = B_1 \sqcup \cdots \sqcup B_M$ such that $|B_j| = G/M$ and $R = \bigotimes_{j=1}^{M} R^{B_j}$. Dependence within modules is allowed, and the module distributions need not be identical. Then every candidate consisting of C product components over all coordinates satisfies

$$\eta = \Delta(R) - W(R, P) \le \sqrt{\frac{\log C}{2M}}. \tag{57}$$

The proof is given in Appendix H.2.

Thus, achieving a fixed positive improvement $\eta > 0$ requires $C \ge \exp(2M\eta^2)$; this is a necessary condition, not a sufficient one. For fixed C, improvement is at most of order $M^{-1/2}$. The two-component candidate in Theorem 5, with fixed $a > 0$ and sufficiently large group width, is compatible with this scaling. This result does not require exponentially many components for every arbitrarily small positive improvement, nor is it equivalent to a component-count lower bound for exact representation.

These two results provide a certifiable structure and a capacity limit on approximate improvement, respectively. The capacity result requires the reference to be a product of equally sized, mutually independent modules and cannot be generalized into a theorem for arbitrary aggregate posteriors. The dimension calibration in Appendix J.8 retains both positive and negative population margins.

I Existing Finite-Space Mechanism and Endpoint Checks

I.1 Transport Tightness in Analytical High-Dimensional Structures

For the soft-module construction in Appendix H.3, push R, P_2 forward through the normalized count map $x \mapsto G^{-1}\sum_g x_g$. Let W_{count} denote the first-order Wasserstein distance between the two pushforward distributions under the absolute-value metric. Then:

$$W_{\mathrm{count}} \le W(R, P_2) \le \mathcal{T}, \qquad \mathcal{T} - W(R, P_2) \le \mathcal{T} - W_{\mathrm{count}}. \tag{I1}$$

The reference normalized-count distribution is obtained by convolving the within-module binomial mixtures and dividing the count by G. The candidate normalized-count distribution is $\frac{1}{2}\mathcal{L}(B_-/G) + \frac{1}{2}\mathcal{L}(B_+/G)$, where $B_\pm \sim \mathrm{Bin}(G, 1/2 \pm a)$, requiring no enumeration of the 2^G states.

Table I1. High-dimensional population transport checks for soft modules

M	b	G	a	$\mathcal{T}$	W_{count}	Upper bound on slack
8	256	2048	0.20	0.145313	0.143104	0.002209
8	1024	8192	0.10	0.072656	0.071475	0.001181
8	4096	32768	0.05	0.036328	0.035729	0.000600
32	4096	131072	0.10	0.086005	0.085854	0.000151

All four settings satisfy $2aA_M - \rho_b > 0$. When $a = 0.05$, the within-group correlation coefficient is 0.01; at the largest dimension, 131072, the projection gives a coupling-slack upper bound of 1.51×10^{-4}. The count distance here is a lower bound, not the exact value of the full distance.

I.2 Finite-Distribution Checks of Asymmetric Endpoints

The original check of the local TV optimum enumerated 437 multinomial count vectors across four designs, including dependence from shared scoring rows and overlapping groups. Table I2 retains only the records of $D - \mathbb{E}\widehat{D}$. It does not report the withdrawn categorywise CDF domination quantities or the numerical values of the old geometric bias bound, nor does it treat the original records as a new numerical check of the current Eq. (G1).

Table I2. Finite-distribution checks of local minima

Design	Number of count vectors	$D - \mathbb{E}\widehat{D}$
Three category vertices	66	0.159325
Two soft centers	41	0.050148
Overlapping groups, shared scoring rows	165	0.022192
Four soft centers	165	0.034245

On the transport side, 249 count vectors across four designs were enumerated. The three-component design checks only one-sided validity and does not use the two-component bias formula.

Table I3. Finite-distribution checks of transport maxima

Design	n	C	Number of count vectors	$\mathcal{T}$	$\mathbb{E}\widehat{\mathcal{T}} - \mathcal{T}$	Two-component bias upper bound β_n
Attainable binomial bias	10	2	11	0.000000	0.103022	0.103022
Three soft source components	8	2	45	0.210000	0.010673	0.115637
Multicategory TV cost	8	2	165	0.374235	0.000496	0.085620
Three candidate components	6	3	28	0.330610	0.024577	Not applicable

The failure budget is 0.2. All enumerated one-sided endpoint failure rates are zero, and the maximum failure rate of the two-component asymmetric intervals is approximately 0.000377. The first row attains the bias upper bound $\gamma\rho_n$. The checks also compare the sorting algorithm with linear programming; the computations use float64, floating-point linear programming, and a tolerance of 2×10^{-8}.

I.3 Paired Comparisons with the Strong CVaR Baseline

Take $a = 0.2$, $c_1 = 2aZ_{\text{cost}}$, $c_2 = 2a(1 - Z_{\text{cost}})$, and two equally weighted components. The atomic condition uses $Z_{\text{cost}} = \text{Bin}(8, 1/2)/8$, and the continuous condition uses $Z_{\text{cost}} \sim \text{Unif}[0, 1]$; both can be realized by posteriors between two Bernoulli candidate components. The first two conditions use $t_0 = 0$, and the third uses the continuous distribution and $t_0 = 0.25$. All methods share the known cost range $B_{\text{cost}} = \gamma = 0.4$.

Each condition uses $n \in \{4096, 16384\}$, with 40 repetitions per setting, for a total of 240 independent scoring streams. Within each stream, the costs and $\delta_j = 0.05/7$ are shared. The table reports excesses of the upper bounds over population transport, not counts of structural replacement certifications.

Table I4. Paired comparisons with strong baselines: mean $U - \mathcal{T}$, lower is better

Condition	n	CVaR envelope	Asymmetric SAA	Integrated anchor	Three-endpoint combination
Eight-module atomic	4096	0.013167	0.010410	0.003557	0.003557
Eight-module atomic	16384	0.006438	0.005189	0.001440	0.001440
Continuous probabilities	4096	0.017835	0.010477	0.004713	0.004713
Continuous probabilities	16384	0.008592	0.005241	0.001934	0.001934
Threshold away from the optimum	4096	0.017885	0.010537	0.013315	0.010537
Threshold away from the optimum	16384	0.008658	0.005315	0.006313	0.005315

None of the 240 streams had an observed upper endpoint below population transport. The balancing remainder vanishes in the atomic condition, and the continuous condition benefits from integral tightening. When the threshold departs from the optimum, SAA dominates the combination.

I.4 Finite-Space Consistency Checks of the Unified TV Hierarchy

The original `check_hierarchy.py` verification batch and its `hierarchy_check_results.json` contain 27 full-support binary product mixtures: 3 specified two-block examples and 24 random examples. The random examples cover 1 to 3 blocks, 1 to 3 source/target components, and nonuniform weights. All hierarchy levels and an independent full-state transport problem are solved to check $\mathcal{T}_{\varnothing} \geq \mathcal{H}_0 \geq \cdots \geq \mathcal{H}_G = W$. The maximum floating-point violation of the hierarchy relations is 2.22×10^{-16}, the maximum difference between the full level and independent OT is 1.11×10^{-16}, and the maximum marginal residual is 4.16×10^{-16}.

In the three two-block Bernoulli examples, $V(p_1, p_2) := \mathrm{Bern}(p_1) \otimes \mathrm{Bern}(p_2)$; whenever two components are listed on the same side, they have equal weights.

Source mixture	Candidate mixture	Direct transport	$\mathcal{H}_0$	$\mathcal{H}_1$	$W = \mathcal{H}_2$
$V(.5, .5)$	$V(.1, .1), V(.9, .9)$	0.40	0.40	0.20	0.16
$V(.1, .1), V(.9, .9)$	$V(.5, .5)$	0.40	0.16	0.16	0.16
$V(.1, .1), V(.9, .9)$	$V(.7, .7)$	0.40	0.26	0.20	0.20

These calculations check only the constructions, ordering, and full-level equality. They are not statistical coverage experiments and do not include rigorous rounding certification. The historical review of Sections 7 to 8 reran the same script on a copy, yielding JSON identical to the original archive. This reproduces the original records and does not add 27 independent cases. The source script and original results are preserved among the historical materials in the reproduction package.

I.5 Mechanism Comparisons for Stopping-Tree Transport

The following cost checks use complete finite populations and exclude statistical radii. First consider a three-block soft-posterior mechanism example: $h \in \{-1, +1\}$ has equal weights, and the candidate roots are $V_h = \mathrm{Bern}(1/2 + 0.2h)^{\otimes 2} \otimes \mathrm{Bern}(1/2)$. The first two blocks of each source posterior associated with a root agree with that root, while the third-block probabilities are $0.2, 0.5, 0.8$, with

conditional weights $1/4, 1/2, 1/4$. Exactly split the root's third block into two equally weighted leaves with probabilities $0.2, 0.8$.

Quantity on the same finite population	Value
Original direct component transport	0.050000
Component transport forced to use both leaves	0.050000
Tree transport retaining the parent node	0
Existing prefix aggregation $\mathcal{H}_0$	0

The middle posterior can stop at the parent node, while the two extreme posteriors can be sent to exactly matching leaves, giving zero tree cost. Since $R = P$ and the three posteriors within each true root average to exactly V_h, the existing $\mathcal{H}_0$ in the current manuscript is also zero. This example shows that retaining the parent node improves on forced refinement and the original direct baseline, but does not establish improvement over the existing full-prefix aggregation.

Now take full-support source posteriors

$$\begin{aligned} Q_- &= \mathrm{Bern}(\varepsilon)^{\otimes 3}, \qquad Q_+ = \mathrm{Bern}(1-\varepsilon)^{\otimes 3}, \\ R &= \tfrac{1}{2}Q_- + \tfrac{1}{2}Q_+, \qquad P = \mathrm{Bern}(1/2)^{\otimes 3}, \\ & 0 < \varepsilon < 1/2. \end{aligned} \tag{I2}$$

There is only one candidate root, so $\mathcal{H}_0 = \mathsf{Pref}(R, P)$. By exchangeability and the prefix conditional probabilities, the cost of the first step is zero, and the unnormalized expected TV at each of the next two steps is $\frac{1}{2}(1-2\varepsilon)^2$. Consequently, $\mathcal{H}_0 = (1-2\varepsilon)^2/3$, independently of the coordinate order.

Perform exact coordinatewise binary refinement of the fair product candidate, stopping mass $1/4$ at each of $00*$ and $11*$, where $*$ denotes the fair third block that has not yet been conditioned. Allocate mass $1/8$ to each of the remaining four point leaves $010, 011, 100, 101$. Source Q_- allocates $1/4, 1/8, 1/8$ to $00*, 010, 100$, and Q_+ allocates symmetrically. This valid plan has cost $1/4 + \varepsilon/3$. When $\varepsilon = 0.01$:

Cost	Value
Direct component transport	0.490000
Original prefix aggregation $\mathcal{H}_0$	0.320133
The feasible parent-stopping plan above	0.253333
Full-state transport $W(R, P)$	0.240100

The full distance is $1/4 - \varepsilon + \varepsilon^2$: relative to the fair product, each of the two extreme states has excess mass $3/8 - 3\varepsilon/2 + 3\varepsilon^2/2$, which is transported along single-bit flips to its three neighboring states. The total cost is the total excess mass divided by 3, and the minimum nonzero distance also gives a lower bound. At $\varepsilon = 0.01$, this parent-stopping plan strictly reduces the original $\mathcal{H}_0$, and the jointly expanded $\mathcal{H}_0^{\mathscr{T}}$ is at most the cost of this plan. This does not claim domination of deeper state levels. The candidate P itself belongs to the original product family, so this example checks only upper-bound tightening and cannot yield a structural replacement certificate. The source posteriors and candidate all have full support; point leaves in the tree are used only for TV.

These numerical values have been checked for float64 consistency using an independent full-state LP and a tree-flow LP. The tree plan and strict comparison in the table also have the analytical witnesses given above. These checks do not replace finite-sample coverage or rigorous rounding certification.

J Branch Endpoints and Configurations of the Current Experiments

J.1 All Comparison Columns for the Three-Block Experiments

Using the endpoint notation of Appendix J.6, every table entry is the mean over 20 repetitions per setting. The recorded fields `C_min` and `U_min` correspond to C_{new} and U_{new}, respectively.

ϵ	n	$\widehat{\mathcal{T}}$	C_{prefix}	C_{tree}	U_{prefix}	U_{tree}	U_{new}
0.0	4096	0.051474	0.002683	0.002573	0.028909	0.028800	0.028800
0.0	65536	0.050369	0.000573	0.000560	0.007130	0.007117	0.007117
0.0	262144	0.050227	0.000341	0.000334	0.003619	0.003612	0.003612
0.03	4096	0.057364	0.012373	0.013148	0.038599	0.039375	0.038599
0.03	65536	0.055346	0.010366	0.010558	0.016923	0.017115	0.016923
0.03	262144	0.055243	0.010260	0.010346	0.013538	0.013624	0.013538
0.09	4096	0.066463	0.031294	0.032022	0.057521	0.058249	0.057521
0.09	65536	0.065468	0.030429	0.030581	0.036986	0.037138	0.036986
0.09	262144	0.065194	0.030122	0.030268	0.033401	0.033546	0.033401

J.2 Retained Strong Statistical Endpoints

Default SAA denotes U_{SAA} with $B_{\mathrm{cost}} = 1$, while small-range SAA denotes U_{small} from Appendix G.11; the strong combined baseline is U_{base} from Appendix J.6. The following table likewise reports means over 20 repetitions per setting.

ϵ	n	Default SAA	Small-range SAA	anchor	CVaR	Strong combined baseline
0.0	4096	0.077701	0.061091	0.062268	0.072894	0.061091
0.0	65536	0.056926	0.052774	0.052847	0.055649	0.052774
0.0	262144	0.053505	0.051429	0.051447	0.052796	0.051429
0.03	4096	0.083591	0.066980	0.067925	0.076727	0.066980
0.03	65536	0.061903	0.057750	0.057765	0.060306	0.057750
0.03	262144	0.058521	0.056445	0.056434	0.057634	0.056434
0.09	4096	0.092690	0.076080	0.076779	0.087120	0.076080
0.09	65536	0.072025	0.067872	0.067827	0.070437	0.067827
0.09	262144	0.068472	0.066396	0.066354	0.067666	0.066354

These columns share the same scoring stream and event budget; the mean of the minima is not the minimum of the means. All per-run tree and prefix costs and upper endpoints are provided in the CSV files. Failure gaps from the complete codebook audit are stored separately in the high-dimensional records.

J.3 Implementation and Random-Stream Conventions

Fitting-data generation, scoring-data generation, and parameter initialization use independent SeedSequence substreams. High-dimensional audits allocate seeds by configuration and repetition index starting from 202609240000; three-block audits allocate seeds starting from 202609230000. Fixed code, array precision, batch sizes, and environment versions are used for reconstruction. Candidates or lower-bound witnesses must not be reselected based on scoring results. The two-center objective subject to category-simplex constraints, transport sorting, tree LP, and full-state LP all have independent consistency checks. High-dimensional runtimes are the original records of the complete codebook audits. Three-block records retain the original batch runtimes and are not used as runtimes for a pure upper-endpoint implementation after interface removal.

J.4 Common Experimental Protocol and Numerical Conventions

The current experiments examine whether objects that are learned and then frozen can pass codebook certification, and whether the additional couplings reduce costs and final upper endpoints under different reference structures. All experiments use synthetic posteriors; none uses a trained RSSM. The incompletely reproduced summary from the early V0 version has been moved to a separate working archive and does not support the results in this section.

Configurations and seeds were fixed before the original batches were run, and all configurations and repetitions are retained. Appendix J.5 uses complete codebook certification with a total budget of $\delta = 0.05$ per run; Appendices J.6-J.7 retain the upper-endpoint comparisons from the original scoring streams, with budget allocations specified in the respective subsections. All methods share the same reference, candidate, scoring stream, and upper-endpoint event budget. All results use float64; the complete audits declare a numerical pass when $L - U > 10^{-9}$, without claiming rigorous rounding certification or simultaneous coverage across the entire batch.

J.5 Complete Synthetic Audits of Objects Learned and Then Frozen

We reconstruct the local learning setup of v10 using new independent random streams, rather than retroactively validating the original 80 historical runs. The coordinates of M independent modules, each containing b coordinates, are randomly permuted. Two distinct categories are selected for each coordinate, and the full probability vectors are generated as

$$p_{zg} = \frac{\sigma}{K}\mathbf{1}_K + (1-\sigma)e_{k_{zg}}, \qquad q_{og} = (1-\zeta)p_{Z_{j(g)},g} + \zeta N_{og}, \tag{59}$$

where $K = 4$, $z \in \{-1, +1\}$, $k_{-1,g} \neq k_{+1,g}$ belong to $[K]$, e_k is a standard basis vector, and $j(g)$ is the module index. Each observation o independently draws module labels $Z_j(o)$, which are also independent and fair across modules; Eq. (59) abbreviates $Z_j = Z_j(o)$. The variables $N_{og} \sim \mathrm{Dirichlet}(\mathbf{1}_K)$ are independent across observations and coordinates. The strong, soft, and independent conditions use $(\sigma, \zeta) = (0.02, 0.01), (0.30, 0.08), (1, 0.15)$, respectively.

Each run uses 512 independent fitting observations to select the category feature with the largest variance. Disjoint groups of at least 8 coordinates are extracted from anchor neighborhoods whose absolute correlation exceeds 0.8. Within each group, two centers are fitted using farthest-point initialization followed by 8 rounds of assignment and mean updates; after sorting along a prespecified direction, the centers are concatenated to form the candidates. Uncovered coordinates use their fitting means. Candidate weights are obtained from the nearest-candidate frequencies among the fitting rows, with exact ties split equally and weights clipped to $[1/512, 511/512]$. These rules, groups, codebooks, and candidates are all frozen before scoring.

Independent scoring observations are generated in batches of 128 rows, and the current codebook lower endpoint in Eq. (43) is used. The upper endpoint is the minimum of the common endpoint for direct transport, the anchor endpoint, the CVaR endpoint, and the small-range SAA endpoint. The eight budget categories $\delta_H, \delta_e, \delta_r, \delta_T, \delta_F, \delta_Y, \delta_2, \delta_{\mathrm{small}}$ each receive $0.05/8$; δ_T is again divided equally. We fix $t_0 = 0$. The model provides the valid range bound $B_{\mathrm{cost}} \leq (1-\zeta)(1-\sigma) + \zeta$ before scoring, without using the sample range. The two-center codebook objective is computed exactly using Eq. (12).

Table J1. Current complete audits: 10 new independent repetitions per setting.

Condition	M	G	n	Numerical passes	Mean L	Mean U	Mean $L-U$
Strong	1	256	4096	10/10	0.421227	0.046297	0.374931
Strong	4	1024	4096	10/10	0.421636	0.316734	0.104903
Strong	8	2048	4096	10/10	0.421931	0.364084	0.057847
Strong	8	2048	16384	10/10	0.440397	0.358375	0.082022
Strong	16	4096	16384	10/10	0.440284	0.394974	0.045310
Strong	32	4096	4096	0/10	0.411506	0.425221	-0.013715
Soft	8	2048	16384	10/10	0.260855	0.257131	0.003724

Condition	M	G	n	Numerical passes	Mean L	Mean U	Mean $L-U$
Independent	8	2048	4096	0/10	0.000000	0.048010	-0.048010

Both the strong and soft conditions recovered the corresponding numbers of groups. The independent condition yielded no valid groups and used the valid zero lower bound. Of the 80 runs, 60 achieved a numerical pass; both the negative gaps for 32 modules and the independent controls are retained. The key lower-bound components are shown below.

Table J2. Mean components of the current lower endpoint.

Condition / scale	$\widehat{D}$	$\widehat{e}$	τ_2	q_n	r_e	Mean $L-U$
strong, M=32, n=4096	0.479041	0.003169	0.035330	0.025749	0.003288	-0.013715
soft, M=8, n=16384	0.320552	0.025352	0.024958	0.008559	0.000827	0.003724

This experiment connects local grouping, codebook and candidate learning, and the current statistical certification procedure. However, the upper bound still uses the direct component baseline and strong statistical endpoints; prefix or stopping-tree expansions were not run on these high-dimensional objects. Thus, this experiment does not establish additional gains from the new couplings on learned high-dimensional objects, nor does it provide a grouping-learning guarantee for arbitrary posteriors.

J.6 Settings and Complete Results of the Perturbation Coupling Comparisons

We begin with the original three-block mechanism example: $h \in \{-1, +1\}$ has equal weights, and the candidate roots are $V_h = \mathrm{Bern}(1/2 + 0.2h)^{\otimes 2} \otimes \mathrm{Bern}(1/2)$. The first two blocks of the source posterior agree with the root, while the third-block probability is $0.2 + \epsilon, 0.5 + \epsilon, 0.8 + \epsilon$, with conditional probabilities $1/4, 1/2, 1/4$. The original batch prespecified $\epsilon \in \{0, 0.03, 0.09\}$ and $n \in \{4096, 65536, 262144\}$, with 20 independent scoring runs per setting.

The tree exactly splits the candidate's third block into equally weighted leaves with probabilities $0.2, 0.8$, while retaining the parent node. For each scoring stream, the direct component cost, full-prefix $\mathcal{H}_0$, and stopping-tree LP are solved separately. The actual combined cost is the minimum of the three; it is not labeled as a solved joint tree-expanded hierarchy. All new endpoints retain the same anchor, CVaR, and small-range SAA endpoints as the baseline.

To preserve the existing upper-endpoint values and the comparison under equal budgets, the five budget categories $\delta_T, \delta_F, \delta_Y, \delta_2, \delta_{\mathrm{small}}$ each remain at $0.05/7$, totaling $0.05 \times 5/7$. The budget cap remains 0.05, the unused allocation is not redistributed, and δ_T is again divided equally. The small range is derived from the frozen, complete finite observation support and computed as the population range of each candidate cost. This coefficient is used only for the original component endpoint and is not transferred to the common W event.

The experimental notation in this subsection, Appendix J.7, and Appendix J is standardized as follows:

$$
\begin{aligned}
C_{\mathrm{prefix}} &= \mathcal{H}_0(\widehat{R}_n, P), \qquad C_{\mathrm{tree}} = \mathcal{T}_{\mathscr{T}}(\widehat{R}_n, P), \\
C_{\mathrm{new}} &= \min\{\widehat{\mathcal{T}}, C_{\mathrm{prefix}}, C_{\mathrm{tree}}\}, \qquad B_{\mathrm{stat}} = \min\{U_{\mathrm{small}}, U_{\mathrm{anchor}}, U_{\mathrm{CVaR}}\}, \\
U_{\mathrm{base}} &= \min\{1, \widehat{\mathcal{T}} + r_T^W, B_{\mathrm{stat}}\}, \\
U_j &= \min\{1, C_j + r_T^W, B_{\mathrm{stat}}\}, \quad j \in \{\mathrm{prefix}, \mathrm{tree}, \mathrm{new}\}.
\end{aligned}
$$

The first two quantities define the optimal values of the corresponding optimization problems; the tables report their float64 solver values, following the numerical convention in Appendix J.4. Each of $U_{\mathrm{prefix}}, U_{\mathrm{tree}}, U_{\mathrm{new}}$ has already been minimized against the strong statistical endpoints; U_{base} is $U_{\mathrm{TV,base}}$ from Appendix G.11 with the small-range branch enabled. For each run, the "cost reduction" is $\widehat{\mathcal{T}} - C_{\mathrm{new}}$ and the "final endpoint reduction" is $U_{\mathrm{base}} - U_{\mathrm{new}}$; the tables then report the means of these per-run quantities.

Table J3. Final upper bounds under the same scoring streams and upper-endpoint budgets.

Perturbation ϵ	n	Mean U_{base}	Mean U_{prefix}	Mean U_{tree}	Mean U_{new}	Mean endpoint reduction
0.0	4096	0.061091	0.028909	0.028800	0.028800	0.032291
0.0	65536	0.052774	0.007130	0.007117	0.007117	0.045657
0.0	262144	0.051429	0.003619	0.003612	0.003612	0.047816
0.03	4096	0.066980	0.038599	0.039375	0.038599	0.028381
0.03	65536	0.057750	0.016923	0.017115	0.016923	0.040827
0.03	262144	0.056434	0.013538	0.013624	0.013538	0.042896
0.09	4096	0.076080	0.057521	0.058249	0.057521	0.018559
0.09	65536	0.067827	0.036986	0.037138	0.036986	0.030841
0.09	262144	0.066354	0.033401	0.033546	0.033401	0.032954

Table J4. Cost tightening and final upper-endpoint reduction reported separately, with $n = 65536$ and means over 20 runs.

Perturbation ϵ	$\widehat{\mathcal{T}}$	C_{prefix}	C_{tree}	Cost reduction	Final endpoint reduction
0.0	0.050369	0.000573	0.000560	0.049810	0.045657
0.03	0.055346	0.010366	0.010558	0.044980	0.040827
0.09	0.065468	0.030429	0.030581	0.035039	0.030841

These 180 repetitions retain all upper-endpoint comparison columns. When $\epsilon = 0$, $R = P$, and both the population prefix cost and tree cost are zero, showing that both couplings eliminate the component-transport slack in this example. This example alone does not establish that the stopping tree outperforms the prefix. As sample size increases, the statistical radii shrink, but cost tightening and final upper-endpoint reduction remain distinct quantities.

The codebook lower endpoints were not recomputed alongside these comparisons. Consequently, the conclusions are limited to improvements in costs and upper endpoints; no additional structural replacement passes are reported on this basis.

J.7 Settings and Complete Results of the Reference-Structure Controls

In the first control, the source is a fixed fair three-block product, and the candidate uses the two roots from Appendix J.6. The second source is an equally weighted mixture of $\mathrm{Bern}(0.01)^{\otimes 3}$ and $\mathrm{Bern}(0.99)^{\otimes 3}$, with a fair product as the candidate. The second reference remains correlated, while the candidate itself belongs to the original product family; strict family-level improvement is therefore impossible. To reuse the two-component statistical endpoints, the same product is represented by two identical, equally weighted roots, without changing the candidate distribution.

Both controls use the common upper endpoint, strong statistical endpoints, and the same budgets as in Appendix J.6. The second tree uses complete coordinatewise binary splitting and allows stopping at parent nodes. For each configuration, 100 independent repetitions are retained at each of $n = 4096, 65536$.

Table J5. Costs and upper endpoints under two reference structures, comprising 400 comparisons in total.

Reference / candidate configuration	n	Mean direct cost	Mean prefix cost	Mean tree cost	Mean U_{base}	Mean U_{new}
Independent reference	4096	0.133333	0.133333	0.133333	0.133333	0.133333
Independent reference	65536	0.133333	0.133333	0.133333	0.133333	0.133333
Correlated reference, product candidate	4096	0.490000	0.322193	0.255230	0.490000	0.281456
Correlated reference, product candidate	65536	0.490000	0.320737	0.253889	0.490000	0.260446

These results compare the effects of reference dependence and candidate representation on the upper bound. The current codebook lower endpoints were not computed in this subsection, so these results do not constitute false-pass frequencies for complete certification. The independent-reference configuration in Appendix J.5 retains the complete codebook audit; the remaining reliability claims follow from the theorem in Appendix G.11.

J.8 Dimension Calibration, Mechanisms, and Scope of the Evidence

The analytical full-support categorical soft modules use $a = 0.2, r = 0.2$ and $K_g \equiv K$, retaining $G = Mb$ and the independent fair module labels from Appendix H.3. Define the single-block probability vectors and the full candidate as

$$v_{\pm} = \left(\frac{1-r}{2} + \frac{r}{K} \pm a, \frac{1-r}{2} + \frac{r}{K} \mp a, \frac{r}{K}, \ldots, \frac{r}{K} \right), \qquad P_{\mathrm{cat}} = \tfrac{1}{2} v_{-}^{\otimes G} + \tfrac{1}{2} v_{+}^{\otimes G}.$$

Every block in the jth module of the source posterior uses v_{Z_j}, and $\mathcal{T}$ in the table denotes direct component transport from this specified posterior decomposition to P_{cat}. Since $\mathrm{TV}(v_+, v_-) = 2a$, the module-counting calculation in Appendix H.3 still gives $\mathcal{T} = a - 2aA_M$; using modules as groups and their corresponding two centers gives $D = a, e = 0, \tau_2 = \rho_b$. Table J6 uses these closed-form quantities to compute the population certificate gap $[a - \rho_b]_+ - \mathcal{T}$, without subtracting finite-sample radii.

Table J6. Block-dimension calibration, displaying both positive and negative gaps.

G	K	D_{flat}	M	b	ρ_b	$\mathcal{T}$	Population certificate gap $\ell_2 - \mathcal{T}$
32	32	1024	1	32	0.071043	0.000000	0.128957
32	32	1024	2	16	0.101123	0.100000	-0.001123
32	32	1024	4	8	0.144546	0.125000	-0.069546
32	32	1024	8	4	0.207360	0.145313	-0.145313
128	4	512	4	32	0.071043	0.125000	0.003957
2048	4	8192	8	256	0.024958	0.145313	0.029729

For the multimodule configurations with $G = 32, K = 32$, the 1024 probability entries do not offset the group-width penalty; increasing the scoring sample size cannot remove this population-level loss. These are analytical constructions, not trained RSSMs.

On the 27 finite-space instances rerun in this round using the original hierarchy script, the maximum monotonicity violation was 2.22×10^{-16}, and the maximum discrepancy between the full level and an independent OT computation was 1.11×10^{-16}. These reruns do not increase the number of independent cases. The complete settings, finite-distribution verification of individual endpoints, and stopping-tree mechanism comparisons are provided in Appendix I. Across the 580 three-block upper-endpoint comparisons, the maximum tree-flow marginal residual was $0.00e + 00$. These checks assess numerical consistency only.

The current evidence separately supports high-dimensional codebook baseline audits of objects that are learned and then frozen, and low-dimensional tightening of coupling costs and upper endpoints. Whether combining the additional couplings with the codebook lower bound increases the number of complete certification passes remains to be tested on the same objects and scoring streams. The tightness of general small trees, applicability to trained RSSMs, and task-level gains also remain unverified. Complete per-run results, frozen objects, scoring sufficient statistics, generation seeds, and code are archived together.

K Supplementary Discussion and Scope of Diagnosis

Blockwise KL and entropy logs do not determine cross-block dependence: block probabilities $1/2+ aS_g$ have identical blockwise KL divergences and entropies under independent fair signs and a single sign shared across all blocks, yet yield independent and correlated aggregate references, respectively. For any nonempty coordinate set $A \subseteq [G]$, the directly scorable local co-occurrence quantity is

$$R_A(x_A) = \mathbb{E}_O \prod_{g \in A} q_{Og}(x_g). \tag{60}$$

Transport plans must preserve the candidate's full distribution; matching only the block marginals is insufficient. Our main implementation directly reads the full block probability vectors of the frozen posterior and retains the original aggregate reference.

The TV upper bound uses component transport as its baseline, retains the unified hierarchy and stopping tree as two computational branches, and combines them through the tree-expanded hierarchy. The stopping tree uses exact node splits, retained parent nodes, and valid stopping flows, computationally optimizing the transport plan and stopping masses. Every construction preserves the original reference and candidate marginals, shares statistical corrections under the default range, and guarantees that their combination is no worse under the same budget. Approximate solvers must retain previously available valid feasible costs. The final combination retains the statistical advantages of the anchor, CVaR, and valid small-range endpoints, without guaranteeing strict improvement at every level. The upward bias of the empirical reference itself in high dimensions, as well as the costs of prefix enumeration and optimization, may still prevent certification.

The two sides of the certificate are controlled by different structures. Under exact coverage, the approximation loss of the local compression lower bound is at most $2e + \tau$; partial coverage additionally incurs the omission term in Eq. (18). The slack of the candidate upper bound depends on the restrictions imposed by the feasible coupling construction. The sufficient condition for detection starts from the current population certificate gap. High dimensionality alone does not guarantee detection: group width, compression distortion, and candidate capacity may still exhaust the advantage. TV couplings directly construct plans in the original space and retain the original model and candidate.

The statistical objects are independent observations under the stated conditional law. Repeated generation with the full simulator state fixed corresponds directly to the target only when that state agrees with the paper's conditioning variable c; adjacent trajectory frames generally cannot be treated as independent repetitions either. The scoring conditional law must agree with the stated target conditional law. If they differ, the guarantees in this paper do not directly apply to the target reference.

Combining multiple endpoints on a common statistical event is necessary to retain post-selection coverage.

Synthetic-posterior experiments examine certifiability and endpoint tightening under specific structures. Analytical constructions and finite-space checks test mechanisms such as transport slack, hierarchy monotonicity, and strict stopping-tree gains. The current experiments do not support inferences about the practical gains of these methods on trained RSSMs.

The certificate compares the candidate's distributional error with that of the entire original family. It does not directly imply policy gains, nor does it show that the current prior has attained the optimum within its family. The latter additionally requires a valid upper bound $W(R, Q_{\mathrm{cur}}) \leq U_{\mathrm{cur}}$ for the current prior $Q_{\mathrm{cur}} \in \mathcal{F}$, together with verification that $U_{\mathrm{cur}} - L$ is sufficiently small.