VIEWCERT: VERIFYING MODEL SELECTION UNDER EVALUATOR AMBIGUITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Model selection is usually reported under one evaluator, even when the same public data admit several defensible choices. ViewCert verifies whether a model-selection claim holds for every program in an explicitly declared, mechanically complete typed language while reusing fixed outputs. Unstructured verification is coNP-complete. Factored margins reduce exactly to classical min-sum elimination, with cost exponential only in interaction width τ. ViewCert returns a replayable worst-case program, handles the full aggregation simplex without a grid, and provides an asymptotically valid simultaneous lower curve after search over programs, competitors, weights, and budgets. As a post-outcome algorithmic closure on Open Images, an audited width-two compiler verifies Grounding DINO Base against five competitors over 2^{50} evaluator programs without enumerating their leaves; its worst-case margin is $+11.953$ AP. Prospectively on WMT25, five official metrics certify Shy over Gemini-2.5-Pro, while two registered LLM judges certify the reverse among 15,120 covered comparisons. Across six regimes, ViewCert distinguishes stable selection, certified ranking dependence, and principled non-identification.

1 INTRODUCTION

A benchmark does more than attach a score to a model: its evaluator defines what counts as success. In structured prediction, that definition includes whether labels are exhaustive, how taxonomies propagate, how group boxes match predictions, which reference geometry is used, which translation reference is selected, which metric operationalizes quality, and how scores are aggregated. Several choices can be defensible on the same public data. They can change a ranking without changing a single model output, yet a conventional leaderboard conditions on only one such choice.

Sampling uncertainty does not resolve this ambiguity. A narrow confidence interval under one evaluator shows that its score is precise; it does not show that another admissible evaluator selects the same model. Conversely, failure to certify a common winner is not evidence of a reversal: the available sample may simply be insufficient. A useful analysis must therefore distinguish three outcomes—selection is stable, a ranking changes, or the data do not identify either conclusion—while accounting for the evaluator and competitor chosen after inspecting the results.

ViewCert turns evaluator ambiguity into one verification question: does a model-selection claim remain valid for every evaluator in an explicitly declared, mechanically complete language? Each evaluator is a typed, prediction-independent program. Every model produces one immutable prediction bank; only the measurement program changes. ViewCert returns a certificate over the whole language or a replayable counterexample, without choosing a preferred evaluator or placing a distribution over evaluators.

This question creates two coupled challenges: search and inference. Unstructured verification is coNP-complete, while factored verification reduces exactly to classical min-sum elimination and is fixed-parameter tractable in interaction width. The optimizer is not the contribution: the new object is the mechanically complete typed language, its replayable proof obligation, and the mixed discrete–continuous certificate built around it. Semantic edit atoms make the resulting budgets invariant to meaning-preserving changes in grammar representation. For continuous aggregation choices, an optimal-transport oracle avoids a weight grid. Since the worst program, competitor, and budget are

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.

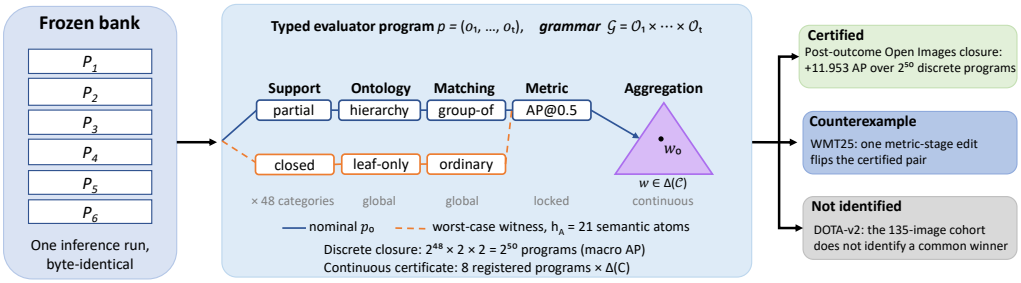


Figure 1: **ViewCert's decision object.** A declared evaluator language acts on one immutable prediction bank and yields a certificate, a replayable counterexample, or *not identified*.

data-selected, we derive one simultaneous post-search confidence event for the resulting breakdown curve.

We evaluate this decision object without creating new annotations. COCO/LVIS varies reference source and category–image support on the same pixels; Open Images varies partial verification, hierarchy propagation, and group-of matching; and DOTA-v2 varies reference geometry on the same aerial predictions. WMT22, a sequentially held-out WMT24 replication, and a separately locked WMT25 transfer vary public references, metrics, normalization, and aggregation over fixed translations. Across these regimes, ViewCert separates large score movement from selection fragility, certifies ranking reversals when both signs are supported, and returns *not identified* when they are not.

Our contributions are:

- **A verifiable decision object.** We define evaluator-program breakdown certificates, a typed language with static safety and mechanical completeness checks, and semantic edit atoms whose budgets are invariant to meaning-preserving grammar refinements.
- **A verification frontier using classical solvers.** We prove coNP-completeness without structure and an exact reduction of factored verification to min-sum elimination, which is exponential only in interaction width and returns an attaining program. An exact simplex reduction and a grid-free transport oracle extend the proof object to continuous aggregation.
- **Post-search inference.** An atomic numerical-delta max-t construction gives simultaneous lower confidence curves after selecting the worst program, competitor, weights, and edit budget. A protocol frozen before 30 million unseen-regime bootstrap draws audits finite-sample coverage and efficiency.
- **Cross-domain evidence without new labels or judgments.** A post-outcome Open Images algorithmic closure verifies Grounding DINO Base over 2^{50} programs with worst-case margin $+11.953$ AP. Prospectively on WMT25, one metric-stage edit yields opposite certified rankings for Shy and Gemini-2.5-Pro among 15,120 covered comparisons.

2 Related Work

Multiverse analyses vary pipelines (Bell et al., 2022), multicriteria benchmarks objectives (Jansen et al., 2024), and Rashomon and predictive-multiplicity studies ask where adequate models agree or disagree (Laberge et al., 2023; Banerjee et al., 2026). Concurrent work certifies a budgeted allocation across evidence-consistent quality tables with an exact two-solve test (Khosravi & Huo, 2026). These methods vary pipelines, models, or quality tables. ViewCert instead verifies typed measurement programs on one frozen output bank and returns the attaining evaluator. Following partial identification (Manski, 2003; Tamer, 2010), it conditions conclusions on a declared evaluator set rather than probabilities or preference weights. Error bars, aggregate targets, and rank intervals (Miller, 2024; Longjohn et al., 2025; Neuhof & Benjamini, 2026) quantify one target; ViewCert asks whether target choice changes the licensed decision.

Robustness and breakdown analysis provide a second connection. Leaderboard perturbations edit comparison records (Huang et al., 2025; Oyarhoseini et al., 2026); SIREN corrects the winner's curse from adaptive tuning on reused items (Xu et al., 2026). ViewCert instead freezes predictions and varies public evaluator semantics and weights. Classical breakdown and directional-inference tools (Masten & Poirier, 2020; Fang & Santos, 2019; Hong & Li, 2018) are compiled into a mechanically complete, noncompensatory certificate that jointly covers the post-search program and competitor. The target is evaluator-robust choice, not tuned-procedure generalization or scalar sensitivity (Appendix Table 5).

The optimization primitive itself is classical. Bucket and variable elimination already solve factored MAP and constraint objectives with time and memory exponential in induced width (Dechter, 1999). We do not claim a new generic graph optimizer. ViewCert's contribution is the auditable object around it: an outcome-independent typed evaluator language, mechanical completeness, representation-stable edit accounting, replayable witnesses, and uncertainty that covers the adversary selected after search.

Formal neural-network verification pursues a related proof-or-counterexample logic. Neural-network verifiers search over inputs using SMT or branch-and-bound (Katz et al., 2017; Wang et al., 2021). ViewCert instead verifies fixed-output model selection over measurement programs; *not identified* is statistical, not a solver timeout.

The closest task-specific precedents concern benchmark semantics in vision and translation. Partial annotation can make proposal evaluation gameable (Chavali et al., 2015), and selecting a reference subset can improve alignment with a downstream use case (Feinglass & Yang, 2024). VisER tests visual support for mentions by large vision-language models (LVLMs) (Hasanebrahimi et al., 2026), while IntroConformal controls factuality risk from internal LVLM signals (Atabuzzaman et al., 2026). These individual-output properties do not imply that registered evaluators identify the same model; ViewCert certifies or refuses that conclusion. Grounding, factuality, identification, and certification are distinct. In translation, BLEU requires an exact implementation contract (Post, 2018), while chrF encodes a different character-level notion of similarity (Popović, 2015). WMT22, WMT24, and WMT25 provide fixed system outputs and multiple public evaluators (Freitag et al., 2022; 2024; Kocmi et al., 2025). Whereas prior WMT work studies agreement with human judgments or stability of human-evaluation protocols (Kocmi et al., 2021; Knowles, 2021), we ask which ordering is identified without treating any registered evaluator as latent truth.

3 Evaluator-program certificates

ViewCert represents evaluator uncertainty through a typed program. Let $\mathcal{M}$ be a frozen candidate roster. Every model m emits one immutable prediction bank P_m. An evaluator is a typed program $p = (o_1, \ldots, o_T)$ whose operators act in a fixed stage order—for example support, ontology, matching, metric, then aggregation. A grammar $\mathcal{G} = \mathcal{O}_1 \times \cdots \times \mathcal{O}_T$ declares the allowed operator at each stage. Static checks reject an operator if it reads model outcomes to choose semantics, changes inference, uses the wrong task type or stage, lacks declared public provenance, or consumes a different prediction-bank hash. Canonical operator hashes identify every program.

Proposition 1 (Mechanical completeness)**.** *A registry contains the complete finite grammar exactly when its canonical program-signature set equals the Cartesian product of the registered operator-signature sets.*

The checker reports missing, extra, invalid, and duplicate cells and prohibits a robustness conclusion after failure. This makes "all admissible views" a machine-checkable claim about a declared language, not an informal list. Completeness is therefore syntactic, not ontological: ViewCert neither asserts that the language contains every defensible evaluator nor identifies a latent correct metric. Expanding the registered language can only weaken a certificate.

The evaluator family combines discrete semantic choices with continuous aggregation. Program p returns component scores s_{pmc} on a common component set $\mathcal{C}$. Aggregation may be any w in the simplex $\Delta(\mathcal{C})$, so

$$S_m(p, w) = \sum_{c \in \mathcal{C}} w_c s_{pmc}, \qquad d_{pmjc} = s_{pmc} - s_{pjc}. \tag{1}$$

The nominal evaluator is (p_0, w_0). We preserve the meaning of changes with the noncompensatory resource vector

$$D(p, w) = \big(h_A(p, p_0),\ \mathrm{TV}(w, w_0),\ \eta(p, p_0)\big), \tag{2}$$

where finite semantic-atom sets are uniquely owned by operator stage, $A(p)$ is their union, $h_A(p, p_0) = |A(p) \triangle A(p_0)|$, and η is a registered public-support shift. Ordinary operator Hamming distance is the special case of one atom per nonnominal categorical choice. We never learn a scalar exchange rate between these units. The atoms are a versioned accounting interface, not an intrinsic semantic distance: their granularity is part of the public registry, dimensions remain incomparable, and Theorem 3 protects only meaning-preserving refactorings of the same atom dictionary.

For candidate m, competitor j, program p, and practical margin $\delta \geq 0$, define its aggregation breakdown radius

$$\beta_{pmj} = \min_{w \in \Delta(\mathcal{C})} \{\mathrm{TV}(w, w_0) : w^\top d_{pmj} \leq \delta\}. \tag{3}$$

It is infinite if reversal is impossible. The exact solver transports mass from the largest margin components to the smallest, stopping at δ. An exchange argument proves optimality, giving $O(|\mathcal{C}| \log |\mathcal{C}|)$ time per program–competitor pair without a weight grid.

A two-program example makes the geometry concrete. With $w_0 = (1/3, 1/3, 1/3)$ and $\delta = 0$, nominal margins $d_{p_0} = (2, 1, -1)$ average to $2/3$. Moving mass $2/9$ from the first to the third component gives $w = (1/9, 1/3, 5/9)$ and zero margin, hence $\beta_{p_0} = 2/9$. One registered edit with $d_{p_1} = (1, -1, -2)$ already loses at uniform weights, so $\beta_{p_1} = 0$; the frontier is $\{(0, 2/9, 0), (1, 0, 0)\}$ with explicit continuous and discrete witnesses.

Each attainable pair yields $q_{pmj} = (h_A(p, p_0), \beta_{pmj}, \eta(p, p_0))$; let $\mathcal{F}_m$ be their Pareto-minimal set.

Theorem 1 (Sharp evaluator-program certificate)**.** *If the typed registry is complete, all programs reuse the same bank and components, and Eq.* (1) *holds, then a counterexample to m exists inside resource budget $b = (H, E, R)$ if and only if some $q \in \mathcal{F}_m$ satisfies $q \leq b$ componentwise.*

Thus every failure has a replayable witness—program hash, competitor, and attaining weights—while absence of a witness is an exact certificate over the declared mixed family. Any sound complete black-box verifier must inspect every unrestricted grammar leaf in the worst case: an unqueried leaf can be changed from all-positive to negative margins without changing its transcript (Appendix A).

Structure makes large implicit families tractable. The black-box bound is avoidable only when a compiler exposes shared sufficient statistics. Write evaluator choices as $x = (x_1, \ldots, x_T)$ and suppose a fixed-aggregation pairwise margin has the auditable decomposition

$$d_{mj}(x) = \sum_{a \in \mathcal{A}} f_a(x_a), \tag{4}$$

where factor a reads only its registered scope x_a. The primal graph connects choices that occur in one factor. Edit cost is a unary factor and is tracked as a separate integer state rather than folded into the margin. The atom registry makes this budget a property of the semantic intervention, not of how many grammar nodes happen to encode it.

Theorem 2 (Computational frontier)**.** *For rational local tables, deciding whether some program has $d_{mj}(x) \leq 0$ is NP-complete, even for binary choices and factors of arity at most three; certifying $d_{mj}(x) > 0$ for every program is coNP-complete. Given an elimination order of induced width τ, maximum option count q, and edit budget H, min-sum variable elimination returns the exact minimum margin and an attaining program in $q^{\tau+1}\,\mathrm{poly}(|\mathcal{A}|, T, H)$ time and $q^{\tau}\,\mathrm{poly}(|\mathcal{A}|, T, H)$ memory. If every component margin $d_c(x)$ has width at most τ, then*

$$\min_{x : h_A(x, x^0) \leq H} \min_{w \in \Delta(\mathcal{C})} w^\top d(x) = \min_{c \in \mathcal{C}} \min_{x : h_A(x, x^0) \leq H} d_c(x), \tag{5}$$

so the implicit programs crossed with the full uncountable simplex are exact with $|\mathcal{C}|$ calls to the same solver.

Theorem 3 (Representation-invariant edit budgets)**.** *Let $R : \mathcal{G} \to \mathcal{G}'$ be a bijection that maps the nominal program x^0 to $R(x^0)$. If R preserves every reconstructed component margin and the global semantic-atom set of every program, then the minimum margin is identical in $\mathcal{G}$ and $\mathcal{G}'$ for every competitor, aggregation vector, and edit budget H. Moreover, R maps attaining programs one-to-one.*

Thus a certificate cannot change merely because one semantic intervention is encoded as one grammar node in one implementation and several nodes in another. The released solver validates every local table, uses deterministic min-fill, emits its induced width and zero optimality gap, and replays the attaining program against the original factors. Equation (5) covers the full simplex; the mixed-TV oracle is separately exact for a complete registered semantic family. We do not claim width-sensitive TV radii. A deterministic scaling control varies variable count T, width τ, options q, and edit budget H using hardware-independent work and state counts (Appendix Table 8). Appendix Fig. 5 visualizes the two exact primitives without overloading the conceptual overview.

Inference must remain valid after searching the breakdown curve. For semantic-edit budget H, let

$$B(H) = \min_{p,j:\, h_A(p,p_0)\le H} \beta_{pmj}. \tag{6}$$

A direct bootstrap of this outer minimum can be anti-conservative near an unknown almost-tie; covering every raw component margin instead can be vacuous. For inference only, censor each atom at the maximal feasible TV: $\bar{\beta}_k = \min\{\beta_k, 1 - \min_{r\in\mathcal{C}} w_{0r}\} \le \beta_k$; impossible reversals map to this finite cap. ViewCert covers these maps first. Let $\widehat{d}$ be the primitive margins, G^* a centered paired-bootstrap process draw, and $\epsilon_n \to 0$ with $\sqrt{n}\epsilon_n \to \infty$. For atom $k = (p, m, j)$, compute

$$Z_k^* = \frac{\bar{\beta}_k(\widehat{d}_k + \epsilon_n G_k^*) - \bar{\beta}_k(\widehat{d}_k)}{\epsilon_n}. \tag{7}$$

If $q_{1-\alpha}$ is the bootstrap quantile of $\max_k Z_k^*/\widehat{\sigma}_k$, set $\ell_k = \max\{0, \widehat{\bar{\beta}}_k - q_{1-\alpha}\widehat{\sigma}_k/\sqrt{n}\}$ and $L(H) = \min_{k:h_k\le H} \ell_k$.

Theorem 4 (Atomic post-search validity)**.** *For a fixed finite atomic family, suppose the joint primitive estimator and centered bootstrap process are consistent, the censored atomic breakdown maps are Hadamard directionally differentiable, the numerical-delta rate conditions hold, and nondegenerate scales are consistently estimated. Then*

$$\liminf_n \Pr\{L(H) \le B(H) \text{ for every registered } H\} \ge 1 - \alpha.$$

The program, competitor, weights, and edit budget may be selected after inspection of the covered family.

Appendix Table 6 separates the system's deterministic exactness, asymptotic inference, and empirical AP calibration. The numerical delta method handles nonsmooth optimal-transport ties (Fang & Santos, 2019; Hong & Li, 2018); the new construction avoids bootstrapping the outer minimum by transforming and jointly covering its atomic certificates. It costs $O(B|\mathcal{G}|(|\mathcal{M}|-1)|\mathcal{C}|\log|\mathcal{C}|)$ for B process draws.

Finite registries arise as a special case. With fixed w_0, each program is an ordinary view v with score S_{mv}. Writing $W(v) = \arg\max_m S_{mv}$, the uniform-winner set is $U = \cap_v W(v)$ and worst-view regret is $\rho(m) = \max_v r(m, v)$ for $r(m, v) = \max_j S_{jv} - S_{mv}$.

Proposition 2 (Finite-family identification)**.** $U = \{m : \rho(m) = 0\}$. *Adding views can only shrink U and increase ρ.*

Proposition 3 (Complete-view necessity for positive certification)**.** *From a proper subset of otherwise unrestricted registered views, no model can be soundly certified as a member of the full U for every completion.*

For candidate rosters $\mathcal{M} \subseteq \mathcal{M}'$ with old scores fixed,

$$\rho_{\mathcal{M}'}(m) \ge \rho_{\mathcal{M}}(m), \qquad U(\mathcal{M}', \mathcal{V}) \cap \mathcal{M} \subseteq U(\mathcal{M}, \mathcal{V}). \tag{8}$$

For finite experiments we retain one joint interval event

$$\Pr\{d_{mjv} \in [L_{mjv}, U_{mjv}]\ \forall v, m \ne j\} \ge 1 - \alpha. \tag{9}$$

It simultaneously licenses common winners, winner-confidence sets, regret, and data-discovered pairwise reversals. Different certified winners imply global view dependence; opposite signed covered margins for one pair imply pairwise ranking dependence; otherwise ViewCert returns *not identified*. The view-wise confidence set is $C(v) = \{m : U_{mjv} \ge 0\ \forall j \ne m\}$ and the uniform set is $C_U = \cap_v C(v)$. Factorial program cubes additionally admit exact Walsh attribution (Appendix A).

Theorem 5 (Joint finite-family validity)**.** *On the event in Eq.* (9)*, every certified winner and opposite-sign pairwise witness is true, every true winner lies in its reported winner-confidence set, and all regret bounds hold simultaneously. Witnesses may be selected after inspecting the covered family.*

4 EXPERIMENTAL PROTOCOL

All studies reuse public human-created annotations, create no labels, and require every evaluator to read the same frozen model outputs. Each design records an explicit outcome boundary: images, categories, prompts, views, candidate sets, resampling units, multiplicity families, and decision rules are frozen before the corresponding outcomes are inspected. Amendments are allowed only before outcome inspection and retain a signed reason, hashes, and the invalidated artifact. The vision roster comprises Grounding DINO Tiny/Base (Liu et al., 2024), OWLv2 Base Ensemble (Minderer et al., 2023), YOLO-World V2.1-L (Cheng et al., 2024), Detic Swin-B (Zhou et al., 2022), and OmDet-Turbo Swin-Tiny (Zhao et al., 2024). Each runs once per image (or DOTA tile); every view reads the same serialized boxes, labels, scores, and source order.

The COCO/LVIS study isolates reference and support choices. Its held-out cohort has 4,509 COCO validation images (Lin et al., 2014) also present in LVIS (Gupta et al., 2019); 300 pilot images are excluded. An exact 63-category public-name crosswalk defines four views: exhaustive COCO; COCO boxes under the LVIS category–image mask; COCO boxes on the strict common mask; and LVIS boxes on that same mask. The last pair isolates reference source at fixed evaluable support. Ten thousand paired image resamples cover 60 pair-by-view margins.

Open Images supplies an independent factorial design. Annotation-only rules select 9,000 validation images and 48 categories from Open Images (Kuznetsova et al., 2020). The complete $2 \times 2 \times 2$ cube varies official partial verification versus closed-world coverage, hierarchy expansion versus leaf-only references, and official group-of matching versus ordinary IoU. The all-official cell is distinguished from seven stress cells; none is latent truth. The score is macro VOC-2010 AP. Ten thousand paired resamples preserve six annotation-derived strata; one 162-contrast family contains 120 pairwise margins and 42 Walsh effects.

For ViewCert, the same cube is compiled as a typed grammar with binary support, ontology, and matching stages and fixed AP and aggregation stages. The checker requires all eight programs. Its continuous layer uses the 48 per-category AP components and the full probability simplex, with TV measured from uniform w_0; it reruns neither models nor a weight grid. The post-search family has 8 programs $\times$ 5 competitors = 40 atomic radii. Before applying it, we hashed $\epsilon_n = n^{-0.35}$, 2,000 Monte Carlo repetitions, 999 process draws, sample sizes 250/1,000/9,000, five unseen exact/local-tie and nonunique-transport regimes, and explicit coverage thresholds.

The DOTA-v2 study isolates reference geometry. Its domain boundary uses 135 validation images and the 11 DOTA-v2 categories with at least 20 non-difficult objects (Xia et al., 2018). Task 1 matches the frozen horizontal prediction to the oriented quadrilateral; Task 2 matches the identical box to its deterministic horizontal envelope. A prediction-blind tiling amendment creates 1,825 overlapping tiles and one merged bank per model. Ten thousand paired image resamples cover 30 margins and six geometry effects. We interpret these intervals as fixed-cohort stability.

The translation transfer comprises three prospectively locked studies. We use the public `mt-metrics-eval` v2 corpus and never read its archived metric scores (Freitag et al., 2022; 2024). The primary WMT22 en–de cohort has 2,037 segments in 181 documents and 15 complete non-reference systems. A sequential held-out rule then selects WMT24 en–de, the most recent post-WMT22 set with two references, at least ten systems, and 30 documents; it has 998 segments, 171 documents, and 26 eligible systems. Each 2×2 cube crosses public references refA/refB with SacreBLEU 2.5.1 BLEU-13a and chrF (Post, 2018; Popović, 2015). Ten thousand paired document resamples and studentized max-t envelopes cover all 420 and 1,300 pair-by-view margins. The WMT24 design and deterministic selection rule were hashed before its text outcomes were read. No translation or judgment is created by this project.

Independently, a score-blind WMT25 lock precedes access to 59 official automatic-score files (Kocmi et al., 2025). A metadata-only gate retains 21 systems complete across 28 language pairs and 2,819 documents. The full grammar has 36 programs: nine registered subsets of the five official metrics times robust or fractional-rank normalization times paragraph-micro or document-macro aggregation. Each program is crossed with the full language-pair simplex, with TV measured from uniform. Ten thousand paired document resamples cover all $36 \times 21 \times 20 = 15{,}120$ ordered program–candidate comparisons. Outputs, metric values, programs, weights, candidate pairs, and stopping rules cannot be changed after the preregistration hash.

Table 1: Frozen detection outcomes. Targets denotes the joint margin/effect family size; Range is within-model AP movement. A negative $\delta^\star$ means no model is certified at $\delta = 0$.

Regime	Views	Targets	Range (AP)	$\delta^\star$	Certificate
COCO/LVIS	4	60	11.14–15.22	2.24	GD-Base common
Open Images	8	162	17.75–27.53	11.84	GD-Base common
DOTA-v2	2	36	0.57–2.42	−2.44	Not identified

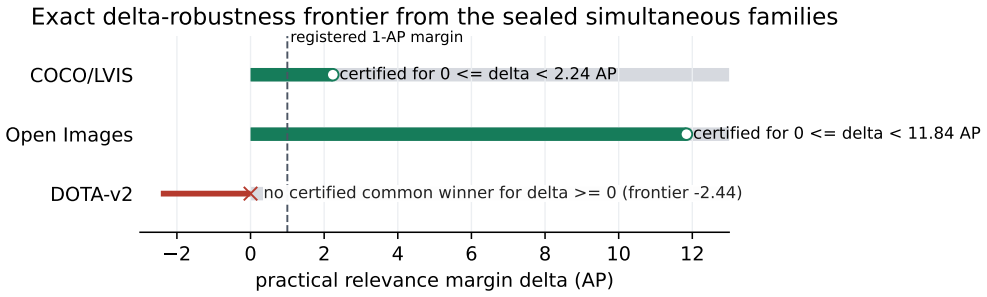


Figure 2: Continuous practical-margin frontiers from the sealed simultaneous families. Open endpoints follow the strict certification inequality.

Every runner passes a common conformance and audit layer. The Open Images evaluator matches 2,000 randomized TensorFlow-reference cases with zero discrepancy; DOTA Task 2 matches 500 public-devkit cases, with separate Task-1 and merge fixtures. Execution is blocked unless cohort, dimensions, coordinates, finite scores, deterministic order, caps, hashes, and artifact identity pass. WMT runners validate line counts, missing-output markers, contiguous document clusters, hashes, and exact SacreBLEU sufficient-statistic reconstruction. Outcome-free Bernoulli and Gaussian simulations first audited the decision family; a later frozen AP-native study directly exercises interpolated AP. Appendix B gives the full construction, amendment, compute, and sensitivity record.

5 Results

The six regimes expose three outcomes that a scalar leaderboard conflates: stable selection despite large score movement, a certified ranking reversal, and non-identification. Table 1 summarizes the detection studies. WMT22 and WMT25 provide the two certified reversals; WMT24 does not support its nominal flip after simultaneous inference.

COCO/LVIS exhibits large score movement without decision fragility. Changing only the view moves every checkpoint by 11.14–15.22 AP. Grounding DINO Base nevertheless leads all four cells (56.66, 65.23, 66.38, and 67.80 AP). Its smallest point margin is 3.640 AP; the 95% simultaneous radius over 60 margins is 1.404 AP, giving singleton winner-confidence sets and a common- winner certificate. The exact frontier is 2.235 AP (Fig. 2). The conclusion survives maxDets $= 300$ and all 12 leave-one-supercategory-out controls. Thus large score shifts do not imply unstable selection.

Open Images independently reproduces that separation in a factorial design. Within-model ranges across eight cells are 17.75–27.53 AP (Fig. 3); Grounding DINO Base wins every cell, from 60.40 AP under all-official semantics to 34.07 AP under all three stress levels. Its 13.814-AP smallest margin exceeds the 1.973-AP simultaneous radius, certifying it through $\delta^\star = 11.841$ AP. All 18 main effects exclude zero; support shifts are largest (-8.43 to -19.30 AP).

The evaluator-program analysis then expands the Open Images cube to a much larger family. Macro AP factorization yields 48 category-specific support choices plus two global choices, or 2^{50} programs. At width two, exact elimination uses 390 local assignments and certifies Grounding DINO Base against five competitors; the worst margin is $+11.953$ AP. Built from frozen scores after inspection, this is an algorithmic closure, not a prospective replication.

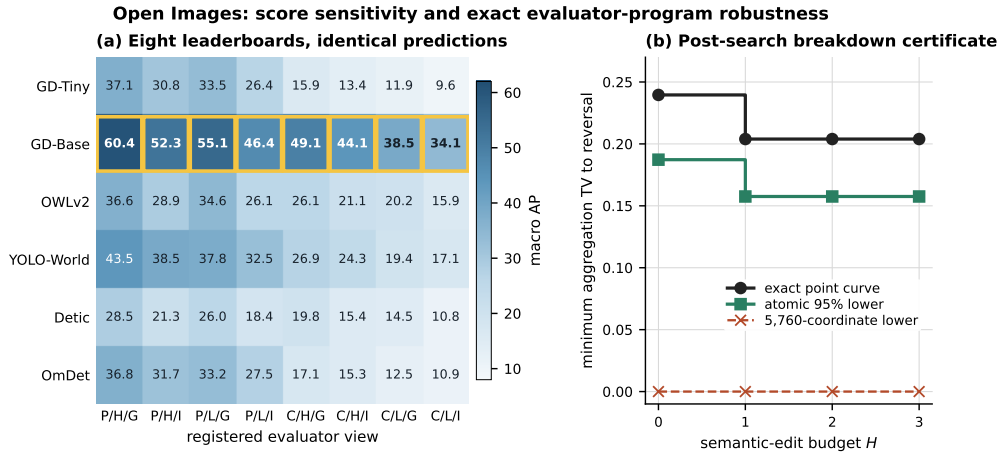


Figure 3: Open Images evaluator sensitivity and breakdown certification. **Left:** macro AP under partial/closed-world (P/C) support, hierarchy/leaf-only (H/L) ontology, and group-of/ordinary-IoU (G/I) matching. **Right:** exact and simultaneous 95% lower curves; raw-coordinate max-t is vacuous.

The separately registered eight-program grammar crossed with the full simplex gives point curve $B(H) = (0.240, 0.204, 0.204, 0.204)$ and simultaneous 95% lower curve $L(H) = (0.187, 0.158, 0.158, 0.158)$. Its witnesses combine reweighting with an optional hierarchy edit. Covering 5,760 raw coordinates instead gives zero throughout; the locked numerical-delta validation attains at least 97.7% coverage and 66.1% of that comparator's mean width. A direct outer-minimum bootstrap undercovers near active-program ties and is barred from claims.

The conclusion survives the registered bootstrap, cap, floor, aggregation, roster, category-deletion, and tie controls (Appendix C). Removing both Grounding DINO checkpoints eliminates the point-common winner, so candidate scope remains explicit. Appendix Table 7 contrasts exhaustive-search, grid, raw-coordinate, and direct-bootstrap baselines.

DOTA-v2 shows a mechanism signal without identification. Horizontal envelopes raise every model by 0.57–2.42 AP. OWLv2 leads both tasks (13.92/16.04 AP), but Grounding DINO Base remains in both confidence sets; with 135 images the familywise radius is 6.354 AP. ViewCert therefore returns *not identified*. Registered sensitivity checks retain the OWLv2 point pattern but do not change that inferential status. The 1,825 overlapping tiles are nested inside 135 independent image clusters, so treating them as independent would be pseudoreplication. More independent images could shrink the radius; this cohort cannot separate structure from finite-sample power.

WMT25 provides a prospective evaluator-program reversal. The nominal five-official-metric program certifies Shy over every system. Its margin over Gemini-2.5-Pro is 0.093 $[0.071, 0.114]$; changing only the metric stage to two registered LLM judges reverses the pair, $[-0.098, -0.030]$ (Fig. 4). Both signs are simultaneous over all 36 programs and 15,120 ordered comparisons. The unchanged-semantics point TV radius is 0.581 but its lower confidence bound is zero. The discrete grammar certifies the reversal; fixed-program simplex analysis supplies the TV radius. We claim the former, not a positive confidence bound for the latter.

WMT22 and WMT24 separate certified ranking dependence from disciplined refusal. WMT22 has three nominal winners and no global-selection certificate, yet its 420-margin envelope certifies `bleu_bestmbr−chrf_bestmbr`: 1.198 $[0.446, 1.950]$ under BLEU/refA but -0.728 $[-1.171, -0.284]$ under chrF/refA, with the same negative sign under chrF/refB. Held-out WMT24 changes nominal winner by reference, but neither global selection nor a pair reversal survives 1,300-margin coverage; ViewCert returns *not identified* (Appendix E).

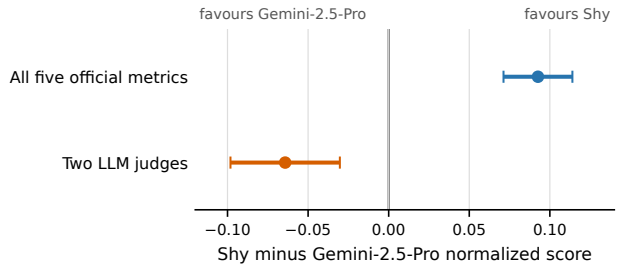


Figure 4: WMT25 simultaneous 95% intervals for Shy minus Gemini-2.5-Pro. One metric-stage edit reverses the pair with all other stages fixed; the envelope covers all 15,120 searched comparisons.

6 Discussion and limitations

ViewCert is a selective proof obligation: it covers the data-selected adversary, certifies only when every simultaneous margin clears the threshold, diagnoses dependence from opposite signs, and otherwise refuses. Grounding, identification, and certification differ. Certificates are conditional on the declared language and roster; completeness is syntactic and public semantic atoms are accounting objects, not a true metric. Exactness requires validated factorization and linear aggregation; width-sensitive TV radii remain open. Inference is asymptotic, while the AP-native audit is empirical.

Controlled comparisons clarify the distinct roles of structured search, continuous aggregation, and post-search inference (Appendix Table 7). Wherever flat enumeration is feasible, structured verification returns exactly the same margin, decision, edit count, and witness across 60 seeded problem–budget cases and 1,680 WMT25 pair–budget cases, while extending to the 2^{50} Open Images family without materializing its leaves. A finite aggregation grid agrees on the simplex vertices but cannot certify the uncountable simplex. At the inferential layer, covering 5,760 raw coordinates drives the Open Images lower curve to zero, whereas atomic coverage yields $(0.187, 0.158, 0.158, 0.158)$; directly bootstrapping the outer minimum undercovers in frozen near-tie regimes. Finally, an ordinary eight-cell multiverse table reports observed variation but certifies neither the implicit family nor the data-selected worst-case witness. The gain is therefore a complete and replayable proof obligation, not a favorable scalar score.

The principal remaining judgment is the boundary of the evaluator language. ViewCert does not remove that choice: for confirmatory claims, the boundary must be fixed before outcomes, versioned, and publicly auditable; post-outcome expansions, such as the Open Images compiler closure, remain algorithmic evidence only. Because expanding the language can only weaken certification, a newly admitted operator may preserve the broader certificate, expose a replayable counterexample, or make the conclusion *not identified*. Reporting official, preregistered, and post-outcome stress-test scopes separately therefore makes this dependence visible. The framework exposes, rather than resolves, the normative choice of evaluator.

7 Conclusion

A leaderboard is a conditional decision, not a property of a model. ViewCert makes that condition explicit and asks whether the decision survives every program in a mechanically complete evaluator language. It returns only what the declared semantics and available data license: a stable winner, a certified reversal, or *not identified*.

Reproducibility statement. The supplement provides proofs, preregistrations, conformance tests, hashes, compute accounting, and commands. The anonymous archive contains the ViewCert API, runners, schemas, and tests; public assets are identified by version and digest, and no labels or judgments are created.

Ethics statement. We use public benchmarks and checkpoints and collect no participant data. Because evaluator choice affects attention, we report unresolved cases, complete view families, and outcome-independent locks.

AI use statement. Generative-AI tools were used in a supporting role to refine author-developed hypotheses and experimental design; to assist with method implementation. They did not supply benchmark labels, annotations, translations, evaluator judgments, or reported empirical measurements. All final scientific decisions, interpretations, theorem statements, proofs, experimental choices, and reported claims were independently determined and verified by the author. All reported numbers were produced by executing the released code on the declared public artifacts. The author reviewed all AI-assisted material and takes responsibility for the complete submission.

References

Md. Atabuzzaman, Christian Alexander, and Chris Thomas. IntroConformal: Conformal factuality guarantees for large vision–language models via introspective signals. *arXiv preprint arXiv:2609.01375*, 2026. URL https://arxiv.org/abs/2609.01375.

Sinjini Banerjee, Tim Marrinan, and Anand D. Sarwate. Measuring consistency via ensemble margin and local prediction variability: Auditing decision systems in the presence of predictive multiplicity. *arXiv preprint arXiv:2609.01397*, 2026. URL https://arxiv.org/abs/2609.01397.

Samuel J. Bell, Onno P. Kampman, Jesse Dodge, and Neil D. Lawrence. Modeling the machine learning multiverse. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/750337e1301941f81ae31a90e0a1c181-Abstract-Conference.html.

Neelima Chavali, Harsh Agrawal, Aroma Mahendru, and Dhruv Batra. Object-proposal evaluation protocol is 'gameable'. *arXiv preprint arXiv:1505.05836*, 2015. URL https://arxiv.org/abs/1505.05836.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. YOLO-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16901–16911, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Cheng_YOLO-World_Real-Time_Open-Vocabulary_Object_Detection_CVPR_2024_paper.html.

Rina Dechter. Bucket elimination: A unifying framework for reasoning. *Artificial Intelligence*, 113 (1–2):41–85, 1999. doi: 10.1016/S0004-3702(99)00059-4.

Zheng Fang and Andres Santos. Inference on directionally differentiable functions. *The Review of Economic Studies*, 86(1):377–412, 2019. doi: 10.1093/restud/rdy049.

Joshua Feinglass and Yezhou Yang. Towards addressing the misalignment of object proposal evaluation for vision–language tasks via semantic grounding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4397–4407, 2024. URL https://openaccess.thecvf.com/content/WACV2024/html/Feinglass_Towards_Addressing_the_Misalignment_of_Object_Proposal_Evaluation_for_Vision-Language_WACV_2024_paper.html.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation*, pp. 46–68. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.wmt-1.2. URL https://aclanthology.org/2022.wmt-1.2/.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pp. 47–81. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.wmt-1.2. URL https://aclanthology.org/2024.wmt-1.2/.

Agrim Gupta, Piotr Dollár, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Afsaneh Hasanebrahimi, Hanxun Huang, Christopher Leckie, and Sarah Erfani. VisER: Visual evidence and reliance for object hallucination detection in LVLMs. *arXiv preprint arXiv:2608.30480*, 2026. URL https://arxiv.org/abs/2608.30480.

Han Hong and Jessie Li. The numerical delta method. *Journal of Econometrics*, 206(2):379–394, 2018. doi: 10.1016/j.jeconom.2018.06.007.

Jenny Y. Huang, Yunyi Shen, Dennis Wei, and Tamara Broderick. Dropping just a handful of preferences can change top large language model rankings. *arXiv preprint arXiv:2508.11847*, 2025. URL https://arxiv.org/abs/2508.11847.

Christoph Jansen, Georg Schollmeyer, Julian Rodemann, Hannah Blocher, and Thomas Augustin. Statistical multicriteria benchmarking via the GSD-front. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b1f140eeee243db24e9e006481b91cf1-Abstract-Conference.html.

Guy Katz, Clark Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In *Proceedings of the 29th International Conference on Computer Aided Verification*, volume 10426 of *Lecture Notes in Computer Science*, pp. 97–117, 2017. doi: 10.1007/978-3-319-63387-9_5.

Hamed Khosravi and Xiaoming Huo. Which LLM for which work? budgeted model allocation under uncertain evaluation. *arXiv preprint arXiv:2608.29560*, 2026. URL https://arxiv.org/abs/2608.29560.

Rebecca Knowles. On the stability of system rankings at WMT. In *Proceedings of the Sixth Conference on Machine Translation*, pp. 464–477. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.wmt-1.56/.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pp. 478–494. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.wmt-1.57/.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinthór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pp. 355–413. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.wmt-1.22. URL https://aclanthology.org/2025.wmt-1.22/.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128:1956–1981, 2020.

Gabriel Laberge, Yann Pequignot, Alexandre Mathieu, Foutse Khomh, and Mario Marchand. Partial order in chaos: Consensus on feature attributions in the rashomon set. *Journal of Machine Learning Research*, 24(364):1–50, 2023. URL https://www.jmlr.org/papers/v24/23-0149.html.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/6319_ECCV_2024_paper.php.

Rachel Longjohn, Giri Gopalan, and Emily Casleton. Statistical uncertainty quantification for aggregate performance metrics in machine learning benchmarks. *arXiv preprint arXiv:2501.04234*, 2025. URL https://arxiv.org/abs/2501.04234.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003. doi: 10.1007/b97478.

Matthew A. Masten and Alexandre Poirier. Inference on breakdown frontiers. *Quantitative Economics*, 11(1):41–111, 2020. doi: 10.3982/QE1288.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024. URL https://arxiv.org/abs/2411.00640.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e6d58fc68c0f3c36ae6e0e64478a69c0-Abstract-Conference.html.

Bitya Neuhof and Yuval Benjamini. Quantifying ranking uncertainty in LLM benchmarks. *arXiv preprint arXiv:2607.16259*, 2026. URL https://arxiv.org/abs/2607.16259.

Hosna Oyarhoseini, Jimmy Lin, and Amir-Hossein Karimi. A unified perturbation framework for analyzing leaderboard stability and manipulation. *arXiv preprint arXiv:2605.15761*, 2026. URL https://arxiv.org/abs/2605.15761.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395. Association for Computational Linguistics, 2015. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049/.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191. Association for Computational Linguistics, 2018. URL https://aclanthology.org/W18-6319/.

Elie Tamer. Partial identification in econometrics. *Annual Review of Economics*, 2:167–195, 2010. doi: 10.1146/annurev.economics.050708.143401.

Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J. Zico Kolter. Beta-CROWN: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://papers.nips.cc/paper/2021/hash/fac7fead96dafceaf80c1daffeae82a4-Abstract.html.

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3974–3983, 2018.

Yang Xu, Jiefu Zhang, Haixiang Sun, Zihan Zhou, Tianyu Cao, and Vaneet Aggarwal. Towards reliable LLM evaluation: Correcting the winner's curse in adaptive benchmarking. *arXiv preprint arXiv:2605.05973*, 2026. URL https://arxiv.org/abs/2605.05973.

Tiancheng Zhao, Peng Liu, Xuan He, Lu Zhang, and Kyusong Lee. Real-time transformer-based open-vocabulary detection with efficient fusion head. *arXiv preprint arXiv:2403.06892*, 2024. URL https://arxiv.org/abs/2403.06892.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, 2022. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/2557_ECCV_2022_paper.php.

A Proofs and edge cases for ViewCert

A.1 Proof of Proposition 1

Each stage option has a canonical signature containing its identifier, type, stage, domain, public provenance, and version. Concatenating one signature per stage is injective over grammar cells. Hence the Cartesian product of option signatures is exactly the set of programs generated by the declared grammar. Set equality with the registry is therefore sufficient. It is also necessary: a missing product signature is an unevaluated grammar cell, an extra signature is outside the language, and repeated signatures violate uniqueness. The implemented checker additionally runs the static type and bank-identity checks before accepting either set. □

A.2 Exact total-variation transport

Moving mass a from component r to component s costs a in total variation and changes the weighted margin by $a(d_s - d_r)$. Suppose a feasible transport uses a donor whose coefficient is below an available donor, or a receiver whose coefficient is above an available receiver. Exchanging that flow for the more extreme donor or receiver leaves cost unchanged and weakly decreases the margin. Repeating yields the greedy order: largest available donor to smallest available receiver. The last transfer may be partial and hits the constraint boundary; any smaller transported mass leaves a positive margin, proving minimality. □

A.3 Proof of Theorem 1

If $q_{pmj} \leq b$, the transport construction supplies its program, competitor, and attaining w, hence an executable counterexample inside the budget. Conversely, take any counterexample (p, j, w) inside b. Definition (3) gives $\beta_{pmj} \leq \mathrm{TV}(w, w_0)$; its semantic and support coordinates are those of p. Thus q_{pmj} is componentwise no larger than the counterexample. It is Pareto-minimal or is dominated by a finite chain ending at a Pareto-minimal point, which also lies inside b. □

The sharp-certificate argument also yields a black-box leaf lower bound. Suppose a sound complete verifier leaves grammar leaf $p^\star$ unevaluated. An all-positive margin surface and a surface identical on every queried leaf but negative only at $p^\star$ give the same transcript and require opposite answers. Therefore every unrestricted leaf must be queried in the worst case; acceleration requires declared shared structure or certified bounds.

A.4 Proof of Theorem 2

For hardness, reduce 3-SAT. Create one binary evaluator choice for every Boolean variable. For each clause a, create a factor on its at most three variables with value zero when the local assignment satisfies the clause and one otherwise. The sum in Eq. (4) has minimum zero if and only if all clauses are simultaneously satisfiable. A complete assignment is a polynomial-size witness, proving NP-completeness of counterexample existence and coNP-completeness of its robust-verification complement.

For the constructive result, each option's unary cost is $|A_t(o) \triangle A_t(o_t^0)|$; place these costs and every local table in the first elimination bucket containing their variables. Eliminating x_t joins the tables in its bucket, takes the minimum over every option of x_t, and retains one state for each assignment to the residual separator and each edit count $0, \dots, H$. Induction over the elimination order shows that each retained state is the exact minimum of all eliminated assignments consistent with its separator and edit count. The final empty-separator table therefore contains the exact global minimum for every feasible edit count; backpointers recover an attaining program. A bucket contains at most $\tau + 1$ variables, hence at most $q^{\tau+1}$ joint assignments. The edit convolutions and number of buckets contribute only polynomial factors in $|\mathcal{A}|$, T, and H; residual tables contain at most $q^{\tau}(H+1)$ states. Finally, a linear function of w over $\Delta(\mathcal{C})$ attains its minimum at a vertex, so $\min_w w^\top d(x) = \min_c d_c(x)$. The two finite minima commute, proving Eq. (5); solving each component gives the stated factor $|\mathcal{C}|$.

□

A.5 Proof of Theorem 3

Margin preservation gives $d_{jc}(R(x)) = d_{jc}(x)$ for every competitor j and component c. Atom-set preservation and the nominal mapping give $h_A(R(x), R(x^0)) = h_A(x, x^0)$. Hence R bijects the feasible programs at every budget H without changing any objective. The two budgeted minima are therefore equal, and their minimizers map in both directions. □

The structured solver passed both empirical and synthetic controls. The Open Images compiler permits an independent partial- versus closed-world support choice for each of 48 categories, crossed with global hierarchy/leaf and group-of/ordinary matching choices. Its 48 factors reconstruct all 40 original candidate–competitor margins over the eight uniform programs to maximum absolute error 1.78×10^{-14} AP. For each competitor, exact elimination verifies all 2^{50} programs at induced width two using 390 local joint assignments and 535 dynamic states, with no materialized leaf. Grounding DINO Base remains the common winner; the global worst case is $+11.953$ AP against YOLO-World and uses 19 closed-world category choices plus both global stress choices (21 registered atoms). This is a post-outcome compiler closure of frozen results, not a prospective replication.

Across 60 seeded seven-choice instances (12 margin surfaces and edit budgets zero through four), variable elimination exactly matches leaf enumeration in minimum margin, decision, edit count, and attaining program. On a width-one chain with 100 binary choices, it covers 2^{100} virtual programs, including 203,366,882,996 within edit budget eight, using 398 local joint assignments and 1,735 dynamic states with no materialized leaf. Three such component graphs crossed with their full simplex require 1,194 local assignments and no program or weight grid. These are synthetic algorithmic controls, not evidence that an official evaluator has this factorization.

A.6 Proof of Theorem 4

For the finite vector of censored atomic maps, consistency of the centered primitive bootstrap and the numerical directional delta method imply joint conditional weak convergence of $(Z_k^*)_k$ to the limiting law of $\sqrt{n}(\hat{\bar{\beta}}_k - \bar{\beta}_k)_k$ (Fang & Santos, 2019; Hong & Li, 2018). Consistent studentization and a regular $(1-\alpha)$ max quantile therefore give

$$\liminf_n \Pr\{\ell_k \le \bar{\beta}_k \ \forall k\} \ge 1 - \alpha.$$

Because $\bar{\beta}_k \le \beta_k$, on this event every registered budget H satisfies

$$L(H) = \min_{k:h_k \le H} \ell_k \le \min_{k:h_k \le H} \beta_k = B(H).$$

All budgets and all data-selected witnesses use the same atomic event, so no additional selection correction is required. □

A.7 Proof of Proposition 2

For a fixed (m, v), $r(m, v) = 0$ if and only if $S_{mv} = \max_j S_{jv}$, which is equivalent to $m \in W(v)$. Therefore

$$\rho(m) = 0 \iff r(m, v) = 0 \ \forall v \iff m \in \bigcap_v W(v) = U(\mathcal{V}).$$

Regret is nonnegative, so if at least one model has zero worst-view regret, zero is the global minimum and its minimizers are exactly $U(\mathcal{V})$. □

A.8 Monotonicity under view expansion

For $\mathcal{V} \subseteq \mathcal{V}'$, the uniform-winner set under $\mathcal{V}'$ intersects every winner set used under $\mathcal{V}$ and possibly more; hence $U(\mathcal{V}') \subseteq U(\mathcal{V})$. Likewise, a maximum over the larger index set cannot be smaller:

$$\rho_{\mathcal{V}'}(m) = \max_{v \in \mathcal{V}'} r(m, v) \geq \max_{v \in \mathcal{V}} r(m, v) = \rho_{\mathcal{V}}(m).$$

A.9 Proof of Proposition 3

Fix any observed score surface on $\mathcal{V}_0$, any proposed model m, and an unobserved view $v^\star \in \mathcal{V} \setminus \mathcal{V}_0$. Choose any competitor $j \neq m$ and complete the surface at $v^\star$ so that $S_{jv^\star} > S_{mv^\star}$; all other missing scores may be arbitrary. This completion agrees with every observation available to the rule, but $m \notin W(v^\star)$ and therefore $m \notin \cap_{v \in \mathcal{V}} W(v)$. Since the construction applies to every observed table and proposed model, a rule sound over all unrestricted completions cannot certify any model as a member of the full uniform-winner set from the proper subset. This does not preclude certifying $U = \varnothing$ when the observed winner sets already have empty intersection, since adding views cannot enlarge that intersection. □

A.10 Proof of Theorem 5

Let E denote the event in Eq. (9). On E, if $L_{mjv} > \delta$, then $d_{mjv} > \delta$. A certified model satisfies this inequality against every competitor, proving (i). If m is a true winner in view v, then $d_{mjv} \geq 0$ for every competitor; interval coverage gives $U_{mjv} \geq d_{mjv} \geq 0$, so $m \in C(v)$, proving (ii). Intersecting this containment over all views proves (iii). For (iv), each certified model is the unique true winner in its corresponding view for $\delta \geq 0$; two different certified model names therefore witness different view selections. Finally, if $L_{mjv} > \delta$ and $U_{mjw} < -\delta$, coverage implies $d_{mjv} > \delta$ and $d_{mjw} < -\delta$ for that pair. Because E covers all pairs and views simultaneously, choosing the witness after inspecting the family does not change the event. Since all five deductions occur on E, their joint probability is at least $\Pr(E) \geq 1 - \alpha$. □

A.11 Monotonicity under candidate-roster expansion

Let $\mathcal{M} \subseteq \mathcal{M}'$ and fix the original score surface. For every original model m and view v,

$$\max_{j \in \mathcal{M}'} S_{jv} - S_{mv} \geq \max_{j \in \mathcal{M}} S_{jv} - S_{mv}.$$

Maximizing over views proves the regret inequality in Eq. (8). If an original model is uniformly optimal against $\mathcal{M}'$, it is also uniformly optimal against the subset $\mathcal{M}$ in every view, proving the set inclusion. The intersection with $\mathcal{M}$ is necessary because a newly added model was absent, not inferior, under the smaller roster.

A.12 Simultaneous regret bounds

On E, $L_{jmv} \leq d_{jmv} \leq U_{jmv}$ for every competitor j. The map $(x_1, \ldots, x_{M-1}) \mapsto \max(0, x_1, \ldots, x_{M-1})$ is coordinatewise monotone, hence

$$\max\{0, \max_{j \neq m} L_{jmv}\} \leq \max\{0, \max_{j \neq m} d_{jmv}\} \leq \max\{0, \max_{j \neq m} U_{jmv}\}.$$

The middle expression is $r(m, v)$. Applying another monotone maximum over views proves the worst-view statement.

A.13 Pairwise overlap is insufficient

Let $W(v_1) = \{a, b\}$, $W(v_2) = \{b, c\}$, and $W(v_3) = \{a, c\}$. Every pair of winner sets intersects, so there is no disjoint-pair witness. Nevertheless $U = \cap_v W(v) = \varnothing$. An implementation based only on pairwise overlap would therefore incorrectly leave open the existence of a uniform winner.

A.14 Pairwise ranking-dependence validity

On the joint event E, suppose the covered interval for one oriented pair satisfies $L_{mjv} > \delta$ in view v and $U_{mjw} < -\delta$ in view w. Coverage gives $d_{mjv} > \delta$ and $d_{mjw} < -\delta$, so m truly outranks j in v and is truly outranked by j in w. The implication holds for every pair and view simultaneously on E. Therefore choosing any witness after inspecting the complete covered family consumes no additional error budget. This establishes pairwise ranking dependence, not dominance over every other candidate and hence not a global-winner change.

A.15 Exact factorial reconstruction

For $z \in \{-1, +1\}^K$, define

$$S_m(z) = \sum_{B \subseteq [K]} \beta_{m,B} \prod_{k \in B} z_k, \qquad \beta_{m,B} = 2^{-K} \sum_{z} S_m(z) \prod_{k \in B} z_k. \tag{10}$$

The parity functions $\chi_B(z) = \prod_{k \in B} z_k$ form an orthogonal basis for real functions on $\{-1, +1\}^K$ under the uniform inner product. Projecting S_m onto this basis yields Eq. (10); Fourier inversion on the finite group gives exact reconstruction. Linearity implies that the coefficient for a pairwise margin $S_m - S_j$ is $\beta_{m,B} - \beta_{j,B}$.

B Reproducibility and artifact chain

B.1 Blocking execution order

The released artifact follows a fail-closed sequence:

1. freeze the image, category, prompt, model, evaluator, inference, and statistical contracts;
2. create one prediction bank per model and dataset, with adapter and manifest hashes embedded in its metadata;
3. audit exact image membership, dimensions, finite coordinates and scores, deterministic ordering, global cap, and artifact identity without loading references or computing a metric;
4. require every bank to pass, then hash all banks, audits, public references, evaluator code, analysis runners, and tests in the benchmark metric lock;
5. compute point scores and per-image sufficient-event ledgers, checking their reconstruction before resampling; and
6. execute the registered simultaneous family and every sensitivity, preserving all intermediate samples needed to reproduce tables and figures.

Any stale hash, missing bank, pre-existing metric output, conformance mismatch, or failed schema gate aborts execution.

B.2 Task-agnostic adoption surface

The repository is packaged as `viewcert` with a dependency-free Python API and command-line entry point. A maintainer supplies a rectangular JSON matrix `view -> model -> score`; ViewCert returns point winner sets, minimax regret, and, when given a valid simultaneous max-absolute radius, common-winner, dependence, confidence-set, and continuous-δ certificates. The software accepts arbitrary finite score units and therefore does not hard-code detection or AP. It deliberately refuses to infer a simultaneous radius from aggregate scores: construction and validation of that uncertainty family remain the benchmark maintainer's responsibility.

The evaluator-program layer adds typed operators, canonical program hashes, Cartesian-product completeness checks, exact mixed-family frontiers, and componentwise budget inversion. These core verifiers remain dependency free. Research-only NumPy runners construct the numerical-delta process and locked coverage experiments from frozen component bootstrap samples.

B.3 Locked evaluator-program inference

The Open Images grammar contains eight programs, 48 aggregation components, and five competitors for the selected candidate. The exact frontier artifact has SHA-256 `7a4a865177d311db7873a78c6ff18cd7deb0069d013bff086e75059310ffcfda`; the 10,000-draw inference artifact has `1819665a4f4bfca638863d184dc88893d8d42ab7ab386c3a9a93057ab5ab2722`. Before independent simulation, the method, tuning rule, unseen regimes, sample sizes, repetition counts, and pass criteria were frozen under protocol hash `4a881257e70c266dc33928f1e85fac1223874f25f7092402f82971cdeaaac531`. The resulting validation artifact hash is `d88ac08c58a07d5b988d45ac38b44d633d67e2d6860e846c66254ec923726c87`. All development variants remain in artifacts; the more favorable direct outer-minimum band is explicitly barred from the paper's confirmatory claim.

The post-outcome Open Images structured closure has SHA-256 `e23693326c9c411cc908f841b2932b3e89321d42f6c85a67472dcc74a685ba5b`. It reconstructs all 40 margins of the original eight-cell grid before independently selecting either registered support policy for each category. Each of five exact width-two certificates covers 2^{50} programs with 390 local assignments and no leaves; the worst margin is $+11.953$ AP. Because this compiler was constructed after outcomes, it strengthens the algorithmic and conformance evidence but is not labeled prospective.

The deterministic synthetic scaling artifact has SHA-256 `7c262a60080cd30b6d08353b573a2c8a6a9e90d5e7af2748f968c34e5482b17c`. It varies T, τ, q, and H and records exact factor-cell, work, and dynamic-state counts rather than machine-dependent timings. Every row materializes zero program leaves; it is an algorithmic control, not evidence that a real evaluator admits the synthetic factorization.

Historical preregistration and artifact identifiers retain the string `viewcert-p` so that their hashes remain unchanged. This immutable identifier predates the final ViewCert name and does not denote a separate method.

Here and below, an ellipsis marks an abbreviated SHA-256 digest; the released artifacts retain the complete digest.

The prospective WMT25 evaluator-program lock has SHA-256 `73f7aeb297cd6a31eca42904fc49c7bcce8afc25f0abffa7d1fd0b96cb88fcd1`. Its score-blind metadata audit, exact point frontier, and 10,000-draw inference artifacts have SHA-256 `d5193953ba3a19d2...634b314`, `1844a2ce51cf88fc...e4690`, and `5d3db699b15fed04...81c95`. The latter covers all 15,120 ordered program–candidate targets and preserves its zero simultaneous lower-radius curve rather than substituting a smaller post-hoc family. The post-outcome structured-compiler audit then reconstructs all 15,120 point margins from 28 language-pair factors with maximum absolute error 1.34×10^{-15}. Across 420 ordered pairs and edit budgets zero through three, all 1,680 exact variable-elimination optima match direct enumeration; there are no status or witness-replay mismatches. This is a conformance check of the already locked 36-cell surface, not a second prospective outcome or a real-domain scaling claim. Four additional Shy–Gemini checks match direct minimization over all 36 programs and all vertices of the 28-language-pair simplex with zero gap; by Eq. (5), this equality covers every weighting in the full simplex without a weight grid.

B.4 Compute accounting

The frozen independent-replication budget contains 64,950 model forward passes: 54,000 Open Images image–model passes and 10,950 DOTA tile–model passes over 1,825 tiles. DOTA tiles are streamed rather than materialized. Metric resampling reuses per-image sufficient-event ledgers and therefore does not rerun a detector. We report pass counts rather than a hardware-specific wall-clock estimate; the readiness audit required at least 16 GiB of GPU memory and a 73.44-GiB incremental storage reserve.

B.5 Evaluator conformance

The Open Images implementation is differentially tested against a pinned TensorFlow reference on 2,000 deterministic randomized cases. The test space includes verified positive, verified negative, unverified, hierarchy-expanded, regular-box, group-of, threshold-equality, duplicate, and exact-score-tie configurations. Separate constructed fixtures test the seven registered counterfactual cells. DOTA Task 2 matches 500 randomized cases from the pinned public development kit. Task 1, inclusive-coordinate conversion, tile merge, difficult-object handling, and accelerated greedy NMS have dedicated parity fixtures. A conformance failure blocks scientific execution rather than becoming a post-hoc limitation.

B.6 Annotation accounting

No project member draws, edits, verifies, adjudicates, or asks a model or human to generate a reference box or label. Public benchmark annotations determine the cohorts, views, denominators, and resampling strata. Deterministic operations—exact public-name crosswalks, hierarchy propagation specified by Open Images, horizontal envelopes specified by DOTA, image tiling, and schema normalization—are versioned transformations, not new annotation. The paper therefore uses the precise phrase *no new manual annotation by this project*, never the inaccurate phrase *no human annotation*.

B.7 Outcome-independent robustness code

Before the Open Images metric lock, the repository fixes code and unit tests for caps 100/300/1,000; analysis floors 0.001/0.005/0.01; 100 seeded within-tie permutations; ordinary and stratified 10,000-replicate paired bootstraps; 5,000 paired without-replacement draws at each of 50/70/90%; support-weighted, leave-one-category-out, and leave-one-selection-stratum-out aggregation; one checkpoint per family; and every leave-one-family-out roster. The full local test suite has 119 passing tests at lock time. DOTA uses a distinct post-primary implementation lock: its sensitivity questions and settings were preregistered, whereas the runner code was completed after the primary result was known and before any sensitivity output. These analyses diagnose the primary result and cannot replace it. The current source tree contains 258 project tests after adding post-lock theorem checks, public score-matrix and δ-frontier checks, AP-native calibration, typed-program completeness, exact mixed-frontier, structured variable-elimination controls, inferential simulation, incomplete-view edge cases, and WMT artifact, pairwise-reversal, metadata-boundary, and mixed-program audits. Of these, 256 run in the base environment and two corpus-score conformance tests require the pinned optional `sacrebleu==2.5.1` dependency. The Open Images metric lock correctly continues to record the 119 tests present when outcomes were sealed.

B.8 Pre-outcome inferential calibration

The first simulation uses correlated Bernoulli per-image contributions in five edge-case scenarios with 60 contrasts. The max-absolute construction attains minimum joint coverage 0.944 (minimum Wilson lower limit 0.928); the max-t alternative attains 0.921 (0.903) and fails the frozen eligibility gate. The second simulation uses oracle multivariate-Gaussian score errors for the exact Open Images and DOTA family sizes. It draws 100,000 realizations for each critical value and 25,000 independent realizations for coverage and power in each of six benchmark-by-correlation configurations. Estimated joint coverage ranges from 0.9496 to 0.9514 at nominal 0.95. Its grid spans cell standard errors 0.25–2 AP, true minimum winner gaps 0.5–5 AP, and both zero- and one-AP relevance margins. Both simulations are explicitly diagnostic of the simultaneous geometry, not proofs of AP-bootstrap regularity.

C Registered robustness diagnostics

Table 2 collects the outcome-preserving and outcome-changing diagnostics in one place. Point diagnostics are not promoted to inferential certificates, and roster deletions are statements about the listed subset only.

Table 2: Registered robustness summary. "GDB" denotes Grounding DINO Base and "OWL" denotes OWLv2 Base Ensemble.

Diagnostic	Open Images	DOTA-v2
Primary simultaneous family	GDB common winner at $\delta = 0$ and 1 AP; radius 1.973 AP over 162 contrasts.	Not identified at $\delta = 0$ and 1 AP; radius 6.354 AP over 36 contrasts.
Cap and score floor	GDB wins 8/8 views at caps 100/300/1,000 and floors 0.005/0.01.	OWL wins 2/2 tasks under the same cap and floor grid.
Alternative resampling	Ordinary bootstrap certifies GDB at 0/1 AP; GDB is uniform winner in 100% of 5,000 draws at each 50/70/90% retention.	OWL wins both tasks in 98.78/99.96/100% of 5,000 draws at 50/70/90% retention.
Support influence	No winner changes in 48 leave-one-category-out or six leave-one-selection-stratum-out runs; maximum score changes 1.396 and 6.987 AP.	No winner changes in 11 leave-one-primary-category-out runs; maximum score change 4.319 AP.
Exact-score ties	Mixed TP/FP blocks contain 0.00666% of 149.4M events; maximum mAP range 1.52×10^{-5} AP over 100 permutations; no winner changes.	Six events occur in mixed blocks; maximum primary-mAP range is zero over 100 permutations.
Candidate roster	Family representatives retain the GDB certificate; removing the Grounding DINO family leaves no pointwise common winner (OWL and YOLO-World are attainable).	Family representatives remain not identified; removing the Grounding DINO family yields an OWL common-winner certificate at 0/1 AP.

For Open Images, support-weighted category aggregation also retains GDB as the pointwise common winner. For DOTA-v2, support-weighted aggregation retains OWL as the point winner in both tasks. These checks diagnose aggregation choices; the primary estimands remain unweighted macro AP over their frozen category sets.

D AP-native simultaneous-coverage audit

The original outcome-free simulations exercise correlated Bernoulli and Gaussian score families but not the nonlinear AP functional. We therefore froze an additional simulation before reading its outputs. Each of 50,000 pseudo-population images receives one latent object. Every model produces a fixed primary detection score, match quality, and optional spurious detection; these values are reused in all four views. A view changes only the match threshold (0.50, 0.60, 0.70, or 0.80). This creates the same separation as the empirical studies: predictions are immutable and the evaluator changes.

For each of three scenarios, 200 outer trials draw 300 images from the fixed pseudo-population. Inside each trial, 400 paired image bootstraps recompute integral interpolated AP for three models and all views. One studentized max-t family covers the 12 pairwise margins. Ground truth is the exact AP surface of the finite pseudo-population, so the recorded event is joint coverage of every true margin.

Table 3: Frozen AP-native coverage results. Wilson intervals quantify Monte Carlo uncertainty across outer trials.

Scenario	Covered	Familywise coverage	Wilson 95% interval
Stable ranking	191/200	0.955	[0.917, 0.976]
Crossing ranking	185/200	0.925	[0.880, 0.954]
Sparse near tie	195/200	0.975	[0.943, 0.989]

The crossing scenario changes its population winner as the matching threshold tightens; the near-tie scenario combines low detection rates and close model scores. All three pass the frozen rule:

estimated coverage is at least 0.90 and the Wilson interval contains 0.95. This audit exercises the implemented code path and nonsmooth AP behavior under realistic pairing. It remains an empirical calibration rather than a theorem covering every possible detection distribution.

E Prospective WMT transfer

The WMT22 outcome boundary was frozen before any score was inspected. Its design has SHA-256 `a53a0779...d5eaea7` before any candidate score, rank, or archived metric-score value was read. The source is the public `mt-metrics-eval` v2 archive, which bundles sources, references, document IDs, and fixed WMT system outputs. Metadata-only inspection was limited to filenames, line counts, missing-output markers, document IDs, and hashes. The score-blind audit has SHA-256 `cc153a9d...74937` and its score-value exposure flag is explicitly false. We compute BLEU and chrF from text using SacreBLEU 2.5.1 rather than consuming the archived metric-score files.

After the WMT22 output was sealed, we defined a sequential held-out replication. A deterministic metadata rule chose the most recent post-WMT22 en–de test set with two references, at least ten system outputs, and at least 30 documents. WMT25 had no eligible two-reference en–de set, so the rule selected WMT24. Its design was frozen at SHA-256 `f8a81128...ff9dfc`; its score-blind metadata audit is `cb4fdb2b...b2eeb32`. This later study cannot alter the primary WMT22 verdict.

All systems share one exact metric and inference contract. For every candidate, one output text is reused in the four cells $\{\text{BLEU}, \text{chrF}\} \times \{\text{refA}, \text{refB}\}$. BLEU uses 13a tokenization, mixed case, exponential smoothing, and no effective-order shortcut. chrF uses character order six, word order zero, $\beta = 2$, no whitespace, and no epsilon smoothing. Per-segment sufficient statistics are summed within documents. A seeded multinomial bootstrap resamples whole documents once per replicate and shares its 10,000 count vectors across every system and view. One studentized max-t critical value covers every unordered system-pair margin in all four cells: 420 margins for WMT22 and 1,300 for WMT24. BLEU and chrF raw units are never pooled; cross-metric claims use interval signs or within-view standardized regret only.

Table 4: Prospective WMT outcomes. "Nominal" lists distinct point winners; selection and pairwise ranking are separate simultaneous certificates. The WMT25 "Seg." entry is inapplicable because its aggregation components are language pairs.

Set	Seg.	Docs	Sys.	Family	Nominal	Certificate
WMT22 en–de	2,037	181	15	420	3 systems	pairwise reversal; selection NI
WMT24 en–de	998	171	26	1,300	2 systems	selection and ranking NI
WMT25, 28 pairs	–	2,819	21	15,120	Shy	one-edit pair reversal

WMT22's max-t critical value is 3.7336. Its only certified reversal pair is `bleu_bestmbr` versus `chrf_bestmbr`: the former wins under BLEU/refA while the latter wins under both chrF/reference cells. BLEU/refB is unresolved for that pair. The global winner-confidence sets are {JDExploreAcademy, Online-B, `bleu_bestmbr`} for BLEU/refA; six systems for BLEU/refB; and the singleton {`chrf_bestmbr`} for both chrF views. WMT24's critical value is 3.9557; none of its nominal winners is globally certified and no covered pair reverses. The sealed result artifacts have SHA-256 `b96e060f...945131` and `e2c2b704...9b51c`; the common derived reversal certificate is `d8bbcefc...14ef2`.

WMT25 provides a separate full-program transfer. A prospective lock (SHA-256 `73f7aeb2...8fcd1`) fixes the complete program grammar, candidate and language-pair rules, uniform nominal weights, the full weighting simplex, all ordered pair targets, 10,000 document resamples, seed, and numerical-delta tuning before automatic-score values are read. The metadata-only audit (`d5193953...b314`, `score_values_exposed=false`) verifies 59 official files and retains 21 systems complete over 28 language pairs and 2,819 documents. Following the released WMT25 evaluation (Kocmi et al., 2025), the grammar crosses nine registered metric subsets of GEMBA-ESA GPT-4.1, GEMBA-ESA Command A, MetricX-24-Hybrid-XL, XCOMET-XL, and CometKiwi-XL with two normalizations and two aggregation operators, giving 36 mechanically complete programs.

Under the nominal all-five-metric robust-scale paragraph-micro program, Shy is the point winner and simultaneously beats all 20 alternatives. Its nearest nominal margin, against Gemini-2.5-Pro, is 0.0926 with simultaneous 95% interval $[0.0712, 0.1140]$. Replacing only the metric stage by the two LLM judges makes the same margin negative, $[-0.0982, -0.0303]$, while all other stages and uniform weights remain fixed. The two-sided max-t envelope covers all $36 \times 21 \times 20 = 15{,}120$ ordered comparisons. The exact point breakdown curve is $(0.581, 0, 0, 0)$, but the locked simultaneous lower curve is $(0, 0, 0, 0)$; the paper therefore claims a signed semantic reversal, not a positive WMT25 radius. The point and inference artifacts have SHA-256 `1844a2ce...4690` and `5d3db699...1c95`, respectively. The sealed inference artifact records 98 signed semantic reversals; the main text reports one simple one-edit witness rather than treating their count as an estimand.

A human-protocol gate prevented an invalid comparison. The WMT20 crowd-versus-expert extension stopped at its metadata gate because the released human allocations need not share an estimand. Seven non-reference systems had complete MQM allocation and at least 500 WMT-z observations, but their WMT-z intersection contained only 12 segments from one document. The locked minimum was 300 segments across 30 documents. The no-go lock (SHA-256 `263adbfa...954fac`) therefore forbids reading score values, imputation, reweighting, or ranking. The project reports the incompatible sampling masks and performs no human-score analysis.

The released workflow reproduces the complete sequence through `audit_wmt_extension_metadata.py`, `evaluate_wmt_view_extension.py`, `audit_wmt24_replication_metadata.py`, `evaluate_wmt24_view_replication.py`, and `derive_wmt_pairwise_reversal_certificates.py`. The WMT25 chain is `audit_wmt25_program_breakdown_metadata.py`, `analyze_wmt25_program_breakdown_frontier.py`, and `analyze_wmt25_program_breakdown_inference.py`. Each output stores source, design, runner, bootstrap-count, and bootstrap-sample hashes. Unit tests compare the vectorized corpus score to SacreBLEU, check document pairing and deterministic seeds, and prove that metadata auditors never serialize numeric score values.

F Scope, guarantees, and direct controls

Table 5: ViewCert does not replace these threads; it addresses the decision problem created by several admissible deterministic evaluators for one immutable output bank.

Thread	Primary object	Output
Multiverse analysis	Defensible analysis pipelines	Distribution or sensitivity of results
Multicriteria benchmarking	Vector of chosen objectives	Dominance/frontier under a preference relation
Rank uncertainty	One chosen leaderboard target	Sampling-based rank intervals
MT metric/protocol audits	Metric agreement or protocol stability	Diagnosis or preferred-metric evidence
Leaderboard data perturbation	Edited preference records	Smallest drop/add/flip changing a fitted ranking
Adaptive benchmarking (SIREN)	Fresh-data performance of a tune–deploy procedure	Selection-aware repeated-split confidence curves
Breakdown-frontier inference	Relaxed identifying assumptions	Confidence frontier for a target conclusion
ViewCert	Typed evaluator programs; fixed outputs	Verified budget or replayable semantic/aggregation counterexample

At fixed $q = 2$, $\tau = 1$, and $H = 8$, increasing T from 8 to 100 raises work from 30 to 398 and states from 79 to 1,735 while the virtual family grows from 2^8 to 2^{100}. At $T = 30$, $q = 2$, and $\tau = 2$, raising H from 0 to 30 leaves the 230 local assignment tables unchanged and raises dynamic states from 115 to 1,827. These deterministic counts expose the dependence on every theorem parameter without turning machine-specific timing into a universal speedup claim.

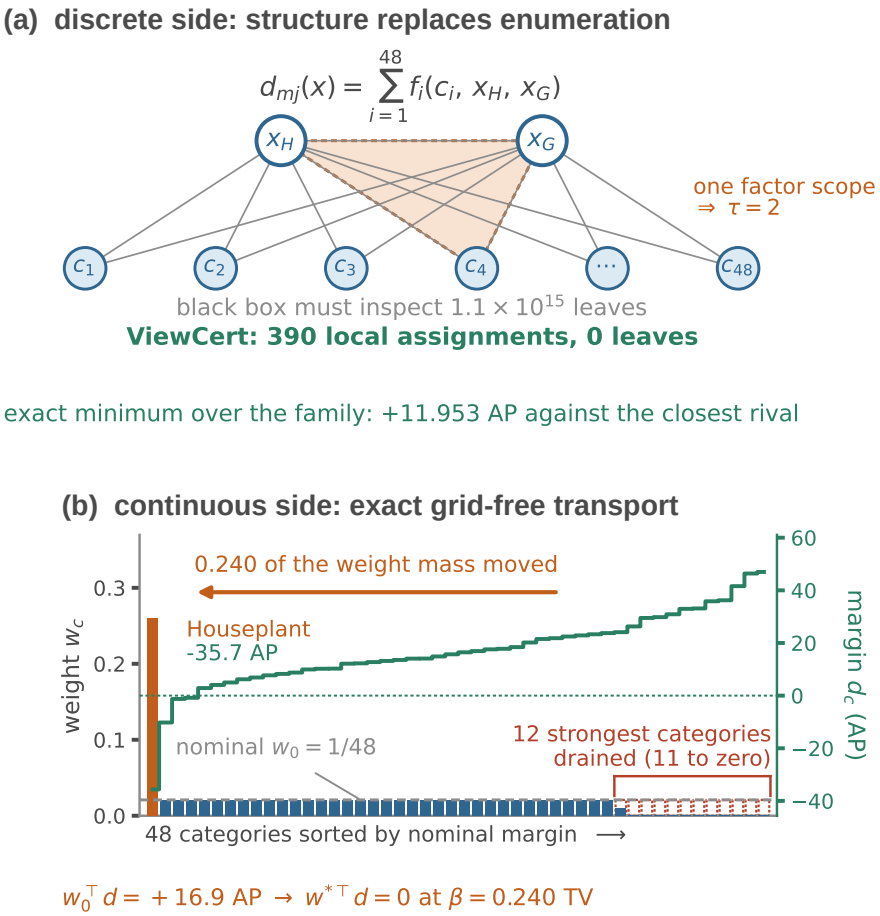


Figure 5: **Exact structured verification.** **(a)** Local factors replace enumeration of 2^{50} programs. **(b)** Optimal transport finds the closest reversing aggregation without a weight grid.

Table 6: Guarantee map. No single notion of guarantee applies uniformly across the four formal layers and the empirical AP-native audit.

Layer	Guarantee	Qualification
Language	Exact signature equality with the Cartesian-product registry	Syntactic completeness for the declared language; not completeness over all defensible evaluators.
Structured optimization	Exact minimum, zero optimality gap, and attaining program	Conditional on complete local tables that reconstruct the frozen pairwise margins; exponential in induced width.
Continuous aggregation	Exact transport radius for a registered program; exact simplex vertex reduction for factored components	Requires linear component aggregation; the width-sensitive result does not cover TV radii for implicit grammars.
Post-search inference	Simultaneous asymptotic lower curve for the selected program, competitor, weights, and budget	Requires the fixed finite atomic family and the consistency, directional-differentiability, rate, and scale conditions of Theorem 4.
AP-native audit	Empirical familywise coverage in three frozen nonlinear AP scenarios	Implementation calibration, not a theorem for every detection distribution.

Table 7: Direct comparators. Each comparison targets a distinct failure mode; no intentionally weakened method is presented as a universal baseline.

Comparator	Controlled evidence	Result
Flat exhaustive enumeration	60 seeded seven-variable problem–budget cases; 1,680 WMT25 pair–budget cases	Identical margin, status, edit count, and witness where enumeration is feasible; flat search is unavailable for the 2^{50} Open Images closure.
Finite aggregation grid	Four WMT25 Shy–Gemini controls enumerate all 36 programs and every simplex vertex	Zero gap to structured simplex verification. A finite interior grid cannot certify the uncountable simplex; transport and vertex reduction are grid-free exact controls.
Raw-coordinate max-t	One envelope over 5,760 Open Images component margins	Lower curve is zero at every budget, versus $(0.187, 0.158, 0.158, 0.158)$ for atomic coverage.
Direct outer-minimum bootstrap	Frozen near-tie and active-program-tie simulations	Undercoverage at the nonregular outer minimum; the method is barred from confirmatory claims rather than substituted after inspection.
Ordinary multiverse table	The complete eight-program Open Images cube	Reports all cell scores and effects, but by itself gives neither an implicit 2^{50} certificate nor post-search coverage for the worst witness.

Table 8: Hardware-independent scalability at $T = 18$ variables and edit budget $H = 6$. "Work" is the number of joint assignments evaluated and "states" is the number of dynamic-program states created; no program leaf is materialized.

Options q	Width τ	Virtual programs	Work	States
2	1	2.62×10^5	70	215
2	2	2.62×10^5	134	409
2	3	2.62×10^5	254	769
2	4	2.62×10^5	478	1,433
3	1	3.87×10^8	156	319
3	2	3.87×10^8	444	901
3	3	3.87×10^8	1,254	2,521
3	4	3.87×10^8	3,522	7,003
4	1	6.87×10^{10}	276	423
4	2	6.87×10^{10}	1,044	1,587
4	3	6.87×10^{10}	3,924	5,907
4	4	6.87×10^{10}	14,676	21,843