CETRA: COMPOSABLE EVIDENCE CERTIFICATES FOR TOOL-ADAPTIVE RISK CONTROL OF MULTIMODAL ASSERTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adaptive multimodal agents select tools using earlier observations, so calibration for one acquisition policy need not survive policy recomposition. We introduce **CETRA**, which converts tool outputs into action–history evidence increments for assertions fixed before acquisition. Cross-fitted calibration bounds logging-reference null means, while unsupported configurations contribute neutral increments. Under independent source-level calibration, stable tools, and explicit full-history conditional null-mean bounds, the product is a nonnegative supermartingale for each successful calibration artifact. Accounting for calibration failure yields anytime false-certification control across admissible acquisition and stopping policies. We evaluate CETRA in CETRA-Sim, ChartQA, and DocVQA. At $\alpha = 0.10$, CETRA passes the policy-wise empirical confidence-bound checks on both real datasets, with worst-policy risk estimates 0.071 and 0.068, while reusing one calibration object per dataset across seven policies. Coverage exceeds static certification by 13.5 and 13.4 percentage points. Policy-specific recalibration achieves higher coverage with separate full allocations, whereas CETRA achieves higher coverage under the evaluated matched total calibration-trajectory budgets. These results characterize the empirical coverage–reuse trade-off while keeping the assumptions required for theoretical certification explicit.

1 INTRODUCTION

Vision–language models increasingly control multimodal agents that gather additional evidence before answering. An agent may inspect an image crop, invoke OCR, extract a table, retrieve context, or execute a calculation, choosing subsequent actions from earlier observations. Recent work on active visual perception and tool-augmented reasoning demonstrates the value of such adaptive acquisition (Srinivasan et al., 2024; Zheng et al., 2026; Liu et al., 2026). The resulting evidence, however, is selected by a policy rather than obtained through a fixed observation process.

This creates a reliability problem beyond final-answer accuracy. A threshold calibrated for a tool in isolation need not remain valid when an agent selectively invokes that tool, compares several outputs, or stops when the evidence appears favorable. Even individually valid tests can become anti-conservative when an agent accepts whichever provides stronger support. Likewise, changing a calibrated procedure's action order, routing rule, tool subset, or stopping behavior can change its false-certification probability. We call this *policy-recomposition failure*. Our target is the probability of certifying an unsupported prespecified assertion, not the error fraction among all certified assertions.

Existing methods address related but distinct certification objects. Budgeted Conformal Evidence Acquisition (BCEA) calibrates the complete post-acquisition mapping, allowing sequential and adaptive decisions within a procedure fixed independently of calibration (Xu et al., 2026). Changing that mapping does not automatically preserve its certificate; additional calibration or simultaneous control over a candidate policy family may be needed. Proof-of-Perception (PoP) instead attaches conformal prediction sets to intermediate perception and reasoning nodes (Fayyazi & Akrami, 2026). Marginal coverage of individual nodes does not, by itself, provide an assertion-level sequential guarantee after adaptive composition. The question we study is therefore different: *under what*

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.

explicit conditions can evidence primitives be calibrated once and reused across changed acquisition policies?

We introduce **CETRA** to provide this primitive-level interface. CETRA converts each selected tool output into a bounded support score and assigns its preceding history to an action-specific calibration cell. Cross-fitted calibration bounds the corresponding logging-reference null mean. A declared conditional mean-domination factor specifies the permitted relationship between that reference mean and the deployment score at the actual observed history. CETRA normalizes the score into a nonnegative evidence increment and accumulates these increments until certification or termination. Configurations without usable calibration support contribute neutral increments rather than extrapolated evidence.

The theoretical guarantee separates finite-sample calibration from the conditional assumption needed for sequential reuse. Under independent source-level calibration and honest cross-fitting, simultaneous upper bounds hold with a specified calibration failure probability. For each successful offline artifact, the declared full-history null-mean bounds and non-anticipating acquisition make the certificate a nonnegative supermartingale. Standard anytime-valid inference then gives the crossing bound (Ramdas et al., 2023); incorporating calibration failure yields the total false-certification guarantee. Cell-average bounds and propensity diagnostics do not establish the conditional mean-domination assumption. The sequential closure argument is standard (Ramdas et al., 2023); CETRA's contribution is the reusable action–history calibration interface and the characterization of its sufficient conditional envelope (Section 4.3), not a certified estimator of that envelope from logging data alone.

This construction separates acquisition from certification. Fixed-order, random, uncertainty-driven, and cost-aware policies can use the same primitive calibrators within a dataset, provided the conditions for non-neutral increments continue to hold. The method does not claim unrestricted validity after arbitrary policy or tool changes. Its practical value is tested through controlled acquisition in CETRA-Sim and heterogeneous chart and document tools on ChartQA and DocVQA, retaining greedy VLM decoding. CETRA passes the policy-wise empirical risk checks and improves coverage over static certification, while separately recalibrated policies achieve higher coverage with their own allocations. Mechanism ablations examine the role of history conditioning and calibration safeguards; matched-budget experiments separately examine reuse under a fixed trajectory allocation.

Our contributions are:

1. We formulate policy-recomposition failure for prespecified assertions, distinguishing primitive-level certification from marginal node coverage and calibration of a fixed complete acquisition procedure.
2. We introduce CETRA, a cross-fitted, support-aware construction of reusable action–history evidence increments. We establish an anytime false-certification bound under explicit sampling, tool-stability, and conditional mean-domination assumptions, accounting for calibration uncertainty, and characterize the sufficient within-cell envelope $\Gamma^\star$ on a fixed history domain.
3. We evaluate CETRA on CETRA-Sim, ChartQA, and DocVQA using policy-wise risk checks, certified coverage, acquisition cost, mechanism ablations, and matched calibration-trajectory budgets. On each real-data benchmark, CETRA and policy-specific recalibration are evaluated on the same seven-policy suite.

2 Related Work

Adaptive multimodal evidence acquisition. ReCoVERR collects follow-up visual evidence to reduce unnecessary abstention (Srinivasan et al., 2024), DeepEyes learns active visual perception through reinforcement learning (Zheng et al., 2026), and Liu et al. (2026) formulate high-resolution inspection as sequential experimental design. These methods motivate adaptive evidence acquisition. CETRA addresses a complementary question: how to certify a prespecified assertion when the acquisition policy is recomposed, rather than how to optimize evidence selection itself.

Conformal risk control and sequential evidence. Conformal Risk Control extends conformal prediction to control expected monotone losses under exchangeability (Angelopoulos et al., 2024). Bai & Jin (2026) construct generalized conformal e-values for controlling bounded risks among

selected predictions; the increment in Eq. (7) is the same betting construction (Waudby-Smith & Ramdas, 2024). In sequential inference, products of nonnegative increments with conditional expectation at most one form test supermartingales, enabling anytime-valid testing (Ramdas et al., 2023). CETRA uses this standard closure property rather than proposing a new optional-stopping theorem. Its construction combines finite-sample bounds on logging-reference null means with an explicitly assumed full-history conditional mean envelope. Cell-average calibration alone does not establish that envelope; the resulting false-certification bound also accounts for calibration failure.

Closest risk-control mechanisms. Budgeted Conformal Evidence Acquisition (BCEA) calibrates the complete post-acquisition mapping, including sequential acquisition and stopping, provided that mapping is fixed independently of the calibration sample (Xu et al., 2026). Selection within a prespecified policy family can also be handled through simultaneous calibration. Thus, complete-policy methods are not restricted to non-adaptive acquisition, nor do they inherently require disjoint labeled datasets for different policies. BCEA's selective-risk objective concerns errors among asserted claims, whereas CETRA controls certification conditional on an assertion being unsupported.

Proof-of-Perception (PoP) equips perception and reasoning nodes with split-conformal prediction sets and uses their uncertainty to guide computation (Fayyazi & Akrami, 2026). Its formal coverage guarantees are marginal and node-wise; assertion-level sequential error control does not follow from those guarantees alone. Conformal Policy Control regulates an optimized policy relative to a safe reference policy under a specified risk constraint (Prinster et al., 2026). Baghi (2026) address policy-induced calibration groups through terminal-pattern calibration and simultaneous validation of complete policy–pattern pairs, showing that a guarantee proved for a grouping map fixed independently of calibration need not transfer once the policy makes that map calibration-dependent.

CETRA instead places the reusable calibration object at the action–history level and accumulates evidence for a prespecified assertion. Its distinction is a primitive-level interface whose enabling condition is identified and estimable rather than merely declared, not a claim that intermediate or complete-policy calibration cannot handle adaptivity: a pattern absent from the logged set has no terminal-pattern object at all, whereas CETRA calibrates action–history primitives that an unseen *composition* of logged primitives can reuse. Our BCEA-style and PoP-style comparisons are adaptations to CETRA's evaluation criterion, not tests of the original methods' guarantees.

3 Problem Formulation

Let (x, c) follow a specified deployment population, where x is the initial multimodal input and c is produced by a fixed assertion rule before evidence acquisition. An externally defined support label $Y(x, c) \in \{0, 1\}$ determines the null

$$H_0^c : \quad Y(x, c) = 0. \tag{1}$$

False-certification risk is the probability of certifying under this null population, not the error rate among certified assertions.

Let $\mathcal{D}$ denote the offline fitting and calibration artifact, independent of the deployment example. Write h_t for the complete deployment history, including x, c, acquired outputs, and any additional information used by the policy or certificate. Define

$$\mathcal{F}_t = \sigma(h_t), \qquad \mathcal{G}_t = \sigma(\mathcal{D}) \vee \mathcal{F}_t, \qquad a_t \sim \pi(\cdot \mid \mathcal{G}_{t-1}). \tag{2}$$

The pre-output information $\mathcal{G}_{t^-}$ includes the selected action and everything available before observing o_t. Action selection is non-anticipating, whereas termination may be a $\mathcal{G}_t$-stopping time $\tau \le T_{\max}$.

Logged trajectories are generated by a behavior policy μ. Changing action order, routing, tool subset, or stopping behavior is *policy recomposition*. For the admissible class Π defined by the conditions below, CETRA seeks one reusable certificate satisfying

$$\sup_{\pi \in \Pi} \Pr_{\mathcal{D}, H_0^c, \pi} \left(\sup_{t \le T_{\max}} M_t^c \ge \frac{1}{\alpha_{\text{seq}}} \right) \le \alpha, \tag{3}$$

where sequential testing and calibration share the total budget α.

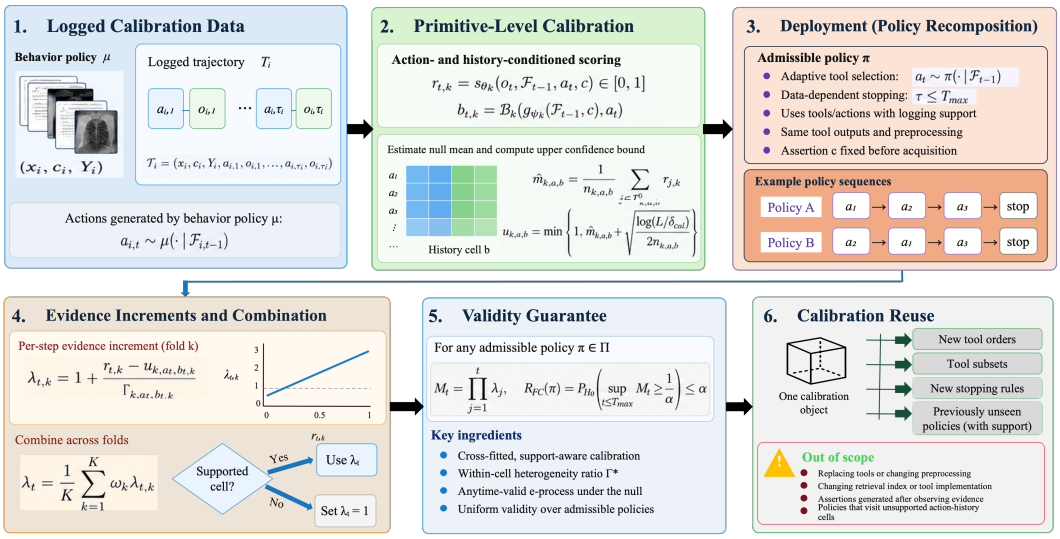

Figure 1: Offline logs calibrate action–history reference means. At deployment, an admissible policy selects tools and CETRA accumulates normalized evidence into an assertion certificate. Unsupported tuples contribute neutral increments; validity requires the stated conditional mean bounds.

4 CETRA

4.1 Cross-Fitted Primitive Calibration

Calibration assumes independent, identically distributed source units; dependent questions, assertions, and rollouts from one source belong to the same unit and fold. For fold k, the scorer s_{θ_k}, history encoder g_{ψ_k}, and finite partition $\mathcal{B}_k$ are fitted without that fold. An invocation produces

$$r^c_{t,k} = s_{\theta_k}(o_t, h_{t-1}, a_t, c) \in [0,1], \qquad b_{t,k} = \mathcal{B}_k(g_{\psi_k}(h_{t-1}, c), a_t). \tag{4}$$

The score is a bounded statistic, not necessarily a calibrated probability.

For each tuple, a fixed rule aggregates a source's matching null invocations into $R_{i,k,a,b} \in [0,1]$. Let $\mathcal{I}^0_{k,a,b}$ contain the contributing held-out sources and $n_{k,a,b} = |\mathcal{I}^0_{k,a,b}|$. CETRA computes

$$\begin{aligned} \widehat{m}_{k,a,b} &= \frac{1}{n_{k,a,b}} \sum_{i \in \mathcal{I}^0_{k,a,b}} R_{i,k,a,b}, \\ u_{k,a,b} &= \min\left\{1, \widehat{m}_{k,a,b} + \sqrt{\frac{\log(L/\delta_{\mathrm{cal}})}{2n_{k,a,b}}}\right\}. \end{aligned} \tag{5}$$

Here L is a deterministic upper bound on all candidate tuples, fixed before calibration, and δ_{cal} is the calibration failure budget. Set $u_{k,a,b} = 1$ when $n_{k,a,b} = 0$. Tuples with $n_{k,a,b} < n_{\min}$ are unsupported. The reference mean $m^{\mu}_{k,a,b}$ is the population mean of this source-level statistic among contributing sources, conditional on the fold's fitted objects.

4.2 Support-Aware Evidence Increments

For non-neutral increments, we require logging support and unchanged tool mechanisms, including their input arguments. For every $\pi \in \Pi$, the substantive assumption is the full-history conditional null-mean bound

$$\mathbb{E}_{H^c_0,\pi}\left[r^c_{t,k} \mid \mathcal{G}_{t-}\right] \le \Gamma_{k,a,b} m^{\mu}_{k,a,b}, \qquad b = b_{t,k}, \quad \Gamma_{k,a,b} \ge 1. \tag{6}$$

This bound conditions on the realized offline artifact as well as the deployment history. Neither cell-average calibration nor action-propensity ratios establish it; Section 4.3 characterizes the sufficient envelope.

For a supported tuple with a finite positive denominator, define

$$\lambda^c_{t,k} = 1 - \eta_{k,a,b} + \eta_{k,a,b} \frac{r^c_{t,k}}{\Gamma_{k,a,b} u_{k,a,b}}, \qquad \eta_{k,a,b} \in [0,1], \tag{7}$$

$$\Lambda^c_t = \sum_{k=1}^{K} \omega_k \lambda^c_{t,k}, \qquad \omega_k \geq 0, \quad \sum_{k=1}^{K} \omega_k = 1. \tag{8}$$

Betting fractions are fitted outside the corresponding calibration fold, and mixture weights are fixed before deployment. Unsupported or unusable tuples receive $\lambda^c_{t,k} = 1$. No independence between tool outputs or fold increments is assumed.

Our score pipeline is deterministic given the offline artifact, observed input, history, and action, so Eq. (6) specializes to a pointwise null-score envelope. Appendix B.6 shows that false crossings require an exceedance of the deployed score envelope; without establishing calibration success, such exceedance do not isolate violations of A5 or certify lower bounds on $\Gamma^\star$.

4.3 Conditional Envelope for Primitive Reuse

Fix the offline artifact, logging law, and null population. Let $q_{k,a}(h)$ be the full-history conditional null score mean on a common supported history domain $\mathcal{H}_{k,a,b}$. For $m^\mu_{k,a,b} > 0$, define

$$\Gamma^\star_{k,a,b} = \max\left\{1, \sup_{h \in \mathcal{H}_{k,a,b}} \frac{q_{k,a}(h)}{m^\mu_{k,a,b}}\right\}. \tag{9}$$

Any finite $\Gamma_{k,a,b} \geq \Gamma^\star_{k,a,b}$ ensures Eq. (6) throughout this domain. This policy-independent characterization is not an estimation guarantee: sample quantiles, empirical maxima, and bootstrap summaries alone do not certify the supremum. The experiments retain declared factors and the two-way budget in Eq. (12); estimation diagnostics and their limitations are discussed in Appendix B.

4.4 Sequential Certification

Starting from $M^c_0 = 1$, update

$$M^c_t = M^c_{t-1} \Lambda^c_t, \tag{10}$$

$$\tau^c_{\mathrm{cert}} = \inf\left\{t \leq T_{\max} : M^c_t \geq \frac{1}{\alpha_{\mathrm{seq}}}\right\}, \tag{11}$$

with $\inf \varnothing = \infty$ and

$$\alpha_{\mathrm{seq}} > 0, \qquad \delta_{\mathrm{cal}} > 0, \qquad \alpha_{\mathrm{seq}} + \delta_{\mathrm{cal}} \leq \alpha < 1. \tag{12}$$

Certification occurs only if the threshold is reached before termination. The policy may use previously observed evidence and certificate values. Unsupported calls are neutral, but subsequent non-neutral calls must still satisfy Eq. (6) under the complete resulting history. Operational diagnostics are detailed in Appendix A.2.

5 Theoretical Guarantees

Under source-level independence, honest cross-fitting, and the prespecified candidate bound L, Hoeffding's inequality and a union bound give

$$\mathcal{E}_{\mathrm{cal}} = \bigcap_{(k,a,b)} \{m^\mu_{k,a,b} \leq u_{k,a,b}\}, \qquad \Pr_{\mathcal{D}}(\mathcal{E}_{\mathrm{cal}}) \geq 1 - \delta_{\mathrm{cal}}. \tag{13}$$

For almost every fixed artifact $d \in \mathcal{E}_{\mathrm{cal}}$, Eq. (6), convex mixing, and the tower property give

$$\mathbb{E}_{H^c_0,\pi}[\Lambda^c_t \mid \mathcal{G}_{t^-}] \leq 1, \tag{14}$$

$$\mathbb{E}_{H^c_0,\pi}[\Lambda^c_t \mid \mathcal{G}_{t-1}] \leq 1, \qquad \pi \in \Pi. \tag{15}$$

where the conditional law fixes $\mathcal{D} = d$. Thus M^c_t is a nonnegative supermartingale under that conditional null law.

Setting	Dataset	Evidence actions
Controlled	CETRA-Sim	Four stochastic probes
Charts	ChartQA	Crop, OCR, DePlot, calculation
Documents	DocVQA	Crop, OCR, layout, table parsing, calculation

Table 1: Evaluation benchmarks and available evidence actions. The acquisition horizon is at most four actions.

Theorem 1 (Anytime-valid policy-composable certification)**.** *Assume the sampling and fitting conditions above, an independent deployment example, non-anticipating policies, stable supported tools, and Eq. (6) uniformly over* Π*. With the predictable increments and risk allocation above,*

$$\sup_{\pi\in\Pi} \Pr_{\mathcal{D},H_0^c,\pi}\left(\sup_{t\le T_{\max}} M_t^c \ge \frac{1}{\alpha_{\mathrm{seq}}}\right) \le \alpha_{\mathrm{seq}} + \delta_{\mathrm{cal}} \le \alpha. \tag{16}$$

Proof sketch. Fix an offline realization $d \in \mathcal{E}_{\mathrm{cal}}$. Conditional increment validity and the tower property give the supermartingale property under the actual deployment filtration. Ville's inequality bounds threshold crossing by α_{seq} for each admissible policy. Integrating over d, and bounding unsuccessful calibration artifacts by probability δ_{cal}, proves the result. This is an unconditional crossing bound; the raw product is not asserted to be an e-process unconditionally over calibration randomness. This closure step uses standard test-supermartingale arguments (Ramdas et al., 2023). Our guarantee uses A5 as a sufficient condition linking logging-reference null means to full-history conditional deployment scores.

The guarantee permits policy recomposition under the stated conditions, but does not cover post-acquisition assertion selection or arbitrary tool drift. Appendix B contains the full assumptions, proofs, finite-library calibration comparison, and the conditional extension in Theorem 2, which requires a separately justified simultaneous upper bound on $\Gamma^\star$. Joint control for prespecified assertions appears in Appendix A.4.

6 Experiments

We evaluate false-certification control under policy recomposition, reuse of primitive calibration across acquisition policies, and the resulting coverage–cost trade-off.

Benchmarks and assertions. Table 1 summarizes the evaluation. CETRA-Sim isolates adaptive selection, evidence dependence, stopping, and logging-to-deployment shift. ChartQA and DocVQA test certification with heterogeneous chart and document tools (Masry et al., 2022; Mathew et al., 2021; Liu et al., 2023). Real-data assertions are fixed before acquisition and assigned support labels using externally specified, dataset-specific reference-answer criteria. Supported assertions come from reference answers or correct initial model predictions; unsupported assertions include naturally incorrect predictions and type-preserving perturbations. Natural-error and perturbation analyses are reported separately in the supplement.

Backbone, fitting, and logging. Real-data experiments use frozen Qwen3-VL-8B-Instruct with greedy VLM decoding (Bai et al., 2025). Cached tool outputs are shared across certification methods. Score-training, calibration, and test partitions are chart- or document-disjoint. Scorers and history partitions are fitted outside their corresponding calibration folds; learned acquisition routers are fitted on score-training trajectories and frozen before calibration. The behavior policy mixes a task policy with uniform exploration over available actions, using exploration weight $\varepsilon = 0.3$. Randomized action selection does not introduce token sampling into VLM generation.

Policies and comparisons. The real-data suite contains seven deployment policies: three fixed action orders, random selection, uncertainty-driven routing, cost-aware routing, and predicted certificate-growth routing. All methods, including BCEA-style complete-policy recalibration, are

evaluated on the same seven-policy suite within each dataset. Two fixed orders are absent as complete sequences from the behavior-policy logs used for CETRA's primitive calibration, but are included in the policy-specific recalibration comparison. CETRA reuses one primitive calibration object across all seven policies, whereas complete-policy recalibration constructs a separate calibration object for each policy.

Baselines comprise static certification, marginal per-tool calibration, PoP-style node-wise calibration (Fayyazi & Akrami, 2026), a product of marginally calibrated tool evidence, BCEA-style complete-policy recalibration, and complete-policy threshold reuse (Xu et al., 2026). The product of marginally calibrated tool evidence is labelled "Marginal-score product"; conditional e-validity after recomposition is not assumed for it. Comparisons share assertions, primitive scores, tools, and test examples, while each method applies its own certification rule. The BCEA-style and PoP-style implementations are adaptations to this evaluation, not tests of the original methods' guarantees under different risk objectives.

Complete-policy recalibration receives a separate calibration allocation per deployment policy. Additional matched-budget comparisons equalize the total number of null calibration trajectories. This is a trajectory-budget comparison, not a claim that complete policies necessarily require disjoint labeled sources. Dataset-specific tools, policies, and allocation protocols are provided in Appendix C.

Metrics and validity criterion. We report false-certification risk conditional on $Y = 0$, certified coverage conditional on $Y = 1$, mean acquisition cost, and calibration demand. For each policy π, let $U_{0.95,\pi}$ denote the one-sided 95% Clopper–Pearson upper bound computed from its false-certification and null test counts. A method passes the empirical validity check for a partition only when

$$\max_{\pi \in \Pi_{\text{test}}} U_{0.95,\pi} \leq \alpha. \tag{17}$$

Mean risk across policies is descriptive only. Primary comparisons use twenty partitions, with target-risk evaluations at $\alpha \in \{0.05, 0.10, 0.20\}$. CETRA allocates $\alpha_{\text{seq}} = 0.8\alpha$ and $\delta_{\text{cal}} = 0.2\alpha$. The empirical check concerns the tested policy suite; it does not establish the full-history mean-domination assumption for arbitrary new policies. Failure to pass the check is not, by itself, evidence that the true risk exceeds the target.

Stress tests and ablations. Stress tests examine changed action orders, omitted or repeated actions, optional stopping, correlated evidence, reduced logging overlap, and corrupted, unavailable, or shifted tools. Mechanism ablations remove history conditioning, shift correction, unsupported-cell protection, cross-fitting, or sequential accumulation. Detailed configurations and secondary analyses are reported in the supplement.

7 RESULTS

We evaluate controlled acquisition in CETRA-Sim and real multimodal evidence on ChartQA and DocVQA. Unless otherwise stated, the target risk is $\alpha = 0.10$. Passing the empirical check means satisfying the policy-wise confidence-bound criterion in Eq. (17); it is not inferred from an averaged risk estimate alone.

Controlled policy recomposition. CETRA-Sim uses seven deployment policies and twenty partitions. Table 2 reports a worst-policy risk of 0.0207 for CETRA, which passes the all-policy empirical check in all twenty partitions. Marginal tool calibration, node-wise correction, the marginal-score product, and complete-policy threshold reuse have worst-policy risk estimates above 0.10. Separately recalibrating each complete policy passes the check. These comparisons distinguish empirical failures of certificate reuse from adaptive acquisition itself.

Coverage, cost, and calibration allocation. CETRA increases coverage from 0.3785 to 0.6317, a gain of 25.3 percentage points over static certification. Policy-specific recalibration achieves 0.8418 coverage at cost 3.88, compared with CETRA's 5.85, showing the utility cost of the reusable construction in this evaluation. CETRA uses $30{,}000$ shared null trajectories, whereas policy-specific recalibration uses $140{,}000$ across seven policies: a 4.67-fold difference in the allocated trajectory budget. Detailed policy-level results and the $\alpha \in \{0.05, 0.20\}$ sweep are in Appendices D and D.1.

Method	Worst-policy risk ↓	Coverage ↑	Cost ↓	Passes check
Static	0.0829	0.3785 ± 0.0031	0.00	Yes
Marginal tool	0.7488	0.8945 ± 0.0021	3.30 ± 0.02	No
Node Bonferroni	0.2385	0.5714 ± 0.0047	5.25 ± 0.02	No
Marginal-score product	0.5030	0.6899 ± 0.0025	5.88 ± 0.01	No
Policy recalibration	0.0833	$\mathbf{0.8418} \pm 0.0022$	3.88 ± 0.01	Yes
Policy reuse	0.1150	0.8399 ± 0.0032	3.91 ± 0.03	No
CETRA	**0.0207**	0.6317 ± 0.0023	5.85 ± 0.01	**Yes**

Table 2: CETRA-Sim at $\alpha = 0.10$. Entries are means over twenty partitions; intervals are *standard errors* of those means, not across-partition standard deviations. The check uses Eq. (17). Policy recalibration receives a separate calibration allocation per policy; bold coverage identifies the highest coverage among methods passing the check.

Method	ChartQA risk / cov.	DocVQA risk / cov.	Passes check
Static	$0.084/0.288$	$0.081/0.324$	Yes
Marginal tool	$0.196/0.618$	$0.181/0.663$	No
PoP-style node-wise	$0.143/0.493$	$0.132/0.536$	No
Marginal-score product	$0.162/0.551$	$0.148/0.584$	No
BCEA-style recalibration	$0.083/\mathbf{0.468}$	$0.079/\mathbf{0.503}$	Yes
BCEA-style policy reuse	$0.147/0.529$	$0.136/0.562$	No
CETRA	$\mathbf{0.071}/0.423$	$\mathbf{0.068}/0.458$	**Yes**

Table 3: Multimodal policy recomposition at $\alpha = 0.10$. Dataset columns report worst-policy empirical risk and certified coverage. The check uses the policy-specific upper bounds in Eq. (17) on both datasets. Bold coverage identifies the highest coverage among methods passing the check.

Tool-kernel shift. In the separate kernel-shift experiment, CETRA's worst-policy risk estimate increases from 0.0214 without shift to 0.0485 at additive shift 0.25, then to 0.1031 and 0.1895 at shifts 0.50 and 0.75. The latter estimates exceed the nominal target. These perturbations probe changes to the evidence-generating mechanism, rather than policy recomposition under the stable-tool conditions of Theorem 1.

7.1 General Multimodal Policy Recomposition

ChartQA and DocVQA combine visual, textual, structural, and computational evidence with fixed-order and adaptive acquisition policies. For each dataset, CETRA constructs one primitive calibration object and reuses it across that dataset's deployment-policy suite. Tool and policy details are provided in Appendix C.3. CETRA passes the empirical check on both datasets, with worst-policy risk estimates of 0.071 and 0.068. Marginal calibration, node-wise certification, marginal-score multiplication, and complete-policy threshold reuse have risk estimates above the target, whereas policy-specific recalibration passes the check. CETRA improves coverage over static certification by 13.5 percentage points on ChartQA and 13.4 points on DocVQA — a floor, since static certification acquires no evidence. Against policy-specific recalibration granted a separate allocation per policy, CETRA is 4.5 and 4.5 points *lower* on coverage, though the ordering reverses at a matched trajectory budget (Section 7.2): a reuse price, not dominance.

Mechanism ablations. Removing history conditioning increases worst-policy risk from 0.071 to 0.151 on ChartQA and from 0.068 to 0.142 on DocVQA. Removing shift correction or unsupported-cell protection also yields estimates above 0.10 on both datasets (Tables 13 and 20). These comparisons support the practical role of the calibration components under the tested policies; they do not establish the assumed full-history conditional mean bound.

Budget	Method	ChartQA risk / cov.	DocVQA risk / cov.
25%	CETRA	0.089/0.332	0.087/0.363
	Policy recalibration	0.101/0.290	0.099/0.322
50%	CETRA	0.078/0.381	0.076/0.418
	Policy recalibration	0.093/0.345	0.090/0.381
100%	CETRA	0.071/0.423	0.068/0.458
	Policy recalibration	0.085/0.401	0.082/0.438

Table 4: Matched total calibration-trajectory budgets at $\alpha = 0.10$. Entries report worst-policy empirical risk and certified coverage. Budget fractions are dataset-specific. Point estimates alone do not establish whether the confidence-bound validity check is passed.

7.2 Calibration Reuse under a Matched Budget

The main complete-policy comparison supplies a separate calibration allocation to each deployment policy. We additionally fix the total number of null calibration trajectories. CETRA uses the shared logging set, while policy-specific recalibration divides the available allocation across seven policies. Table 4 reports three fractions of each dataset's full allocation. CETRA has higher coverage and lower worst-policy risk point estimates at every reported allocation. At the full matched budget, its coverage advantage is 2.2 percentage points on ChartQA and 2.0 points on DocVQA, in contrast to the higher coverage of policy-specific recalibration with separate full allocations. This result concerns the stated trajectory-budget division, not an intrinsic labeled-sample complexity advantage: complete policies need not use disjoint labeled sources. Full comparisons are in Tables 14 and 21.

8 Limitations and Conclusion

Limitations. CETRA's guarantee requires the stated sampling and fitting conditions, stable tool mechanisms, and the full-history conditional null-mean bound in Eq. (6). Cell-average calibration, propensity diagnostics, and mechanism ablations do not establish this bound. For our deterministic score pipeline, it becomes a pointwise null-score envelope, as detailed in Appendix B.6. Section 4.3 characterizes the sufficient envelope $\Gamma^\star$, but the proposed regression diagnostic does not provide a certified upper bound on it. The experiments retain declared factors: their empirical risk checks neither verify the theorem's assumptions nor isolate A5 violations from calibration failure. Finer history partitions can increase calibration demand and unsupported configurations, while conservative normalization may reduce coverage and increase acquisition cost. The empirical findings concern the evaluated populations and policy suites; they do not establish validity for arbitrary future policies, post-acquisition assertions, or unmodeled tool changes. Certificates address support under the specified criterion, not unrestricted response correctness or deployment safety.

Conclusion. We introduced CETRA, a primitive-level interface for certifying evidence gathered by adaptive multimodal agents. Under explicit conditional mean bounds, its cross-fitted calibration construction yields an anytime false-certification bound that accounts for calibration uncertainty. CETRA-Sim, ChartQA, and DocVQA demonstrate empirical reuse across the tested policies and improved coverage over static certification. Separately recalibrated policies retain higher coverage with their own full allocations, while matched-budget comparisons support reuse under the evaluated trajectory allocation rather than an intrinsic labeled-sample complexity advantage. CETRA thus connects reusable evidence primitives to adaptive assertion certification while making the conditions and trade-offs of that reuse explicit.

Reproducibility Statement

Appendix C specifies splits, folds, partition granularity, support threshold, betting-fraction grid, risk allocation, and the behaviour policy; Appendix B gives the shift-factor estimator and its guarantee. Table entries are means over twenty partitions, with standard errors reported alongside them; per-partition, per-policy counts (false-certification count, null denominator, coverage numerator and denominator, acquisition cost, declared Γ) are released with the supplementary material, together with cached tool outputs and primitive scores sufficient to reproduce every table without vision–language model inference.

Use of Large Language Models

Large language models were used for prose editing throughout the manuscript and as a discussion aid. All experiments were implemented and executed by the authors, and every number reported in this paper is the output of those experiments; large language models did not generate, modify, estimate, or interpolate any experimental result. The authors take full responsibility for the entire content of the paper.

Ethics Statement

This work concerns reliability certificates for automated assertions. A CETRA certificate speaks only to support under a specified evidence criterion; it does not establish completeness, causal validity, clinical correctness, or deployment safety, and should not be presented to a downstream user as though it did.

References

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Melika Baghi. Conditional validity for adaptive modality acquisition: When the policy chooses its own calibration group. *arXiv preprint arXiv:2608.15520*, 2026. doi: 10.48550/arXiv.2608.15520.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Tian Bai and Ying Jin. Conformal selective prediction with general risk control. *arXiv preprint arXiv:2603.24704*, 2026. URL https://arxiv.org/abs/2603.24704.

Arya Fayyazi and Haleh Akrami. Proof-of-perception: Certified tool-using multimodal reasoning with compositional conformal guarantees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5144–5153, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Fayyazi_Proof-of-Perception_Certified_Tool-Using_Multimodal_Reasoning_with_Compositional_Conformal_Guarantees_CVPR_2026_paper.html.

Anjie Liu, Ziqin Gong, Yan Song, Yuxiang Chen, Xiaolong Liu, Hengtong Lu, Kaike Zhang, Chen Wei, and Jun Wang. The perceptual bandwidth bottleneck in vision-language models: Active visual reasoning via sequential experimental design. *arXiv preprint arXiv:2605.01345*, 2026. URL https://arxiv.org/abs/2605.01345.

Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. DePlot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10381–10399. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.660. URL https://aclanthology.org/2023.findings-acl.660/.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209, 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Mathew_DocVQA_A_Dataset_for_VQA_on_Document_Images_WACV_2021_paper.html.

Drew Prinster, Clara Fannjiang, Ji Won Park, Kyunghyun Cho, Anqi Liu, Suchi Saria, and Samuel Stanton. Conformal policy control. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306. PMLR, 2026. URL https://arxiv.org/abs/2603.02196.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894.

Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Chandu. Selective "selective prediction": Reducing unnecessary abstention in vision-language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12935–12948. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.767. URL https://aclanthology.org/2024.findings-acl.767/.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

Jian Xu, Yanning Wu, Delu Zeng, John Paisley, and Qibin Zhao. Look again before you abstain: Budgeted conformal evidence acquisition for reliable vision-language model. *arXiv preprint arXiv:2606.16667*, 2026. URL https://arxiv.org/abs/2606.16667.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/cdb347d7516d52a9280ea9d6708911f6-Abstract-Conference.html.