CERTIFY BEFORE YOU INTERVENE: PATHWISE INTERVENTION-CERTIFIED DECODING FOR MULTIMODAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) describe objects, attributes and findings that the image does not show. Inference-time corrections, such as retrieving evidence and regenerating part of the response, reduce these hallucinations on average, but they also turn correct responses into wrong ones and delete useful content. Whether a correction helps depends on the specific revision, which uncertainty and evidence relevance measure only indirectly. We introduce *Pathwise Intervention-Certified Decoding* (PICD), which revises a response only when a calibrated lower bound on the benefit of that revision is positive. A critic, trained on paired continuations that share an input and a prefix, predicts how a revision changes the hallucination loss and the utility of the completed response. PICD calibrates the largest overestimation of this critic along each response, so that its bounds hold for whichever candidate the decoder selects, and recalibrates every round on the states that earlier rounds produce. Under exchangeability, with a probability the operator chooses, every accepted revision on a response's path lowers hallucination loss and stays within a utility tolerance. Across three VLMs and five general-domain and medical benchmarks, the critic predicts which revisions lower hallucination loss better than uncertainty and relevance signals, and PICD keeps the share of responses with a violating revision below each requested rate we test, at the cost of revising fewer responses than an uncalibrated critic.

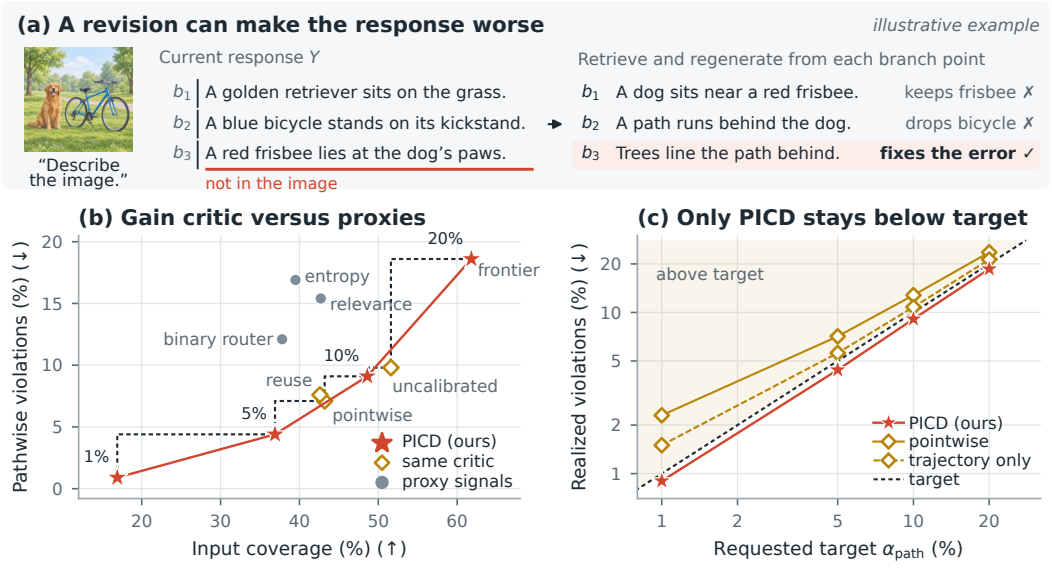

Figure 1: A revision can make a response worse, so PICD certifies its gain before revising and holds the requested violation rate. (a) Illustrative example: the response contains one unsupported claim, and regenerating from different branch points either keeps the error, drops correct content, or fixes only the error. (b) Pathwise violation rate against input coverage for the selectors of Section 4.2 at a 5% target, with PICD at targets of 1, 5, 10 and 20% and the Pareto frontier dashed. (c) Realized against requested violation rate for three calibrations of the same critic. Points are macro-averages over benchmark-backbone groups without intervals.

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.

1 INTRODUCTION

Vision-language models (VLMs) need to know when a correction of their own output will help, because a correction can make a response worse as well as better. VLMs state objects, attributes, relations and clinical findings that the image does not support (Rohrbach et al., 2018; Li et al., 2023; Guan et al., 2024; Sun et al., 2024). For example, a VLM that drafts a chest radiograph report can describe an effusion that the image does not show. Inference-time corrections such as contrastive decoding, visual re-grounding and retrieval reduce such errors on average (Leng et al., 2024; Huang et al., 2024b; Chuang et al., 2024; Lewis et al., 2020). However, language models that revise their own answers without external feedback turn correct answers into wrong ones (Huang et al., 2024a), and irrelevant retrieved context lowers the accuracy of retrieval-augmented models (Yoran et al., 2024). An *operator* who deploys a VLM therefore needs to know, before a revision is applied, whether it will (i) remove the unsupported claim and (ii) keep the content that was correct. Both questions concern the *terminal gain* of a revision, the change in hallucination loss and in utility between the revised response and the response it replaces.

The terminal gain is hard to know in advance, since it concerns a revised response that does not exist yet, and measuring it requires generating that response and scoring it against a reference. Existing methods therefore decide from signals of the current state. Uncertainty-triggered methods revise where the model is unsure (Zou et al., 2025), but an uncertain statement can be correct and a confident one wrong. Adaptive retrieval methods estimate whether evidence is needed or relevant (Asai et al., 2024; Jeong et al., 2024), but relevant evidence can still conflict with the image. Conformal methods bound the risk of an output set or of the final response (Quach et al., 2024; Mishra et al., 2026), but not the change that one revision causes. This mismatch is an opportunity, as a model of the terminal gain, trained offline where revised responses can be generated and scored, would let the decoder revise only where it helps and keep the response everywhere else.

Our key observation is that the terminal gain can be learned from pairs of continuations that share an input and a prefix, and that the error of this prediction can be calibrated. Offline, each candidate revision is generated and scored next to the original continuation, which labels whether it helps, and a critic learns to predict this label from what is known before revising. Across 48,216 candidate revisions, 34.8% would make the response worse, and the critic identifies the revisions that lower hallucination loss with an AUROC of 0.842, compared with 0.752 for the strongest current-state signal. We therefore propose *Pathwise Intervention-Certified Decoding* (PICD), which revises only where a calibrated lower bound on the terminal gain, its *certified gain*, is positive. PICD solves two problems of naive calibration. First, the decoder picks the best of several branch points, which favors those the critic overestimates, so PICD calibrates the largest error along each response. Second, an accepted revision changes the states that later rounds see, so PICD calibrates each round on fresh data replayed through the earlier rounds. Under exchangeability, this gives a finite-sample guarantee over the whole revision path, which the name *pathwise* refers to.

We evaluate PICD on `LLaVA-1.5-7B`, `Qwen2-VL-7B-Instruct` and `InternVL2.5-8B` across four general-domain hallucination benchmarks and radiology report generation on MIMIC-CXR. We find that PICD keeps the *pathwise violation rate*, the fraction of inputs whose revision path contains a revision that fails either constraint, below each of the four requested targets from 1% to 20%, with 4.4% at a 5% target, whereas calibrating each branch point separately exceeds all four targets, with 7.1% at 5%, as Figure 1c shows. Certified revisions also improve the final output, and on MIMIC-CXR the rate of unsupported clinical claims falls from 18.9% for the base decoder to 7.9%. Calibration trades coverage for this control, since without calibration 8.9% of the same critic's revisions are harmful rather than 2.8%, but it revises 51.6% of inputs rather than 36.9%. PICD adds 24.9% to the median latency, as it decodes only the selected revision.

Our contributions are the following.

1. We propose PICD, which certifies the terminal gain of every candidate revision with trajectory-level and round-matched calibration, and prove that all its accepted revisions satisfy both constraints with probability at least $1 - \alpha_{\text{path}}$ (Section 3).
2. We identify the gap between current-state signals and terminal gain, where seven such signals predict the outcome of a revision worse than a critic trained on paired continuations (Section 4.1).
3. We evaluate shared-critic controls, ablations and stress tests, under which the violation rate stays at or below 4.9% when retrieval is damaged but rises to 11.9% under external shift (Section 4.4).

2 Related Work

Inference-time hallucination mitigation. A first challenge is to reduce hallucination without retraining, since fine-tuning a VLM for every deployment is expensive. Contrastive decoding with VCD subtracts the predictions for a distorted image from those for the original (Leng et al., 2024), and OPERA penalizes attention that concentrates on a few summary tokens and rolls decoding back when it does (Huang et al., 2024b). MemVR re-injects visual tokens when uncertainty rises (Zou et al., 2025), VISTA steers hidden activations toward visual information (Li et al., 2025), and Self-Aug selects query-dependent image augmentations (Im et al., 2026). These methods decide when and how to correct from signals of the current state, such as uncertainty, attention or activations.

Adaptive retrieval. A second challenge is to decide when to use external evidence, since retrieved text can mislead as well as inform. Self-RAG learns reflection tokens that trigger retrieval and critique its use (Asai et al., 2024), and Adaptive-RAG routes each query to no, single-step or iterative retrieval by its estimated complexity (Jeong et al., 2024). D^2-RAG, proposed for text-only medical question answering, combines several uncertainty estimates to decide on retrieval and weights relevant and noisy contexts during decoding (Zhang et al., 2026). These methods estimate whether evidence is needed or useful, not the change that a revision based on it causes.

Conformal guarantees for generation. A third challenge is to give distribution-free guarantees for generated text, which conformal prediction gives under exchangeability. Conformal language modeling calibrates sampling and rejection to return a set that contains an acceptable response (Quach et al., 2024), and conformal factuality removes claims until the remainder is correct with high probability (Mohri & Hashimoto, 2024). C-RAG bounds the generation risk of retrieval-augmented models (Kang et al., 2024), and SAFER calibrates both the sampling budget and the filter over sampled answers (Wang et al., 2026a). For VLMs, Sample then Identify controls the risk of sampling and selecting a response (Wang et al., 2025), and CEBC calibrates visual-evidence thresholds to edit or suppress unsupported object mentions (Mishra et al., 2026). ITCR calibrates when to stop expanding a reasoning graph (Wang et al., 2026b). These methods bound the risk of an output or a set of outputs, not the change that one revision causes.

Conformal selection and policy control. The closest challenge is to choose actions with guarantees, since an adaptive choice among candidates invalidates per-candidate calibration. Multivariate conformal selection selects candidates under several criteria while controlling the false discovery rate (Bai et al., 2025), and conformal policy control uses data from a safe reference policy to limit how far an optimized policy may deviate (Prinster et al., 2026). Our work instead certifies the change that one revision causes relative to the response it replaces, and calibrates each round of a sequential policy on the states that earlier rounds create.

3 Pathwise Intervention-Certified Decoding

PICD revises a response only where a calibrated lower bound on the terminal gain of the revision is positive. It proceeds in three steps, which Figure 2 numbers. Step ① fits a critic that predicts the terminal gains from paired continuations (Section 3.2). Step ② calibrates the largest overestimation of this critic along each response, round by round (Section 3.3). Step ③ revises at the branch point with the best certified gain, or keeps the response (Section 3.4). Algorithm 1 in Section A gives the pseudocode for calibration and deployment.

3.1 Setting and Terminal Gain

Let x be an image, q a text query and π_θ a frozen autoregressive VLM that produces the response $Y^{(0)} = \mathrm{Decode}_{\pi_\theta}(x, q)$. For a response Y, a deterministic rule defines the candidate *branch points* $\mathcal{B}(Y)$, where each $b \in \mathcal{B}(Y)$ keeps the prefix $Y_{<b}$ and allows the rest to change. Short-form question answering uses the start of the response, and long-form tasks use sentence or report-section boundaries, as Section A.1 specifies. An *intervention* $\mathcal{I}$ at b retrieves evidence and regenerates the suffix,

$$E_b = \mathrm{Retrieve}(x, q, Y_{<b}), \qquad Y_b^{\mathcal{I}} = Y_{<b} \oplus \mathrm{Decode}_{\pi_\theta}\big(x, q, Y_{<b}, c(E_b)\big), \tag{1}$$

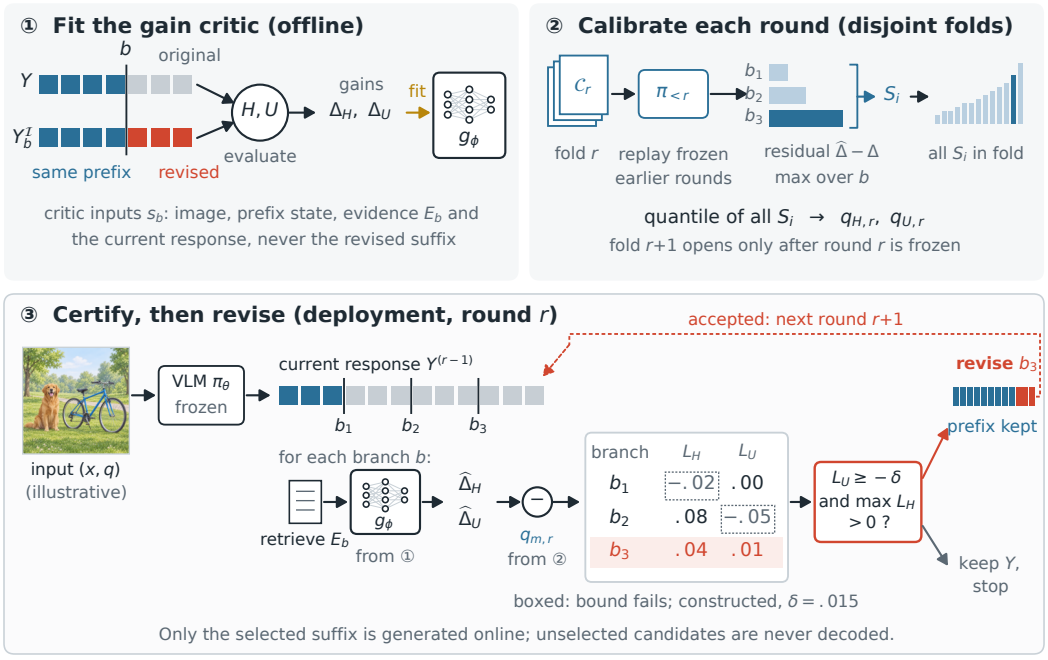

Figure 2: PICD: **(1)** fit the gain critic on paired continuations; **(2)** calibrate trajectory-maximum residuals on disjoint, policy-matched folds; **(3)** select the best branch satisfying $L_H > 0$ and $L_U \geq -\delta_r$, or stop. Only the accepted suffix is generated online. The branch values are constructed examples, and the input image is illustrative rather than a benchmark sample.

where $c(E_b)$ serializes the evidence and $\oplus$ appends the new suffix to the kept prefix. The retriever, serializer, decoder and generation settings are fixed before calibration, so PICD certifies the complete retrieval-and-regeneration action rather than individual evidence items.

Each benchmark d has a hallucination loss $H_d(x, q, Y) \in [0, 1]$, where lower is better, and a task utility $U_d(x, q, Y) \in [0, 1]$, where higher is better, both fixed before calibration and defined in Section C.2. Omitting d, x and q, the terminal gains of an intervention at b are

$$\Delta_H(Y, b) = H(Y) - H\big(Y_b^{\mathcal{I}}\big), \qquad \Delta_U(Y, b) = U\big(Y_b^{\mathcal{I}}\big) - U(Y). \tag{2}$$

The hallucination gain Δ_H is positive when the revision lowers hallucination loss. The utility gain Δ_U is negative when the revision loses useful content. Given a tolerance $\delta \geq 0$, an intervention is *acceptable* if $\Delta_H(Y, b) > 0$ and $\Delta_U(Y, b) \geq -\delta$, and *harmful* if $\Delta_H(Y, b) < 0$ or $\Delta_U(Y, b) < -\delta$.

3.2 Learning the Terminal Gain

The terminal gains depend on a revised response that does not exist when the decoder decides, so PICD learns to predict them offline. For every fitting input and branch point, PICD generates the original and the revised continuation from the same prefix and scores both with Equation (2). A critic g_ϕ predicts both gains, $(\widehat{\Delta}_H(Y, b), \widehat{\Delta}_U(Y, b)) = g_\phi(s_b)$. Its input s_b contains the decoder state at the branch point, the visual representation, the retrieved evidence and the current response, which are all available before the revision is decoded. The revised continuations serve only as training labels and never as critic inputs. The critic is an MLP with two hidden layers trained with a Huber loss on both gains and an auxiliary loss on the sign of Δ_H, and it is frozen before calibration (Section A.3).

3.3 Certifying the Terminal Gain

The critic only estimates the gains, and choosing the largest of several estimates favors branch points whose gain it overestimates. PICD therefore calibrates the largest overestimation along each response. For objective $m \in \{H, U\}$ and calibration response i, the score and the resulting lower

bound are

$$S_i^m = \max_{b \in \mathcal{B}(Y_i)} \big[\widehat{\Delta}_m(Y_i, b) - \Delta_m(Y_i, b)\big], \qquad L_m(Y, b) = \widehat{\Delta}_m(Y, b) - q_m, \tag{3}$$

where q_m is the split-conformal quantile of $\{S_i^m\}_{i=1}^n$ at level $1 - \alpha_m$. The score S_i^m is the worst overestimation over all branch points of one response, and the margin q_m exceeds the score of a new response with probability at least $1 - \alpha_m$. On that event every branch point satisfies $\Delta_m \geq L_m$ at once, so the bound also holds for the branch point that the decoder selects. We call L_H and L_U the *certified gains*, and Section A.4 gives the finite-sample conventions.

An accepted intervention changes the prefixes, evidence and candidates that later rounds see, so a margin calibrated on unrevised responses is not guaranteed to hold in round two. PICD allows at most R interventions and splits the calibration data into disjoint folds $\mathcal{C}_1, \ldots, \mathcal{C}_R$. Before round r is calibrated, the policies frozen in rounds $1, \ldots, r-1$ are applied to every input of $\mathcal{C}_r$, and the scores of Equation (3) are computed on the responses they produce. This yields the round-specific margins $q_{H,r}$ and $q_{U,r}$, and the round-r policy is frozen before $\mathcal{C}_{r+1}$ is opened. Figure 6 in the appendix works through both steps with numbers, including a case where reusing the first-round margin fails.

3.4 Decision Rule

In round r, PICD keeps the branch points whose certified utility gain is within the tolerance δ_r and selects the one with the largest certified hallucination gain,

$$\mathcal{F}_r(Y) = \big\{b \in \mathcal{B}(Y) : L_U(Y, b) \geq -\delta_r\big\}, \qquad b^\star = \arg\max_{b \in \mathcal{F}_r(Y)} L_H(Y, b), \tag{4}$$

and it revises at $b^\star$ if and only if $L_H(Y, b^\star) > 0$. If $\mathcal{F}_r(Y)$ is empty or the best certified hallucination gain is not positive, PICD keeps the current response and stops. Otherwise it decodes only the selected suffix, so the revision path is $Y^{(0)} \to Y^{(1)} \to \cdots \to Y^{(R)}$, and unused rounds leave the response unchanged. Uncertainty, retrieval relevance and output risk may enter the critic as features, but none of them triggers an intervention on its own.

3.5 Pathwise Guarantee

Theorem 1 (Pathwise intervention validity)**.** *Assume that, conditional on the fitting and validation data and the earlier calibration folds, the inputs of each current fold and a fresh test input are exchangeable and the pipeline that produces the scores is fixed, with disjoint policy-matched folds and matching calibration and deployment procedures as specified in Section B.1. If the prespecified error levels satisfy*

$$\textstyle\sum_{r=1}^{R} (\alpha_{H,r} + \alpha_{U,r}) \leq \alpha_{\mathrm{path}}, \tag{5}$$

then, with probability at least $1 - \alpha_{\mathrm{path}}$ *over the calibration data and the test input, every accepted revision satisfies*

$$H\big(Y^{(r)}\big) < H\big(Y^{(r-1)}\big), \qquad U\big(Y^{(r)}\big) \geq U\big(Y^{(r-1)}\big) - \delta_r, \tag{6}$$

where $Y^{(r)} = Y^{(r-1)}$ *for rounds without a revision. On the same event,* $H(Y^{(R)}) \leq H(Y^{(0)})$ *and* $U(Y^{(R)}) \geq U(Y^{(0)}) - \sum_{r=1}^{R} \delta_r$.

Proof sketch. Conditional on the data used before round r, the earlier policies are a fixed map from an input to its current response, and applying one fixed map to exchangeable inputs keeps them exchangeable. The maximum in Equation (3) then makes split-conformal coverage hold for all round-r branch points at once, so any accepted revision meets both true constraints on this event. A union bound over the two objectives and the R rounds gives Theorem 1 without assuming independence across rounds, and Section B gives the complete proof.

Scope. The guarantee is marginal over the calibration data and a fresh input, and it is relative to the fixed evaluators H and U. It does not establish factual correctness, correctness of the evaluators, coverage for each subgroup, or validity under external distribution shift.

4 Experiments

We ask whether candidate revisions differ in their terminal effect and whether current-state signals predict it (Section 4.1), whether calibration holds the requested violation rate (Section 4.2), whether

certified revisions improve the final output (Section 4.3), which components matter when retrieval, calibration data or the tolerance change (Section 4.4), and where PICD still fails (Section 4.5).

Models and benchmarks. We use three frozen VLMs, `LLaVA-1.5-7B`, `Qwen2-VL-7B-Instruct` and `InternVL2.5-8B`, with greedy decoding. The general-domain benchmarks are POPE (Li et al., 2023), CHAIR (Rohrbach et al., 2018), MMHal-Bench (Sun et al., 2024) and HallusionBench (Guan et al., 2024), and the medical task is report generation on 3,269 held-out MIMIC-CXR studies (Johnson et al., 2019). PICD allows one revision for short-form question answering and two for long-form generation, with a total utility tolerance of $\delta_{\text{tot}} = 0.03$ and $\alpha_{\text{path}} = 0.05$ unless stated otherwise.

Retrieval and partitions. Critic fitting, validation, calibration folds, retrieval index and test data are disjoint, and all medical partitions separate patients. Test examples, their annotations and near duplicates are excluded from the retrieval index.

Baselines. Retrieval baselines are always-retrieve RAG, a cross-encoder reranker followed by revision, entropy-gated retrieval, a router adapted from D^2-RAG with the same capacity as our critic, and TCD, which triggers retrieval by entropy and filters evidence with a conformal test on embedding density. Decoding baselines are VISTA and CEBC, run with their official implementations. The shared-critic controls use the same critic without calibration, with pointwise calibration of each branch point, with trajectory-level calibration that is not matched to later rounds, and with the first-round margins reused in every round, which isolates each calibration choice.

Metrics. Input coverage is the fraction of test inputs that receive at least one revision. The harmful-intervention rate (HIR) is the fraction of accepted revisions with $\Delta_H < 0$ or $\Delta_U < -\delta_r$. The pathwise violation rate (PVR), which Theorem 1 bounds, is the fraction of test inputs whose path contains any accepted revision with $\Delta_H \leq 0$ or $\Delta_U < -\delta_r$. Rates are computed within each benchmark-backbone group and then macro-averaged, and Section C.3 gives the formulas. All values are point estimates, and Section C gives all settings.

4.1 Terminal Effects and Their Prediction

To test whether revisions differ in their effect, we generate and score every candidate revision offline, whether or not PICD would accept it. Figure 3a gives the share of acceptable, neutral and harmful outcomes, and Figure 3b the AUROC of each predictor for a positive hallucination gain, where a point further right separates helpful revisions better. **More than a third of candidate revisions are harmful.** Of 48,216 candidates, 43.7% are acceptable, 21.5% are neutral because they leave hallucination loss unchanged, and 34.8% would make the response worse if applied without selection. **Current-state signals predict the terminal gain poorly.** The critic reaches an AUROC of 0.842, and the best of seven current-state signals, a learned binary utility router, reaches 0.752. Predictive entropy and retrieval similarity reach 0.662 and 0.691, so uncertainty and evidence relevance carry less information about the outcome of a revision than paired supervision does.

4.2 Holding the Requested Violation Rate

The central question is whether the certified gain keeps harmful revisions below the rate the operator requests. Figure 1b plots each selector's PVR against its coverage, where points lower and further right are better, and Figure 1c plots the realized against the requested PVR, where points below the diagonal hold the target. **Only PICD stays below all four targets.** PICD reaches 0.9, 4.4, 9.1 and 18.6% at targets of 1, 5, 10 and 20%. Pointwise calibration ignores that the decoder selects among branch points and exceeds all four targets, and trajectory-level calibration without round matching exceeds them as well, with 5.6% at 5%. Reusing the first-round margins in round two reaches 7.6% at the 5% target. **The critic dominates the proxies, and calibration moves along its curve.** Entropy, retrieval relevance and the binary router lie above the frontier at every coverage, whereas all variants of the gain critic lie on or next to it. The uncalibrated critic reaches 9.8% at 51.6% coverage, close to PICD at the 10% target, so calibration contributes control over the risk level rather than a better trade-off at matched coverage. Figure 4 gives the exact operating points, including the constraint-violation rate (CVR), which counts accepted revisions that fail either constraint.

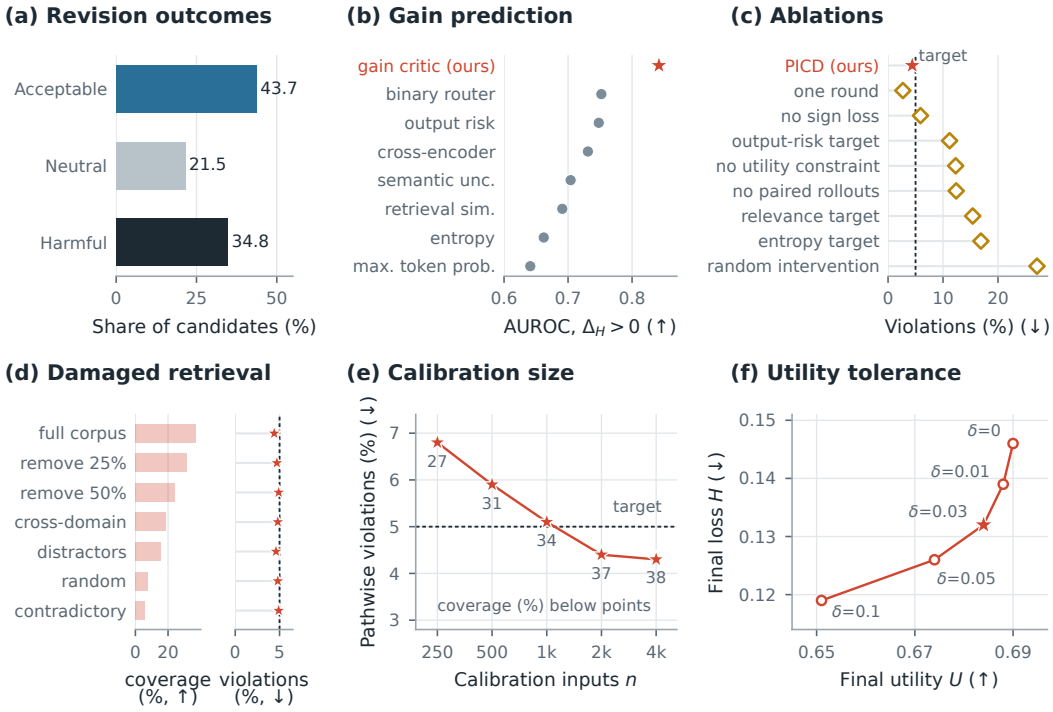

Figure 3: Current-state signals miss the terminal gain, ablations raise violations, and damaged retrieval mainly lowers coverage. (a) Outcomes of all 48,216 candidate revisions. (b) AUROC for predicting $\Delta_H > 0$. (c) PVR of ablations at a 5% target (diamonds) and PICD (star). (d) Coverage and PVR under seven retrieval conditions, which are distinct and not ordered by severity. (e) PVR against the number of calibration inputs n, with coverage in % below each point. (f) Final hallucination loss against utility for total tolerances δ_{tot} from 0 to 0.10, with the main setting starred. Colors and markers follow Figure 1, and all points are macro-averages without intervals.

4.3 Final Output Quality

We next ask whether revisions that pass the certificate improve the final response, which Figure 5 reports per benchmark. **PICD improves on the strongest external baseline.** Compared with CEBC, PICD raises POPE F1 from 92.1 to 93.2 and the MMHal score from 3.31 to 3.56, and lowers CHAIR_S from 20.7 to 15.9. On MIMIC-CXR, the unsupported-claim rate falls from 18.9% for the base decoder to 7.9%, while RadGraph, CheXbert and utility rise. **Calibration costs little output quality.** The uncalibrated critic applies more revisions and reaches similar aggregate scores, but 8.9% of its revisions are harmful compared with 2.8% for PICD. Per backbone, PICD has the lowest HIR everywhere, but on `InternVL2.5-8B` the uncalibrated critic ties on CHAIR_S and is ahead on CHAIR_I and HallusionBench, as Section D shows. In a blinded study of 300 output pairs, annotators prefer PICD over the base decoder in 78.0% of pairs, compared with 65.0% for CEBC over the same decoder, and rate its factuality at 4.31 versus 4.08 for CEBC.

4.4 Components, Robustness and Cost

Finally, we ask which components the violation rate depends on and how PICD behaves when its inputs degrade. Figure 3c gives the PVR of each ablation at the 5% target, Figure 3d coverage and PVR under damaged retrieval, Figure 3e PVR against calibration size, and Figure 3f the loss-utility trade-off over the tolerance. **The target and the supervision both matter.** Training the critic on entropy instead of the terminal gain raises PVR from 4.4 to 16.9%, and training it without paired continuations raises it to 12.4%. Dropping the utility constraint lowers hallucination loss from 0.132 to 0.118 but lowers utility from 0.684 to 0.641 and raises HIR to 11.6%, so a lower loss alone does not make a better policy. A single round reaches 2.7% PVR at 29.4% coverage, and the two

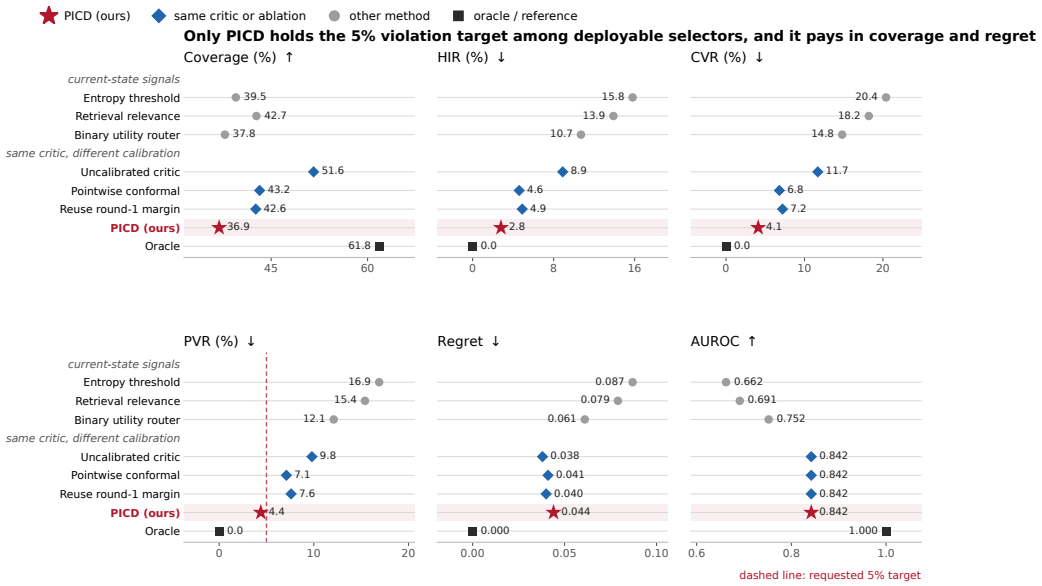

Figure 4: Only PICD stays below the 5% target, and it pays with coverage and regret. Operating points of all selectors at $\alpha_{\text{path}} = 0.05$, the numbers behind Figure 1b. Each column places the rows between their minimum and maximum, with dots further right being better, and prints the exact value. Coverage and PVR count inputs, HIR and CVR count accepted revisions, the vertical tick marks the 5% target, and regret is measured against the oracle, which sees the realized gains and cannot be deployed. Rates are macro-averaged.

configurations also differ in hallucination loss (0.143 versus 0.132). **Damaged retrieval mainly lowers coverage.** Removing half of the relevant items lowers coverage from 36.9 to 24.5%, and contradictory evidence lowers it to 5.9% while HIR rises from 2.8 to 5.1%, and PVR stays between 4.4 and 4.9% under all seven conditions. Changing retrieval at test time breaks the matched-pipeline assumption, so these runs measure selectivity rather than the guarantee. **Small calibration sets exceed the target in these runs.** With 2,000 or more calibration inputs PVR stays below 5%, whereas 250, 500 and 1,000 inputs give 6.8, 5.9 and 5.1%, which the marginal guarantee allows for a single realized calibration set. The tolerance sets the trade-off, since raising δ_{tot} from 0 to 0.10 lowers hallucination loss from 0.146 to 0.119 and utility from 0.690 to 0.651. **Cost.** PICD raises median latency from 410 to 512 ms per example on an A100 GPU, because the critic scores every branch point but only the selected suffix is decoded. Offline, paired continuations cost 42 GPU hours per backbone and calibration 31 GPU hours.

4.5 Remaining Failures

To see where accepted revisions still fail, we assign each analyzed residual failure of an accepted revision one primary source. **Most residual failures come from the critic, the evaluator and retrieval.** The critic overestimates the gain in 31% of these cases, the automatic evaluator and human judgment disagree in 26%, retrieval is missing or contradictory in 23%, the branch point lies after the unsupported claim in 12%, and regeneration of the suffix is unstable in 8%. These shares describe the analyzed failures rather than independent risks, and evaluator disagreement does not by itself establish harm under H and U. On MIMIC-CXR, PICD is lowest for all six findings, yet 10.1% of consolidation and 11.7% of pneumothorax statements remain unsupported (Figure 8).

5 Discussion and Conclusion

Risk control and coverage. At the 5% target, PICD revises 36.9% of inputs, whereas the uncalibrated critic revises 51.6% and the outcome-informed oracle 61.8%. Calibration sets where on the critic's risk-coverage curve the operator sits, so improving the critic, rather than the calibration, would move the curve toward the oracle, which we leave to future work. **Evaluator-relative guar-**

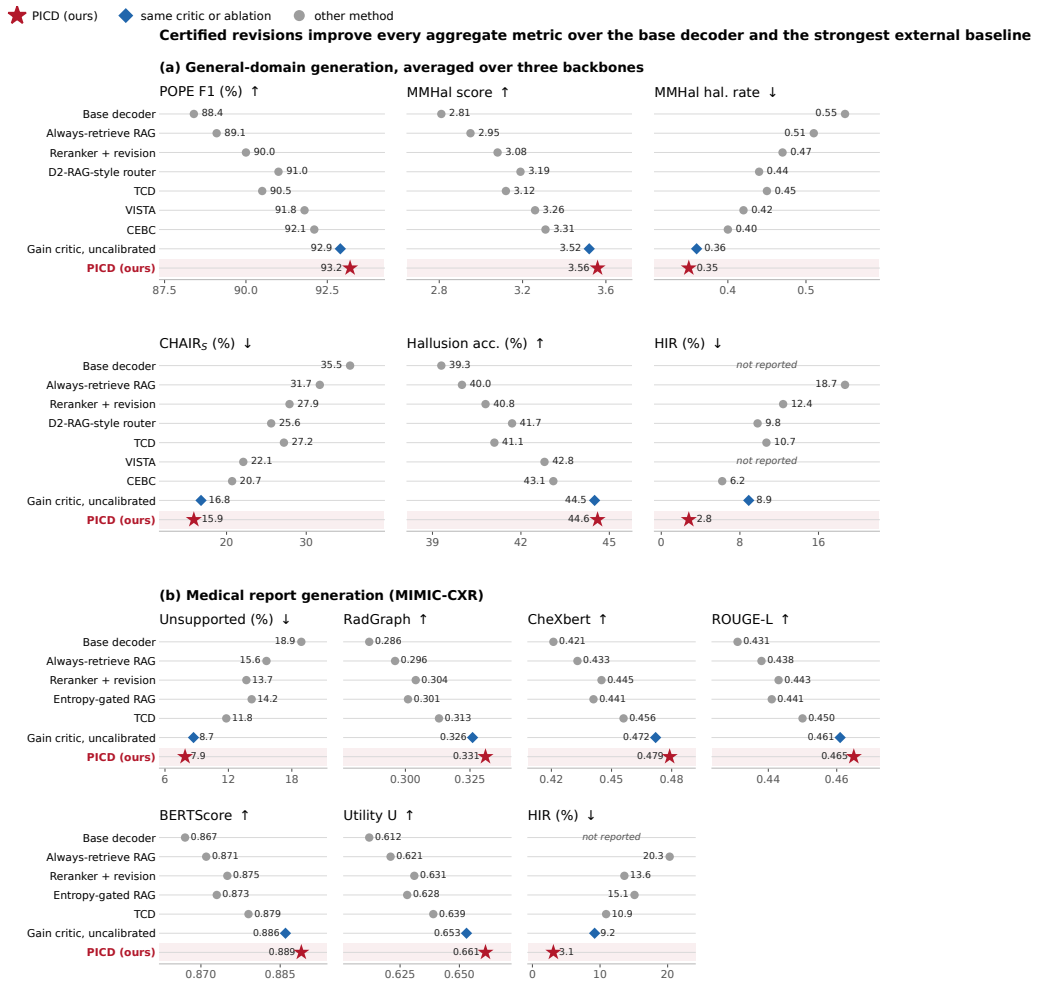

Figure 5: Certified revisions improve every aggregate metric over the base decoder and the strongest external baseline. (a) General-domain results averaged over the three backbones. (b) MIMIC-CXR report generation. POPE F1, CHAIR$_S$, HallusionBench accuracy, the unsupported-claim rate and HIR are percentages, and the MMHal hallucination rate is a fraction. Dots further right are better in every column, bold marks the best value, and a dash marks a metric that is not reported.

antees. The certificate holds with respect to the functionals H and U, and for MIMIC-CXR the loss counts clinical statements that do not match the reference report rather than checking the radiograph. Human raters agree with each other at Krippendorff's α between 0.69 and 0.74, and a natural extension calibrates against human judgments of revisions. **Beyond exchangeability.** Round matching handles the shift that PICD's own revisions cause, but not shift in the deployment data, under which PVR rises from 4.4 to 11.9%. Rare concepts also receive less coverage than frequent ones (28.4 versus 39.8%), and weighted conformal methods or shift detection are natural next steps for both. **Richer interventions.** We certify one family of interventions, retrieval followed by regeneration of the suffix. The critic and the certificate apply to any fixed action whose effect can be generated and scored offline, such as re-grounding on image regions, which we leave to future work.

Conclusion. We introduce PICD, which revises a VLM's response only where a calibrated lower bound on the terminal gain of the revision is positive. Across three VLMs and five benchmarks, PICD stays below all four requested violation rates we test and lowers unsupported claims in radiology reports from 18.9 to 7.9%. We encourage evaluating inference-time corrections by the change they cause in the completed response, not by the risk of the response they start from.

Ethics Statement

This work studies inference-time methods that reduce unsupported content in vision-language generation. The certificate is not a probability that the final response is correct, since it bounds the effect of a specified intervention with respect to fixed evaluators, and errors outside these evaluators, distribution shift, or failures of the model and the retriever may remain. All datasets were used under their licenses, data-use agreements and institutional requirements. Test examples and their annotations were excluded from the retrieval indices, medical partitions separate patients, and retrieved evidence keeps its source identifiers for auditing. No patient data or identifying information are released.

The expert reader study was conducted under `[INSTITUTIONAL APPROVAL OR EXEMPTION IDENTIFIER]`. Participants were informed of the procedure, outputs were shown in randomized and blinded order, and no generated report was used for patient care. The medical experiments evaluate the method in a safety-critical setting and are not evidence that it is ready for autonomous clinical use. Deployment in such settings would require prospective validation, monitoring for shift, assessment per subgroup and review by experts. Retrieval corpora and automatic evaluators may encode historical or demographic biases, and because rare concepts receive less coverage, the absence of a certified revision is not evidence that a response is safe.

Use of Large Language Models

Generative AI tools were used for prose editing and manuscript organization. They were not used to execute experiments, access raw data or model outputs, or generate, modify, estimate, or interpolate any reported numerical result. The authors independently verified the mathematical derivations, implementation, citations, experiments, and analysis, and claims, and take full responsibility for the entire content of the paper.

Reproducibility Statement

Section A specifies the branch-point rule, the intervention, the critic, the calibration and the decision procedure, and Section B gives all assumptions and proofs. Section C lists the benchmarks, models, partitions, retrieval indices, evaluators, metrics and baseline settings, and Section D reports every number behind the figures. The figure sources and the values behind every plot are provided with the supplementary material.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Tian Bai, Yue Zhao, Xiang Yu, and Archer Y. Yang. Multivariate conformal selection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 2535–2559. PMLR, 2025. URL https://proceedings.mlr.press/v267/bai25d.html.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Th6NyL07na.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_CVPR_2024_paper.html.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=IkmD3fKBPQ.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, June 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_OPERA_Alleviating_Hallucination_in_Multi-Modal_Large_Language_Models_via_Over-Trust_CVPR_2024_paper.html.

Eun Woo Im, Muhammad Kashif Ali, and Vivek Gupta. Self-Aug: Query and entropy adaptive decoding for large vision-language models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Z1zHgGGG3B.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, and Pranav Rajpurkar. RadGraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/c8ffe9a587b126f152ed3d89a146b445-Abstract-round1.html.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.389. URL https://aclanthology.org/2024.naacl-long.389/.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6: 317, 2019. doi: 10.1038/s41597-019-0322-0. URL https://doi.org/10.1038/s41597-019-0322-0.

Mintong Kang, Nezihe Merve Gürel, Ning Yu, Dawn Song, and Bo Li. C-RAG: Certified generation risks for retrieval-augmented language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22963–23000. PMLR, 2024. URL https://proceedings.mlr.press/v235/kang24a.html.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and

Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35799–35819. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25ca.html.

Ashish Mishra, Tarun Kumar, Arpit Shah, Suparna Bhattacharya, and Martin Foltin. CEBC: Conformal evidence-bounded control for low-hallucination vision–language generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46193–46206, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.2142. URL https://aclanthology.org/2026.acl-long.2142/.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36029–36047. PMLR, 2024. URL https://proceedings.mlr.press/v235/mohri24a.html.

Drew Prinster, Clara Fannjiang, Ji Won Park, Kyunghyun Cho, Anqi Liu, Suchi Saria, and Samuel Don Stanton. Conformal policy control. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=rbwy4E0Kyz.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pzUhfQ74c5.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 1500–1519. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.117. URL https://aclanthology.org/2020.emnlp-main.117/.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.775. URL https://aclanthology.org/2024.findings-acl.775/.

Qingni Wang, Tiantian Geng, Zhiyuan Wang, Teng Wang, Bo Fu, and Feng Zheng. Sample then identify: A general framework for risk control and assessment in multimodal large language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9WYMDgxDac.

Qingni Wang, Yue Fan, and Xin Eric Wang. SAFER: Risk-constrained sample-then-filter in large language models. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/582e9771ac8527cb6390e5e9444a0fee-Paper-Conference.pdf.

Ting Wang, Yuanjie Shi, Yan Yan, and Huan Zhang. Inference-time conformal reasoning with valid factuality control for large language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026b. doi: 10.48550/arXiv.2606.08831. URL https://arxiv.org/abs/2606.08831.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ZS4m74kZpH.

Jinshuo Zhang, Xiaoding Zhou, Weiyu Zhang, Guoqiang Chen, Ying Lian, Xiaoyang Meng, Yonghe Chen, Hongjiao Guan, Jiasheng Si, and Wenpeng Lu. D^2-RAG: Dual-decision retrieval-augmented generation via multi-dimensional uncertainty and utility-aware decoding. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15355–15382, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.754. URL https://aclanthology.org/2026.findings-acl.754/.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and Xuming Hu. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80873–80899. PMLR, 2025. URL https://proceedings.mlr.press/v267/zou25e.html.