STILL CORRECT, ALREADY AT RISK: PRE-ONSET HALLUCINATION FORECASTING IN VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing approaches primarily predict hallucination risk in vision–language models (VLMs) before generation or detect local errors during decoding. We study the forecasting opportunity between these settings: predicting future hallucinations after generation has begun but while the output is still visually supported. We formalize this task as **Pre-Onset Hallucination Forecasting**: continuously predicting whether the remaining response will contain a hallucination using only currently available information. We find that still-supported generation provides predictive information beyond the generation-start state, enabling warnings several tokens before the first error, with gains that vary non-monotonically with the cumulative monitoring horizon. To study this predictive ability, we construct **PREHALLU**, comprising approximately 85K generation trajectories and more than 180K unique continuations sampled from supported prefixes, for studying hallucination onset and alternative futures. We introduce **RISKTRACE**, which combines current multimodal evidence with the observed history for continuous response-level early warning. Dynamic monitoring outperforms generation-start prediction on InstructBLIP, LLaVA-NeXT, and Qwen3-VL; on InstructBLIP, monitoring with the same frozen scorer raises response-level AUROC from 75.43% to 83.58%. Matched-false-alarm comparisons and lead-time analyses further show that pre-onset forecasting improves future-hallucination prediction and provides an opportunity for verification and intervention before the error appears.

1 INTRODUCTION

Vision–language models (VLMs) are advancing rapidly and are now widely used for visual understanding, instruction following, and open-ended generation (Dai et al., 2023; Li et al., 2024b;a; Bai et al., 2025b;a). Yet stronger capabilities do not guarantee faithful grounding in the input image: a model can still introduce unsupported objects, attributes, or relations (Li et al., 2023; Gunjal et al., 2024). As illustrated in Figure 1, a model may first describe a couple cutting a cake correctly and only later introduce an unsupported attribute, such as calling the woman's dress "yellow." Before this first error appears, the generated prefix remains supported by visual evidence. This interval offers an opportunity to warn before unsupported content reaches the user. We therefore ask whether the generation trajectory reveals future hallucination risk while the emitted prefix is still supported.

Two closely related lines of work estimate response-level risk before generation (Kogilathota et al., 2026) or detect local errors during decoding (Park et al., 2025; Duan et al., 2025); early-window methods also use the first few generated tokens (Zollicoffer et al., 2025). Our question concerns a different prediction target and timing: while generation is still supported, can the observed prefix predict a hallucination that may occur several tokens later? We introduce *Pre-Onset Hallucination Forecasting* to study this question. The predictor updates as the response unfolds, uses only information available at the current prefix, and forecasts whether the remaining response will contain a hallucination. A warning succeeds only if it precedes the first unsupported token.

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.

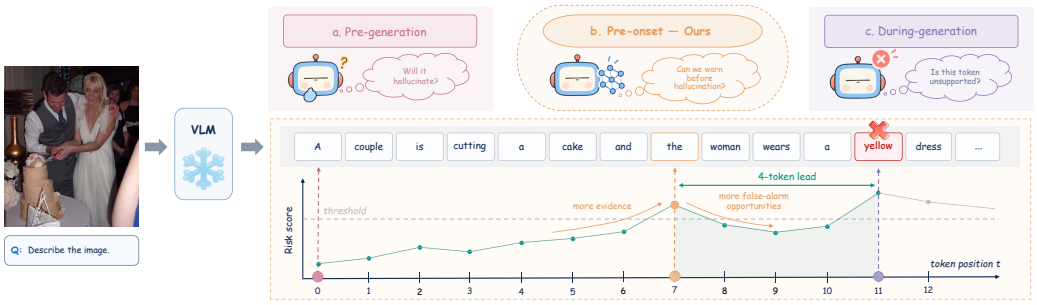

Figure 1: **Forecasting hallucination before its first occurrence.** (a) Pre-generation prediction estimates response-level risk from the initial state. (b) Pre-onset forecasting (ours) exploits an evolving, still-supported prefix to predict hallucination later in the response. (c) Local decoding-time detection assesses whether the token at the current position is unsupported. The curve shows per-prefix scores before cumulative aggregation; each state precedes the token above it. The example warns four supported tokens before the first unsupported token, "yellow."

We find that still-supported generation provides predictive evidence beyond the initial state. With the scorer held fixed, later prefixes improve discrimination between responses that will and will not hallucinate. The gain is not uniform across monitoring horizons: discrimination initially decreases before improving as monitoring extends. Thus, forecasting benefits from evidence revealed during generation, but simply observing more prefixes does not guarantee better predictions.

Repeated monitoring creates additional opportunities for both early warning and false alarms. We therefore interpret current evidence in its historical context and assess forecasting gains at matched response-level false-alarm costs.

To study this setting, we construct PREHALLU, comprising approximately 85K VLM generation trajectories and more than 180K unique continuations sampled from still-supported prefixes. Onset-annotated trajectories define the pre-onset evaluation interval, while prefix-conditioned continuations expose multiple possible futures from the same observed state. We further introduce RISKTRACE, a trajectory-based forecaster that combines current evidence with the observed history and selectively strengthens the base risk score before aggregating predictions into an irreversible response-level alarm.

On InstructBLIP/Vicuna-7B, pre-onset monitoring improves response-level AUROC from 75.43% to 83.58% with the scorer held fixed. At a matched clean-response false-alarm rate of approximately 5%, monitoring identifies 993 additional hallucinating responses. At this operating point, successful warnings leave a median of 12 supported tokens before onset, demonstrating prediction beyond the immediately upcoming token. Consistent gains on LLaVA-NeXT and Qwen3-VL extend this finding across VLM backbones and evaluation sets. Alarm-triggered visual grounding further reduces object hallucination relative to reject-and-resample on two held-out Qwen3-VL sets.

Our contributions are threefold:

- **Pre-onset forecasting and its predictive value.** We introduce Pre-Onset Hallucination Forecasting and show that future hallucination can become more predictable during still-supported generation, with gains that depend non-monotonically on the cumulative monitoring horizon.
- **A dataset for pre-onset forecasting.** We construct PREHALLU, containing approximately 85K VLM generation trajectories and more than 180K unique prefix-conditioned continuations, enabling controlled study of future events from observed prefixes.
- **A trajectory-based monitor and empirical validation.** We develop RISKTRACE to combine current and historical evidence for response-level early warning, and validate its forecasting value through controlled comparisons, lead-time analysis, and cross-model experiments.

2 Related Work

Risk estimation and decoding-time detection. Model-internal representations support both response-level and local risk estimation. HALP uses visual and multimodal representations before the first response token, whereas MTRE aggregates early-token evidence for response-level prediction (Kogilathota et al., 2026; Zollicoffer et al., 2025). TruthPrInt uses the preceding hidden state to predict hallucination in an upcoming object token (Duan et al., 2025). HalLoc, VADE, and UniProbe examine token-level signals through localization, visual attention, or multiple internal representations (Park et al., 2025; Prabhakaran et al., 2025; Samuel et al., 2026). The distinction is not whether a detector runs during decoding: repeatedly estimating local token risk does not by itself specify a forecast of a later response-level event. Our evaluation tests whether these evolving observations anticipate the first hallucination, and charges each clean response for any alarm across its full trajectory. This compares forecasting performance at a common response-level false-alarm cost.

Hallucination evaluation and supervision. VLM hallucination has been studied at several granularities, including unsupported objects, attributes, relations, and claims (Liu et al., 2024; Bai et al., 2024). Object-level evaluations test unsupported mentions in captions or object presence through discriminative questions (Rohrbach et al., 2018; Li et al., 2023; Lovenia et al., 2024). Broader benchmarks and evaluators assess visual faithfulness across generative and discriminative tasks (Wang et al., 2023a; Sun et al., 2024; Wang et al., 2023b), while diagnostic datasets probe visual illusions, misleading context, and visual perturbations (Guan et al., 2024; Wu et al., 2024b; Liu et al., 2025). Span annotations and atomic-fact evaluation provide more localized supervision (Gunjal et al., 2024; Jing et al., 2024). These resources establish what counts as unsupported content at different granularities. For forecasting, we also need to know when that content first appears and which futures can follow the same supported prefix. PreHallu pairs onset-annotated trajectories with prefix-conditioned continuations to support these temporal questions.

Hallucination mitigation and correction. Inference-time approaches improve visual grounding through contrastive decoding, attention-guided adjustments, and generation control (Leng et al., 2024; Huang et al., 2024; Chen et al., 2024; Wang et al., 2024; Yang et al., 2024; Woo et al., 2024; Gong et al., 2024; Yue et al., 2024). Training-based approaches use hallucination-aware supervision, corrective feedback, and preference optimization to improve response faithfulness (Jiang et al., 2024; Zhou et al., 2023; Sun et al., 2024; Yu et al., 2024; Xie et al., 2024), while post-generation systems verify or revise completed responses (Yin et al., 2024; Wu et al., 2024a). Forecasting is complementary to these mechanisms: a warning can determine when to invoke verification or correction rather than prescribing how to modify the generated content. Our alarm-triggered grounding experiments examine this connection.

3 Pre-Onset Forecasting: Task and Data

We define pre-onset forecasting and response-level evaluation, then introduce PreHallu; Section 4 presents the monitor.

3.1 Task formulation

Let I denote an image, q a prompt, and $x_{1:T}$ a generated response. At time t, the visible prefix is $x_{1:t}$ and the decoder predicts the next token; $t = 0$ precedes the first response token. Define

$$\tau = \min\{j : x_j \text{ belongs to an annotated unsupported span}\}, \qquad Y = \mathbb{I}\{\tau < \infty\}. \tag{1}$$

Responses with no annotated hallucination have $\tau = \infty$; supported prefixes satisfy $t < \tau$.

Let $\mathcal{F}_t$ contain the image, prompt, emitted prefix, and current decoder features. The predictor outputs $r_t = f(\mathcal{F}_t)$ to score future hallucination from a supported prefix. Onset annotations define the evaluation interval; the monitor uses only the observed prefix and its decoder features, conditioned on the image and prompt.

Let $\mathcal{T}_i$ contain the monitored prefix times for response i, including $t = 0$. Its eligible times and response score are

$$\mathcal{L}_i = \begin{cases} \{t \in \mathcal{T}_i : t < \tau_i\}, & Y_i = 1, \\ \mathcal{T}_i, & Y_i = 0, \end{cases} \qquad R_i = \max_{t \in \mathcal{L}_i} r_{i,t}. \tag{2}$$

Clean responses are monitored until natural termination. At threshold θ, pre-onset recall and the clean-response false-alarm rate are

$$\text{Recall}_{\text{pre}}(\theta) = \Pr(R_i \geq \theta \mid Y_i = 1), \tag{3}$$

$$\text{FAR}_{\text{clean}}(\theta) = \Pr(R_i \geq \theta \mid Y_i = 0). \tag{4}$$

We report AUROC for R_i and recall at matched empirical clean FAR, using each method's evaluation ROC. Thresholds frozen on development data and their transfer performance are reported separately.

3.2 PREHALLU: CONSTRUCTION AND ANNOTATION

PREHALLU provides two complementary units: onset-annotated trajectories delimit the pre-onset prediction interval, and prefix-conditioned continuations expose alternative futures from the same supported state. Figure 2 illustrates continuation construction and summarizes the inventory.

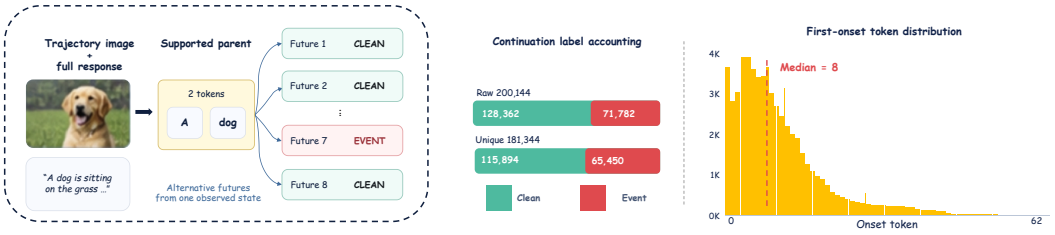

Figure 2: **PREHALLU connects supported prefixes to alternative futures.** Left (schematic): with the image, prompt, and supported prefix fixed, multiple continuations are sampled and labeled as clean or hallucinating, enabling the study of future risk from the same observed state. Center: 200,144 raw continuation records yield 181,344 unique continuations after response-ID deduplication. Right: zero-based first-onset indices in the deduplicated event continuations, with median 8.

3.2.1 COMPLETE TRAJECTORIES AND ONSET ANNOTATION

For onset-based analysis, we sample complete responses, label whether each contains a hallucination, and, if so, locate the first unsupported span and map it to the generating model's token positions. An annotated trajectory thus retains the input, full response, event label, and onset, enabling comparison of initial prediction and later supported prefixes on the same response.

3.2.2 PREFIX-CONDITIONED CONTINUATIONS

We select parent prefixes before annotated onset, hold the image, prompt, and emitted text fixed, and resample multiple suffixes, labeling each future event. These *continuations* represent alternative futures from one observed state, beyond the suffix realized in the parent trajectory. Continuations of the same prefix share their input and split assignment.

Scale and quality control. PREHALLU contains approximately 85K trajectory records and 181,344 unique continuations from 22,668 parent prefixes, with eight continuations per prefix. Open-ended responses use model assessment and adjudication; the COCO/CHAIR subset uses object annotations and matching rules. Training and evaluation are grouped by image, with derived prefixes and continuations inheriting the parent image's split. Appendix A details the inventory, label scope, and quality audits.

Onset-annotated trajectories support training and evaluation, while prefix-conditioned continuations reveal alternative futures from the same observed state and support continuation-assisted forecasting.

4 RISKTRACE: PRE-ONSET MONITORING WITH HISTORICAL EVIDENCE

RISKTRACE turns visible-prefix evidence into response-level warnings through history-informed scoring, calibrated evidence fusion, and cumulative alarm generation. We first construct a prefix score s_t, then express the alarm interface relative to its initial value $b_0 = s_0$. Figure 3(a)–(b) summarizes

evidence extraction and this online interface. The InstructBLIP scorer below reads a frozen VLM's observed trajectory without generating auxiliary continuations.

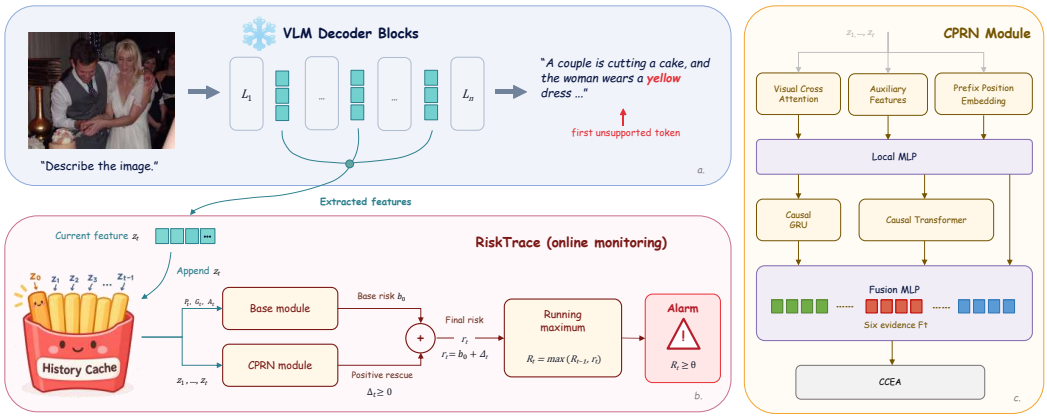

Figure 3: **Prefix evidence and the online alarm interface.** (a) A frozen VLM supplies evidence from the visible prefix. (b) An initial anchor and nonnegative evidence gains feed an irreversible alarm statistic. (c) An illustrative multi-head design for prefix-risk modeling; the evaluated scorer is defined in Sections 4.1–4.2.

4.1 Prefix evidence and historical context

Let z_t be the visible state after t response tokens have been emitted and before the next token is generated. The history cache is initialized with z_0 and retains $z_{0:t}$ as new states arrive. In InstructBLIP, the frozen final normalization and LM head recover a greedy candidate and its top-1 probability from the current decoder state. A trained probe combines candidate lexical features with causal trajectory features. To distinguish isolated high scores from persistent changes, the base path incorporates score increments, deviations from moving averages, high-score persistence, and a sparse scan, producing P_t.

Complementary paths summarize historical variation (H_t), attend to salient prefix states (A_t), model clean-response histories (N_t), and scan for deviations from clean references at comparable initial risk and generation position (C_t). With a high-score protection flag G_t, they form

$$\mathbf{c}_t = [P_t, G_t, A_t, H_t, N_t, C_t]^\top, \qquad G_t = \mathbb{I}\{P_t \geq \kappa_{\text{tail}}\}. \tag{5}$$

The protection threshold κ_{tail} is frozen before evaluation.

The attention branch projects causal history features, adds positional encodings, and applies a GRU. Critical pooling selects a salient state and aggregates support from the observed history; A_t averages member-wise running peaks at dynamic positions. We form the dynamic base score $B_t = \text{TailFuse}(P_t, G_t, A_t)$, which incorporates this historical support while preserving the base ordering in the protected upper tail. Its piecewise definition is given in Appendix C.

4.2 Calibrated evidence fusion and prefix scoring

The evidence branches differ in scale and capture complementary aspects of a prefix. Our *Calibrated Causal Evidence Aggregator* (CCEA) combines them on a common training-reference scale. The feature map Φ records component strengths, side-branch differences from P_t, and their dispersion; Ψ_{Train} maps these features to empirical ranks and rank logits:

$$\begin{aligned} \mu_t &= \text{CCEA}(\mathbf{c}_t) = \sigma\big(\mathbf{w}^\top \Psi_{\text{Train}}(\Phi(\mathbf{c}_t)) + \beta\big)\,, \\ m_t &= \widehat{F}_{\text{meta,OOF}}(\mu_t), \qquad \eta_t = \widehat{F}_{\text{base,clean}}(B_t). \end{aligned} \tag{6}$$

Here σ is the sigmoid. The rank m_t measures fused evidence relative to training out-of-fold (OOF) outputs, while η_t measures base-score extremity among clean training responses. Both references remain frozen.

We strengthen the base score when it is high relative to clean responses and the fused evidence is strong:

$$\delta_t = \lambda \mathbb{I}\{\eta_t \geq q_B\} \left[\frac{m_t - q_m}{1 - q_m}\right]_+, \qquad s_t = B_t + \delta_t. \tag{7}$$

Here $[u]_+ = \max(u, 0)$, $\lambda \geq 0$ sets the increment scale, and q_B and $q_m < 1$ are frozen thresholds. Other prefixes retain their base scores. CCEA supplies the evidence μ_t; the gated increment δ_t then produces the complete prefix score s_t.

4.3 Initial anchoring and online alarms

We express the alarm interface in Figure 3(b) as an initial anchor plus a nonnegative gain:

$$b_0 = s_0, \qquad \Delta_t = [s_t - b_0]_+, \qquad r_t = b_0 + \Delta_t = \max(b_0, s_t). \tag{8}$$

Here b_0 is the complete scorer's initial output, and Δ_t is its positive gain over that anchor, distinct from the internal increment δ_t. Because monitoring includes $t = 0$, this representation leaves the cumulative alarm statistic unchanged. Starting from $R_{-1} = -\infty$, we update

$$R_t = \max(R_{t-1}, r_t), \qquad t_{\text{alarm}} = \inf\{t : R_t \geq \theta\}. \tag{9}$$

An issued alarm remains active. The deployment threshold θ is selected from complete clean development trajectories, applying the empirical false-alarm budget to whole responses.

Training and freezing. Current and historical probes use event/clean supervision, while critical attention uses response-level multiple-instance supervision. Image-grouped OOF predictions supply response-level fusion features and reference outputs; the fusion model uses event labels with higher weight on clean examples and applies the same feature schema at each prefix. Development data select candidates, gates, and alarm thresholds. At inference, all parameters and reference distributions remain frozen; only prefix evidence and cumulative risk change. Appendix C gives the component definitions, training details, and anchoring equivalence.

5 Experiments

Experimental setup. We evaluate RiskTrace on InstructBLIP/Vicuna-7B and test cross-model replication on LLaVA-NeXT and three disjoint Qwen3-VL sets. The InstructBLIP study uses a separate onset-annotated benchmark; PreHallu supports the cross-model and continuation studies. Comparisons use the same scorer at generation start, optimized static controls, and pre-onset monitoring. We report response-level AUROC and recall at a 5% empirical clean-FAR budget (R@5), with 95% paired image-cluster bootstrap intervals. We instantiate the forecasting framework for each backbone; the LLaVA predictor additionally uses prefix-conditioned continuations. Data splits and baseline selection are detailed in Appendix D.1.

Predictive gains beyond generation start. Table 1 first tests the effect of observing additional eligible prefixes with the scorer held fixed. InstructBLIP AUROC rises from 75.43% to 83.58%, with a 95% gain interval of $[7.06, 9.21]$ percentage points (pp). At the same 48/970 clean false alarms, dynamic monitoring identifies 4,282/9,020 event responses versus 3,289 at generation start. It thus warns on 993 additional responses at the same response-level false-alarm cost, a recall gain of 11.01 pp.

The independent static control tests whether optimizing specifically for generation-start prediction closes the gap. We select it from 124 candidates on Development, spanning classifier families, model capacities, and feature representations, then freeze the choice before Fresh evaluation. It reaches 75.82% AUROC and 37.04% R@5, below RiskTrace by 7.76 and 10.43 pp, respectively. The AUROC gain interval $[6.80, 8.80]$ excludes zero. Thus the advantage remains when the initial predictor is optimized specifically for generation-start classification.

The gain is not confined to the 5% operating point (Figure 4(a)). At a 2% FAR cap, recall rises from 22.58% to 30.08%; at 10%, it rises from 46.77% to 61.36%. All seven evaluated budgets between 2% and 20% have positive simultaneous 95% lower bounds for gains over the same-stream initial baseline, covering both conservative and more permissive alarm budgets.

Table 1: **Initial prediction, static controls, and pre-onset monitoring.** Values are percentages; gains are AUROC pp. (a) compares predictors on the same InstructBLIP set; HALP reports the mean across runs. (b) compares dynamic monitoring with the same predictor at generation start.

(a) InstructBLIP main setting

Method	AUROC ↑	R@5 ↑
Uncertainty baseline	65.74	17.58
HALP (adapted)	70.93	–
RISKTRACE (initial)	75.43	36.46
Optimized static baseline	75.82	37.04
RISKTRACE (dynamic)	**83.58**	**47.47**

(b) Replication across models and evaluation sets

Set	Event/clean	Dynamic AUROC	Initial AUROC	Gain	R@5 (dyn./init.)
InstructBLIP	9020/970	83.58	75.43	+8.14	47.47/36.46
LLaVA-NeXT Fresh	364/836	81.12	79.09	+2.03	35.44/34.07
Qwen3 Fresh	217/1783	64.27	60.99	+3.28	21.66/11.06
Qwen3 Fresh2	438/3562	61.61	57.98	+3.63	15.53/8.45
Qwen3 Fresh3	416/3584	61.04	57.34	+3.71	16.83/7.69

The advantage extends beyond InstructBLIP. Qwen3-VL gains 3.28, 3.63, and 3.71 AUROC pp on its three disjoint sets; all six intervals for AUROC and R@5 gains lie above zero. LLaVA-NeXT also improves over the same predictor at generation start on an independent, cross-domain Fresh set using continuation-assisted evidence. Together, these results establish a consistent predictive advantage from still-supported generation across the evaluated VLM backbones.

The phenomenon also appears with a different evidence construction. In a complementary experiment (Table D.3), a frozen token-peak predictor on Qwen3-VL Fresh raises R@5 from 11.06% at generation start to 29.49%, a gain of 18.43 pp with interval $[11.11, 25.61]$. The gain extends low-FAR forecasting benefits to another evidence construction: token-level signals from supported positions can anticipate a later response-level hallucination.

Dynamic predictability and warning lead time. We hold InstructBLIP responses and the scorer fixed and vary only the cumulative monitoring horizon. In Figure 4(b), AUROC falls from 75.43% at $t = 0$ to 72.99% at $t = 1$, then reaches 77.87% at $t = 10$ and 83.59% at $t = 20$, versus 83.58% for full monitoring. A response's running maximum cannot decrease, but event and clean responses can change relative order. Cumulative discrimination therefore need not improve monotonically; more observations alone do not imply better warnings. By $t = 20$, AUROC is already within 0.01 pp of the full-monitoring result. The trajectory thus has an early decline, a later gain, and a plateau, distinguishing useful temporal evidence from a uniform benefit of longer observation.

At the same InstructBLIP matched-FAR5 threshold, the monitor identifies 39.37% of hallucinating responses with at least five supported tokens remaining before onset, and 28.98% with at least ten (Figure 4(c)). Both percentages use all event responses as the denominator. Let L count supported tokens emitted after the alarm and before onset. Among timely alarms, the median is $L = 12$, with quartiles 6 and 26. Unlike the monitoring horizon, which limits how long a response is observed, L measures the warning interval remaining after an alarm. These results demonstrate forecasting beyond the last decision immediately before onset.

Onset-stratified analysis complements these gaps by identifying which events are caught. Among InstructBLIP responses with onset at cache positions 11–20, the monitor catches 1,384 of 2,273 events; at positions 21 and later, it catches 1,909 of 4,718. Thus, coverage extends to responses with an extended supported prefix. Onset position describes the available observation interval, while the gap L measures how much of that interval remains after the alarm. Together, they locate both the events reached by monitoring and the opportunity for intervention before the error appears.

RISKTRACE component analysis. We remove upstream evidence paths and retrain the remaining fusion model under the same Train-OOF protocol (Table 2). Full RISKTRACE improves over current evidence alone by 3.16 AUROC pp and 5.06 R@5 pp. Removing critical-state attention reduces

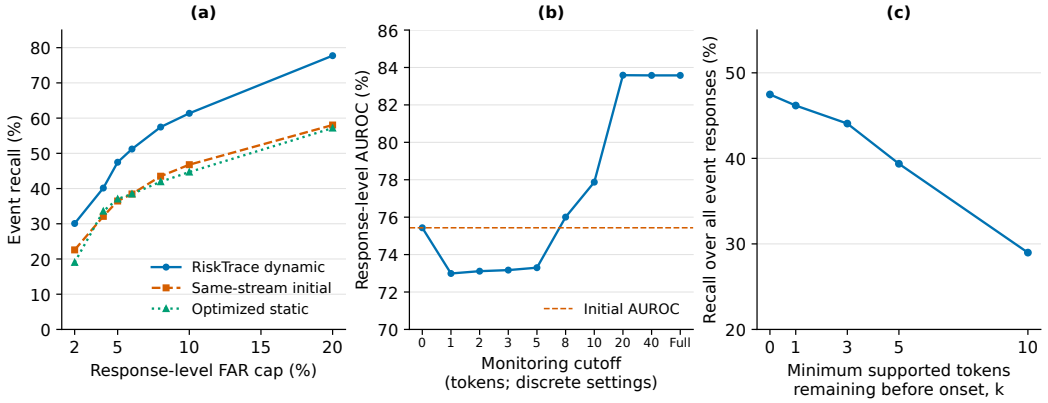

Figure 4: **Predictive gains, observation horizons, and warning lead time.** (a) Recall gains persist across response-level FAR budgets. (b) AUROC varies non-monotonically with the cumulative monitoring cutoff. (c) Recall with at least k supported tokens remaining before onset, at one fixed matched-FAR5 threshold (realized FAR 4.95%). The denominator is all 9,020 event responses; $k = 0$ includes every pre-onset alarm.

AUROC to 81.34%; removing both attention and history reduces R@5 to 42.14%. Historical paths therefore improve the use of current evidence under repeated monitoring.

Table 2: **Upstream-path ablations on InstructBLIP.** Each ablated component is removed from both base-score construction and fusion inputs.

Variant	AUROC (%)	R@5 (%)
Full RISKTRACE	**83.58**	**47.47**
Without critical attention	81.34	46.02
Without ordinary history	83.49	46.43
Without attention and history	80.76	42.14
Current evidence only	80.42	42.42

Ablations also distinguish evidence extraction from final aggregation (full results in Appendix F). Removing ordinary history changes AUROC by only 0.09 pp but reduces R@5 by 1.04 pp, indicating a larger effect in the low-FAR region than in overall ranking. Relative to the history-supported base, final fusion raises R@5 from 45.32% to 47.47%, while AUROC changes from 83.39% to 83.58%. The outer fusion's principal observed contribution is therefore low-FAR recall. A signed-residual alternative attains slightly higher AUROC but lower R@5, illustrating why both metrics matter for the alarm task.

A physical-truncation test directly checks prefix-only computation. Across 9,990 responses, we select prefixes by response-ID hash, remove all later states from the input files, and recompute predictions. This removes 172,239 future states from 9,951 responses. The maximum final-score difference is 4.623×10^{-4}, with no response exceeding the prespecified tolerance of 5×10^{-4}. Thus, the historical evidence used by the monitor is recoverable from the observed prefix itself. This computational check complements the prediction and lead-time results: the measured forecasting benefit is available before the future states are observed.

From early warnings to intervention. Beyond matched-FAR comparisons, we test whether early warnings persist when thresholds are fixed before evaluation. For InstructBLIP, each threshold is selected on Development at a nominal 5% clean-FAR target and transferred unchanged to Fresh. Table 3 reports both event recall and realized FAR: this tests a fixed decision rule, whereas Table 1 compares prediction at matched false-alarm costs.

With its frozen threshold, the dynamic monitor identifies nearly half of hallucinating responses at 5.36% realized FAR. Successful warnings have a median lead of 13 supported tokens; 41.10% of all event responses are caught with at least five supported tokens remaining, and 30.51% with at least ten.

Table 3: **Early warning with frozen thresholds on InstructBLIP.** Thresholds are selected on Development and are not retuned on Fresh. Realized FAR varies across methods; values are percentages.

Method	Event recall	Realized clean FAR
RISKTRACE (initial)	29.63	3.61
Optimized static baseline	34.21	4.23
RISKTRACE (dynamic)	49.28	5.36

The latter two recalls include missed events in their denominator. Thus, substantial advance coverage remains when the alarm threshold is selected before evaluation, not only when operating points are read from the evaluation ROC.

We next test whether an alarm can usefully trigger a grounding action. On Qwen3-VL Fresh2 and Fresh3, each containing 4,000 images, we compare original responses, reject-and-resample, and frozen visual-evidence selection. Within each set, rejection and visual selection act on the same alarms, so the comparison includes an active resampling baseline rather than only unchanged responses. Table 4 shows that visual selection reduces CHAIRs beyond rejection on both sets.

Table 4: **Object hallucination after alarm-triggered intervention.** CHAIRs (%, lower is better) is evaluated over all responses in each set; the two interventions use the same alarm set within each evaluation.

Set	Original	Reject-and-resample	Visual selection
Qwen3 Fresh2	10.950	10.175	**9.775**
Qwen3 Fresh3	10.400	9.875	**9.475**

The additional reduction is 0.40 pp on each set, with paired-bootstrap 95% improvement intervals of $[0.225, 0.625]$ and $[0.075, 0.726]$ pp, respectively. Mean response length changes by only 0.04 and 0.03 words relative to rejection, so these gains occur without substantial average shortening. Together, the fixed-threshold and intervention results connect pre-onset forecasting to a usable decision rule and targeted visual grounding. Appendices G–H provide the full deployment and intervention evaluations, including broader quality metrics and action costs.

6 Discussion and Limitations

Supported generation provides a useful forecasting window between initial prediction and local error detection. The observed non-monotonicity characterizes forecasting performance under cumulative monitoring, rather than intrinsic information in the VLM.

The evaluation uses machine-generated labels, with Qwen3 results covering object hallucination. Human onset audits would strengthen annotation validity. The image-disjoint InstructBLIP evaluation set was inspected during prior research; a newly sealed confirmation set would test robustness to research adaptation. Predictors and computational costs also vary across backbones. Compute-matched and task-aligned early-window comparisons would further clarify the sources of predictive gain and the cost of deploying these warnings.

7 Conclusion

Future hallucinations can become more predictable before the first unsupported token appears. We formalize this opportunity as pre-onset forecasting, construct PREHALLU to study supported prefixes and their possible futures, and develop RISKTRACE to turn current and historical evidence into response-level warnings. Gains over initial prediction recur across InstructBLIP, LLaVA-NeXT, and Qwen3-VL; lead-time analysis shows that useful warnings extend beyond the step immediately before onset. Our results establish still-supported generation as a useful forecasting window for hallucinations that have yet to appear.

AI USE STATEMENT

Generative AI assisted with writing, translation, methods, equations, and result analysis; model-based data generation and annotation are described in Section 3.2.

ETHICS STATEMENT

This work concerns the reliability of generated visual descriptions. Machine-generated support labels can be incomplete or incorrect, and an alarm does not establish that a response is false. Data documentation should retain image provenance and licensing conditions and distinguish inherited labels from newly generated ones.

REPRODUCIBILITY STATEMENT

Appendices A–B document data construction and evaluation. Appendix C describes the method and training configuration; Appendices D–H report additional results and their evidence sources. The available materials contain result summaries, selected scripts, and audit records. Core inference code, complete raw trajectories, hidden states, and model weights are not publicly released.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024. doi: 10.48550/arXiv.2404.18930.

Zhaorun Chen, Zhenying Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Jun Zhao, Weimin Wang, Boyi Li, Pascale Fung, and Steven C. H. Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jinhao Duan, Fei Kong, Hao Cheng, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. TruthPrInt: Mitigating large vision-language models object hallucination via latent truthful-guided pre-intervention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. DAMRO: Dive into the attention mechanism of LVLM to reduce object hallucination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7696–7712, 2024. doi: 10.18653/v1/2024.emnlp-main.439.

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, 2024. doi: 10.1109/CVPR52733.2024.01363.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18135–18143, 2024. doi: 10.1609/aaai.v38i16.29771.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, 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*, 2024.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27026–27036, 2024. doi: 10.1109/CVPR52733.2024.02553.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. doi: 10.18653/v1/2024.findings-emnlp.290.

Sai Akhil Kogilathota, Sripadha Vallabha E G, Luzhe Sun, and Jiawei Zhou. HALP: Detecting hallucinations in vision-language models without generating a single token. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, 2026.

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*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024a. URL https://arxiv.org/abs/2408.03326.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024b. URL https://arxiv.org/abs/2407.07895.

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, 2023. doi: 10.18653/v1/2023.emnlp-main.20.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024. doi: 10.48550/arXiv.2402.00253.

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhan Kang, and Xirong Li. PhD: A ChatGPT-prompted visual hallucination evaluation dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. doi: 10.1109/CVPR52734.2025.01849.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. Negative object presence evaluation (NOPE) to measure object hallucination in vision-language models. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research*, pp. 37–58, 2024. doi: 10.18653/v1/2024.alvr-1.4.

EunKyu Park, Minyeong Kim, and Gunhee Kim. HalLoc: Token-level localization of hallucinations for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. doi: 10.1109/CVPR52734.2025.02782.

Vishnu Prabhakaran, Purav Aggarwal, Vinay Kumar Verma, Gokul Swamy, and Anoop Saladi. VADE: Visual attention guided hallucination detection and elimination. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. doi: 10.18653/v1/2025.findings-acl.773.

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, 2018. doi: 10.18653/v1/D18-1437.

Dvir Samuel, Guy Bar-Shalom, Fabrizio Frasca, Ethan Fetaya, Yftah Ziser, Gal Chechik, and Haggai Maron. UniProbe: A learnable token-level hallucination detector for large VLMs using multi-structural internal representations. *arXiv preprint arXiv:2608.10835*, 2026.

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, 2024. doi: 10.18653/v1/2024.findings-acl.775.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023a. doi: 10.48550/arXiv.2311.07397.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, Jitao Sang, and Haoyu Tang. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023b. doi: 10.48550/arXiv.2308.15126.

Xintong Wang, Jingheng Pan, Ding Liang, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15840–15853, 2024. doi: 10.18653/v1/2024.findings-acl.937.

Sangmin Woo, Jae-Hyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. RITUAL: Random image transformations as a universal anti-hallucination lever in large vision language models. *arXiv preprint arXiv:2405.17821*, 2024. doi: 10.48550/arXiv.2405.17821.

Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. Logical closed loop: Uncovering object hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6944–6962, 2024a. doi: 10.18653/v1/2024.findings-acl.414.

Xiyang Wu, Tianrui Guan, Dianqi Li, Shuaiyi Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan Boyd-Graber, Tianyi Zhou, and Dinesh Manocha. AutoHallusion: Automatic generation of hallucination benchmarks for vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024b. doi: 10.18653/v1/2024.findings-emnlp.493.

Yuxi Xie, Guanzhen Li, Xu Xiao, and Min-Yen Kan. V-DPO: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13258–13273, 2024. doi: 10.18653/v1/2024.findings-emnlp.775.

Dingchen Yang, Bowen Cao, Guang Chen, and Changjun Jiang. Pensieve: Retrospect-then-compare mitigates visual hallucination. *arXiv preprint arXiv:2403.14401*, 2024. doi: 10.48550/arXiv.2403.14401.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12), 2024. doi: 10.1007/s11432-024-4251-x.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, pp. 13807–13816, 2024. doi: 10.1109/CVPR52733.2024.01310.

Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an EOS decision perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 11766–11781, 2024. doi: 10.18653/v1/2024.acl-long.633.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023. doi: 10.48550/arXiv.2310.00754.

Geigh Zollicoffer, Minh Vu, and Manish Bhattarai. MTRE: Multi-token reliability estimation for hallucination detection in VLMs. *arXiv preprint arXiv:2505.11741*, 2025.

A PREHALLU CONSTRUCTION AND ANNOTATION

A.1 CONSTRUCTION PROCEDURE

Figure 2 describes prefix-conditioned sampling and its statistics. Here we provide the complete-trajectory annotation procedure, inventory, splits, and audits.

An onset-annotated trajectory records the image, prompt, complete generated response, response-level event label, and first unsupported span, when present. A *continuation* records a suffix resampled from a supported parent prefix and whether that future suffix contains a hallucination. Multiple continuations from one prefix are alternative futures, not independent images or repeated observations of the parent trajectory's realized suffix.

A.2 SCALE AND INVENTORY

PREHALLU comprises complete response trajectories and prefix-conditioned continuations. The 85,000 trajectories are counted by summing collection inventories. There are 200,144 raw continuation records, or 181,344 after global response-ID deduplication, corresponding to 22,668 parent prefixes. Each parent has exactly eight continuations with distinct response IDs. Deduplication removes 18,800 repeated records; the main text reports over 180K unique continuations.

Trajectory counts exclude 4,051 action candidates, 2,392 utility comparisons, 21,600 sealed-test continuations whose labels remain unopened, and 31,168 historical InstructBLIP records used separately. The 85K total is reconciled at collection level; global response/image deduplication and full onset-annotation coverage across these collections remain unverified.

Data sources and study roles. The inventory and the main InstructBLIP benchmark serve distinct studies:

- The historical InstructBLIP records support the main static–dynamic comparisons, temporal analyses, and component ablations; their splits are reported in Appendix D.1.
- The LLaVA-NeXT and Qwen3-VL evaluation collections in PREHALLU support cross-model replication. The LLaVA Confirmation and Fresh sets are distinct, as are the three Qwen3-VL Fresh sets.
- Prefix-conditioned continuations expose alternative futures from a common observed state and support continuation-assisted forecasting. They remain separate data units from the complete trajectories.

Table A.1: PREHALLU inventory. Counts refer to trajectories or sampled continuations, not unique images. Collections can share image sources; continuations inherit their parent image's split.

Collection	Responses
LLaVA, Qwen2.5-VL, Qwen3-VL short-response Train/Development	50,400
Detailed-response supplement	4,200
Additional complete LLaVA responses	18,000
LLaVA Confirmation and independent Fresh	2,400
Qwen3-VL COCO/CHAIR Fresh	10,000
Total response trajectories	85,000
Raw prefix-conditioned continuation records	200,144
Continuations after response-ID deduplication	181,344
Unique parent prefixes	22,668

A.3 SPLITS AND QUALITY CONTROL

Training and evaluation are grouped by image, and derived prefixes and continuations inherit that assignment. The September 23, 2026 continuation audit covers six source collections and 3,600 unique images. Cross-split intersections of image SHA-256 hashes, parent-prefix identities, and

response IDs are all empty. Identical text or suffix-token sequences can nevertheless occur for different images; the split audit tests isolation by image and parent identity.

Table A.2: Deduplicated continuation counts by split. Unique continuations are counted by response ID; all counts come from the record-level inventory audit.

Split	Raw	Unique	Parents	Images	Event/clean
Train	125,360	106,560	13,320	1,200	37,732/68,828
Confirmation	37,424	37,424	4,678	1,200	13,880/23,544
Fresh	37,360	37,360	4,670	1,200	13,838/23,522
Total	200,144	181,344	22,668	3,600	65,450/115,894

The audit checks parent-key bindings, label matches, empty continuations, and uncertain labels. All six raw collections match their final labels, with no empty continuations or `UNSURE` labels. A two-annotator blind audit with adjudication has not been completed. Source-image licenses are not uniformly embedded in the manifest, and redistribution conditions require source-level verification before release.

A.4 Annotation sources and scope

Blinded assessment models label open-ended responses, with model adjudication resolving disagreements. These machine pseudo-labels address broad visual support at the response level. Structured COCO/CHAIR populations instead use object annotations and mentions to identify object hallucination and onset. For the latter, a supported prefix denotes the absence of an annotated object hallucination. Result tables identify the evaluation sets and label sources.

Onset is first localized to a character span and then mapped to the generating model's tokenizer. A response with no annotated unsupported span has no onset position but still contributes its complete trajectory to clean-response FAR. For event responses, only prefixes strictly before onset enter pre-onset evaluation; post-onset tokens and the onset annotation itself are not monitor inputs. Each trajectory retains its image identifier, prompt, generation configuration, token sequence, label source, and parent-prefix index to support replay and audit.

A.5 Prefix selection and future continuations

For an onset-annotated trajectory, we select eligible prefixes before its first unsupported token and restart generation from the same image, prompt, and prefix. A continuation's parent key points to the observed prefix, not the complete response's realized suffix. Different continuations therefore represent alternative futures. Their labels support study of the conditional future distribution. Image-grouped assignments propagate to parent responses, prefixes, and continuations across training, development, and evaluation.

One traceable example starts from the parent prefix `A train`. Of its eight continuations, seven receive clean machine labels and one an event label, illustrating different futures from the same observed prefix rather than an error already in that prefix. Deduplicated event continuations have median zero-based onset index 8, quartiles 4/14, and range 0–62. These are recorded response-level onset indices, not warning lead times or distances from the parent prefix. Eight samples per parent provide repeated observations of possible futures from a fixed prefix.

B Task and Evaluation Protocol

B.1 Causal information constraints

The passive predictor reads only the image, prompt, generated prefix, and current decoder features. It does not read the realized future suffix, annotated onset, or final response length. Onset is an evaluation variable, not an input, and an event alarm succeeds only before τ. Auxiliary continuations sampled from the current prefix constitute additional information and computation; these results are reported separately from passive monitoring of the observed trajectory.

B.2 Response-level evaluation

Clean responses are monitored until natural termination, and any alarm counts toward clean-response FAR. For event responses, only pre-onset monitoring times are eligible. We report response-level AUROC and recall at approximately 5% matched clean FAR on the evaluation ROC. Development-frozen threshold transfer is a separate evaluation.

For budget cap α, low-FAR recall on the discrete ROC is

$$\mathrm{R@}\alpha = \max_{\theta:\, \widehat{\mathrm{FAR}}(\theta) \le \alpha} \widehat{\mathrm{Recall}}(\theta). \tag{10}$$

Discrete thresholds and ties can produce realized FAR below the nominal budget. Actual counts must therefore be retained; R@5 does not always mean exactly 5% FAR. In the BLIP same-stream comparison, both methods have 48 clean false alarms at R@5, whereas the optimized static control has 47. Inference uses paired bootstrap over image clusters, resampling all responses from one image together. Percentages are rounded from raw proportions multiplied by 100; percentage-point differences are computed from unrounded values whenever available.

B.3 Alarms under repeated observation

The monitor updates its score at each observed prefix, and an alarm remains active once triggered. Clean-response FAR is thus the probability of at least one threshold crossing over a full response, not a token-level false-positive rate. An event is caught only if at least one crossing occurs before onset. This matches the use of an online warning system and explains why transient high scores can incur additional costs as observation continues. All methods are compared under the same response-level protocol.

B.4 Minimum-lead-time evaluation

The main text uses one-based output-token onset τ_i and emitted-prefix length a_i at alarm time. To align with the frozen supplementary report, we define $L_i = \tau_i - a_i - 1$ as the number of supported tokens that can still be emitted before onset. $L_i = 0$ means the next token is the first unsupported one, but the alarm still precedes it. Counting generation steps including that token would instead give $L_i + 1$; we keep these units distinct.

Event responses without timely alarms are excluded from conditional lead-time quantiles but remain in the recall denominator. At a fixed threshold θ,

$$\mathrm{Recall}_{L \ge k}(\theta) = \frac{1}{N_1} \sum_{i:Y_i=1} \mathbb{I}\{a_i(\theta) < \tau_i,\ \tau_i - a_i(\theta) - 1 \ge k\}. \tag{11}$$

All values of k use the same development-frozen threshold or the same matched-FAR threshold derived from complete clean trajectories; thresholds are not retuned for a minimum gap. In the frozen implementation, state s precedes output token s, and onset has zero-based index j. Thus $s \le j$ corresponds to $a < \tau$, and the reported $j - s$ equals L. These indices bind evaluation records and are not predictor features.

C RiskTrace Implementation Details

C.1 Backbone-specific implementations

Table C.1 links each evaluation to its evidence path. All predictors score supported prefixes under the same response-level forecasting protocol. The main InstructBLIP results use the six-component scorer defined below. Figure 3(c) illustrates a multi-head design; its risk, peak-identity, and hazard heads are not components of this evaluated scorer. CCEA in the main method denotes training-reference transforms and frozen logistic fusion.

For Qwen3, four decoder depths query same-image visual tokens. A cross-layer Transformer constructs current evidence c_t, and a unidirectional GRU encodes its history, including changes from the initial state, innovations, and running summaries. Initial and dynamic heads produce b_0 and u_t:

$$\Delta_t = [u_t - b_0]_+, \qquad r_t = b_0 + \Delta_t = \max(b_0, u_t). \tag{12}$$

Table C.1: Predictor implementations used in the main evaluations.

Backbone	Evidence and aggregation	Evaluation
InstructBLIP	Current probe, causal history, critical attention, and reference-calibrated residual fusion	9,990 historical Fresh responses
Qwen3-VL	Multilayer visual alignment, cross-layer Transformer, temporal GRU, and initial/token risk heads	Fresh, Fresh2, Fresh3
LLaVA-NeXT	Prefix-conditioned auxiliary continuations and causal risk aggregation	LLaVA Fresh

The response score is the maximum over eligible prefix scores. Qwen3's alarm-triggered action experiments threshold the dynamic head u_t at its development-selected operating point. The main-text InstructBLIP construction instead updates its internal base score B_t throughout generation and anchors the alarm interface at the complete initial score s_0.

C.2 INSTRUCTBLIP EVIDENCE CONSTRUCTION

The current probe reconstructs a greedy candidate and top-1 probability from the layer-32 decision state through the frozen final normalization and LM head. A trained probe combines candidate lexical features with causal trajectory features. History, sparse scans, and latent fusion then form P_t. The InstructBLIP predictor uses six components: the pre-protection base score P_t, protection flag G_t, critical attention A_t, ordinary history H_t, clean-reference history N_t, and conditional scan C_t. Learned weights and training-reference distributions are frozen together and reused at every evaluation prefix.

C.3 SIX-COMPONENT FUSION

Tail protection defines the dynamic base B_t before the internal increment δ_t is added. With the frozen weight $\omega \in [0, 1]$, the main-text operator $\mathrm{TailFuse}$ is

$$\begin{aligned} M_t &= \mathrm{clip}((1-\omega)\rho_P(P_t) + \omega\rho_A(A_t),\, 0,\, 1)\,, \\ B_t &= \begin{cases} 1 + P_t, & G_t = 1, \\ \max(P_t, M_t), & G_t = 0. \end{cases} \end{aligned} \tag{13}$$

Here $G_t = \mathbb{I}\{P_t \geq \kappa_{\mathrm{tail}}\}$. Since both ranks lie in $[0, 1]$, clipping does not change their convex mixture. Each rank map uses its corresponding scores on clean training responses. For base scores in $[0, 1]$, protection places the upper tail in a separate score interval while preserving its base ordering.

The 12 entries in $\Phi(\mathbf{c}_t)$ comprise six raw components, four differences of A_t, H_t, N_t, C_t from P_t, and the standard deviation and range of these four side signals. Frozen Train references map each entry to an empirical rank; concatenating clipped rank-logits gives 24 fusion inputs. The configuration uses source-conditional ranks with pooled fallback when a source lacks a local training reference. Logistic fusion uses $C = 0.1$, clean-example weight 8, and five meta-OOF folds grouped by image hash.

The logistic output μ_t is ranked against all training OOF outputs to form m_t, whereas B_t is ranked against the training clean-response base-score reference to form η_t. The configuration has $\lambda = 2$, $q_B = 0.8$, and $q_m = 0.5$:

$$s_t = B_t + 2\,\mathbb{I}\{\eta_t \geq 0.8\} \left[\frac{m_t - 0.5}{1 - 0.5}\right]_+ . \tag{14}$$

References, classifiers, and gates are fixed before evaluation. The resulting score s_t is real-valued rather than a calibrated probability.

C.4 Initial-anchor representation and alarm equivalence

Let $s_0, \ldots, s_t$ be the complete scorer's outputs, with monitoring starting at $t = 0$. Equation (8) defines $b_0 = s_0$, $\Delta_j = [s_j - s_0]_+$, and $r_j = \max(s_0, s_j)$. Hence

$$\begin{aligned} R_t &= \max_{0 \le j \le t} r_j = \max\left\{ s_0, \max_{0 \le j \le t} s_j \right\} \\ &= \max_{0 \le j \le t} s_j. \end{aligned} \tag{15}$$

The final equality follows because the monitored set contains $j = 0$. For any fixed threshold θ, both representations therefore have the same first crossing, response-level score, and cumulative-horizon ranking. Anchoring is an equivalent alarm interface, not an additional learned component or an extra source of performance gain.

The initial anchor is the complete score s_0, not the internal base B_0. Likewise, $\Delta_t = [B_t + \delta_t - s_0]_+$ is distinct from the gated increment δ_t. Individual scores r_t and s_t can differ below the anchor; only their cumulative maxima are equivalent. Analyses of unaggregated prefix scores refer to s_t. The history cache includes z_0, while $z_{1:t}$ denotes subsequent updates; the initial state remains available when computing s_t.

C.5 Attention-branch training objective

The attention branch constructs causal history features and standardizes them with frozen training means and scales before GRU encoding. At dynamic positions, A_t averages member-wise running maxima of prefix logits. Strict replay sets $A_0 = -30$ and uses the same training-reference rank map thereafter.

The branch uses response-level multiple-instance supervision. Let $\mathcal{B}_i = \mathcal{L}_i \cap \{t \ge 1\}$ be the eligible dynamic positions. For prefix logits $\ell_{i,t}$ in a nonempty bag, its smooth response summary is

$$\ell_i^{\mathrm{bag}} = \gamma \log \left(\frac{1}{|\mathcal{B}_i|} \sum_{t \in \mathcal{B}_i} \exp(\ell_{i,t}/\gamma) \right). \tag{16}$$

Here $\gamma > 0$ is the pooling temperature. The objective combines binary cross-entropy on Y_i with a pairwise term. Hard-clean sampling emphasizes the score region relevant to low FAR.

C.6 Training, freezing, and threshold transfer

InstructBLIP components first produce OOF predictions on image-grouped training folds, which the fusion head then consumes as one feature row per response. Online inference constructs the same feature schema at each observed prefix. Final models are fit on all training data, while OOF outputs supply the reference distribution. Training references, empirical-rank transforms, and hard-clean references are fixed before evaluation. Evaluation data are used for scoring with these fixed references. Development data select fusion hyperparameters and the alarm threshold, and the same frozen scoring procedure runs at every evaluation prefix. Matched-FAR points in Table 1 are read from each method's evaluation ROC; development-frozen results directly transfer one selected threshold. They answer different questions and are reported separately.

C.7 Passive monitoring versus auxiliary continuations

The main RiskTrace result uses passive monitoring of the emitted prefix and decoder features. LLaVA's continuation-assisted predictor additionally samples possible futures from the current prefix; its generation costs are reported in Appendix G.

C.8 Online execution

The monitor caches initial evidence at $t = 0$. Each subsequent prefix updates the history and complete score s_t, then forms the anchored score r_t and updates the running maximum R_t. The first threshold crossing triggers an irreversible alarm, which can route verification, abstention, resampling, or visual grounding.

D EXTENDED EXPERIMENTS AND TASK-ALIGNED COMPARISONS

D.1 EVALUATION SETS AND STATISTICAL SETUP

The main InstructBLIP/Vicuna-7B setting contains 17,303 training, 3,875 development, and 9,990 Fresh evaluation responses. Evaluation comprises 9,020 event and 970 clean responses across 2,500 image clusters. Its images are disjoint from training and development, although this external Fresh set has been inspected in prior project work. These historical data are excluded from the 85K PREHALLU inventory. LLaVA-NeXT Fresh contains 400 images each from COCO, Flickr30k, and NoCaps. Qwen3-VL uses three disjoint COCO sets totaling 10,000 images. Label sources and data construction are detailed in Appendix A.

The main comparisons are the final scorer's first score, independently optimized static predictors, and cumulative pre-onset monitoring. BLIP also includes a development-selected uncertainty baseline and a HALP MLP adapted to Vicuna's hidden dimension (Kogilathota et al., 2026). Predictors are model-specific. LLaVA uses an average of 5.766 auxiliary continuations per prefix; generation budgets and costs are given in Appendix G.

AUROC and recall are evaluated at the response level. R@5 uses the empirical evaluation ROC under a clean-FAR cap of 5%; realized false-alarm counts are retained for discrete thresholds. Confidence intervals use paired bootstrap over image clusters, keeping responses from the same image together. Matched-FAR comparisons assess ranking and low-FAR recall, while threshold transfer applies a development-selected threshold unchanged to evaluation data (Appendices B and G).

D.2 TRAINING AND SELECTING STATIC CONTROLS

The same-stream $t = 0$ comparison holds the scorer fixed and changes only whether later prefixes are observed. The independent static control instead allows optimization specifically for prediction before generation. The 124 BLIP candidates span logistic regression, HistGradientBoosting, ExtraTrees, RBF-SVM, MLPs of different capacities, raw/rank/hybrid features, and rank ensembles. Inputs include eligible $t = 0$ components, multilayer logit-lens features, entropy, top-1 probability, and margins, without later tokens, onset, or response length. Selection uses five-fold image-grouped OOF evaluation on 3,875 Development responses. The selected model, balanced logistic regression on score features with $C = 10$, is frozen before Fresh evaluation.

Its Fresh AUROC is 75.82%, versus 83.58% for RISKTRACE, with paired image-cluster 95% gain interval $[6.80, 8.80]$ pp. At R@5, the static control has 47/970 clean alarms and 3,341/9,020 event detections; dynamic monitoring has 48/970 and 4,282/9,020. The recall gain is 10.43 pp, with interval $[3.48, 15.79]$. "Strong static" denotes selection from the prespecified candidates on development data, with the resulting choice fixed for Fresh. The dimension-adapted HALP MLP yields 70.93% mean AUROC over six seeds, with standard deviation 0.45 pp. It serves a different comparison role from the same-stream initial score and optimized static control.

D.3 RESULTS ACROSS FALSE-ALARM BUDGETS

Table D.1 reports initial, optimized-static, and dynamic recall. All seven simultaneous 95% lower bounds on gains over the same-stream initial baseline are positive. Against the independent static control, all seven point estimates are positive, but the simultaneous lower bound at the 4% budget is -0.20 pp. The statistical conclusions of these comparisons are therefore different.

D.4 CROSS-MODEL REPLICATION AND LLAVA BASELINE ALIGNMENT

The `own-t0` baseline in the original LLaVA Fresh v129 report differs from the first score of the final dynamic stream. The supplementary audit retains both on the same 1,200 responses (Table D.2). The main text consistently uses the final stream's first score, giving LLaVA gains of 2.03 AUROC pp and 1.37 R@5 pp. The original 2.43/9.62 gains are not same-stream comparisons at equal realized FAR.

The original dynamic-minus-own-t0 intervals are $[1.21, 3.75]$ AUROC pp and $[-2.17, 13.57]$ R@5 pp for the originally reported baseline. A separate LLaVA Confirmation v95 set has 1,200 images, 353 event responses, and 847 clean responses. Dynamic/reported-initial AUROC is 81.30/78.90%,

Table D.1: BLIP across FAR budgets. Parentheses give clean false-alarm counts out of 970. The last column is the simultaneous 95% lower bound (pp) on dynamic-minus-initial recall.

FAR cap	Initial recall	Strong static	Dynamic recall	Lower bound
2%	22.58 (18)	19.04 (19)	30.08 (19)	1.48
4%	32.11 (38)	33.58 (38)	40.14 (38)	2.03
5%	36.46 (48)	37.04 (47)	47.47 (48)	5.00
6%	38.46 (58)	38.46 (57)	51.22 (58)	6.75
8%	43.50 (77)	41.93 (77)	57.46 (76)	7.95
10%	46.77 (97)	44.68 (97)	61.36 (97)	8.58
20%	58.04 (194)	57.17 (194)	77.73 (190)	13.68

Table D.2: LLaVA Fresh baseline identity and realized false alarms. All rows use the same evaluation set.

Score	AUROC	R@5	FAR	Event/clean alarms
Reported own-t0	78.69	25.82	2.15	94/18
Final stream's first score r_0	79.09	34.07	4.90	124/41
Dynamic running maximum	81.12	35.44	4.90	129/41

with gain interval $[1.15, 3.77]$; R@5 is 26.06/23.23%, with interval $[1.18, 4.66]$. We report the two evaluation sets separately.

For Qwen3 Fresh/Fresh2/Fresh3, dynamic-minus-initial AUROC gains are 3.28, 3.63, and 3.71 pp, with 95% intervals $[1.64, 5.28]$, $[2.27, 5.05]$, and $[2.39, 5.12]$. R@5 gains are 10.60, 7.08, and 9.13 pp, with intervals $[5.38, 14.49]$, $[4.13, 9.77]$, and $[5.75, 11.88]$. The sets contain 10,000 disjoint images in total, providing repeated evidence of dynamic-prediction gains with the Qwen3 predictor in Table C.1.

D.5 Task-aligned comparisons with existing detectors

Table D.3 records the training stage and comparator of HALP-, MTRE-, and TruthPrInt-style experiments (Kogilathota et al., 2026; Zollicoffer et al., 2025; Duan et al., 2025). They use different data and budgets, so absolute AUROC values do not form a common leaderboard. They test how existing forms of evidence adapt to pre-onset event forecasting.

Table D.3: Task-aligned supplementary comparisons. Values are method/baseline; parentheses identify the evaluation stage.

Method and stage	Baseline	AUROC	R@5
LLaVA endpoint (Train-OOF)	HALP-compatible static	80.45/61.57	26.75/10.64
LLaVA MTRE-causal (Train-OOF)	Independent static	60.61/55.33	10.10/7.46
Qwen3 future-event adaptation (Selection)	Same-head initial	59.40/52.63	2.86/5.71
Qwen3 token-peak (Selection)	Frozen initial	68.29/56.02	31.43/8.57
Qwen3 token-peak (Fresh)	Initial	64.56/60.99	29.49/11.06

Both LLaVA Train-OOF experiments use 12K responses and 6K images. The HALP-compatible comparison has AUROC/R@5 gain intervals of $[17.80, 20.05]$/$[13.64, 18.14]$ pp. MTRE-causal gains over independent static prediction have intervals $[3.88, 6.60]$/$[0.79, 4.60]$, but gains over its own initial head are not significant. Qwen3 future-event adaptation has Selection AUROC gain interval $[0.57, 12.94]$, while its low-FAR recall decreases. These distinct adaptation outcomes require considering both metrics.

On independent Qwen3 Fresh, token-peak improves R@5 by 18.43 pp, with interval $[11.11, 25.61]$. Its AUROC gain is 3.57 pp, with interval $[-1.62, 8.67]$ crossing zero. The token-peak predictor therefore provides additional evidence of low-FAR predictive value. Selection and Fresh are reported separately; Appendix G gives realized FAR under frozen thresholds.

E Temporal Analysis and Prefix Information Audits

E.1 Cumulative monitoring horizon

Table E.1 varies the cumulative cutoff for the same BLIP responses and frozen scorer. Event trajectories retain only eligible pre-onset states; clean trajectories are subject to the same cutoff. The reported quantity is response-level discrimination of cumulative maximum risk, not token classification at a single position or a direct measure of intrinsic risk information.

Table E.1: BLIP cumulative monitoring horizons.

Cutoff	AUROC (%)	R@5 (%)
0	75.43	36.46
1	72.99	36.39
2	73.11	36.46
3	73.17	36.65
5	73.30	37.24
8	76.00	40.32
10	77.87	42.73
20	83.59	47.57
40	83.58	47.47
Full	83.58	47.47

Although a response's running maximum never decreases, ranking can worsen when more clean responses receive higher scores. This explains why the early AUROC decline is compatible with irreversible alarms. Performance reaches a plateau by $t = 20$, within 0.01 AUROC pp of full monitoring.

E.2 Pre-onset gaps and coverage

This analysis reads frozen per-prefix scores without retraining, generating responses, or searching candidates. Table E.2 fixes the first alarm at each model's original matched-FAR5 threshold. Conditional quantiles include only timely alarms; recall denominators contain all event responses. The gap L counts supported tokens as defined in Appendix B.

Table E.2: Warning gaps and overall coverage. Recall values are percentages; gap quartiles are conditional on timely alarms.

Set	Recall	L: Q1/med./Q3	$L \geq 1$	$L \geq 3$	$L \geq 5$	$L \geq 10$
InstructBLIP	47.47	6/12/26	46.18	44.07	39.37	28.98
LLaVA Fresh	35.44	2/4/7	31.59	24.45	17.03	6.59
Qwen3 Fresh	21.66	0/2/19.5	11.52	10.60	10.60	10.14
Qwen3 Fresh2	15.53	0/0.5/19	7.76	7.53	7.31	6.39

Of BLIP's 4,282 timely event alarms, 3,551 have $L \geq 5$ and 2,614 have $L \geq 10$. Events without a timely alarm remain in the denominator of 9,020. Gaps differ across models, with some Qwen3 alarms occurring at the last eligible state. Table E.2 distinguishes overall coverage from coverage at positive lead times.

E.3 Stratification by onset position

Early-onset responses offer shorter eligible observation intervals. Stratified coverage identifies which events are caught; conditional gaps describe how early. Counts are reported for each stratum to show their different sample sizes.

E.4 Physical truncation and frozen-state audit

The main BLIP replay recomputes all six frozen components and fusion scores at each eligible prefix while preserving SHA-256 hashes of 25 weight and reference files. Every dynamic response score

Table E.3: Timely coverage by onset cache position in the frozen report. Entries are alarms/events within a stratum. Appendix B maps cache indices to output-token positions.

Set	0–2	3–5	6–10	11–20	21+
InstructBLIP	6/44	300/732	683/1253	1384/2273	1909/4718
LLaVA Fresh	36/78	39/92	30/111	21/74	3/9
Qwen3 Fresh	0/7	4/18	9/40	17/81	17/71
Qwen3 Fresh2	1/14	8/30	16/83	18/156	25/155

includes the same-stream initial score, with zero violations. A separate black-box test selects prefixes solely by response-ID hash, without labels, onset, or scores, physically removes later states from input files, and rescores the retained prefix.

Table E.4: BLIP physical-truncation audit. Differences compare the same prefix rescored before and after truncation.

Audit item	Result
Responses audited / with suffix states removed	9,990 / 9,951
Original / retained states	346,477 / 174,238
Future states removed	172,239
Maximum critical-attention difference	1.688×10^{-4}
Maximum final-score difference	4.623×10^{-4}
Prespecified final-score tolerance	5×10^{-4}
Responses exceeding final-score tolerance	0

The 39 responses with no removed suffix are excluded from the effective truncation count. Small attention differences are recorded as floating-point effects from altered padding or matrix shapes. The other reported base, ordinary-history, natural-history, and conditional-scan components have zero difference. The test checks the implemented scorer's invariance to future states.

E.5 Repeated observations and permutation controls

Copying r_0 to every eligible position and taking its maximum yields r_0 for every response. This equal-look control excludes gains from simply repeating an unchanged initial score, but does not test temporal ordering of new features. Permuting an already computed score sequence also leaves its maximum unchanged, so post-scoring permutation invariance does not establish the value of history order. Alternative-network ablations that change actual history inputs are reported separately in Appendix F.

F Complete Ablations and Alternative Models

F.1 BLIP upstream-path and fusion-layer ablations

Upstream ablations remove a module from both base-score construction and fusion inputs, then retrain the remaining fusion model with the same Train-OOF protocol. Beyond the main-text variants, sparse current evidence achieves 80.72% AUROC and 43.18% R@5; removing ordinary history gives 83.49/46.43%, and current evidence alone gives 80.42/42.42%. These are component comparisons within the main setting.

Table F.1 changes only the final meta/fusion rule while upstream checkpoints remain frozen. Removing its history slot does not remove all historical information because the base score can still contain a history path. Table F.1 reports overall ranking, low-FAR recall, and transferred-threshold performance for each fusion rule.

The full method's principal additional benefit over the base champion is R@5 (45.32% to 47.47%), with a smaller AUROC change. A signed residual yields higher AUROC but slightly lower R@5, exposing a tradeoff between overall ranking and recall in the low-FAR region.

Table F.1: BLIP frozen-fusion ablations. The last two columns are realized values after development-threshold transfer, not matched-FAR comparisons.

Variant	AUROC	R@5	Transfer recall	FAR
Base champion	83.39	45.32	45.75	5.15
Full RISKTRACE	83.58	47.47	49.30	5.36
Remove attention slot	83.51	45.85	47.66	5.36
Remove history slot	83.58	47.64	48.94	5.26
Remove natural-history slot	83.51	45.10	40.52	4.23
Remove conditional-scan slot	83.62	47.48	45.94	4.85
No source-conditional rank	83.58	47.52	49.30	5.36
No clean-tail gate	83.58	47.54	49.37	5.36
No meta-tail threshold	83.61	46.02	45.47	4.74
Signed residual	83.74	47.13	43.71	4.12

F.2 DEVELOPMENT ABLATIONS OF A UNIFIED SEQUENCE NETWORK

Cross-backbone ablations evaluate an end-to-end sequence network alongside the main InstructBLIP predictor. Table F.2 reports 10 Qwen3 variants evaluated through Development fit-to-selection, separately from the Fresh evaluations.

Table F.2: Qwen3 Development ablations of an alternative sequence network.

Variant	AUROC	R@5
Primary dynamic network	62.88	20.00
No visual attention	57.67	25.71
No current evidence	50.00	0.00
No history	58.96	20.00
No innovation	59.40	17.14
No clean-tail	67.11	20.00
No initial anchor	60.29	17.14
Shuffled history	59.32	20.00
Random history	60.64	14.29
Parameter-matched static (repeated initial state)	47.23	2.86

Against the parameter-matched static model, the primary dynamic network has AUROC/R@5 gain intervals $[1.31, 29.96]$/$[5.41, 32.15]$ pp; intervals against its own initial score cross zero. The no-running-maximum run is excluded pending verification of its evaluation protocol.

In the same series, InstructBLIP Development AUROC is 55.19% for the primary dynamic network and 57.72% for parameter-matched static. Current-only, no-history, and no-visual-attention variants achieve 54.34, 54.93, and 57.78%; shuffled and random history give 54.18 and 55.64%. Seven of eight planned variants have results; the fixed-gate run has no valid evaluation. The LLaVA no-auxiliary series has four completed variants with zero-filled auxiliary channels and is reported separately from auxiliary-enabled experiments.

F.3 LARGER TRAJECTORY SETS AND ALTERNATIVE EVIDENCE

A LLaVA full-hidden + visual GRU was evaluated with five-fold image-grouped OOF on 18,000 responses, 6,000 images, and 247,300 eligible prefix states. Dynamic AUROC is 58.78%, versus 60.07% for independent static prediction, a -1.30 pp difference with 95% interval $[-1.85, -0.78]$. All five fold differences are negative; R@5 is 8.73% versus 9.75%. More hidden-state trajectories in a sequence model do not automatically yield pre-onset gains.

By contrast, LLaVA rollout/peak variants give positive results on the same 1,200-image Train-OOF population. Fixed fusion, multi-FAR, multi-horizon, and fixed-gain variants achieve AUROC 82.89, 82.65, 82.74, and 83.05%, respectively, against static AUROC 80.05%. Gain intervals for the latter three are $[1.46, 3.90]$, $[1.38, 4.08]$, and $[1.50, 4.58]$ pp, but low-FAR gains are less stable. These variants use the same Train data.

A post-hoc three-domain analysis of the fixed-gain variant gives AUROC gains of 2.79, 1.61, and 4.65 pp on COCO, Flickr30k, and NoCaps. The Flickr30k interval and all three R@5 gain intervals cross zero. This analysis locates where evidence helps within the Train-OOF population. Another fusion variant gains only about 0.08 AUROC pp on Qwen3, with an interval crossing zero, while a later LLaVA static-miss hazard model gives 56.45% dynamic versus 59.15% static AUROC. Together, the positive and negative results identify which evidence constructions yield predictive gains in the evaluated settings.

G Threshold Transfer and Computational Cost

G.1 External evaluation of frozen thresholds

Table G.1 applies thresholds selected during training or development directly to evaluation data and reports realized FAR. This tests threshold transfer by reporting both event coverage and realized false-alarm cost under a fixed deployment rule.

Table G.1: Transfer of nominal FAR5 thresholds. The BLIP dynamic row uses the first-alarm audit in the lead-time supplement.

Set and scorer	Event recall	Clean FAR
BLIP same-stream initial	29.63	3.61
BLIP optimized static	34.21	4.23
BLIP dynamic (first-alarm audit)	49.28	5.36
LLaVA dynamic	40.38	6.34
Qwen3 Fresh learned dynamic	17.05	2.64
Qwen3 Fresh token-peak	32.72	7.12
Qwen3 Fresh2 learned dynamic	12.79	3.14
Qwen3 Fresh2 token-peak	27.17	6.18

The original BLIP online registry records 4,447 event alarms (49.30%); the later first-alarm supplement from the immutable score stream records 4,445 (49.28%). Both record 52/970 clean alarms. Pending per-response reconciliation of the two boundary counts, we retain each artifact's provenance rather than interpret the discrepancy as a performance change. Both agree on the main matched-FAR5 result of 4,282/9,020. Table F.1 uses the original ablation registry; Table G.1 uses the first-alarm supplement.

At these frozen thresholds, BLIP has median conditional gap 13 and overall recall 41.10/30.51% for $L \geq 5/10$; LLaVA has median 5 and coverage 20.88/7.97%. Qwen3 Fresh learned dynamic has median 0 and coverage 5.53/5.07%; Fresh2 has median 0 and coverage 4.11/4.11%. These results further distinguish event coverage from warnings with positive gaps.

G.2 Auxiliary continuation budgets and runtime

LLaVA costs are measured on one RTX 4090 using 100 label-blind Train responses. For each budget, 10 responses are used for warmup followed by 100 timed responses. Responses contain 3.95 eligible prefixes on average. Continuations have a maximum of 64 tokens, temperature 0.7, and top-p 0.9. An independent audit covers all 500 budget records. Table G.2 measures only the VLM generation subsystem, excluding the external judge; it is not end-to-end detector latency.

Table G.2: LLaVA auxiliary-generation cost. Memory is peak allocated GPU memory; throughput is measured in responses per second.

Rollouts/prefix	Forwards/prefix	Mean s/resp.	p90 s	Resp./s	GB
0	1.00	1.057	1.218	0.946	16.26
1	14.53	2.094	2.492	0.478	15.92
2	30.30	4.229	5.257	0.236	15.92
4	60.37	8.391	10.195	0.119	15.92
8	118.32	16.559	20.260	0.060	15.92

Fresh v129 averages 5.766 rollouts per prefix. Mixing measured four- and eight-rollout budgets using the historical escalation rate gives approximately 11.997 seconds per response as an interpolated generation-cost estimate. Available BLIP alarm-head measurements are approximately 1.17–1.29 milliseconds per prefix, excluding VLM forward passes and feature construction. Qwen3's available offline measurements include waiting time; online detector latency was not measured in that record.

G.3 Teacher–student and budget tradeoffs

On OOF records from the same 1,200 LLaVA Train images, a no-rollout hidden student gives dynamic/initial AUROC of 55.55/56.41%. Four fixed rollouts give 77.63/75.01%; four rollouts with sham distillation give 78.56/74.77%; an adaptive four-to-eight policy gives 81.13/78.40%; and the full eight-rollout teacher gives 82.09/79.91%. Lower continuation budgets retain some dynamic ranking ability, but the zero-rollout student has not achieved that advantage.

The adaptive student's Train-OOF R@5 is 17.17%, below its own initial 37.40%. Budget–AUROC tradeoffs therefore do not substitute for low-FAR validation, and costs, AUROC, and recall from different evaluation sets cannot be assembled into one performance curve.

H Alarm-Triggered Visual Grounding Actions

We test the utility of alarms by comparing original responses, reject-and-resample, and a frozen visual-evidence selector on the same alarm set. On Fresh2, CHAIRs decreases from 10.175% for reject-and-resample to 9.775% for the selector, with paired-bootstrap improvement interval $[0.225, 0.625]$ pp and two-sided exact-test $p = 0.000145$. On Fresh3, CHAIRs decreases from 9.875% to 9.475%, with improvement interval $[0.075, 0.726]$ and $p = 0.02590$. Each population uses a frozen selector from the same family of visual-evidence selection methods. These results demonstrate the value of routing alarms to object-grounding interventions.

Table H.1: Alarm-triggered visual grounding. Lower CHAIRs is better.

Population	Original	Reject	Selector	Selector–Reject
Qwen3 Fresh2	10.950	10.175	**9.775**	-0.400
Qwen3 Fresh3	10.400	9.875	**9.475**	-0.400

H.1 Action costs and response quality

Triggered-action latency is approximately 3.70 and 3.40 seconds on the two Qwen3 populations. Object F1 changes by -0.350 and -0.134 pp. These metrics characterize the tradeoff between reduced object hallucination, content coverage, and action cost.

H.2 Other completed action experiments

On the first Qwen3 Fresh set, reject-and-resample reduces CHAIRs from 10.85% to 9.25%, with difference interval $[-2.35, -0.85]$ pp; the visual action does not outperform rejection there. On Fresh2/Fresh3, the selector improves/worsens 17/1 and 31/15 images relative to rejection. Mean response length decreases by only 0.04/0.03 Unicode words, with no empty responses or fixed refusals. Object hallucination decreases with almost unchanged response length; Object F1 and semantic-quality metrics quantify the accompanying content effects.

An InstructBLIP action study uses 399 adaptive Development images. Each image receives one plain, two rejection, and two visual candidates, totaling 1,995 candidates. A frozen visual judge evaluates anonymized candidates in three rotated rounds. Relative to plain responses, hallucination falls from 59.90% to 49.37%, with difference interval $[-15.29, -6.02]$ pp: 70 events are repaired and 28 clean responses harmed, for 42 net repairs. Blinded correct-content utility improves by 33.58 pp, with interval $[26.32, 40.11]$. This Development set has already influenced method design, and the noninferiority test against the preceding action variant did not pass. We report it as a Development study.