CAMERA SCENE GRAMMARS: HISTORY-CONDITIONED SEMANTIC RELIABILITY FOR FIXED-VIEW SURVEILLANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Surveillance streams are monitored through scalar quality scores, yet a single score cannot say which observations remain dependable, whether a degradation has removed a person, introduced a phantom one, or turned a parked vehicle into a moving one. We argue that a fixed camera carries a resource that isolated-image quality assessment lacks, namely its own earlier observations of the same view, and we formulate history-conditioned semantic reliability as judging a degraded stream against that history without any contemporaneous clean reference. A camera scene grammar G_c stores persistent anchors, class-conditioned occupancy, track continuity, and relation context estimated from undistorted history. Comparing the stream with its grammar yields three typed evidence channels, missing expected content, unsupported observations, and inadmissible relations, together with a coverage mask marking where history cannot support a judgment, and a grammar-conditioned student amortizes the comparison without opinion scores. We give an exact consensus error decomposition under shared generator bias, conditional monotonicity and stability guarantees, and an identifiability limit on what history can recover. Under a camera-disjoint protocol on the 50-camera D-CQA corpus with independently annotated object-level errors, the scalar readout reaches 0.903 SRCC, attribution reaches 0.881, 0.902, and 0.858 AUROC for dropped, hallucinated, and misrelated content, and matched-history baselines and swapped or permuted grammars recover only part of the gain, isolating camera-specific semantic structure as its source. We release the graph annotation layer D-CQA-SG.

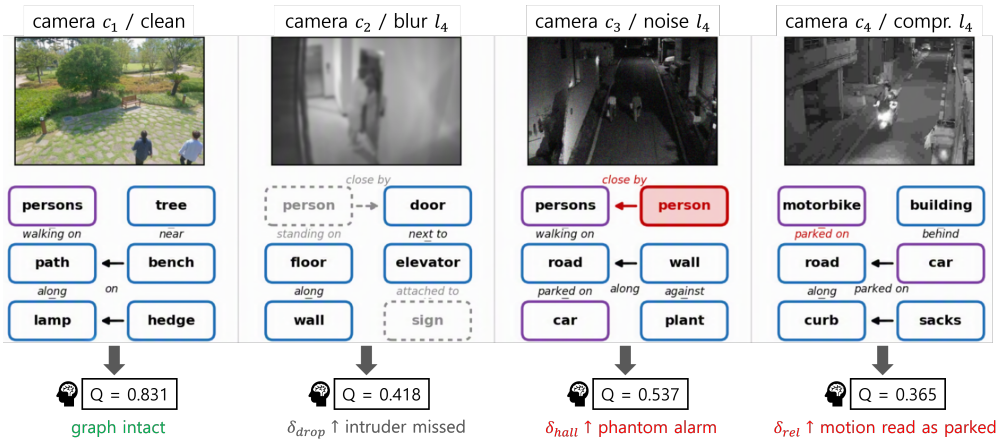

Figure 1: Four fixed cameras, their scene graphs, and the scalar readout Q. Purple marks agents, blue statics, dashed gray drops, red hallucinations and flips. One readout cannot tell the three opposite collapses apart.

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

A surveillance stream can fail in several operationally different ways. Blur can remove a pedestrian from a detector output, sensor noise can introduce a spurious one, and compression can change the relation assigned to an otherwise correct object, as the three degraded cameras in Fig. 1 show. A scalar quality score ranks the overall impairment, and the readouts in the figure do differ, yet no single number, including the Q of our own method, can say whether the loss was a missed intruder, a phantom alarm, or motion read as parked. Even a perfect opinion score leaves an operator unable to tell an unsupported alarm from a missing expected observation, which is why the primary output of this paper is a graph rather than a number.

Existing paradigms do not produce this output. Scalar no-reference regressors compress every semantic consequence of degradation into one number (Wu et al., 2022; 2023; 2024b), and quality localization indicates where the signal is damaged without saying what the damage did to the objects inside (Ying et al., 2020). Robust scene graph generation predicts graphs that survive corruption (Zhang et al., 2024), but its objective is graph prediction rather than an estimate of how much each current observation should be trusted. Single-scene anomaly detection and camera tampering detection do exploit the history of a fixed camera, yet they emit a frame level alarm and never say which object it is about (Ramachandra et al., 2022; Ribnick et al., 2006).

Fixed cameras make the problem different from quality assessment of an isolated image. Earlier observations provide repeated views of the same geometry, the locations where agents usually appear, and the transitions they usually make, a camera-specific source of evidence that is neither a clean copy of the current frame nor a generic prior, and that can be built before deployment and queried without any undistorted view of the current scene. We study the resulting question, which current observations are consistent with this history and where the available evidence is insufficient, as history-conditioned semantic reliability. A camera scene grammar G_c summarizes anchored objects, class-conditioned occupancy, track continuity, and relation context. Comparing the stream with G_c yields a structured output over observed nodes, expected but unmatched node slots, and relations, each with a consistency score, its evidence source, and a validity mask, while a scalar readout supports conventional comparison, and degraded versions of training sources supervise a student that conditions on the grammar and a causal window. Historical regularity alone cannot establish that an observation is false, since a real intruder may enter a previously unused location and an occluder may hide a valid anchor, so the formulation separates measurable inconsistency from semantic correctness and evaluates the latter against independent annotations. Our contributions are fourfold as follows.

- **A camera-conditioned reliability representation.** We make the camera the unit of assessment and define a typed output with explicit expected-node slots and evidence coverage, so that a missing object has a place to be reported and an unsupported judgment is marked rather than guessed.
- **Grammar comparison and compatible distillation.** We define bounded drop, hallucination, and relation channels over the grammar, and a student whose training and deployment inputs obey the same historical constraint, trained from clean and degraded pairs without opinion scores.
- **Theory that states what history can and cannot recover.** We give an exact consensus error decomposition under shared generator bias, conditional monotonicity and stability of the channels, and an observational identifiability limit for objects never observed in the current event.
- **An evaluation that separates history from structure.** Under camera-disjoint folds and independent object-level annotations we compare against baselines given the same history, intervene on the camera prior by swapping and permuting grammars, and test valid rare events, public fixed-camera benchmarks, and moving cameras.

2 RELATED WORK

Perceptual and machine-oriented quality. BRISQUE and NIQE assess quality from natural scene statistics, the latter without any opinion supervision (Mittal et al., 2012; 2013), and learned

regressors such as MUSIQ, CLIP-IQA, FAST-VQA, and DOVER map a frame or clip to a predicted opinion score (Ke et al., 2021; Wang et al., 2023; Wu et al., 2022; 2023), while Q-Align trains a multimodal model on discretized quality levels (Wu et al., 2024b), and a CCTV-specific benchmark relates perceptual quality to image entropy (Han et al., 2025). Localization variants attribute the score spatially (Ying et al., 2020) and pseudo-reference methods aggravate the distortion of the test image itself (Min et al., 2018), but a quality map does not identify a missing object or an inadmissible predicate. A parallel line asks how degradation affects machine consumers rather than viewers, from corruption benchmarks (Hendrycks & Dietterich, 2019; Michaelis et al., 2019) and coding for machines (Duan et al., 2020) to recent machine-oriented quality models that predict downstream loss (Li et al., 2025; Wang et al., 2025). Our distinction is the use of camera history to produce typed object-level evidence, and we compare against a machine-oriented model under matched history.

Scene graphs and scene-specific normality. Scene graph generation predicts objects and pairwise relations (Krishna et al., 2017; Xu et al., 2017; Zellers et al., 2018; Cong et al., 2023), and robust generators such as HiKER-SGG keep predicting well under corruption (Zhang et al., 2024). They operate without a clean image at inference and serve as our components, but their objective is graph prediction rather than an estimate of how much each observation should be trusted. Background modeling learns per-pixel regularity of a fixed view (Stauffer & Grimson, 1999; Zivkovic, 2004), single-scene anomaly detection learns frame level normality from unlabeled footage (Hasan et al., 2016; Liu et al., 2018; Ramachandra et al., 2022), and camera tampering detectors compare the live frame with a stored view (Ribnick et al., 2006; Sağlam & Temizel, 2009). Closest to us, graph-based anomaly detection scores interactions (Mumcu et al., 2025), and C-SARD and BUS-SARD learn scene-specific relation normality from summaries or triplet densities (Lai et al., 2025; Schween et al., 2026). CSG uses related contextual evidence but asks whether degradation has compromised an observation, and since legitimate unusual events also violate scene priors, these methods serve as controls for that ambiguity. The term grammar follows stochastic image grammars (Zhu & Mumford, 2006).

Judgment with vision-language models. Zero-shot vision-language models can rate the plausibility of an image or an extracted graph (Wu et al., 2024a). Their prior is camera agnostic, so a phantom pedestrian beside a wall is plausible in general even when this camera has never observed a person there, and we evaluate a VLM both without history and with the same historical summary CSG receives. Table 4 in Appendix A places all of these families along the facets the introduction motivates.

3 Camera Scene Grammars

3.1 Observation Setting and Output

A camera c observes a single fixed view and emits a stream of frames. A clean historical sample $\mathcal{H}_c$ precedes every evaluation window and is partitioned into a grammar subset $\mathcal{H}_c^G$ and a disjoint calibration subset $\mathcal{H}_c^B$. At time t the system receives the causal window $x_{t-L+1:t}$ and the cached track state T_{t-1}, and no contemporaneous clean observation is available. Clean and degraded pairs exist only for training and for independent evaluation, and a test camera contributes its earlier history for unsupervised adaptation alone. A frozen scene graph generator g_k produces nodes (a_v, b_v, s_v) with class, box, and confidence and relations (v, p, v', s_r) with a predicate p from a compact surveillance vocabulary (Appendix L), and we run K heterogeneous generators to separate what the scene contains from what one model believes. For each generator the candidate set contains the current detections, the persistent anchor slots of the grammar, and the previously observed tracks eligible for continuation. Unmatched slots stay in the output as hypotheses with predicted locations, because a missing object has no current detection to which a score could otherwise be attached, and the output records a consistency score and an evidence mask for every candidate and relation without assigning a confirmed identity to an unmatched slot. The pipeline in Fig. 2 estimates G_c from $\mathcal{H}_c$, compares the stream against it, and reads out object-level reliability.

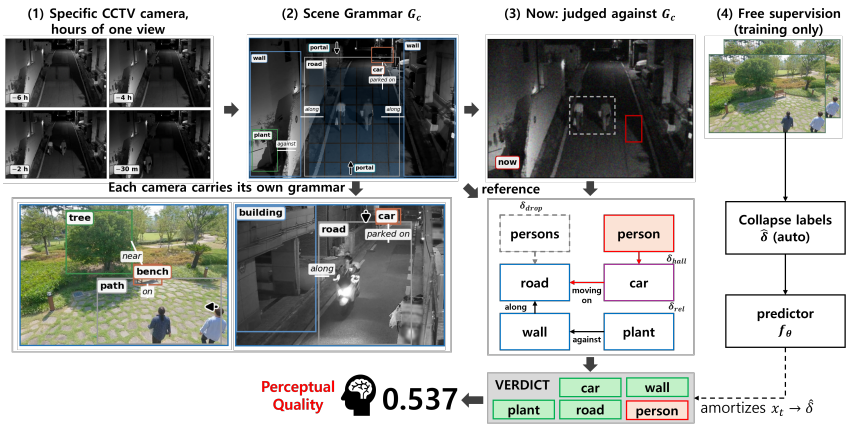

Figure 2: The CSG pipeline. The camera's own history (1) is distilled into its grammar G_c (2), the current window is judged against it (3) into a per-object verdict and the scalar Q, and clean and degraded pairs from training cameras supervise the student (4). Appendix B shows the training and deployment information flow.

3.2 Estimating Camera-Specific Context

The view is fixed, so image coordinates are shared across all of $\mathcal{H}_c$ and semantics can be anchored to them. We write $G_c = (M_c, P_c, A_c, b_c)$ with $M_c = (U_c, \Pi_c)$ holding anchor nodes and anchor relations. Detections in $\mathcal{H}_c^G$ are clustered by compatible class and spatial overlap on a canonical grid $\mathcal{Z}$, and a cluster u becomes an anchor when its persistence $\pi(u)$, the fraction of sampled history frames supporting it, satisfies $\pi(u) \geq \tau_s$ and its class passes a fixed-structure eligibility rule. The rule excludes movable classes such as people and vehicles, so a frequently parked car is never converted into fixed structure. Each anchor stores its class, canonical region ρ_u, and mean confidence $\bar{s}_u$, and $\Pi_c(u, u')$ contains the predicates supported between two anchors at frequency at least τ_r, measured over the historical opportunities in which both endpoints were supported. The **occupancy model** is a distribution over location conditional on agent class,

$$p_c(z \mid a) \;=\; \frac{n_c(a, z) + \alpha}{\sum_{z' \in \mathcal{Z}} n_c(a, z') + \alpha|\mathcal{Z}|}, \tag{1}$$

with $n_c(a, z)$ the historical count of class a in cell z. It is not the probability that an agent at z exists, and $1 - p_c(z \mid a)$ approaches one at perfectly ordinary locations on a fine grid, so we do not use it as a penalty directly. Let q_{ca} be the lower β quantile of the occupancy densities evaluated at the class-a locations observed in $\mathcal{H}_c^B$. The **rarity deficit**

$$h_c(v) \;=\; \max\Big\{0,\; 1 - \frac{p_c(z_v \mid a_v)}{\max(q_{ca}, \epsilon)}\Big\} \tag{2}$$

is zero at any location as common as the calibration history, positive only where this camera has rarely or never seen the class, and unavailable for a class with insufficient calibration support. It measures historical rarity, which is evidence to weigh rather than proof of hallucination. The **agent grammar** A_c holds a portal distribution φ_c of where tracks are born and die, a right-censored track survival function S_c, and conditional relation frequencies. For an eligible track of age d its continuation weight combines $S_c(d + \Delta)/\max(S_c(d), \epsilon)$ with the probability that its predicted position remains in view, and eligibility is decided from T_{t-1} before the current detections are read. An unmatched eligible track therefore signals interrupted support for a previously observed agent, and it cannot diagnose an agent that was missed throughout the event. For dynamic relations the context κ contains the endpoint classes, their anchor or spatial cell, the previous predicate, and a coarse velocity bin, and smoothed frequencies $R_c(p \mid \kappa)$ define the admissible predicates, so a moving rider read as parked is inadmissible even though both predicates are individually plausible. Motion predicates require temporal evidence, and contexts with insufficient history remain unscored.

3.3 TYPED CONSISTENCY EVIDENCE

For each generator we match current detections to anchor and track slots by one-to-one maximum-weight bipartite assignment under class compatibility and $\mathrm{IoU} \geq \eta$, with deterministic tie breaking, so a detection can support at most one expected slot. Each channel $h \in \{\mathrm{drop}, \mathrm{hall}, \mathrm{rel}\}$ is a bounded ratio over its valid candidate set $\mathcal{I}_h^k$ with evidence terms $e_{hi}^k \in [0,1]$ and nonnegative masses a_{hi}^k,

$$\delta_h^k \;=\; \frac{\sum_{i \in \mathcal{I}_h^k} a_{hi}^k \, e_{hi}^k}{\epsilon + \sum_{i \in \mathcal{I}_h^k} a_{hi}^k}. \tag{3}$$

The **drop** channel places unmatched indicators on anchors and eligible tracks, weighted by historical confidence and continuation support. The **hallucination** channel places $h_c(v)$ on agent detections and unmatched support on fixed-structure detections, weighted by current confidence, and it is an unsupported-observation score that includes genuine scene changes. The **relation** channel places an inadmissibility indicator on each predicted predicate between matched endpoints, weighted by relation confidence, with admissibility from Π_c for anchor pairs and R_c for dynamic contexts. Relation queries are formed from eligible ordered endpoint pairs before the predicted predicates are inspected, so a missing predicate or an unavailable context lowers the coverage of the channel rather than counting as evidence that the relation is correct. Masses carry fixed operational class weights ω with $\omega_{\mathcal{A}} > \omega_{\mathcal{S}}$, since a dropped agent is a missed intruder while a dropped static node is a nuisance, and these weights are specified before evaluation and never fitted to opinion scores. The coverage of a channel is the fraction of its candidate queries with valid evidence, an empty channel returns an unavailable mask, and every candidate retains its type, evidence score, and coverage alongside the aggregate. Over the K_h members that support a channel we report the consensus $\bar{\delta}_h$ and, for $K_h \geq 2$, the sample dispersion S_K^2, whose relation to consensus error Proposition 1 makes precise. Finally, the baseline $b_c = (\mu_c, \sigma_c)$ is estimated by running the frozen grammar on $\mathcal{H}_c^B$, and

$$\hat{\delta}_h \;=\; \frac{\bar{\delta}_h - \mu_{ch}}{\max(\sigma_{ch}, \sigma_{\min})}, \qquad Q \;=\; 1 - \sigma\Big(\textstyle\sum_h w_h \hat{\delta}_h\Big), \quad w_h \geq 0, \tag{4}$$

standardizes each channel against the camera's own clean statistics and exposes a scalar for comparison with scalar protocols. A night camera is judged against night statistics with no regime label, Q equals one half at the clean mean by construction and is not a calibrated opinion score, and Q is available only when every required channel is supported.

3.4 GRAMMAR-CONDITIONED DISTILLATION

Running K generators and the full comparison at every frame is accurate but heavy, and the training corpus supplies the labels for a lighter predictor. The full teacher $T(x_{t-L+1:t}, G_c, T_{t-1})$ uses the same information at training and deployment. The student f_θ takes the causal image window, grammar tokens, and expected-node queries, its visual tokens attend to the grammar, and anchor, track, and learned detection queries predict classes, boxes, candidate consistency scores, and relation scores while a channel head predicts the teacher's aggregate evidence. For a training pair (x^0, x^ℓ) of an original and a distorted window from a training camera, the teacher supplies a deployment-consistent target and the clean-to-degraded graph correspondence supplies an additional automatic error label y^{pair} with its own validity mask. Training minimizes

$$\mathcal{L} \;=\; \frac{\big\| m_T \odot (f_\theta^\delta - T^\delta) \big\|_1}{\max(1, \|m_T\|_1)} \;+\; \lambda_d \mathcal{L}_{\mathrm{distill}} \;+\; \lambda_p \mathcal{L}_{\mathrm{pair}} \;+\; \lambda_b \mathcal{L}_{\mathrm{det}}, \tag{5}$$

where the masked distillation term transfers candidate consistency and validity, the pair term predicts the masked automatic error labels, and the detection term combines class cross-entropy with box regression to teacher targets, each normalized by its valid target count. No opinion score enters the objective, and thresholds and stopping rules use training and validation cameras only. Automatic supervision and pretrained generators are still required, so the method is opinion-free rather than annotation-free. At deployment the full path runs the ensemble and Equations (3) to (4), while the fast path evaluates f_θ once and still consults G_c, which is always available because it belongs to the camera. The scalar exists for comparison, the graph is the deliverable.

4 THEORETICAL ANALYSIS

The grammar measures consistency with historical and recent evidence, and this section states what that measurement guarantees and where it stops. Proofs and the scope of each statement are given in Appendix C.

Proposition 1 (Consensus error under shared bias) *For a target deviation d suppose $\delta^k = d + b + \xi_k$ with shared bias b, $\mathbb{E}[\xi_k] = 0$, finite second moments, and $\Sigma_{kl} = \mathbb{E}[\xi_k \xi_l^\top]$. Then*

$$\mathbb{E}\,\|\bar{\delta} - d\|_2^2 \;=\; \|b\|_2^2 + \frac{1}{K^2}\sum_{k,l=1}^{K} \operatorname{tr}(\Sigma_{kl}), \tag{6}$$

which equals $\|b\|_2^2 + v\{\varrho + (1-\varrho)/K\}$ when each residual has total variance v and each distinct pair has covariance trace ϱv, and the sample dispersion satisfies $\mathbb{E}[S_K^2] = (1-\varrho)v$.

Reading. Averaging removes the independent component at rate $1/K$ and nothing else. Shared bias survives averaging entirely, positive correlation slows the decay, and the dispersion measures only the uncorrelated part, so neither low disagreement nor a larger ensemble identifies degradation as the cause of an alert. This is why the ensemble is heterogeneous, since seeds of one architecture share failure modes and hence a large ϱ, and why S_K^2 is reported beside the consensus rather than folded into it (Appendix G).

Proposition 2 (Conditional monotonicity and stability) *Consider a channel over a fixed eligible set I, $D(e,s) = \sum_{i\in I}\omega_i s_i e_i \,/\, (\epsilon + \sum_{i\in I}\omega_i s_i)$ with $\epsilon > 0$, $0 \le e_i, s_i \le 1$, and $\omega_i \ge 0$. For fixed s and ω, if every $\mathbb{E}[e_i(\ell)]$ is non-decreasing in severity ℓ, then $\mathbb{E}[D(e(\ell), s)]$ is non-decreasing. Separately, for fixed e, if $\|s' - s\|_\infty \le \tau$ and the frozen baseline scale satisfies $\sigma \ge \sigma_{\min} > 0$, then*

$$\left|\frac{D(e,s') - \mu}{\sigma} - \frac{D(e,s) - \mu}{\sigma}\right| \;\le\; \frac{2\sum_{i\in I}\omega_i}{\epsilon\,\sigma_{\min}}\,\tau. \tag{7}$$

Reading. With the candidate set and confidence masses frozen, the denominator of Equation (3) is a constant, so monotonicity follows from linearity of expectation without exchanging the expectation of a ratio for a ratio of expectations, and a bounded jitter of the confidences moves the standardized channel by a proportional amount whose constant depends on the total operational weight rather than on an absolute. Both statements are conditional, threshold crossings, changed assignments, scene changes, and camera motion fall outside them, and Fig. 6 in Appendix C tests each premise under a fixed candidate protocol.

Proposition 3 (Observational identifiability limit) *Let O_t contain the camera history and all observations through time t, and let Y_t indicate a transient object that is missed. If two admissible latent scenes produce the same O_t with opposite values of Y_t, any decision rule based only on O_t has worst-case error at least $1/2$ over those scenes.*

Reading. A genuinely empty region and an object never observed during the current event can satisfy this condition, and no occupancy prior resolves it. A previously observed track supplies conditional continuity evidence, but its predicted survival does not establish current presence. The output therefore reports evidence of inconsistency, keeps unsupported cases as unresolved through the coverage mask, and leaves semantic correctness to be measured against independent annotations, which is the protocol of Section 5.

5 EXPERIMENTS

5.1 DATA AND PROTOCOL

All experiments use the D-CQA corpus (Anonymous, 2026; Han et al., 2025), 50 fixed cameras recorded for about one hour each, with 3,000 source frames and clips distorted by five degradation types at four severity levels and annotated with opinion scores (Appendix N). Within each recording the first 30 minutes form $\mathcal{H}_c^G$, the next 10 minutes form $\mathcal{H}_c^B$, and the final 20 minutes supply evaluation sources, so no evaluated frame contributes to its own reference. Cameras are divided into five

Table 1: Human axis and object-level attribution on D-CQA under the camera-disjoint protocol. ft marks fine-tuning on training-camera opinion scores, matched-history rows receive the same history as CSG, AUROC is scored per error mode against independent annotations, a dash marks an output the method does not produce, and brackets give 95% camera-bootstrap intervals on SRCC.

		SRCC	PLCC	KRCC	RMSE	drop	hall	rel
no history	BRISQUE (Mittal et al., 2012)	0.512	0.538	0.356	0.198	–	–	–
	NIQE (Mittal et al., 2013)	0.488	0.509	0.339	0.204	–	–	–
	FAST-VQA (Wu et al., 2022)	0.716	0.728	0.522	0.151	–	–	–
	DOVER (Wu et al., 2023)	0.741	0.752	0.545	0.143	–	–	–
	Q-Align (Wu et al., 2024b)	0.758	0.769	0.561	0.139	–	–	–
	DOVER ft [0.786, 0.821]	0.804	0.815	0.611	0.126	–	–	–
	VLM plausibility	0.583	0.601	0.412	0.182	0.612	0.641	0.596
matched history	DOVER, history standardized	0.834	0.846	0.639	0.118	–	–	–
	Feature distance to history	0.853	0.862	0.657	0.113	–	0.823	–
	Confidence and track consistency	0.821	0.833	0.624	0.121	0.792	0.816	–
	History-conditioned VLM	0.844	0.855	0.648	0.116	0.771	0.817	0.742
	C-SARD (Lai et al., 2025)	0.861	0.870	0.668	0.111	–	0.842	0.803
	BUSSARD (Schween et al., 2026) [0.858, 0.890]	0.875	0.884	0.683	0.106	–	0.863	0.831
	RA-MIQA, hist. standardized (Wang et al., 2025)	0.878	0.887	0.686	0.104	–	–	–
ours	CSG fast [0.857, 0.883]	0.871	0.880	0.678	0.104	0.842	0.867	0.815
	CSG full [0.891, 0.914]	**0.903**	**0.911**	**0.719**	**0.089**	**0.881**	**0.902**	**0.858**

folds of ten, and in each fold 30 cameras train every learned component, 10 select hyperparameters and stopping, and 10 are tested, so every camera is tested exactly once and a test camera contributes nothing but its own history. Because teacher-generated labels cannot validate the teacher, attribution is scored against an independent annotation set of 1,000 sources with four corrupted variants each, labeled by two annotators for missing objects, spurious detections, and incorrect predicates without seeing any method output (Appendix D). Baselines span four families, listed in Table 1. Scalar no-reference models cover classical, learned, and multimodal generations, and the strongest is also fine-tuned on training-camera opinion scores. Matched-history baselines receive exactly $\mathcal{H}_c^G$ and $\mathcal{H}_c^B$ and use them for score standardization, for a camera-specific feature memory, for confidence and track continuity without spatial or relation constraints, as a historical summary for a VLM, or, for C-SARD, BUSSARD, and RA-MIQA, in their documented form (Appendix E). The machine axis correlates predicted degradation with signed clean-to-corrupted losses of an independently chosen detector and tracker on contiguous clips (Zhao et al., 2024; Cao et al., 2023), and is repeated on the static public benchmarks MOT20 and UA-DETRAC (Dendorfer et al., 2020; Wen et al., 2020) under the corruption suite of Hendrycks & Dietterich (2019). Learned components are averaged over three seeds, intervals come from a paired camera bootstrap, and hyperparameters are in Appendix L.

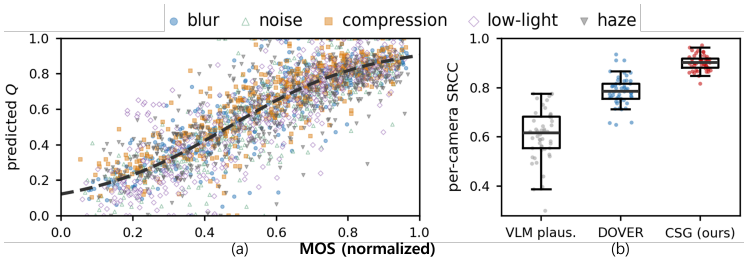

Figure 3: Main results. (a) Predicted quality against MOS on D-CQA by distortion type. (b) Per-camera SRCC over the 50 cameras.

5.2 Human Axis and Object-Level Attribution

On the human axis, CSG full reaches 0.903 SRCC, the fast path retains 0.871 without the ensemble at deployment, and the strongest baseline without history, DOVER fine-tuned on training cameras,

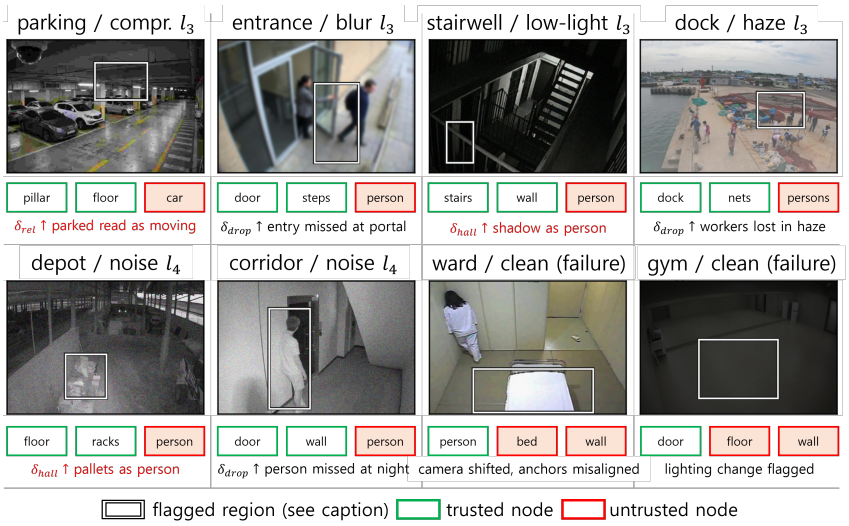

Figure 4: Verdicts on eight cameras, six successes across the three evidence channels and two failures, a camera shift and a lighting change.

reaches 0.804 with non-overlapping intervals in Table 1. Giving baselines the same history closes part of that gap. Standardizing DOVER against its own clean statistics per camera lifts it to 0.834, a camera-specific feature memory reaches 0.853, and the scene-specific relation models C-SARD and BUSSARD reach 0.861 and 0.875, with RA-MIQA at 0.878 the strongest, so roughly three quarters of the original margin is attributable to history itself and the remainder to what CSG does with it, a remainder that is significant at $p < 0.001$ under the paired bootstrap. The right block of the table is the capability claim. Attribution is undefined for scalar methods, the relation models score relations but not dropped or hallucinated nodes, and the VLM can be prompted per node but reaches only 0.612 and 0.641 AUROC without history and 0.771 and 0.817 with it, because a generic prior cannot hold what is normal for this camera. CSG full attains 0.881, 0.902, and 0.858 AUROC on dropped, hallucinated, and misrelated content against annotations that no generator produced. The per-camera view in Fig. 3 (b) shows where the margin over history-free methods comes from, since they rank frames reasonably within a camera but carry a per-camera score offset that degrades the pooled ranking, whereas CSG standardizes against b_c and its pooled and per-camera correlations nearly coincide. The scatter in Fig. 3 (a) is widest under low light, where graph extraction is least stable.

Table 2: Machine axis. Spearman correlation between predicted degradation and signed detector (mAP) and tracker (IDF1) loss on contiguous clips.

	D-CQA		MOT20		UA-DETRAC	
	mAP	IDF1	mAP	IDF1	mAP	IDF1
DOVER ft	0.672	0.649	0.615	0.591	0.637	0.604
DOVER, history standardized	0.761	0.735	0.706	0.681	0.724	0.697
History-conditioned VLM	0.769	0.742	0.711	0.684	0.728	0.701
BUSSARD	0.807	0.788	0.751	0.729	0.769	0.744
RA-MIQA, history standardized	0.828	0.801	0.774	0.752	0.792	0.769
CSG fast	0.812	0.789	0.771	0.748	0.786	0.759
CSG full	**0.845**	**0.826**	**0.803**	**0.781**	**0.821**	**0.794**

5.3 Machine Axis

On predicting downstream loss, CSG full correlates at 0.845 with detector mAP loss and 0.826 with tracker IDF1 loss on D-CQA in Table 2, above the machine-oriented RA-MIQA at 0.828 even after that model is standardized against camera history, because the evidence channels measure exactly

the failure modes that break detection and tracking. The ordering is preserved on MOT20 and UA-DETRAC, where the cameras are public and the degradations are the standard corruption suite, so the result does not depend on the corpus we authored, and there ground-truth boxes define dropped nodes exactly, giving a drop AUROC of 0.852 and 0.866 independent of any generator (Appendix F).

5.4 Camera Correspondence and Ablations

The fixed-view hypothesis is tested most cleanly by holding the test observations fixed and intervening on the prior, reported in the left block of Table 3. Replacing G_c with another camera's grammar drops SRCC to 0.754 and hallucination AUROC to 0.736, which tests camera correspondence, permuting the anchored regions and occupancy support while keeping every stored value drops them to 0.812 and 0.781, which tests spatial organization, and removing camera conditioning from a retrained student lands at 0.798, so both the identity and the layout of the history are load-bearing. Translating the test view without rebuilding the grammar costs 0.097 SRCC, while applying the same transform to history and view recovers to 0.889, which separates sensitivity to incorrect correspondence from a generic loss under resampling. The component ablation in the right block shows a clean division of labor. Anchors carry the drop channel, occupancy carries the hallucination channel, the dynamic relation context R_c carries the relation channel, and standardization b_c costs the most SRCC because unstandardized deviations mix day and night regimes across cameras. Operational weights barely move SRCC yet raise drop AUROC by 0.022, so they serve the verdict rather than the opinion score. Appendix G adds the history length, which saturates by twenty minutes, the moving-camera transfer, where the advantage vanishes, and the ensemble composition predicted by Proposition 1.

Table 3: Interventions on the camera prior with test observations held fixed (left) and component ablations (right). AUROC per error mode.

intervention	SRCC	drop	hall	rel	ablation	SRCC	drop	hall	rel
correct camera grammar	**0.903**	**0.881**	**0.902**	**0.858**	CSG full	**0.903**	**0.881**	**0.902**	**0.858**
other camera's grammar	0.754	0.719	0.736	0.708	w/o M_c (anchors)	0.856	0.742	0.874	0.769
spatially permuted grammar	0.812	0.772	0.781	0.743	w/o P_c (occupancy)	0.869	0.869	0.781	0.851
no camera conditioning	0.798	0.754	0.762	0.731	w/o A_c (tracks)	0.881	0.823	0.893	0.849
misaligned test view	0.806	0.746	0.789	0.738	w/o R_c (dynamic rel.)	0.887	0.879	0.899	0.796
realigned history and view	0.889	0.865	0.887	0.842	w/o b_c (standardize)	0.837	0.851	0.868	0.827
					$K = 1$ (no ensemble)	0.874	0.833	0.851	0.808
					uniform ω	0.895	0.859	0.897	0.856

5.5 Rare Events and Qualitative Results

A reliability estimator must not accuse valid rare events. On a challenge set of 36 clean clips with first arrivals, agents in previously unused locations, occlusion, moved objects, lighting changes, and camera displacement, with thresholds fixed on validation cameras, CSG rejects 8.4% of valid events and marks 6.1% unresolved through the coverage mask, against 21.7% rejected by BUSSARD and 12.9% by the confidence and track baseline, neither of which can abstain (Appendix H). Per-object verdicts across eight cameras are collected in Figure 4, whose six successes cover all three channels, from a parked car read as moving to a pedestrian dropped on a portal while an eligible track was active. The two failures share a root, since a small shift of the ward camera misaligns the anchors and a lighting change in the gym is flagged during the transient because b_c absorbs stable regimes rather than abrupt ones. Appendix J adds six cameras and a legitimate scene change flagged as collapse, Appendix K reports a gated sliding-window update that removes most such false flags, and Appendix M reports deployment cost.

6 Conclusion

This paper turned the history of a fixed surveillance camera into a reference for semantic reliability. A camera scene grammar preserves missing, unsupported, and misrelated content as separate evidence channels with explicit coverage, is amortized by a student trained without opinion scores, and its gain is isolated from history alone through matched baselines and interventions on the camera prior. The approach assumes a stable view, so displacement, lighting transitions, and scene changes are flagged like corruption, and an online grammar update (Appendix K) is our next step.

AI USE STATEMENT

Generative AI tools were used only for proofreading and language editing, including grammar correction and minor improvements to the readability of the manuscript. They were not used for the development of research ideas, methodologies, experiments, code, data analysis, result interpretation, or the creation of figures and tables. All scientific content and conclusions were developed independently by the authors, and all AI-assisted language edits were reviewed and verified by the authors. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work involves surveillance video, which raises privacy and dual-use concerns. The study assesses the reliability of scene observations rather than identifying or profiling individuals. The source footage in D-CQA was collected from cameras operated with the authorization of the corresponding site owners, faces and license plates are blurred at source, and no identity labels or person-specific annotations are included in D-CQA-SG or the grammars. A grammar stores spatial structure and aggregate activity statistics, and it is not designed for identification or re-identification. Removing direct identifiers does not guarantee anonymity, since location and activity patterns can remain sensitive, so the release documents authorization, access terms, and retention. A rare valid event must not be treated as false solely because it violates historical regularity, which is why the evaluation includes false alarms on clean unusual events and the coverage of unresolved outputs. The dataset and method are intended for research, and their use should comply with applicable privacy and data-protection requirements and receive ethical review before deployment.

REPRODUCIBILITY STATEMENT

Code, grammar estimation scripts, the D-CQA-SG annotation layer, the independent annotation set, camera and source split manifests, corruption configurations, checkpoints, and scripts that regenerate every table and figure will be released upon publication. Section 5.1 describes the protocol, Appendix D the annotation rules, Appendix E the information and tuning restrictions of every baseline, Appendix L the hyperparameters, ensemble, student, and VLM protocol, and Appendix C the complete proofs.

REFERENCES

Anonymous. D-CQA: A distorted CCTV quality assessment database, 2026. Under review. Self-citation anonymized for double-blind review.

Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric SORT: Rethinking SORT for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9686–9696, 2023.

Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. RelTR: Relation transformer for scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11169–11183, 2023. doi: 10.1109/TPAMI.2023.3268066.

Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. MOT20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020.

Ling-Yu Duan, Jiaying Liu, Wenhan Yang, Tiejun Huang, and Wen Gao. Video coding for machines: A paradigm of collaborative compression and intelligent analytics. *IEEE Transactions on Image Processing*, 29:8680–8695, 2020.

Yujin Han, Jiwoo Kang, Sanghoon Lee, and Taewan Kim. Understanding perceptual quality in CCTV images: A benchmark dataset and entropy-based insights. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2025.

Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. In *CVPR*, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.

Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The konstanz natural video database (KoNViD-1k). In *International Conference on Quality of Multimedia Experience*, pp. 1–6, 2017. doi: 10.1109/QoMEX.2017.7965673.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *ICCV*, 2021.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. doi: 10.1007/s11263-016-0981-7.

Yu-Chen Lai, Motoharu Sonogashira, Itthisak Phueaksri, and Yasutomo Kawanishi. Scene-specific anomalous relationship detection using scene graph summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3995–4003, 2025.

Chunyi Li, Yuan Tian, Xiaoyue Ling, Zicheng Zhang, Haodong Duan, Haoning Wu, Ziheng Jia, Xiaohong Liu, Xiongkuo Min, Guo Lu, Weisi Lin, and Guangtao Zhai. Image quality assessment: From human to machine preference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7570–7581, 2025.

Rongjie Li, Songyang Zhang, and Xuming He. SGTR: End-to-end scene graph generation with transformer. In *CVPR*, 2022.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *ECCV*, 2024.

Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection – a new baseline. In *CVPR*, 2018.

Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

Xiongkuo Min, Guangtao Zhai, Ke Gu, Yutao Liu, and Xiaokang Yang. Blind image quality estimation via distortion aggravation. *IEEE Transactions on Broadcasting*, 64(2):508–517, 2018.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. doi: 10.1109/TIP.2012.2214050.

Anish Mittal, Rajiv Soundararajan, and Alan Conrad Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.

Furkan Mumcu, Michael J. Jones, Yasin Yilmaz, and Anoop Cherian. ComplexVAD: Detecting interaction anomalies in video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pp. 1003–1012, 2025.

Bharathkumar Ramachandra, Michael J. Jones, and Ranga Raju Vatsavai. A survey of single-scene video anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (5):2293–2312, 2022.

Evan Ribnick, Stefan Atev, Osama Masoud, Nikolaos Papanikolopoulos, and Richard Voyles. Real-time detection of camera tampering. In *IEEE International Conference on Video and Signal Based Surveillance (AVSS)*, 2006.

Ali Sağlam and Alptekin Temizel. Real-time adaptive camera tamper detection for video surveillance. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2009.

Melissa Schween, Mathis Kruse, and Bodo Rosenhahn. BUSSARD: Normalizing flows for bijective universal scene-specific anomalous relationship detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2):612–627, 2019. doi: 10.1109/TIP.2018.2869673.

Chris Stauffer and W. Eric L. Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 246–252, 1999.

Jianyi Wang, Kelvin C.K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *AAAI*, 2023.

Xiaoqi Wang, Yun Zhang, and Weisi Lin. Image quality assessment for machines: Paradigm, large-scale database, and models, 2025. URL `https://arxiv.org/abs/2508.19850`.

Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *Computer Vision and Image Understanding*, 193:102907, 2020.

Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. FAST-VQA: Efficient end-to-end video quality assessment with fragment sampling. In *European Conference on Computer Vision*, 2022.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-Bench: A benchmark for general-purpose foundation models on low-level vision. In *ICLR*, 2024a.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54015–54029. PMLR, 2024b.

Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017.

Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, 2018.

Ce Zhang, Simon Stepputtis, Joseph Campbell, Katia Sycara, and Yaqi Xie. HiKER-SGG: Hierarchical knowledge enhanced robust scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28233–28243, 2024.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16965–16974, 2024.

Song-Chun Zhu and David Mumford. A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision*, 2(4):259–362, 2006.

Zoran Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *International Conference on Pattern Recognition*, volume 2, pp. 28–31, 2004. doi: 10.1109/ICPR.2004.1333992.

APPENDIX

Appendix A compares CSG with related families, Appendix B shows the training and deployment information flow, Appendix C gives proofs and their empirical checks, Appendix D documents the independent annotation set, Appendix E the matched-history baselines, Appendix F the public benchmarks, Appendix G the history length, moving-camera transfer, and ensemble composition, Appendix H the rare-event challenge set, Appendix I and J further visual results, Appendix K a grammar update rule, Appendix L implementation, Appendix M runtime, Appendix N the corpus, and Appendix O limitations.

A FACET COMPARISON WITH RELATED FAMILIES

The families reviewed in Section 2 are placed next to CSG in Table 4. Scene-specific relation normality is the closest relative, since it is stateful, reference-free, and label-free, but it scores relations rather than objects and treats every departure from normality alike, whereas CSG separates three evidence types, reports coverage, and evaluates against degradation-induced error rather than anomaly.

Table 4: Facet comparison. What means the method says which content fails, Trust means per-object reliability, No-ref means no clean reference at deployment, Free means no opinion labels, and Stateful means a per-camera model.

	What	Trust	No-ref	Free	Stateful	Output
Scalar NR-VQA (Wu et al., 2022; 2023; 2024b)	✗	✗	✓	✗	✗	score
Quality localization (Ying et al., 2020)	*where*	✗	✓	✗	✗	score, map
Machine-oriented IQA (Wang et al., 2025)	✗	✗	✓	✓	✗	score
Robust SGG (Zhang et al., 2024)	✓	✗	✓	✓	✗	graph
Background modeling (Stauffer & Grimson, 1999)	✗	✗	✓	✓	✓	pixel mask
Single-scene anomaly det. (Liu et al., 2018)	*when*	✗	✓	✓	✓	frame score
Scene-specific relation normality (Schween et al., 2026)	*rel.*	✗	✓	✓	✓	relation score
Camera tampering det. (Sağlam & Temizel, 2009)	✗	✗	✓	✓	✓	alarm
VLM plausibility (Wu et al., 2024a)	✓	✗	✓	✓	✗	text, score
CSG (ours)	✓	✓	✓	✓	✓	graph, coverage, score

B TRAINING AND DEPLOYMENT INFORMATION FLOW

The schematic in Fig. 5 makes explicit that clean and degraded pairs exist only on training cameras, that both the teacher and the student consume earlier history of the same camera, that no contemporaneous clean frame is used during evaluation, and that expected-node slots and coverage remain part of the structured output on both paths.

C PROOFS AND SCOPE OF THE GUARANTEES

The propositions concern the observable information and the score aggregation of Section 3, and they do not assume that the ensemble supplies ground truth.

Proof of Proposition 1. Let $\bar{\xi} = K^{-1} \sum_k \xi_k$, so $\mathbb{E}[\bar{\xi}] = 0$ and

$$\mathbb{E}\|\bar{\delta} - d\|_2^2 = \mathbb{E}\|b + \bar{\xi}\|_2^2 = \|b\|_2^2 + 2b^\top \mathbb{E}[\bar{\xi}] + \mathbb{E}\|\bar{\xi}\|_2^2 = \|b\|_2^2 + \frac{1}{K^2} \sum_{k,l=1}^{K} \operatorname{tr}\big(\mathbb{E}[\xi_k \xi_l^\top]\big), \tag{8}$$

where the trace identity follows from $\xi_k^\top \xi_l = \operatorname{tr}(\xi_l \xi_k^\top)$. If diagonal covariance traces equal v and off-diagonal traces equal ϱv, the sum has K diagonal and $K(K-1)$ off-diagonal terms, giving

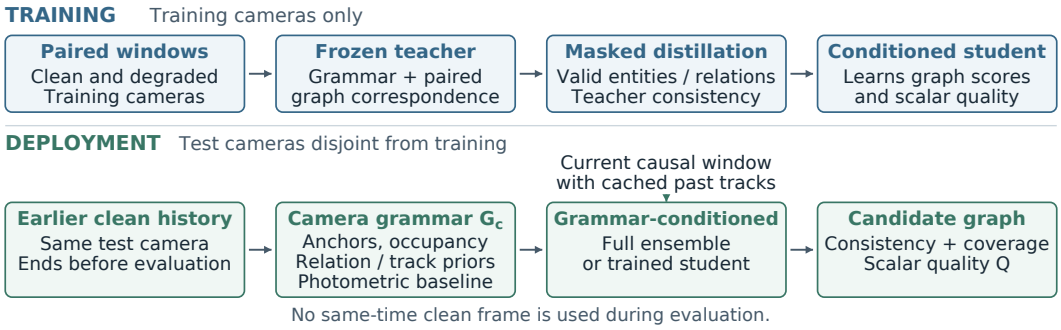

Figure 5: Information flow. Training uses paired windows of training cameras only, deployment uses the causal window, cached tracks, and the earlier history of the test camera, and the full ensemble and the trained student produce the same structured output.

$\{Kv + K(K-1)\varrho v\}/K^2 = v\{\varrho + (1-\varrho)/K\}$. No independence is used in the general identity. For the dispersion, $S_K^2 = (K-1)^{-1}\sum_k \|\delta^k - \bar{\delta}\|_2^2 = (K-1)^{-1}\sum_k \|\xi_k - \bar{\xi}\|_2^2$, the bias cancels, and $\mathbb{E}[S_K^2] = (K-1)^{-1}\{\sum_k \mathbb{E}\|\xi_k\|^2 - K\mathbb{E}\|\bar{\xi}\|^2\} = (1-\varrho)v$. □

Under the equicorrelation specialization a valid covariance requires $-1/(K-1) \le \varrho \le 1$. Positive correlation reduces the measured dispersion without removing the correlated component of consensus error, so the dispersion cannot estimate total error and is reported as a disagreement statistic. The identity is an error decomposition relative to a specified target, not a procedure for estimating the unknown bias without independent evidence, which is why attribution is scored against independent annotations.

Proof of Proposition 2. For monotonicity let $A = \epsilon + \sum_{i\in I}\omega_i s_i$, a positive constant independent of severity because the eligible set, weights, and confidences are fixed. Linearity of expectation gives $\mathbb{E}[D(e(\ell), s)] = A^{-1}\sum_{i\in I}\omega_i s_i\,\mathbb{E}[e_i(\ell)]$, a nonnegative combination of non-decreasing functions, and no independence between error terms is needed. For stability write $N = \sum_i \omega_i s_i e_i$, $A = \epsilon + \sum_i \omega_i s_i$, $W = \sum_i \omega_i$, and primes for the quantities under s'. The perturbation condition gives $|N' - N| \le W\tau$ and $|A' - A| \le W\tau$, and since $0 \le N/A \le 1$ and $A, A' \ge \epsilon$,

$$\left|\frac{N'}{A'} - \frac{N}{A}\right| \;\le\; \frac{|N'-N|}{A'} + \frac{N}{A}\,\frac{|A-A'|}{A'} \;\le\; \frac{2W\tau}{\epsilon}. \tag{9}$$

Subtracting a common frozen mean leaves the difference unchanged, and dividing by $\sigma \ge \sigma_{\min}$ gives Equation (7). If I is empty or $W = 0$, both scores are zero and the bound holds. □

The constant depends on total operational weight and is not absolute in the graph size. Matching, visibility decisions, candidate construction, and relation selection must remain fixed during the comparison, and if any of them changes when confidence is perturbed the complete-system response requires empirical evaluation, which is what the fixed-candidate protocol of Fig. 6 provides.

Proof of Proposition 3. Let w_0 and w_1 be the two admissible worlds with the same observation o and $Y_t(w_0) = 0$, $Y_t(w_1) = 1$. A randomized rule returns one with probability $q(o) \in [0,1]$, so its error probabilities in the two worlds are $q(o)$ and $1 - q(o)$ and $\max\{q(o), 1-q(o)\} \ge 1/2$. A deterministic rule is wrong in at least one world, and the same conclusion holds when the worlds induce an identical distribution of observations by integrating $q(O_t)$ under that distribution. □

The proposition is conditional on the observation model admitting such an ambiguity and is not a claim that every corrupted frame is uninformative. A previously observed track, an independently established visibility condition, or another sensor may constrain the admissible worlds. A historical spatial prior alone does not verify an object that was never observed during the current event, and a continuity score for an active track retains its interpretation as conditional evidence that includes legitimate exit or occlusion.

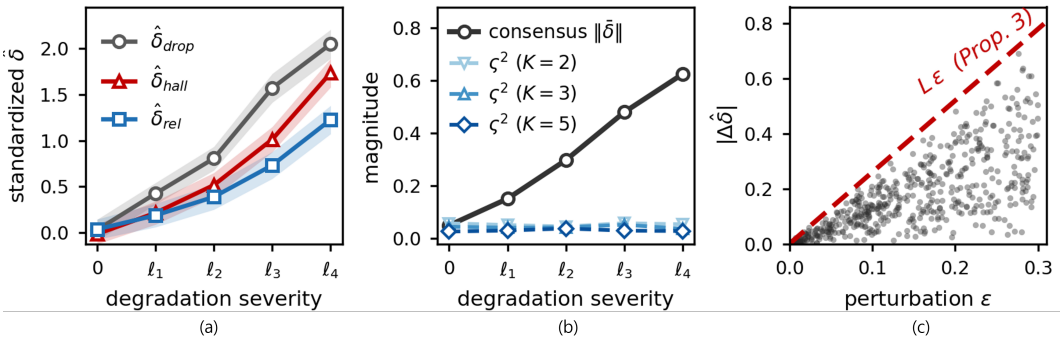

Figure 6: Empirical checks under a fixed candidate protocol. (a) Standardized channels against severity with eligible sets and confidence masses frozen, the premise of Proposition 2. (b) Consensus against dispersion S_K^2 for $K = 2, 3, 5$, where the dispersion stays flat across severity while the consensus rises, consistent with Proposition 1. (c) Response of the standardized channel to injected confidence jitter against the bound of Equation (7).

D Independent Annotation Set

The annotation set contains 1,000 pristine sources balanced at 20 per camera, each with four corrupted variants chosen from the twenty corruption conditions by a balanced incomplete design, so that every condition appears four times per camera, for 4,000 source-variant pairs. Two annotators label the source objects, classes, spatial support, and eligible relations, and then the corresponding corrupted observations, without seeing method scores or teacher graphs, and a separate adjudication step resolves disagreement. A drop is a source object that the evaluated graph fails to recover under the declared matching rule, a hallucination is a predicted object with no corresponding source object, and a relation error is an incorrect predicate between eligible matched endpoints. An unresolved state covers cases in which the corrupted observation cannot support a defensible judgment, and semantic category errors are distinguished from localization errors. Source annotations supply the candidate universe for missing objects, while deployment candidates remain restricted to grammar anchors and supported track hypotheses, so attribution is reported both conditionally on eligible candidates and as coverage of all source objects. Agreement before adjudication is $\kappa = 0.81$ for objects and 0.74 for relations, prevalence is 0.183, 0.097, and 0.141 for drops, hallucinations, and relation errors, 4.6% of candidates are unresolved and excluded, and 7.2% of source objects fall outside the deployment candidate universe and count toward the coverage limitation of Proposition 3. AUPRC of CSG full on the three modes is 0.742, 0.688, and 0.611, and camera-macro AUROC over the 46 cameras with both outcome classes is within 0.008 of the pooled values in Table 1. Teacher-generated training labels are stored separately from these evaluation labels.

E Matched-History Baselines

Every matched-history baseline receives exactly $\mathcal{H}_c^G$ and $\mathcal{H}_c^B$, and any per-camera calibration is fitted on $\mathcal{H}_c^B$ with the direction and variance floor fixed in advance. History-standardized DOVER and RA-MIQA estimate their clean score location and scale per camera and are standardized as in Equation (4). The feature-distance baseline builds a camera memory of frozen DINOv2 features from $\mathcal{H}_c^G$ and scores a window by its nearest-neighbor distance, with a node version that compares region features of the same candidates used by the graph methods. The confidence and track baseline receives the same graph predictions and trajectories as CSG but omits spatial and relation constraints, combining detection confidence and continuity with weights fitted on validation cameras. C-SARD and BUSSARD use their released scene-specific relation scoring on the common vocabulary and historical partition, their relation anomaly scores are evaluated directly, and the node and scalar reductions are fixed on validation cameras. The history-conditioned VLM receives a deterministic summary of the camera history under a fixed context budget of 16 images and the expected-node slots, and its camera-agnostic counterpart receives only the current window and graph. Opinion labels never tune any matched-history baseline.

Table 5: Machine axis and node attribution on public fixed-camera benchmarks.

	MOT20				UA-DETRAC			
	mAP	IDF1	drop	hall	mAP	IDF1	drop	hall
RA-MIQA, hist. std.	0.774	0.752	–	–	0.792	0.769	–	–
History-conditioned VLM	0.711	0.684	0.694	0.716	0.728	0.701	0.708	0.731
CSG fast	0.771	0.748	0.819	0.833	0.786	0.759	0.834	0.851
CSG full	**0.803**	**0.781**	**0.852**	**0.869**	**0.821**	**0.794**	**0.866**	**0.882**

F Public Fixed-Camera Benchmarks

The machine-axis protocol is repeated on the four training sequences of MOT20, all recorded by fixed elevated cameras (Dendorfer et al., 2020), and on twenty UA-DETRAC traffic sequences from distinct positions (Wen et al., 2020). The first two thirds of each sequence form the history, split in the same 3 to 1 ratio into grammar and calibration subsets, and the remainder is evaluated under the fifteen corruptions at five severities of Hendrycks & Dietterich (2019). Because ground-truth boxes exist, a dropped node is a ground-truth object with no detection above the matching threshold and a hallucinated node is a detection matching no ground-truth object, so node AUROC on these benchmarks is independent of any generator consensus. Table 5 preserves every ordering of Table 2, and the lower absolute correlations on MOT20 reflect its crowd density, under which the occupancy density is nearly uniform over the walkable area and h_c loses resolution.

Table 6: History length ablation with $\mathcal{H}_c^B$ fixed.

history	anchors / camera	tracks / camera	SRCC	drop	hall	rel
30 min	11.4	212	**0.903**	**0.881**	**0.902**	**0.858**
20 min	11.1	143	0.896	0.874	0.897	0.851
10 min	10.2	71	0.878	0.851	0.879	0.829
5 min	8.7	36	0.842	0.803	0.846	0.791
2 min	6.1	14	0.791	0.742	0.788	0.733

G History Length, Moving Cameras, and Ensemble Composition

Varying the length of $\mathcal{H}_c^G$ with $\mathcal{H}_c^B$ held fixed, every history-dependent quantity is rebuilt at each condition, and Table 6 reports the supported anchors and tracks alongside accuracy. Ten minutes recover 0.878 SRCC and twenty minutes are within 0.007 of the full history, so the grammar saturates quickly and suits a newly installed camera, while two minutes do not reveal the portals and occupancy of a scene. Transferred to the moving-camera corpora KoNViD-1k and LIVE-VQC (Hosu et al., 2017; Sinno & Bovik, 2019), where no persistent view exists and the grammar degenerates to a single-window prior, the CSG advantage of $+0.025$ over the strongest matched baseline on D-CQA turns into -0.013 and -0.014 in Table 7. This comparison cannot isolate camera motion from every other difference between corpora, which is why the interventions of Table 3 carry the causal claim, but it confirms that the method offers nothing where its premise fails. Table 8 compares the hetero-

Table 7: Transfer to moving-camera corpora. Δ is CSG minus the strongest listed baseline.

	FAST-VQA	DOVER	RA-MIQA	CSG	Δ
D-CQA (fixed)	0.716	0.741	0.878	**0.903**	$+0.025$
KoNViD-1k (moving)	**0.784**	0.779	0.780	0.771	-0.013
LIVE-VQC (moving)	0.749	**0.752**	0.751	0.738	-0.014

geneous ensemble with five seeds of one architecture at equal cost. The homogeneous dispersion rises with severity, so its members fail together and S_K^2 leaks severity, whereas the heterogeneous dispersion stays flat as Proposition 1 predicts for a smaller ϱ.

Table 8: Heterogeneous versus homogeneous ensemble at $K = 5$. The last column is the SRCC between S_K^2 and the severity level.

	SRCC AUROC	drop AUROC	hall SRCC(S_K^2, severity)	
five seeds of RelTR	0.884	0.857	0.869	0.42
five heterogeneous generators (default)	**0.903**	**0.881**	**0.902**	0.06

H Rare-Event Challenge Set

The challenge set contains 36 clean clips collected from the D-CQA cameras independently of any model score, covering valid first arrivals, agents in previously unused locations, prolonged stationary agents, partial occlusion, relocated objects, legitimate scene changes, abrupt lighting changes, and camera displacement, each annotated for whether the observation itself is valid. Thresholds are fixed on validation cameras. The rejection rate of valid events, false alarms per clip, and the fraction of unresolved outputs are reported in Table 9, and selective processing at matched retained coverage is compared against random selection, detector confidence, and history-standardized quality. CSG is the only method that can abstain, and its unresolved outputs concentrate on first arrivals in previously unused locations, which is the case Proposition 3 identifies.

Table 9: Rare-event challenge set at thresholds fixed on validation cameras.

	valid events rejected	false alarms / clip	unresolved fraction
DOVER, history standardized	0.163	0.52	0.000
Confidence and track consistency	0.129	0.44	0.000
BUSSARD	0.217	0.58	0.000
CSG full	**0.084**	**0.31**	0.061

I Learned Grammars

The occupancy densities and portals of eight cameras are drawn over a representative frame in Fig. 7. A street camera concentrates agent mass on the sidewalk and crosswalk, a concourse spreads it over



Figure 7: Learned grammars on eight cameras (top left to bottom right: street, concourse, yard, parking, playground, shop, night storage, entrance). Heat is the occupancy density p_c, black segments mark portals.

the floor, a parking garage confines it to the driving lane, and a night storage room keeps it near the door, while the portals mark the borders and doorways through which tracks are born and die. Because the grammar is estimated from unlabeled history with frozen generators, these maps are obtained for a new camera without annotation.

J Extended Qualitative Results

Six further cameras complement the main grid in Fig. 8, where a clerk is dropped under blur, a phantom pedestrian appears beside a crowd, a moving car is read as parked, structure is lost in low light, and sacks are hallucinated as a person in a night storage room. The last panel is a third failure type, a legitimate rearrangement of the playroom flagged because the grammar has never seen the new boxes. Together with the main paper this completes the taxonomy of grammar staleness, scene change, camera shift, and lighting transition, which Appendix K addresses.

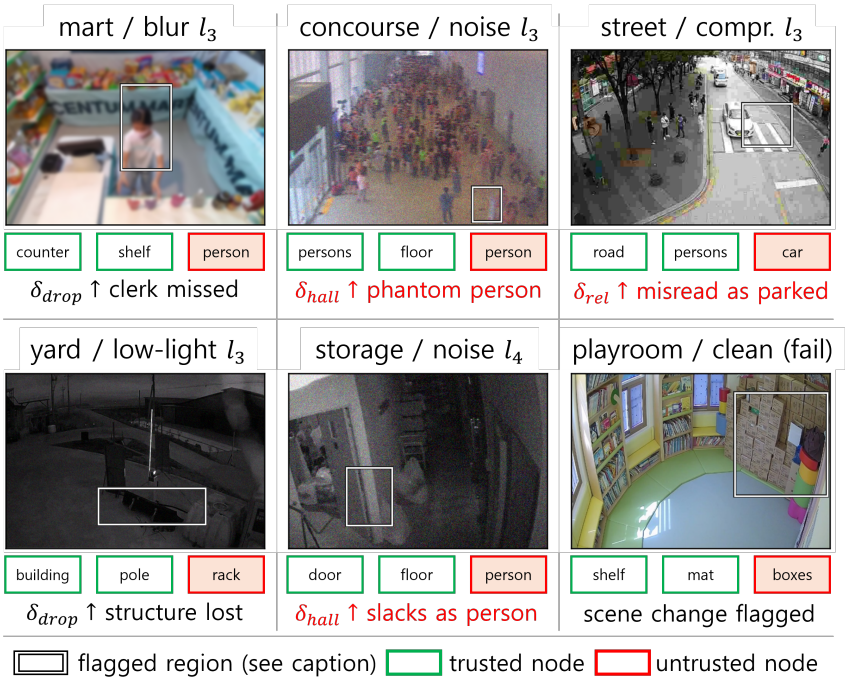

Figure 8: Additional verdicts on six cameras, the last a scene-change failure.

K Grammar Staleness and a Gated Update

The three failure types share a root, a grammar estimated once and never revised. As a first remedy we re-estimate M_c, P_c, and A_c on a sliding window of the most recent 30 minutes of frames whose Q exceeds 0.8, so only frames the current grammar already trusts are admitted, and we gate the update by a change-point test on the drop channel, so that a sustained rise of $\hat{\delta}_{\mathrm{drop}}$ over 90 seconds with a flat hallucination channel is treated as a scene change rather than a corruption. Evaluated on the transition clips of the challenge set, the gated update lowers the false flag rate after the transition settles from 0.412 to 0.097 while SRCC on the regular D-CQA evaluation moves from 0.903 to 0.901, whereas an ungated update reaches 0.175 but falls to 0.884 SRCC because a slowly worsening corruption is absorbed into the grammar. A principled online rule that distinguishes the two cases is the direction named in Section 6.

L Implementation Details

Grammar construction. Historical detections above a threshold selected on validation cameras are mapped to normalized image coordinates, and a deterministic assignment matches each frame to accumulated anchor clusters by compatible class and maximum total IoU. A cluster is eligible as fixed structure only if its class has fixed-structure semantics and its normalized centers and sizes are stable over time. Each generator g_k keeps its own grammar statistics in a shared vocabulary and coordinate system, so no member's confidence is compared with another member's baseline, and the notation G_c suppresses this indexing. Tracks are initialized only from earlier detections and cached before the current window is processed, survival times are right-censored at recording ends, and candidate tracks expire after a validation-selected timeout. The quantile q_{ca} requires at least 30 supported class observations in $\mathcal{H}_c^B$, and a class below that count has its occupancy mask set to zero. The dynamic relation context is $\kappa = (a_i, a_j, z_i, z_j, p_{t-1}, \nu_t)$ with ν_t a velocity bin from earlier tracked positions, smoothed counts define $R_c(p \mid \kappa) = (n_c(p, \kappa) + \alpha_r)/(\sum_{p'} n_c(p', \kappa) + \alpha_r|\mathcal{P}|)$, the admissible set contains predicates above τ_r^A with the most frequent one retained if the set would be empty, a context requires at least 30 opportunities, and backoff removes spatial cells first and the previous predicate second in a fixed order that never removes the temporal-evidence requirement for motion predicates. The vocabulary contains twelve predicates grouped into support (on, standing on, parked on), proximity (close by, next to, near, against), alignment (along, behind, attached to), and motion (moving on, crossing), and the mapping from generator vocabularies to this set, including which predicates require temporal augmentation, is fixed before any experiment.

Ensemble. The five members are RelTR (Cong et al., 2023), SGTR (Li et al., 2022), Neural Motifs (Zellers et al., 2018), and two two-stage pipelines that pair an RT-DETR detector (Zhao et al., 2024) or a Grounding DINO detector (Liu et al., 2024) with a lightweight relation head trained on Visual Genome. The detector and tracker used for the machine axis are chosen from families outside the ensemble, and the teaching ensemble never produces the labels against which it is scored.

Student. The student uses a ResNet-18 visual encoder, a two-layer temporal transformer over eight causally sampled frames, and a two-layer query decoder at hidden width 256. Grammar tokens contain class embeddings, canonical boxes, persistence, occupancy lookup features, continuation support, and validity masks, expected anchor and track queries are concatenated with 100 learned detection queries, and cross-attention receives both visual and grammar tokens. Candidate heads emit class, box, three evidence scores, and evidence validity, and relation heads act on pairs of matched candidate embeddings with available context. The candidate distillation loss is masked binary cross-entropy with soft teacher targets plus a validity term, the pair loss is binary cross-entropy on drop, hallucination, and relation labels supported by clean-to-degraded correspondence with uncertain correspondences excluded, and the detection loss combines class cross-entropy with L1 and generalized IoU box terms after query assignment. The pair head is retained as a diagnostic and is not treated as validation of the teacher. Table 10 lists every hyperparameter with its selection rule. All experiments run on a single NVIDIA A100 40GB GPU.

VLM protocol. Both VLM conditions use Qwen2.5-VL-72B at 1024 pixel input with greedy decoding, the same prompt, and a fixed serialization of candidates. The prompt asks for one record per candidate with an identifier, a support score in $[0, 1]$, evidence availability, and one sentence of justification, allows an unavailable answer, and instructs the model not to mark a rare location as false solely because it is uncommon. The history-conditioned condition additionally receives the deterministic history summary and expected-node slots of Appendix E. Parsing failures count toward missing coverage, prompt tuning uses validation cameras only, and per-node outputs are scored on the same candidate universe as every other method.

M Runtime

At deployment the full path runs at about 640 ms per frame and the fast path at 14 ms including grammar lookup, measured on one NVIDIA A100 40GB at 1280×720 input with batch size one after warmup and with synchronization, and the stage breakdown in Table 11 shows that the ensemble accounts for 92.2 percent of the full path, which is exactly the cost the fast path removes.

Table 10: Hyperparameters and selection rules. Sensitivity of SRCC to each grammar parameter around its default is given in the last column as the spread over the searched range, and all spreads are below 0.015.

symbol	meaning	value	selection rule	spread
τ_s	anchor persistence	0.8	validation cameras, $[0.6, 0.95]$	0.007
τ_r	anchor relation frequency	0.6	validation cameras, $[0.3, 0.8]$	0.009
τ_r^A	dynamic relation admissibility	0.1	validation cameras, $[0.05, 0.3]$	0.006
η	matching IoU	0.5	fixed, $[0.3, 0.7]$	0.006
$\lvert\mathcal{Z}\rvert$	occupancy grid	16×16	validation cameras, 8^2 to 32^2	0.014
α, α_r	smoothing	1	fixed, $[0.25, 4]$	0.002
β	occupancy quantile	0.05	validation cameras, $[0.02, 0.1]$	0.005
ϵ	numerical floor	10^{-6}	fixed	–
$\sigma_{\min}$	baseline floor	0.05	validation cameras	0.004
$\omega_{\mathcal{A}}/\omega_{\mathcal{S}}$	operational asymmetry	4	operational choice, 1 to 8	0.004
w	readout weights	$(1, 1, 1)/3$	fixed, no MOS fitting	–
K	ensemble size	5	1 to 7	0.013
L	student window	8 frames	causal sampling	–
$(\lambda_d, \lambda_p, \lambda_b)$	loss weights	(1, 0.5, 1)	validation cameras	–
	optimizer, learning rate	AdamW, 10^{-4}	training and validation only	–
	batch, epochs, seeds	16, 50, 3	stop on validation loss	–

The grammar is estimated once per camera offline in about eleven minutes of wall time over 1,800 sampled frames, and its stored size is about 0.4 MB per camera.

Table 11: Per-frame runtime on a single GPU.

stage	full path (ms)	share (%)	fast path (ms)
scene graph ensemble ($K = 5$)	590	92.2	–
student f_θ	–	–	12
grammar matching and channels	44	6.9	–
grammar lookup and verdict	6	0.9	2
total	640	100.0	14

N THE D-CQA CORPUS AND D-CQA-SG

D-CQA is a distorted CCTV quality assessment corpus under separate review (Anonymous, 2026). It contains 50 fixed surveillance cameras covering streets, parking facilities, transit concourses, retail interiors, industrial yards, corridors, and night storage, each recorded for about one hour at 15 fps. From each camera 60 source frames and clips are sampled from the final 20 minutes, for 3,000 sources, and each is distorted by Gaussian blur, sensor noise, compression, low light, and haze at four severity levels, for 60,000 distorted items counted against the released manifest. Opinion scores were collected from 24 raters per item in a controlled viewing environment following ITU-R BT.500, screened by the standard outlier rejection, and normalized to $[0, 1]$. Contiguous clips carry spatial corruption applied consistently across frames with track identities preserved, and video baselines receive these clips rather than repeated frames. The history footage carries no opinion scores and is never evaluated. D-CQA-SG is the graph annotation layer released with this paper and provides, for every item, the consensus graph of the pristine source, the per-candidate evidence labels of every distorted version generated by the procedure of Section 3.4, the independent annotation set of Appendix D, and the per-camera grammars. Faces and license plates are blurred at source, and no identity annotations are included.

O LIMITATIONS

Beyond grammar staleness, the method inherits the class vocabulary and the shared failure modes of its generators, so an object class no generator recognizes never enters the grammar and Proposition 1 protects against uncorrelated errors only. An agent never observed during the current event cannot be attributed as missing from history alone (Proposition 3), and such cases are reported as a coverage limitation rather than removed from the denominator. The rarity deficit loses resolution where agents can be anywhere, as the MOT20 results in Appendix F show. Until independent calibration is established for a deployment, the output should support review and selective processing rather than automatically suppress rare detections.