PLAUSIBILITY IS NOT QUALITY: OBSERVABILITY-AWARE OPTICAL EVIDENCE FOR SURVEILLANCE IMAGE ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models now synthesize surveillance frames that are sharper than authentic footage yet depict scenes that could not physically occur, with reflections lacking sources and shadows lacking objects. Image quality assessment offers no protection because these frames are not degraded, and we show that perceptual quality and perceived plausibility are dissociable perceptual axes. A third property, observability, governs both, since a shadow may be cast by an object outside the frame, so an assessor must distinguish an observed violation from insufficient evidence. We formalize plausibility as the Perceived Plausibility Score (PPS), a continuous human-anchored measure whose protocol also records whether each optical cue was judged decidable, and we construct GenCQA-DB, approximately 55,000 source-grouped authentic and generated CCTV images with controlled reflection, shadow, and illumination violations. We propose PACT, a Physical and Appearance Cross-map Transformer that pairs each optical residual map with a cue confidence map encoding evidence availability and fuses them with appearance entropy maps through cross-map attention. PACT attains an SRCC of 0.874 against 0.791 for an equally supervised image baseline under camera-disjoint evaluation, retains 0.836 when cameras and generator are both held out, and follows human judgments when the evidence for a violation fades instead of mistaking a vanished residual for a consistent scene.

1 INTRODUCTION

Surveillance imagery is consulted as a record of what physically happened, and generative and editing models now put this evidentiary role at risk. Compression and limited spatial detail can degrade the appearance of an otherwise plausible scene, whereas a sharp image can contain a reflection that does not match the person standing in front of the glass. These contrasting cases, illustrated in Fig. 1 (a) and (b), motivate a distinction between perceptual quality, which concerns how an image looks, and perceived optical plausibility, which concerns how consistent its visible optical relationships appear. Realistic texture is evidence for the first and says little about the second, so plausibility is not quality, and the axis that conventional quality assessment cannot see is the one that matters once imagery is treated as evidence.

Reliable plausibility assessment also depends on whether the image contains enough evidence to support a judgment at all. A shadow may be cast by an object outside the camera view, and an apparent lighting discrepancy may be explained by a light source the camera never sees. The nighttime scene in Fig. 1 (c) illustrates this ambiguity, since enlarging the region reveals local detail without resolving the lighting configuration. An assessor must therefore distinguish an observed discrepancy from insufficient evidence, and we refer to the availability and reliability of such evidence as observability. Observability also connects the two axes above, because degradation rarely creates an optical inconsistency but can remove the evidence needed to detect one.

Existing assessment leaves this combination unaddressed. Image quality models learn human judgments of visual appearance (Ke et al., 2021; Wu et al., 2024), AGIN collects naturalness ratings (Chen et al., 2023), GOBench assesses optical authenticity (Zhu et al., 2025), and HEIE combines implausibility scores, heatmaps, and explanations (Yang et al., 2025), but these efforts score

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.

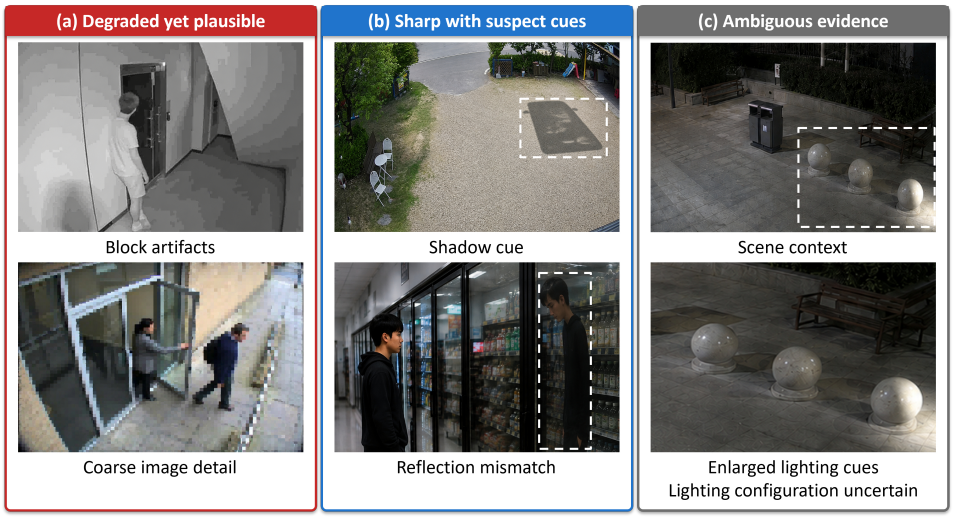

Figure 1: Visual quality and optical evidence in surveillance imagery. (a) Degraded images can remain plausible. (b) Sharp images can contain suspect optical cues. (c) Enlarged lighting cues can remain ambiguous when the lighting configuration is uncertain. These cases motivate assessing optical consistency together with the availability of the evidence that supports it.

the appearance axis, treat plausibility as a holistic impression, or delegate optical judgments to a vision-language model, and none represents whether the evidence for a judgment is present at all, which is costly on footage that is routinely dark, compressed, and partially occluded. We ask whether explicitly representing optical evidence together with its observability improves the prediction of human plausibility judgments beyond appearance statistics and beyond image models trained with the same supervision.

We study this question through the Perceived Plausibility Score (PPS), a continuous human rating of how optically plausible the visible scene appears, and we propose PACT, a Physical and Appearance Cross-map Transformer that predicts it. PACT combines appearance entropy maps with residual maps that measure how far observed reflections, shadows, and illumination depart from approximate optical models, pairs each residual with a confidence map of cue availability and reliability so that missing support is distinguished from a small discrepancy, and fuses all maps through cross-map attention. Under camera-disjoint evaluation PACT reaches a PPS SRCC of 0.874 against 0.791 for an equally supervised image baseline, and under joint camera and generator holdout the two reach 0.836 and 0.702, so the margin widens when neither scene nor generator has been seen. Our contributions are threefold.

- We present GenCQA-DB, a surveillance database with source-grouped authentic and generated imagery and a human-rated subset that supports the joint study of quality, plausibility, and observability.
- We introduce an observability-aware optical evidence representation with explicit cue confidence, together with PACT, which integrates this evidence with appearance statistics through cross-map attention.
- We specify camera-disjoint evaluation, joint camera and generator holdout, and source-matched interventions that isolate the contribution of optical evidence under changes in scene content and visibility.

2 Related Work

Perceptual assessment of generated imagery has grown along two lines. No-reference quality models predict human judgments from scene statistics, learned representations, or language priors (Mittal et al., 2012; 2013; Su et al., 2020; Ke et al., 2021; Yang et al., 2022; Wu et al., 2024), entropy descriptors track quality judgments on authentic CCTV frames (Han et al., 2025), and databases for gener-

ated content add naturalness and rationality dimensions (Li et al., 2023; 2024; Wang et al., 2023b; Chen et al., 2023). Physics benchmarks such as PhyBench, Science-T2I, PhysBench, GOBench, PICABench, and PhyEditBench corroborate that optical violations are a dominant failure mode of current generators (Meng et al., 2024; Li et al., 2025; Chow et al., 2025; Zhu et al., 2025; Pu et al., 2025; Guo et al., 2026). Yet the former line measures how an image looks or how natural it feels as a whole, the latter relies on vision-language judges or rubric scores over generated images alone, and neither represents whether a given optical cue is observable. GenCQA-DB pairs authentic and generated surveillance frames, and PACT predicts human plausibility ratings while encoding whether each cue can be evaluated from the available observations.

Forensic localization exposes edits through learned manipulation evidence (Liu et al., 2022; Guo et al., 2023; Guillaro et al., 2023), and explainable evaluators such as HEIE add heatmaps and textual explanations (Yang et al., 2025). The reliability these methods express concerns the detector rather than the scene, and an edited region can remain optically plausible, so we separate plausibility from manipulation detection and use editing masks only as auxiliary supervision. Model-based forensics exploits lighting and shadow inconsistencies (Johnson & Farid, 2005), mirror geometry underlies reflection forensics and mirror-aware generation (Fei et al., 2026; Dhiman et al., 2025), and Lambertian images occupy a low-dimensional subspace (Basri & Jacobs, 2003). Each of these models presupposes that its inputs are observable, a precondition that surveillance imagery frequently violates, and applying them regardless turns missing evidence into spurious residual. We therefore represent observability rather than assume it, and an extended discussion of related work is provided in Appendix A.

3 GenCQA-DB and the Perceived Plausibility Score

3.1 Database Construction

GenCQA-DB extends the distortion-free surveillance benchmark CQA-DB (Han et al., 2025) to the generative setting with approximately 55,000 images. Authentic frames are drawn from deployed surveillance cameras spanning four scene types defined by environment and time of day, and every frame is tagged with its source camera so that images sharing a viewpoint form a source group. Generated counterparts are anchored to the same viewpoints through multiple generation and editing pipelines, so the two streams share operating statistics such as resolution, compression, and lighting and differ primarily in optical validity. Each generated image targets exactly one of three optical factors, namely mirror reflection, cast shadow, and scene illumination, at a graded strength level, and the manipulated region is recorded as a binary mask during generation. Source grouping is what makes the evaluation protocols of Sec. 5 possible, since a camera-disjoint split assigns entire source groups to one partition and a generator holdout withholds one pipeline in full. Because authentic surveillance footage is routinely dark, compressed, and partially occluded, the database also spans a wide range of evidence availability without any deliberate intervention, which is the property the observability analysis below relies on. Composition statistics and representative samples are provided in Appendix B.

3.2 Subjective Study Protocol

Human ratings are collected on a subset of 4,800 images sampled with stratification over scene type, optical factor, strength level, and source group, so that every source group and every factor is represented while the remaining images carry only generation labels and masks. Two independent sessions follow the ITU-R BT.500 recommendation (ITU-R, 2019) with disjoint participant groups. In the quality session, participants rate perceptual quality on a five-point scale under the standard IQA instruction, yielding a mean opinion score (MOS) per image. In the plausibility session, participants answer a different question, namely how optically possible the depicted scene appears, on a continuous slider, and are explicitly instructed to disregard noise, blur, and compression. The Perceived Plausibility Score of image i is the trimmed mean of its normalized plausibility ratings,

$$\mathrm{PPS}_i = \frac{1}{|\mathcal{K}_i|} \sum_{k \in \mathcal{K}_i} \tilde{s}_{i,k}, \tag{1}$$

where $\mathcal{K}_i$ is the set of valid ratings after discarding scores beyond $\pm 2\sigma$ and $\tilde{s}_{i,k} \in [0, 1]$. After rating each image, participants in the plausibility session additionally answer, for each of the three

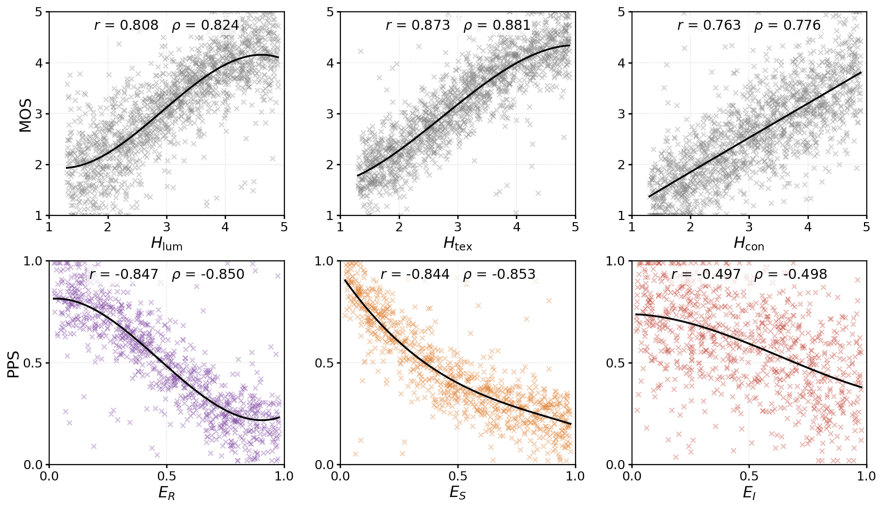

Figure 2: Subjective analysis on GenCQA-DB. The top row relates quality MOS to the luminance, texture, and contrast entropies, and the bottom row relates PPS to the image-level violation magnitudes E_R, E_S, and E_I on the corresponding factor subsets.

cues, whether the cue appears consistent, appears inconsistent, or cannot be decided from the image. The majority response to this third question yields a human judgeability label per image and cue, which serves as the reference for evidence availability in the analysis below and for validating the confidence maps of Sec. 4. Separating the sessions by participant group prevents carry-over between the quality and plausibility instructions, which matters because the central claim of this work is that the two constructs are dissociable. Inter-observer consistency reached a Spearman rank correlation of 0.912 in the quality session and 0.874 in the plausibility session, and the judgeability responses reached a Fleiss κ of 0.71.

3.3 Human Perceptual Analysis

The subjective annotations reveal how the two perceptual axes relate to the two factor families, as summarized in Fig. 2. In the top row, quality MOS correlates strongly with the luminance, texture, and contrast entropies, with Pearson r of 0.808, 0.873, and 0.763 and Spearman ρ of 0.824, 0.881, and 0.776, and the luminance and texture relations saturate at high entropy, indicating diminishing perceptual returns once scene detail is visible. This replicates the appearance sensitivity previously observed on authentic surveillance imagery and justifies carrying the three entropy maps into PACT as the appearance evidence stream.

The bottom row plots PPS against the image-level violation magnitudes E_R, E_S, and E_I defined in Sec. 4 on the corresponding factor subsets. Perceived plausibility declines with violation strength for every factor, yet both the sensitivity and the shape of the decline are factor-specific. Shadow violations are punished immediately, with PPS dropping steeply at small magnitudes ($r = -0.844$, $\rho = -0.853$), and reflection violations are tolerated up to a threshold and then collapse ($r = -0.847$, $\rho = -0.850$), consistent with observers overlooking subtle geometric offsets until the mirrored content visibly departs from the scene. Illumination violations show a much weaker relation ($r = -0.497$, $\rho = -0.498$), which a naive reading would take as insensitivity to lighting, but the judgeability labels say otherwise. Participants marked the illumination cue as not decidable in 38% of that subset, against 11% for reflection and 14% for shadow, and restricting the analysis to decidable images raises the illumination correlation to $\rho = -0.803$ while the remaining images show almost none ($\rho = -0.186$), whereas the same stratification moves the other two factors by less than 0.03. Lighting inconsistencies admit benign explanations such as unseen light sources far more often than an orphaned shadow does, so the weak aggregate relation reflects the available evidence rather than human perception. Observers respond to optical violations as a graded signal whenever the evidence permits a judgment, which motivates the continuous PPS, the per-factor residual maps, and the separate confidence maps that PACT consumes.

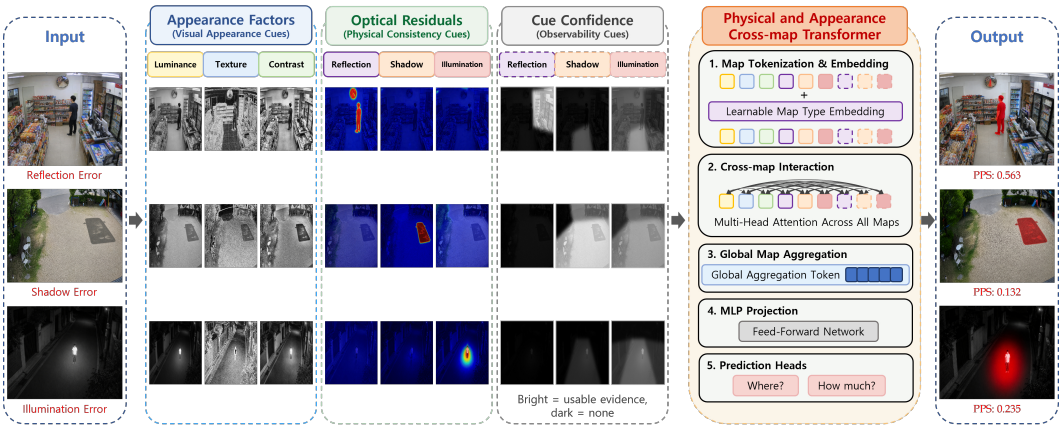

Figure 3: Overview of PACT. Appearance entropy maps, optical residual maps, and cue confidence maps are tokenized with learnable map-type embeddings, fused by cross-map attention, and decoded by localization and scoring heads.

Finally, the two rows jointly exhibit the dissociation that Fig. 1 anticipates. On the authentic subset, MOS and PPS are strongly coupled with a Spearman correlation of 0.814, whereas on the generated subset the coupling collapses to 0.217 because appearance statistics remain intact while optical validity varies freely. Quality carries almost no information about plausibility exactly where plausibility matters most, and this gap, together with the evidence availability that governs whether a violation can be perceived at all, is what PACT is designed to represent.

4 PACT, AN OBSERVABILITY-AWARE CROSS-MAP TRANSFORMER

PACT converts an input frame into three families of maps, as shown in Fig. 3. Three appearance entropy maps carry the statistics on which quality judgments rest, three optical residual maps measure how far observed reflections, shadows, and illumination depart from approximate optical models, and three cue confidence maps record whether each measurement could be made and how far it can be trusted. The nine maps are tokenized into a shared space, exchanged through cross-map attention, and decoded by heads that answer where a violation resides and how much it erodes plausibility. On authentic imagery the residuals stay near the level real footage naturally exhibits and appearance evidence governs the prediction, and where a cue is not observable the confidence map lets the model withhold judgment instead of reading a missing measurement as consistency.

Appearance factor maps. The frame I is resized to 384×384, each of the luminance, texture, and contrast response channels is partitioned into 16×16 patches, and the Shannon entropy of each patch is computed from its normalized K-bin histogram with $K = 64$,

$$A_c(p) = -\sum_{k=1}^{K} h_{c,k}(p)\, \log_2 h_{c,k}(p), \qquad c \in \{\mathrm{lum}, \mathrm{tex}, \mathrm{con}\}. \tag{2}$$

The maps on the 24×24 grid summarize the descriptors that Sec. 3.3 showed to track quality judgments, and their generation is illustrated in Appendix C.

Optical residual maps. Each factor instantiates one compact optical model, renders the expected state, and converts the discrepancy with the observation into a spatial residual. Residuals are mapped into $[0, 1]$ by a saturating normalization ϕ_f calibrated on authentic frames, and the pipeline is illustrated together with estimation details in Appendix D.

Reflection. The reflective surface is a mirror plane π with $\mathbf{n}^\top \mathbf{x} + d = 0$, the planar geometry that mirror-aware generation adopts as its supervision target (Dhiman et al., 2025). A scene point maps to $\mathbf{x}' = \mathbf{x} - 2(\mathbf{n}^\top \mathbf{x} + d)\,\mathbf{n}$, its projection $\Pi(\mathbf{x}')$ predicts the image position of the reflection, and the residual compares observed reflected content with content warped from the prediction,

$$M_R(\mathbf{u}) = \phi_R(\|I(\mathbf{u}) - I(w_\pi(\mathbf{u}))\|_1)\,, \tag{3}$$

where w_π is the induced warp, turning the reflection cue of forgery detection (Fei et al., 2026) from a verdict into a residual.

Shadow. Cast and attached shadow constraints are imposed under a point light whose direction is estimated from object-shadow correspondences (Johnson & Farid, 2005). Projecting each occluder silhouette along that direction onto the ground yields the expected support S^{pred}, and

$$M_S(\mathbf{u}) = \phi_S(|\mathbb{1}_{S^{\mathrm{obs}}}(\mathbf{u}) - \mathbb{1}_{S^{\mathrm{pred}}}(\mathbf{u})| \cdot g(\mathbf{u})), \tag{4}$$

where g softens the disagreement by local shadow contrast. An orphaned shadow raises M_S over its full support because no occluder projects onto it under any admissible direction.

Illumination. Following the low-dimensional structure of Lambertian images under distant lighting (Basri & Jacobs, 2003), the scene is partitioned into regions j with normals $\mathbf{n}_j$, and the predicted shading under ambient term a, albedo p_j, and light direction $\mathbf{l}$ is $Y_j^{\mathrm{pred}} = a + p_j \max(0, \mathbf{n}_j^\top \mathbf{l})$. The residual $M_I(\mathbf{u}) = \phi_I(|Y(\mathbf{u}) - Y_{j(\mathbf{u})}^{\mathrm{pred}}|)$ accumulates over a pedestrian who glows without any light source.

Cue confidence maps. Every residual is conditional on assumptions a surveillance frame may not satisfy, since reflection needs a reflective surface with correspondences, shadow needs a visible occluder and ground plane, and illumination needs enough shaded surface to fit the lighting model. When these fail, a model consuming the residual alone cannot tell a consistent scene from an absent measurement. We therefore attach to each factor

$$C_f(\mathbf{u}) = a_f(\mathbf{u})\, r_f, \qquad f \in \{R, S, I\}, \tag{5}$$

where the availability $a_f(\mathbf{u}) \in [0, 1]$ marks the support of the cue and the reliability $r_f \in [0, 1]$ summarizes how well the fitted model explains the observations it was fitted to, namely the inlier ratio of the plane fit, the agreement of light-direction votes across occluders, and the goodness of fit of the Lambertian model. Where a cue is absent, both residual and confidence are zero, so the pair states that nothing could be measured rather than that nothing was wrong. Two rules keep confidence from absorbing the violations it qualifies. Reliability is computed from the consensus set of the robust fit, so one element that contradicts the scene lowers the residual and leaves the confidence intact, whereas a scene with too few consistent elements, such as a lone occluder with no second shadow to vote, receives low confidence in line with the human judgment that it cannot be decided. Confidence is also computed identically for authentic and generated frames and contains no parameters fitted to plausibility labels, so it cannot act as a generator detector, which the interventions of Sec. 5 verify. This is the representation Fig. 1(c) calls for, since the lighting configuration is not recoverable from the enlarged spheres, so the illumination residual there cannot be trusted and C_I says so.

Image-level magnitudes. For each factor, the violation magnitude used in Sec. 3.3 is the mean of the top 2% of residual responses,

$$E_f = \frac{1}{|\Omega_f|} \sum_{\mathbf{u} \in \Omega_f} M_f(\mathbf{u}), \qquad \Omega_f = \text{top } 2\% \text{ of } M_f, \tag{6}$$

and the observability O_f is the mean of C_f over Ω_f, with $O_f = 0$ when the cue is absent. The optical models are deliberately low-dimensional. Their role is not to reconstruct the scene but to expose where the image resists optical explanation and where that question cannot be asked, and this generator-agnostic construction is what the holdout experiments probe.

Cross-map transformer. The nine maps are tokenized on the shared 24×24 grid, and each token receives a $D = 384$ dimensional projection, a positional embedding, and a learnable map-type embedding $\mathbf{e}_m$ identifying its source map,

$$\mathbf{z}_i^m = W_m \,\mathrm{vec}(x_i^m) + \mathbf{e}_{\mathrm{pos}(i)} + \mathbf{e}_m. \tag{7}$$

The map-type embedding lets statistically incommensurable evidence, entropy, residual, and confidence, coexist in one attention space. A global token $\mathbf{g}$ is prepended and the $9 \times 576 + 1$ tokens pass through $N = 6$ blocks of multi-head self-attention across all maps, so appearance context can discount a reflection residual in a low-texture region where the warp is unreliable and a confidence token can suppress an illumination residual whose fit was poorly constrained. Because confidence enters as tokens rather than a fixed gate, the model learns how much a given level of evidence should moderate a given discrepancy, and Sec. 5.4 compares this choice against multiplicative gating.

Prediction heads and training objectives. The localization head upsamples patch tokens into an edit-support probability map $\hat{V}$ supervised by the generation masks with $\mathcal{L}_{\text{loc}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}}$, with authentic frames supplying empty masks to suppress false alarms. It is an auxiliary task that grounds spatial evidence rather than defining plausibility, since an edited region need not be optically implausible. The scoring head regresses $\widehat{\text{PPS}} = \sigma(\mathbf{w}^\top \mathbf{g})$ from the global token together with an auxiliary $\widehat{\text{MOS}}$ that anchors the appearance stream to the quality construct,

$$\mathcal{L}_{\text{score}} = \|\widehat{\text{PPS}} - \text{PPS}\|_1 + \|\widehat{\text{MOS}} - \text{MOS}\|_1 + \gamma\, \mathcal{L}_{\text{rank}}, \tag{8}$$

where $\mathcal{L}_{\text{rank}}$ preserves the ordering of human scores, and the total objective is $\mathcal{L} = \mathcal{L}_{\text{score}} + \lambda\, \mathcal{L}_{\text{loc}}$, with training details in Appendix G.

5 Experiments

5.1 Experimental Setup

Evaluation follows the source grouping of Sec. 3.1. Under the camera-disjoint protocol, source groups are assigned to training, validation, and test partitions in an 8:1:1 ratio with stratification over scene type, so no test image shares a camera with a training image. Under the joint camera and generator holdout, the test partition additionally contains only images from the withheld generation pipeline. All numbers average five partitions with seeds $\{0, 1, 2, 3, 4\}$, with per-partition standard deviation within ± 0.010 SRCC, and we report Spearman (SRCC), Pearson (PLCC), and Kendall (KRCC) correlations against PPS, with PLCC after the standard four-parameter logistic mapping. Zero-shot no-reference IQA models (Mittal et al., 2012; Ke et al., 2021; Yang et al., 2022; Zhang et al., 2023; Wu et al., 2024) and VLM judges are evaluated with released weights, the latter with a plausibility instruction matched to the human study wording, a forensic detector (Guillaro et al., 2023) contributes its integrity score, and the equally supervised image baselines, marked (ft), are retrained on GenCQA-DB with PPS labels under identical partitions so that any remaining gap reflects the evidence representation rather than the training signal. Training details are in Appendix G.

Table 1: PPS prediction on GenCQA-DB under camera-disjoint evaluation and under joint camera and generator holdout. (ft) denotes retraining on GenCQA-DB with PPS labels under the same partitions. Best results are in bold and second best are underlined.

		Camera-disjoint			Joint holdout	
Family	Method	SRCC	PLCC	KRCC	SRCC	PLCC
Zero-shot IQA	BRISQUE	0.214	0.238	0.146	0.203	0.221
	MUSIQ	0.312	0.334	0.214	0.297	0.315
	MANIQA	0.345	0.361	0.237	0.329	0.342
	LIQE	0.362	0.377	0.249	0.348	0.359
	Q-Align	0.401	0.418	0.276	0.388	0.402
VLM judge	GPT judge	0.512	0.505	0.361	0.498	0.489
	Gemini judge	0.487	0.492	0.339	0.471	0.476
Forensic	TruFor (score)	0.402	0.421	0.278	0.351	0.366
Equally supervised	MANIQA (ft)	0.741	0.756	0.552	0.618	0.631
	Q-Align (ft)	<u>0.791</u>	<u>0.803</u>	<u>0.601</u>	<u>0.702</u>	<u>0.714</u>
Ours	PACT	**0.874**	**0.886**	**0.692**	**0.836**	**0.848**

5.2 Comparison under Camera and Generator Holdout

The comparison across families is reported in Table 1. Zero-shot IQA models correlate weakly with perceived plausibility, with even the strongest reaching an SRCC of 0.401, which is the quantitative form of the dissociation thesis. VLM judges land in between, consistent with prior reports that they reason about physical commonsense yet cannot ground light transport at the pixel level. Retraining on PPS labels helps substantially, yet both equally supervised baselines remain below PACT, which reaches 0.874 under camera-disjoint evaluation and improves over the strongest baseline by 0.083.

The joint holdout columns isolate generalization. Both equally supervised baselines lose roughly 0.1 SRCC when generator and cameras change together, indicating that part of what they learned is generator-specific appearance, whereas PACT loses 0.038 because its residuals derive from scene geometry and photometry rather than generator statistics. Across factor subsets PACT reaches SRCC of 0.869, 0.891, and 0.847 for reflection, shadow, and illumination, and the illumination value is where observability matters most, since the same model without confidence maps reaches 0.803 on that subset, mirroring Sec. 3.3. On the authentic subset, the auxiliary MOS output reaches an SRCC of 0.891, within 0.006 of the strongest dedicated IQA model retrained on the same labels (Appendix H), so one model covers both regimes of a surveillance pipeline.

5.3 Observability Analysis and Source-matched Interventions

Using the image-level observability O_f of Eq. (6) as a score for whether participants marked a cue as not decidable yields an AUROC of 0.912, 0.897, and 0.884 for reflection, shadow, and illumination, so an analytically computed quantity with no plausibility supervision recovers the human notion of decidability. Stratifying the camera-disjoint test set into observability terciles, PACT reaches SRCC of 0.902, 0.871, and 0.796 from high to low against 0.822, 0.790, and 0.641 for Q-Align (ft), so the margin widens from 0.080 to 0.155 as evidence becomes scarce, with full results in Appendix E.

Source-matched interventions change the evidence while the camera stays fixed, with full results in Table 4 of Appendix F. A content intervention inserts one violation into an authentic frame, and a visibility intervention degrades a violated frame by downscaling, recompression, and darkening so that the violation remains but its evidence fades. Both sets received human ratings under the protocol of Sec. 3.2. Under content interventions every model with optical residuals reproduces most of the human drop of 0.312. Under visibility interventions human PPS rises by only 0.184, since participants increasingly answer that the cue cannot be decided, and the design choices of Sec. 4 separate. The model without confidence raises its prediction by 0.356, nearly twice the human change, because a residual that fades is read as consistency, multiplicative gating reduces this to 0.231, and confidence as tokens gives 0.176, within 0.01 of the human value, while the per-pair correlation with human changes rises from 0.421 to 0.741. Averaged over the intervened factor, O_f falls from 0.71 to 0.29 under visibility interventions and stays at 0.70 under content interventions, so confidence responds to the loss of evidence and not to the presence of a violation, as the first design rule of Sec. 4 requires.

Table 2: Ablation studies. PPS columns report SRCC under camera-disjoint evaluation and joint holdout, and MOS reports SRCC on the authentic subset. Shaded rows concern the observability mechanism.

Variant	PPS (camera)	PPS (joint)	MOS
PACT (full)	**0.874**	**0.836**	**0.891**
w/o optical residual stream	0.694	0.612	0.883
w/o appearance stream	0.802	0.781	0.641
w/o confidence maps	0.851	0.784	0.889
confidence as multiplicative gate	0.859	0.803	0.890
late fusion instead of cross-map attention	0.831	0.779	0.872
w/o map-type embedding	0.845	0.802	0.879
w/o localization head	0.856	0.821	0.885
w/o ranking loss	0.861	0.826	0.887

5.4 Ablation Studies

Each design choice is dissected in Table 2. Removing the optical residual stream costs 0.180 SRCC under camera-disjoint evaluation and 0.224 under joint holdout while leaving quality prediction intact, and removing the appearance stream costs 0.250 on MOS while plausibility degrades moderately, confirming that the two streams carry complementary constructs. The confidence maps show the signature of an observability mechanism, since their removal costs 0.023 under camera-disjoint evaluation but 0.052 under joint holdout, where unseen cameras and an unseen generator shift the distribution of available evidence. A multiplicative gate recovers only part of this gap, in line with

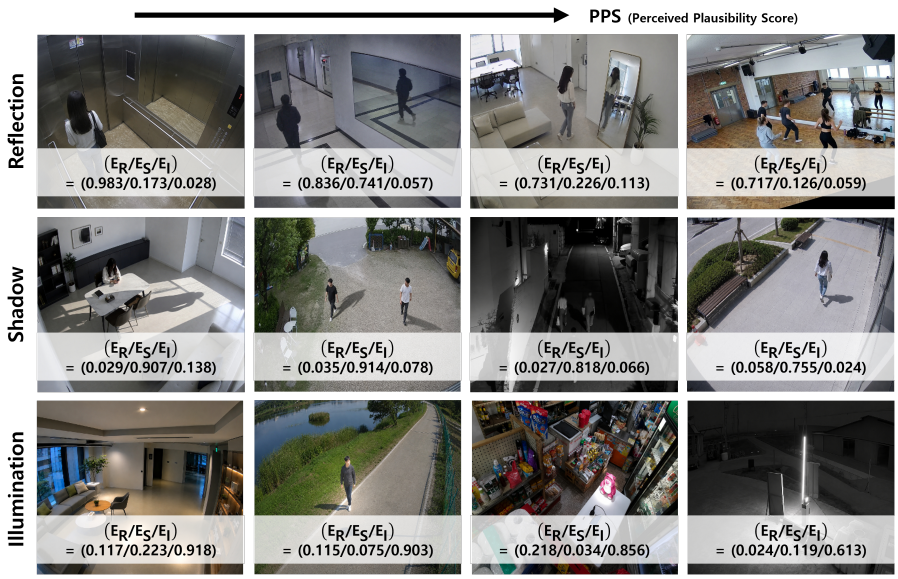

Figure 4: Qualitative results on the three factor subsets ordered by predicted PPS from left to right, with per-factor violation magnitudes $(E_R/E_S/E_I)$. The dominant magnitude identifies the violated factor in every case.

the interventions above, which supports letting the model learn how evidence should moderate a discrepancy. Late fusion and dropping the map-type embedding each degrade plausibility prediction, and the localization head and ranking loss contribute smaller but consistent gains.

5.5 Localization and Qualitative Analysis

The edit-support head improves IoU by 0.171 and AUPRC by 0.168 over the strongest forgery localization baseline retrained on GenCQA-DB masks (Appendix I), which we attribute to the residuals indicating where the scene resists optical explanation rather than where editing traces survive. Qualitative results in Fig. 4 show that the dominant magnitude attributes each image to the correct violation type and that the nighttime scenes receive intermediate scores because low observability moderates their residuals, the behavior Fig. 1(c) asks for, with more examples in Appendix J. Finally, human ratings couple MOS and PPS on real frames ($\rho = 0.814$) and decouple them on generated frames ($\rho = 0.217$), and PACT predictions reproduce this structure (ρ of 0.752 and 0.195, Appendix K), indicating that the model has learned the perceptual geometry of the two axes rather than a shortcut between them.

6 Conclusion

We argued that plausibility is not quality, and that observability determines whether either can be judged from a surveillance image at all. Human ratings on GenCQA-DB bear this out, since the two axes decouple on generated frames and the apparent insensitivity of observers to lighting violations disappears once images they judged undecidable are set aside. PACT operationalizes these findings by pairing each optical residual with a cue confidence map, so that a missing measurement is never mistaken for a consistent scene, and it outperforms equally supervised image models under camera and generator holdout by a margin that widens as evidence becomes scarce, while following human judgments when the evidence for a violation fades.

Other optical factors such as contact and occlusion order await their own residual and confidence pairs, and learning observability from human judgeability labels or extending the formulation to video are natural next steps. As generative models erase the statistical traces that forensics has relied on, whether an image is optically possible, and whether it offers enough evidence to tell, may be what endures.

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. Separately, generative and editing models were used as research subjects to construct the AI-generated portion of GenCQA-DB, as described in Sec. 3.1, which is the object of study rather than a writing aid. 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 imagery and the deliberate creation of manipulated CCTV frames, which raises privacy and dual-use concerns that we address as follows. The authentic footage underlying GenCQA-DB was collected from cameras operated with the authorization of the corresponding site owners, personally identifiable visual information such as faces and license plates is blurred, and no identity labels or person-specific annotations are included. The proposed measure assesses scene-level physical plausibility and is not designed for person identification or re-identification. The manipulated images in GenCQA-DB are created solely to study how physical violations affect human perception and to train detectors of such violations. To limit misuse, the generation pipelines and manipulation parameters will not be released, the database will be distributed for research purposes under a usage agreement, and every generated image is released with its violation annotation so that the data cannot be passed off as authentic. The subjective studies were conducted with informed consent under our institutional guidelines, participants were compensated, and no personal information beyond basic demographics was collected. We acknowledge that plausibility assessment could in principle be inverted to screen generated content for undetectable manipulations, and we believe the defensive value of a public benchmark and assessment model outweighs this risk, consistent with the established practice of the forensics community. Users of the dataset and method should comply with applicable privacy, data-protection, and ethical requirements and conduct appropriate ethical review before deployment in real-world surveillance settings.

REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate reproducibility. The PACT implementation, the physical consistency map extraction code, the PPS and violation-mask annotations of GenCQA-DB, and the configuration files used to produce the reported results will be released in a public repository upon publication, subject to the controlled-distribution terms described in the ethics statement. Sec. 5.1 describes the evaluation protocol, including the data splits, seeds, and metrics. Appendix D provides the estimation and calibration details of the three physics models, Appendix G the complete training hyperparameters, and Appendix B the database composition. The subjective study protocol, including instructions and rating interfaces for both sessions, will be documented in the released materials to allow replication of the PPS collection procedure.

REFERENCES

Ronen Basri and David W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003.

Zijian Chen, Wei Sun, Haoning Wu, Zicheng Zhang, Jun Jia, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. Exploring the naturalness of AI-generated images. *arXiv preprint arXiv:2312.05476*, 2023.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. PhysBench: Benchmarking and enhancing vision-language models for physical world understanding. In *International Conference on Learning Representations (ICLR)*, 2025.

Ankit Dhiman, Manan Shah, and R. Venkatesh Babu. MirrorVerse: Pushing diffusion models to realistically reflect the world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Hongyan Fei, Zexi Jia, Chuanwei Huang, Jinchao Zhang, and Jie Zhou. Exploring specular reflection inconsistency for generalizable face forgery detection. In *International Conference on Learning Representations (ICLR)*, 2026.

Robert T. Frankot and Rama Chellappa. A method for enforcing integrability in shape from shading algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(4):439–451, 1988.

Cindy M. Goral, Kenneth E. Torrance, Donald P. Greenberg, and Bennett Battaile. Modeling the interaction of light between diffuse surfaces. In *Proceedings of the 11th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pp. 213–222, 1984.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20606–20615, 2023.

Shengbin Guo, Shaokang He, Chaoyue Meng, Shengpeng Xiao, Xunzhi Xiang, Shaofeng Zhang, and Qi Fan. PhyEditBench: A real-world multi-stage benchmark for physics-aware image editing. *arXiv preprint arXiv:2606.26551*, 2026.

Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3155–3165, 2023.

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 (ICCV) Workshops*, pp. 3443–3450, 2025.

ITU-R. Recommendation ITU-R BT.500-14: Methodologies for the subjective assessment of the quality of television images. International Telecommunication Union, 2019.

Micah K. Johnson and Hany Farid. Exposing digital forgeries by detecting inconsistencies in lighting. In *Proceedings of the ACM Multimedia and Security Workshop*, pp. 1–10, 2005.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5148–5157, 2021.

Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. AGIQA-3K: An open database for AI-generated image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. doi: 10.1109/TCSVT.2023.3319020.

Chunyi Li, Tengchuan Kou, Yixuan Gao, Yuqin Cao, Wei Sun, Zicheng Zhang, Yingjie Zhou, Zhichao Zhang, Weixia Zhang, Haoning Wu, et al. AIGIQA-20K: A large database for AI-generated image quality assessment. *arXiv preprint arXiv:2404.03407*, 2024.

Jialuo Li, Wenhao Chai, Xingyu Fu, Haiyang Xu, and Saining Xie. Science-T2I: Addressing scientific illusions in image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2734–2744, 2025.

Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7505–7517, 2022.

Fanqing Meng, Wenqi Shao, Lixin Luo, Yahong Wang, Yiran Chen, Quanfeng Lu, Yue Yang, Tianshuo Yang, Kaipeng Zhang, Yu Qiao, and Ping Luo. PhyBench: A physical commonsense benchmark for evaluating text-to-image models. *arXiv preprint arXiv:2406.11802*, 2024.

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.

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

Yuandong Pu, Le Zhuo, Songhao Han, Jinbo Xing, Kaiwen Zhu, Shuo Cao, Bin Fu, Si Liu, Hongsheng Li, Yu Qiao, Wenlong Zhang, Xi Chen, and Yihao Liu. PICABench: How far are we from physically realistic image editing? *arXiv preprint arXiv:2510.17681*, 2025.

Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3667–3676, 2020.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2555–2563, 2023a.

Jiarui Wang, Huiyu Duan, Jing Liu, Shi Chen, Xiongkuo Min, and Guangtao Zhai. AIGCIQA2023: A large-scale image quality assessment database for AI generated images: From the perspectives of quality, authenticity and correspondence. In *CAAI International Conference on Artificial Intelligence*, pp. 46–57, 2023b.

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 (ICML)*, pp. 54015–54029, 2024.

Fan Yang, Ru Chen, Haoning Wu, et al. HEIE: MLLM-based hierarchical explainable AIGC image implausibility evaluator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2411.17261. Author list and venue to be verified.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1191–1200, 2022.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14071–14081, 2023.

Xiaorong Zhu, Ziheng Jia, Jiarui Wang, Xiangyu Zhao, Haodong Duan, Xiongkuo Min, Jia Wang, Zicheng Zhang, and Guangtao Zhai. GOBench: Benchmarking geometric optics generation and understanding of MLLMs. *arXiv preprint arXiv:2506.00991*, 2025.

Appendix

The appendix provides supplementary material for the main paper. Appendix A extends the discussion of related work, Appendix B details the composition of GenCQA-DB, Appendix C illustrates the appearance map generation pipeline, and Appendix D gives the estimation and calibration details of the three optical models together with the computation of the cue confidence maps. Appendix E reports the validation of the confidence maps against human judgeability and the stratified results, Appendix F describes the source-matched intervention protocol, Appendix G lists the training hyperparameters, Appendix H reports quality assessment on authentic imagery, Appendix I reports edit-support localization, Appendix J presents additional qualitative results, and Appendix K visualizes the dissociation structure in human ratings and model predictions.

A Extended Related Work

This section expands the positioning summarized in Sec. 2.

Perceptual quality and naturalness assessment. No-reference image quality assessment predicts human judgments from natural scene statistics (Mittal et al., 2012; 2013), learned visual representations (Su et al., 2020; Ke et al., 2021; Yang et al., 2022), and rating models built on language priors (Wu et al., 2024), and entropy descriptors of luminance, texture, and contrast track human quality judgments of authentic CCTV frames (Han et al., 2025). Human assessment of generated imagery extends this scope through AGIQA-3K, AIGIQA-20K, and AIGCIQA2023 (Li et al., 2023; 2024; Wang et al., 2023b), and AGIN adds naturalness ratings with technical and rationality dimensions (Chen et al., 2023). Each of these measures how an image looks or how natural it feels as a whole rather than whether its visible optical relationships are consistent. Quality nonetheless remains relevant to our target because degradation can obscure the cues on which a plausibility judgment rests, so we compare against quality predictors only when they receive the same plausibility supervision and partitions as our model.

Physical consistency in image generation and editing. PhyBench evaluates physical commonsense in text conditioned generation (Meng et al., 2024), Science-T2I learns SciScore for scientific consistency (Li et al., 2025), PhysBench probes the physical world understanding of vision-language models (Chow et al., 2025), GOBench isolates geometric optics with human judgments of optical authenticity (Zhu et al., 2025), and PICABench and PhyEditBench evaluate the physical effects that editing must preserve (Pu et al., 2025; Guo et al., 2026). These benchmarks corroborate our premise that optical violations are a dominant failure mode of current generators, but their evaluators are largely vision-language judges or rubric scores applied to generated images alone, and none represents whether a given cue is observable. GenCQA-DB pairs authentic and generated surveillance frames, and PACT predicts human plausibility ratings while encoding whether each cue can be evaluated from the available observations.

Manipulation localization and explainable assessment. Image forensics localizes alterations through learned manipulation evidence (Liu et al., 2022; Guo et al., 2023). TruFor outputs an integrity score, a localization map, and a reliability map for its own predictions (Guillaro et al., 2023), and HEIE produces implausibility scores, heatmaps, and textual explanations (Yang et al., 2025). The reliability these methods express concerns the detector rather than the scene, and their outputs answer a different question, since an edited region can remain optically plausible and a visible discrepancy need not coincide with the extent of an edit. We therefore separate perceived plausibility from manipulation detection, use editing masks only as auxiliary supervision, and attach confidence to the optical cue itself rather than to the prediction.

Observability-aware optical evidence. Model-based forensics exploits lighting and shadow inconsistencies (Johnson & Farid, 2005), planar mirror geometry underlies reflection forensics and mirror-aware generation (Fei et al., 2026; Dhiman et al., 2025), and Lambertian images under distant lighting occupy a low-dimensional subspace (Basri & Jacobs, 2003). Each model presupposes that its inputs are observable, whether as object-reflection correspondences, a visible occluder and ground plane, or enough shaded surface to fit a lighting model. Surveillance imagery frequently

violates these preconditions, as Fig. 1(c) illustrates, and applying the models regardless turns missing evidence into spurious residual. We therefore treat observability as a quantity to be represented rather than assumed, and PACT carries it alongside each residual so that a small discrepancy backed by usable evidence is distinguished from a measurement that could not be made.

B GenCQA-DB Composition Details

All images span four scene types defined by environment and illumination, namely Indoor-Day (34.7%), Indoor-Night (22.5%), Outdoor-Day (22.6%), and Outdoor-Night (20.2%), for an overall composition of 57.2% indoor and 42.8% outdoor scenes. The violation distribution is balanced across the three optical factors (shadow 34.1%, reflection 32.3%, illumination 31.6%). Authentic frames originate from 186 source cameras, each forming one source group, and every generated image inherits the source group of the authentic frame it is anchored to, which is what allows the camera-disjoint partitions of Sec. 5.1. The human-rated subset of 4,800 images covers all source groups, with at least 18 rated images per group, and its composition follows the same scene and factor proportions as the full database. The composition statistics together with representative samples from the four scene types are shown in Fig. 5.



Figure 5: Overview of GenCQA-DB. The left panels report scene composition and violation-type statistics, and the right panels show representative samples from the four scene types.

C Appearance Map Generation

The generation pipeline of the appearance factor maps is illustrated in Fig. 6. Patch-wise entropy is computed on the luminance, texture, and contrast responses following Eq. (2) and assembled into three appearance maps on the shared token grid.

D Optical Model Estimation, Calibration, and Confidence

The residual map generation is illustrated in Fig. 7. The saturating normalization is $\phi_f(t) = t/(t + \tau_f)$, where τ_f is calibrated per factor as the 95th percentile of the residuals measured on authentic frames, so that residuals are expressed relative to the level that real surveillance imagery naturally exhibits. For the reflection model, the mirror plane parameters are estimated with RANSAC over object-reflection correspondences, so a single inconsistent reflection cannot corrupt the estimate. For the shadow model, the light direction is estimated from occluding contours and object-shadow correspondences that vote consistently, and surface orientation where needed is recovered by integrability-constrained shape from shading (Frankot & Chellappa, 1988). For the illumination model, the parameters $(a, \{p_j\}, 1)$ are obtained by least squares over all regions. We deliberately

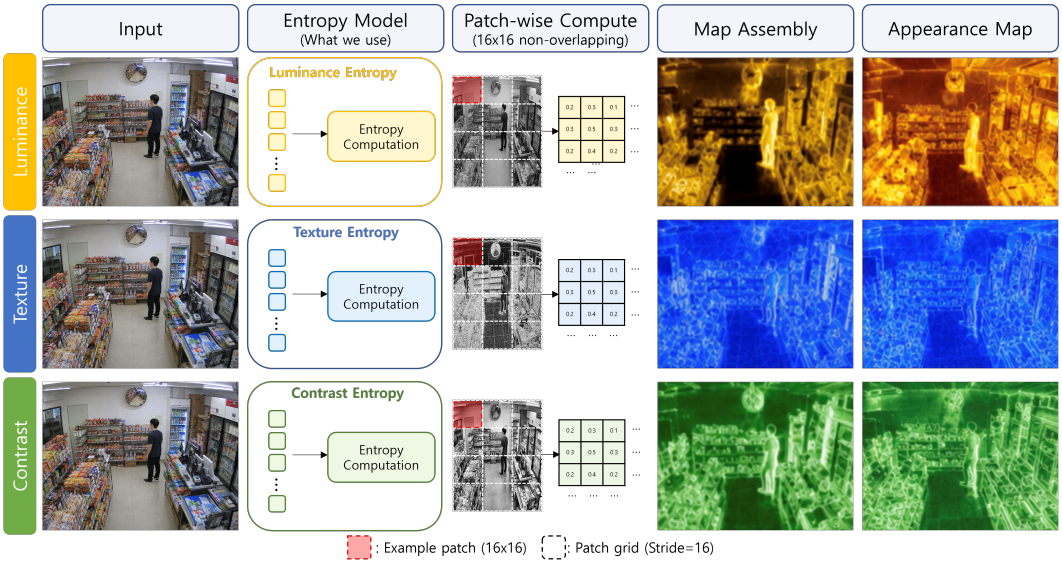

Figure 6: Appearance factor map generation. Patch-wise entropy is computed on luminance, texture, and contrast responses and assembled into three appearance maps.

exclude interreflection, whose faithful treatment requires global illumination models (Goral et al., 1984), because the diffuse term already exposes the gross violations that generative pipelines produce while keeping the estimation stable on low-quality footage.

The cue confidence maps of Eq. (5) are computed from the same estimation stage. The availability term a_f is a soft mask obtained by blurring the binary support of the cue with a Gaussian of 8 pixels on the 384×384 input, so that confidence decays smoothly at the boundary of the supported region. For reflection, the support is the detected reflective surface, and the reliability r_R is the RANSAC inlier ratio after the final refit. For shadow, the support is the union of the observed and predicted shadow regions, and r_S is the fraction of light-direction votes within 15° of the consensus direction, which is zero when fewer than two occluders vote. For illumination, the support is the union of regions whose recovered normal has a residual below 0.15 and whose luminance range exceeds 12 gray levels, and r_I is one minus the normalized root mean square error of the Lambertian fit on those regions, clipped to $[0, 1]$. All thresholds were fixed on authentic frames before any plausibility label was consulted, and none is tuned on GenCQA-DB ratings.

E Observability Validation and Stratified Results

The image-level observability O_f is compared against the human judgeability labels of Sec. 3.2 in Table 3. The upper block reports the AUROC of O_f as a score for the label that a cue cannot be decided, together with the fraction of images so labeled on each factor subset. The lower block reports PPS prediction on observability terciles of the camera-disjoint test set, where each image is assigned to a tercile by the O_f of its violated factor. The margin between PACT and the strongest equally supervised baseline grows from 0.080 on the high tercile to 0.155 on the low tercile, and the model without confidence maps loses most of its advantage precisely on the low tercile.

F Source-matched Intervention Protocol

Both intervention sets are built from held-out source groups that appear in none of the training partitions. For content interventions, 400 authentic frames are paired with a version of the same frame into which one violation of medium strength is inserted by a pipeline seen during training, so that the only change between the two members of a pair is the presence of the violation. For visibility interventions, 400 violated frames are paired with a degraded version produced by downscaling to

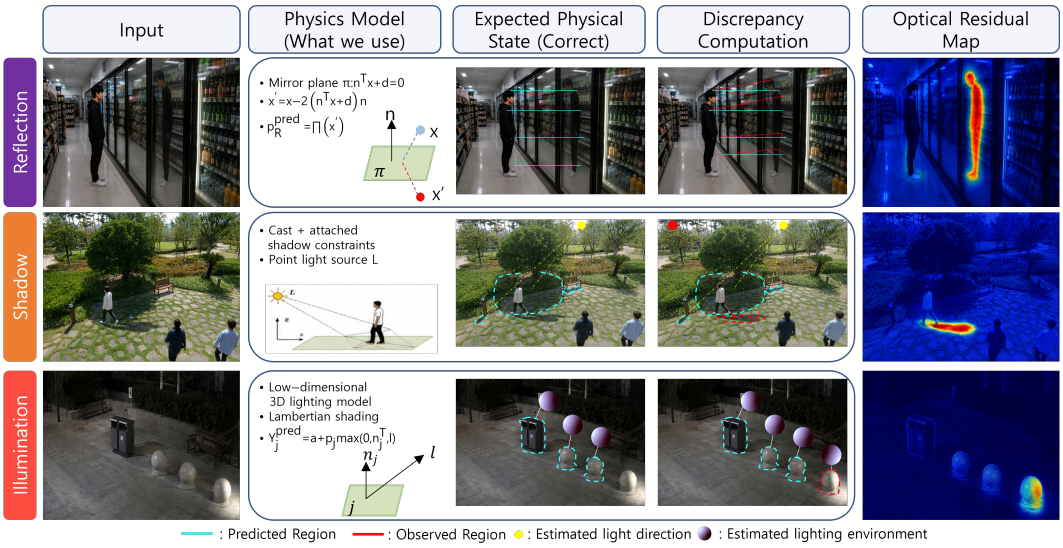

Figure 7: Optical residual map generation. Each factor renders the expected state from estimated scene parameters and maps the discrepancy with the observation into a residual map. The same estimation stage yields the availability and reliability terms of the cue confidence maps.

Table 3: Validation of cue confidence against human judgeability and PPS prediction by observability tercile.

	Reflection	Shadow	Illumination
Not decidable fraction	11%	14%	38%
AUROC of O_f	0.912	0.897	0.884
	High O_f	Middle O_f	Low O_f
Q-Align (ft)	0.822	0.790	0.641
PACT w/o confidence	0.889	0.848	0.702
PACT	**0.902**	**0.871**	**0.796**

one half or one quarter resolution, JPEG recompression at quality 30 or 15, and gamma darkening with exponents 1.6 or 2.2, with one level of each operation sampled per pair, so that the violation remains present while its evidence fades. All 1,600 images were rated in the plausibility session protocol of Sec. 3.2, including the judgeability question, by participants disjoint from those who rated the main subset. Per-pair changes are computed as the score of the intervened member minus the score of the original member, and pair SRCC is the Spearman correlation between predicted and human per-pair changes. Full results are reported in Table 4.

G Training Details

PACT is trained for 50 epochs with AdamW, a learning rate of 10^{-4} under cosine decay, and a batch size of 32 on four A6000 GPUs. The loss weights are $\lambda = 0.5$ and $\gamma = 0.1$, and the transformer uses $N = 6$ blocks with 6 attention heads and embedding dimension $D = 384$ over the $9 \times 576 + 1$ token sequence. The equally supervised baselines are fine-tuned end to end on the same partitions with the same loss on PPS, learning rate 10^{-5}, and early stopping on the validation SRCC.

Table 4: Source-matched interventions. Content interventions insert a violation into an authentic frame and visibility interventions degrade a violated frame from the same camera. ΔPPS is the mean change of the human or predicted score, and pair SRCC measures how well per-pair predicted changes track the human changes.

	Content intervention		Visibility intervention	
	ΔPPS	Pair SRCC	ΔPPS	Pair SRCC
Human	−0.312	–	+0.184	–
Q-Align (ft)	−0.203	0.612	+0.091	0.338
PACT w/o confidence	−0.287	0.793	+0.356	0.421
PACT, gated confidence	−0.281	0.784	+0.231	0.627
PACT	**−0.295**	**0.812**	**+0.176**	**0.741**

H Quality Assessment on Authentic Imagery

The auxiliary MOS output of PACT is evaluated on the authentic subset against dedicated IQA models (Ke et al., 2021; Yang et al., 2022; Wang et al., 2023a; Zhang et al., 2023; Wu et al., 2024) retrained on the same quality labels, as reported in Table 5.

Table 5: Quality MOS prediction on the authentic subset.

Method	SRCC	PLCC	KRCC
MUSIQ	0.861	0.874	0.672
MANIQA	0.879	0.890	0.694
CLIP-IQA+	0.868	0.881	0.681
LIQE	0.884	0.895	0.701
Q-Align	**0.897**	**0.908**	**0.716**
PACT (ours)	0.891	0.902	0.709

I Edit-support Localization Results

Localization performance against forgery localization baselines (Liu et al., 2022; Guo et al., 2023; Guillaro et al., 2023) retrained on GenCQA-DB masks is reported in Table 6.

Table 6: Edit-support localization on GenCQA-DB.

Method	IoU	AUPRC
PSCC-Net	0.376	0.458
HiFi-Net	0.401	0.487
TruFor	0.412	0.503
PACT (ours)	**0.583**	**0.671**

J Additional Qualitative Results

Additional qualitative results across the three violation types are shown in Fig. 8, ordered by predicted PPS with per-factor violation magnitudes.



Figure 8: Additional qualitative results ordered by predicted PPS with per-factor violation magnitudes $(E_R/E_S/E_I)$.

K Dissociation Structure in Ratings and Predictions

The dissociation structure in human ratings and its reproduction by PACT are visualized in Fig. 9. Human ratings couple MOS and PPS on real frames and decouple them on generated frames, and PACT predictions reproduce the same structure.

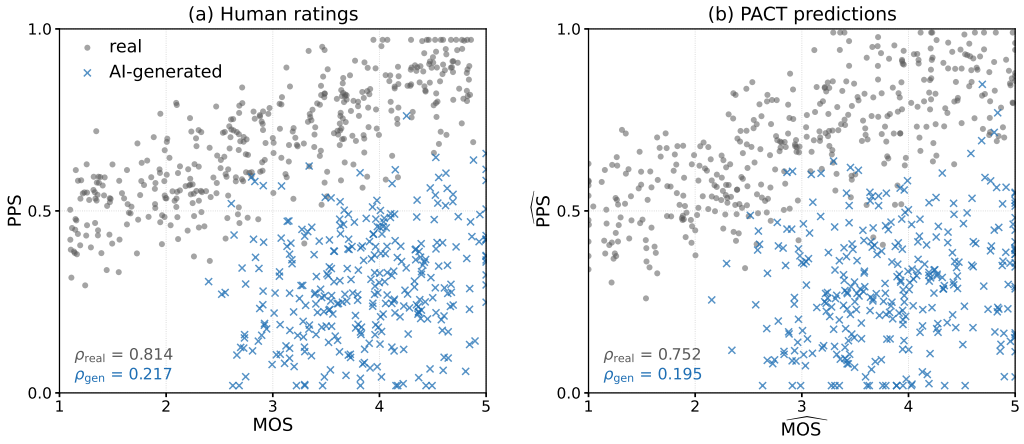

Figure 9: Dissociation structure in human ratings and PACT predictions. Real frames couple the two axes and generated frames decouple them.