ENTROPY FLOW GRAPHS: ANALYTIC SPATIOTEMPORAL GRAPH REPRESENTATIONS FOR NO REFERENCE VIDEO QUALITY ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Video quality failures have spatiotemporal structure, yet no reference video quality assessment (NR-VQA) collapses every failure into a single scalar, and no existing benchmark provides human judgments of quality below the frame level in video. We address both gaps. First, we introduce the Entropy Flow Graph (EFG), a quality representation that fixes its structure analytically and confines learning to three matrices. Region nodes carry entropies of five cues, spatial edges measure the conditional entropy deviation of boundary joint statistics from natural scenes, and motion compensated temporal edges let the aligned past of the video serve as its own reference, so where and when a degradation strikes are read off the edges themselves. The analytic structure carries frame rate and permutation invariance by construction, a dominance decomposition that separates global from transient degradations, and a consistency bound that fixes the histogram bin count. Second, we introduce D-CQA, $60{,}000$ clips with human region level quality scores and traced degradation extents, the first benchmark for spatiotemporal quality localization. On D-CQA EFG attains a global SRCC of 0.897 against 0.831 for the strongest holistic baseline with a localization AUC of 0.823, transfers without retraining to KoNViD-1k and LIVE-VQC, and, serialized as context, lifts the region level quality reasoning of a multimodal LLM by 12.6 points.

1 INTRODUCTION

A video rarely fails everywhere at once. In Fig. 1 blur softens every region of every frame, a noise burst strikes one region for a single instant, and a pixelated patch persists in one corner throughout, yet the frame level pooling that dominates no reference video quality assessment (NR-VQA) returns three nearly identical scalars, 0.412, 0.443, and 0.431. What pooling erases is exactly what a downstream system needs first, where a degradation lives and when it strikes.

Graph structure has entered quality assessment from several directions, but none reaches this need. GraphIQA (Sun et al., 2023) builds graphs over sample features that never touch the image plane, a second family passes messages over patches, viewports, points, or whole frames with learned similarity edges (Liu & Guo, 2025; Xu et al., 2021; Fu et al., 2022; Shan et al., 2024; Pan et al., 2021), and PANDA (Janjua et al., 2026) grounds a graph in regions but for pairwise comparison of static images against a presumed anchor. Holistic NR-VQA models such as FAST-VQA (Wu et al., 2022) and DOVER (Wu et al., 2023b) output a single opaque score. No existing representation is at once region first, temporal, free of any reference, and equipped with closed form edge semantics (Table 6 of Appendix E), and no public video corpus provides human quality judgments below the frame level.

We close both gaps. The Entropy Flow Graph (EFG) inverts the usual recipe and fixes the structure analytically while confining learning to a minimal readout. Nodes are image regions carrying entropies of five cues, spatial edges are weighted by conditional entropy deviation, the closed form departure of a region's joint statistics with its neighbors from natural scene regularities, and temporal edges are weighted by motion compensated conditional entropy, so the aligned past of the video acts as its own reference. Because every weight has a fixed meaning, the edges themselves are the localization output, spatial edges lighting up where a degradation lives and temporal edges when it

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.

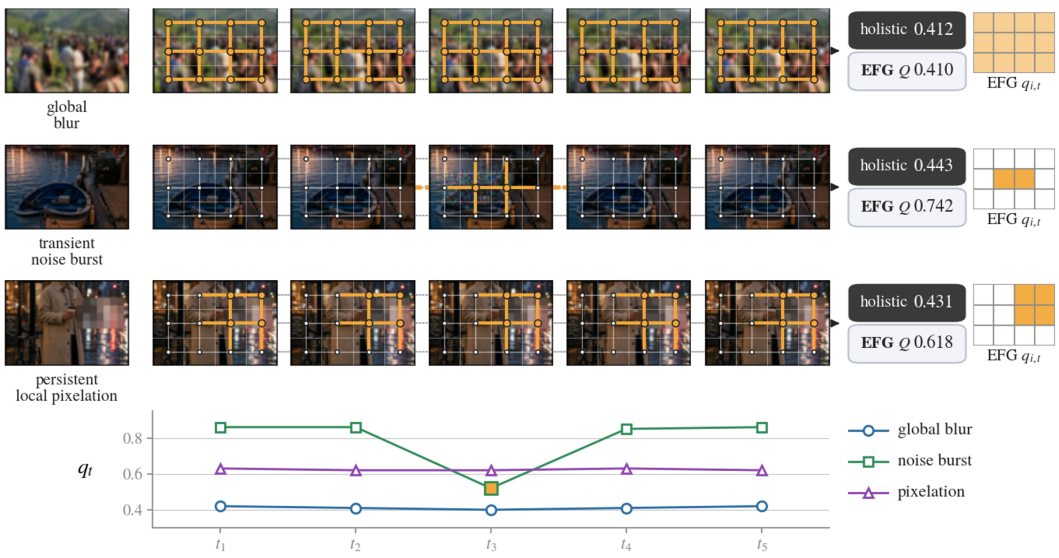


Figure 1: Three videos, two readings. Each row shows one failure type with the entropy flow graph drawn on five resampled frames, solid edges spatial and dashed edges temporal, highlighted where they depart from their clean value. A holistic model returns one nearly identical number per video (dark boxes), whereas EFG returns Q, the region scores $q_{i,t}$, and the frame scores q_t.

appears, as Fig. 1 shows, and a readout with three weight matrices turns the graph into region, frame, and global scores in one pass. The analytic structure carries frame rate and permutation invariance by construction, a dominance decomposition that separates spatial from temporal evidence, and an estimator consistency bound that fixes the histogram bin count. Evaluating such a representation requires labels that did not exist, so we also collected them. Our contributions are the following.

- **Representation and analysis.** We formulate NR-VQA as inference over a spatiotemporal entropy graph whose nodes and edges are closed form information functionals, with invariances by construction, a dominance decomposition, and a consistency bound yielding the bin count rule $B^2 \leq n/4$.
- **Instantiation with localization.** A three matrix readout emits region, frame, and global scores in a single pass, and the edge weights double as where and when maps with no threshold and no extra parameters.
- **D-CQA benchmark.** We introduce 60,000 clips from 3,000 sources with human region level quality scores and degradation masks, released with fixed splits and metrics. On it EFG attains 0.897 global SRCC against 0.831 for the strongest holistic baseline and 0.823 localization AUC, and transfers to KoNViD-1k and LIVE-VQC without retraining.
- **EFG as analytic context for multimodal LLMs.** Serialized into structured context, the graph lifts the region level reasoning of GPT-4o (OpenAI, 2024) by 12.6 points and transfers more faithfully than the learned categorical graph of Janjua et al. (2026).

2 Related Work

Graphs and holistic models in quality assessment. The graph families of Section 1 play three roles, feature space graphs over samples (Sun et al., 2023), unit graphs over patches, viewports, points, or frames (Liu & Guo, 2025; Xu et al., 2021; Fu et al., 2022; Shan et al., 2024; Pan et al., 2021), and the region grounded pairwise graph of Janjua et al. (2026). In every case the edges are learned labels or similarities without closed form meaning, and none localizes in video (Table 6). Holistic NR-VQA regresses one score from embeddings, from recurrent aggregation (Li et al., 2019) to FAST-VQA (Wu et al., 2022), DOVER (Wu et al., 2023b), and multimodal scoring in Q-Align (Wu et al., 2024), so localization is outside its design.

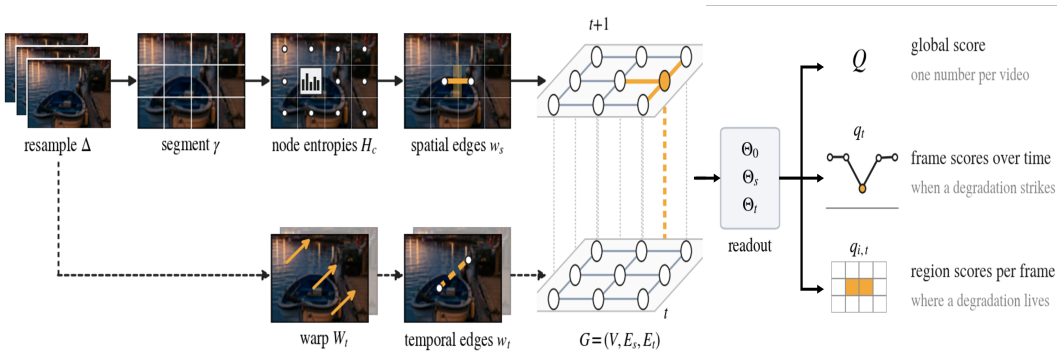


Figure 2: Construction of the entropy flow graph. The spatial lane (solid) resamples, segments, computes five cue node entropies H_c, and weights spatial edges by CED w_s. The temporal lane (dashed) warps each frame onto the next and weights temporal edges by conditional entropy w_t. A readout with three matrices returns Q, q_t, and $q_{i,t}$.

Entropy evidence. A classical line established entropy departures from natural scene statistics as degradation evidence, in spatial and spectral entropies (Liu et al., 2014), NSS features (Mittal et al., 2012; 2013), and the free energy view (Gu et al., 2015), but always pooled into global descriptors. EFG distributes the same evidence over a graph so it becomes structured and localized.

Benchmarks and structured context. KoNViD-1k (Hosu et al., 2017) and LIVE-VQC (Sinno & Bovik, 2019) carry one opinion score per video and CQA-DB (Han et al., 2025) provides image level scores for CCTV content, so none offers human judgments below the frame level, the gap D-CQA fills. Janjua et al. (2026) showed that a predicted distortion graph fed as reasoning context improves the region wise reasoning of multimodal LLMs. Their scaffold is a learned categorical graph, and we test whether analytic structure transfers better.

3 Entropy Flow Graph

3.1 Overview and Notation

EFG casts NR-VQA as inference over a spatiotemporal graph built from closed form information measures, constructed along the two lanes of Fig. 2. Uniform time resampling of the input video x with interval Δ yields frames $I_1, \ldots, I_T$, which anchors every temporal quantity to physical time rather than frame indices. A segmentation operator γ, a semantic segmenter snapped to a regular grid or a plain grid, partitions each frame into regions $\{m_{i,t}\}_{i=1}^{N_t}$, each of which becomes a node $v_{i,t}$. Spatial edges E_s connect regions that share a boundary within a frame, temporal edges E_t connect corresponding regions of consecutive resampled frames after warping, and $G = (V, E_s, E_t)$ is the entropy flow graph. Every node attribute and edge weight is an entropy of a local cue distribution, so learning enters only through the readout of Section 3.5.

3.2 Region Nodes with Five Cue Entropy Attributes

Each node carries the entropy of five cues, luminance, texture (local gradient orientation and magnitude), contrast (locally normalized responses), kinematic (temporal gradient magnitude after the motion compensation of Section 3.4), and spectral (band energy ratios of a local frequency decomposition). Let $b_c(p)$ denote the quantized response of cue $c \in \mathcal{C}$ at pixel p with B bins. The within region cue distribution is the normalized histogram

$$p_c(b \mid v_{i,t}) = \frac{1}{|m_{i,t}|} \sum_{p \in m_{i,t}} \mathbf{1}[\, b_c(p) = b \,], \tag{1}$$

and each cue contributes its Shannon entropy

$$H_c(v) = -\sum_{b=1}^{B} p_c(b \mid v) \log p_c(b \mid v), \tag{2}$$

so that the node attribute vector is $\phi(v_{i,t}) = [\, H_c(v_{i,t}) \,]_{c \in \mathcal{C}}$. This follows the entropy evidence of Section 2 but keeps each cue attached to a region rather than pooled globally.

3.3 Spatial Edges from Conditional Entropy Deviation

A spatial edge $(i, j, t) \in E_s$ exists whenever $m_{i,t}$ and $m_{j,t}$ share a boundary. Along a band of width β pixels around the boundary we form, for each cue, the joint distribution of paired responses

$$p_c^{ij}(b, b') = \frac{1}{|\mathcal{B}_{ij}|} \sum_{(p,q) \in \mathcal{B}_{ij}} \mathbf{1}[\, b_c(p) = b,\ b_c(q) = b' \,], \tag{3}$$

where $\mathcal{B}_{ij}$ is the set of pixel pairs facing each other across the boundary, and the conditional entropy of region i given region j

$$H_c(v_i \mid v_j) = -\sum_{b,b'} p_c^{ij}(b, b') \log \frac{p_c^{ij}(b, b')}{\sum_{\tilde{b}} p_c^{ij}(\tilde{b}, b')}. \tag{4}$$

Natural scenes exhibit lawful joint statistics between adjacent regions, which we model by a fixed regressor $\hat{H}_c(\cdot)$ fitted once on undistorted content and frozen (Appendix A.1). The spatial edge weight is the conditional entropy deviation (CED)

$$\begin{aligned} w_s(i, j, t) &= \frac{1}{2|\mathcal{C}| \log B} \sum_{c \in \mathcal{C}} \big(d_c(i \mid j) + d_c(j \mid i) \big), \\ d_c(i \mid j) &= \big| H_c(v_i \mid v_j) - \hat{H}_c(v_j) \big|, \end{aligned} \tag{5}$$

which is symmetric, normalized to $[0, 1]$, and large exactly where the joint statistics depart from the natural regularity. A global degradation lifts w_s on most edges while a local one lifts only the edges incident to it, the spatial half of the dominance decomposition of Section 4.

3.4 Temporal Edges from Motion Compensated Conditional Entropy

A temporal edge connects $v_{i,t}$ to the region of maximal overlap at $t{+}1$ once frame t is aligned by a two stage warp W_t, a global translation from phase correlation followed by residual dense optical flow. Writing $\tilde{v}_{i,t} = W_t(v_{i,t})$, the temporal edge weight is the cue averaged conditional entropy of the present given the aligned past,

$$w_t(i, t) = \frac{1}{|\mathcal{C}| \log B} \sum_{c \in \mathcal{C}} H_c(v_{i,t+1} \mid \tilde{v}_{i,t}), \tag{6}$$

with pixels lacking a valid correspondence masked out. The past of the video serves as its own reference, which is what frees EFG of any reference. Fig. 3 shows why W_t is indispensable. Under a camera pan the raw residual yields w_t up to 0.269 while the warped residual stays at most 0.038, and an injected photometric transient survives as a localized glow, the temporal half of the dominance decomposition. By design w_t also responds to genuine content change that the warp cannot absorb, a limit discussed in Section 6.

3.5 Lightweight Graph Readout and Multilevel Scores

Because the edges already localize the evidence, the readout stays small. Node states are initialized as $h_{i,t}^{(0)} = [\, \phi(v_{i,t}),\ \bar{w}_s(i, t),\ \bar{w}_t(i, t) \,]$ with the mean incident spatial and temporal weights, and L rounds of edge weighted message passing update them as

$$\begin{aligned} h_{i,t}^{(l+1)} = \rho\Big(\Theta_0\, h_{i,t}^{(l)} &+ \sum_{j \in \mathcal{N}_s(i,t)} \tilde{w}_s(i, j, t)\, \Theta_s\, h_{j,t}^{(l)} \\ &+ \tilde{w}_t(i, t{-}1)\, \Theta_t\, h_{i^-,t-1}^{(l)} + \tilde{w}_t(i, t)\, \Theta_t\, h_{i^+,t+1}^{(l)} \Big), \end{aligned} \tag{7}$$

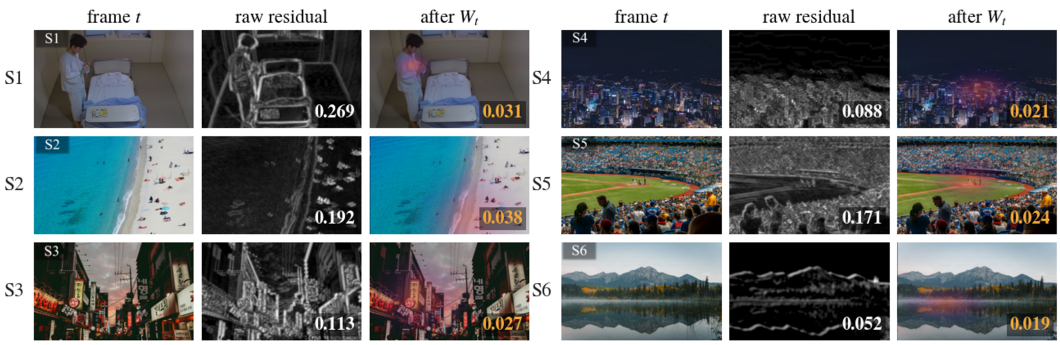


Figure 3: Motion compensation on six scenes. Each row shows the frame, the raw inter frame residual, and the residual after the warp W_t, with w_t in the corner (white before, amber after). Camera motion alone yields w_t up to 0.269, the warp brings every scene to at most 0.038, and an injected transient survives as a localized glow.

where $\mathcal{N}_s(i,t)$ are the spatial neighbors, i^- and i^+ the regions matched at $t{-}1$ and $t{+}1$ (dropped at $t = 1$ and $t = T$), $\tilde{w}$ weights normalized by the sum of incident weights of the same type, ρ the ReLU, and $\Theta_0, \Theta_s, \Theta_t$ the only weight matrices in the model. Region, frame, and global scores follow in one pass,

$$q_{i,t} = \sigma\left(\theta^\top h_{i,t}^{(L)}\right), \qquad q_t = \sum_{i=1}^{N_t} \omega_{i,t}\, q_{i,t}, \tag{8}$$

$$Q = (1-\eta)\,\frac{1}{T}\sum_{t=1}^{T} q_t + \eta\, \mathrm{P}_{10}\big(\{q_t\}_{t=1}^{T}\big), \tag{9}$$

with σ the logistic function, $\omega_{i,t}$ area proportional weights, and P_{10} the tenth percentile, so that η blends mean opinion with the worst one to three seconds of a 10 to 30 second clip ($\eta = 0.3$ from the sweep of Table 5). Training minimizes

$$\mathcal{L} = \big(Q - y\big)^2 + \lambda_r \sum_{(i,t)\in\mathcal{R}} \big(q_{i,t} - y_{i,t}\big)^2, \tag{10}$$

where y is the global label and $\mathcal{R}$ the set of regions with labels $y_{i,t}$, empty for global only corpora. Algorithm 1 in Appendix A summarizes the procedure. Graph construction is linear in the pixel count for fixed B, and the three small matrices keep the model far below the parameter count of holistic NR-VQA backbones.

3.6 Localization Output from Edge Maps

Rasterizing $w_s(i,j,t)$ onto boundaries and region supports yields a spatial map $M_s(t)$ that marks where the joint statistics break, and stacking $w_t(i,t)$ over time yields a temporal profile $M_t(i)$ that marks when a region departs from its aligned past. Both maps need no threshold and no extra parameters and are evaluated against the annotated degradation extents in Section 5.

4 Theoretical Analysis

Because every node attribute and edge weight of G is a closed form functional of local histograms, properties that a learned graph could only be trained toward hold for EFG either by construction or as consequences of standard entropy estimation theory. We state them here with the idea of each argument. Complete proofs and the assumptions each one requires are given in Appendix C, and Fig. 4 verifies Propositions 1, 3, and 5 empirically.

4.1 Invariances by Construction

Proposition 1 (Frame Rate Invariance). *Let two encodings of the same underlying video be provided at frame rates $f_1, f_2 \geq 1/\Delta$ such that both contain frames at every sampling instant $k\Delta$,*

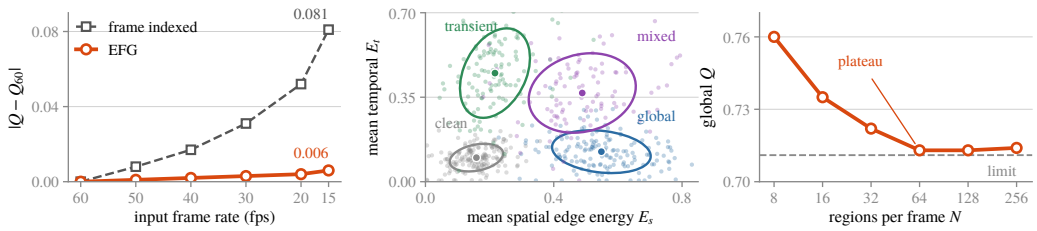


Figure 4: Empirical verification. Left, frame rate invariance, EFG drifts by at most 0.006 down to 15 fps against 0.081 for a frame indexed variant. Middle, dominance decomposition (Prop. 3), the four clip classes separate in the plane of mean spatial versus temporal edge energy with overlapping tails. Right, refinement stability (Prop. 5), Q plateaus beyond 64 regions per frame.

$k = 1, \ldots, T$. *Then the two entropy flow graphs coincide, and so do all region, frame, and global outputs.*

Proposition 2 (Permutation Invariance). *Let π_t be any relabeling of the region indices within each frame. Then q_t and Q are unchanged, and the region scores are permuted by the same π_t.*

Idea. Both statements are structural. The construction reads the input only through the resampled frames and the readout is a deterministic map of G, which gives Proposition 1, and relabeling acts on G as a graph isomorphism under which the shared parameter readout of Eq. (7) is equivariant, which gives Proposition 2. Proposition 1 still has a practical consequence that frame indexed methods lack. When the rate condition fails and the nearest frame is taken, Appendix C.1 bounds the output change by the modulus of continuity of the cue fields at lag $1/(2f)$, and the left panel of Fig. 4 shows the resampled EFG within 0.006 of its reference down to 15 fps while a frame indexed variant drifts by 0.081.

4.2 Dominance Decomposition

Proposition 3 (Spatial Temporal Dominance Decomposition). *Let g be a degradation operator applied to the resampled frames. (i) If g is time constant and acts per pixel as a bijection on the cue bins, then every temporal weight w_t in Eq. (6) equals its clean value. (ii) If g acts only on region $m_{i,t+1}$ at a single time step, then exactly the spatial edges incident to $v_{i,t+1}$ and the temporal edges incident to it change, and all remaining weights of G are unchanged.*

Idea. A bin bijection relabels the support of the joint histogram and leaves conditional entropies unchanged, and locality of Eqs. (3) to (6) gives (ii). Both statements are exact only in idealized cases. A persistent monotone tone mapping that respects bin boundaries satisfies (i), but blur, noise, compression, and haze are neither bin bijective nor perfectly localized, so for them the separation holds in expectation rather than pointwise. This is the overlapping tail structure in the middle panel of Fig. 4, and it is also why persistent noise, whose realizations are independent across frames, still lifts w_t and keeps the temporal edges informative even when no transient event occurs, the effect quantified in Section 5.3. The three rows of Fig. 1 are the three regimes of the proposition, spatial edges alone for global blur, spatial and temporal edges together for the transient burst, and spatial edges alone again for persistent pixelation.

4.3 Estimator Consistency and the Bin Count Rule

Proposition 4 (Plug In Consistency). *Assume the samples used to estimate each histogram are i.i.d. Let $\hat{w}$ denote any weight of G computed from n samples, namely $n = |m_{i,t}|$ for node entropies, $n = |\mathcal{B}_{ij}|$ for spatial edges, and the count of validly matched pixels for temporal edges. Then $\hat{w} \to w$ almost surely as $n \to \infty$, and*

$$\left| \mathbb{E}[\hat{w}] - w \right| \le \frac{B^2}{2n \log B}, \qquad \mathrm{sd}(\hat{w}) \le \frac{c_2}{\sqrt{n}}, \tag{11}$$

with c_2 depending only on the cue set and $\log B$.

Table 1: Main results on D-CQA and transfer without retraining (SRCC), means over three seeds. † frame by frame adaptation, ‡ first frame anchor adaptation. Dashes mark capabilities a family lacks by design.

	D-CQA				Transfer without retraining (SRCC)		
Method	SRCC	PLCC	Region SRCC	Loc. AUC	CQA-DB	KoNViD-1k	LIVE-VQC
NIQE (Mittal et al., 2013)	0.583	0.591	–	–	–	–	–
VSFA (Li et al., 2019)	0.762	0.774	–	–	–	–	–
GraphIQA† (Sun et al., 2023)	0.786	0.798	–	–	–	–	–
FAST-VQA (Wu et al., 2022)	0.818	0.826	–	–	0.809	0.784	0.741
DOVER (Wu et al., 2023b)	0.831	0.842	–	–	0.822	0.801	0.766
PANDA‡ (Janjua et al., 2026)	0.694	0.703	0.518	0.647	–	–	–
EFG (ours)	**0.897**	**0.905**	**0.764**	**0.823**	**0.861**	**0.842**	**0.795**

Idea. The bias expansion of Miller and Madow bounds the plug in bias of an entropy over $K \leq B^2$ cells by $(K-1)/(2n)$ nats, and a bounded difference argument gives the fluctuation. The i.i.d. assumption is an idealization, since neighboring pixels are correlated and the effective sample size is smaller than the pixel count, so the bounds should be read as scaling laws in B and n.

Bin count rule. Requiring the worst case bias of an unnormalized joint entropy, $B^2/(2n)$ nats, to stay below $1/8$ nat (about 6% of the range $\log 8$ at $B = 8$) gives $B^2 \leq n/4$. The smallest sample in the graph is the boundary band of a spatial edge. At the operating point of Section 5.1, a 16×16 grid on 540p frames with band width $\beta = 6$ yields $|\mathcal{B}_{ij}|$ between about 260 and 360 pixel pairs, so the rule admits $B \leq 8$ for every spatial edge, which fixes $B = 8$. Node histograms have $n = |m_{i,t}| \approx 2{,}000$ pixels and would admit $B \leq 22$, so $B = 16$ is safe for nodes but not for edges, and $B = 32$ violates the rule everywhere, which is the drop seen in Fig. 7 (c).

Proposition 5 (Segmentation Refinement Stability). *If γ' refines γ with maximal region diameter δ and every region retains at least $n_{\min}$ pixels, then $|Q(\gamma') - Q(\gamma)| \leq C_1\, \omega(\delta) + C_2\, B^2/n_{\min}$, where ω is the modulus of continuity of the cue fields.* The bound combines the total variation perturbation of split histograms with the estimation term of Proposition 4. The predicted plateau appears in the right panel of Fig. 4 beyond 64 regions per frame, and the full statement and proof are in Appendix C.5.

5 Experiments

5.1 Benchmark and Setup

Region level claims require region level human labels, which no existing video corpus provides, so we collected D-CQA. It extends the CCTV domain of CQA-DB (Han et al., 2025) to video with 3,000 source clips of 10 to 30 seconds, each degraded by five distortion types (blur, noise, compression, darkening, haze) at four severity levels, yielding 60,000 clips of which 40% carry a global, 35% a persistent local, and 25% a transient local degradation. Twenty four raters scored every 4×4 tile of the sampled frames on a five point ACR scale, at least five ratings per tile, and traced the spatial extent and onset of each degradation. Tile means define $y_{i,t}$, the clip mean defines y, and inter rater agreement is 0.861 SRCC at the tile level and 0.912 at the clip level. Appendix B gives the recording, session, and split details.

We report SRCC and PLCC for global scores, SRCC over labeled tiles, and the area under the ROC curve of M_s against the traced extents (Loc. AUC), and we evaluate the global score without retraining on KoNViD-1k (Hosu et al., 2017), LIVE-VQC (Sinno & Bovik, 2019), and the image level CQA-DB, where each image is a clip with $T = 1$ and the temporal terms of Eq. (7) dropped. EFG uses $\Delta = 100$ ms, SAM masks snapped to a 16×16 grid (Kirillov et al., 2023), $B = 8$, $\beta = 6$, $L = 2$, $\eta = 0.3$, and $\lambda_r = 0.3$. All numbers are means over three seeds with standard deviations below 0.006. Baselines are retrained on the D-CQA training split under the identical protocol, with GraphIQA applied frame by frame and PANDA anchored on the first frame.

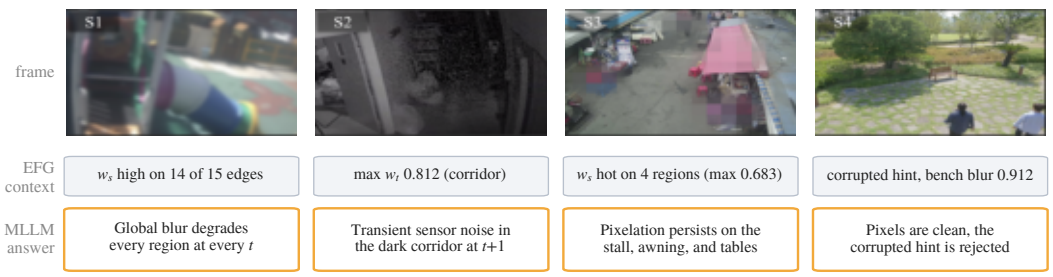


Figure 5: EFG as context for a multimodal LLM. Four test clips with the hottest serialized entry of the frozen graph and the model answer. Spatial evidence names the region, temporal evidence names the time step, and a deliberately corrupted entry (S4) is overridden by pixel evidence.

Table 2: Key ablations on D-CQA, deltas against the full model. The complete table with all cue removals and per type effects is Table 4.

Variant	ΔGlobal SRCC	ΔRegion SRCC	ΔLoc. AUC
w/o temporal edges E_t	-0.171	-0.148	-0.202
w/o spatial edges E_s	-0.093	-0.187	-0.174
w/o motion compensation W	-0.118	-0.061	-0.129
w/o kinematic cue	-0.087	-0.052	-0.066
readout depth $L = 1$	-0.024	-0.029	-0.011

5.2 Results

EFG outperforms every baseline on every measure in Table 1. On the global score it reaches 0.897 SRCC against 0.831 for DOVER, the strongest holistic model, with 0.08M learnable parameters against tens of millions for the holistic backbones (Appendix H). Only the region grounded PANDA adaptation can produce region scores or localization at all, and its first frame anchor degrades as content drifts, leaving it at 0.518 region SRCC and 0.647 AUC where EFG attains 0.764 and 0.823. Every dashed entry is a capability the corresponding family lacks by design, the gap of Table 6. Because nothing in the graph is fit to the training distribution, EFG also transfers without retraining, holding 0.861 SRCC on CQA-DB, 0.842 on KoNViD-1k, and 0.795 on LIVE-VQC, ahead of the holistic baselines on all three targets even though the in the wild sets contain authentic distortions never seen in D-CQA. Models trained in domain on KoNViD-1k reach higher SRCC, for example 0.909 for DOVER (Wu et al., 2023b), which we do not claim to match. The node states are also organized, a t-SNE embedding separates the five distortion types more clearly than the holistic backbone and orders severity levels monotonically without metric learning (Appendix F).

5.3 Ablation

Table 2 accounts for each design choice. Removing the temporal edges is the most damaging cut at -0.171 global SRCC and -0.202 AUC, and not only because of transient events. Persistent noise and compression produce realizations that are independent or block shifted across frames, so their conditional entropy given the aligned past stays high at every step, and the per type breakdown in Table 4 shows the loss concentrated on noise (-0.214) and compression (-0.163) rather than blur (-0.052). Removing the spatial edges hurts the region score most at -0.187, so the two halves of the dominance decomposition of Section 4 appear as complementary failure modes. Disabling motion compensation costs -0.118, the quantitative echo of the spurious hot edges in Fig. 3, the kinematic cue is the most valuable of the five, and the readout saturates by $L = 2$.

5.4 EFG as Context for Multimodal LLMs and Localization

Serializing the frozen graph into the prompt of GPT-4o (OpenAI, 2024) as structured text, with the instruction that the graph is a hint and pixel evidence has priority, tests whether analytic structure transfers faithfully. Table 3 shows EFG context lifting region level distortion identification by 12.6

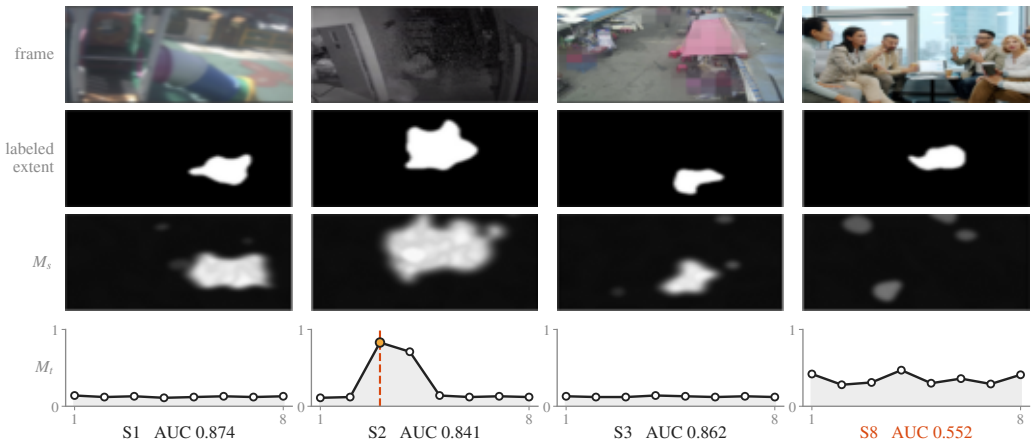


Figure 6: Localization on four scenes, showing the degraded frame, the traced extent, the spatial map M_s, and the temporal profile M_t with the labeled onset dashed. Persistent degradations keep M_t flat, the transient burst (S2) spikes it at the onset, and S8 is the failure case where busy texture masks a local patch.

Table 3: MLLM showcase on 600 D-CQA test clips with GPT-4o. Hint rejection is the rate at which a deliberately corrupted context entry is overridden by pixel evidence.

Context	Distortion ID acc.	Region SRCC	Hint rejection (%)
none	0.412	0.318	–
learned graph (Janjua et al., 2026)	0.487	0.402	58.3
EFG (ours)	**0.538**	**0.451**	**76.4**

points over the model without context and by 5.1 points over the learned categorical graph of Janjua et al. (2026) serialized under the identical format, with region quality SRCC in the same order, and the ordering holds for the open weight Qwen2.5-VL 72B (0.371, 0.428, 0.469). The cases in Fig. 5 show the mechanism, the model grounds its answer in the hottest serialized edge, naming the region for spatial evidence and the time step for temporal evidence, and it rejects a deliberately corrupted EFG hint in 76.4 percent of probes against 58.3 for the learned graph, which suggests that numerically meaningful weights are easier to check against pixels than categorical labels. The maps behind the 0.823 localization AUC are shown in Fig. 6. Persistent degradations keep M_t quiet while M_s covers the traced extent, the transient burst spikes M_t at the labeled onset, and the failure case shows busy texture masking a local patch, the same limit that caps the region score in Table 1. Complete graphs over real frame pairs are drawn in Appendix G.

6 Conclusion

We introduced the Entropy Flow Graph, a quality representation that fixes its structure analytically and confines learning to three matrices, and D-CQA, the first video corpus with human region level quality labels and degradation extents. Spatial edges weighted by conditional entropy deviation tell where a degradation lives and motion compensated temporal edges tell when it strikes, so the representation is region first, temporal, free of any reference, and localizing at once. On D-CQA EFG reaches 0.897 global SRCC and 0.823 localization AUC, transfers to in the wild benchmarks without retraining, and lifts the region level reasoning of a multimodal LLM by 12.6 points as serialized context.

The remaining limits are texture masking of local degradations, the response of temporal edges to genuine content change, dependence on the segmenter, and scene cuts that break the temporal self reference. Richer node semantics, flow confidence weighting of w_t, joint training of the segmenter, and cut aware temporal edges are natural remedies, and extending the D-CQA annotation protocol to authentic distortions is the next step.

AI USE STATEMENT

In this work, we used generative AI tools solely as a language aid. A large language model was used to correct grammar and to improve the fluency of English prose in passages written by the authors, and every suggested edit was reviewed and accepted or rejected by the authors.

ETHICS STATEMENT

This work studies quality assessment on video, and part of the evaluation uses surveillance footage in which people appear incidentally. The region level labels of D-CQA were collected in a subjective study with twenty four adult raters recruited from the authors' institution, who gave written informed consent, were compensated at the local hourly rate, and judged imaging quality only. The study protocol was approved by the institutional review board of the authors' institution. No attribute of any individual appearing in the footage is inferred or annotated, the footage was collected under the terms of the source corpora, individuals are neither identified nor tracked, and the released material consists of split definitions, region masks, quality annotations, and evaluation code rather than any new recording. Faces and other identifying detail are avoided in the qualitative figures wherever the content is degraded for demonstration.

Quality assessment for surveillance carries a dual use consideration, since a method that reports where and when a camera is degraded can improve the reliability of a monitoring system regardless of the purpose of that system. EFG signals fidelity and localizes degradations. It does not detect, recognize, or characterize people or activities, and it produces no biometric output. The region masks and quality labels used here describe imaging conditions rather than persons, so the fairness risk is concentrated in whether the model degrades unevenly across content types, which is why we report the per category breakdown of Appendix D and note the reduced margin on heavily textured content in Section 6. We declare no conflicts of interest and no sponsorship that influenced the design, execution, or reporting of this study.

REPRODUCIBILITY STATEMENT

Every component of EFG is deterministic given its inputs, and the pieces needed to reproduce our numbers are distributed across the paper as follows. The construction is specified by Eqs. (1) to (10) in Section 3 and by Algorithm 1 in Appendix A, which lists the full procedure together with the constants used in every experiment. All hyperparameters, the segmenter, the datasets, the splits, and the evaluation measures are stated in Section 5.1 and Appendix B, and the sweep grid behind the reported operating point is given in Table 5. Complete proofs of every theoretical claim, including the assumptions each one requires, are provided in Appendix C. The learned component consists of three matrices trained with three fixed seeds whose scores are averaged, and the runtime profile that characterizes the compute needed to repeat the experiments appears in Appendix H. We release the fixed splits, the region level localization metrics, the graph construction and readout implementation, and the serialization templates and prompts of the multimodal protocol as anonymous supplementary material.

REFERENCES

Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2011.

Baoliang Chen, Lingyu Zhu, Guo Li, Fangbo Lu, Hongfei Fan, and Shiqi Wang. Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):1903–1916, 2022.

Jun Fu, Chen Hou, Wei Zhou, Jiahua Xu, and Zhibo Chen. Adaptive hypergraph convolutional network for no-reference 360-degree image quality assessment. In *ACM International Conference on Multimedia (ACM MM)*, pp. 961–969, 2022.

Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. Using free energy principle for blind image quality assessment. *IEEE Transactions on Multimedia*, 17(1):50–63, 2015.

Yujin Han, Jiwoo Kang, Sanghoon Lee, and Taewan Kim. Understanding perceptual quality in CCTV images: A benchmark dataset and entropy-based insights. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3443–3450, 2025.

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 (QoMEX)*, pp. 1–6, 2017.

Muhammad Kamran Janjua, Abdul Wahab, and Bahador Rashidi. Panoptic pairwise distortion graph. In *International Conference on Learning Representations (ICLR)*, 2026.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12):5923–5938, 2019.

Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *ACM International Conference on Multimedia (ACM MM)*, pp. 2351–2359, 2019.

Lixiong Liu, Bao Liu, Hua Huang, and Alan Conrad Bovik. No-reference image quality assessment based on spatial and spectral entropies. *Signal Processing: Image Communication*, 29(8):856–863, 2014.

Yufan Liu and Jiefeng Guo. Image quality assessment by enabling inter-patch message passing via graph convolutional networks. *Neural Computing and Applications*, 37(8):6145–6161, 2025.

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

OpenAI. GPT-4o system card. `https://openai.com/index/gpt-4o-system-card/`, 2024.

Da Pan, Xueting Wang, Ping Shi, and Shaode Yu. No-reference video quality assessment based on modeling temporal-memory effects. *Displays*, 70:102075, 2021.

Ziyu Shan, Qi Yang, Rui Ye, Yujie Zhang, Yiling Xu, Xiaozhong Xu, and Shan Liu. GPA-Net: No-reference point cloud quality assessment with multi-task graph convolutional network. *IEEE Transactions on Visualization and Computer Graphics*, 30(8):4955–4967, 2024.

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

Simeng Sun, Tao Yu, Jiahua Xu, Wei Zhou, and Zhibo Chen. GraphIQA: Learning distortion graph representations for blind image quality assessment. *IEEE Transactions on Multimedia*, 25:2912–2925, 2023.

Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. A deep learning based no-reference quality assessment model for UGC videos. In *ACM International Conference on Multimedia (ACM MM)*, pp. 856–865, 2022.

Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan Conrad Bovik. UGC-VQA: Benchmarking blind video quality assessment for user generated content. *IEEE Transactions on Image Processing*, 30:4449–4464, 2021a.

Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan Conrad Bovik. RAPIQUE: Rapid and accurate video quality prediction of user generated content. *IEEE Open Journal of Signal Processing*, 2:425–440, 2021b.

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 (ECCV)*, pp. 538–554, 2022.

Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, Jinwei Gu, and Weisi Lin. Neighbourhood representative sampling for efficient end-to-end video quality assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15185–15202, 2023a.

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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, 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 *International Conference on Machine Learning (ICML)*, pp. 54015–54029, 2024.

Jiahua Xu, Wei Zhou, and Zhibo Chen. Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):1724–1737, 2021.

Appendix

A Full Procedure and Pristine Regressor

Algorithm 1 lists the complete construction and scoring procedure with the constants of Section 5.1.

Algorithm 1 EFG construction and scoring

Require: video x, interval Δ, segmenter γ, bins B, band β
1: resample x at interval Δ into $I_1, \dots, I_T$
2: segment each frame, $\{m_{i,t}\} \leftarrow \gamma(I_t)$
3: compute cue responses and node entropies $\phi(v_{i,t})$ by Eqs. (1) and (2)
4: **for** $t = 1, \dots, T$ **do**
5: form boundary joints and spatial weights w_s by Eqs. (3) to (5)
6: **end for**
7: **for** $t = 1, \dots, T-1$ **do**
8: estimate W_t (global translation, then residual flow), warp, match regions by overlap
9: compute temporal weights w_t by Eq. (6) with invalid pixels masked
10: **end for**
11: assemble $G = (V, E_s, E_t)$ and run L rounds of Eq. (7)
12: output $q_{i,t}$, q_t, Q by Eqs. (8) and (9) and the edge maps of Section 3.6

A.1 Pristine Regressor Fitting

For each cue $c \in \mathcal{C}$ the pristine conditional entropy $\hat{H}_c(\cdot)$ is a ridge regressor fit on boundary pairs drawn from the 500 images of BSDS500 (Arbeláez et al., 2011), which are held out from all quality experiments. The regressor takes the marginal entropy $H_c(v_j)$ of the conditioning region as its sole input and is frozen before any quality data is seen, so it contributes no learnable parameters to the quality pipeline. On held out BSDS500 boundaries the fit reaches R^2 between 0.812 (spectral) and 0.874 (luminance) with residual standard deviation between 0.061 and 0.083 nats, about 3 to 4 percent of the range $\log 8$. Refitting $\hat{H}_c$ on 200 clean CCTV frames of D-CQA instead of BSDS500 changes the D-CQA global SRCC by $+0.004$ and the localization AUC by $+0.006$, so the natural scene prior transfers to the CCTV domain with a negligible loss.

B D-CQA Details

The $3{,}000$ source clips were recorded by fixed CCTV cameras at 1920×1080 and 30 fps and span the six content categories of Fig. 7 (d). Local degradations cover between 4 and 25 percent of the frame, and transient ones last one to three seconds with the onset placed uniformly inside the clip. Ratings were collected in controlled sessions of at most 40 minutes on calibrated displays at $3H$ viewing distance, with training and outlier screening following ITU BT.500. Each rater saw a clip once at full resolution and then scored its 4×4 tiles on frames sampled every second, followed by tracing the degradation extent and marking its onset. Fixed train, validation, and test splits of $70/10/20$ percent are disjoint by source clip so that no scene appears in more than one split.

C Complete Proofs

Throughout the proofs we use the classical continuity of Shannon entropy. For two distributions p, q on B bins with total variation $\delta = \frac{1}{2}\|p - q\|_1 \le \frac{1}{2}$,

$$\big|H(p) - H(q)\big| \;\le\; \delta \log(B-1) + h_b(\delta), \tag{12}$$

where h_b is the binary entropy, so entropy is uniformly continuous in total variation with modulus $\Lambda_B(\delta) = \delta \log(B-1) + h_b(\delta)$.

C.1 Proof of Proposition 1 (Frame Rate Invariance)

Let two encodings of the same underlying video contain frames at every sampling instant $k\Delta$. The resampler selects $I_k = x(k\Delta)$, and by assumption both encodings return the identical frame at every k, so the resampled sequences coincide. Every node attribute in Eqs. (1) and (2) is a deterministic functional of a single resampled frame, every spatial weight in Eqs. (3) to (5) of a single frame and the frozen regressor, and every temporal weight in Eq. (6) of a consecutive pair together with the warp estimated from that same pair, so G is a deterministic functional of $(I_1, \dots, I_T)$ and coincides for the two encodings. The readout of Eqs. (7) to (9) is a deterministic map of G, which proves equality of all outputs. When the rate condition fails and the nearest available frame is taken, the selected frame differs from $x(k\Delta)$ by a time offset of at most $1/(2f)$, every cue histogram moves by at most $C\,\omega(1/f)$ in total variation for the modulus of continuity ω of the cue fields, Eq. (12) turns this into an entropy perturbation $\Lambda_B(C\omega(1/f))$, and the Lipschitz readout inflates it by a constant, which vanishes as $f \to \infty$. □

C.2 Proof of Proposition 2 (Permutation Invariance)

Let π_t relabel the region indices of frame t and write Π for the induced relabeling of nodes. Histograms in Eq. (1) are attached to pixel sets and do not read the index, and boundary adjacency and warped overlap correspondence are properties of pixel sets as well, so Π acts on G as a graph isomorphism that permutes node attributes and transports edge weights. We prove by induction over the layer l that $h^{(l)}_{\pi_t(i),t}[\text{relabeled}] = h^{(l)}_{i,t}[\text{original}]$. The base case is the initialization, which stacks ϕ and mean incident weights, all transported by Π. For the step, the sums in Eq. (7) run over $\mathcal{N}_s(\pi_t(i), t) = \pi_t(\mathcal{N}_s(i, t))$ and over the transported temporal neighbors with transported normalized weights, and the matrices $\Theta_0, \Theta_s, \Theta_t$ are shared, so the two sides agree. The head in Eq. (8) is shared across nodes, hence region scores are permuted by π_t, the area weights $\omega_{i,t}$ are transported with the regions so q_t is invariant, and Eq. (9) depends only on the multiset $\{q_t\}$. □

C.3 Proof of Proposition 3 (Dominance Decomposition)

For (i), let g act per pixel as a bijection σ on the cue bins and identically at t and $t{+}1$. The joint histogram of the degraded pair in Eq. (6) is the pushforward of the clean joint histogram under $\sigma \times \sigma$, a relabeling of the support cells, and Shannon entropy is invariant under support relabeling, so every conditional entropy and hence every w_t equals its clean value. For (ii), every quantity in Eqs. (3) to (6) is a functional of the pixels of its incident regions and boundary bands alone. If g modifies only $m_{i,t+1}$, the affected weights are exactly the spatial edges whose band intersects $m_{i,t+1}$, which are the edges incident to $v_{i,t+1}$, and the temporal edges that read frame $t{+}1$ at region i, namely (i, t) and, when a later frame exists, $(i, t{+}1)$. All other weights are computed from unmodified pixels and are unchanged. General degradations are neither bin bijective nor perfectly localized, so beyond these exact statements the separation holds in expectation, which is the overlap structure visible in the middle panel of Fig. 4. □

C.4 Proof of Proposition 4 (Plug In Consistency)

Let $\hat{p}$ be the empirical histogram of n i.i.d. samples over at most $K \le B^2$ occupied cells and $\hat{H} = H(\hat{p})$ the plug in estimate. The expansion of Miller and Madow gives

$$\mathbb{E}\big[\hat{H}\big] - H \;=\; -\frac{K-1}{2n} + O\big(n^{-2}\big)\,, \tag{13}$$

so the bias magnitude of an unnormalized joint entropy is at most $B^2/(2n)$ nats. For the fluctuation, replacing one sample changes $\hat{p}$ by at most $2/n$ in total variation, Eq. (12) bounds the resulting change of $\hat{H}$ by $\Lambda_B(1/n) \le c\,(1 + \log B)/n$, and the bounded difference inequality of McDiarmid yields $\mathrm{sd}(\hat{H}) \le c_2/\sqrt{n}$ with sub Gaussian concentration. Conditional entropies are differences of joint and marginal plug in estimates, and the marginal has $K \le B$ cells, so its bias is of lower order and the joint term dominates. Each weight in Eqs. (5) and (6) is an average over cues of such conditional entropies divided by $\log B$, so the bias of $\hat{w}$ is at most $B^2/(2n \log B)$ and the fluctuation constant is scaled by $1/\log B$, which proves Eq. (11). Almost sure convergence follows from the strong law applied cellwise and continuity of H. The bin count rule of Section 4.3 follows from

Table 4: Complete ablation on D-CQA including all cue removals and the effect of removing E_t per distortion type.

Variant	ΔGlobal SRCC	ΔRegion SRCC	ΔLoc. AUC
w/o temporal edges E_t	-0.171	-0.148	-0.202
w/o spatial edges E_s	-0.093	-0.187	-0.174
w/o motion compensation W	-0.118	-0.061	-0.129
w/o kinematic cue	-0.087	-0.052	-0.066
w/o texture cue	-0.056	-0.049	-0.043
w/o luminance cue	-0.041	-0.036	-0.028
w/o spectral cue	-0.038	-0.031	-0.025
w/o contrast cue	-0.032	-0.027	-0.021
readout depth $L = 1$	-0.024	-0.029	-0.011
readout depth $L = 3$	$+0.003$	$+0.004$	$+0.002$
w/o E_t, blur clips only	-0.052	-0.041	-0.063
w/o E_t, noise clips only	-0.214	-0.186	-0.247
w/o E_t, compression clips only	-0.163	-0.142	-0.198
w/o E_t, darkening clips only	-0.071	-0.058	-0.082
w/o E_t, haze clips only	-0.088	-0.069	-0.097

requiring the worst case unnormalized bias $B^2/(2n)$ to stay below $1/8$ nat, that is $B^2 \le n/4$, and its empirical bite is the drop beyond the two marked boundaries in Fig. 7 (c). □

C.5 Proof of Proposition 5 (Segmentation Refinement Stability)

We first state the proposition in full. *Let γ' refine γ so that every region of γ is a union of regions of γ', let δ be the maximal region diameter under γ', and let every region retain at least $n_{\min}$ pixels. Then*

$$\big|\, Q(\gamma') - Q(\gamma) \,\big| \;\le\; C_1\, \omega(\delta) \;+\; C_2\, \frac{B^2}{n_{\min}}, \tag{14}$$

where ω is the modulus of continuity of the cue fields and C_1, C_2 depend only on the readout Lipschitz constants.

Let region R of γ split into regions $r_1, \dots, r_m$ of γ'. The parent histogram is the area weighted mixture

$$p_c(\cdot \,|\, R) = \sum_{u=1}^{m} \frac{|r_u|}{|R|}\, p_c(\cdot \,|\, r_u), \tag{15}$$

and since every pixel of r_u lies within distance δ of any other, piecewise continuity of the cue fields gives $\mathrm{TV}(p_c(\cdot|r_u),\, p_c(\cdot|R)) \le C\, \omega(\delta)$. Applying Eq. (12) bounds every population level entropy and conditional entropy discrepancy between the two partitions by $\Lambda_B(C\omega(\delta)) \le C_1'\, \omega(\delta)$ for $\omega(\delta)$ small. At finite sample sizes each plug in estimate additionally deviates from its population value by the bias and fluctuation of Proposition 4, which is at most $B^2/(2n_{\min} \log B)$ in expectation since every region of γ' retains $n_{\min}$ pixels. The area weighted aggregation of Eq. (8) is an average and does not inflate either term, and the readout is Lipschitz with constant bounded by $\prod_l \|\Theta^{(l)}\|$, so absorbing all constants into C_1, C_2 yields Eq. (14). The empirical plateau in the right panel of Fig. 4 matches the predicted $\omega(\delta) \to 0$ decay with the $B^2/n_{\min}$ floor. □

D Sensitivity and Breakdown

The breadth of the tuning and breakdown experiments is summarized in Fig. 7 and Tables 4 and 5. Panel (a) of Fig. 7 shows all 320 sampled configurations over the search space of Table 5 as a corner plot of the three headline metrics. The ten best configurations concentrate in the joint upper right corner of every pairwise view, so no metric trades off against another and a single configuration serves all three tasks. Panel (b) reports the exhaustive per type and per severity grid, where the global SRCC decays monotonically with severity for every distortion type and stays above 0.842 in the hardest cell. Panel (c) sweeps the resampling interval Δ and the bin count B. Both curves are flat around the operating point $\Delta = 100$ ms and $B = 8$, beyond the boundary band bound at

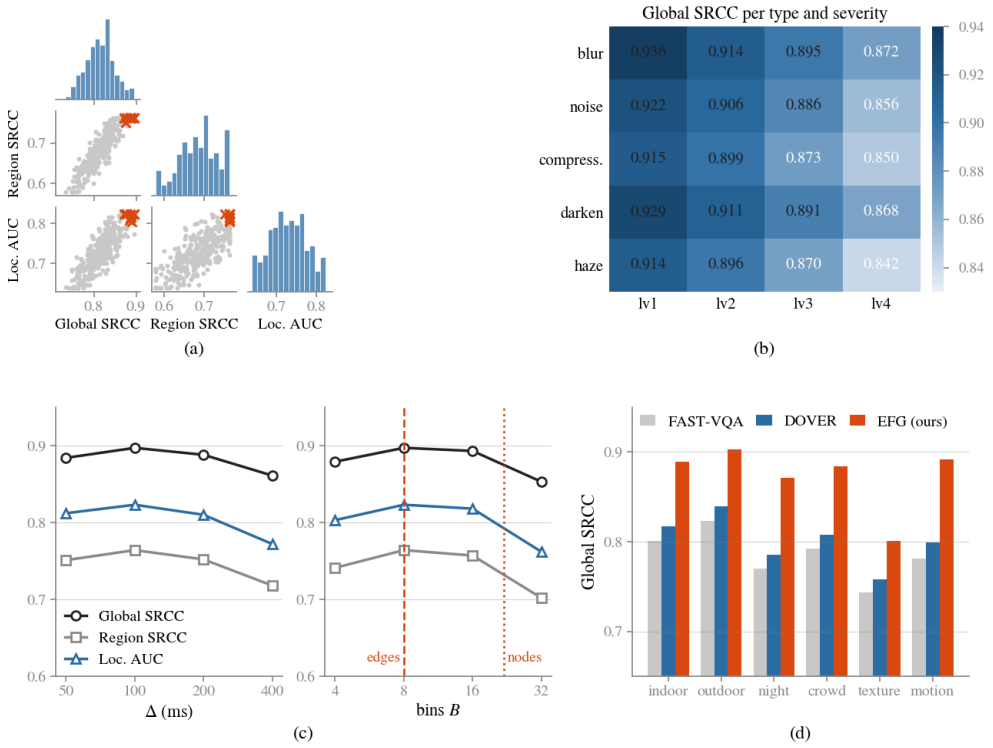


Figure 7: Sensitivity and breakdown. (a) Sweep corner plot with the ten best configurations marked. (b) Global SRCC per type and severity. (c) Δ and B sensitivity, with the bin count boundaries $B^2 \leq n/4$ for boundary bands (dashed, $B = 8$) and for nodes (dotted, $B = 22$). (d) Per category comparison.

Table 5: Search space for the 320 sampled configurations. Configurations are sampled uniformly from the grid and evaluated with three seeds each.

Parameter	Values
resampling interval Δ (ms)	50, 100, 200, 400
bin count B	4, 8, 16, 32
boundary band β (px)	3, 6, 9
readout depth L	1, 2, 3
pooling blend η	0, 0.3, 0.6
region loss weight λ_r	0.1, 0.3, 1.0

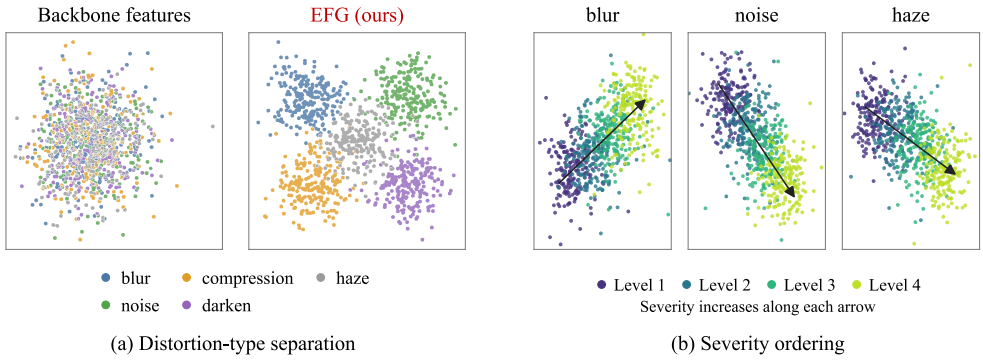
$B = 8$ the region score and localization AUC begin to fall, and beyond the node bound at $B = 22$ every measure drops, the estimation penalty predicted by the bin count rule of Section 4.3. Panel (d) breaks the comparison down by content category. EFG leads in every category, and the margin is smallest on heavily textured content, the same texture masking limit seen in Fig. 6 and discussed in Section 6. Table 4 completes the ablation of Table 2 with every cue removal and with the effect of removing the temporal edges on each distortion type, which is largest on noise and compression and smallest on blur.

E Facet Comparison

Table 6 compares the graph based and holistic families of Section 2 facet by facet. Region means nodes are image regions, Temporal means inter frame structure is modeled, Analytic means node and edge semantics are closed form, NR means no reference or anchor is needed, and Localize means where or when is output.

Table 6: Facet comparison of graph based and holistic quality models.

Method	Region	Temporal	Analytic	NR	Localize
GraphIQA (Sun et al., 2023)	✗	✗	✗	✓	✗
IQG (Liu & Guo, 2025)	✗	✗	✗	✓	✗
VGCN (Xu et al., 2021)	✗	✗	✗	✓	✗
AHGCN (Fu et al., 2022)	✗	✗	✗	✓	✗
GPA-Net (Shan et al., 2024)	✗	✗	✗	✓	✗
TMES (Pan et al., 2021)	✗	✓	✗	✓	✗
PANDA (Janjua et al., 2026)	✓	✗	✗	✗	✓
FAST-VQA (Wu et al., 2022)	✗	✓	✗	✓	✗
DOVER (Wu et al., 2023b)	✗	✓	✗	✓	✗
Q-Align (Wu et al., 2024)	✗	✓	✗	✓	✗
EFG (ours)	✓	✓	✓	✓	✓



(a) Distortion-type separation (b) Severity ordering

Figure 8: Representation analysis. (a) Distortion type separation in backbone and EFG embeddings. (b) Severity ordering within EFG clusters, arrows indicate increasing severity.

F Extended Representation Analysis

Fig. 8 shows the type separation against the backbone and the severity ordering within representative EFG clusters. The EFG embedding separates the five distortion types more clearly than the backbone representation, while the blur, noise, and haze clusters exhibit monotonic severity progression along type specific directions. Clusters overlap at their boundaries and stray points cross them, consistent with the expectation level separation of Proposition 3 rather than a hard partition.

G Dense Graph Examples

Fig. 9 draws complete entropy flow graphs over two real frame pairs. In the park scene only the two walkers carry hot temporal edges, which is the content change response of w_t discussed in Section 3.4 rather than a degradation. In the terminal scene the concourse and a traveler dominate while the departure board sits at an intermediate weight as its display updates. Every weight is readable directly off the graph, the same structure the MLLM consumes in Section 5.4.

H Efficiency

The landscape of parameters against accuracy in Fig. 10 (a) places EFG among fifteen representative models spanning three tiers, hand crafted statistics (Mittal et al., 2012; 2013; Korhonen, 2019; Tu et al., 2021a;b), deep holistic regressors (Li et al., 2019; Sun et al., 2022; Wu et al., 2023a; 2022; 2023b; Pan et al., 2021; Chen et al., 2022; Sun et al., 2023; Janjua et al., 2026), and a scorer built on a large multimodal model (Wu et al., 2024). EFG occupies the empty upper left corner, matching hand crafted models in learnable parameters at 0.08M while exceeding every deep and multimodal

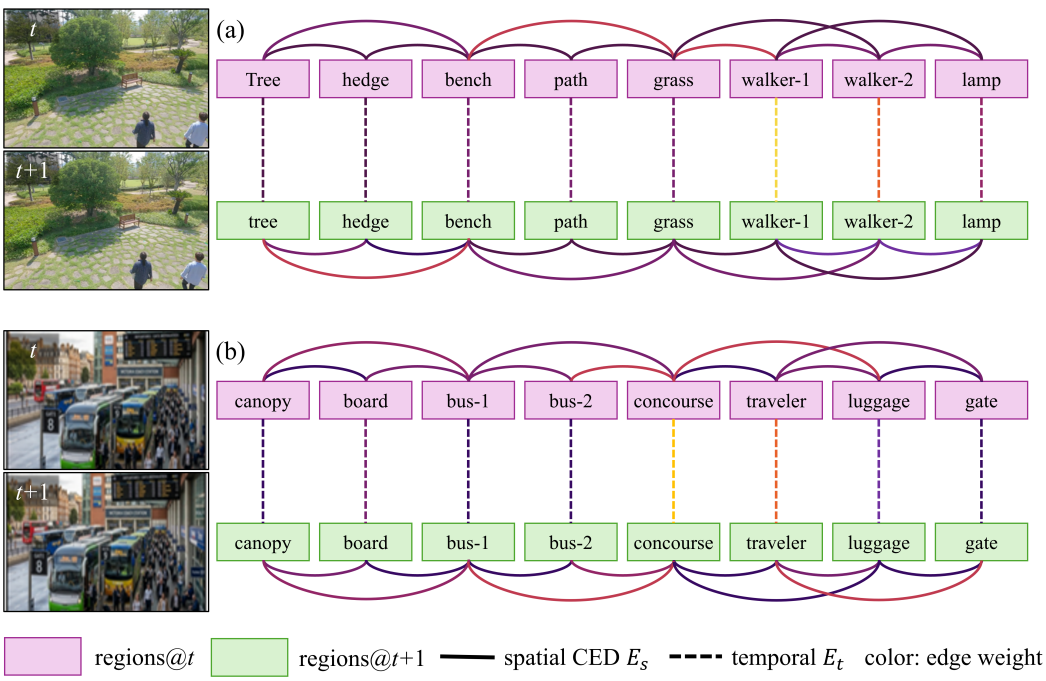


Figure 9: Dense entropy flow graphs on two real frame pairs.

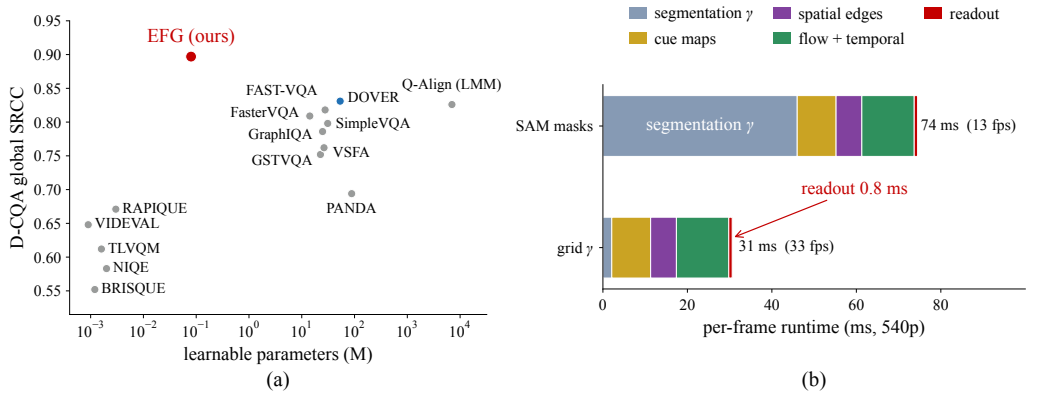


Figure 10: Efficiency. (a) Parameters against accuracy over fifteen models. (b) Per frame runtime for two segmenters.

baseline in accuracy, which substantiates the parameter claim of Section 5.2 quantitatively. The runtime breakdown in Fig. 10 (b) profiles the pipeline per frame at 540p for two segmenter choices. The readout accounts for 0.8 ms and is negligible, graph construction itself runs at interactive rates, and the segmenter dominates the budget, so swapping SAM masks for the grid γ moves the pipeline from 13 to 33 fps at the stability cost quantified in the right panel of Fig. 4, which provides a concrete path to streaming deployment.