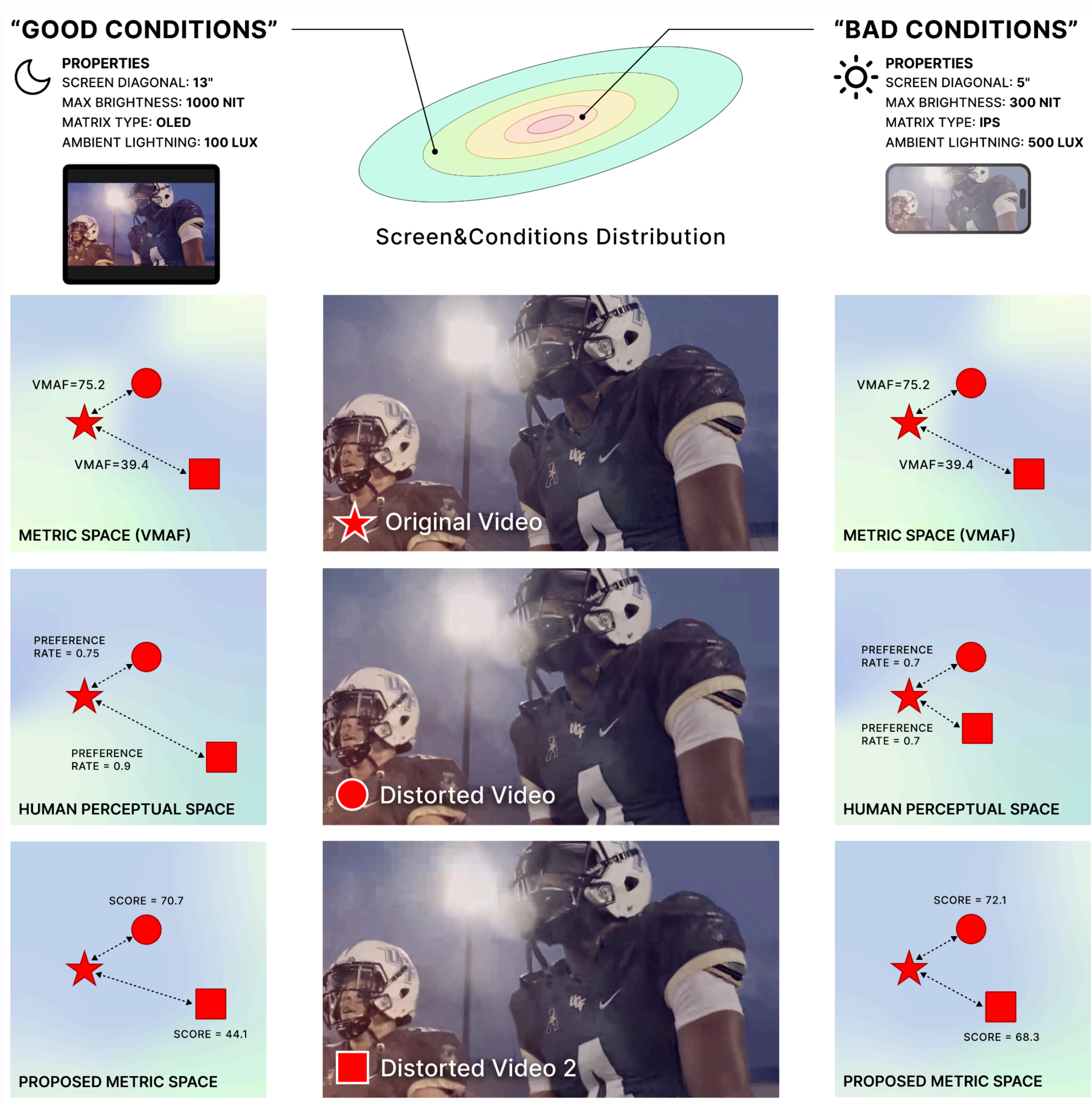
Learning Flexible Generalization in Video Quality Assessment by Bringing Device and Viewing Condition Distributions

Nikolay Safonov[1,2] and Dmitriy Vatolin[1]

[1] MSU Institute for Artificial Intelligence, Moscow, Russia [2] AI Center, Lomonosov Moscow State University, Moscow, Russia

1 Motivation



The same compressed video can look acceptable in one setup and visibly distorted in another.

Objective VQA aims to predict human-perceived video quality. The key difficulty is that perception is not invariant: screen size, resolution, display brightness and ambient lighting change the visibility of compression artifacts. Existing VQA datasets and metrics often assume a uniform viewing environment, which limits their reliability on mobile devices.

4 Blade-Chest

Bradley-Terry and Elo models aggregate pairwise preferences into global scores, but do not use the viewing condition z of each comparison.

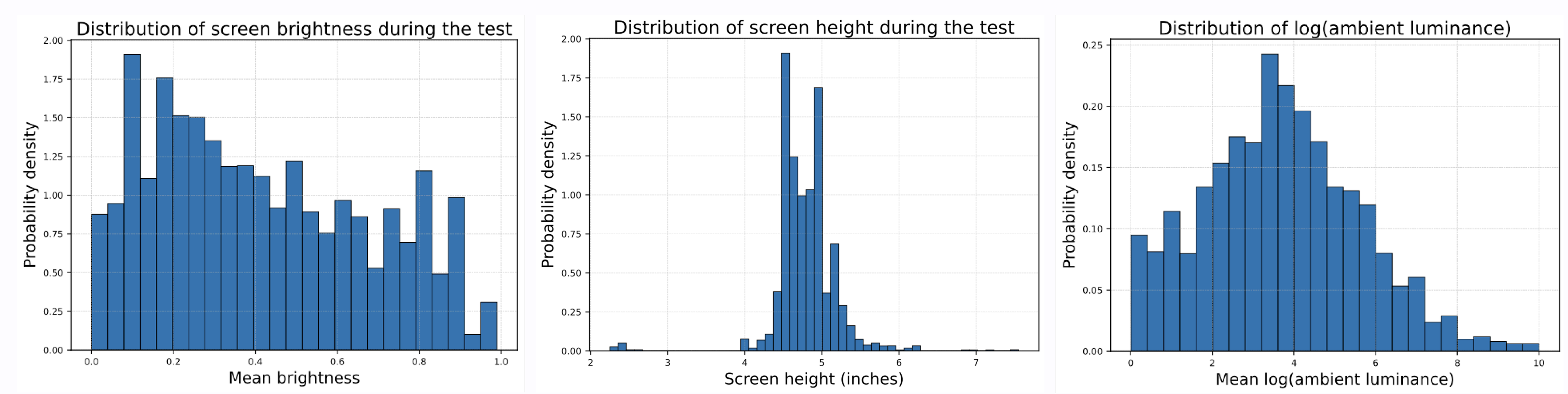
We adopt Blade-Chest because it naturally incorporates viewing conditions into score aggregation.

$$P(i \succ j \mid z) = \sigma\left(\left\|f_{c(q_i,z)} - f_{b(q_j,z)}\right\|^2 - \left\|f_{c(q_j,z)} - f_{b(q_i,z)}\right\|^2\right)$$

- Latent subjective video-quality scores: q_i, q_j;
- Viewing condition vector z: screen size, resolution, brightness, ambient luminance and display type;
- Condition-aware transformations f_c, f_b optimized with EM.

2 Dataset

We developed an Android application for crowd-sourced pairwise quality assessment. During each test, it records the device model, screen specifications, brightness, ambient luminance and orientation.

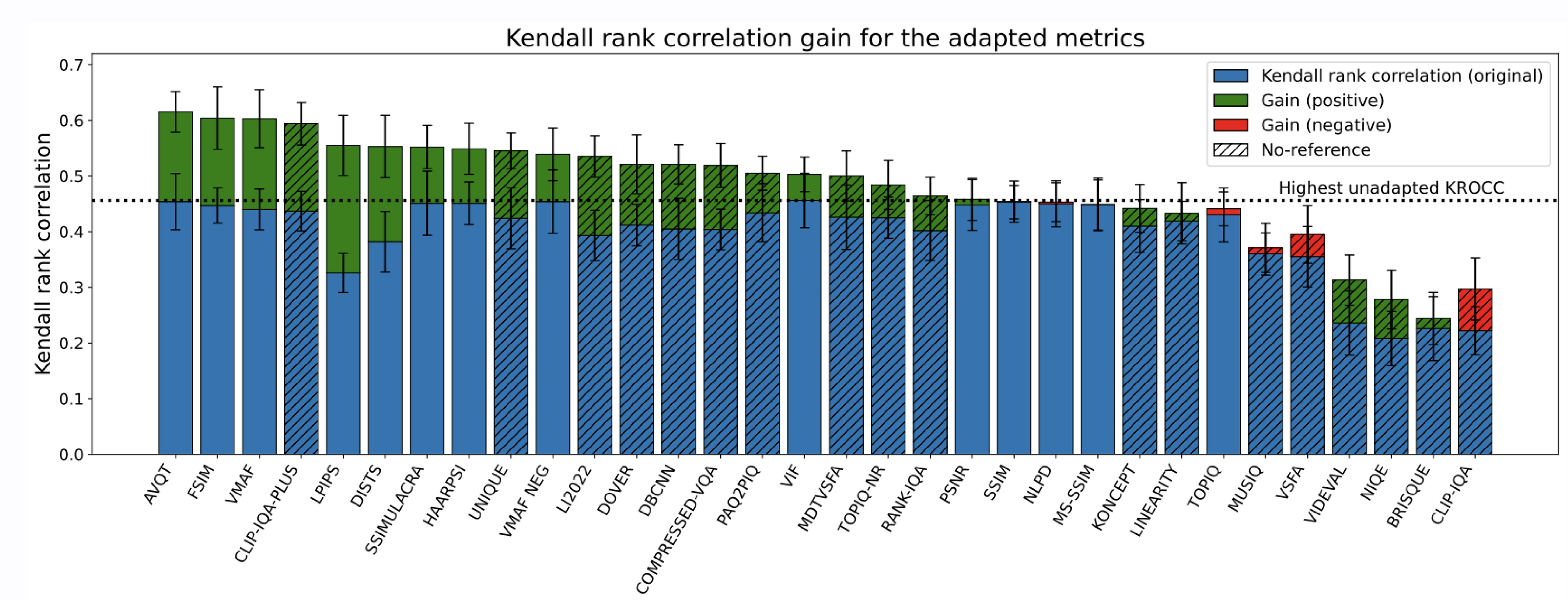


Viewing-condition distributions: screen brightness, screen height and ambient luminance.

Participants compare compressed versions of the same source video in full-screen mode and choose the better version or an equal-quality option. Each task contains 12 comparisons, including 2 control pairs; at least 15 judgments are collected for every pair. The dataset covers SDR/HDR compressed content viewed on 300+ Android devices.

25	650	250K+	300+
source videos	distorted videos	annotations	Android devices

5 Metrics Gain

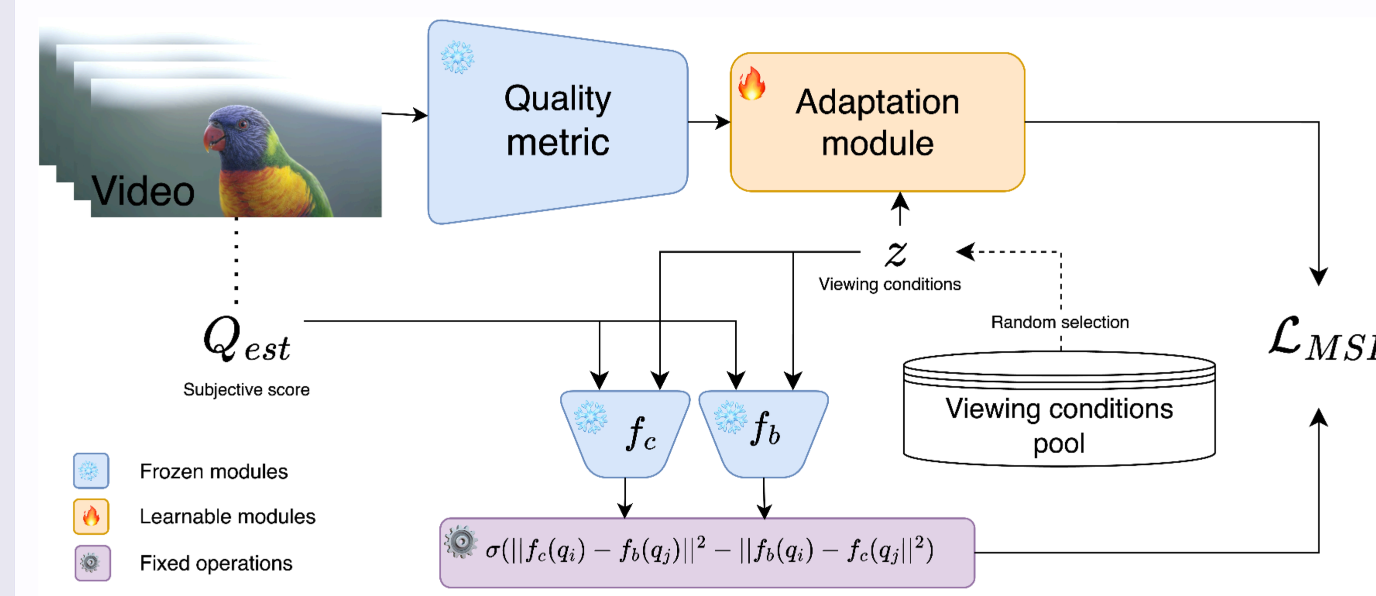


Adapted metrics improve Kendall rank correlation on pairwise judgments.

Pearson and Spearman correlations require reliable aggregated subjective scores. In our dataset, each assessor labels only a small part of the content, so fixed-condition scores are sparse. Kendall rank correlation instead measures the share of correctly ordered pairs and therefore matches the pairwise preference protocol.

- Adaptation improves most evaluated objective metrics;
- Held-out-device tests show generalization to unseen smartphone models;
- Residual noise comes from human variability and unmodeled viewing effects.

3 Architecture



Two-stage condition-aware framework: score aggregation followed by metric adaptation.

Stage 1

Blade-Chest aggregation

Infers subjective scores from pairwise votes while accounting for viewing condition vector z.

Stage 2

Metric adaptation

Maps a metric prediction and target conditions to condition-specific video quality.

Phase 1 estimates latent subjective quality scores and condition-aware transformations from sparse pairwise votes. Phase 2 trains a lightweight fully connected adaptation module for each metric; the original metric is not retrained.

The condition pool samples physically plausible viewing-condition vectors beyond the exact recorded sessions. These simulated conditions augment sparse mobile observations and improve generalization to unseen device and environment states.

6 Conclusions

VQA should model both the encoded video and the viewing conditions under which quality is perceived.

- Large-scale mobile VQA dataset with 250K+ comparisons on 300+ devices;
- Condition-aware aggregation and adaptation for objective quality metrics;
- Condition-pool sampling improves robustness to unseen viewing setups.

videoprocessing.github.io/device-viewing-conditions

nikolay.safonov@graphics.cs.msu.ru