Improvements on Subjective Experiment Data Analysis Process: An Update

Zhi Li, Christos Bampis, Lukas Krasula, *Netflix*
Lucjan Janowski, *AGH*
Ioannis Katsavounidis, *Facebook*

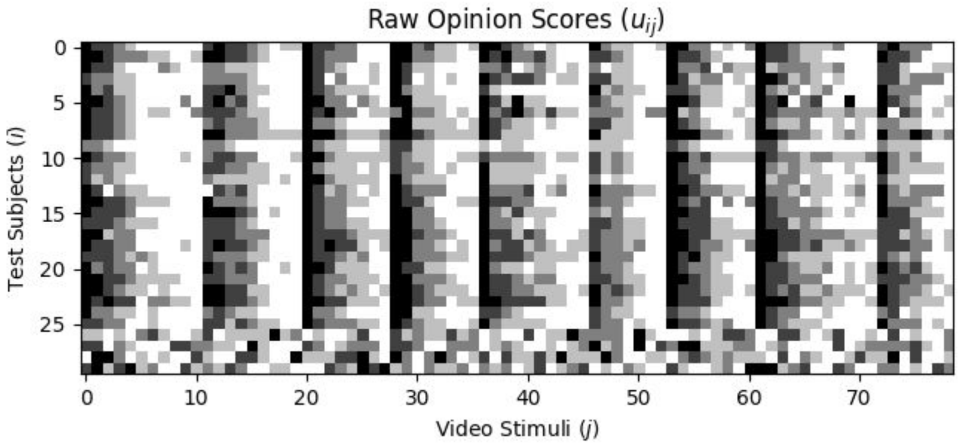
Q19 Interim Meeting, VQEG Fall 2020

Outline

- Background and motivation
- Proposed methodology
- Progress since ITU-T SG12 C470 (April 2020)
    - New comparison results with BT.500 / P.913
    - Interpreting the limitations of BT.500 / P.913
    - Update on the calculation of confidence intervals
    - Runtime analysis
- ITU Proposals

Raw opinion scores are noisy and unreliable

Black: lowest score
White: highest score



Subject bias

Last 4 Outliers

- Would MOS or DMOS be good enough?
- Corrective mechanisms
  - Subject outlier rejection
  - Subject bias removal

*NFLX Public Dataset

Prior Art: Subject Outlier Rejection (ITU-R BT.500)

For each test presentation, calculate the mean, $\bar{u}_{jkr}$, standard deviation, S_{jkr}, and kurtosis coefficient, β_{2jkr}, where β_{2jkr} is given by:

$$\beta_{2jkr} = \frac{m_4}{(m_2)^2} \quad \text{with} \quad m_x = \frac{\sum_{i=1}^{N}(u_{ijkr} - \bar{u}_{ijkr})^x}{N} \qquad (5)$$

For each observer, i, find P_i and Q_i, i.e.:

for $j, k, r = 1, 1, 1$ to J, K, R

if $2 \le \beta_{2jkr} \le 4$, then:

if $u_{ijkr} \ge \bar{u}_{jkr} + 2\,S_{jkr}$ then $P_i = P_i + 1$

if $u_{ijkr} \le \bar{u}_{jkr} - 2\,S_{jkr}$ then $Q_i = Q_i + 1$

else:

if $u_{ijkr} \ge \bar{u}_{jkr} + \sqrt{20}\,S_{jkr}$ then $P_i = P_i + 1$

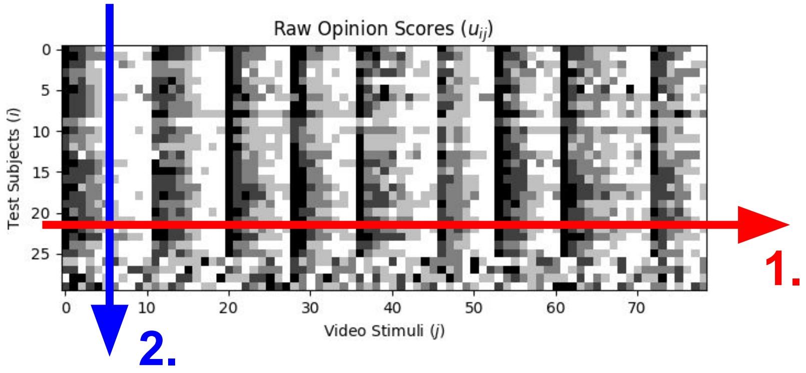
if $u_{ijkr} \le \bar{u}_{jkr} - \sqrt{20}\,S_{jkr}$ then $Q_i = Q_i + 1$

If $\frac{P_i + Q_i}{J \cdot K \cdot R} > 0.05$ and $\left|\frac{P_i - Q_i}{P_i + Q_i}\right| < 0.3$ then reject observer i

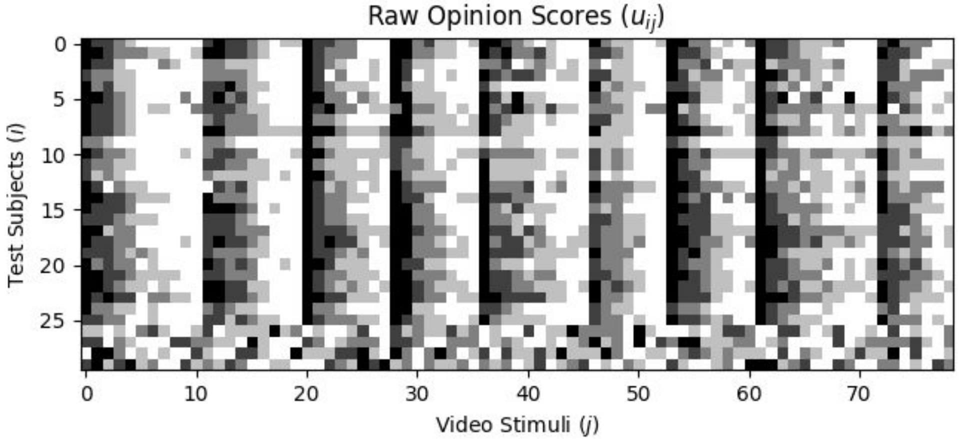
with:

- N: number of observers
- J: number of test conditions including the reference
- K: number of test images or sequences
- R: number of repetitions
- L: number of test presentations (in most cases the number of presentations will be equal to $J \cdot K \cdot R$, however it is noted that some assessments may be conducted with unequal numbers of sequences for each test condition).

1. Video by video, the algorithm counts the number of instances when a subject's opinion score deviates by a few sigmas (i.e. std's)
2. Subject by subject, if the occurrences are more than a fraction, reject the subject



Limitations of BT.500-Style Subject Outlier Rejection



Black: lowest score
White: highest score

Subject bias

Last 4 are outliers

- All scores corresponding to rejected subjects are discarded - an overkill
- Often only identifies a subset of outliers
  - In the example above, only subjects #26, #28, #29 were rejected*
- Hard-coded parameters / thresholds:
  - Not very interpretable
  - May not be suitable for all conditions

*To be discussed in a later slide why only 3 out 4 outliers detected

Prior Art: Subject Bias Removal (ITU-T P.913)

First, estimate the MOS for each PVS:

$$\mu_{\psi_j} = \frac{1}{I_j}\sum_{i=1}^{I_j} o_{ij}$$

where:

- o_{ij} is the observed rating for subject i and PVS j;
- I_j is the number of subjects that rated PVS j;
- μ_{ψ_j} estimates the MOS for PVS j, given the source stimuli and subjects in the experiment.

Second, estimate subject bias:

$$\mu_{\Delta_i} = \sum_{j=1}^{J_i}\left(o_{ij} - \mu_{\psi_j}\right)$$

where:

- μ_{Δ_i} estimates the overall shift between the ith subject's scores and the true values (i.e., opinion bias)
- J_i is the number of PVSs rated by subject i.

Third, calculate the normalized ratings by removing subject bias from each rating:

$$r_{ij} = o_{ij} - \mu_{\Delta_i}$$

where:

- r_{ij} is the normalized rating for subject i and PVS j.

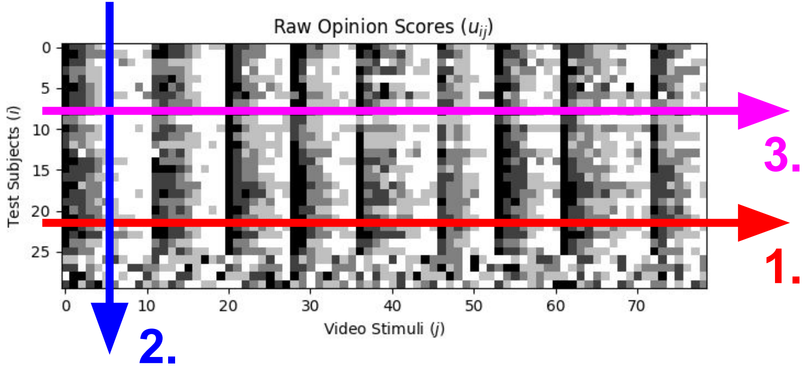
MOS and DMOS are then calculated normally. This normalization does not impact MOS:

$$\mu_{\psi_j} = \frac{1}{I_j}\sum_{i=1}^{I_j} r_{ij} = \frac{1}{I_j}\sum_{i=1}^{I_j} o_{ij}$$

where:

- μ_{ψ_i} estimates the MOS of PVS j.

1. Video by video, estimate MOS by averaging over subjects
2. Subject by subject, estimate subject bias by comparing against MOS
3. Video by video, estimate MOS again based on bias-removed opinion scores (often combined with BT.500-style subject rejection)



Can we do better?

Can we do better?

- Insight #1:
  - The subject outliers do NOT have to be rejected as a whole
  - Instead, we can model them as having large “inconsistencies”
    - “Soft” rejection
    - Avoid hard decision and hard coded parameters
- Insight #2:
  - The subject bias removal and subject outlier rejection do NOT have to be carried out in separate steps, which leads to sub-optimality
  - Instead, we can incorporate “bias” and “inconsistency” in one model and jointly solve the model parameters in one step
- Our proposal:
  - A simple yet effective model to account for:
    - Subject bias
    - Subject inconsistency
      - Outliers as a special case with very large inconsistencies
  - Jointly solve the model parameters via maximum likelihood estimation (MLE)

Other Considerations (for Standardization)

- Strike a delicate balance between reality and model simplicity
    - Other candidates:
        - PVS/Content ambiguity [Janowski&Pinson’15, Li&Bampis’17]
        - Environmental influences
        - Continuous vs. discrete scales [Janowski et al’19]
        - Fringe effect of scales
    - The proposed model accounts for two of the most dominant effects
        - Subject bias
        - Subject inconsistency

Other Considerations (for Standardization)

Old Standard

New Standard

- For easy acceptance, new standard should NOT be a *paradigm shift*
  - A good approach is to encompass prior standard as a special case
- Solution must be intuitive
  - Each model parameter should carry explicit meaning
  - Each solution step should be highly interpretable
- Solution should be widely applicable to different subjective methodologies
  - ACR / ACR-HR
  - DCR (DSIS)
  - Continuous (DSCQS) / discrete scales
- Solution should be *fast* and *stable*

Outline

- Background and motivation
- Proposed methodology
- Progress since ITU-T SG12 C470
  - New comparison results with BT.500 / P.913
  - Interpreting the limitations of BT.500 / P.913
  - Update on the calculation of confidence intervals
  - Runtime analysis

Proposed Model*

Raw Opinion Score / True Quality / Subject Bias / Subject Inconsistency

$$U_{ijr} = \psi_j + \Delta_i + \upsilon_i X$$

- U_{ijr} - Opinion score of subject i, stimulus j and repetition r
- ψ_j - true quality of stimulus j
- Δ_i - bias of subject i
- υ_i - inconsistency (std) of subject i
- X - i.i.d. normal random variables, $X \sim N(0, 1)$

*The model is a simplified version of [Li&Bampis'17] without considering the ambiguity of content. Compared to the original, the solution to the simplified model is faster and more stable.

Solving the Model Parameters via Maximum Likelihood Estimation (MLE)

- Given raw opinion scores $\{U_{ijr}\}$
- The task is to solve for the free parameters $\theta = (\{\psi_j\}, \{\Delta_i\}, \{v_i\})$
- Define log-likelihood function $l(\theta)$

$$l(\theta) = \log P(U_{ijr} | \{\psi_j\}, \{\Delta_i\}, \{v_i\})$$

- Numerically solve to maximize the log-likelihood function

$$\hat{\theta} = \arg\max l(\theta)$$

- Example problem size:
  - # observations: 300 (stimuli) * 30 (subjects) = 9000
  - # parameters:
    - True quality scores 300
    - Subject bias 30
    - Subject inconsistency 30

Proposed Solver

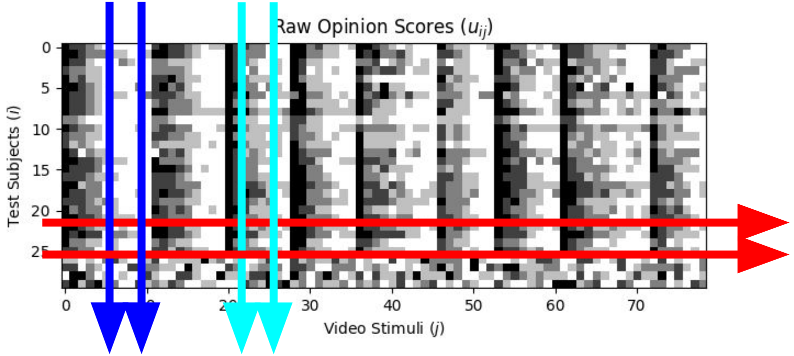
- Input:
  - u_{ijr} for subject $i = 1, \ldots, I$, stimulus $j = 1, \ldots, J$ and repetition $r = 1, \ldots, R$.
  - Stop threshold ψ^{thr}.
- Initialize $\{\psi_j\} \leftarrow \{MOS_j\}$, where $MOS_j = (\sum_{ir} 1)^{-1} \sum_{ir} u_{ijr}$.
- Initialize $\{\Delta_i\} \leftarrow \{BIAS_i\}$, where $BIAS_i = (\sum_{jr} 1)^{-1} \sum_{jr} (u_{ijr} - MOS_j)$.
- Loop:
  - $\{\psi_j^{prev}\} \leftarrow \{\psi_j\}$.
  - $\epsilon_{ijr} \leftarrow u_{ijr} - \psi_j - \Delta_i$ for $i = 1, \ldots, I$, $j = 1, \ldots, J$ and $r = 1, \ldots, R$.
  - $v_i \leftarrow \sigma_i\{\epsilon_{ijr}\}$, where $\sigma_i\{\epsilon_{ijr}\} = \sqrt{(\sum_{jr} 1)^{-1} \sum_{jr} (\epsilon_{ijr} - \epsilon_i)^2 - \epsilon_i^2}$ and $\epsilon_i = (\sum_{jr} 1)^{-1} \sum_{jr} \epsilon_{ijr}$, for $i = 1, \ldots, I$.
  - $\psi_j \leftarrow (\sum_{ir} v_i^{-2})^{-1} \sum_{ir} v_i^{-2} (u_{ijr} - \Delta_i)$, for $j = 1, \ldots, J$.
  - $\Delta_i \leftarrow \left(\sum_{jr} 1\right)^{-1} \sum_{jr} (u_{ijr} - \psi_j)$, for $i = 1, \ldots, I$.
  - If $\sqrt{\sum_{j=1}^{J} (\psi_j - \psi_j^{prev})^2} < \psi^{thr}$, break.
- Output: $\{\psi_j\}$, $\{\Delta_i\}$, $\{v_i\}$.

Alternating Projection (AP) Solver

1. Video by video, estimate MOS by averaging over subjects
2. Subject by subject, estimate subject bias by comparing against the MOS

In a loop:

a. Subject by subject, estimate subject inconsistency as the std of the residue of raw scores
b. Repeat step 1 (with weighting).
c. Repeat step 2.
d. If solution stabilizes, break



Proposed Solver - Interpretation

- Strong intuition behind the updating steps

$$\psi_j = \frac{\sum_{ir} \upsilon_i^{-2} (u_{ijr} - \Delta_i)}{\sum_{ir} \upsilon_i^{-2}}$$

"Subject Consistency"

$$\Delta_i = \frac{\sum_{jr} (u_{ijr} - \psi_j)}{\sum_{jr} 1}$$

$$\upsilon_i = \sigma_i\{\epsilon_{ijr}\}$$

True quality are weighted by "subject consistency" (υ_i^{-2}) after the subject bias (Δ_i) is removed. The "subject consistency" is the inverse of the (squared) subject inconsistency (υ_i^2)*.

Subject bias (Δ_i) as the mean of the opinion scores after the true quality (ψ_j) removed.

Subject inconsistency as the standard deviation of the estimation residue (ϵ_i).

- P.913 subject bias removal is a **special case** of the proposed solver in the following sense:
  - The P.913 solver is not iterative
  - The true quality in P.913 is not weighted by "subject consistency"

*In practical implementation, we add a small ε to make the denominator non-zero.

Summary of Each Method Compared

- BT.500 - keep or reject subjects
- P.913 - remove subject bias, keep or reject subjects
- Proposed AP - weigh subjects by consistency

When Will the Proposed Method Be Most Valuable?

- When faced with uncertainties
  - Crowdsourcing
  - Cross-lab study
  - Analyzing a new technology
  - New rating scale may confuse subjects
  - Inexperienced person designing the test
  - Media contain multiple confounding impairments
  - Distracting test environment
  - Unusual experiment design may have unintended consequences
- When BT.500 and P.913 give contradictory subject rejections

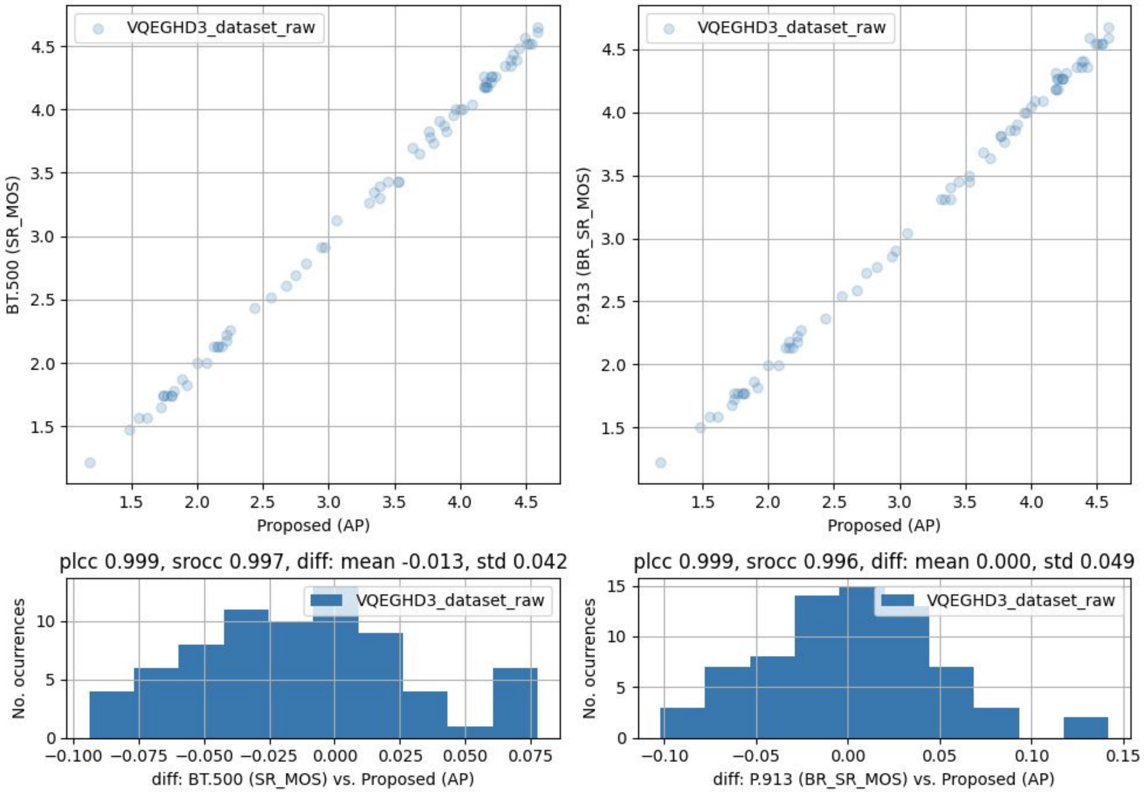
Outline

- Background and motivation
- Proposed methodology
- Progress since ITU-T SG12 C470
  - New comparison results with BT.500 / P.913
    - Scatter plots between proposed and BT.500/P.913
    - Analyze a crowdsourcing dataset: correlation with partial data
    - Cross-lab study on VQEG FRTV Phase I datasets
  - Interpreting the limitations of BT.500 / P.913
  - Update on the calculation of confidence intervals
  - Runtime analysis

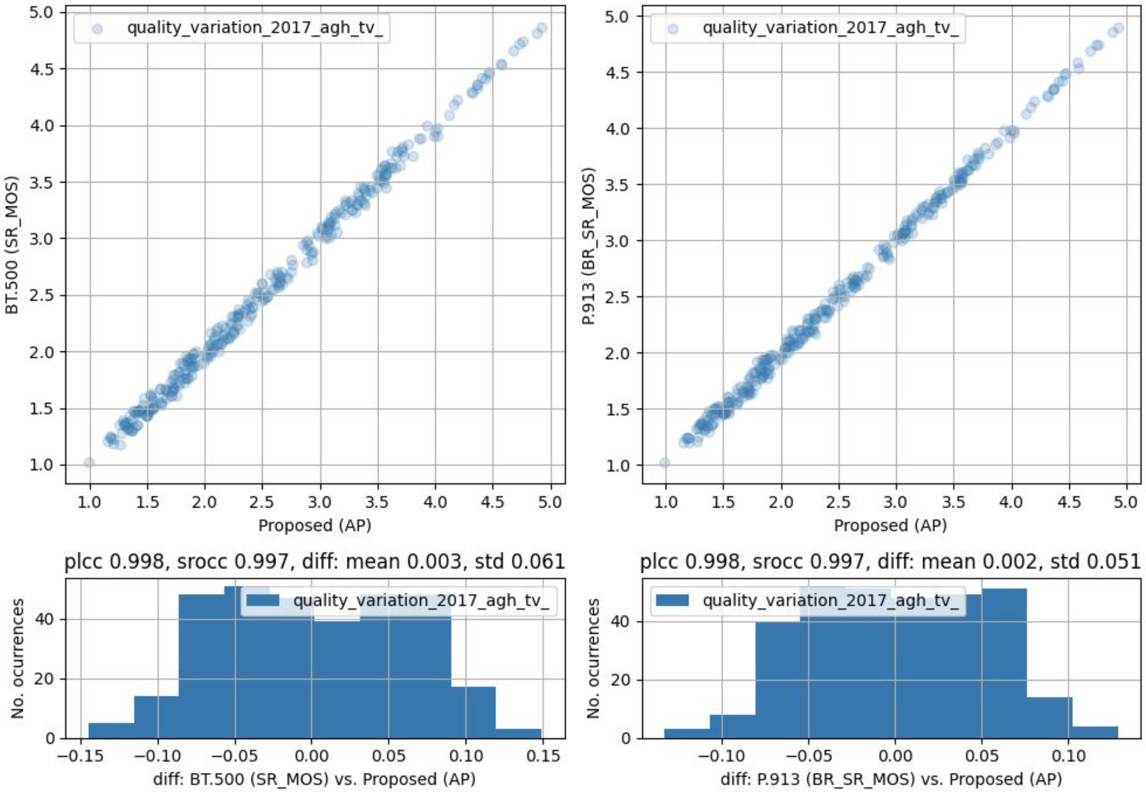
Scatter plots:
Proposed AP vs. BT.500
Proposed AP vs. P.913

More Datasets in the Appendix

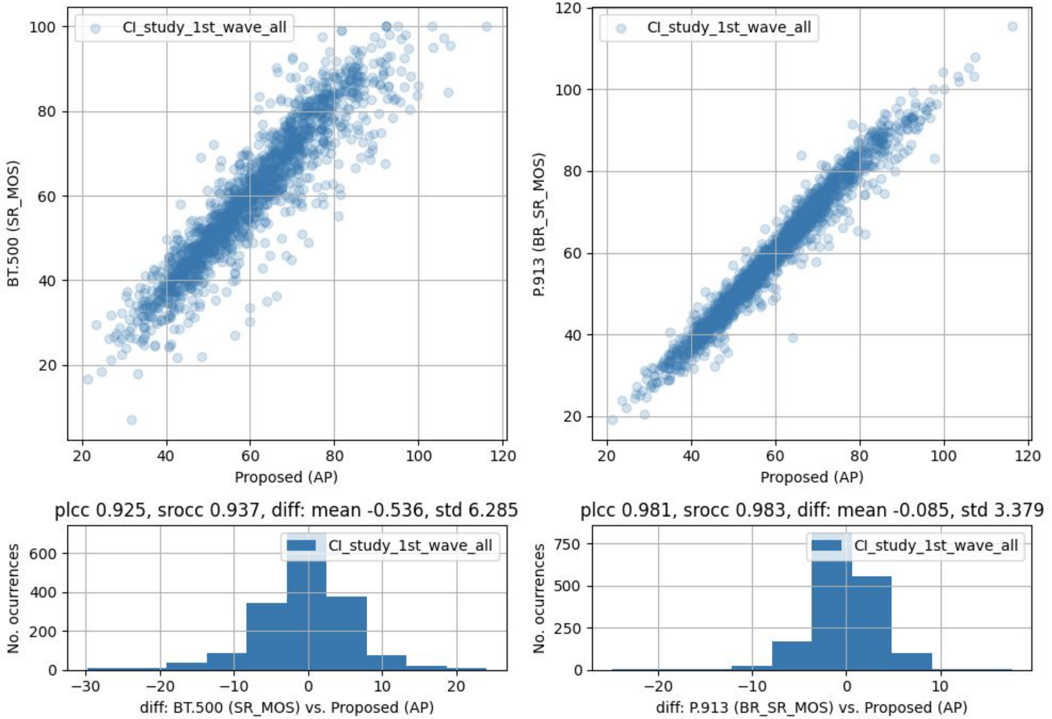
Recovered Quality Score - Proposed vs. BT.500/P.913 VQEG HD3 (Lab Study)



Recovered Quality Score - Proposed vs. BT.500/P.913 Quality Variation 2017 AGH TV Dataset (Lab Study)

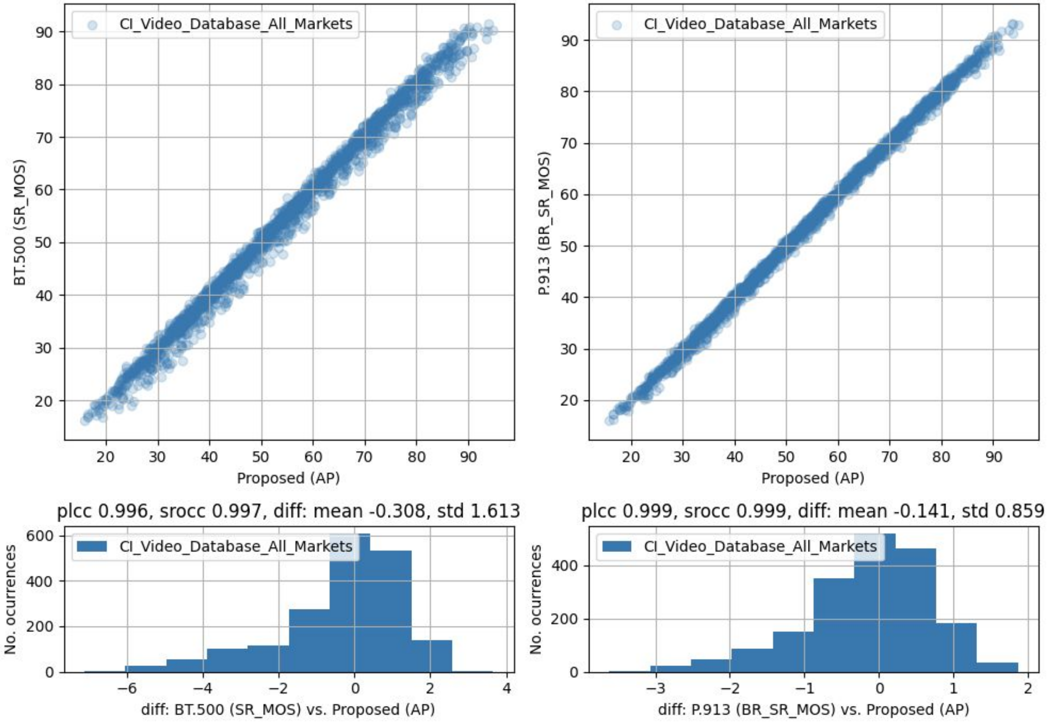


Recovered Quality Score - Proposed vs. BT.500/P.913 CI Study 1st Wave Dataset* (Crowdsourcing Study)



*The 1st wave dataset is a small pre-test, with certain PVSs having only one raw score.

Recovered Quality Score - Proposed vs. BT.500/P.913 CI Study 2nd Wave Dataset* (Crowdsourcing Study)



*The 2nd wave dataset is a large test with a minimum 108 scores per PVS.

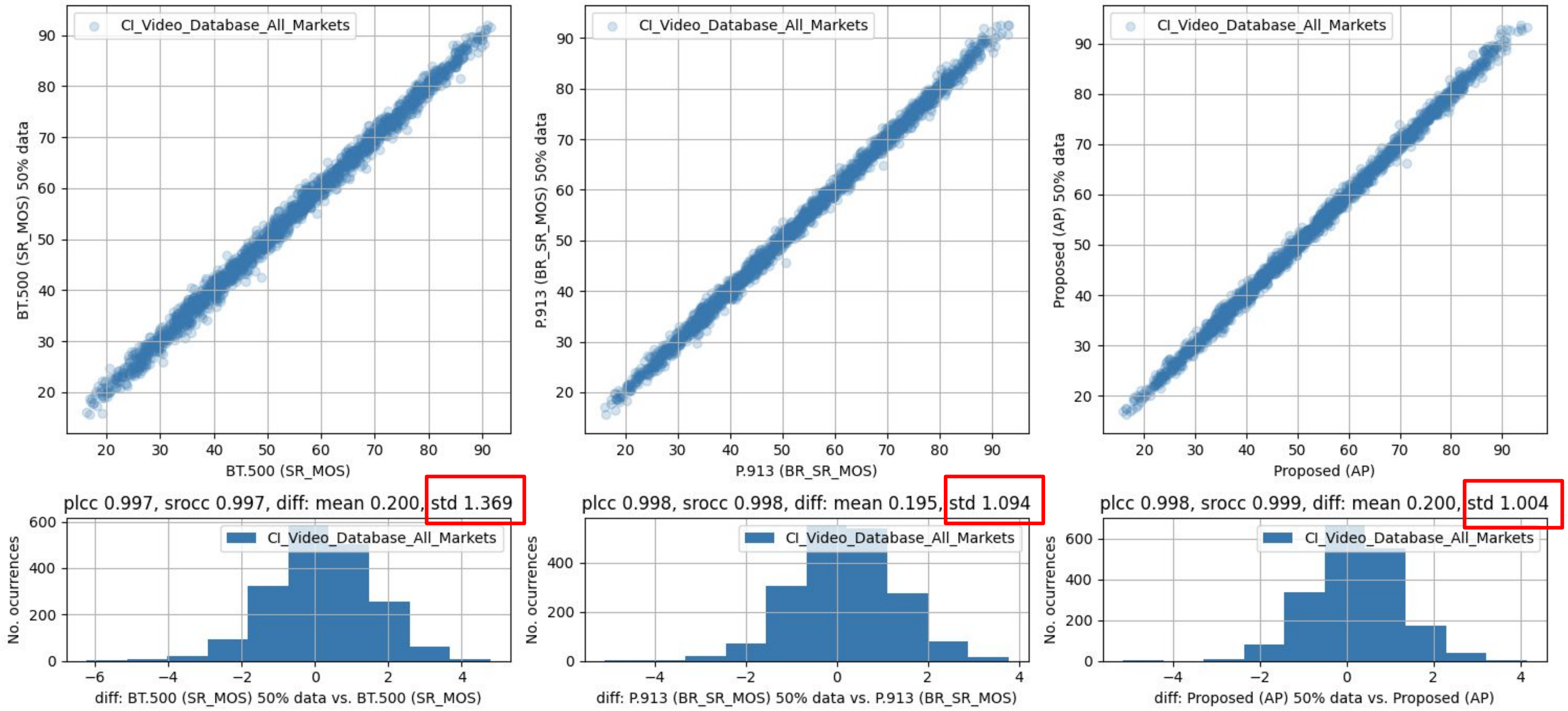
Observations on the correlation between the proposed and BT.500 / P.913 methods

- For lab results, the recovered scores from the proposed and BT.500/P.913 are highly correlated
- For crowdsourcing results, the recovered score from the proposed and BT.500/P.913 can deviate, but will improve as the subject size increases

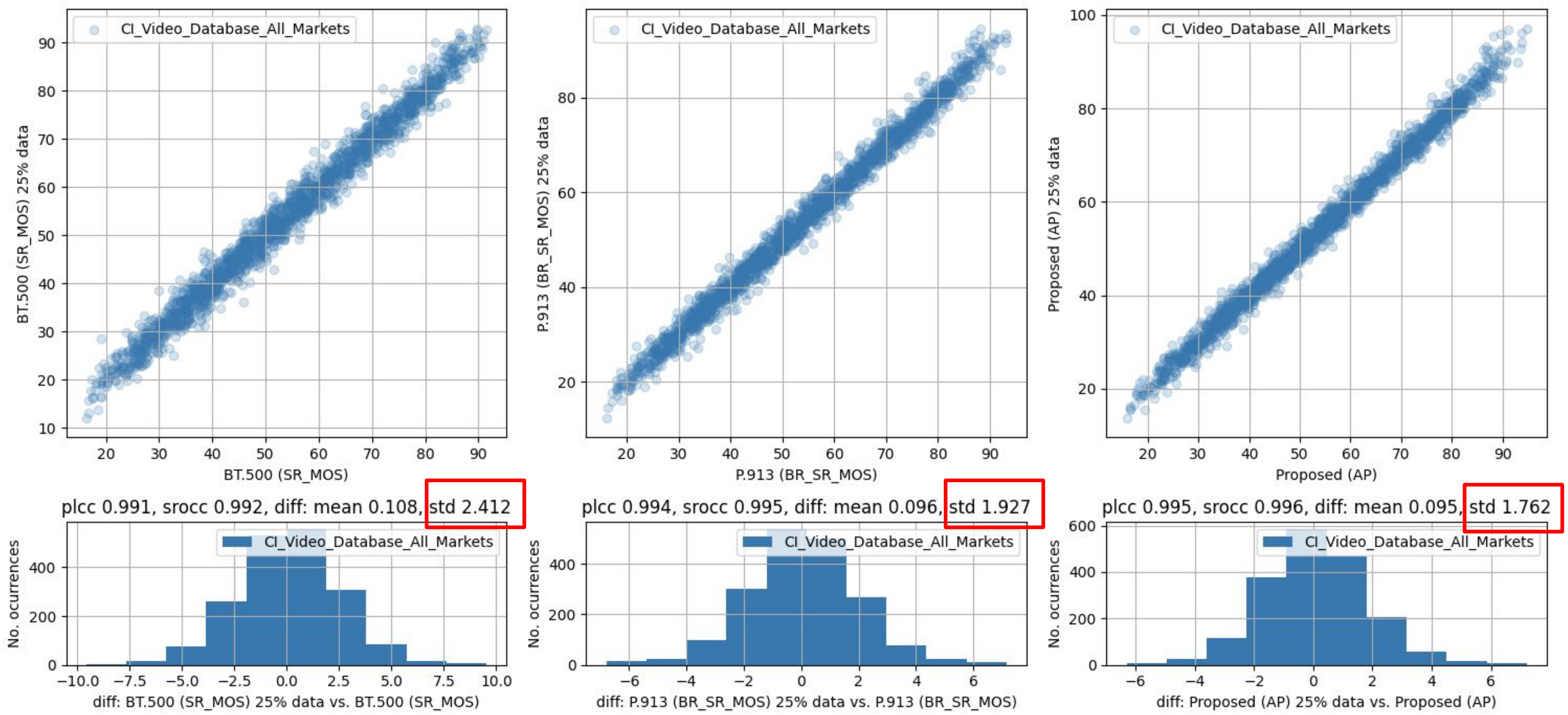
Analyze a crowdsourcing dataset: Scheme X with y% data, vs. Scheme X with 100% data

X = BT.500, P.913, Proposed
y = 50, 25, 10

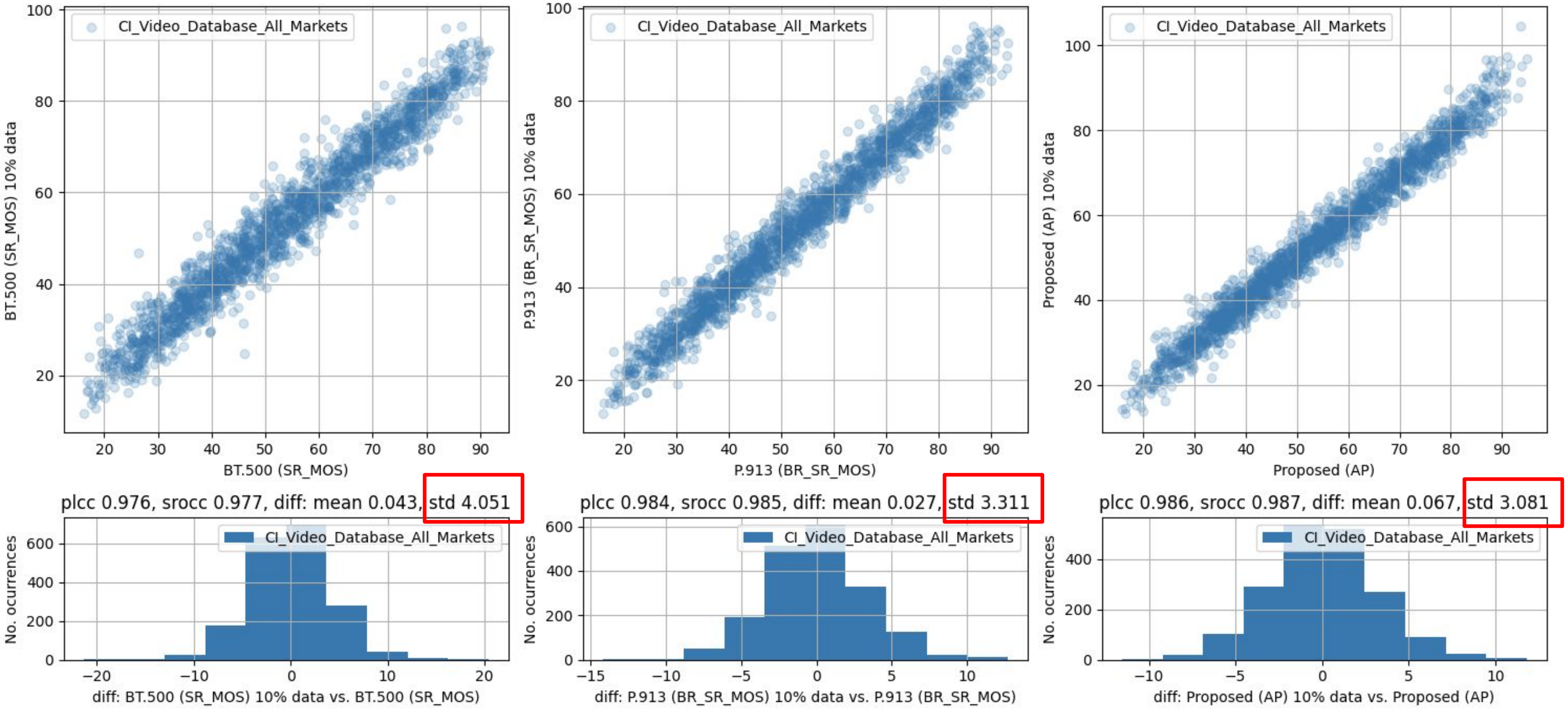
CI Study 2nd Wave Dataset* (Crowdsourcing Study)
100% Data vs. 50% Data



CI Study 2nd Wave Dataset* (Crowdsourcing Study)
100% Data vs. 25% Data



CI Study 2nd Wave Dataset* (Crowdsourcing Study)
100% Data vs. 10% Data



Observations on the correlation between scores recovered from the full data and partial data

- On the crowdsourcing dataset we have tested, the proposed AP method is doing better (yielding higher correlation and lower variance compared to the final score) than P.913, and P.913 is doing better than BT.500.

Analyze VQEG FRTV Phase I Datasets: Cross-Lab Comparison

VQEG FRTV Phase I Datasets

- Four datasets:
  - 525 Line Low
  - 525 Line High
  - 625 Line Low
  - 625 Line High
- In total 8 labs participated in the test
- Each dataset is evaluated by 4 of the 8 labs
- We evaluate the resulting Pearson Linear CC (PLCC) across labs

PLCC Across Labs
VQEG FRTV Phase I - 525 Line Low

Lab	1	4	6	8
1	1.0	0.944	0.9438	0.9485
4		1.0	**0.9577**	0.9411
6			1.0	0.9443
8				1.0

BT.500

Lab	1	4	6	8
1	1.0	0.9504	0.9427	0.95
4		1.0	0.9556	0.9406
6			1.0	0.9447
8				1.0

P.913

Lab	1	4	6	8
1	1.0	**0.9523**	**0.9492**	**0.9588**
4		1.0	0.9574	**0.9454**
6			1.0	**0.9487**
8				1.0

Proposed AP

Colored: Best among three methods

PLCC Across Labs
VQEG FRTV Phase I - 525 Line High

Lab	1	4	6	8
1	1.0	0.8908	0.9002	0.9151
4		1.0	**0.8815**	**0.8505**
6			1.0	**0.8756**
8				1.0

BT.500

Lab	1	4	6	8
1	1.0	0.8889	0.903	0.9063
4		1.0	0.8679	0.834
6			1.0	0.8747
8				1.0

P.913

Lab	1	4	6	8
1	1.0	**0.9068**	**0.9118**	**0.9155**
4		1.0	0.8763	0.8231
6			1.0	0.8331
8				1.0

Proposed AP

Colored: Best among three methods

PLCC Across Labs
VQEG FRTV Phase I - 625 Line Low

Lab	2	3	5	7
2	1.0	0.7435	0.913	0.9149
3		1.0	0.8125	0.7055
5			1.0	0.9004
7				1.0

BT.500

Lab	2	3	5	7
2	1.0	0.7649	0.9084	0.9043
3		1.0	0.8408	0.7559
5			1.0	0.9055
7				1.0

P.913

Lab	2	3	5	7
2	1.0	**0.8149**	**0.9264**	**0.923**
3		1.0	**0.875**	**0.8047**
5			1.0	**0.9184**
7				1.0

Proposed AP

Colored: Best among three methods

PLCC Across Labs
VQEG FRTV Phase I - 625 Line High

Lab	2	3	5	7
2	1.0	0.7904	**0.8538**	**0.8182**
3		1.0	0.818	0.8363
5			1.0	0.8694
7				1.0

BT.500

Lab	2	3	5	7
2	1.0	0.7646	0.7949	0.7377
3		1.0	**0.8263**	0.8341
5			1.0	0.8495
7				1.0

P.913

Lab	2	3	5	7
2	1.0	**0.8296**	0.8184	0.8004
3		1.0	0.8254	**0.8604**
5			1.0	**0.8742**
7				1.0

Proposed AP

Colored: Best among three methods

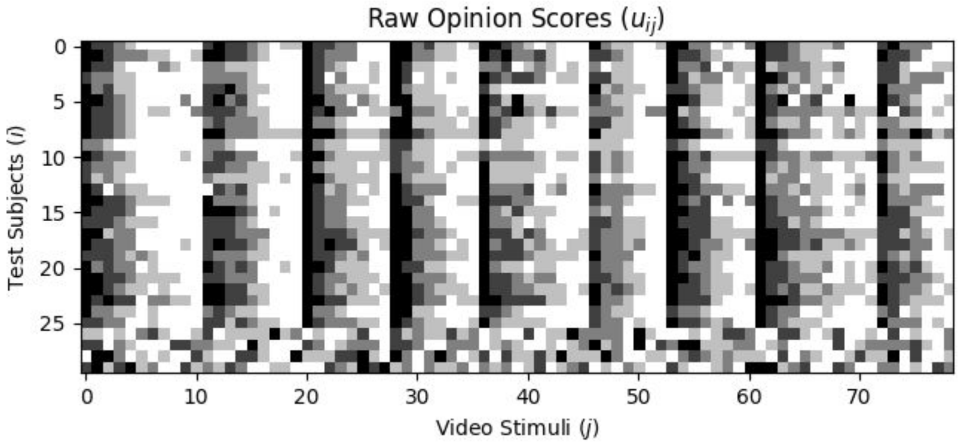
Observations from VQEG FRTV Phase I Dataset

- Statistically, the proposed AP method yields better consistency (higher PLCC) across labs than BT.500 and P.913

Outline

- Background and motivation
- Proposed methodology
- Progress since ITU-T SG12 C470
    - New comparison results with BT.500 / P.913
    - Interpreting the limitations of BT.500 / P.913
    - Update on the calculation of confidence intervals
    - Runtime analysis

NFLX Public Dataset



Black: lowest score
White: highest score

Last 4 Outliers

NFLX Public Dataset (BT.500 subject rejection)

For each test presentation, calculate the mean, $\bar{u}_{jkr}$, standard deviation, S_{jkr}, and kurtosis coefficient, β_{2jkr}, where β_{2jkr} is given by:

$$\beta_{2jkr} = \frac{m_4}{(m_2)^2} \quad \text{with} \quad m_x = \frac{\sum_{i=1}^{N} (u_{ijkr} - \bar{u}_{ijkr})^x}{N} \tag{5}$$

For each observer, i, find P_i and Q_i, i.e.:

for $j, k, r = 1, 1, 1$ to J, K, R

if $2 \leq \beta_{2jkr} \leq 4$, then:

if $u_{ijkr} \geq \bar{u}_{jkr} + 2\,S_{jkr}$ then $P_i = P_i + 1$

if $u_{ijkr} \leq \bar{u}_{jkr} - 2\,S_{jkr}$ then $Q_i = Q_i + 1$

else:

if $u_{ijkr} \geq \bar{u}_{jkr} + \sqrt{20}\;S_{jkr}$ then $P_i = P_i + 1$

if $u_{ijkr} \leq \bar{u}_{jkr} - \sqrt{20}\;S_{jkr}$ then $Q_i = Q_i + 1$

If $\frac{P_i + Q_i}{J \cdot K \cdot R} > 0.05$ and $\left|\frac{P_i - Q_i}{P_i + Q_i}\right| < 0.3$ then reject observer i

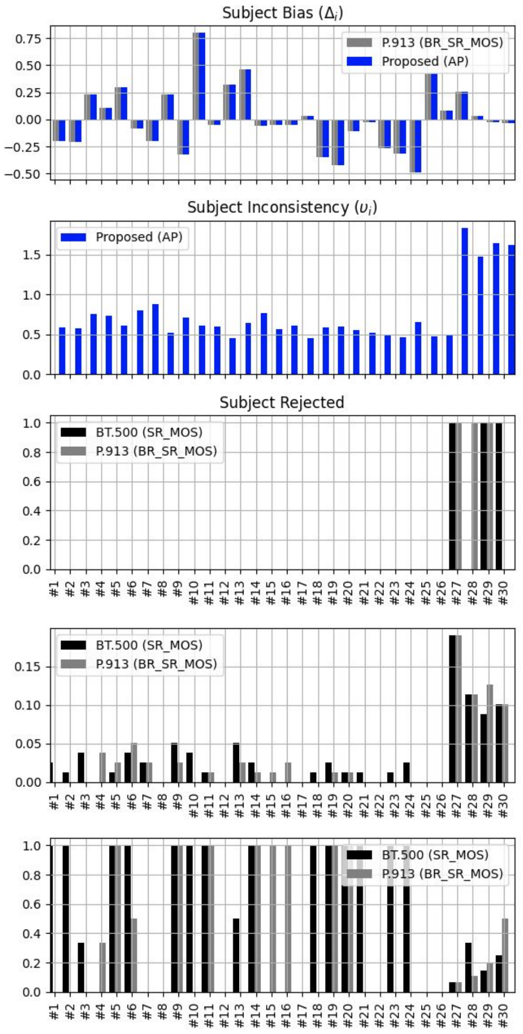
with:

- N: number of observers
- J: number of test conditions including the reference
- K: number of test images or sequences
- R: number of repetitions
- L: number of test presentations (in most cases the number of presentations will be equal to $J \cdot K \cdot R$, however it is noted that some assessments may be conducted with unequal numbers of sequences for each test condition).

Rejected

$$\frac{P_i + Q_i}{J \cdot K \cdot R}$$

$$\left|\frac{P_i - Q_i}{P_i + Q_i}\right|$$



its4s_NTIA Dataset (BT.500 subject rejection)

For each test presentation, calculate the mean, $\bar{u}_{jkr}$, standard deviation, S_{jkr}, and kurtosis coefficient, β_{2jkr}, where β_{2jkr} is given by:

$$\beta_{2jkr} = \frac{m_4}{(m_2)^2} \quad \text{with} \quad m_x = \frac{\sum_{i=1}^{N} (u_{ijkr} - \bar{u}_{ijkr})^x}{N} \tag{5}$$

For each observer, i, find P_i and Q_i, i.e.:

for $j, k, r = 1, 1, 1$ to J, K, R

if $2 \le \beta_{2jkr} \le 4$, then:

if $u_{ijkr} \ge \bar{u}_{jkr} + 2\, S_{jkr}$ then $P_i = P_i + 1$

if $u_{ijkr} \le \bar{u}_{jkr} - 2\, S_{jkr}$ then $Q_i = Q_i + 1$

else:

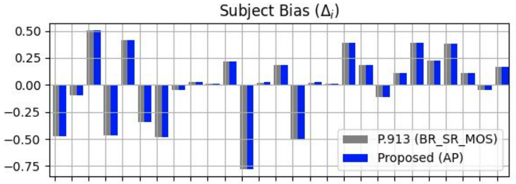
if $u_{ijkr} \ge \bar{u}_{jkr} + \sqrt{20}\, S_{jkr}$ then $P_i = P_i + 1$

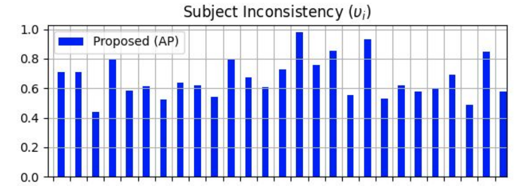
if $u_{ijkr} \le \bar{u}_{jkr} - \sqrt{20}\, S_{jkr}$ then $Q_i = Q_i + 1$

If $\frac{P_i + Q_i}{J \cdot K \cdot R} > 0.05$ and $\left|\frac{P_i - Q_i}{P_i + Q_i}\right| < 0.3$ then reject observer i

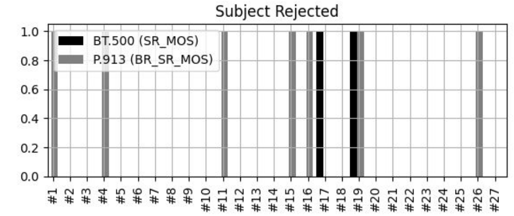
with:

- N: number of observers
- J: number of test conditions including the reference
- K: number of test images or sequences
- R: number of repetitions
- L: number of test presentations (in most cases the number of presentations will be equal to $J \cdot K \cdot R$, however it is noted that some assessments may be conducted with unequal numbers of sequences for each test condition).

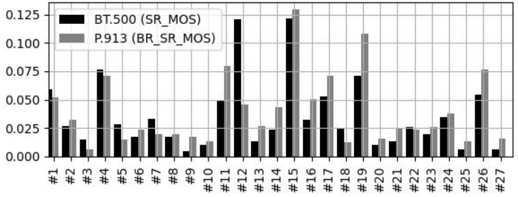




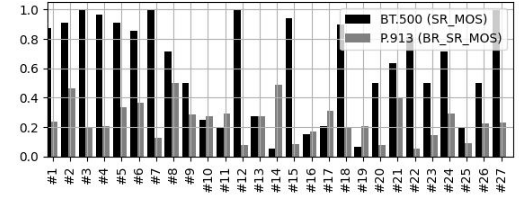
Rejected



$$\frac{P_i + Q_i}{J \cdot K \cdot R}$$



$$\left|\frac{P_i - Q_i}{P_i + Q_i}\right|$$



Observations on BT.500 and P.913

$$\text{If} \quad \frac{P_i + Q_i}{J \cdot K \cdot R} > 0.05 \quad \text{and} \quad \left| \frac{P_i - Q_i}{P_i + Q_i} \right| < 0.3 \quad \text{then reject observer } i$$

- The hard-coded rules (the outlier % detection and skewness detection) can cause missing outliers
- BT.500 and P.913 often yield contradictory results
- P.913’s rejection result is more consistent than BT.500 with the subjects with large inconsistency predicted by the proposed AP method

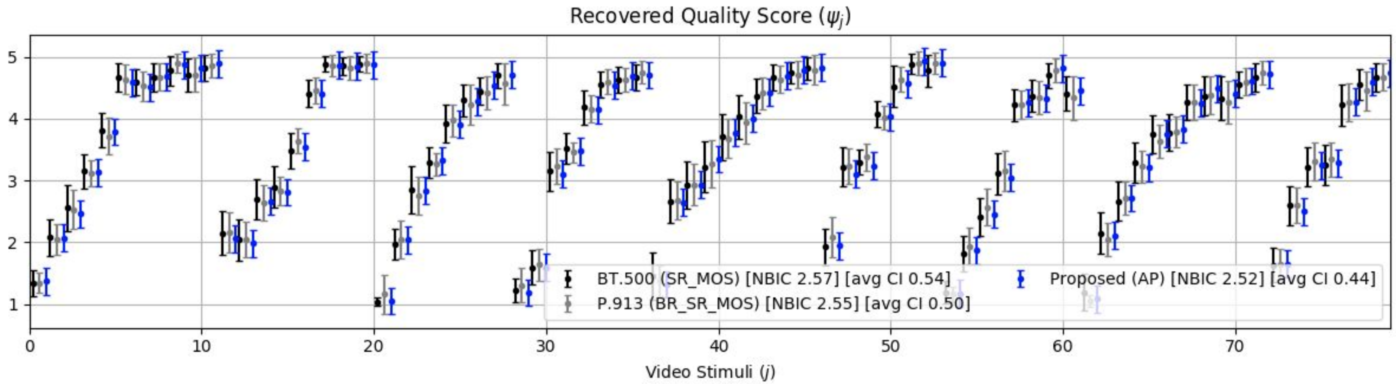
Outline

- Background and motivation
- Proposed methodology
- Progress since ITU-T SG12 C470
    - New comparison results with BT.500 / P.913
    - Interpreting the limitations of BT.500 / P.913
    - Update on the calculation of confidence intervals
    - Runtime analysis

Confidence Intervals (CI) of Quality Scores

- Estimated CI based on Cramer-Rao bound:

$$CI(\psi_j) = \psi_j \pm 1.96 \frac{1}{\sqrt{\sum_{ir} v_i^{-2}}},$$



The previously proposed method results in equal CI lengths for all quality scores !

Revised CI Formula - AP vs. AP2

$$CI(\psi_j) = \psi_j \pm 1.96 \frac{1}{\sqrt{\sum_{ir} \upsilon_i^{-2}}},$$

$$\upsilon_i = \sigma_i(\{\epsilon_{ijr}\})$$

$$\sigma_i(\{\epsilon_{ijr}\}) = \sqrt{(\sum_{jr} 1)^{-1} \sum_{jr} (\epsilon_{ijr} - \epsilon_i)^2 - \epsilon_i^2},$$

$$\epsilon_i = (\sum_{jr} 1)^{-1} \sum_{jr} \epsilon_{ijr}.$$

AP

$$CI_2(\psi_j) = \psi_j \pm 1.96 \frac{\upsilon_j}{\sqrt{\sum_{ir} 1}}.$$

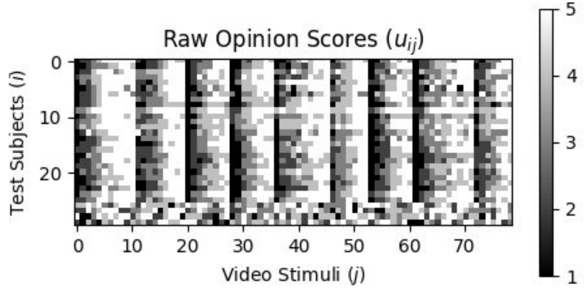
$$\upsilon_j = \sigma_j(\{\epsilon_{ijr}\})$$

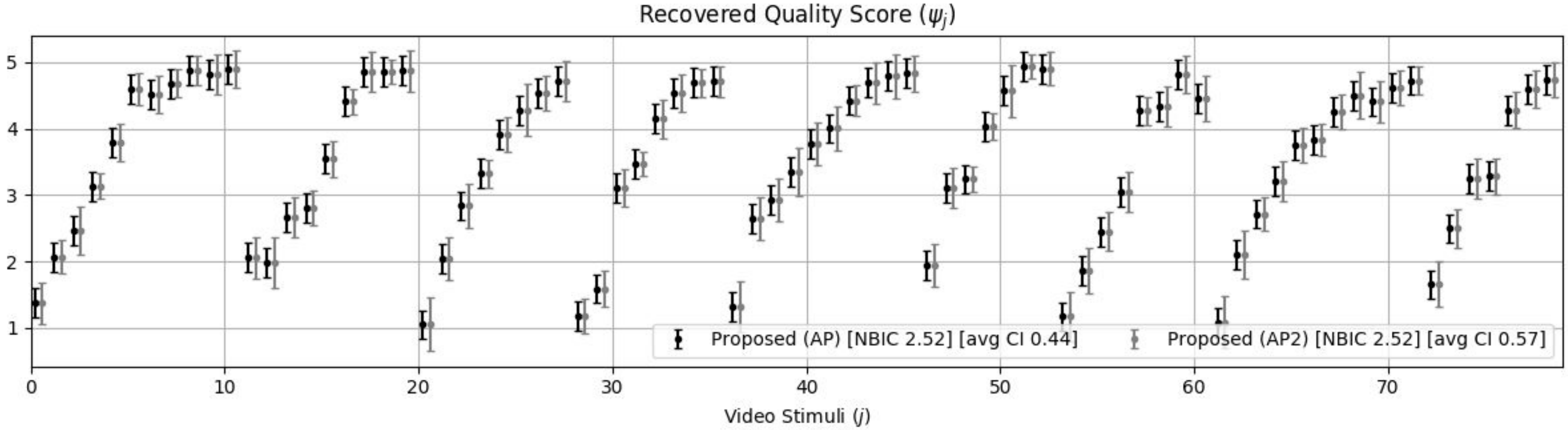
$$\sigma_j(\{\epsilon_{ijr}\}) = \sqrt{(\sum_{ir} 1)^{-1} \sum_{ir} (\epsilon_{ijr} - \epsilon_j)^2 - \epsilon_j^2},$$

$$\epsilon_j = (\sum_{ir} 1)^{-1} \sum_{ir} \epsilon_{ijr}.$$

AP2

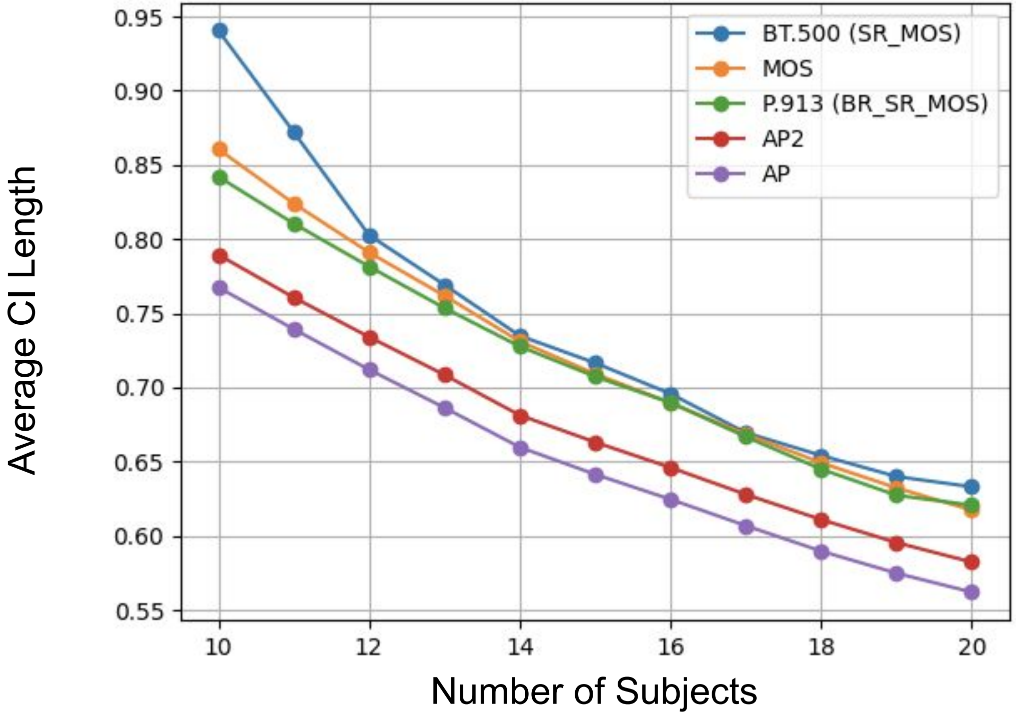
NFLX_pub - AP vs. AP2





AP has equal CI length while AP2 has unequal CI length

Average CI length vs. Subject numbers
Quality Variation 2017 AGH TV Dataset (Lab Study)



*Each point is an average of 100x randomly selecting a given number of subjects.

Average CI length

Dataset	MOS	BT.500	P.913	NR/AP (AP2)
VQEG HD3	0.59	0.60	0.49	**0.46** (0.47)
NFLX Public	0.62	0.54	0.5	**0.44** (0.57)
HDTV Ph1 Exp1	0.50	0.61	0.48	**0.46** (0.46)
HDTV Ph1 Exp2	0.57	0.57	0.53	**0.48** (0.49)
HDTV Ph1 Exp3	0.56	0.59	0.52	**0.48** (0.48)
HDTV Ph1 Exp4	0.63	0.63	0.52	**0.47** (0.49)
HDTV Ph1 Exp5	0.57	0.57	0.53	**0.49** (0.5)
HDTV Ph1 Exp6	0.50	0.51	0.48	**0.45** (0.45)
ITU-T Supp23 Exp1	0.61	0.61	0.56	**0.47** (0.50)
MM2 1	0.59	0.60	0.57	**0.53** (0.55)
MM2 2	1.21	1.21	1.12	**0.88** (0.99)
MM2 3	0.47	0.48	0.45	**0.42** (0.43)
MM2 4	0.58	0.59	0.54	**0.48** (0.51)
MM2 5	0.63	0.65	0.58	**0.52** (0.56)
MM2 6	0.62	0.70	0.59	**0.56** (0.57)
MM2 7	0.60	0.61	0.57	**0.55** (0.55)
MM2 8	0.76	0.76	0.71	**0.66** (0.68)
MM2 9	0.84	0.85	0.74	**0.68** (0.71)
MM2 10	0.77	0.83	0.73	**0.70** (0.70)
its4s2	0.82	0.82	0.66	**0.60** (0.64)
its4s AGH	0.68	0.68	0.61	**0.56** (0.60)
its4s NTIA	0.57	0.58	0.54	**0.48** (0.50)

Table 2: Average length of confidence intervals of the estimated quality scores reported on the compared methods on public datasets. The NR and AP methods produce identical results. MOS: arithmetic mean of all opinion scores; NR: Newton-Raphson; AP: Alternating Projection. NR and AP methods produce identical results. In the last column, two CI lengths are reported: the CI based on (8) and (in the parenthesis) the alternative CI based on (11).

Confidence Interval Validation*

Dataset	MOS	NR			AP (AP2)		
	ψ_j	ψ_j	Δ_i	υ_i	ψ_j	Δ_i	υ_i
VQEG HD3	93.3	93.6	93.9	93.0	93.2 (93.5)	94.4	91.9
NFLX Public	94.2	93.7	94.5	93.1	93.5 (97.5)	94.1	92.3
HDTV Ph1 Exp1	93.9	94.1	93.9	93.1	93.8 (93.2)	94.2	91.3
HDTV Ph1 Exp2	93.8	94.0	94.5	92.5	93.8 (94.1)	94.0	91.2
HDTV Ph1 Exp3	93.9	93.9	94.4	92.5	93.7 (93.6)	94.1	90.6
HDTV Ph1 Exp4	93.8	94.0	94.3	91.9	93.8 (94.1)	94.1	90.9
HDTV Ph1 Exp5	93.8	94.1	94.2	92.2	93.9 (93.8)	94.2	90.9
HDTV Ph1 Exp6	93.8	94.0	94.4	92.6	93.9 (93.6)	94.0	91.0
ITU-T Supp23 Exp1	93.8	94.0	94.4	91.2	93.8 (94.5)	94.9	90.0
MM2 1	93.5	92.8	95.4	92.6	92.5 (93.8)	94.0	91.6
MM2 2	92.1	81.5	92.9	80.0	68.1 (87.5)	92.1	75.4
MM2 3	94.4	93.6	95.1	93.4	93.4 (94.1)	94.2	92.0
MM2 4	93.2	93.6	95.6	93.0	93.2 (94.7)	95.1	92.0
MM2 5	93.2	93.2	95.7	92.7	91.8 (94.3)	95.3	91.4
MM2 6	93.6	93.3	95.2	92.8	93.0 (93.8)	94.1	91.4
MM2 7	93.6	93.3	95.2	92.8	92.9 (93.2)	94.2	91.9
MM2 8	93.0	92.4	95.4	88.8	92.2 (92.6)	94.5	87.0
MM2 9	93.2	93.3	94.8	89.1	92.8 (93.3)	94.2	88.1
MM2 10	93.2	93.1	95.7	89.7	92.8 (92.3)	94.5	87.9
its4s2	93.1	94.1	94.6	60.6	94.1 (94.0)	94.2	59.2
its4s AGH	93.6	94.0	94.4	90.4	94.0 (94.8)	94.4	89.7
its4s NTIA	93.9	94.4	94.7	86.1	94.3 (95.0)	95.1	85.6

Table 3: Average confidence interval coverage (CI%) reported on public datasets. For each proposed solver and each dataset, run the solver to estimate the parameters. Treat the estimated parameters as "synthetic" parameters, run simulations to generate synthetic samples according to the model (1) (except for MOS, whose samples are generated according to (12)). Run the solver again on the synthetic data to yield the "recovered" parameters and their confidence intervals. The reported "CI%" is the percentage of occurrences when the synthetic ground truth falls within the confidence interval. For each dataset, the simulation is run 100 times with different seeds. Note that for both MOS and the proposed NR and AP methods, the CI% is slightly below 95%, due to the underlying Gaussian assumption used instead of the legitimate Student's t-distrubtion. (MOS: plain mean opinion score; NR: Newton-Raphson; AP: Alternating Projection.) For ψ_j of AP, two CI% are reported: the CI based on (8) and (in the parenthesis) the alternative CI calculated based on (11).

*Using synthetic data generated from each model, the CI should match closely to 95%.

Outline

- Background and motivation
- Proposed methodology
- Progress since ITU-T SG12 C470
    - New comparison results with BT.500 / P.913
    - Interpreting the limitations of BT.500 / P.913
    - Update on the calculation of confidence intervals
    - Runtime analysis

Runtime Analysis

Dataset	Mean Runtime (seconds)					No. Iterations	
	MOS	BT.500	P.913	NR	AP	NR	AP
VQEG HD3	5.2e-4	1.5e-2	1.5e-2	2.1e-1	4.3e-3	26.2	12.1
NFLX Public	5.7e-4	1.8e-2	1.9e-2	2.8e-1	4.5e-3	34.5	11.8
HDTV Ph1 Exp1	7.7e-4	3.3e-2	3.4e-2	2.0e-1	4.6e-3	23.4	10.3
HDTV Ph1 Exp2	7.8e-4	3.3e-2	3.4e-2	2.8e-1	4.9e-3	33.2	11.3
HDTV Ph1 Exp3	7.8e-4	3.3e-2	3.4e-2	2.5e-1	4.7e-3	29.4	10.7
HDTV Ph1 Exp4	7.6e-4	3.3e-2	3.4e-2	3.3e-1	5.0e-3	38.3	11.5
HDTV Ph1 Exp5	7.8e-4	3.3e-2	3.4e-2	2.7e-1	4.7e-3	31.3	10.8
HDTV Ph1 Exp6	7.6e-4	3.3e-2	3.4e-2	2.2e-1	4.6e-3	25.8	10.7
ITU-T Supp23 Exp1	8.1e-4	3.5e-2	3.5e-2	3.4e-1	5.0e-3	36.0	11.6
MM2 1	4.9e-4	1.3e-2	1.3e-2	2.1e-1	4.3e-3	27.4	12.4
MM2 2	4.0e-4	1.0e-2	1.1e-2	5.8e-1	1.4e-2	78.0	54.9
MM2 3	5.3e-4	1.3e-2	1.4e-2	1.8e-1	4.2e-3	23.3	11.6
MM2 4	5.0e-4	1.3e-2	1.4e-2	2.6e-1	4.6e-3	33.4	13.8
MM2 5	5.0e-4	1.3e-2	1.4e-2	2.9e-1	6.0e-3	37.3	19.3
MM2 6	4.8e-4	1.2e-2	1.3e-2	2.2e-1	4.3e-3	28.8	13.1
MM2 7	4.8e-4	1.2e-2	1.3e-2	2.0e-1	4.2e-3	25.6	12.3
MM2 8	4.3e-4	1.1e-2	1.1e-2	2.7e-1	5.5e-3	35.3	18.7
MM2 9	4.3e-4	1.1e-2	1.2e-2	2.8e-1	5.1e-3	36.5	16.8
MM2 10	4.3e-4	1.1e-2	1.2e-2	2.3e-1	4.8e-3	29.8	15.4
its4s2	3.3e-3	2.5e-1	2.5e-1	1.1e+0	1.3e-2	49.8	13.3
its4s AGH	8.7e-4	4.1e-2	4.2e-2	3.5e-1	5.3e-3	39.4	11.6
its4s NTIA	2.6e-3	1.6e-1	1.6e-1	6.4e-1	1.1e-2	46.2	11.3

Table 4: Average runtime in seconds and number of iterations (for NR and AP) reported on public datasets. For each proposed solver and each dataset, run the solver to estimate the parameters. Treat the estimated parameters and the "synthetic" parameters, run simulations to generate synthetic samples according to the model (1) (except for MOS, whose samples are generated according to (12)). Run the solver again on the synthetic data. For each dataset, the simulation is run 100 times with different seeds, and the mean is reported. For NR and AP, also reported are the number of iterations. (MOS: plain mean opinion score; NR: Newton-Raphson; AP: Alternating Projection.)

Conclusions

- Recommendations for subject experiment data analysis process such as ITU-R BT.500 and ITU-T P.913 are not without their own limitations
- We propose new model and the corresponding MLE-based solver, which can be considered as a generalization of P.913, with the following advantages:
  - Better model-data fit
  - Tighter confidence intervals (hence less #subjects required)
  - Better robustness against subject outliers
  - Negligible runtime increase - similar to BT.500/P.913
  - Absence of hard coded parameters / thresholds
  - Auxiliary information on test subjects
- We propose to standardize the AP method (with confidence interval calculated as in AP2) in recommendation P.913 (and in the future, BT.500)

Publication & Open Source Code

Home / Electronic Imaging, Human Vision and Electronic Imaging 2020

A Simple Model for Subject Behavior in Subjective Experiments

Download Article:
Download
(PDF 1,450.2 kb)

Authors: Li, Zhi; Bampis, Christos G.; Janowski, Lucjan; Katsavounidis, Ioannis
Source: Electronic Imaging, Human Vision and Electronic Imaging 2020, pp. 131-1-131-14(14)
Publisher: Society for Imaging Science and Technology
DOI: https://doi.org/10.2352/ISSN.2470-1173.2020.11.HVEI-131

< previous article | view table of contents | next article > | ADD TO FAVOURITES

Abstract | References | Citations | Supplementary Data | Article Media | Metrics | Suggestions

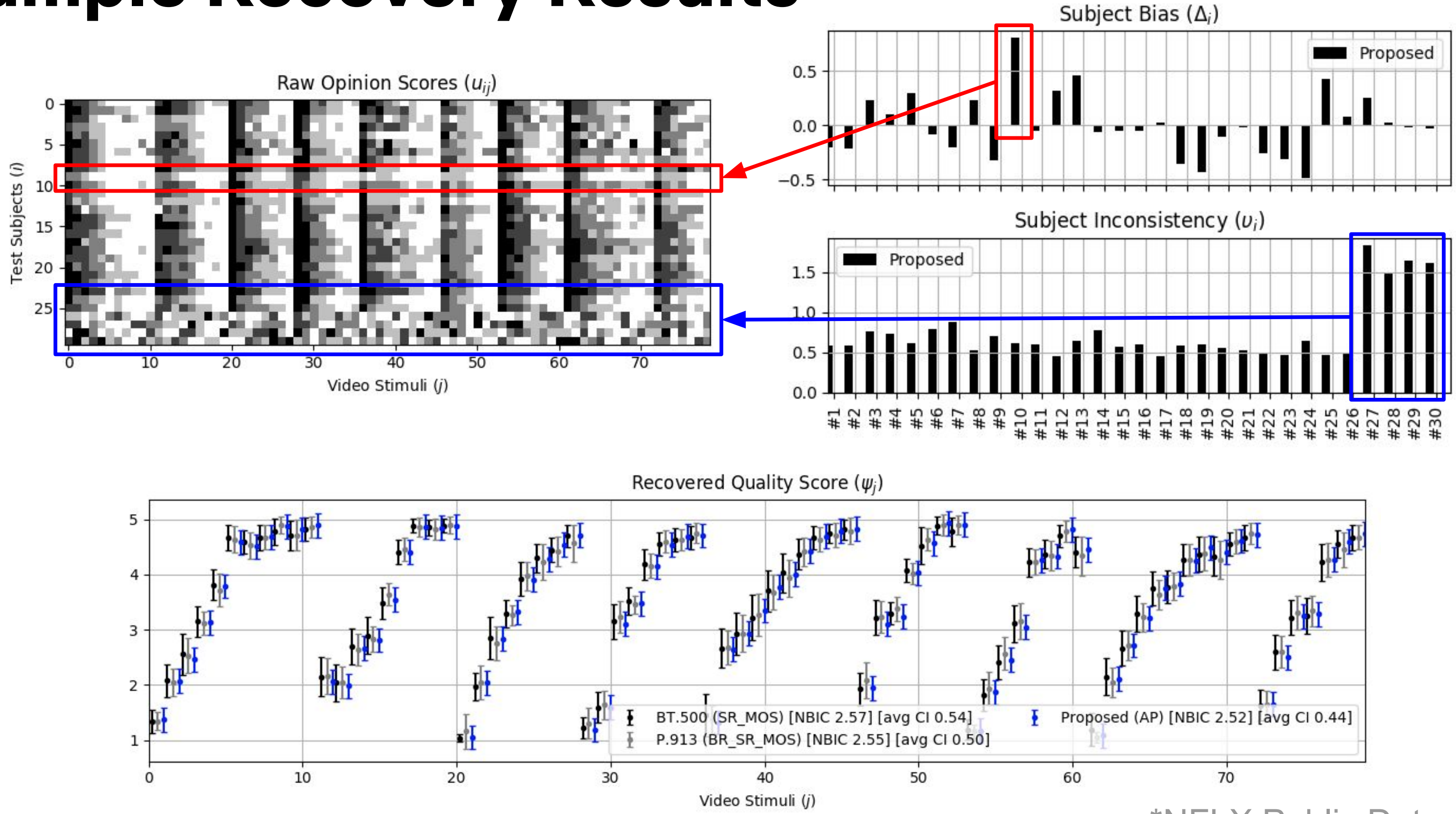
In a subjective experiment to evaluate the perceptual audiovisual quality of multimedia and television services, raw opinion scores offered by subjects are often noisy and unreliable. Recommendations such as ITU-R BT.500, ITU-T P.910 and ITU-T P.913 standardize post-processing procedures to clean up the raw opinion scores, using techniques such as subject outlier rejection and bias removal. In this paper, we analyze the prior standardized techniques to demonstrate their weaknesses. As an alternative, we propose a simple model to account for two of the most dominant behaviors of subject inaccuracy: bias (aka systematic error) and inconsistency (aka random error). We further show that this model can also effectively deal with inattentive subjects that give random scores. We propose to use maximum likelihood estimation (MLE) to jointly estimate the model parameters, and present two numeric solvers: the first based on the Newton-Raphson method, and the second based on alternating projection. We show that the second solver can be considered as a generalization of the subject bias removal procedure in ITU-T P.913. We compare the proposed methods with the standardized techniques using real datasets and synthetic simulations, and demonstrate that the proposed methods have advantages in better model-data fit, tighter confidence intervals, better robustness against subject outliers, shorter runtime, the absence of hard coded parameters and thresholds, and auxiliary information on test subjects. The source code for this work is open-sourced at https://github.com/Netflix/sureal.

Keywords: maximum likelihood estimation; mean opinion score; subject rejection; subjective experiment

Publication available at: https://www.ingentaconnect.com/contentone/ist/ei/pre-prints/content-ei2020-hvei-131
Source code - *free and open-sourced* - can be found at https://github.com/Netflix/sureal

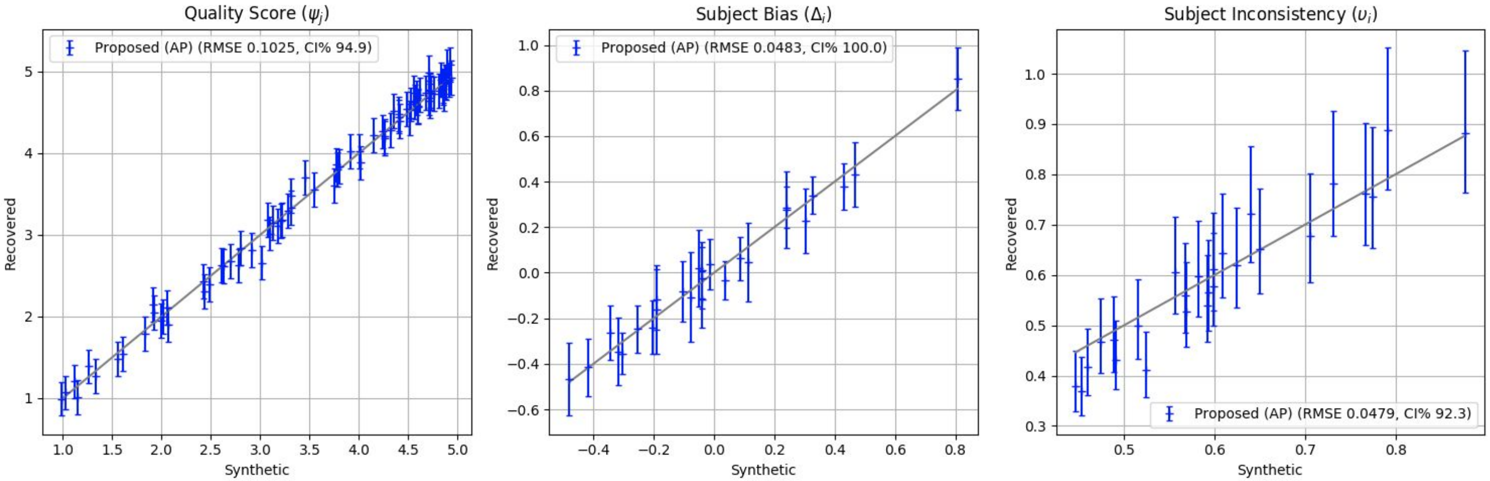
Backup Slides

Sample Recovery Results



*NFLX Public Dataset

Solver Accuracy Validation Using Synthetic Data



- Synthetic data generation
  - Take NFLX Public dataset, run solver to estimate parameters
  - Treat the estimated parameters as synthetic “ground truth”, run simulations to generate synthetic samples according to the proposed model
  - Run solver on the generated samples to recover the parameters again

SR: subject rejection; BR: bias removal; MOS: mean opinion score; RMSE: root mean squared error

Model-Data Fit Validation Using Bayesian Information Criterion

- BIC is a criterion for model fitting, balancing between:
  - The degree of freedom (number of parameters)
  - The goodness of fit (log-likelihood function)

- Use "normalized" BIC (NBIC) to compare across datasets

$$NBIC = \frac{log\,(n)|\theta| - 2L(\theta)}{n}$$

  - $|\theta|$ - the number of model parameters
  - n - the number of observations (i.e. raw opinion scores)
  - $L(\theta)$ - log-likelihood function

Normalized Bayesian Information Criterion (NBIC)*

Table NBIC: Normalized Bayesian Information Criterion (NBIC) reported on the compared methods on public datasets. The NR and AP methods produce identical results. (MOS: plain mean opinion score; NR: Newton-Raphson; AP: Alternating Projection.)

Dataset	MOS	BT.500	P.913	NR/AP
VQEG HD3	2.75	2.74	2.39	**2.30**
NFLX Public	2.97	2.57	2.55	**2.52**
HDTV Ph1 Exp1	2.45	2.46	2.38	**2.20**
HDTV Ph1 Exp2	2.72	2.72	2.52	**2.32**
HDTV Ph1 Exp3	2.72	2.71	2.37	**2.29**
HDTV Ph1 Exp4	2.96	2.96	2.51	**2.27**
HDTV Ph1 Exp5	2.77	2.77	2.47	**2.33**
HDTV Ph1 Exp6	2.51	2.49	2.32	**2.16**
ITU-T Supp23 Exp1	2.91	2.91	2.35	**2.31**
MM2 1	2.80	2.78	2.83	**2.74**
MM2 2	3.89	3.89	3.52	**3.13**
MM2 3	2.48	2.47	2.45	**2.41**
MM2 4	2.74	2.73	2.62	**2.47**
MM2 5	2.90	2.82	2.67	**2.64**
MM2 6	2.81	2.74	2.74	**2.72**
MM2 7	2.73	2.72	2.76	**2.67**
MM2 8	3.00	2.92	2.88	**2.70**
MM2 9	3.27	3.21	2.95	**2.79**
MM2 10	3.04	3.05	2.98	**2.82**
its4s2	3.63	3.63	2.96	**2.59**
its4s AGH	3.15	3.05	2.77	**2.64**
its4s NTIA	2.94	2.91	2.53	**2.38**

*The model with the smallest NBIC is preferred.

Confidence intervals of Estimated Quality Scores

Table CI: Average length of confidence intervals of the estimated quality scores reported on the compared methods on public datasets. The NR and AP methods produce identical results. (MOS: plain mean opinion score; NR: Newton-Raphson; AP: Alternating Projection.)

Dataset	MOS	BT.500	P.913	NR/AP
VQEG HD3	0.59	0.60	0.49	**0.46**
NFLX Public	0.62	0.54	0.5	**0.44**
HDTV Ph1 Exp1	0.50	0.61	0.48	**0.46**
HDTV Ph1 Exp2	0.57	0.57	0.53	**0.48**
HDTV Ph1 Exp3	0.56	0.59	0.52	**0.48**
HDTV Ph1 Exp4	0.63	0.63	0.52	**0.47**
HDTV Ph1 Exp5	0.57	0.57	0.53	**0.49**
HDTV Ph1 Exp6	0.50	0.51	0.48	**0.45**
ITU-T Supp23 Exp1	0.61	0.61	0.56	**0.47**
MM2 1	0.59	0.60	0.57	**0.53**
MM2 2	1.21	1.21	1.12	**0.88**
MM2 3	0.47	0.48	0.45	**0.42**
MM2 4	0.58	0.59	0.54	**0.48**
MM2 5	0.63	0.65	0.58	**0.52**
MM2 6	0.62	0.70	0.59	**0.56**
MM2 7	0.60	0.61	0.57	**0.55**
MM2 8	0.76	0.76	0.71	**0.66**
MM2 9	0.84	0.85	0.74	**0.68**
MM2 10	0.77	0.83	0.73	**0.70**
its4s2	0.82	0.82	0.66	**0.60**
its4s AGH	0.68	0.68	0.61	**0.56**
its4s NTIA	0.57	0.58	0.54	**0.48**

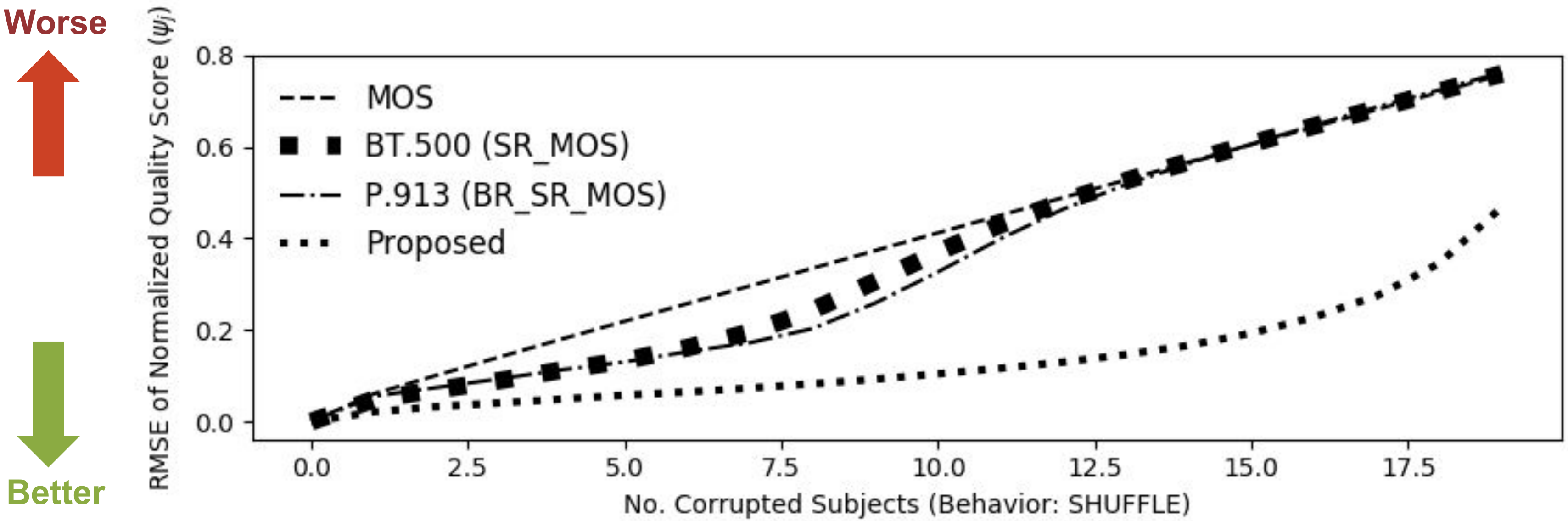
Confidence Interval Validation*

Table CI%: Average confidence interval coverage (CI%) reported on public datasets. For each proposed solver and each dataset, run the solver to estimate the parameters. Treat the estimated parameters and the "synthetic" parameters, run simulations to generate synthetic samples according to the model (1) (except for MOS, whose samples are generated according to (7)). Run the solver again on the synthetic data to yield the "recovered" parameters and their confidence intervals. The reported "CI%" is the percentage of occurrences when the synthetic ground truth falls within the confidence interval. For each dataset, the simulation is run 100 times with different seeds. Note that for both MOS and the proposed NR and AP methods, the CI% is slightly below 95%, due to the underlying Gaussian assumption used instead of the legitimate Student's *t*-distrubtion. (MOS: plain mean opinion score; NR: Newton-Raphson; AP: Alternating Projection.)

Dataset	MOS	NR			AP		
	ψ_j	ψ_j	Δ_i	υ_i	ψ_j	Δ_i	υ_i
VQEG HD3	93.3	93.6	93.9	93.0	93.2	94.4	91.9
NFLX Public	94.2	93.7	94.5	93.1	93.5	94.1	92.3
HDTV Ph1 Exp1	93.9	94.1	93.9	93.1	93.8	94.2	91.3
HDTV Ph1 Exp2	93.8	94.0	94.5	92.5	93.8	94.0	91.2
HDTV Ph1 Exp3	93.9	93.9	94.4	92.5	93.7	94.1	90.6
HDTV Ph1 Exp4	93.8	94.0	94.3	91.9	93.8	94.1	90.9
HDTV Ph1 Exp5	93.8	94.1	94.2	92.2	93.9	94.2	90.9
HDTV Ph1 Exp6	93.8	94.0	94.4	92.6	93.9	94.0	91.0
ITU-T Supp23 Exp1	93.8	94.0	94.4	91.2	93.8	94.9	90.0
MM2 1	93.5	92.8	95.4	92.6	92.5	94.0	91.6
MM2 2	92.1	81.5	92.9	80.0	68.1	92.1	75.4
MM2 3	94.4	93.6	95.1	93.4	93.4	94.2	92.0
MM2 4	93.2	93.6	95.6	93.0	93.2	95.1	92.0
MM2 5	93.2	93.2	95.7	92.7	91.8	95.3	91.4
MM2 6	93.6	93.3	95.2	92.8	93.0	94.1	91.4
MM2 7	93.6	93.3	95.2	92.8	92.9	94.2	91.9
MM2 8	93.0	92.4	95.4	88.8	92.2	94.5	87.0
MM2 9	93.2	93.3	94.8	89.1	92.8	94.2	88.1
MM2 10	93.2	93.1	95.7	89.7	92.8	94.5	87.9
its4s2	93.1	94.1	94.6	60.6	94.1	94.2	59.2
its4s AGH	93.6	94.0	94.4	90.4	94.0	94.4	89.7
its4s NTIA	93.9	94.4	94.7	86.1	94.3	95.1	85.6

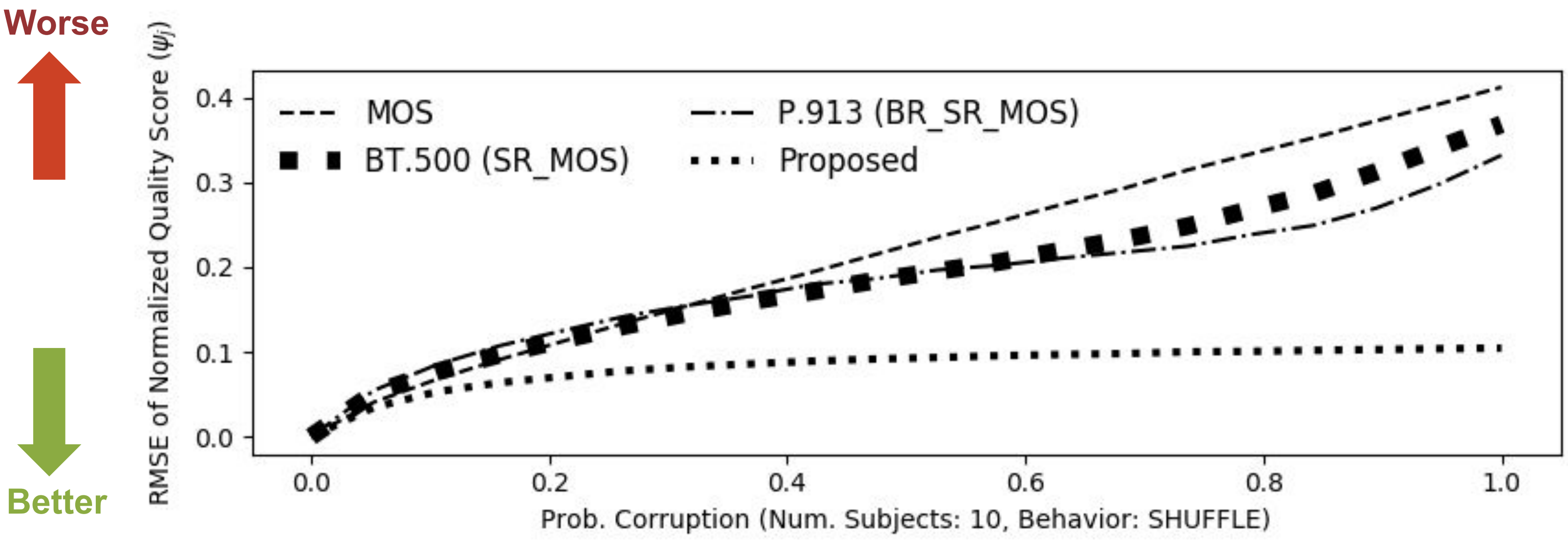
*Using synthetic data generated from each model, the CI should match closely to 95%.

Robustness Against Subject Outliers



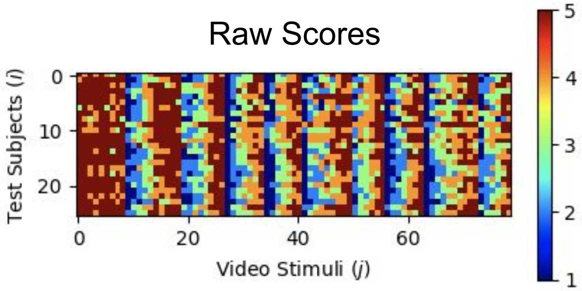
Random behavior: a subject's scores are shuffled among themselves
Y-axis: RMSE with respect to clean case
SR: subject rejection; BR: bias removal; MOS: mean opinion score; RMSE: root mean squared error

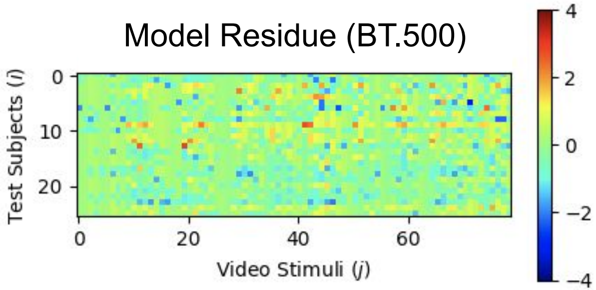
Robustness Against Increasing Corruption Probability

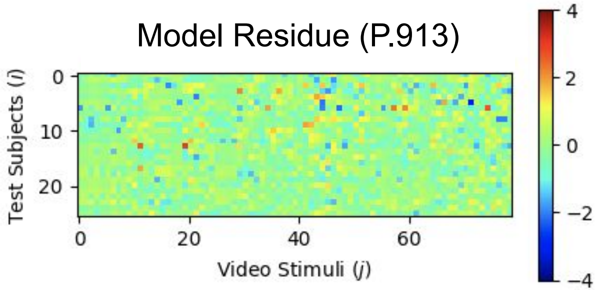


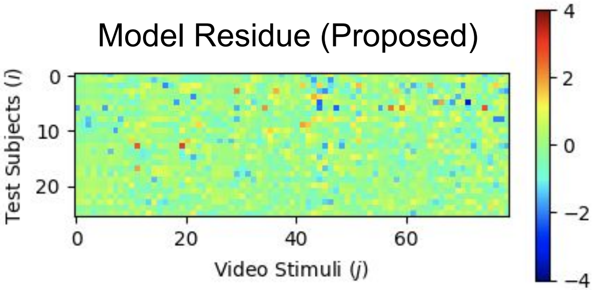
10 random subjects are corrupted, with corruption probability varying from 0.0 to 1.0
Y-axis: RMSE with respect to clean case
SR: subject rejection; BR: bias removal; MOS: mean opinion score; RMSE: root mean squared error

Model Residue Visualization
NFLX Public Dataset (Lab Study)





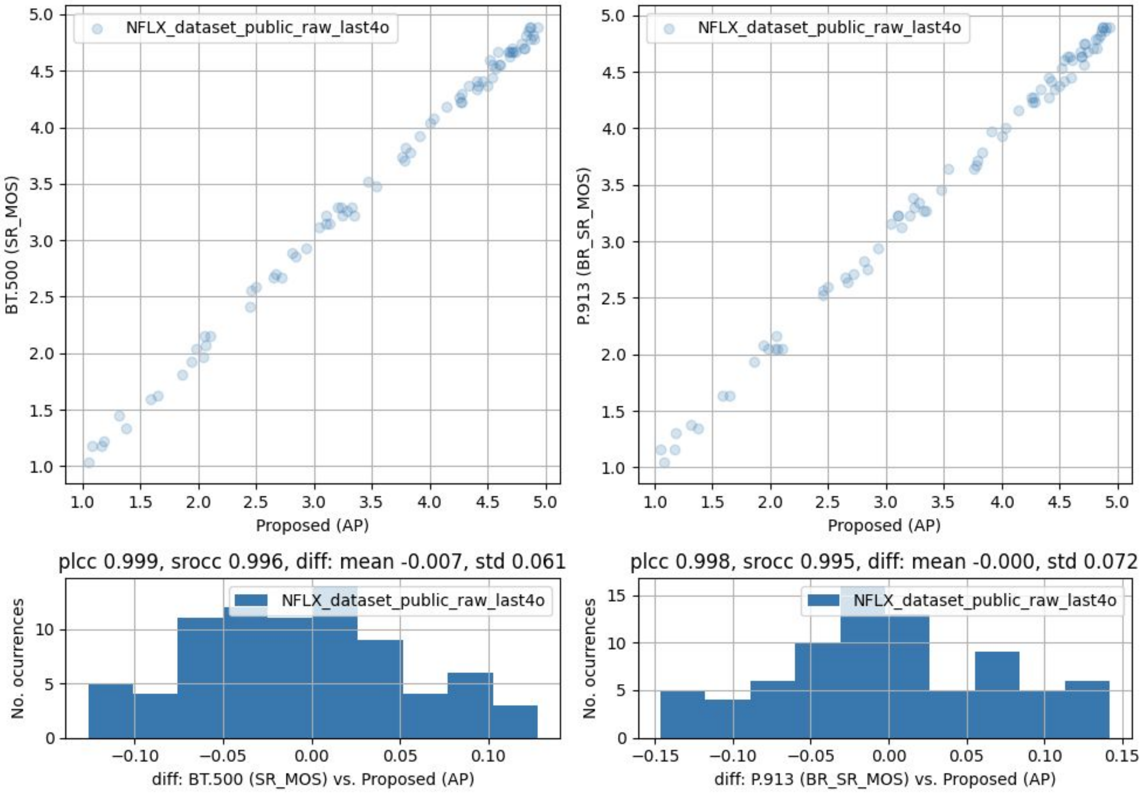




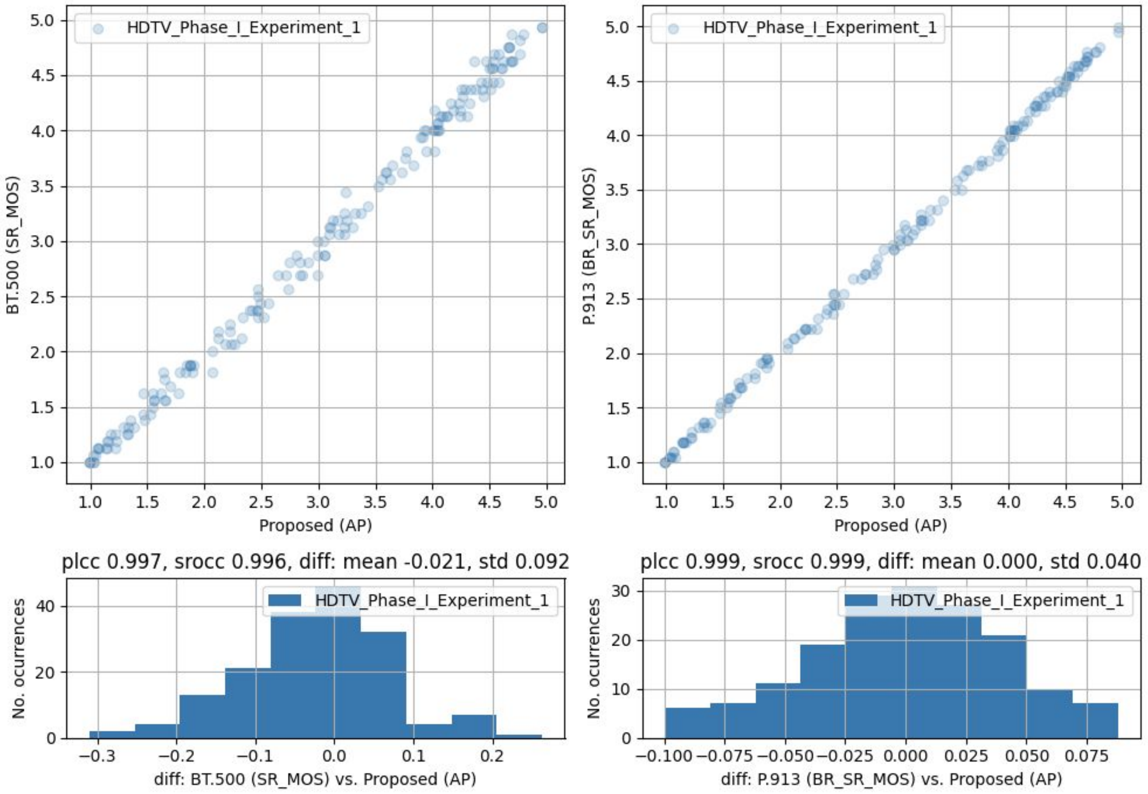
For a good scheme, the residue should look like random noise

Scatter plot: Proposed vs. BT.500/P.913 - More Datasets

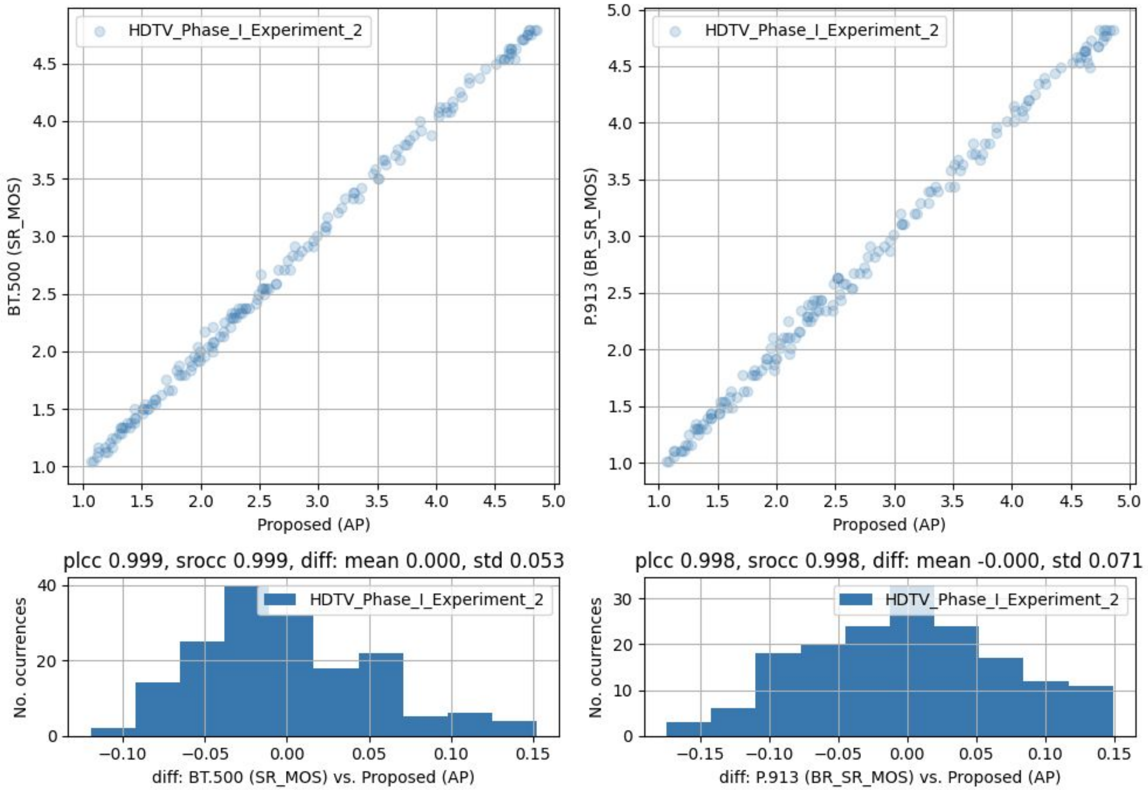
Recovered Quality Score - Proposed vs. BT.500/P.913 NFLX Public Dataset (Lab Study)



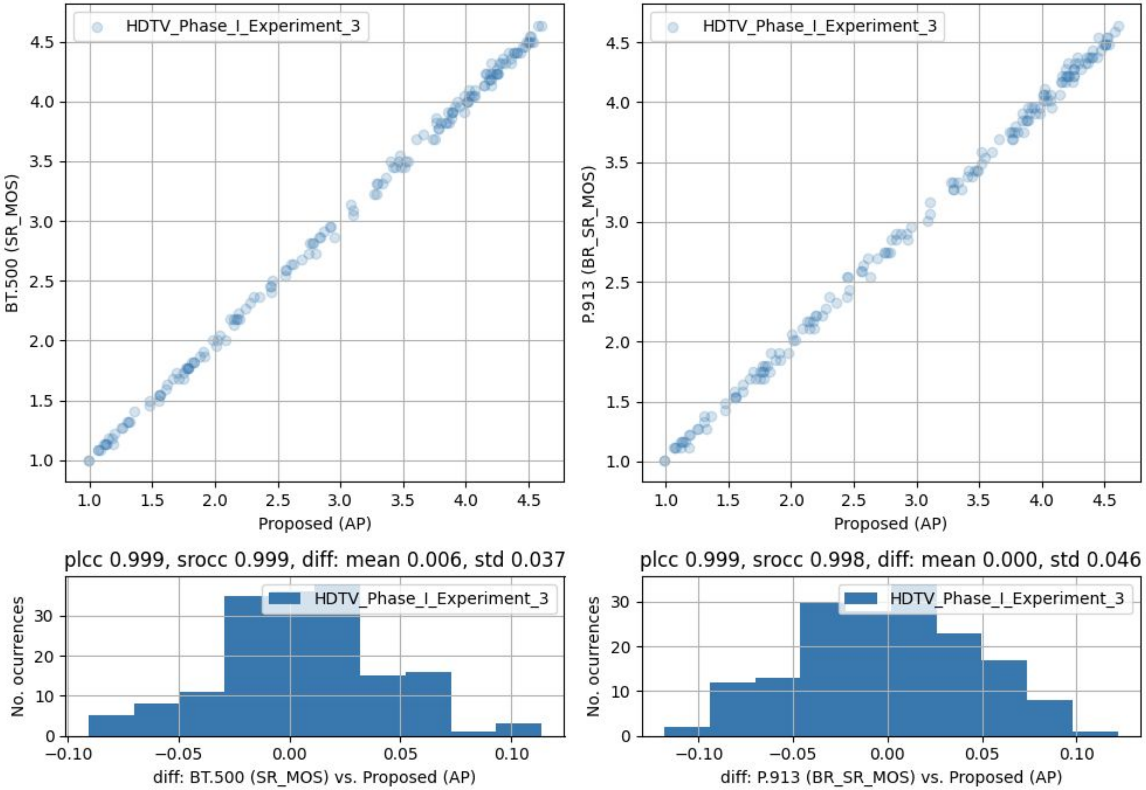
Recovered Quality Score - Proposed vs. BT.500/P.913 VQEG HDTV Phase I Exp 1 (Lab Study)



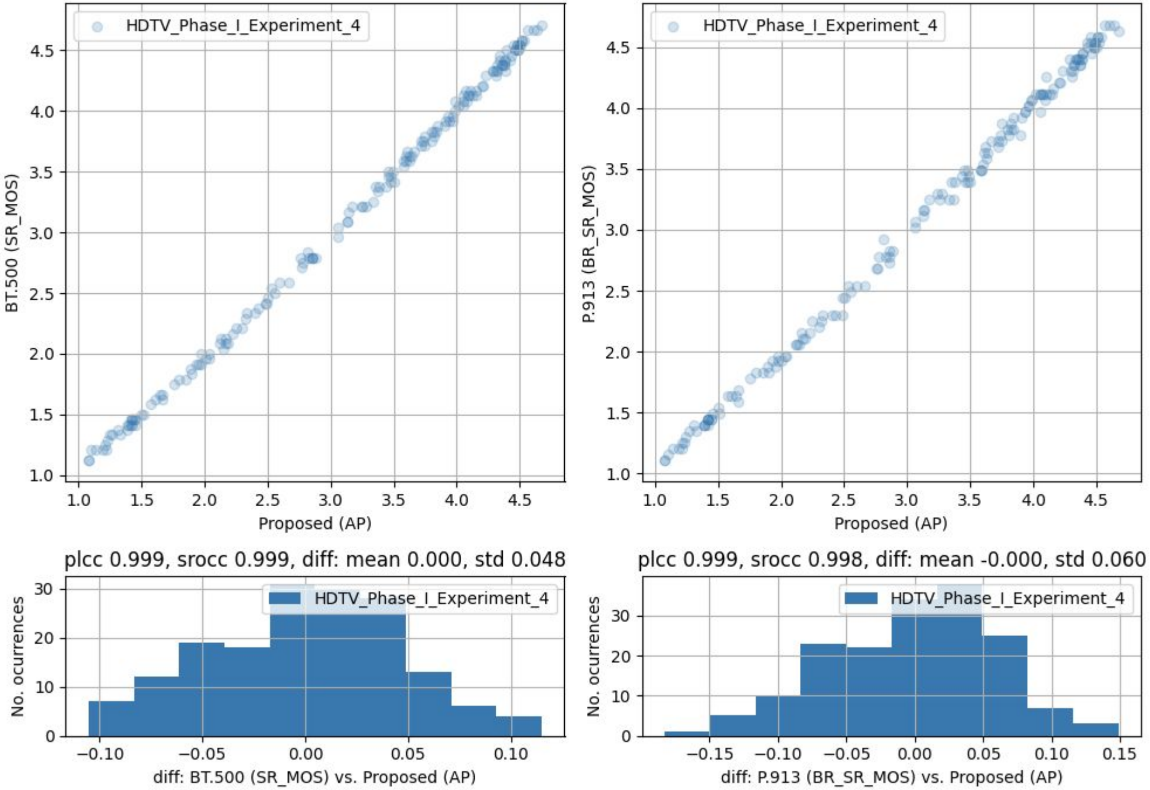
Recovered Quality Score - Proposed vs. BT.500/P.913 VQEG HDTV Phase I Exp 2 (Lab Study)



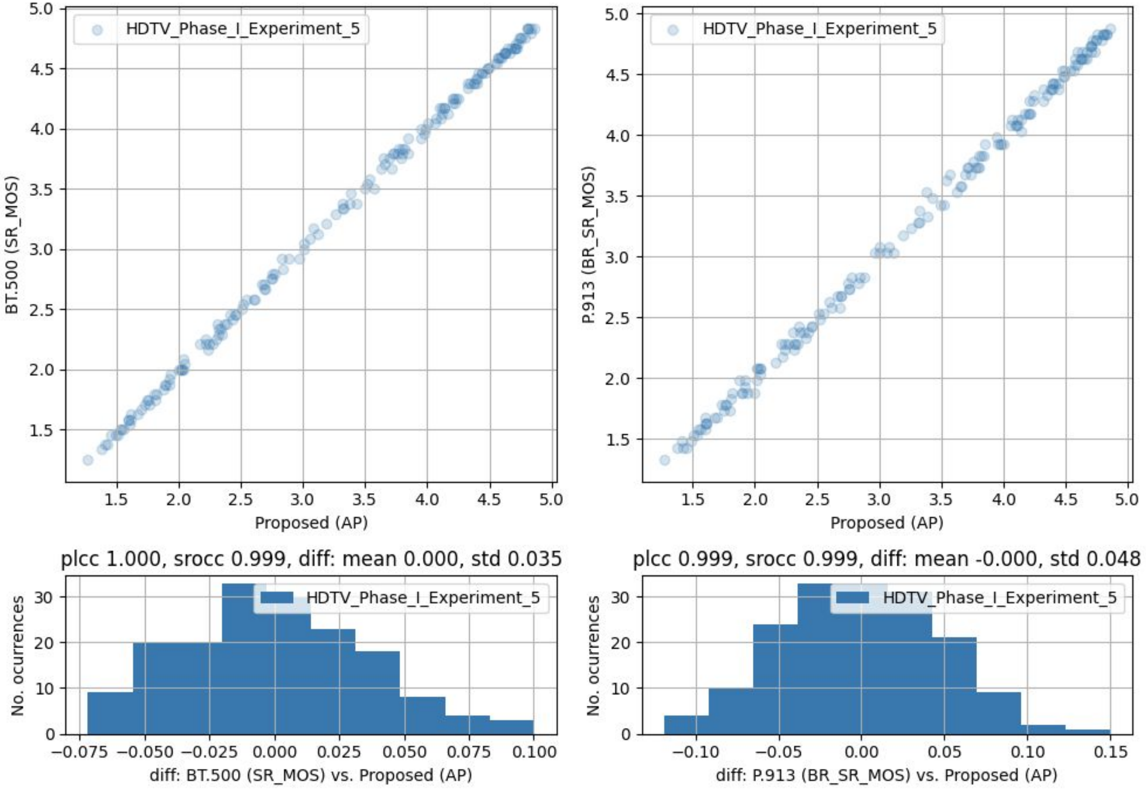
Recovered Quality Score - Proposed vs. BT.500/P.913 VQEG HDTV Phase I Exp 3 (Lab Study)



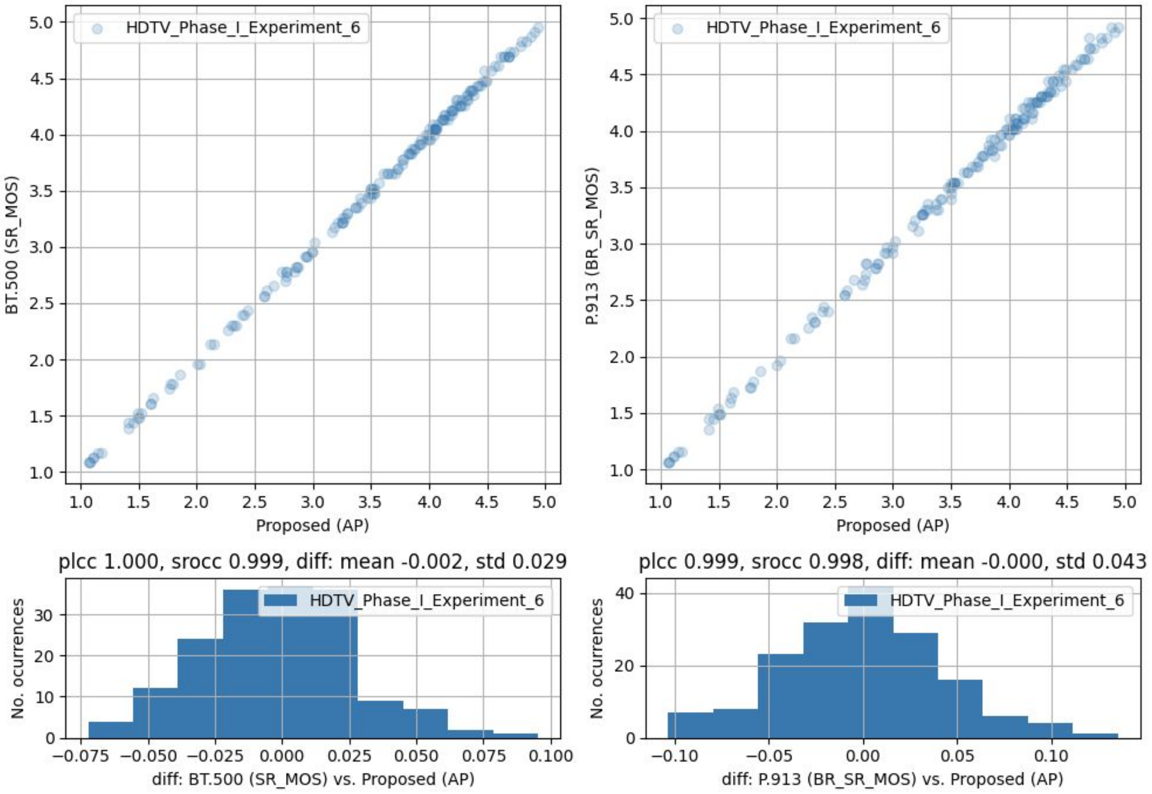
Recovered Quality Score - Proposed vs. BT.500/P.913 VQEG HDTV Phase I Exp 4 (Lab Study)



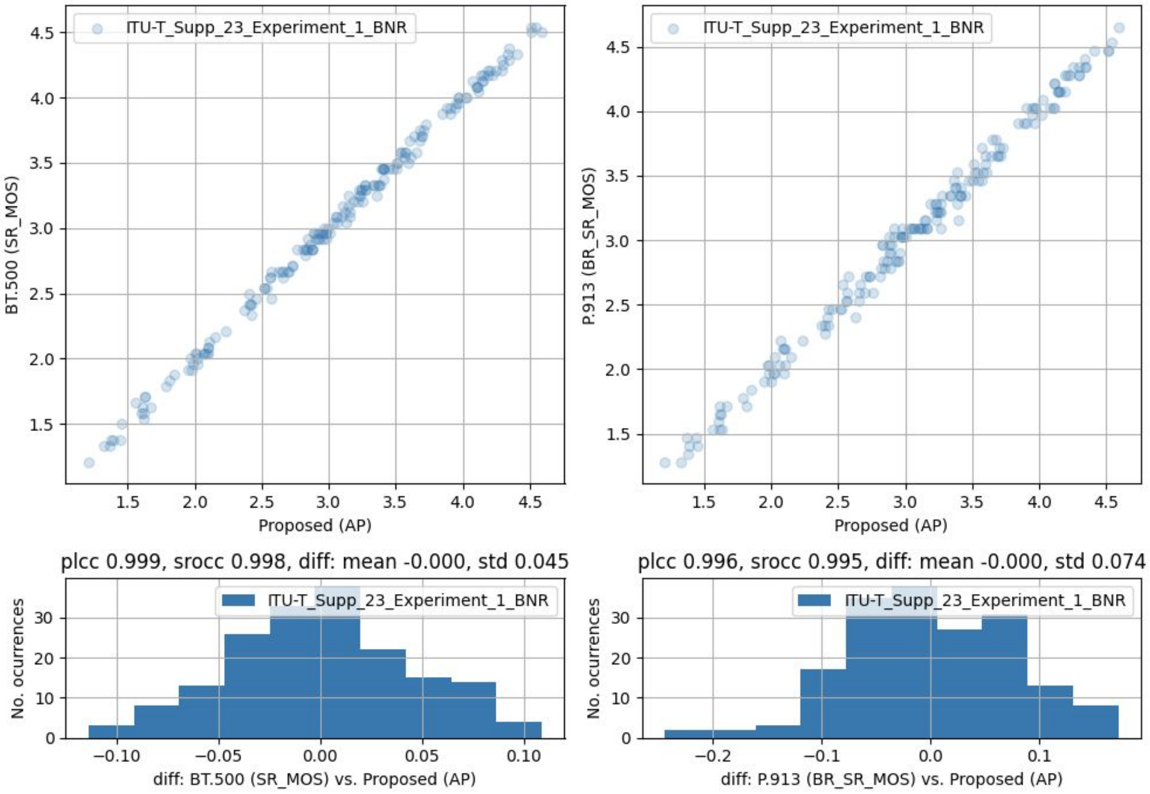
Recovered Quality Score - Proposed vs. BT.500/P.913 VQEG HDTV Phase I Exp 5 (Lab Study)



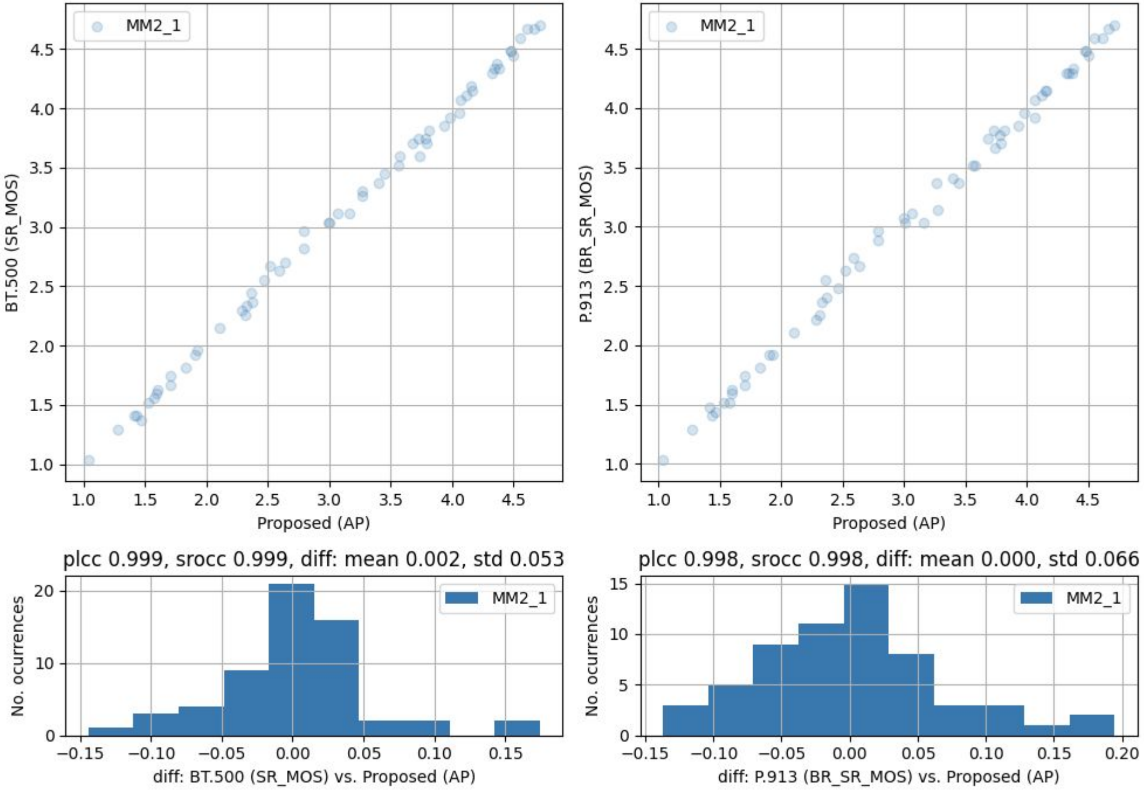
Recovered Quality Score - Proposed vs. BT.500/P.913 VQEG HDTV Phase I Exp 6 (Lab Study)



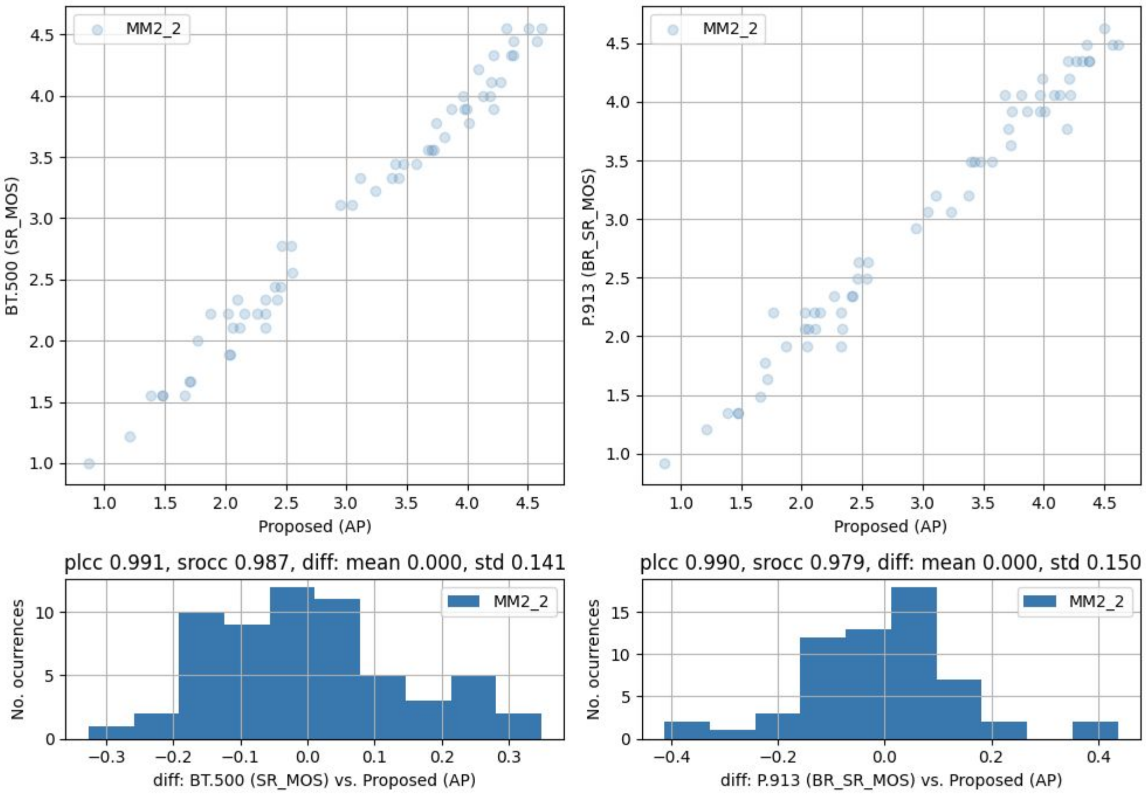
Recovered Quality Score - Proposed vs. BT.500/P.913 ITU-T Supp 23 Experiment 1 BNR - Audio (Lab Study)



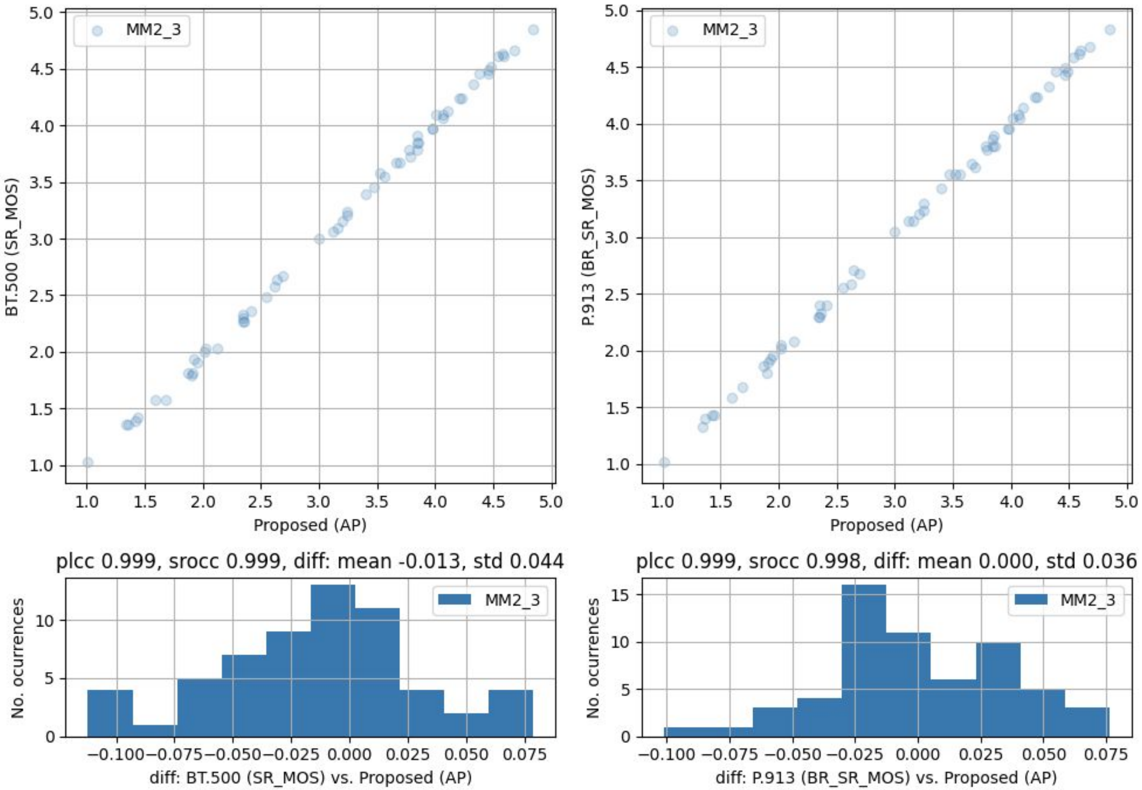
Recovered Quality Score - Proposed vs. BT.500/P.913 MM2 1 (Lab Study)



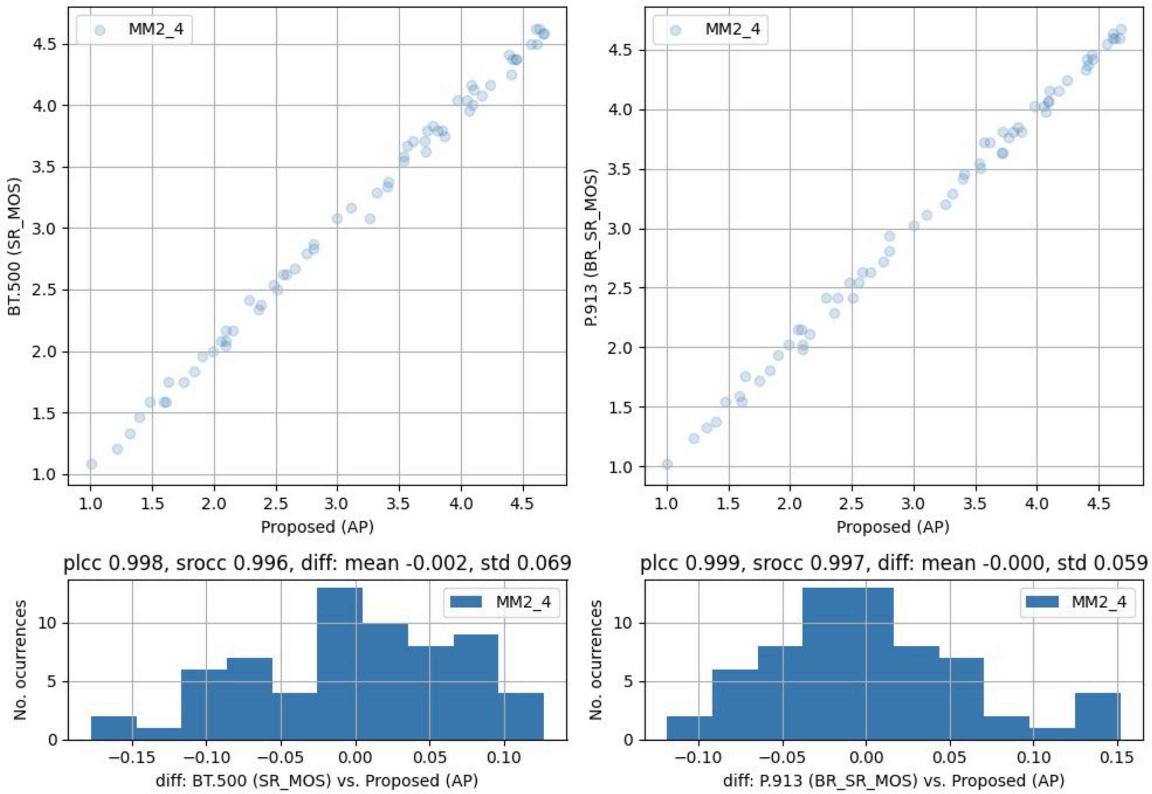
Recovered Quality Score - Proposed vs. BT.500/P.913 MM2 2 (Lab Study)



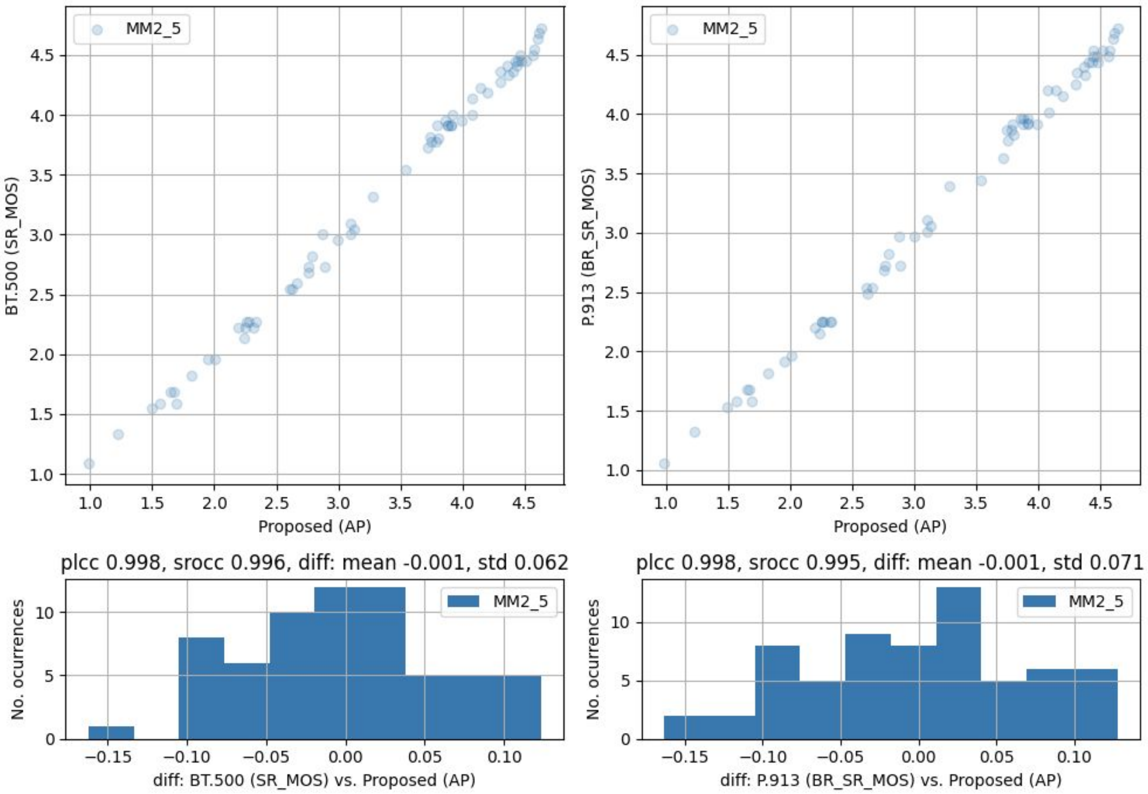
Recovered Quality Score - Proposed vs. BT.500/P.913 MM2 3 (Lab Study)



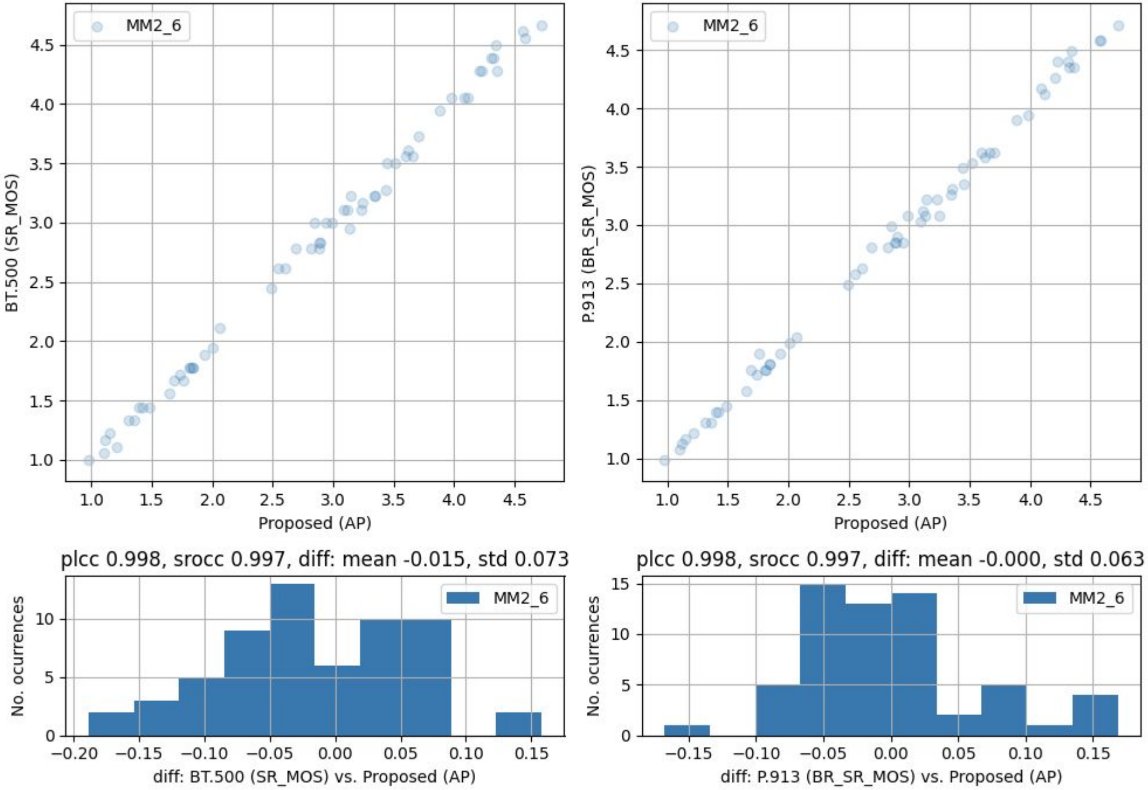
Recovered Quality Score - Proposed vs. BT.500/P.913 MM2 4 (Lab Study)



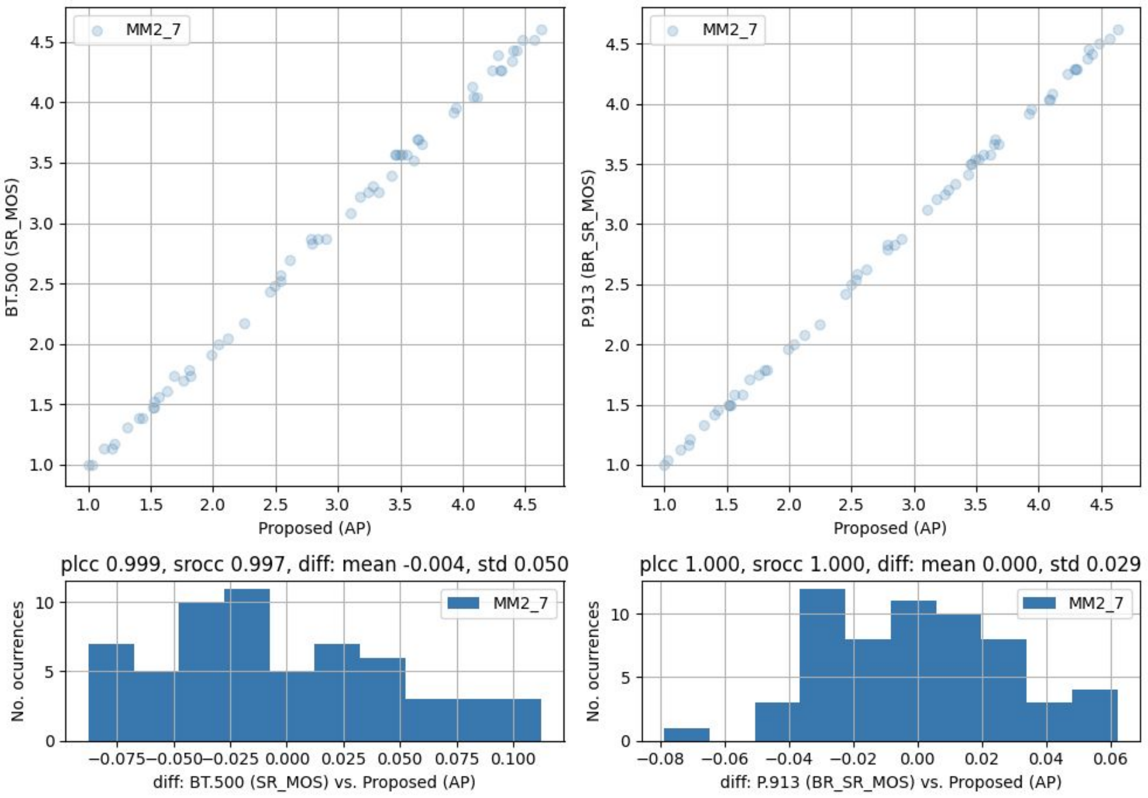
Recovered Quality Score - Proposed vs. BT.500/P.913 MM2 5 (Lab Study)



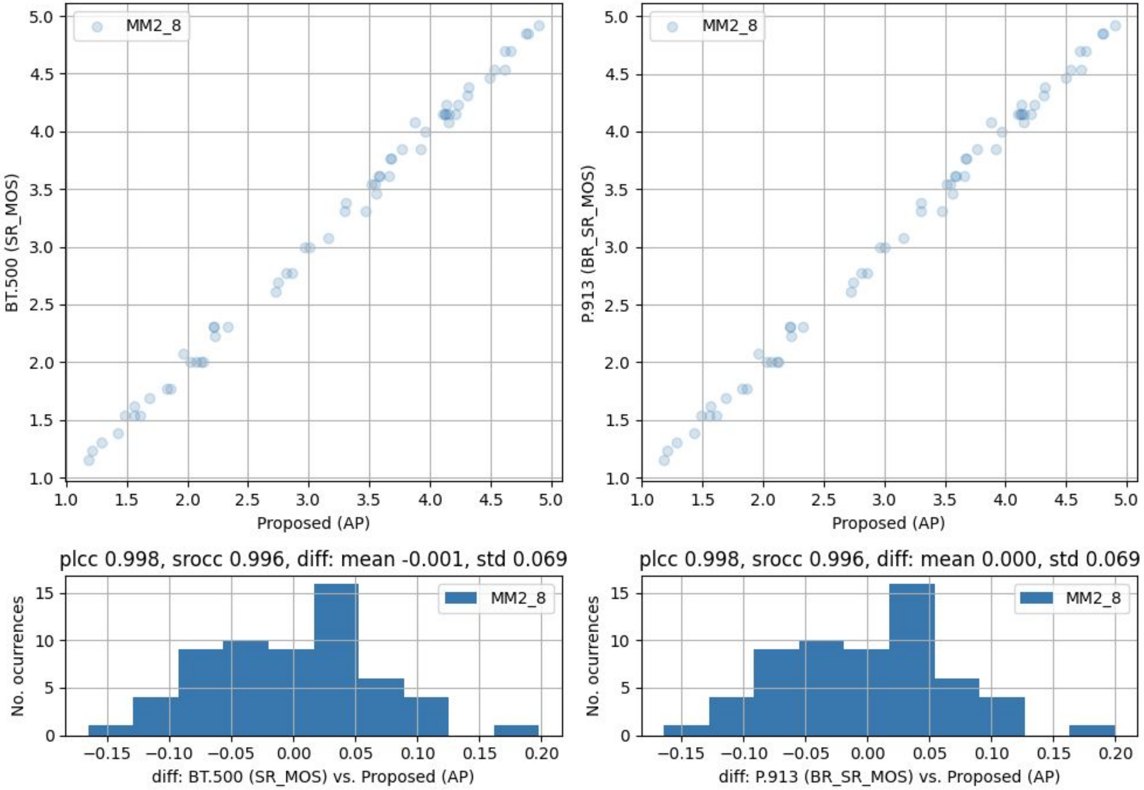
Recovered Quality Score - Proposed vs. BT.500/P.913 MM2 6 (Lab Study)



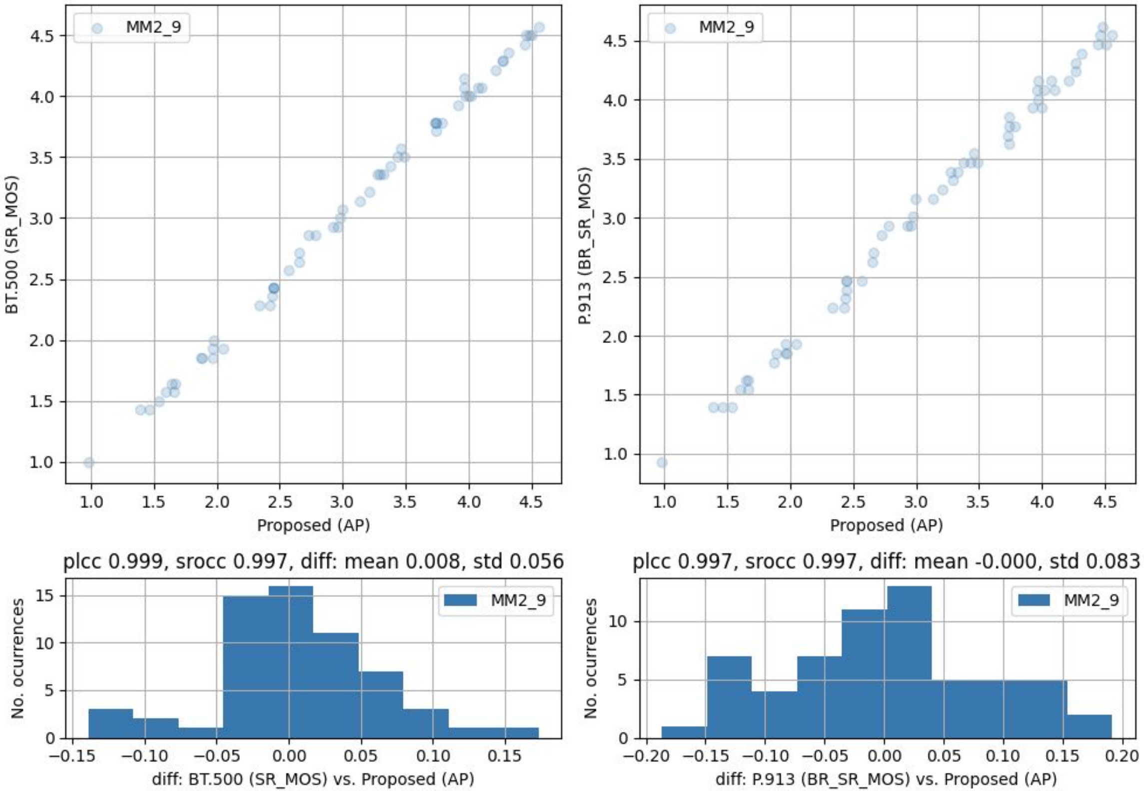
Recovered Quality Score - Proposed vs. BT.500/P.913 MM2 7 (Lab Study)



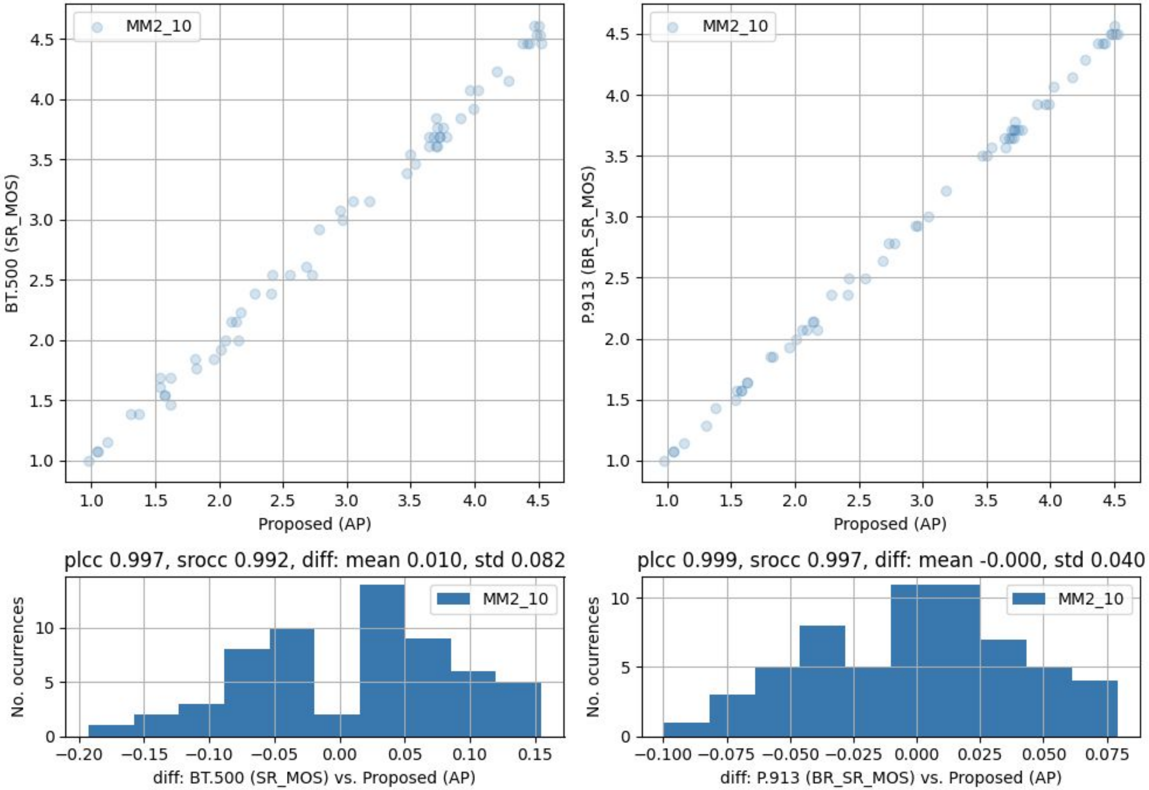
Recovered Quality Score - Proposed vs. BT.500/P.913 MM2 8 (Lab Study)



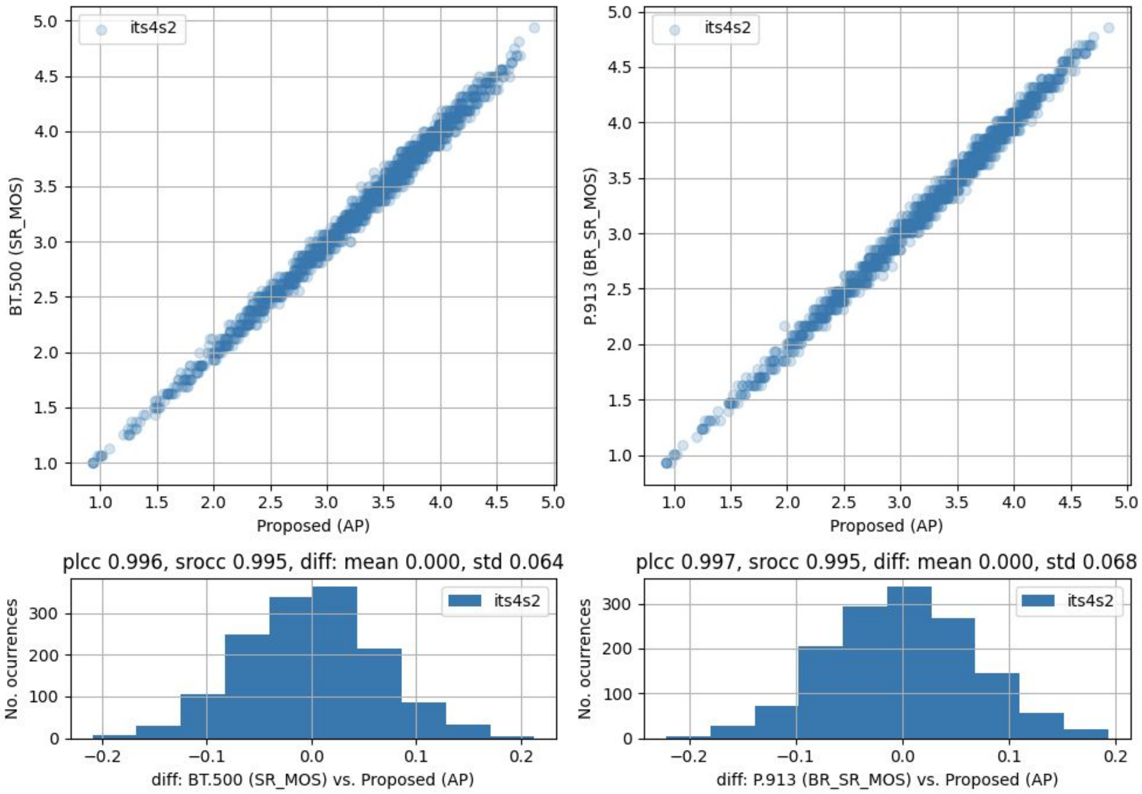
Recovered Quality Score - Proposed vs. BT.500/P.913 MM2 9 (Lab Study)



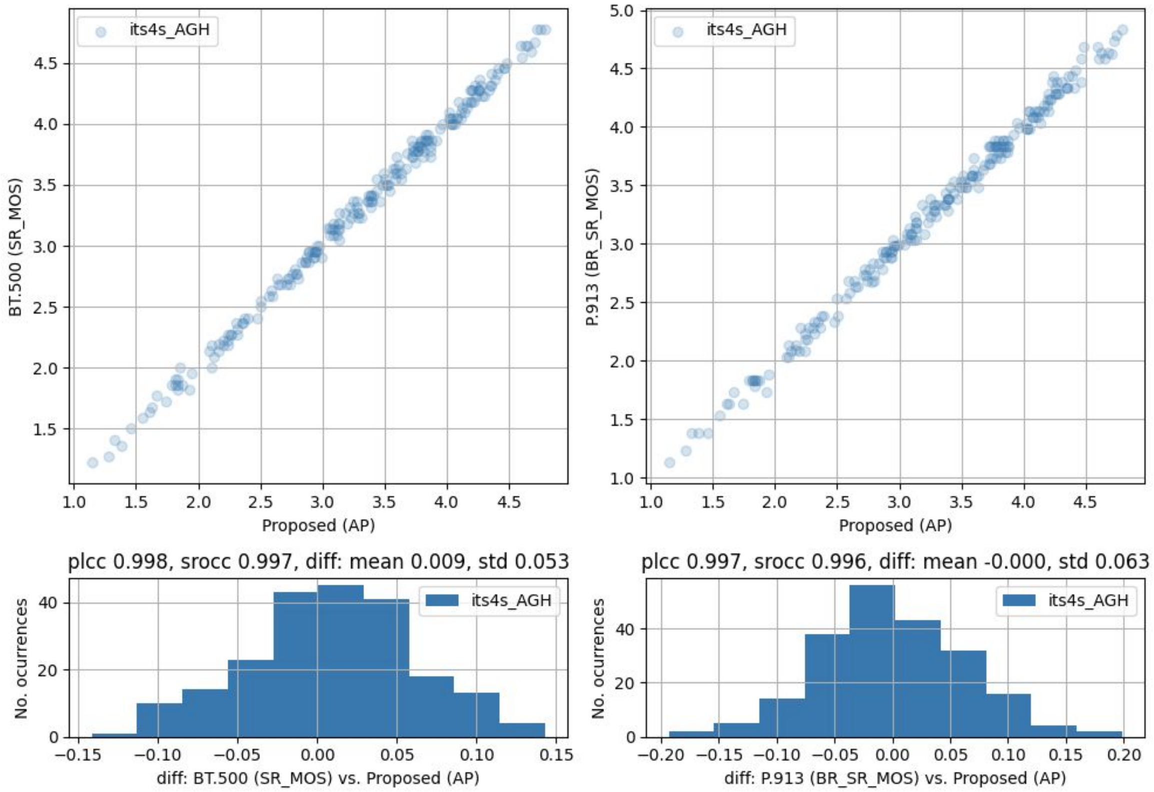
Recovered Quality Score - Proposed vs. BT.500/P.913 MM2 10 (Lab Study)



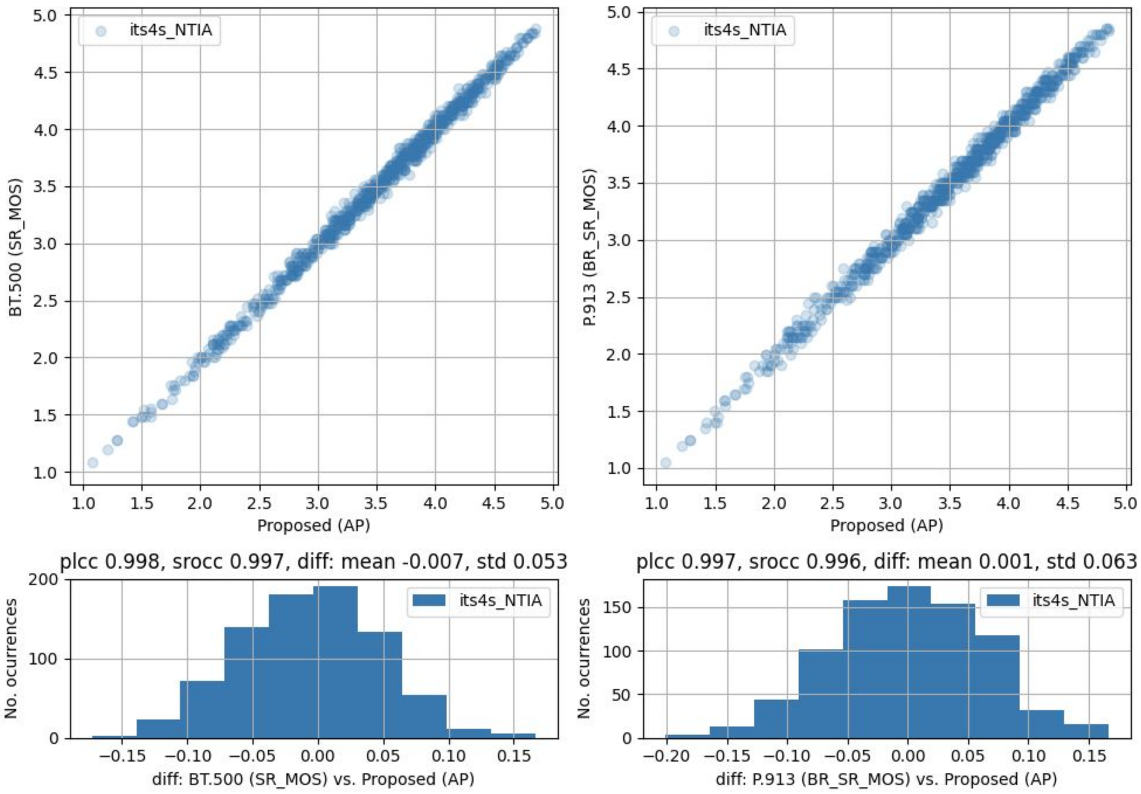
Recovered Quality Score - Proposed vs. BT.500/P.913 its4s2 (Lab Study)



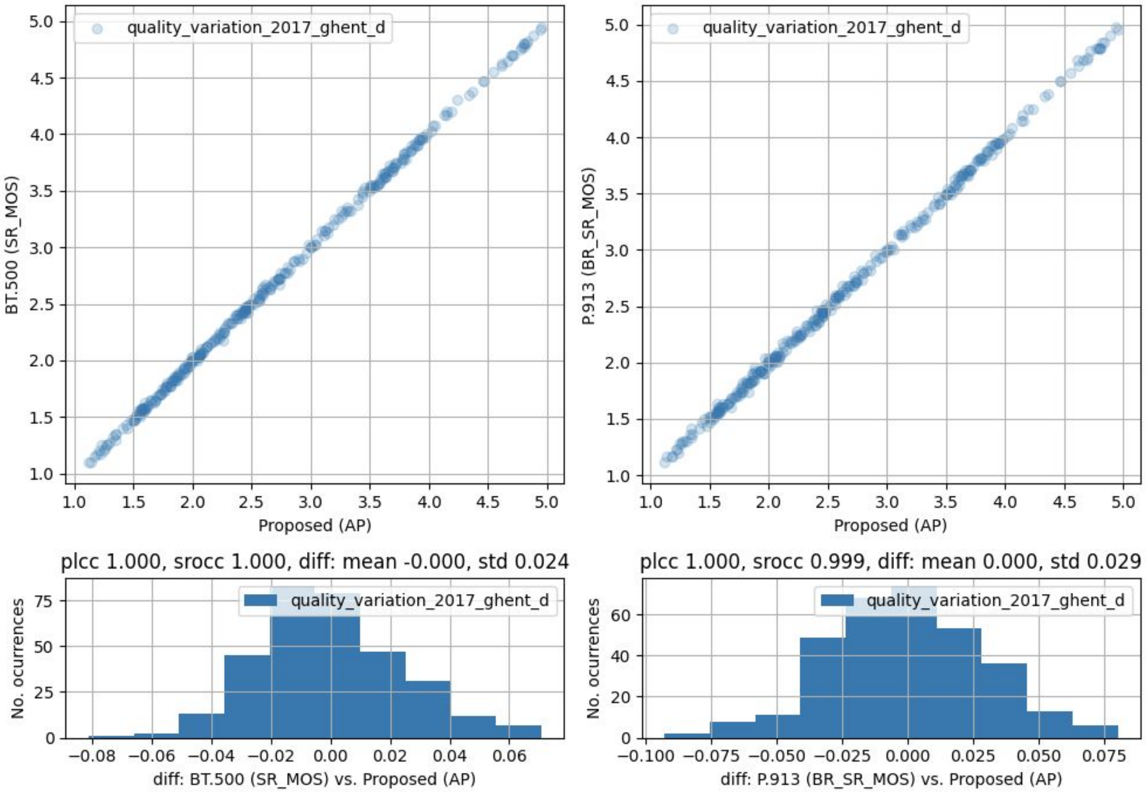
Recovered Quality Score - Proposed vs. BT.500/P.913 its4s2 AGH (Lab Study)



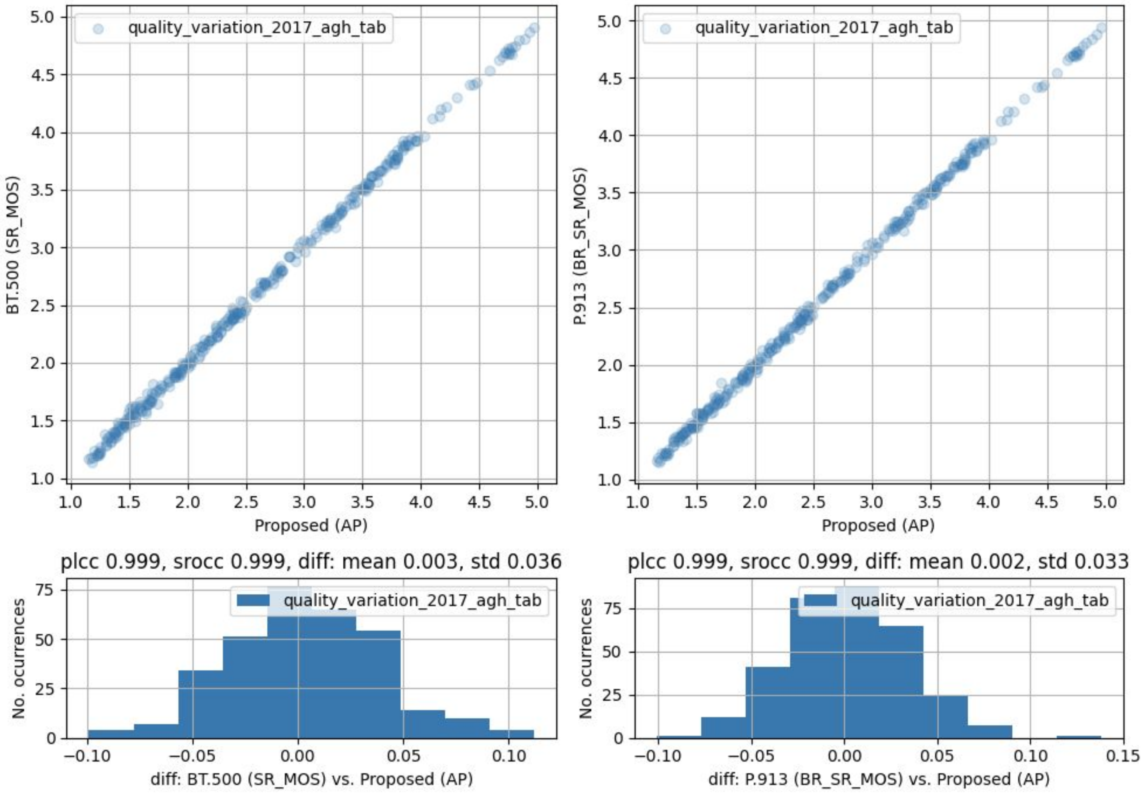
Recovered Quality Score - Proposed vs. BT.500/P.913 its4s2 AGH NTIA (Lab Study)



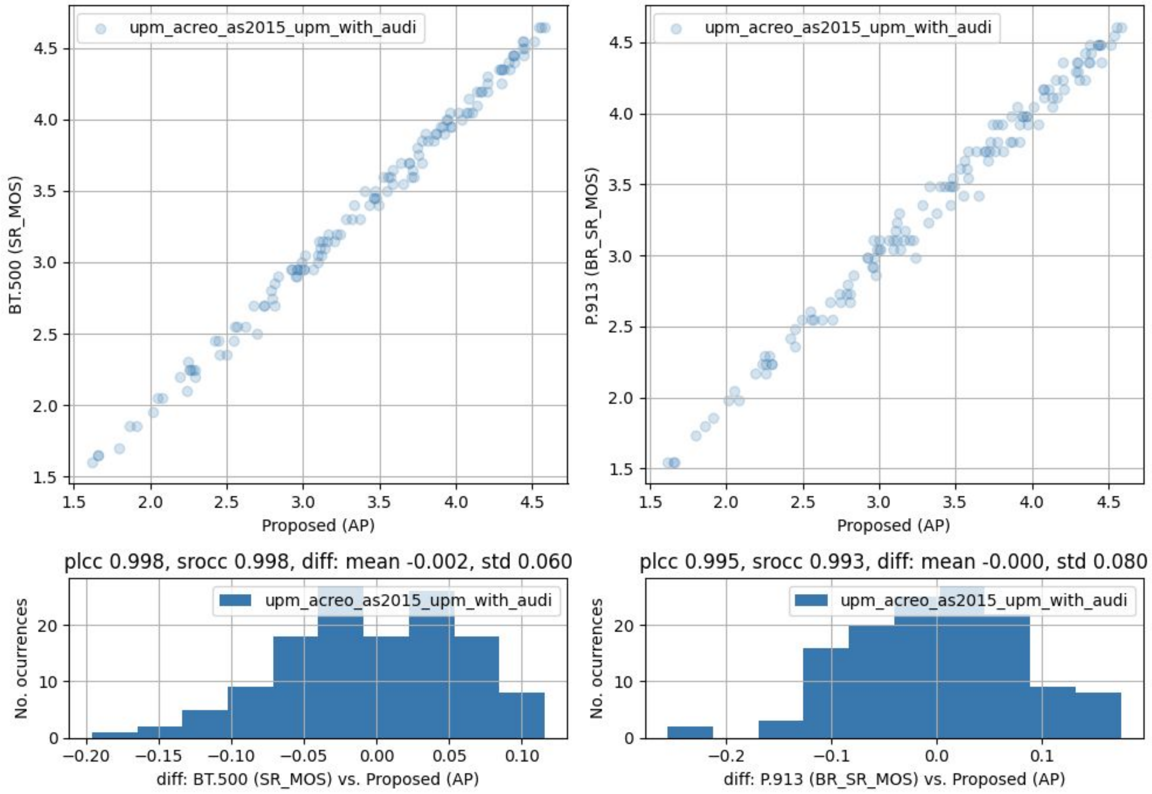
Recovered Quality Score - Proposed vs. BT.500/P.913 Quality Variation 2017 Ghent Dataset (Lab Study)



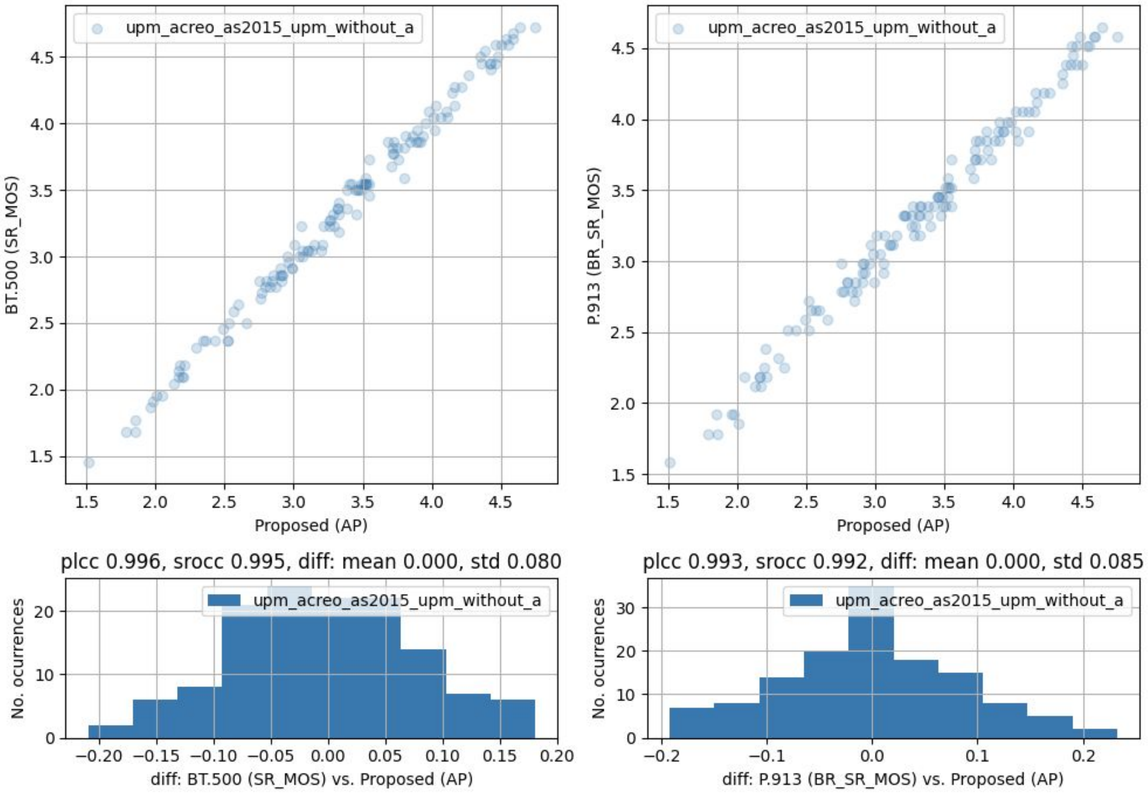
Recovered Quality Score - Proposed vs. BT.500/P.913 Quality Variation 2017 AGH Tablet Dataset (Lab Study)



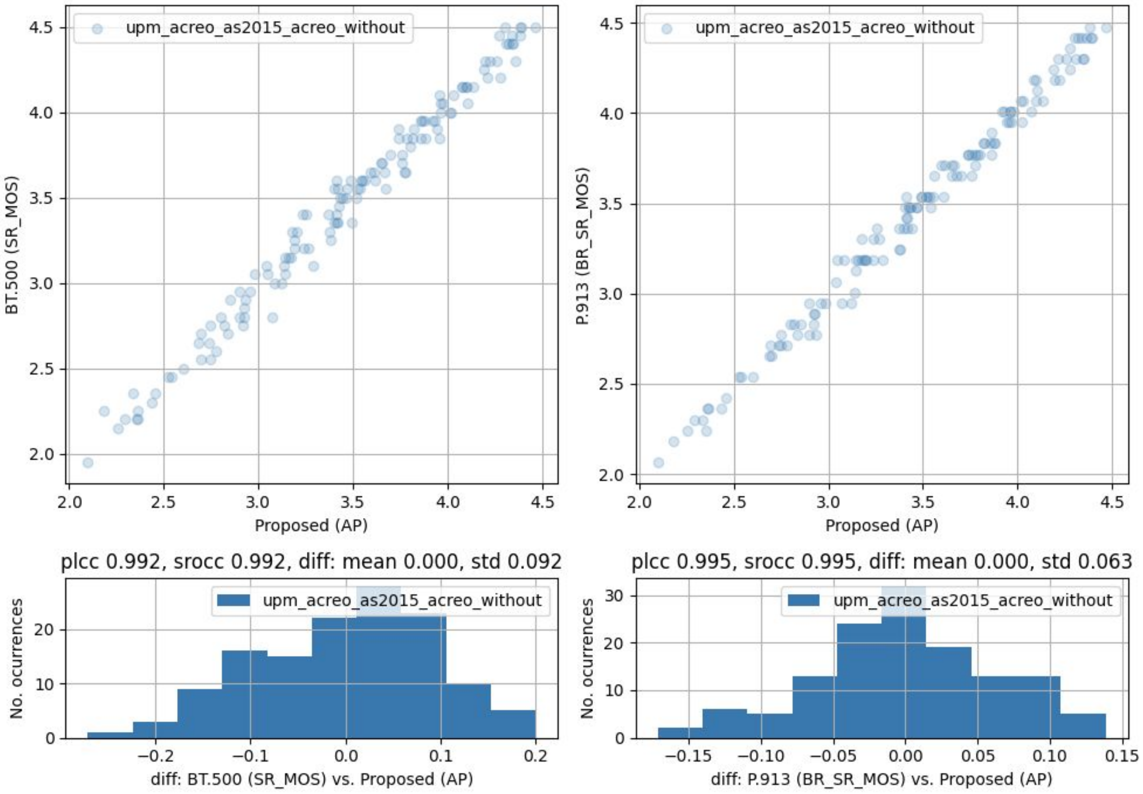
Recovered Quality Score - Proposed vs. BT.500/P.913 AS2015 UPM w/ audio (Lab Study)



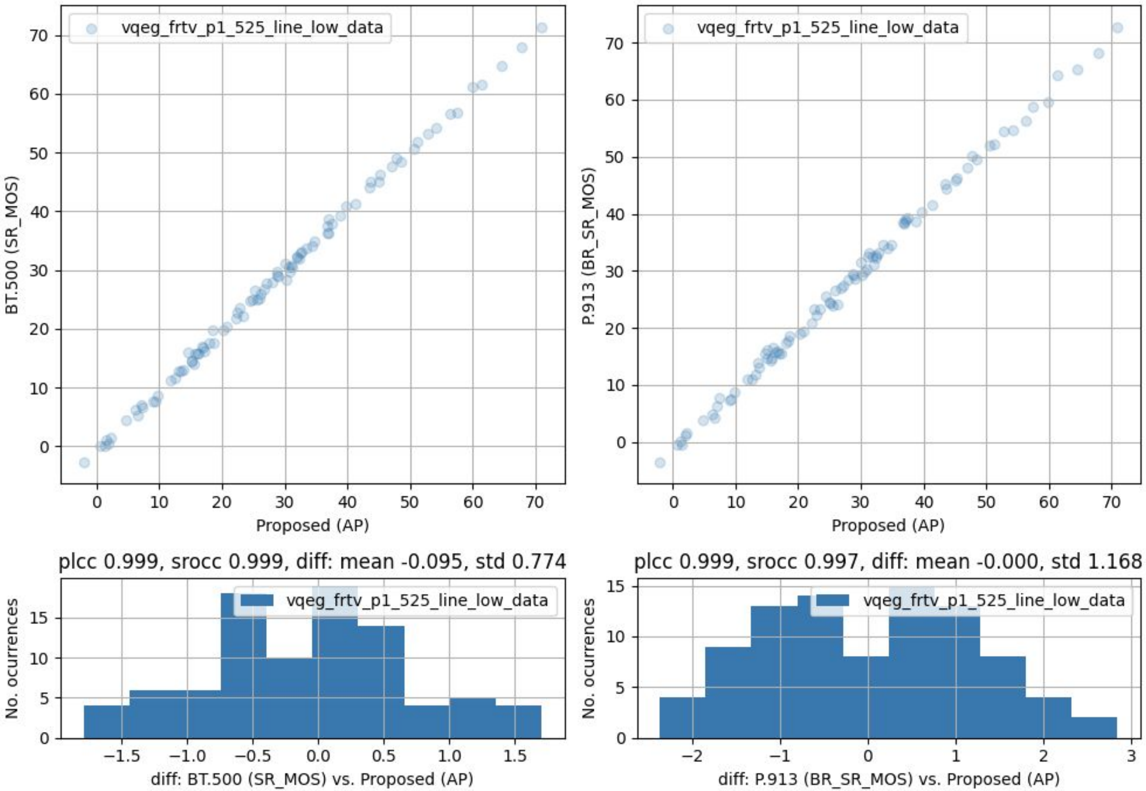
Recovered Quality Score - Proposed vs. BT.500/P.913 AS2015 UPM w/o audio (Lab Study)



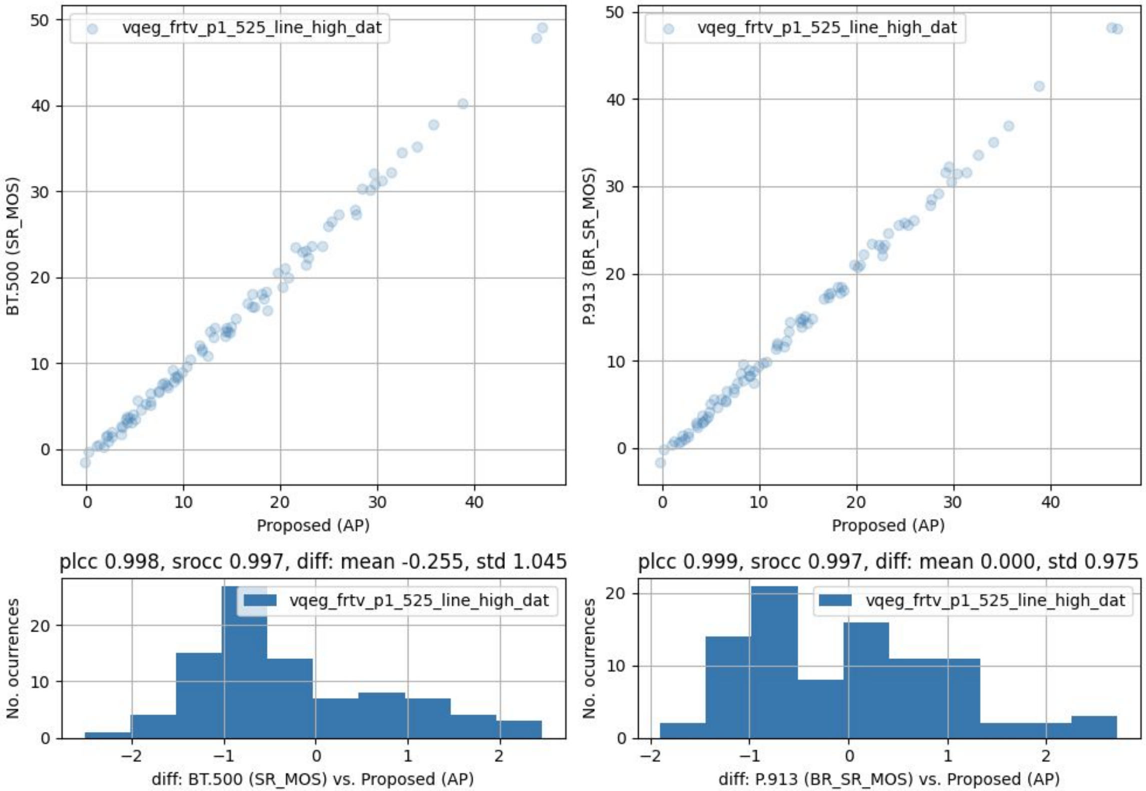
Recovered Quality Score - Proposed vs. BT.500/P.913 AS2015 ACREO w/o audio (Lab Study)



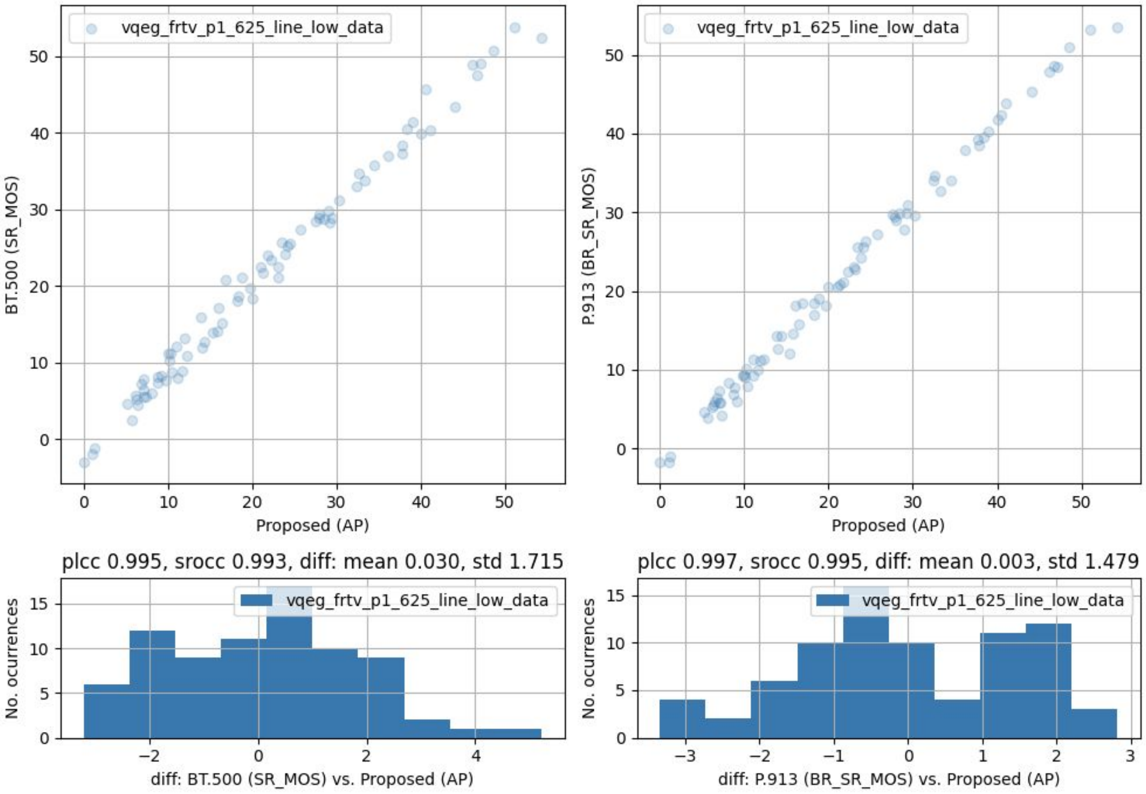
Recovered Quality Score - Proposed vs. BT.500/P.913 VQEG FRTV Phase I 525 line low (Lab Study)



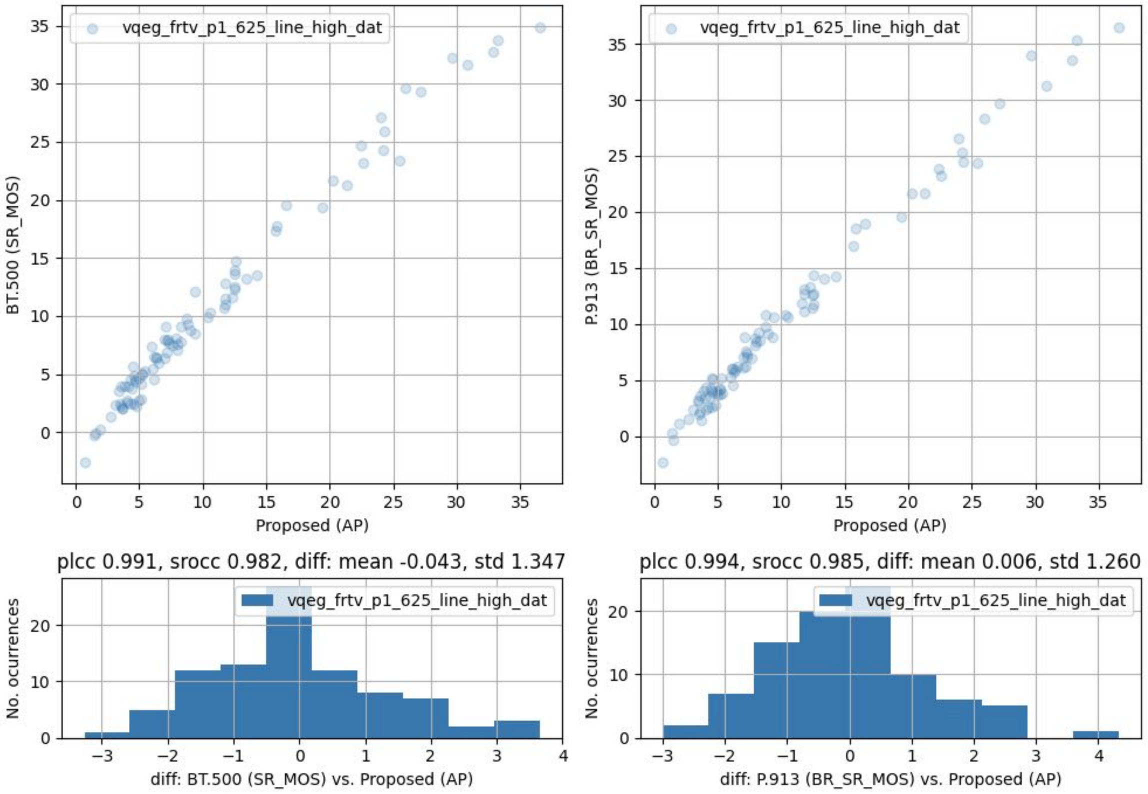
Recovered Quality Score - Proposed vs. BT.500/P.913 VQEG FRTV Phase I 525 line high (Lab Study)



Recovered Quality Score - Proposed vs. BT.500/P.913 VQEG FRTV Phase I 625 line low (Lab Study)

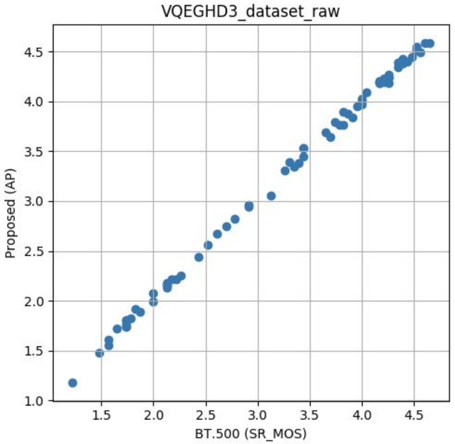


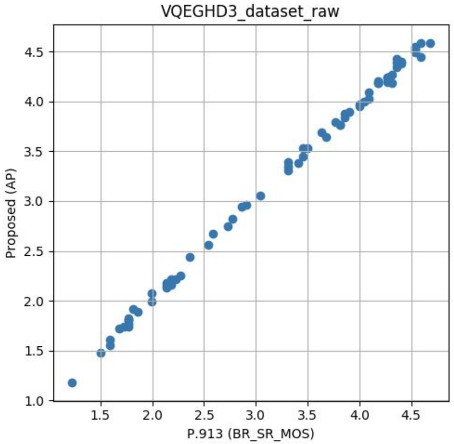
Recovered Quality Score - Proposed vs. BT.500/P.913 VQEG FRTV Phase I 625 line high (Lab Study)

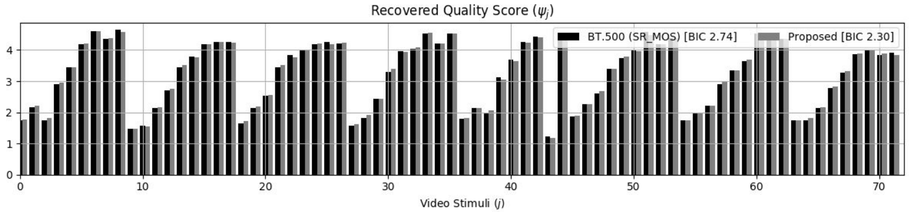


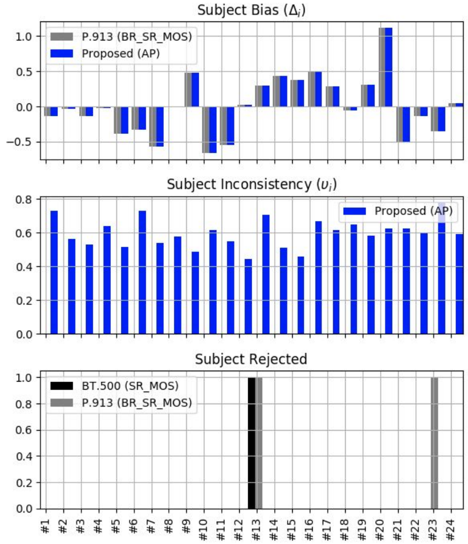
Recovered Result by the Proposed Method - More Datasets

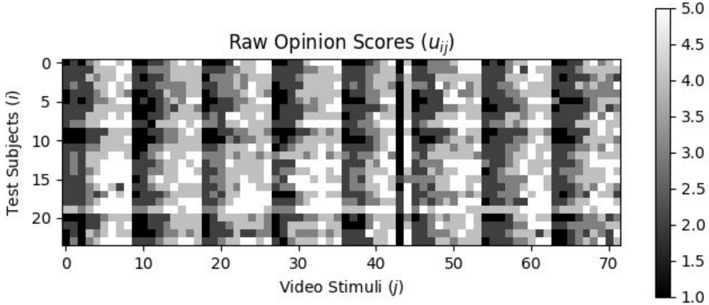
VQEGHD3_dataset_raw



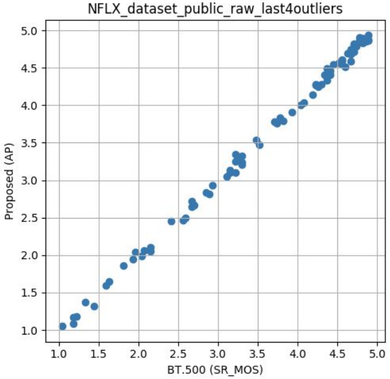


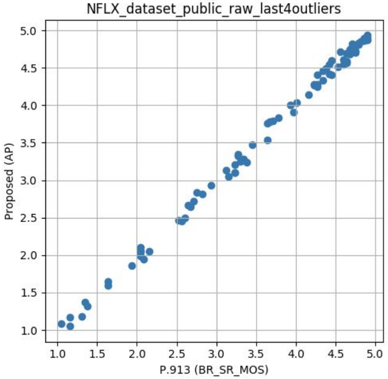


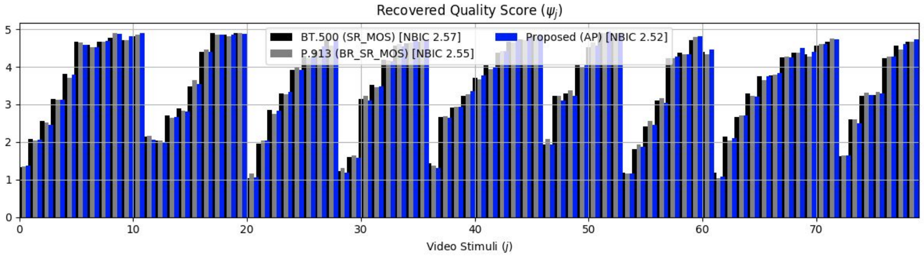


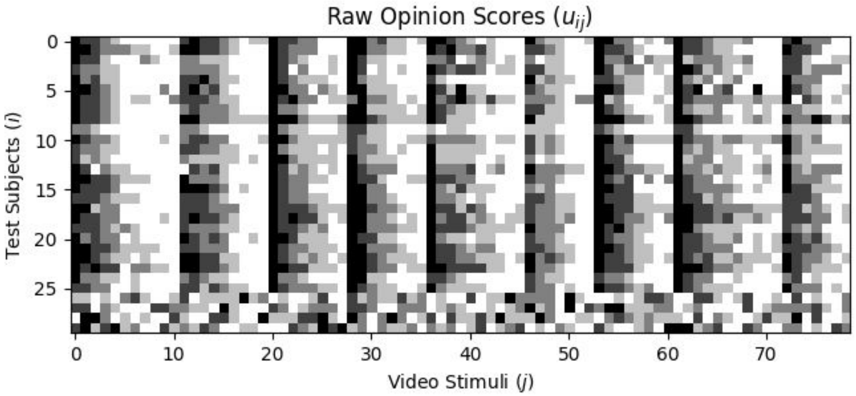


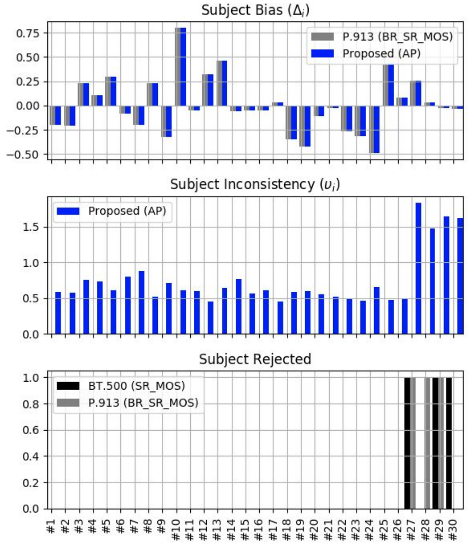
NFLX_public_last4outliers



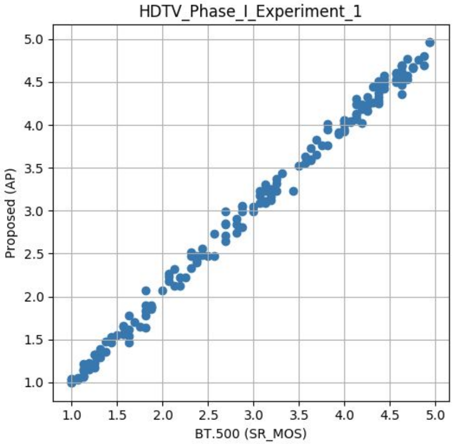


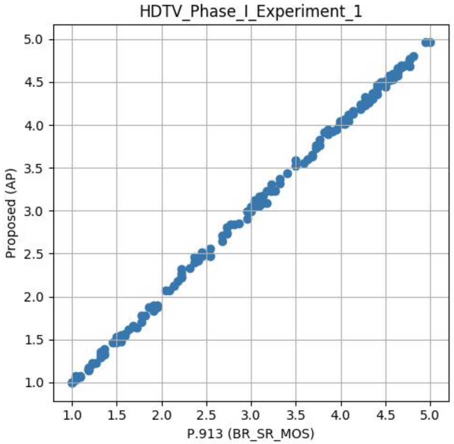


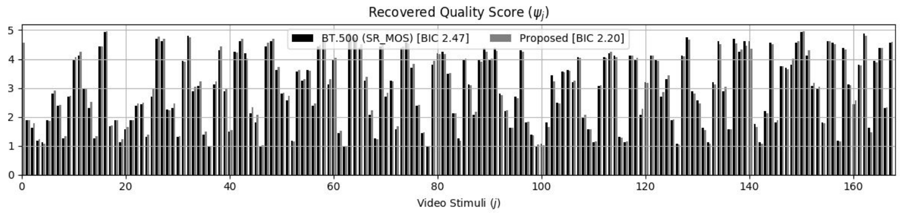


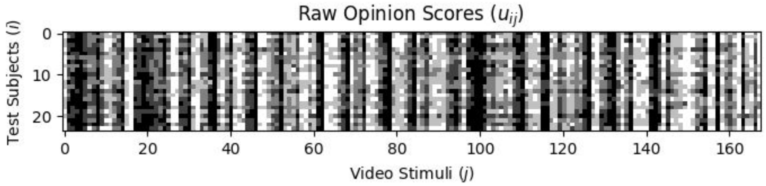


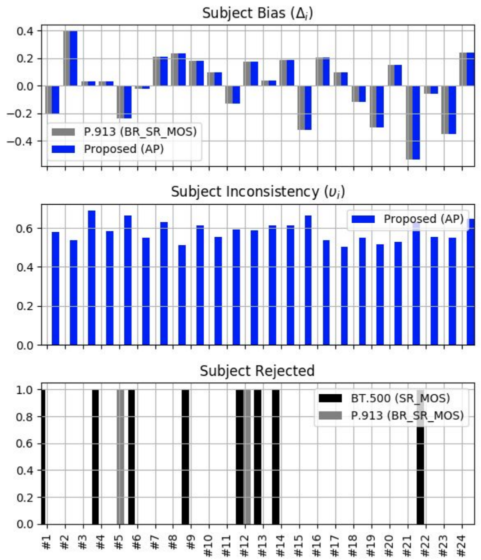
HDTV_Phase_I_Experiment_1



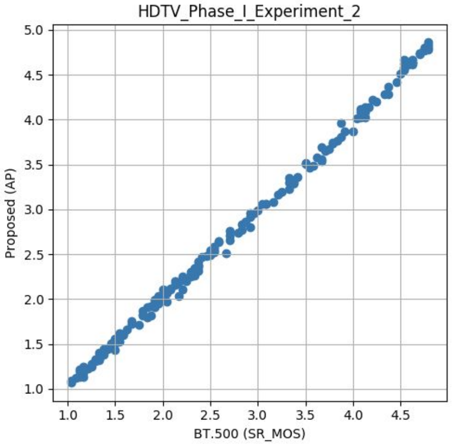


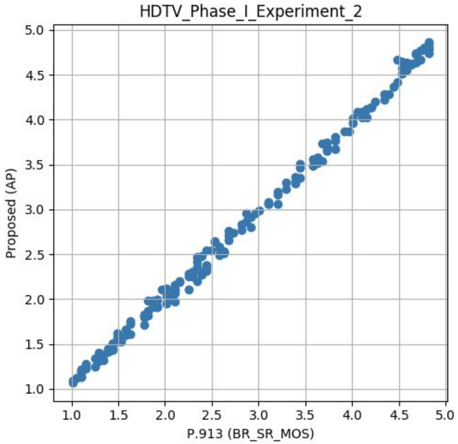


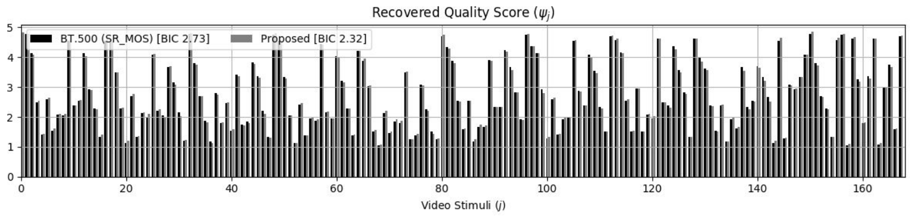


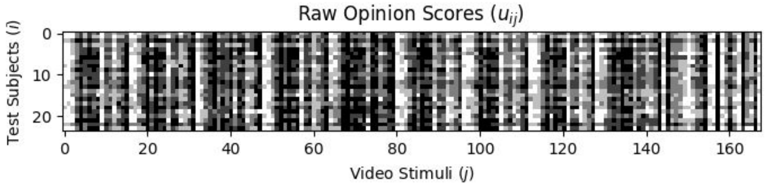


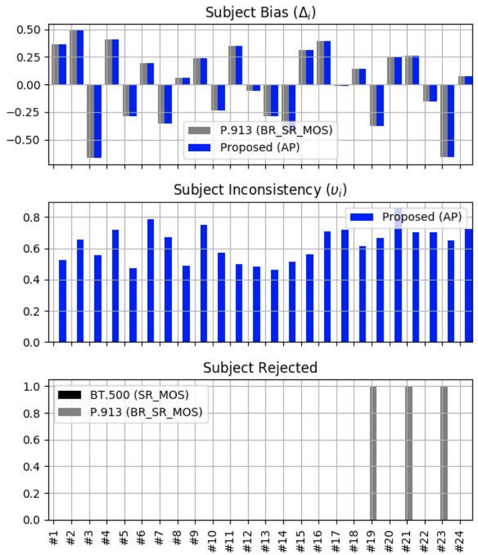
HDTV_Phase_I_Experiment_2



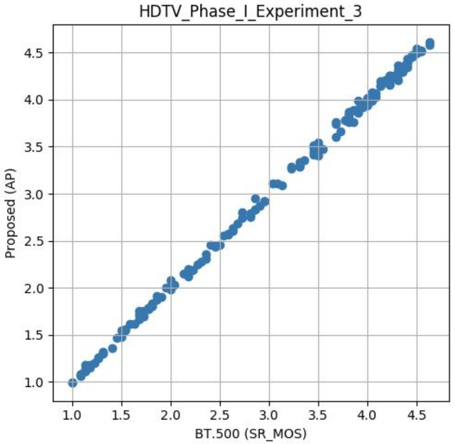


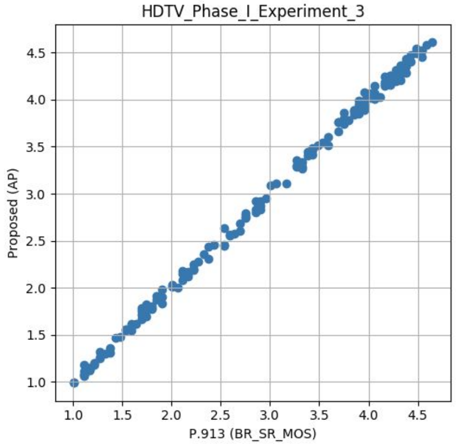


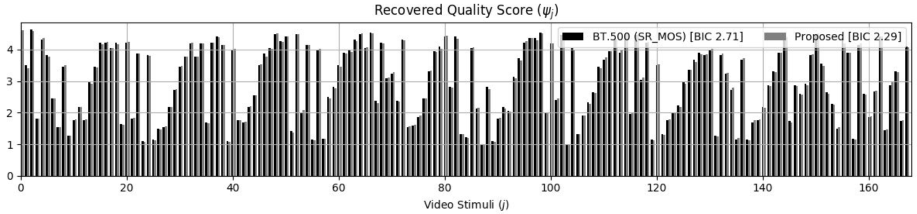


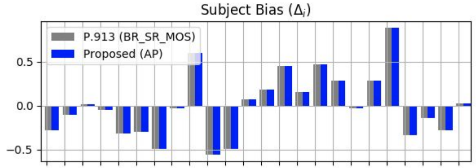


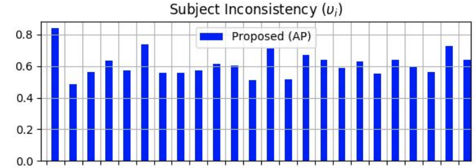
HDTV_Phase_I_Experiment_3

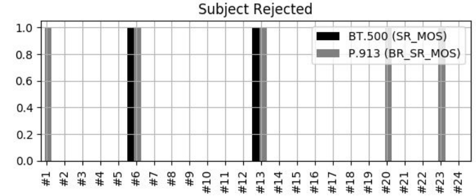


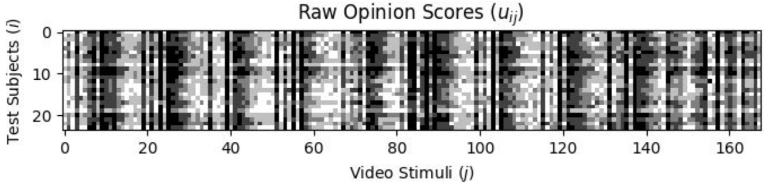




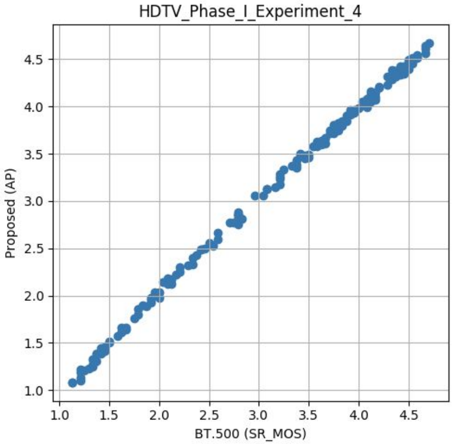


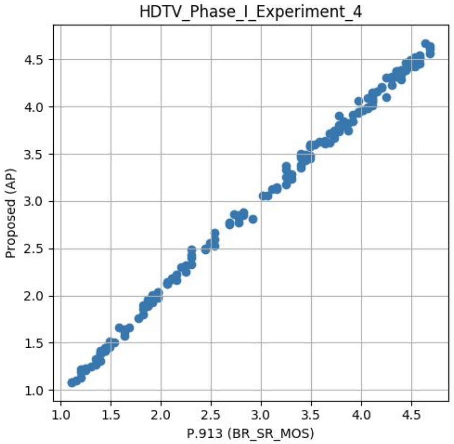


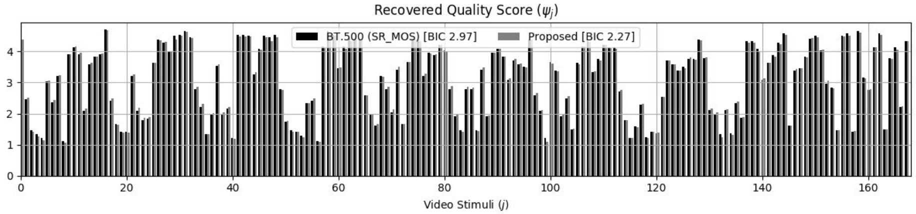


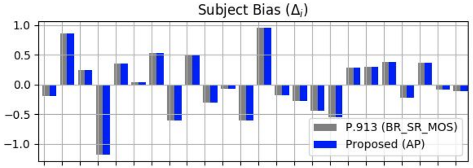


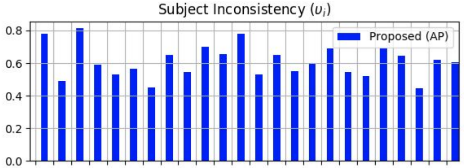
HDTV_Phase_I_Experiment_4

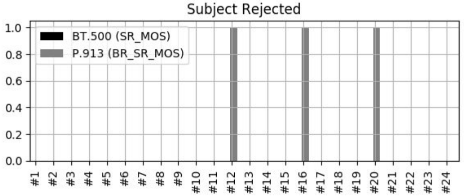


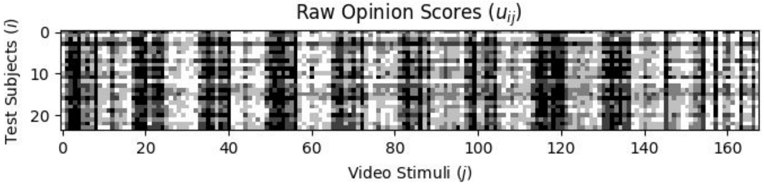




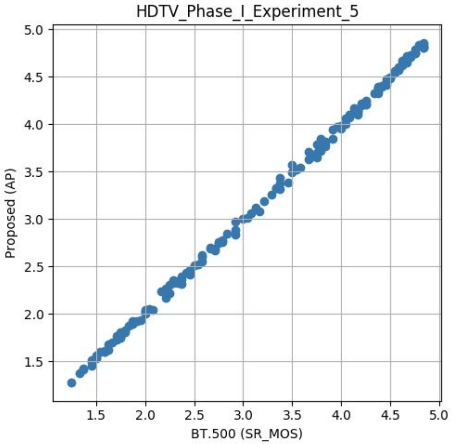


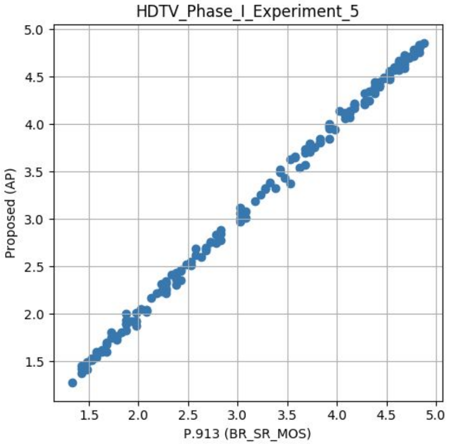


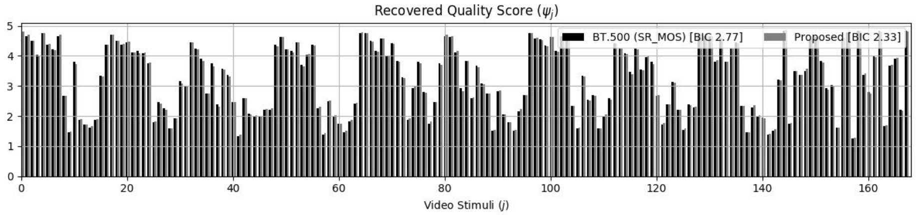


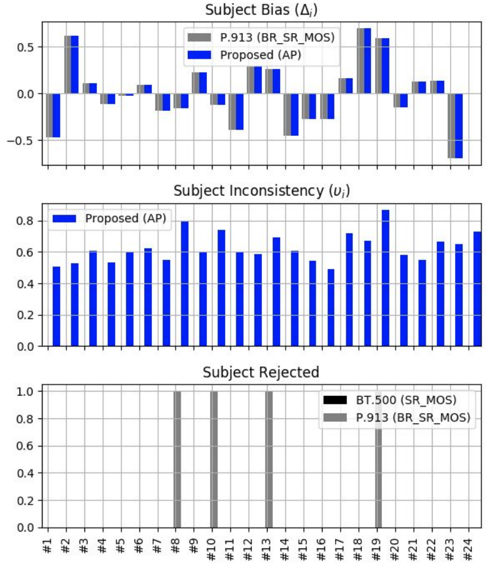


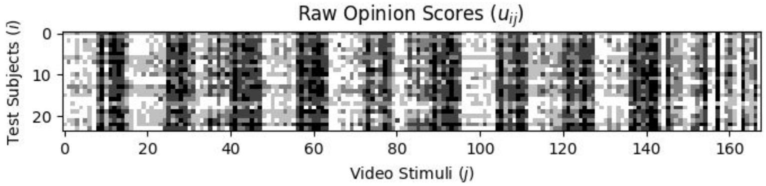
HDTV_Phase_I_Experiment_5



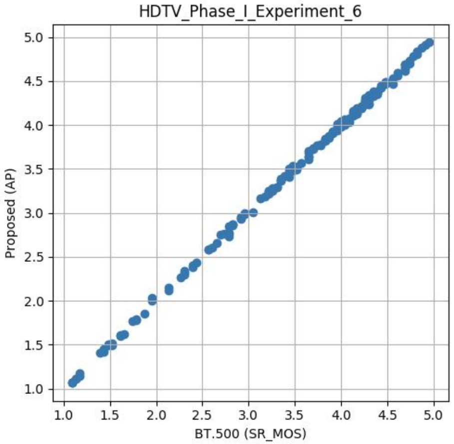


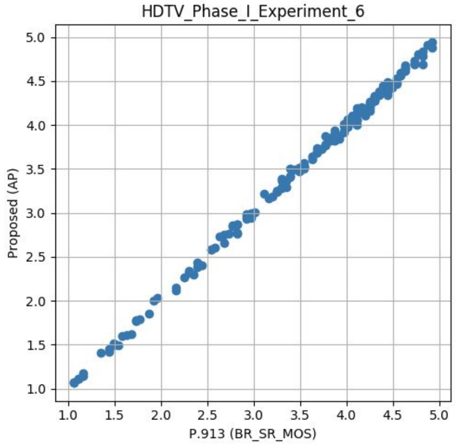


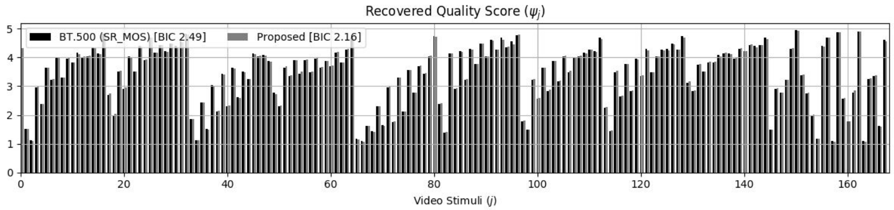


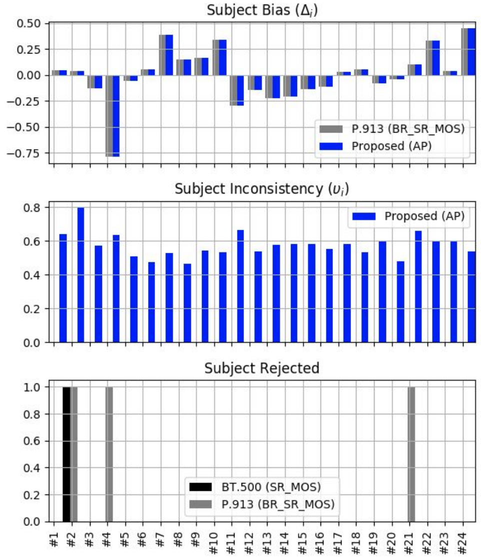


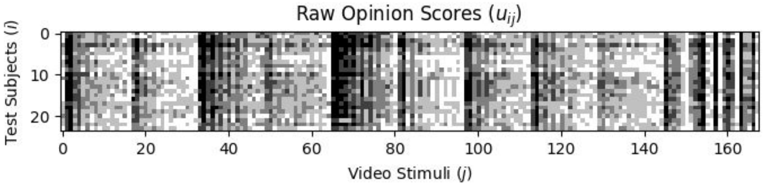
HDTV_Phase_I_Experiment_6



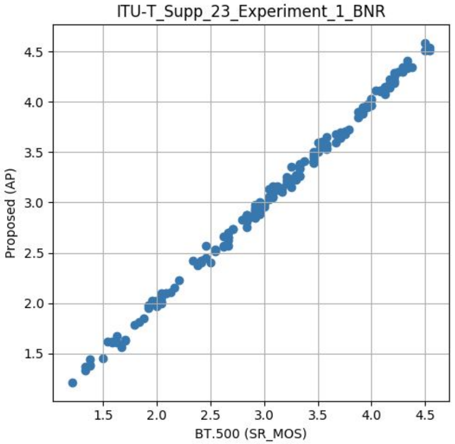


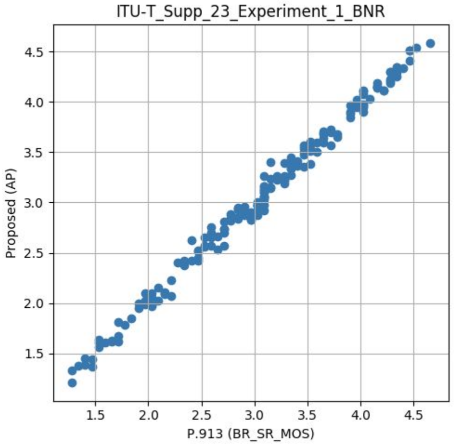


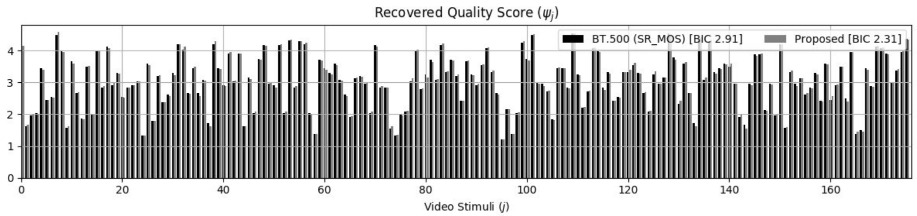


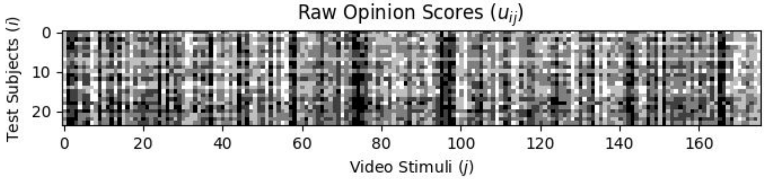


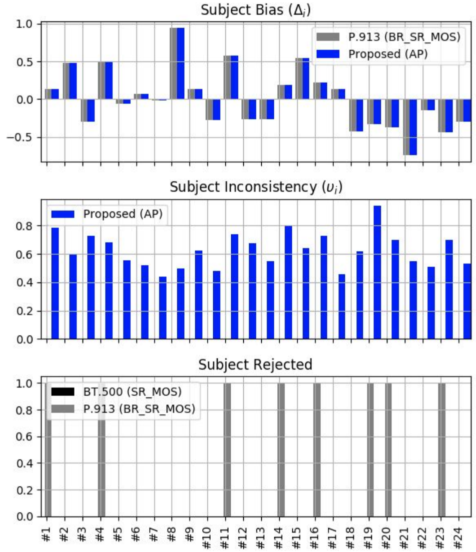
ITU-T_Supp_23_Experiment_1_BNR



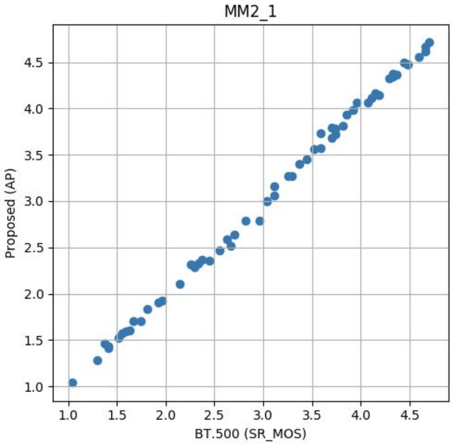


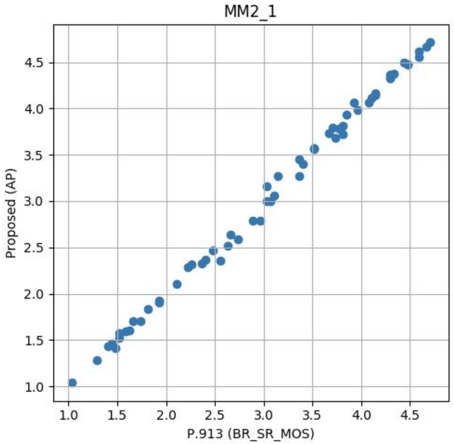


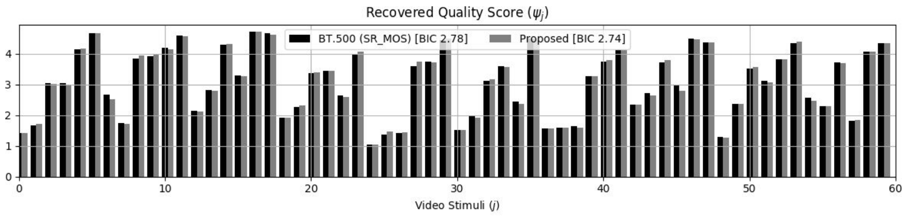


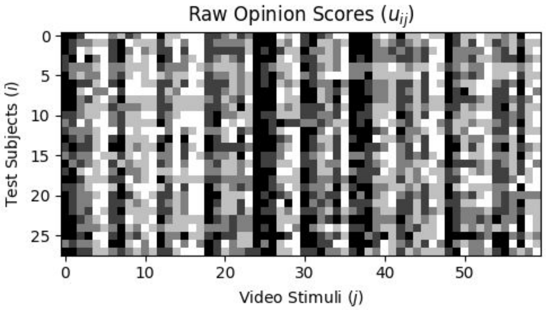


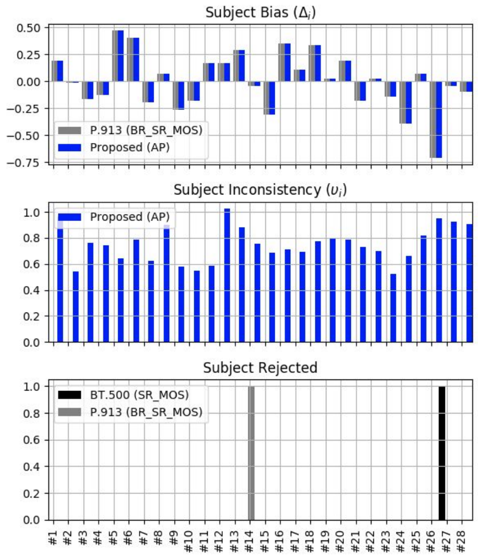
MM2_1


MM2_1
Proposed (AP)
BT.500 (SR_MOS)

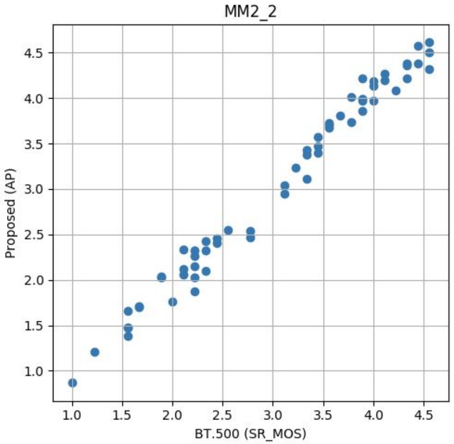

MM2_1
Proposed (AP)
P.913 (BR_SR_MOS)

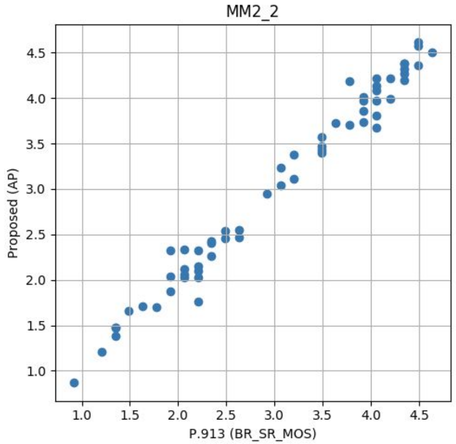

Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 2.78]
Proposed [BIC 2.74]
Video Stimuli (j)

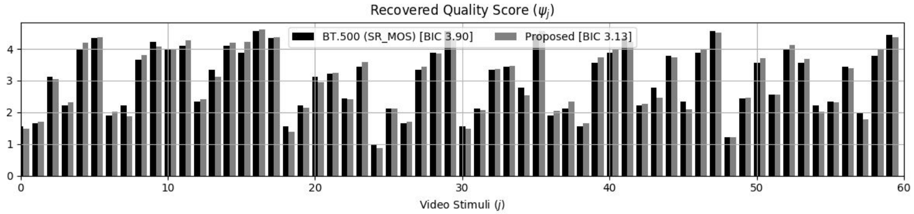

Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)

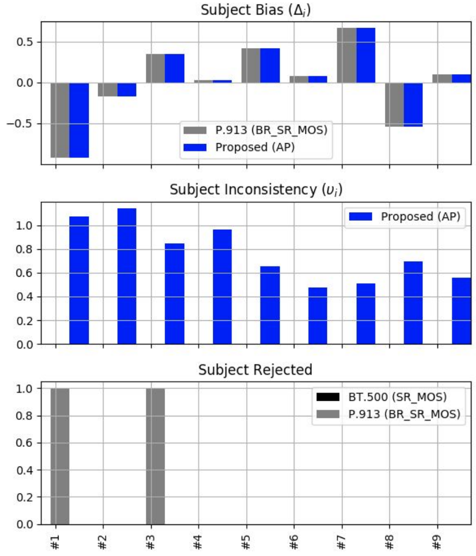

Subject Bias (Δ_i)
P.913 (BR_SR_MOS)
Proposed (AP)
Subject Inconsistency (υ_i)
Proposed (AP)
Subject Rejected
BT.500 (SR_MOS)
P.913 (BR_SR_MOS)

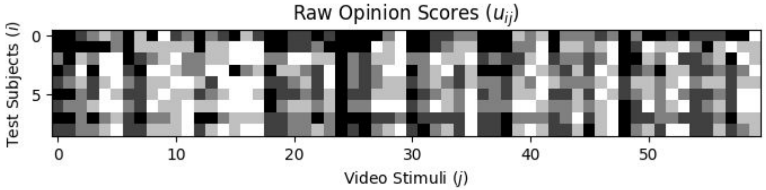
MM2_2


MM2_2
Proposed (AP)
BT.500 (SR_MOS)

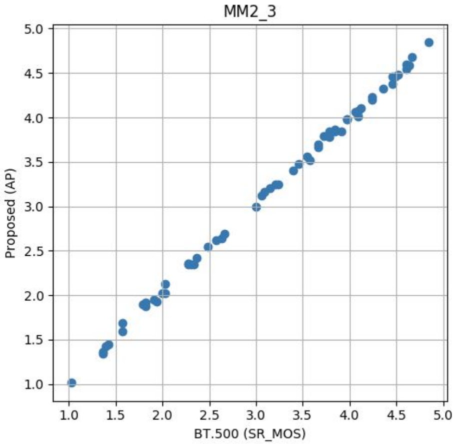

MM2_2
Proposed (AP)
P.913 (BR_SR_MOS)

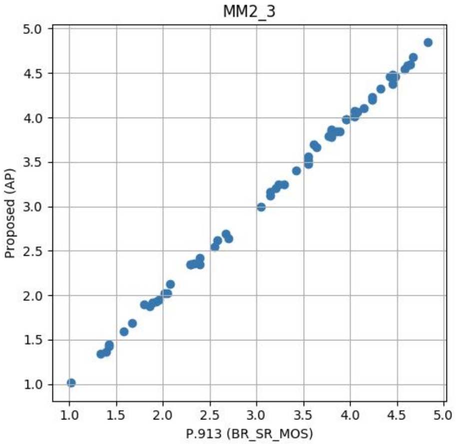

Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 3.90]
Proposed [BIC 3.13]
Video Stimuli (j)

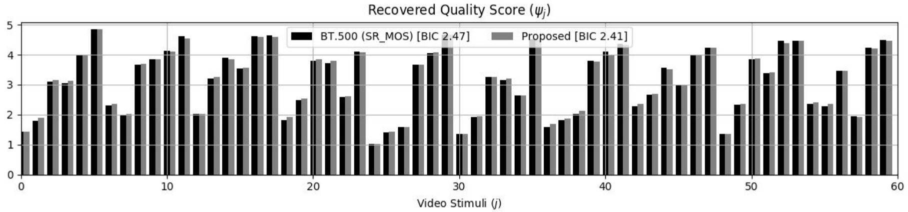

Subject Bias (Δ_i)
P.913 (BR_SR_MOS)
Proposed (AP)
Subject Inconsistency (υ_i)
Proposed (AP)
Subject Rejected
BT.500 (SR_MOS)
P.913 (BR_SR_MOS)
#1 #2 #3 #4 #5 #6 #7 #8 #9

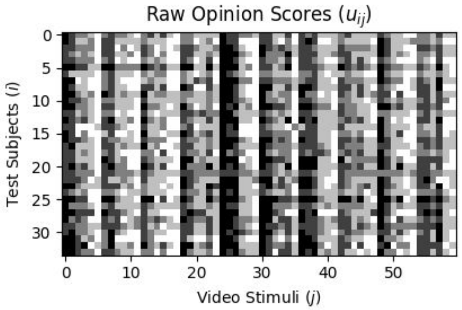

Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)

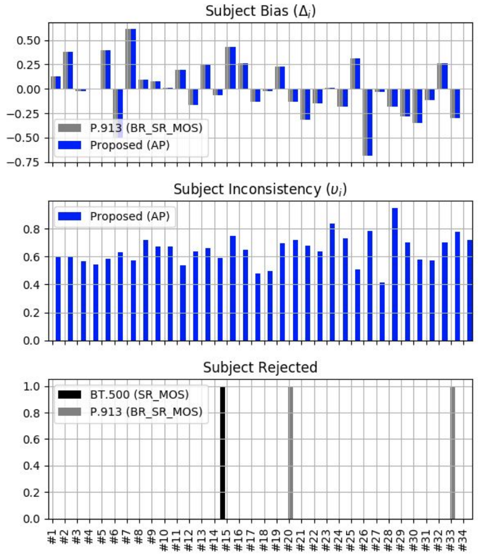
MM2_3


MM2_3
Proposed (AP)
BT.500 (SR_MOS)

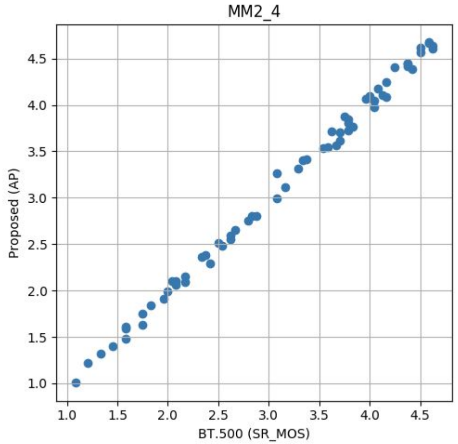

MM2_3
Proposed (AP)
P.913 (BR_SR_MOS)

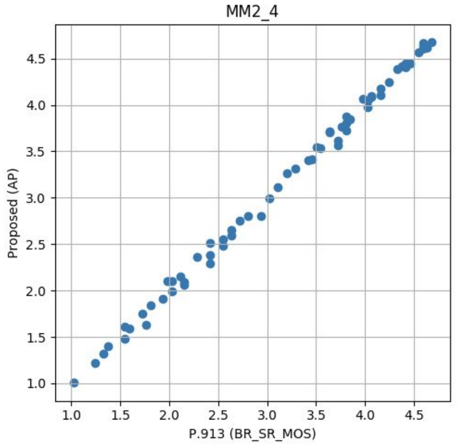

Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 2.47]
Proposed [BIC 2.41]
Video Stimuli (j)

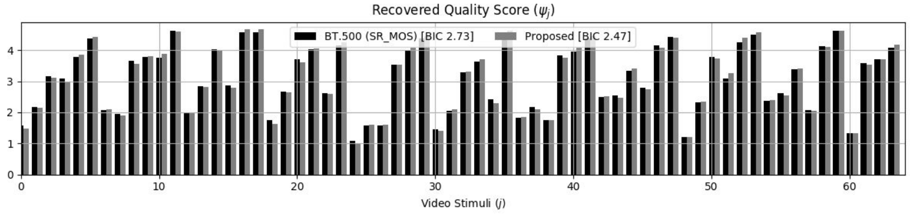

Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)

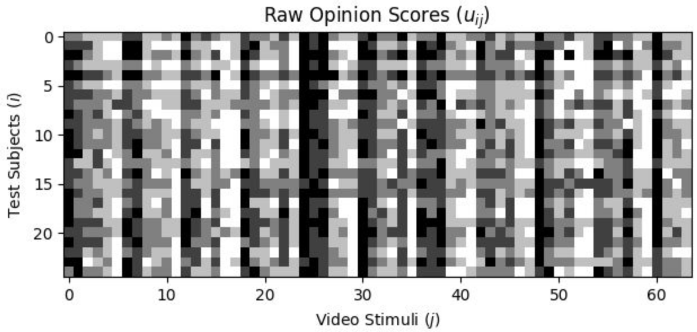

Subject Bias (Δ_i)
P.913 (BR_SR_MOS)
Proposed (AP)
Subject Inconsistency (υ_i)
Proposed (AP)
Subject Rejected
BT.500 (SR_MOS)
P.913 (BR_SR_MOS)

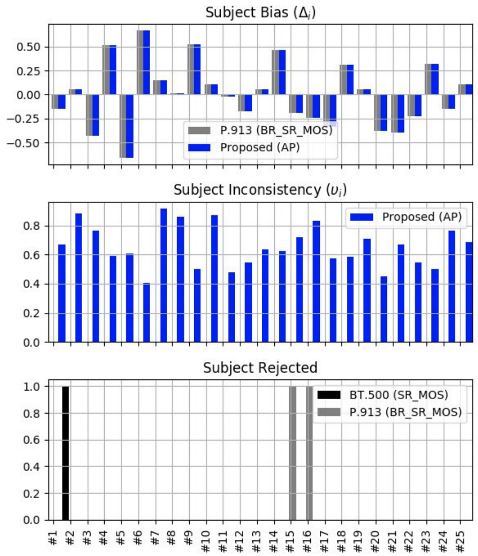
MM2_4


MM2_4
Proposed (AP)
BT.500 (SR_MOS)

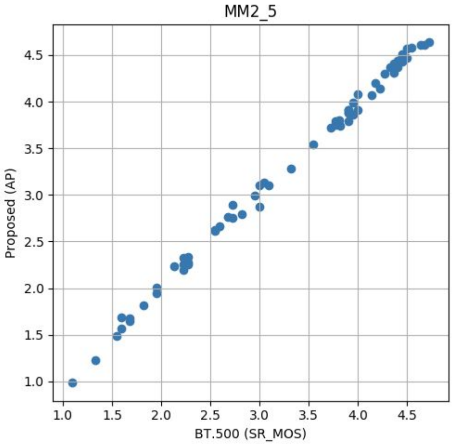

MM2_4
Proposed (AP)
P.913 (BR_SR_MOS)

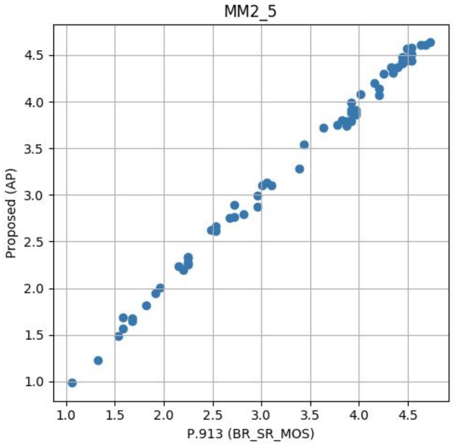

Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 2.73]
Proposed [BIC 2.47]
Video Stimuli (j)

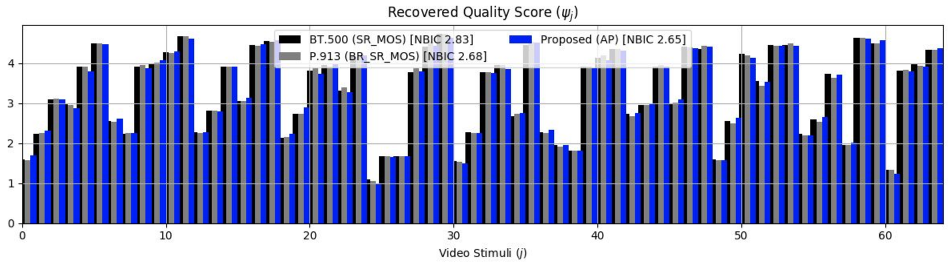

Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)

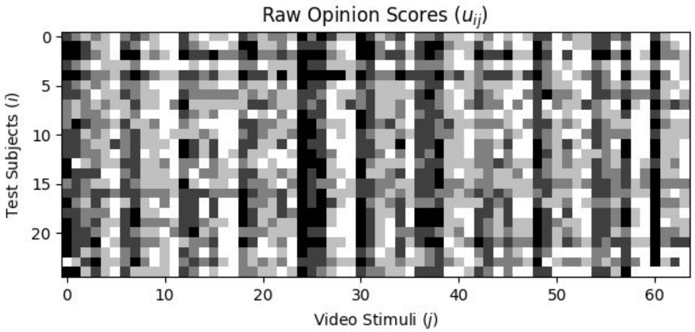

Subject Bias (Δ_i)
P.913 (BR_SR_MOS)
Proposed (AP)
Subject Inconsistency (υ_i)
Proposed (AP)
Subject Rejected
BT.500 (SR_MOS)
P.913 (BR_SR_MOS)
#1 #2 #3 #4 #5 #6 #7 #8 #9 #10 #11 #12 #13 #14 #15 #16 #17 #18 #19 #20 #21 #22 #23 #24 #25

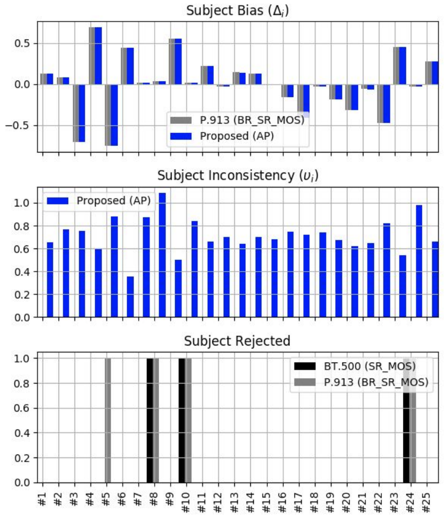
MM2_5

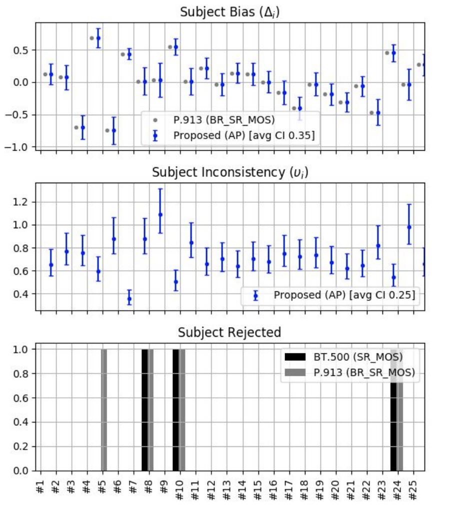

MM2_5
Proposed (AP)
BT.500 (SR_MOS)


MM2_5
Proposed (AP)
P.913 (BR_SR_MOS)

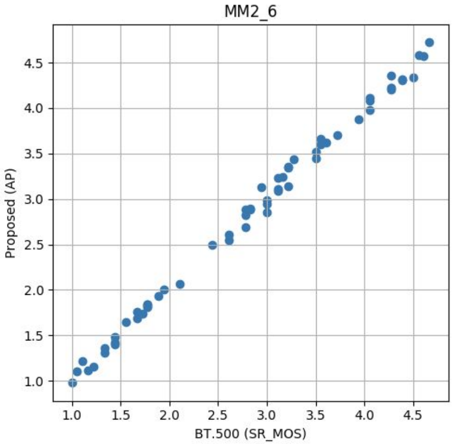

Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [NBIC 2.83]
P.913 (BR_SR_MOS) [NBIC 2.68]
Proposed (AP) [NBIC 2.65]
Video Stimuli (j)

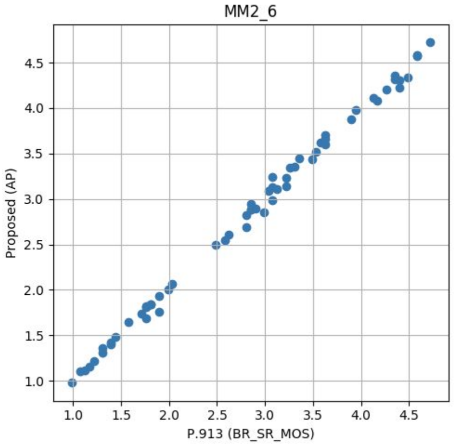

Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)

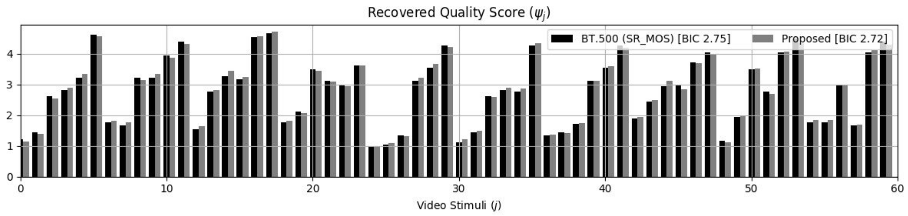

Subject Bias (Δ_i)
P.913 (BR_SR_MOS)
Proposed (AP)
Subject Inconsistency (υ_i)
Proposed (AP)
Subject Rejected
BT.500 (SR_MOS)
P.913 (BR_SR_MOS)

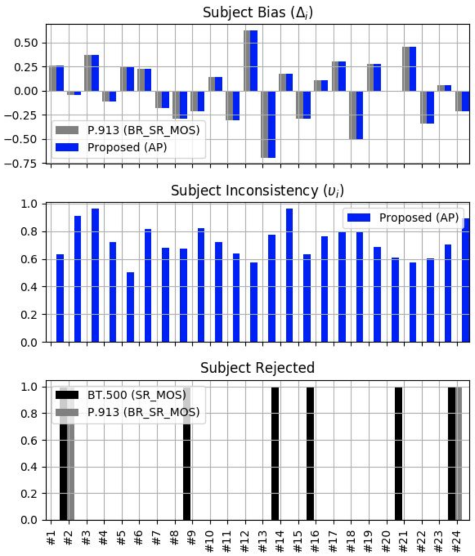

Subject Bias (Δ_i)
P.913 (BR_SR_MOS)
Proposed (AP) [avg CI 0.35]
Subject Inconsistency (υ_i)
Proposed (AP) [avg CI 0.25]
Subject Rejected
BT.500 (SR_MOS)
P.913 (BR_SR_MOS)

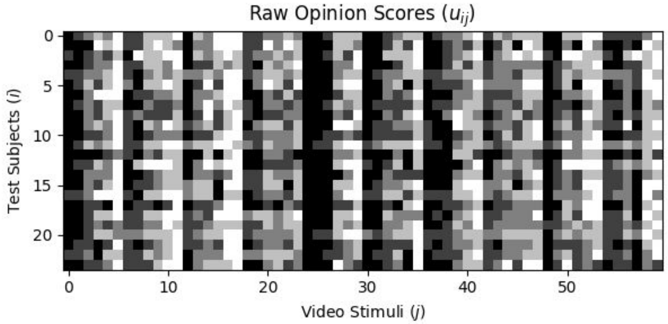
MM2_6


MM2_6
Proposed (AP)
BT.500 (SR_MOS)

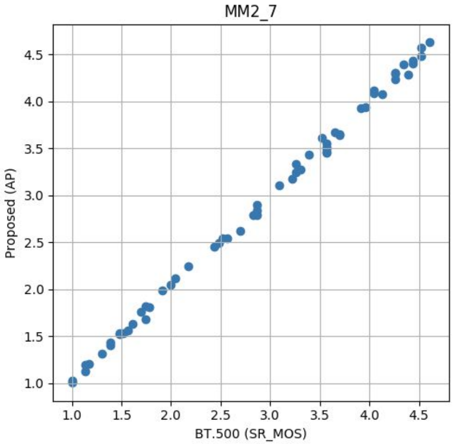

MM2_6
Proposed (AP)
P.913 (BR_SR_MOS)

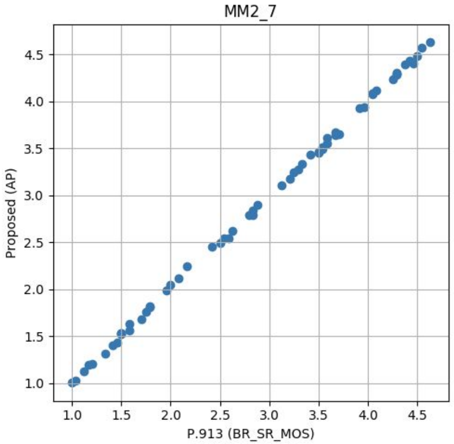

Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 2.75]
Proposed [BIC 2.72]
Video Stimuli (j)

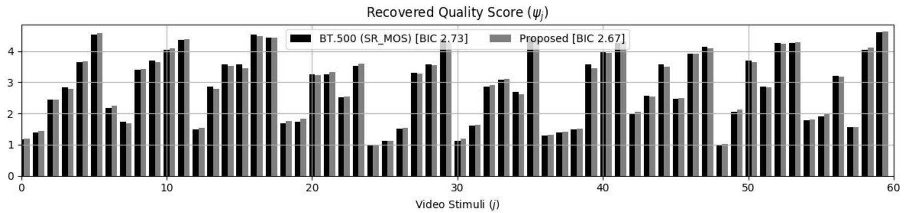

Subject Bias (Δ_i)
P.913 (BR_SR_MOS)
Proposed (AP)
Subject Inconsistency (υ_i)
Proposed (AP)
Subject Rejected
BT.500 (SR_MOS)
P.913 (BR_SR_MOS)
#1 #2 #3 #4 #5 #6 #7 #8 #9 #10 #11 #12 #13 #14 #15 #16 #17 #18 #19 #20 #21 #22 #23 #24

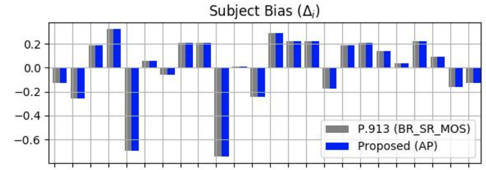

Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)

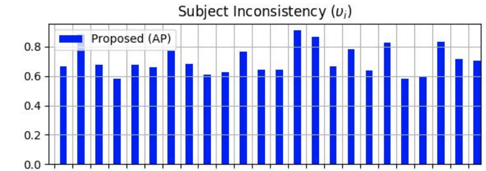
MM2_7

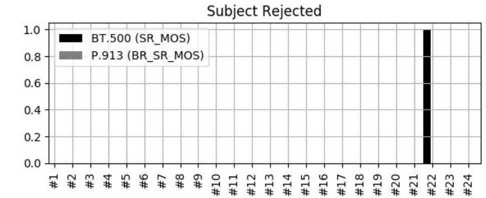

MM2_7
Proposed (AP)
BT.500 (SR_MOS)

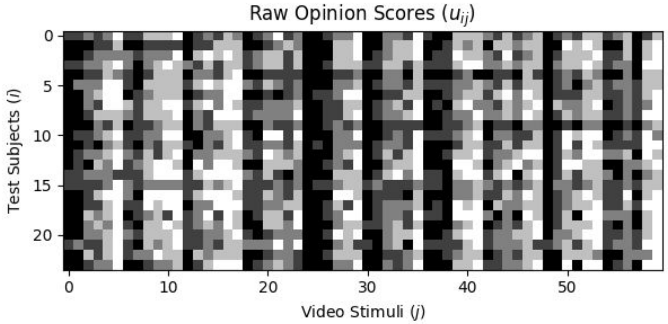

MM2_7
Proposed (AP)
P.913 (BR_SR_MOS)


Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 2.73]
Proposed [BIC 2.67]
Video Stimuli (j)

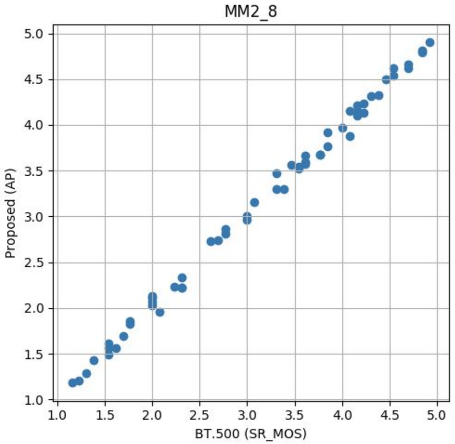

Subject Bias (Δ_i)
P.913 (BR_SR_MOS)
Proposed (AP)

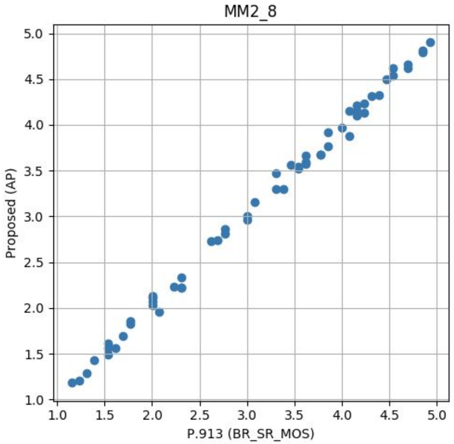

Subject Inconsistency (υ_i)
Proposed (AP)

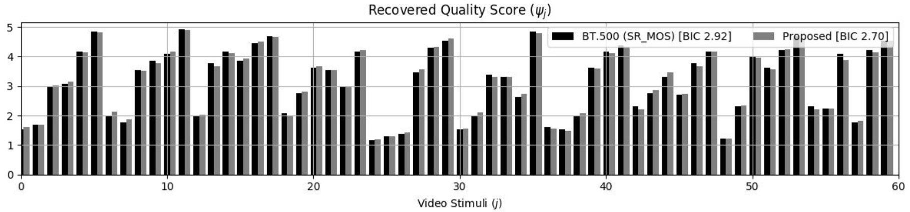

Subject Rejected
BT.500 (SR_MOS)
P.913 (BR_SR_MOS)
#1 #2 #3 #4 #5 #6 #7 #8 #9 #10 #11 #12 #13 #14 #15 #16 #17 #18 #19 #20 #21 #22 #23 #24

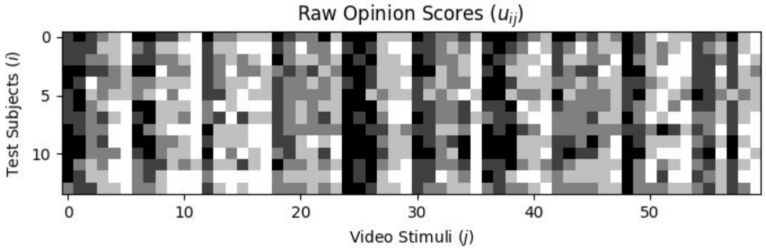

Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)

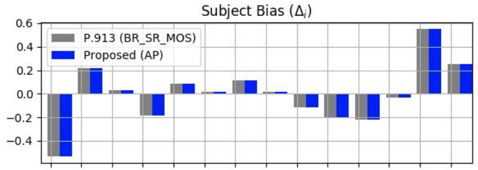
MM2_8

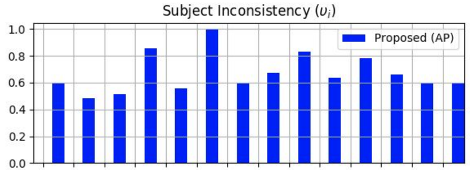

MM2_8
Proposed (AP)
BT.500 (SR_MOS)

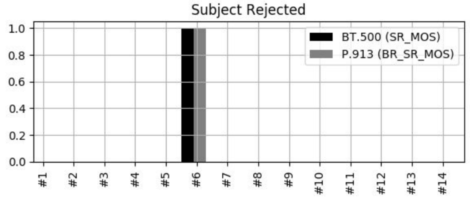

MM2_8
Proposed (AP)
P.913 (BR_SR_MOS)


Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 2.92]
Proposed [BIC 2.70]
Video Stimuli (j)

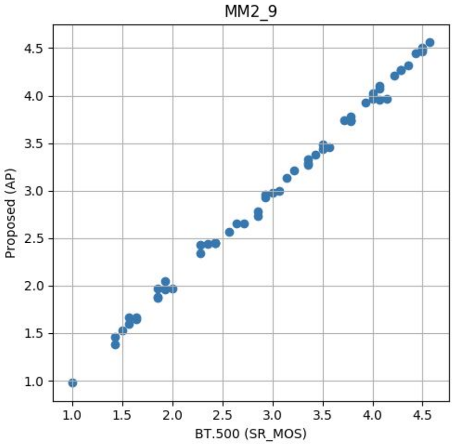

Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)

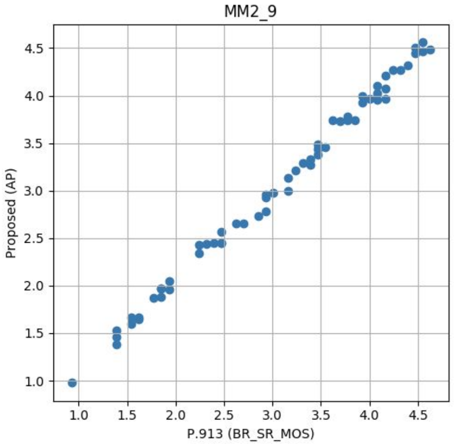

Subject Bias (Δ_i)
P.913 (BR_SR_MOS)
Proposed (AP)

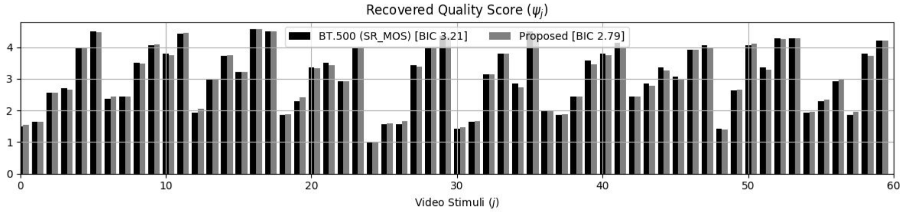

Subject Inconsistency (υ_i)
Proposed (AP)

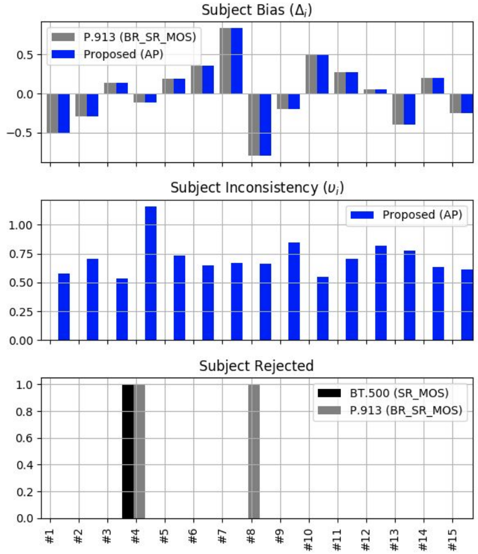

Subject Rejected
BT.500 (SR_MOS)
P.913 (BR_SR_MOS)
#1
#2
#3
#4
#5
#6
#7
#8
#9
#10
#11
#12
#13
#14

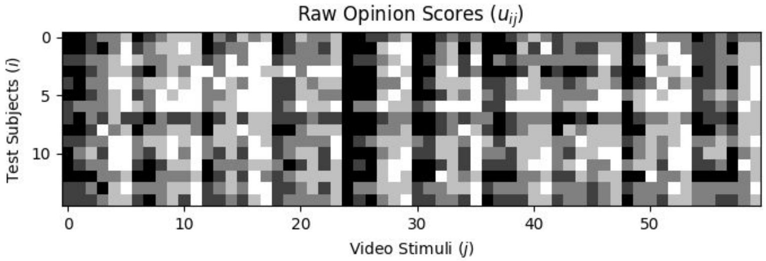
MM2_9


MM2_9
Proposed (AP)
BT.500 (SR_MOS)

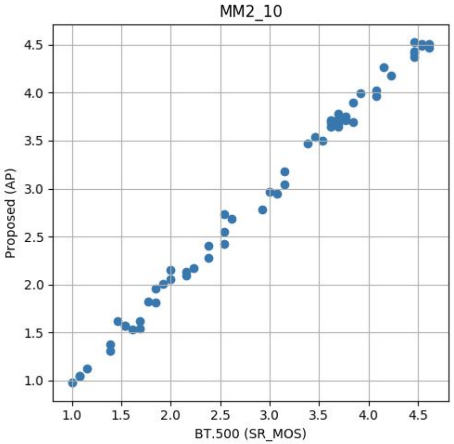

MM2_9
Proposed (AP)
P.913 (BR_SR_MOS)

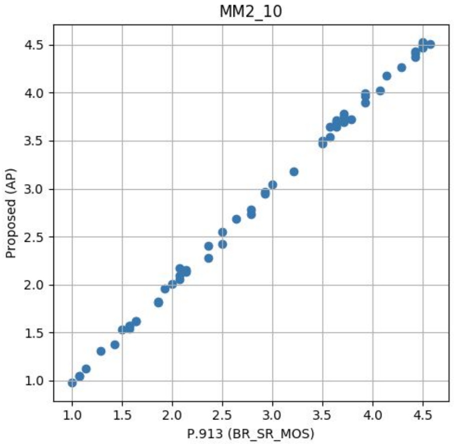

Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 3.21]
Proposed [BIC 2.79]
Video Stimuli (j)

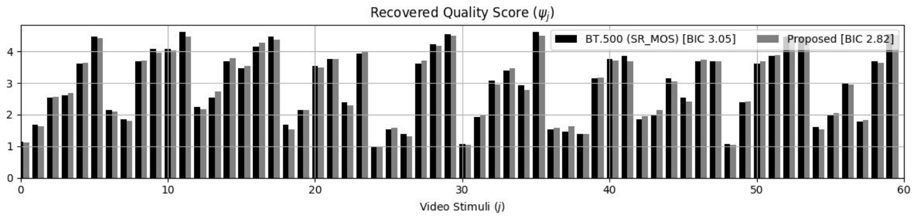

Subject Bias (Δ_i)
P.913 (BR_SR_MOS)
Proposed (AP)
Subject Inconsistency (υ_i)
Proposed (AP)
Subject Rejected
BT.500 (SR_MOS)
P.913 (BR_SR_MOS)
#1 #2 #3 #4 #5 #6 #7 #8 #9 #10 #11 #12 #13 #14 #15

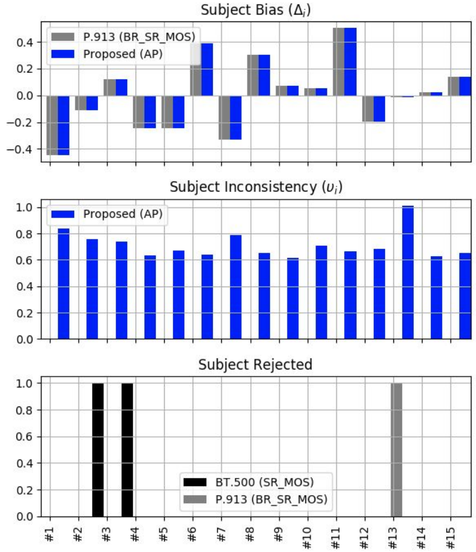

Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)

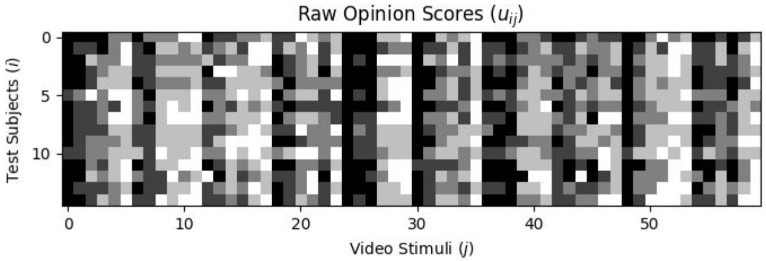
MM2_10


MM2_10
Proposed (AP)
BT.500 (SR_MOS)

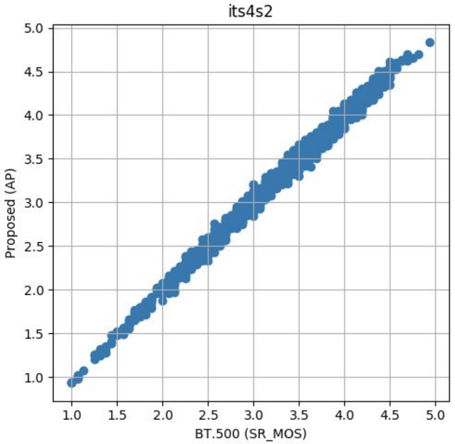

MM2_10
Proposed (AP)
P.913 (BR_SR_MOS)


Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 3.05]
Proposed [BIC 2.82]
Video Stimuli (j)

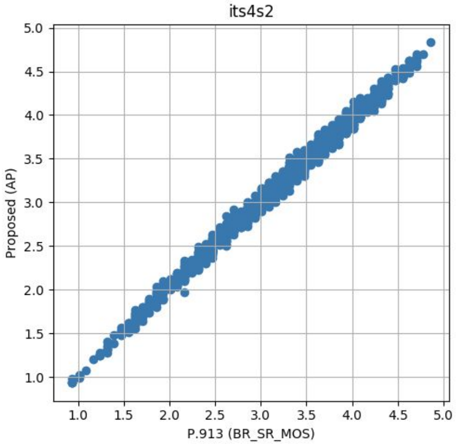

Subject Bias (Δ_i)
P.913 (BR_SR_MOS)
Proposed (AP)
Subject Inconsistency (υ_i)
Proposed (AP)
Subject Rejected
BT.500 (SR_MOS)
P.913 (BR_SR_MOS)
#1 #2 #3 #4 #5 #6 #7 #8 #9 #10 #11 #12 #13 #14 #15


Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)

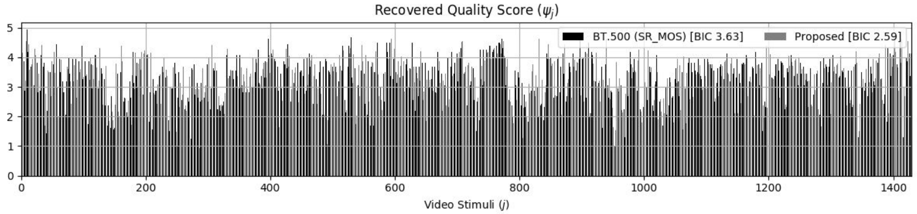
its4s2



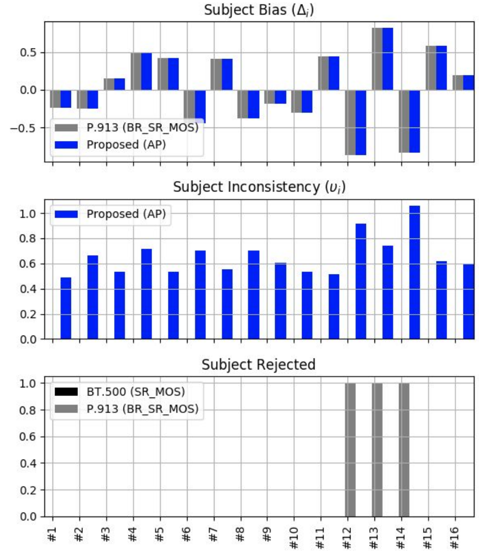
its4s2
Proposed (AP)
BT.500 (SR_MOS)



its4s2
Proposed (AP)
P.913 (BR_SR_MOS)



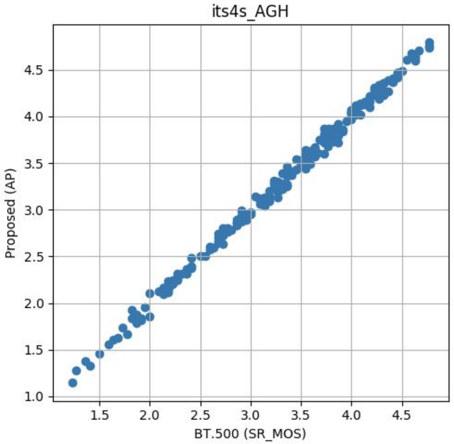
Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 3.63]
Proposed [BIC 2.59]
Video Stimuli (j)

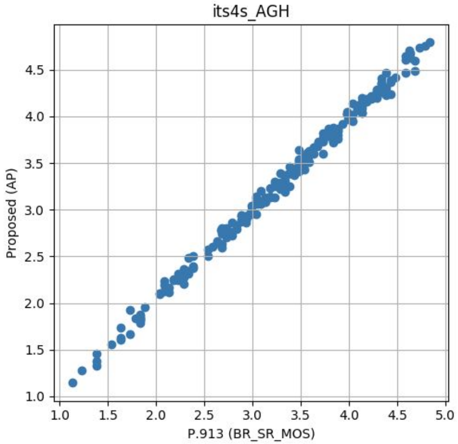


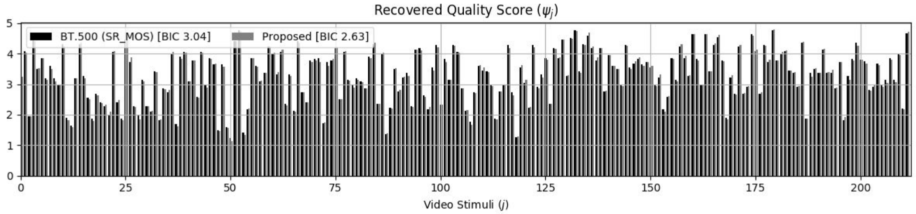
Subject Bias (Δ_i)
P.913 (BR_SR_MOS)
Proposed (AP)
Subject Inconsistency (υ_i)
Proposed (AP)
Subject Rejected
BT.500 (SR_MOS)
P.913 (BR_SR_MOS)
#1 #2 #3 #4 #5 #6 #7 #8 #9 #10 #11 #12 #13 #14 #15 #16

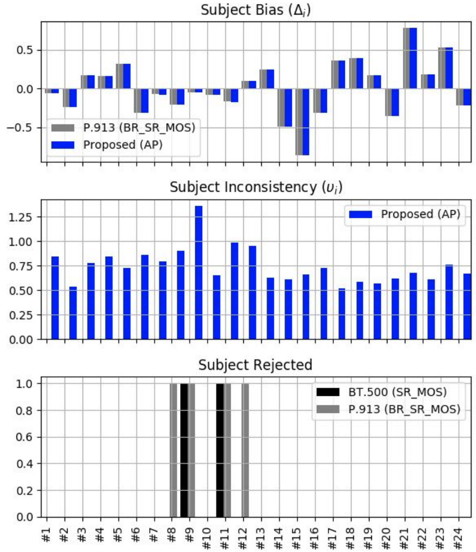
Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)

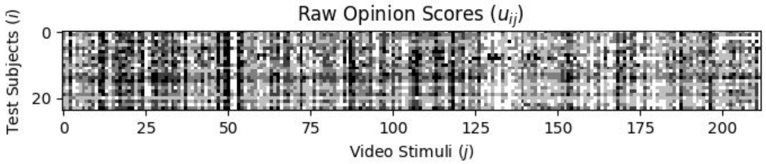
its4s_AGH


its4s_AGH
Proposed (AP)
BT.500 (SR_MOS)

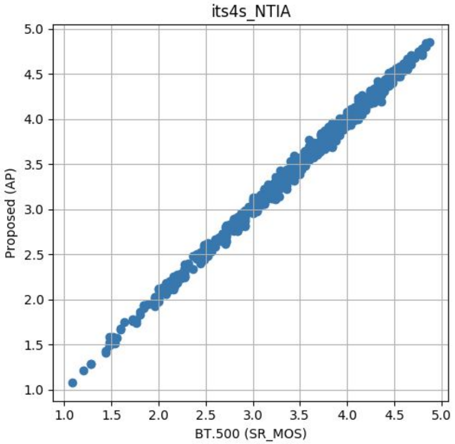

its4s_AGH
Proposed (AP)
P.913 (BR_SR_MOS)

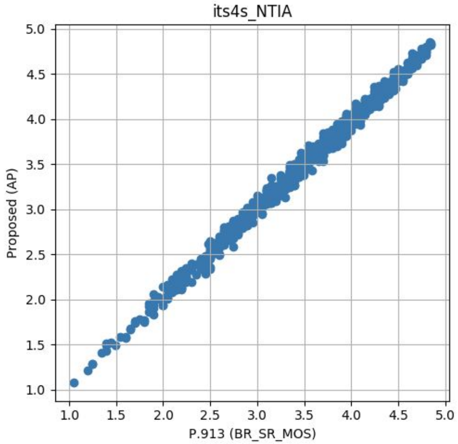

Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 3.04]
Proposed [BIC 2.63]
Video Stimuli (j)

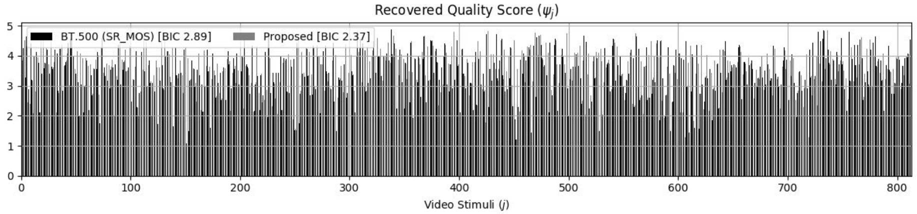

Subject Bias (Δ_i)
P.913 (BR_SR_MOS)
Proposed (AP)
Subject Inconsistency (υ_i)
Proposed (AP)
Subject Rejected
BT.500 (SR_MOS)
P.913 (BR_SR_MOS)
#1 #2 #3 #4 #5 #6 #7 #8 #9 #10 #11 #12 #13 #14 #15 #16 #17 #18 #19 #20 #21 #22 #23 #24

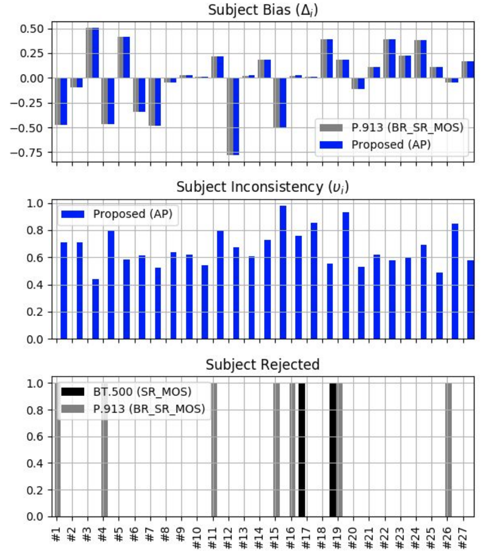

Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)

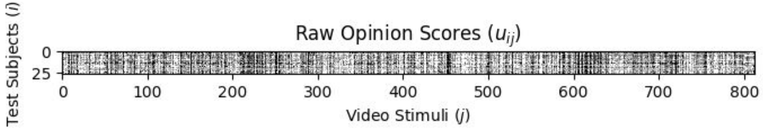
its4s_NTIA


its4s_NTIA
Proposed (AP)
BT.500 (SR_MOS)

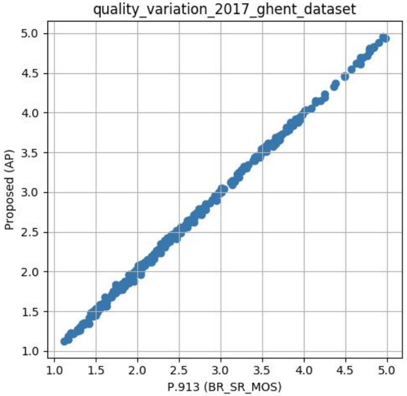

its4s_NTIA
Proposed (AP)
P.913 (BR_SR_MOS)

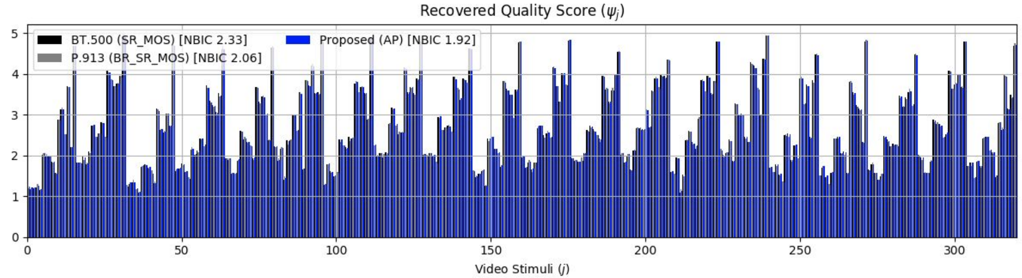

Recovered Quality Score (ψ_j)
BT.500 (SR_MOS) [BIC 2.89]
Proposed [BIC 2.37]
Video Stimuli (j)

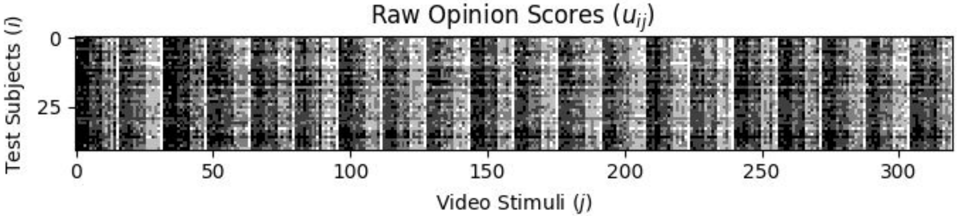

Subject Bias (Δ_i)
P.913 (BR_SR_MOS)
Proposed (AP)
Subject Inconsistency (υ_i)
Proposed (AP)
Subject Rejected
BT.500 (SR_MOS)
P.913 (BR_SR_MOS)
#1 #2 #3 #4 #5 #6 #7 #8 #9 #10 #11 #12 #13 #14 #15 #16 #17 #18 #19 #20 #21 #22 #23 #24 #25 #26 #27

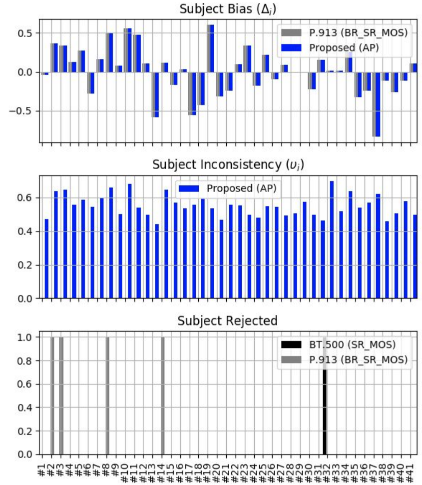

Raw Opinion Scores (u_{ij})
Test Subjects (i)
Video Stimuli (j)

quality_variation_2017_ghent

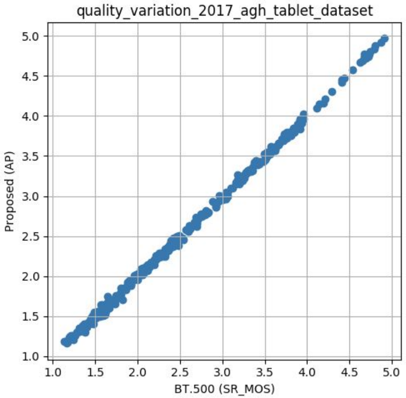


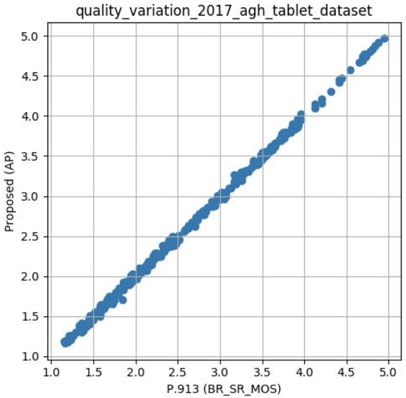


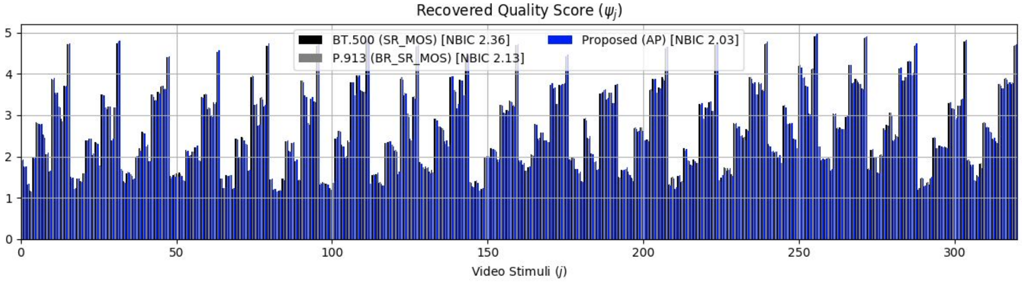


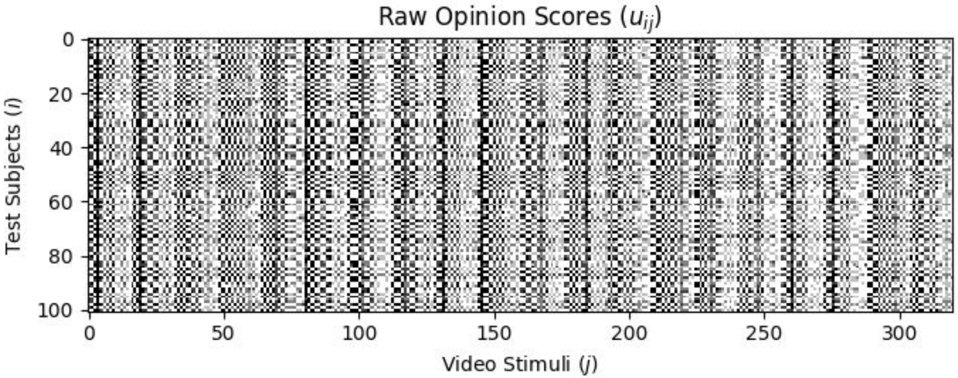


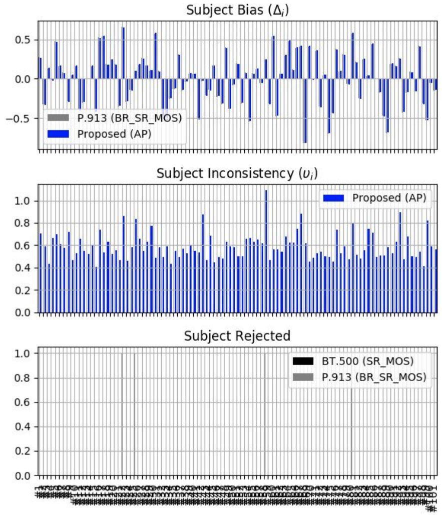
quality_variation_2017_agh_tablet



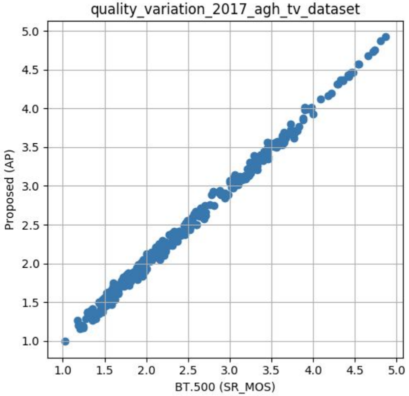


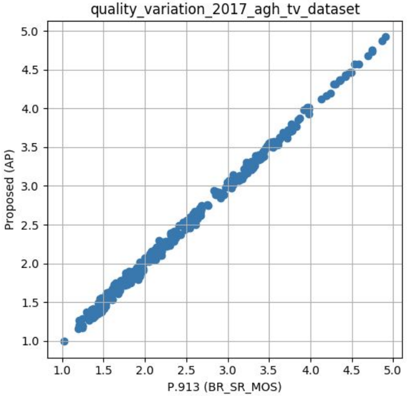


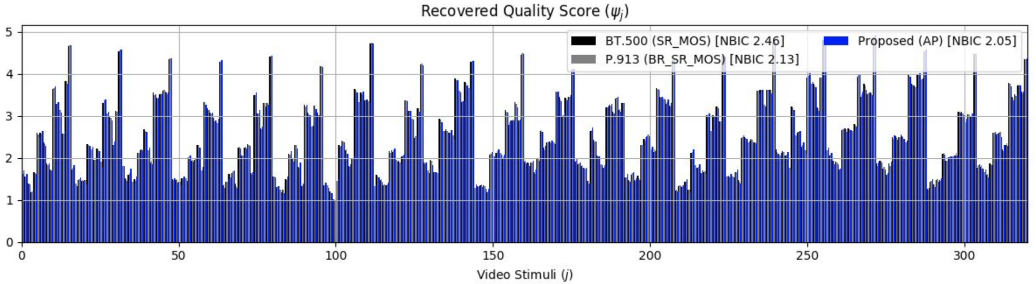


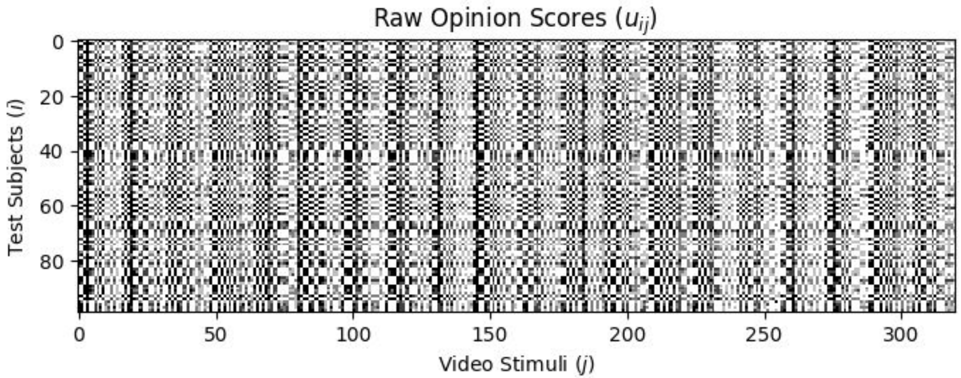


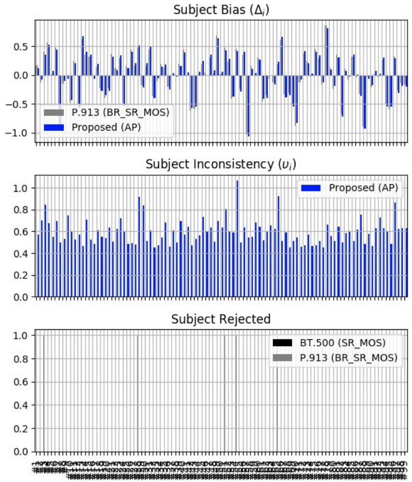
quality_variation_2017_agh_tv



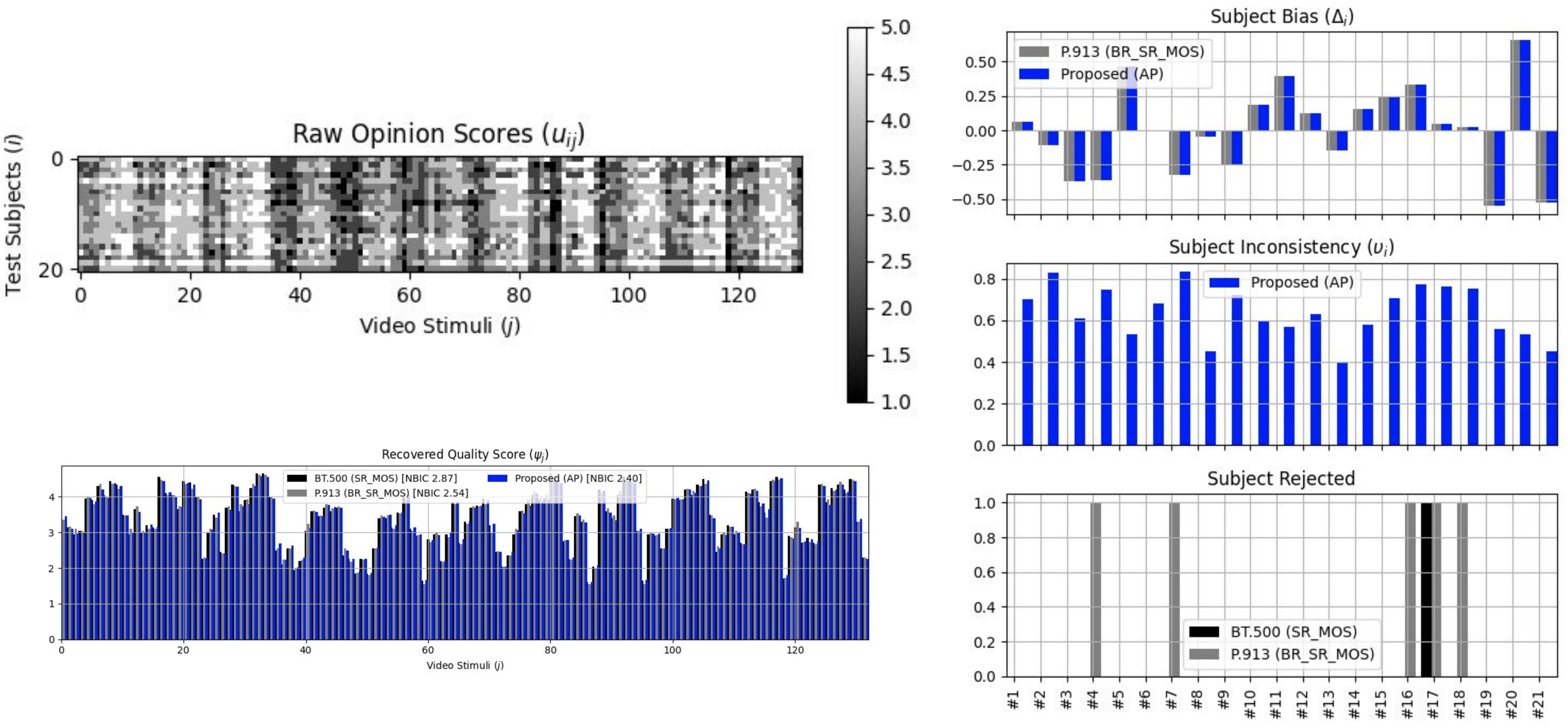




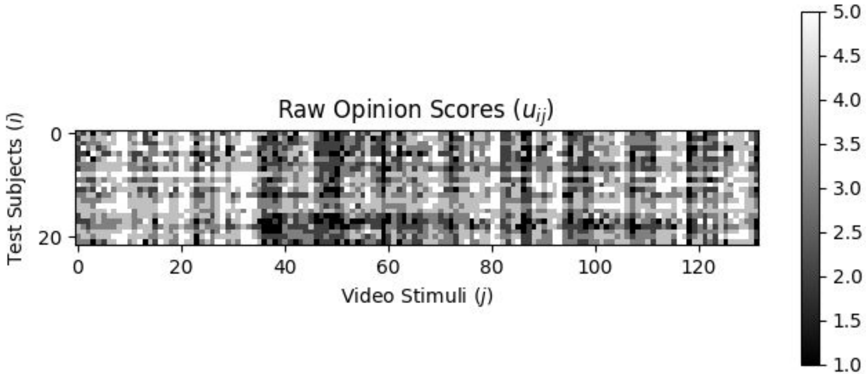


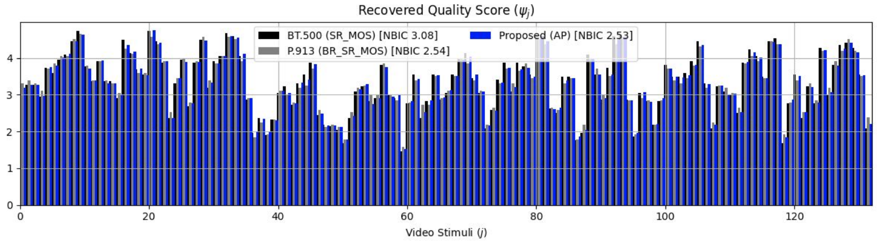


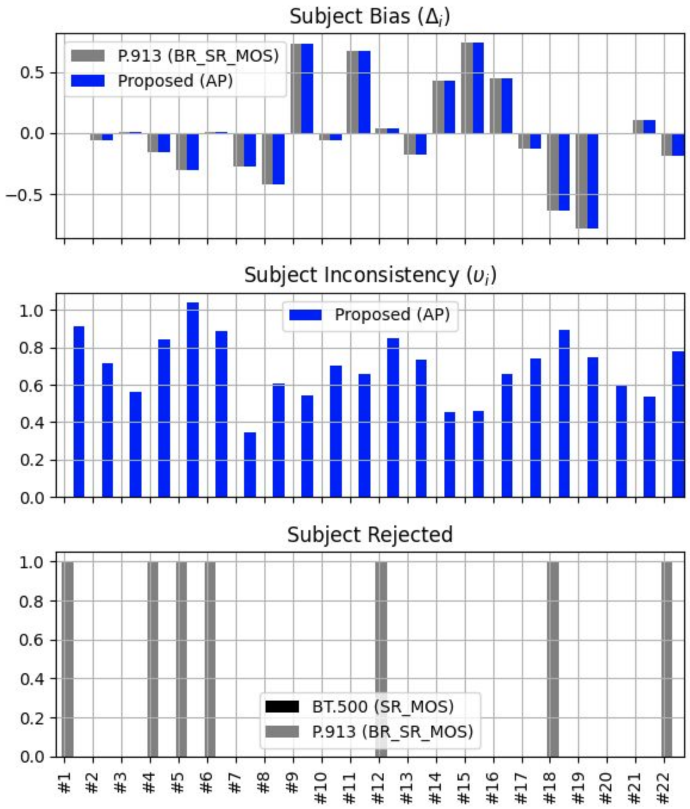
upm_acreo_as2015_upm_with_audio_dataset



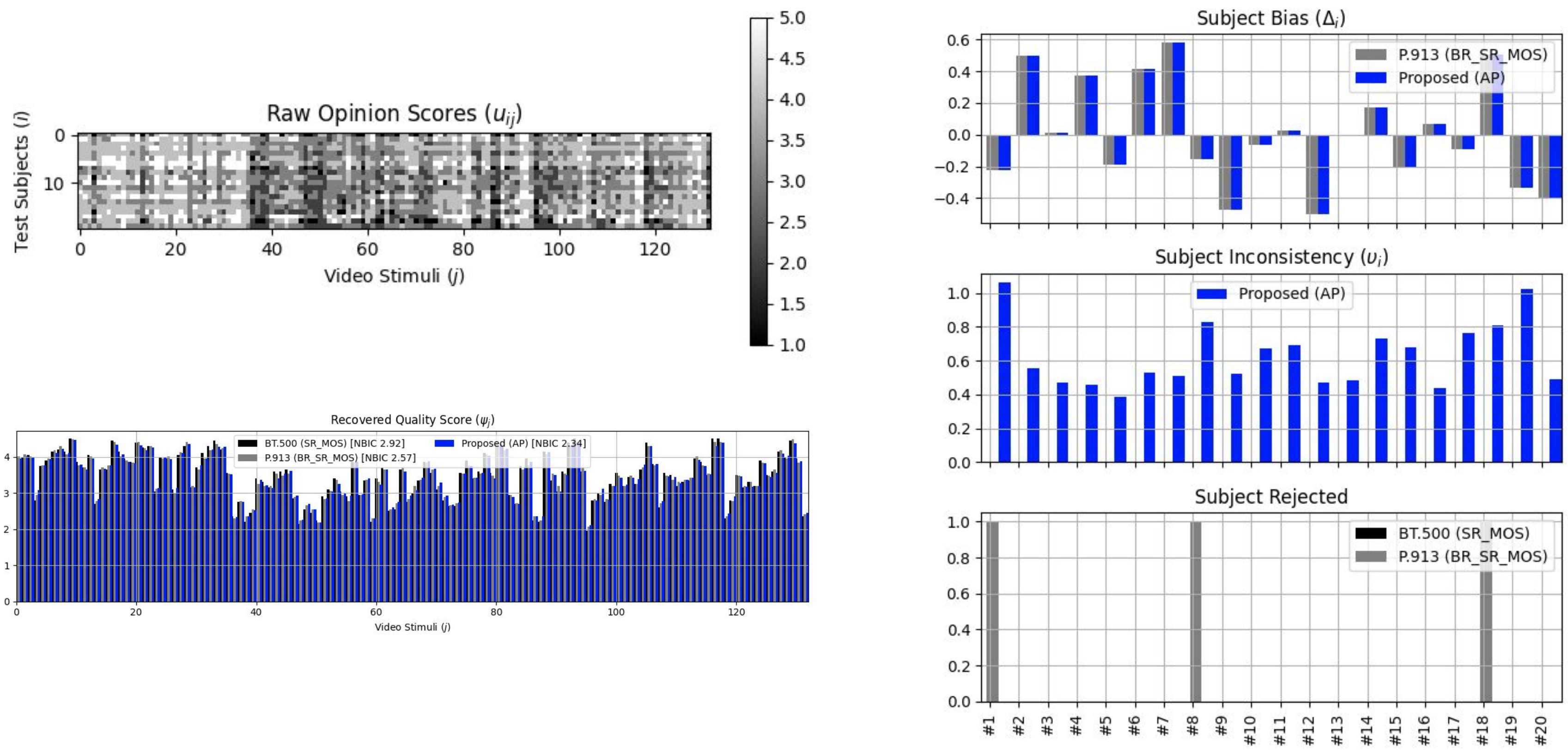
upm_acreo_as2015_upm_without_audio_dataset



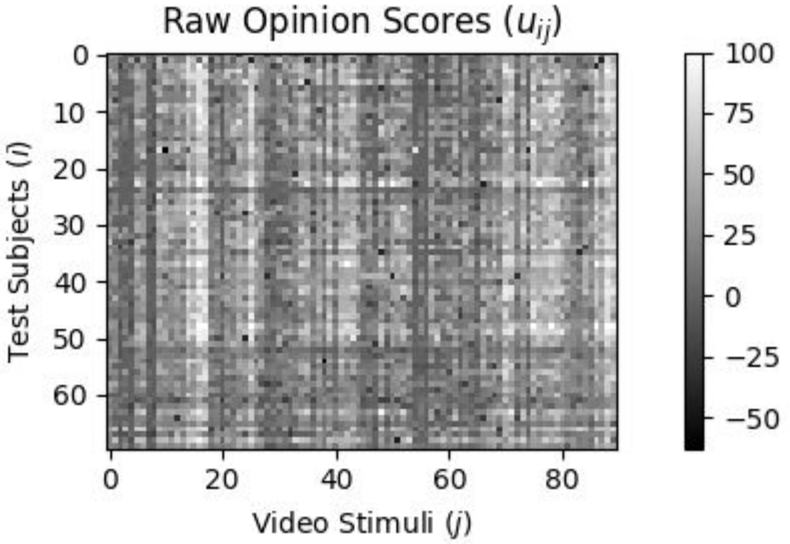


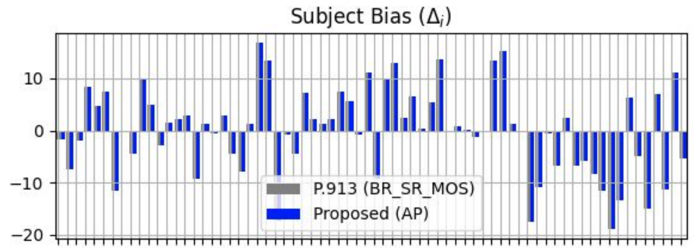


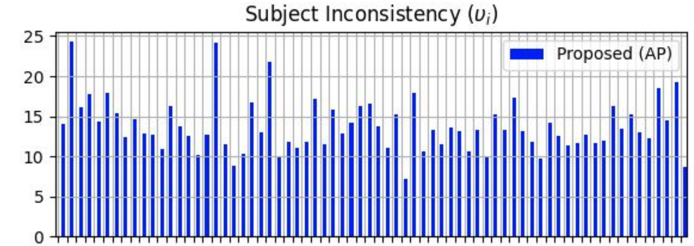
upm_acreo_as2015_acreo_without_audio_dataset

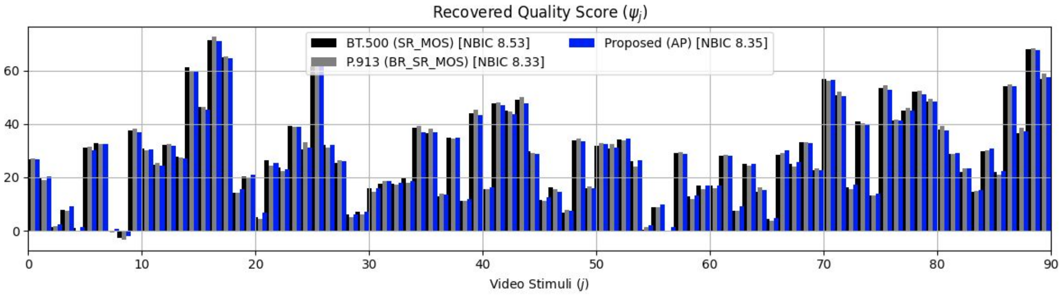


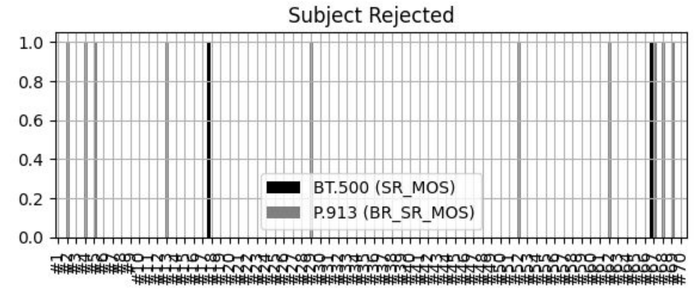
vqeg_frtv_p1_525_line_low_dataset



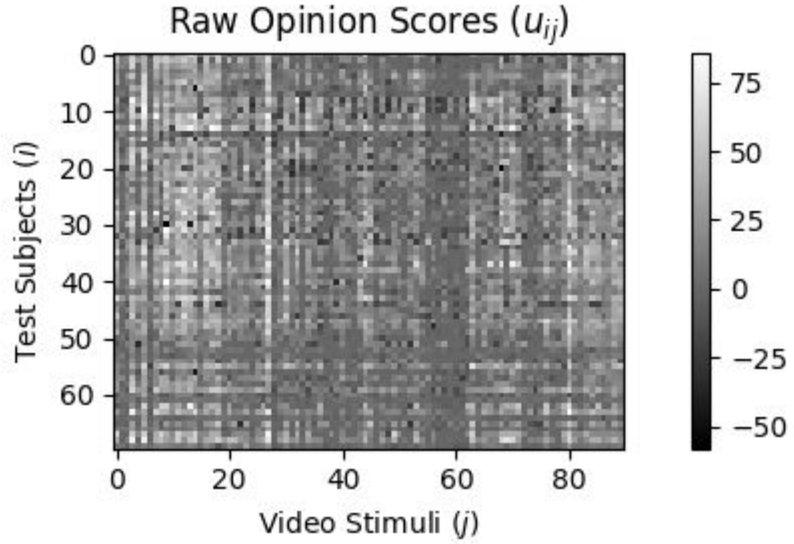


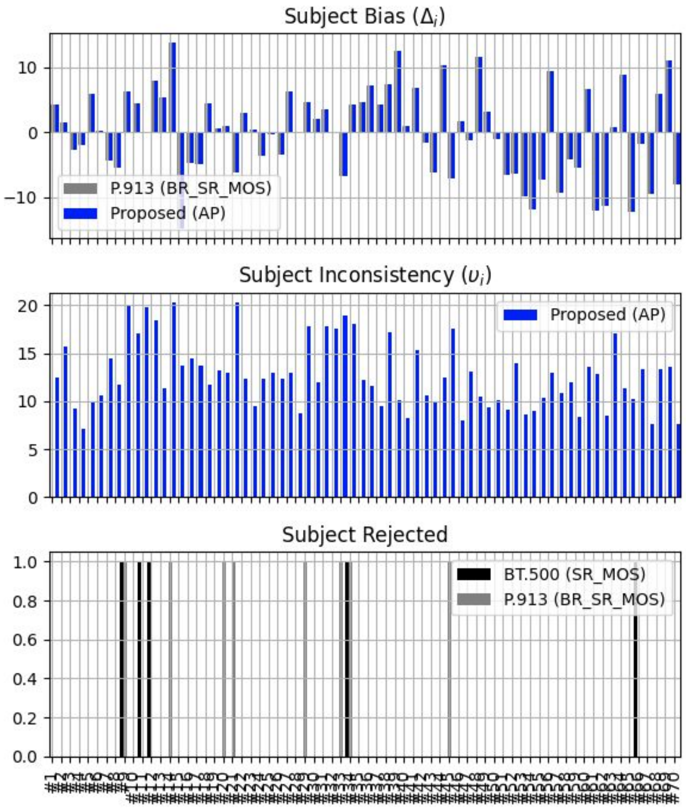


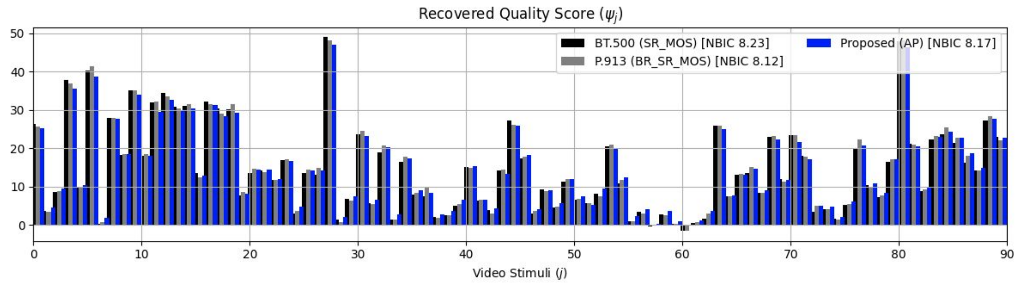




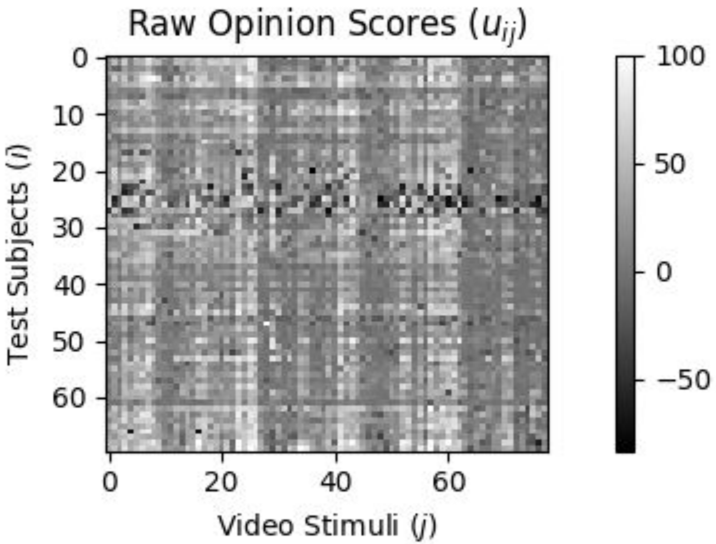
vqeg_frtv_p1_525_line_high_dataset

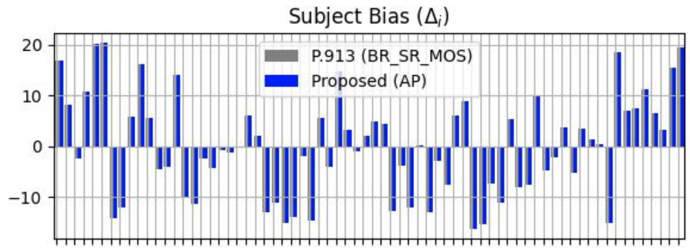


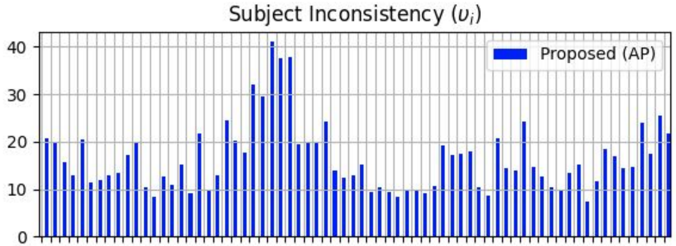


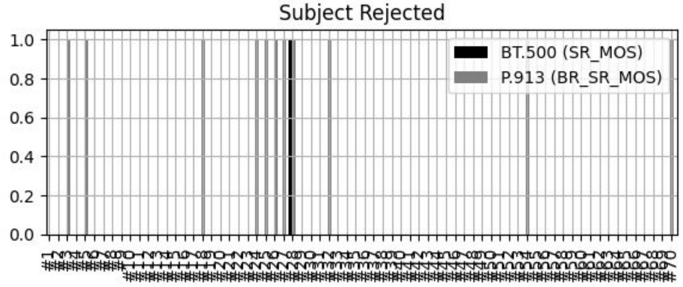


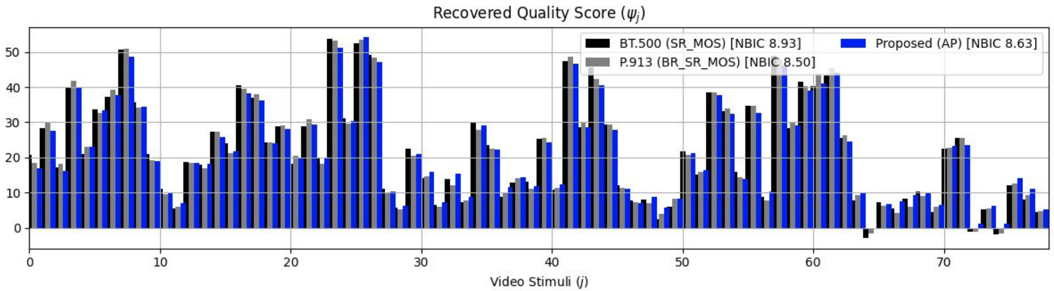
vqeg_frtv_p1_625_line_low_dataset











vqeg_frtv_p1_625_line_high_dataset