A Tale of Two Datasets

Learning from subjectively evaluating CAMBI*

Zhi Li & Lukas Krasula

Video and Image Quality
Encoding Technologies

*Details about the CAMBI banding detection algorithm will be presented in Tuesday's NORM session.

VQEG December 2021

{zli, lkrasula}@netflix.com

https://tinyurl.com/2cheb485

Banding (aka false contouring) is false staircase-like edges in otherwise smooth transitions in a picture.

One of the most prominent causes for banding is the quantization in lossy video compression.

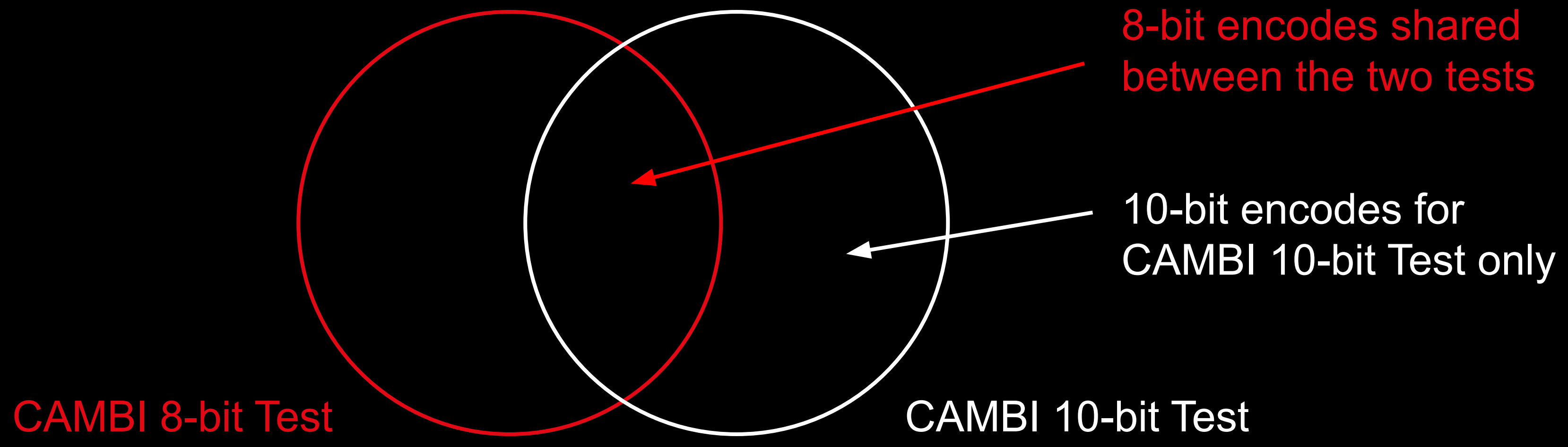
Another significant factor for banding visibility is the bit depth (e.g. 8- vs 10-bit) to represent a video signal.

8-bit depth

10-bit depth

Over the process of developing CAMBI, we have conducted two subjective tests to collect data to support algorithm tuning and validation.

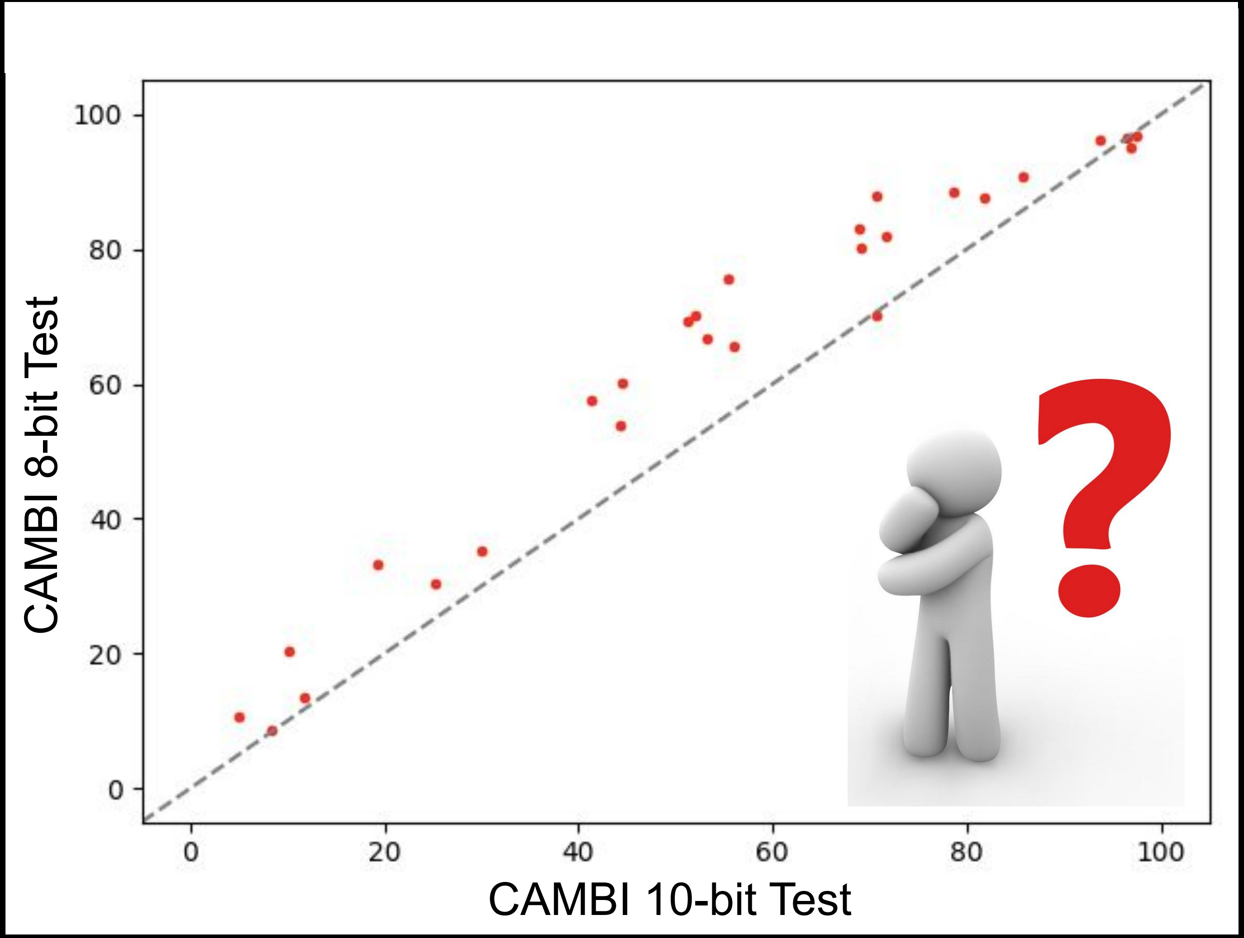
The CAMBI 8-bit Test uses only 8-bit encodes; the CAMBI 10-bit Test includes 10-bit encodes, but also a subset of 8-bit encodes from the 8-bit Test.



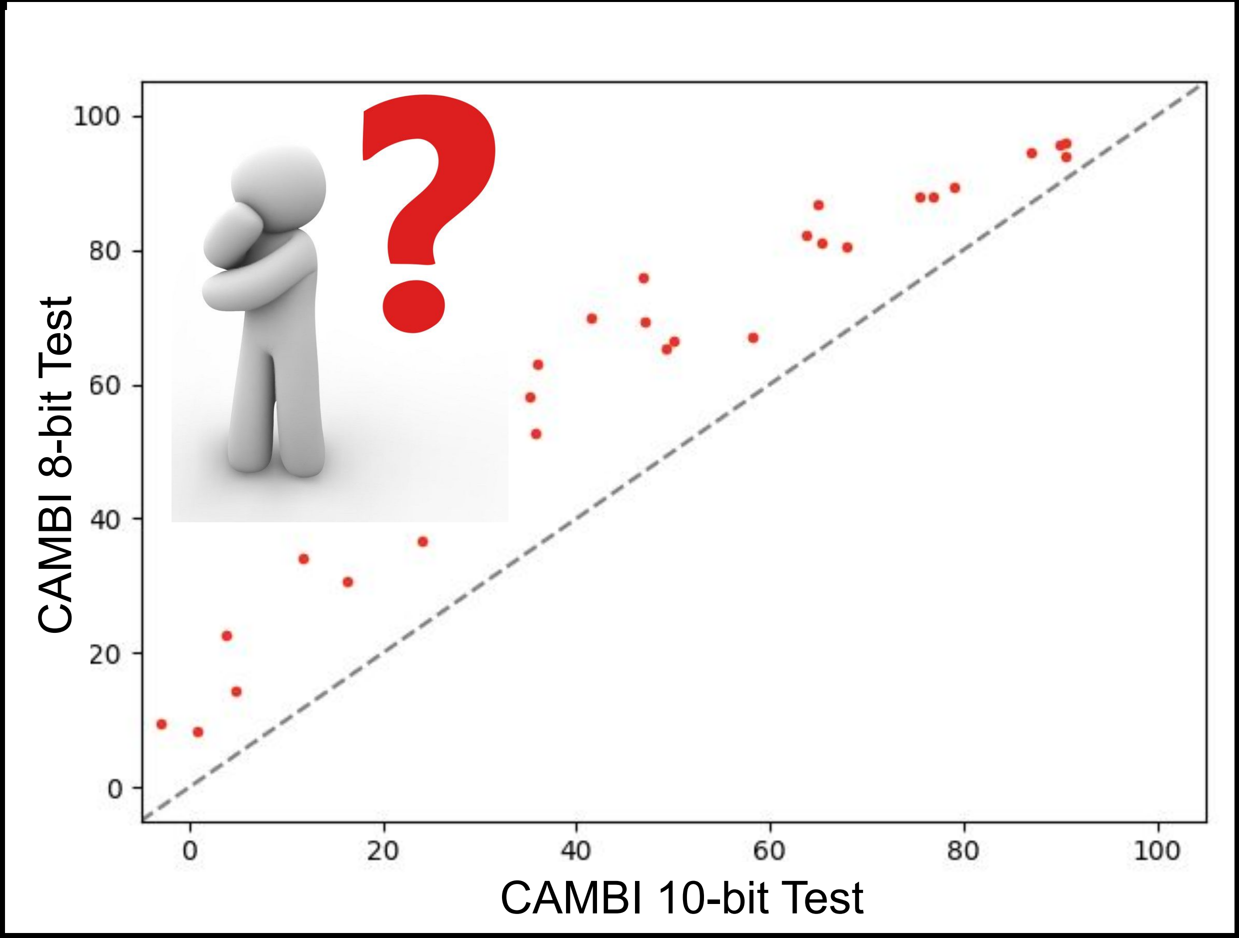
For data analysis, we use two techniques to calculate the MOS:

- Bias-subtracted MOS: ITU-T P.913 Section 12.4
- Bias-subtracted consistency-weighted MOS: recently standardized in ITU-T P.913 Section 12.6 and ITU-T P.910 Annex E (prepublished)

Recovered MOS for the 8-bit encodes shared across two datasets



Analysis using bias-subtracted MOS

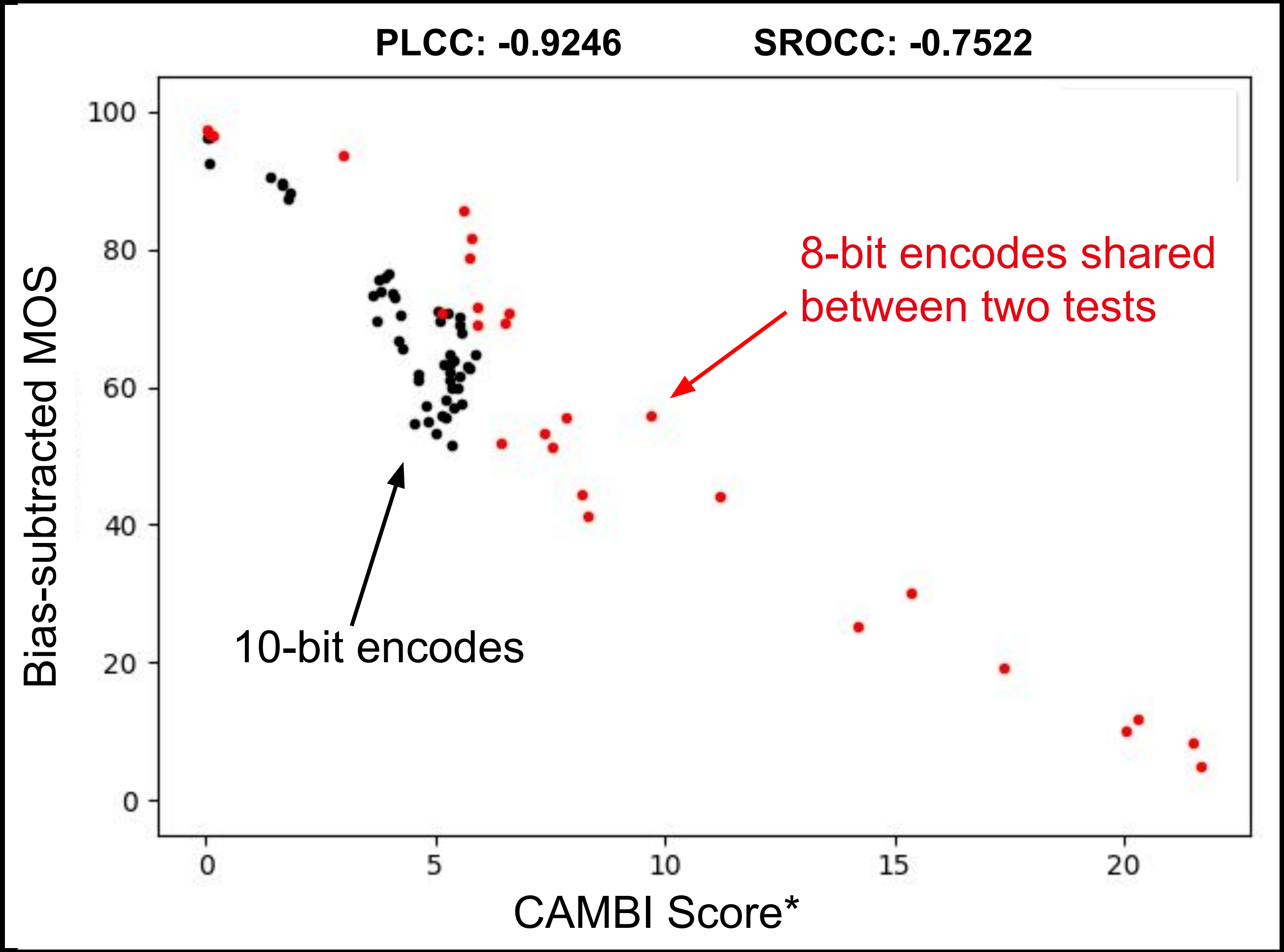


Analysis using bias-subtracted consistency-weighted MOS

Two puzzles:

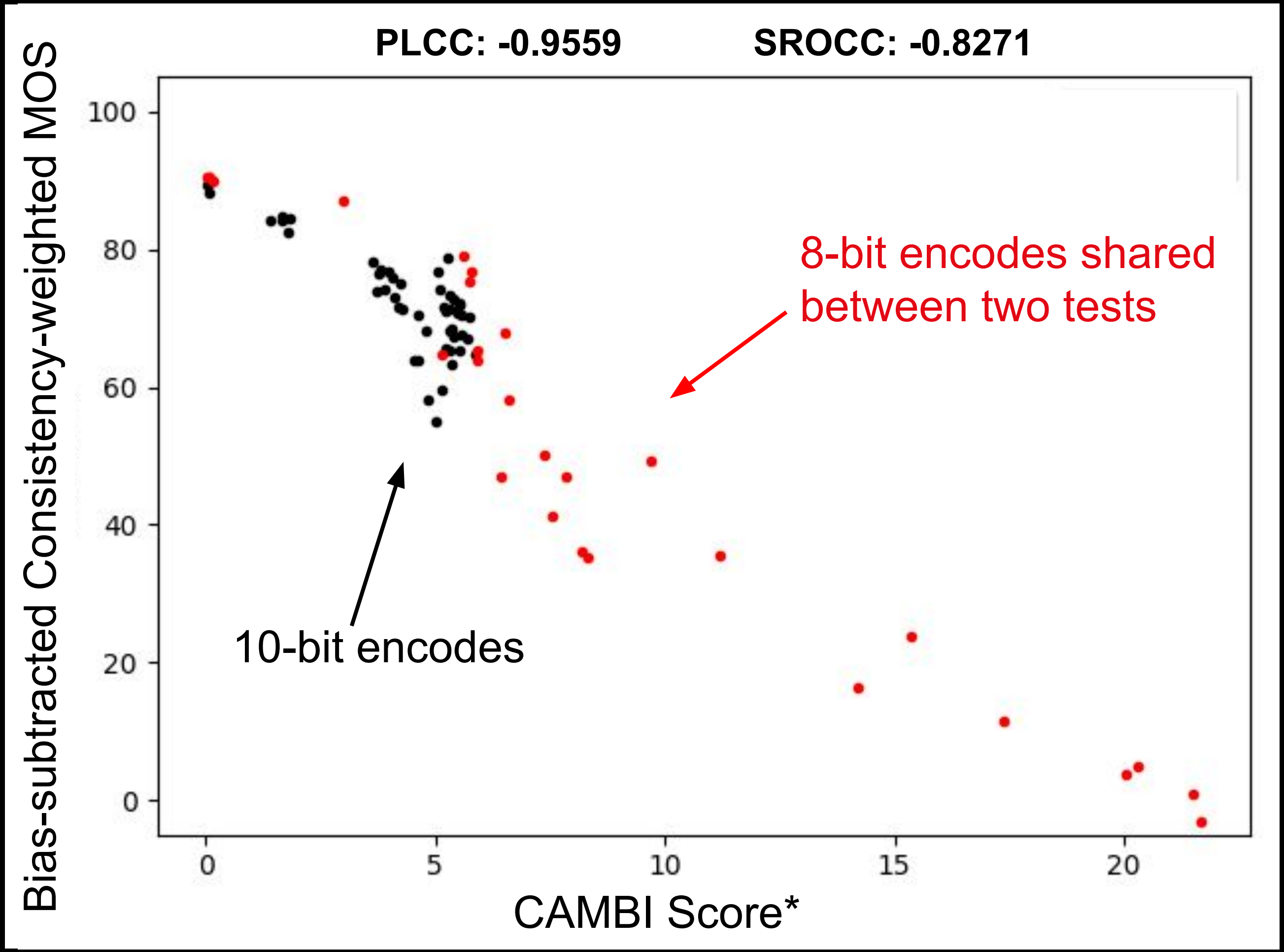
- Why do the shared 8-bit encodes receive lower scores in the CAMBI 10-bit Test than in the CAMBI 8-bit Test?
- Why does the analysis using bias-subtracted consistency-weighted MOS further encourage this behavior?

Inspecting the whole CAMBI 10-bit dataset: Bias-subtracted MOS vs. CAMBI score



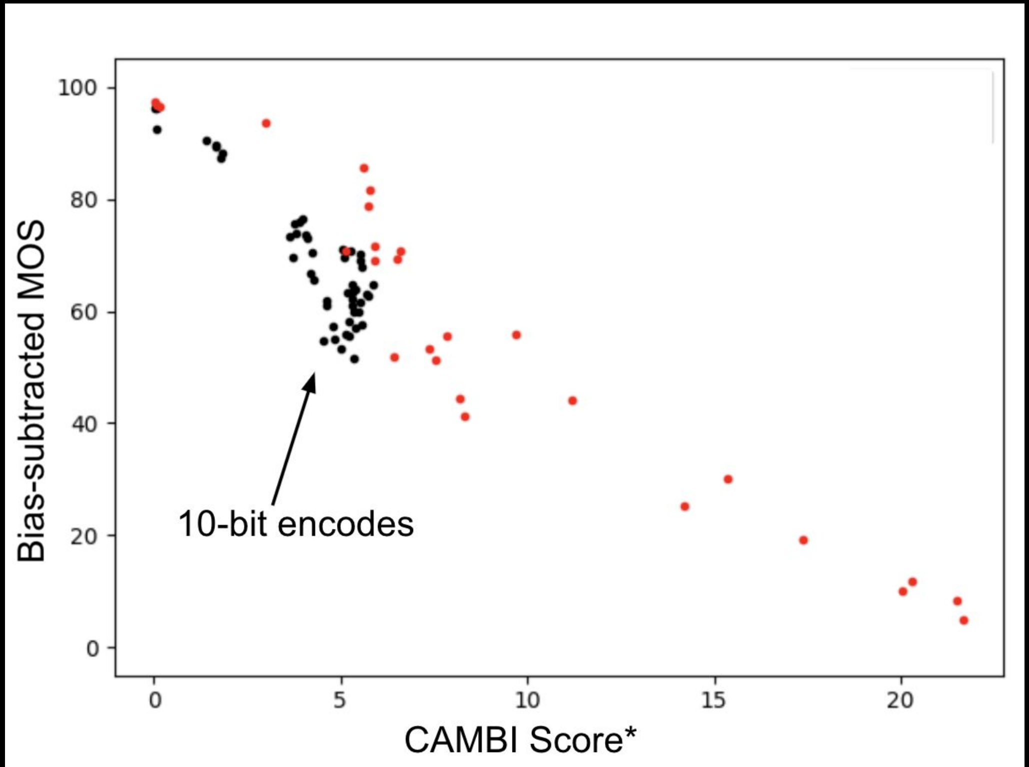
*Interpreting CAMBI score: 0 means no banding; 24 is severe banding (unwatchable); around 5 is where banding starts to become slightly annoying.

Inspecting the whole CAMBI 10-bit dataset: Bias-subtracted consistency-weighted MOS vs. CAMBI score



*Interpreting CAMBI score: 0 means no banding; 24 is severe banding (unwatchable); around 5 is where banding starts to become slightly annoying.

Observation #1: the perceptual quality of the 10-bit encodes in the CAMBI 10-bit Test dataset is very concentrated in a small region.



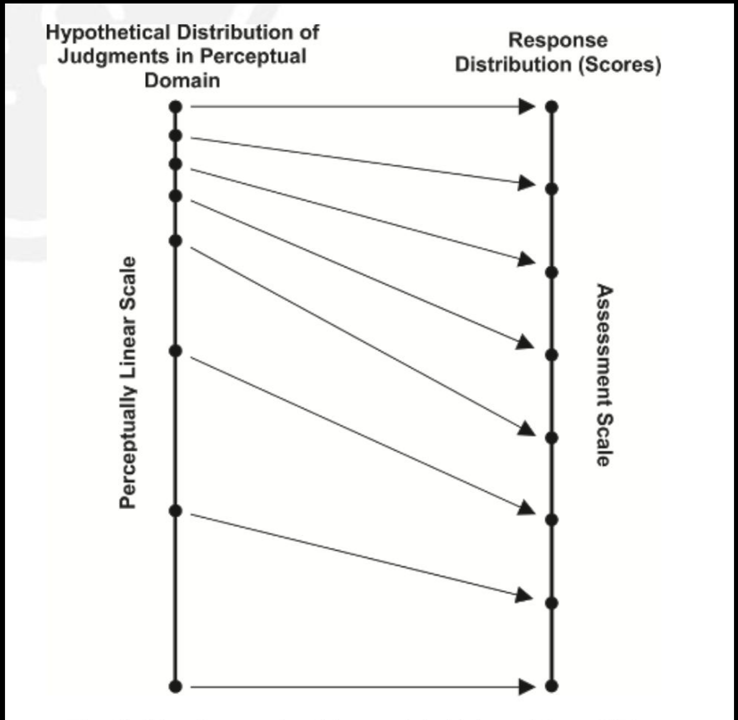
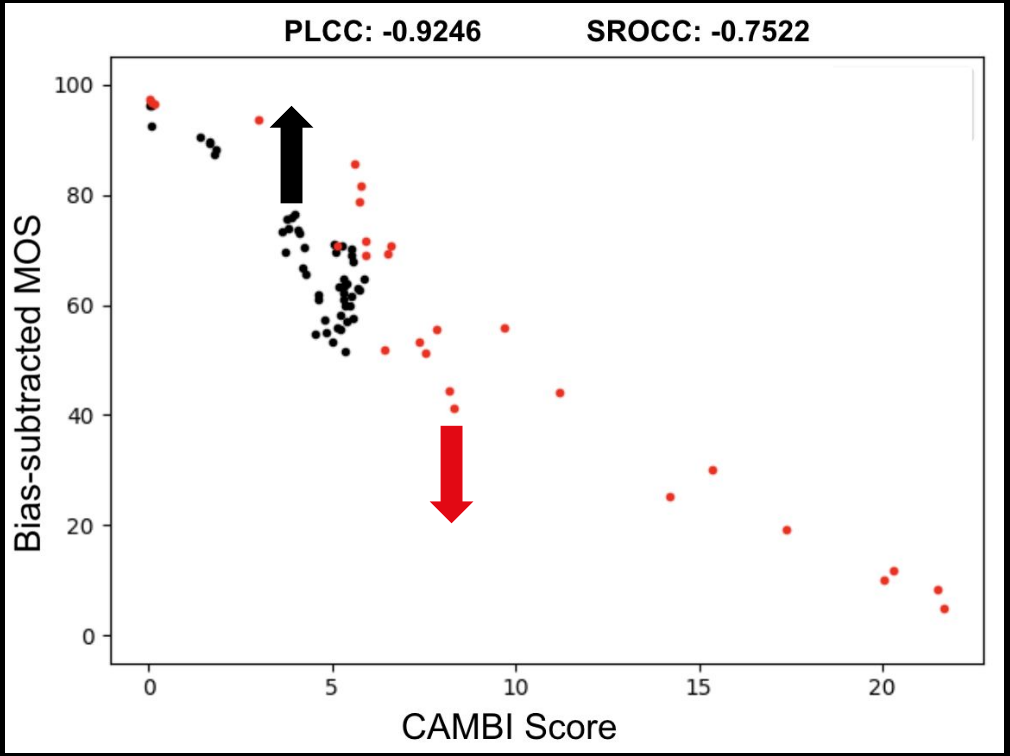


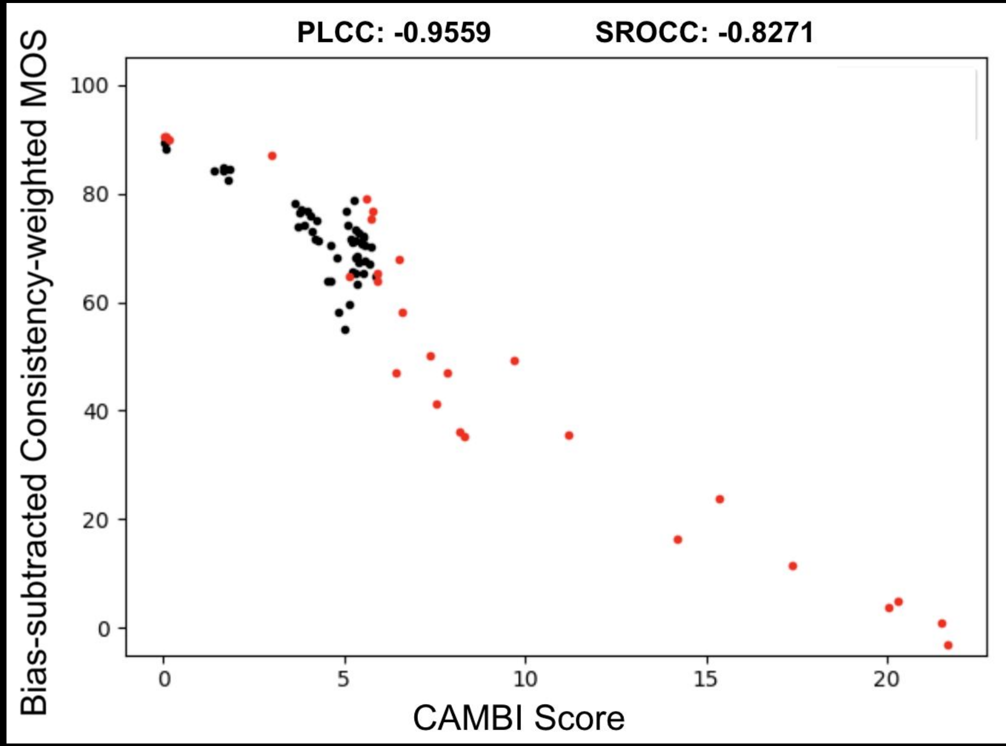
Fig. 5. Stimulus spacing bias model. (Adapted from [8].)

[Zielinski & Rumsey '08]

This encourages stimulus spacing bias, pushing down the scores of the 8-bit encodes.

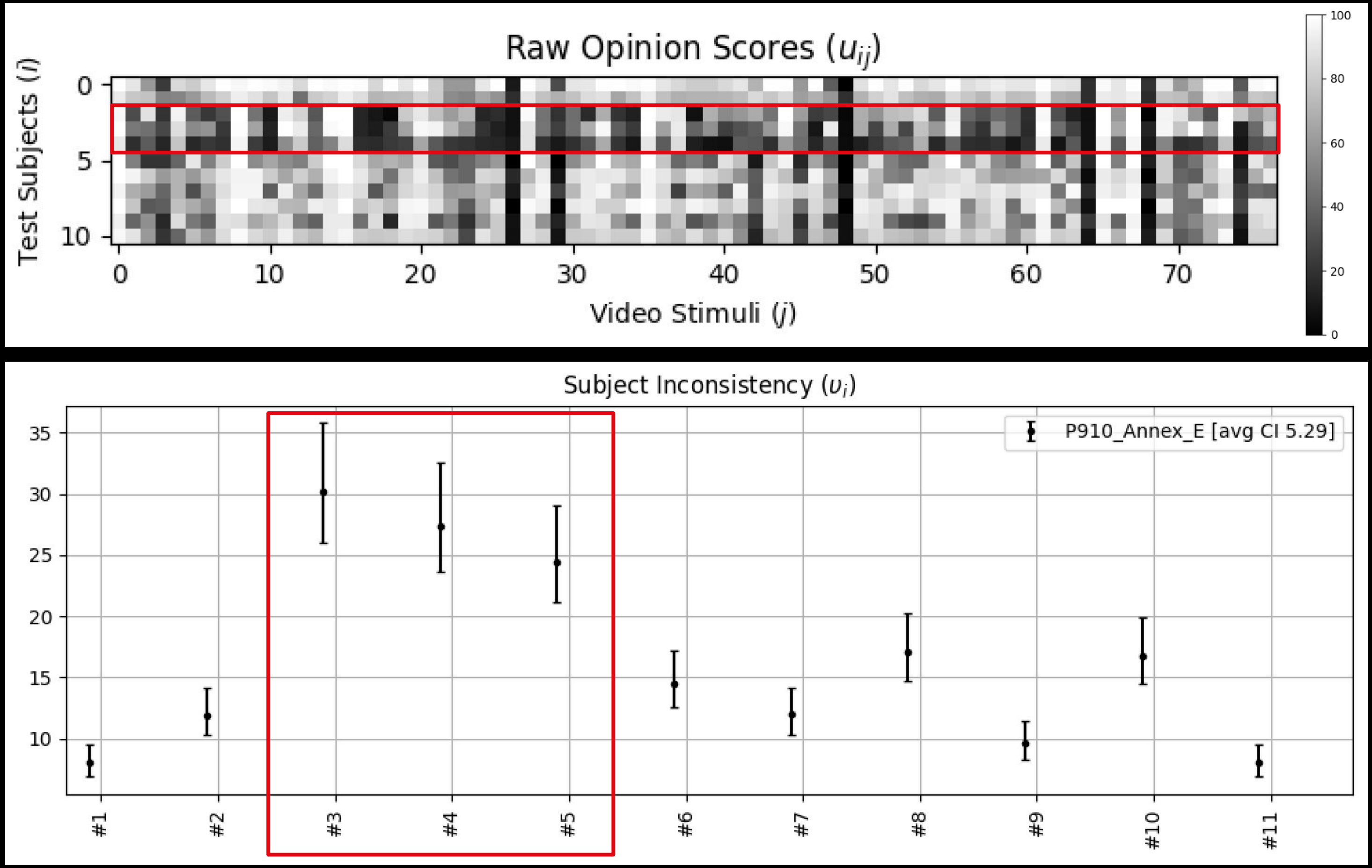
Observation #2: the pure effect of the consistency-weighting is to bring up the 10-bit encodes' scores and bring down the 8-bit encodes' scores. (Coincidentally or not, the correlation between the MOS and the CAMBI scores also improves.)





... How does it manage to achieve this?

This is accomplished by giving unequal weights proportional to subjects' consistency.



Subjects #3, #4, #5 produce scores of large variability (high random error), leading to regression to the mean. Consistency-weighting reduces this effect.

Lessons learned

- Some subjective tests inevitably have perceptually unbalanced stimuli. This could result in stimulus spacing bias, and introduce systematic error and random error to the dataset.
- Applying data analysis technique in P.913 Section 12.6 (or P.910 Annex E) could mitigate the random error introduced, by weighing subjects by their consistency ("soft rejection").
- Because this technique adjusts scores locally, it could not eliminate the systematic error, which is global.

The end

{zli, lkrasula}@netflix.com

Test parameters

- 9 contents, 3 resolutions (4K, QHD, FHD), AV1 encoder, 3 QPs (12, 20, 32)
- CAMBI 8-bit Test: 86 8-bit videos, 23 observers
- CAMBI 10-bit Test: 77 videos (50 10-bit and 27 8-bit videos), 11 observers

Bias-Subtracted MOS - ITU-T P.913 Section 12.4

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.

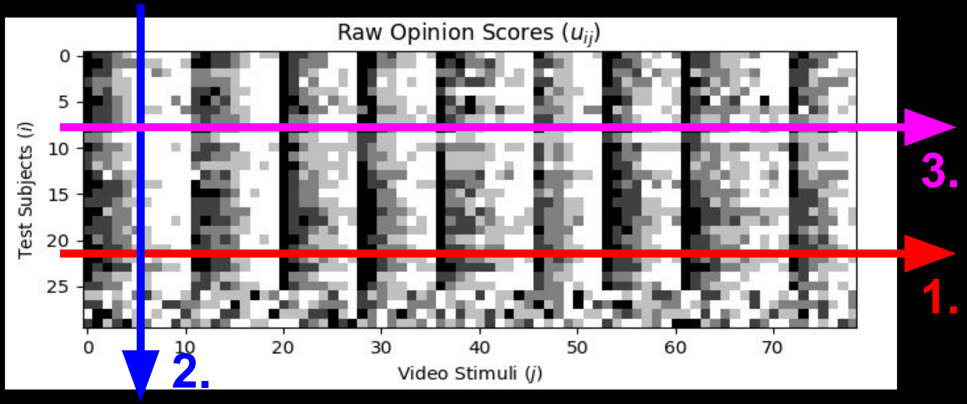
MOS and DMOS are then calculated normally. This normalization does not impact MOS:

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

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)



Bias-Subtracted Consistency-Weighted MOS
- ITU-T P.913 Section 12.6 and ITU-T P.910 Annex E (Prepublished)

- 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: $\{\psi_j\}$, $\{\Delta_i\}$, $\{v_i\}$.

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