Politecnico di Torino

Subjective and objective quality evaluation of high-resolution images

LOHIC FOTIO TIOTSOP

JEG-HYBRID: VQEG MEETING, 15 DECEMBER 2025

Research Context

- **High-resolution content (4K+, low compression) is now widely produced**

- Actual viewing resolution is **not fully controlled** by the producer:
  - Heterogeneous devices (from HD to 4K)
  - Platform-specific downscaling
  - Adaptive bitrate streaming (resolution varies with bandwidth)

- **State-of-the-art IQA models are DNN-based,** but their predictions are not always resolution-invariant [1]
  - Their learned features strongly depend on training-scale statistics

This scenario raises several open research questions

[1]. Wiedemann, O., Hosu, V., Su, S. and Saupe, D., 2023. Konx: cross-resolution image quality assessment. *Quality and User Experience*, *8*(1), p.8.

Research Questions

How should we conduct subjective testing on lightly compressed native high-resolution content?

- **ACR ([2] [3]) or pairwise comparison:** we chose pairwise comparison
- **How to select stimuli for a pairwise comparison test?** We formed **IQA scores-based clusters of images** and chose between and within them
- **What is the appropriate display resolution for evaluation?** We chose Full HD (most common)

How well do native-resolution DNN-IQA scores predict perceived quality on Full HD displays?

- To what extent objective and subjective scores are correlated in the considered scenario?
- Does the computation of IQA score at display resolution improve correlation with human judgments?
- Do humans perceive clear quality differences **between clusters** defined by IQA metrics?
- Do small objective score differences (**within a cluster**) reflect similar perceived quality?

[2] Huang, H., Wan, Q. and Korhonen, J., 2024, April. High resolution image quality database. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3105-3109). IEEE.
[3] Hosu, V., Agnolucci, L., Wiedemann, O., Iso, D. and Saupe, D., 2024, September. UHD-IQA benchmark database: Pushing the boundaries of blind photo quality assessment. In European Conference on Computer Vision (pp. 467-482). Cham: Springer Nature Switzerland.

Dataset & Metrics

22,300 high resolution, lightly compressed images gathered from professional photography websites

- **Resolution:** typically, more than 4k and up to 8k
- **Compression level:** QP>=98

Objective annotation of the dataset

- 11 state-of-the-art DNN-based IQA metrics (From 2014 to 2024)
- Computed at the **native** and at **Full HD** resolution

Objective ranking of image qualities

- Let r_{ij} be the ranking of image i according to metric j
- Note: r_{ij} belongs to [1, 22300]
- We defined:

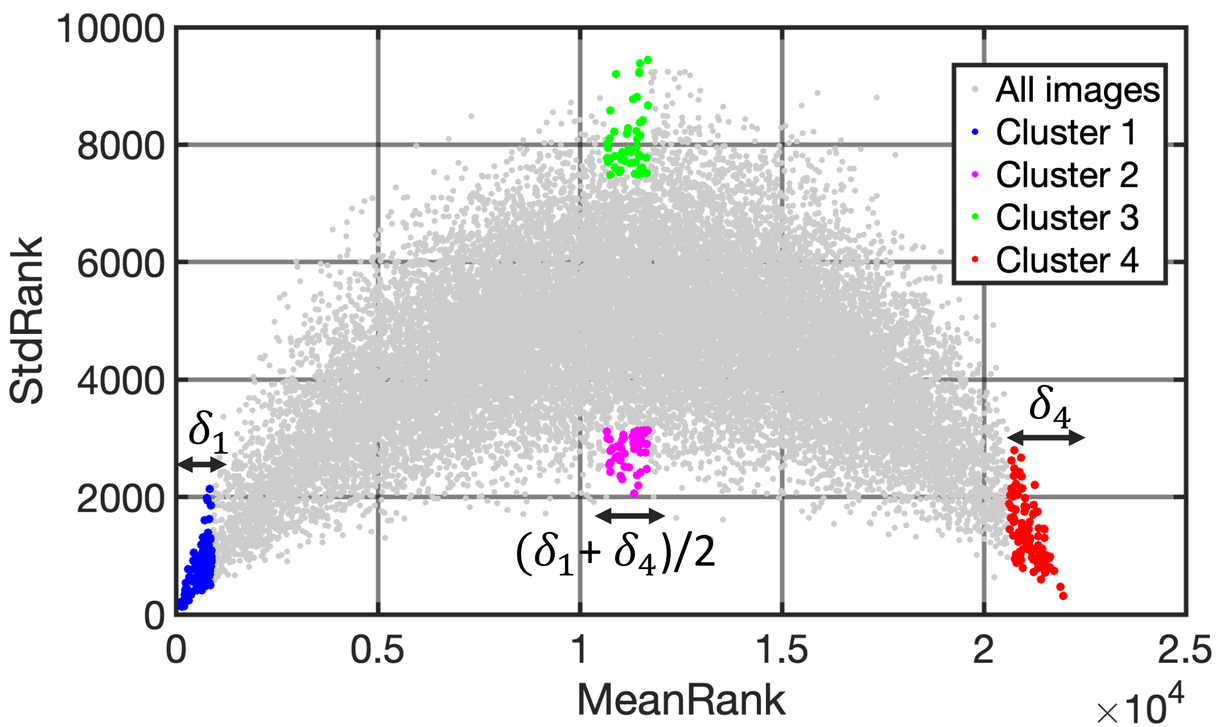
$$\boldsymbol{MeanRank_i = AVG_j(r_{ij})}$$

$$\boldsymbol{StdRank_i = STD_j(r_{ij})}$$

Clusters Definition

4 clusters based on MeanRank and StdRank

- **Cluster 1**: the 100 images with **the smallest MeanRank values**
- **Cluster 2**: 50 images with **middle MeanRank values** and **low StdRank values**
- **Cluster 3**: 50 images with **middle MeanRank values** and **high StdRank values**
- **Cluster 4**: 100 images with the **largest MeanRank values**



Subjective test

A total of 300 stimuli

- **Cluster 1**: 100 images
- **Cluster 2**: 50 images
- **Cluster 3**: 50 images
- **Cluster 4**: 100 images

Image pairs

- a total of **1000 pairs**
- Between and within clusters
- Selected in a way to enable a stable fitting of Thurstone case 5 model

27 subjects have participated so far

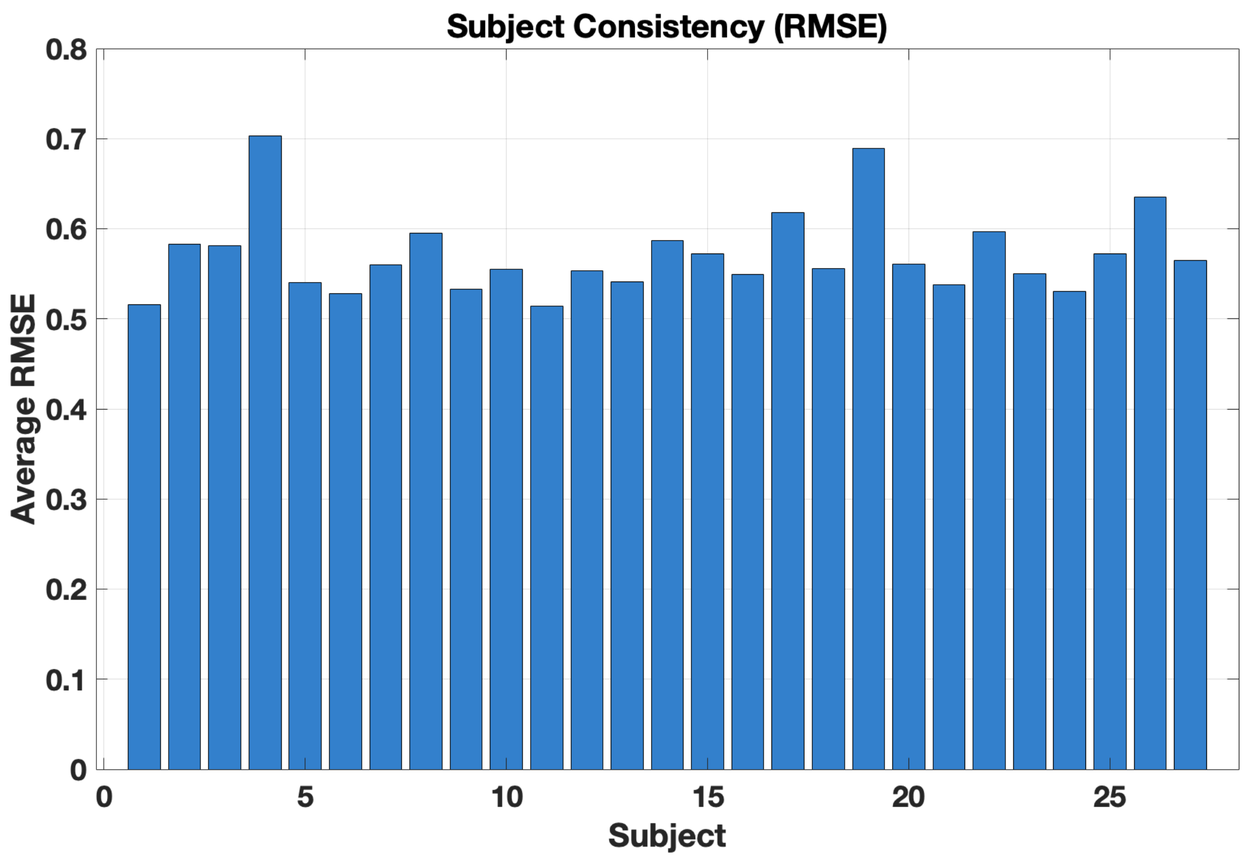
- a total of **300 pairs presented** (different from subject to subject)
- One image shown at a time
- Free iteration between the two images of a pair
- No constraint on decision time
- Choose the one with better quality

50 pairs (Anchors set) presented to all subjects

- Subjects' consistency analysis
- Outliers' detection

Subjects’ consistency on the Anchor set

For each of the 50 pairs, subjects select an image (0 = first, 1 = second). The chart reports the average RMSE between each subject’s choices and all others.

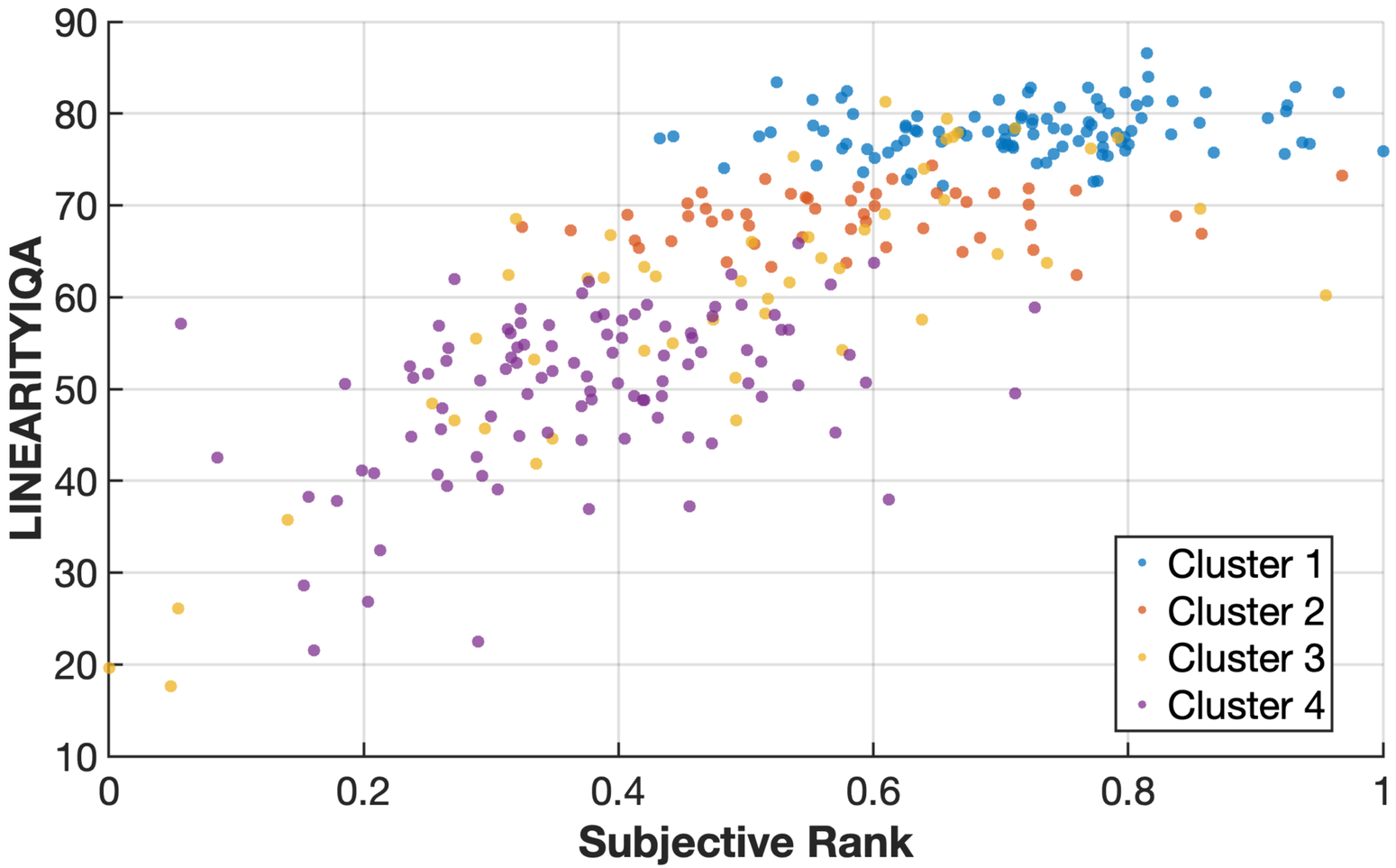


Metrics Accuracy

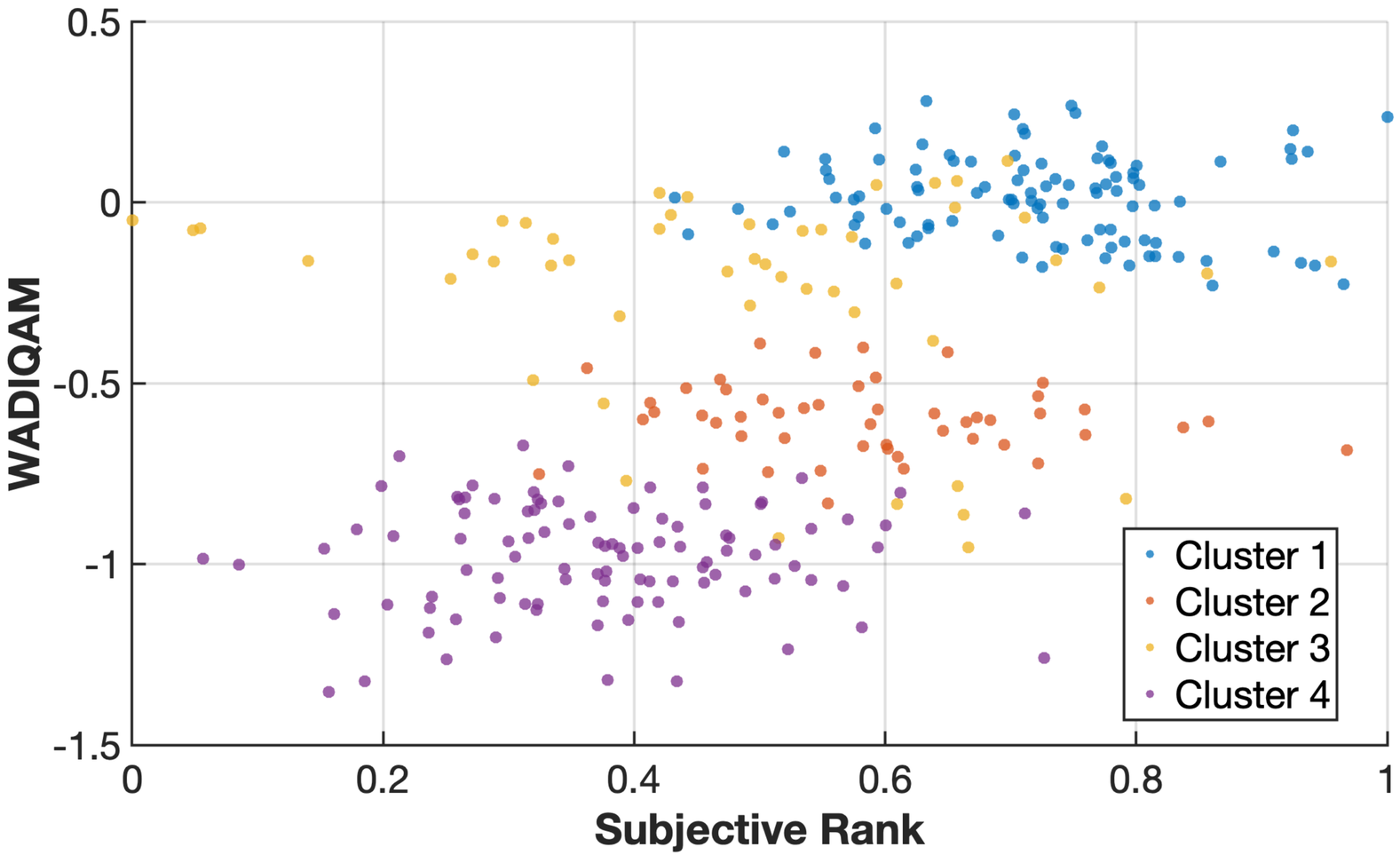
Metric	Pearson	Spearman	Spearman (Scaled)	Pearson (Scaled)
LINEARITYIQA	0.794	0.792	**0.808**	**0.804**
CLIPIQA_	0.690	0.703	**0.756**	**0.746**
TOPIQ	0.688	0.709	**0.759**	**0.756**
CONTRIQUE	0.685	**0.724**	**0.720**	0.698
ARNIQA	**0.677**	**0.677**	0.676	0.662
QUALICLIP	0.660	**0.699**	**0.675**	0.663
DBCNN	0.643	0.661	**0.751**	**0.738**
LIQE	0.638	0.689	**0.749**	**0.736**
HYPERIQA	0.615	0.636	**0.737**	**0.725**
CNNIQA	0.583	0.590	**0.702**	**0.668**
WADIQAM	0.577	0.606	**0.730**	**0.715**

Table 1: Pearson and Spearman correlations between the subjective ranking (Thurstone) and the objective metrics computed at the original resolution and at the subjective-test resolution (Scaled). Bold numbers indicate the best value for each metric.

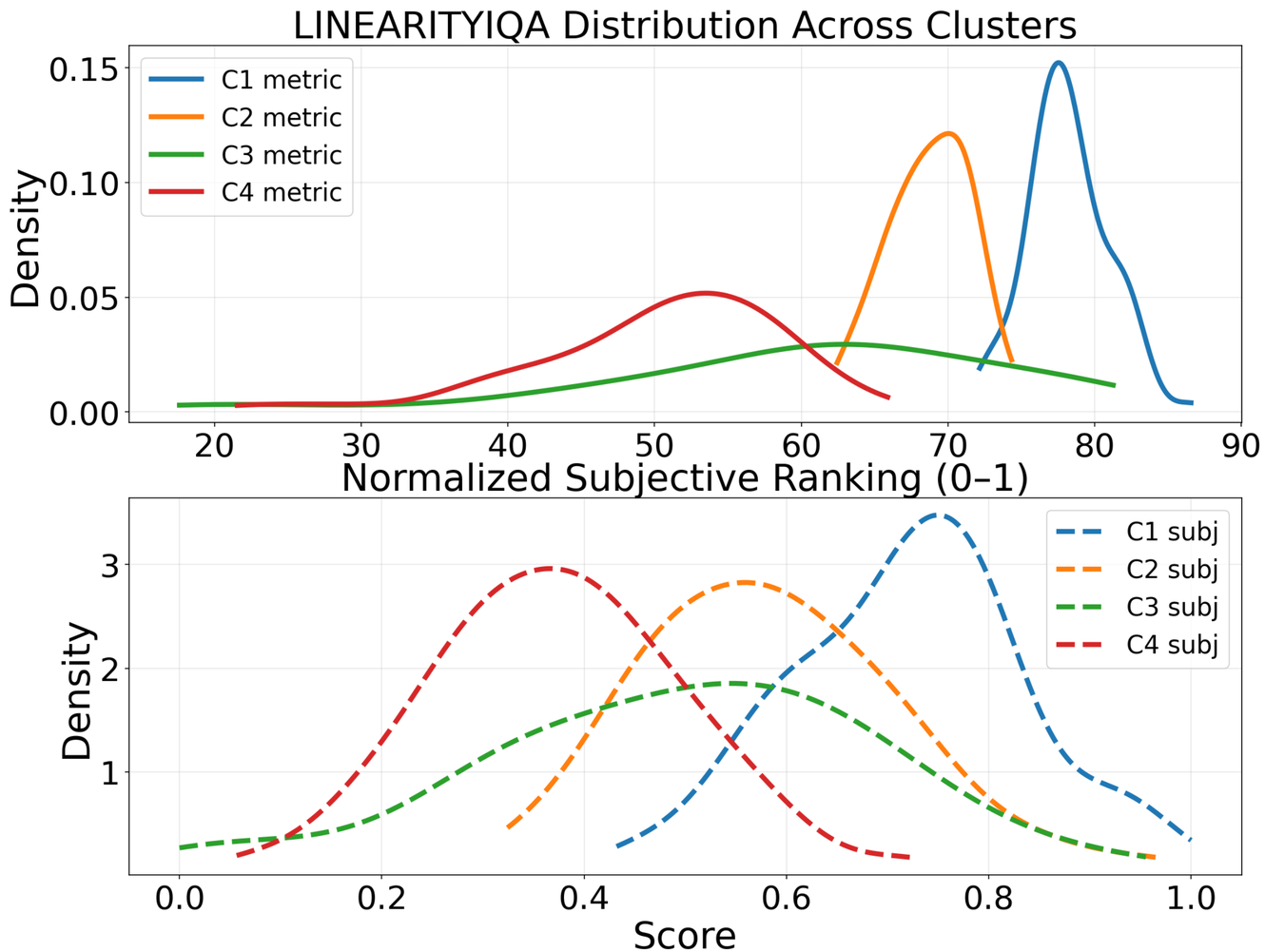
Best Performance



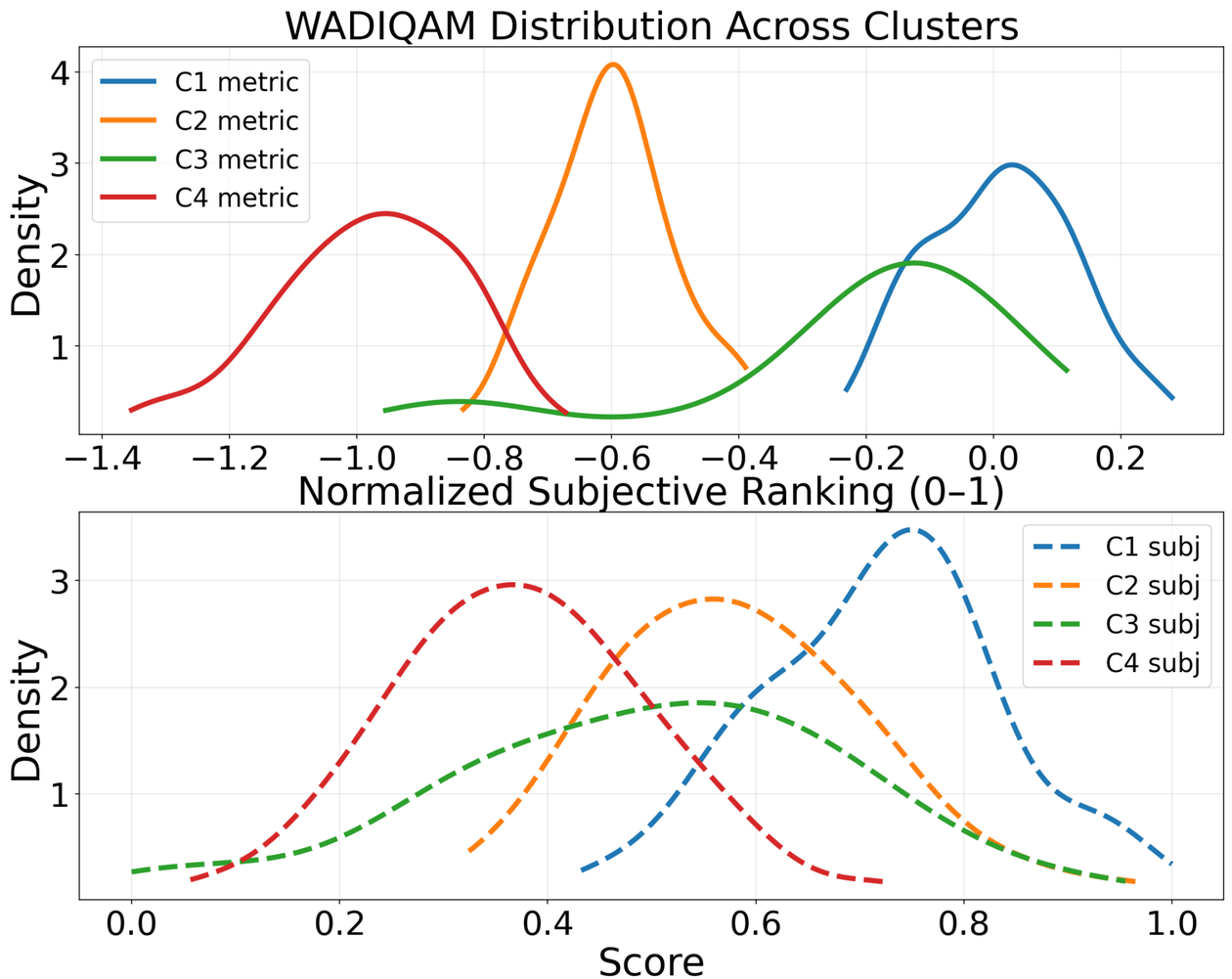
Worst Performance



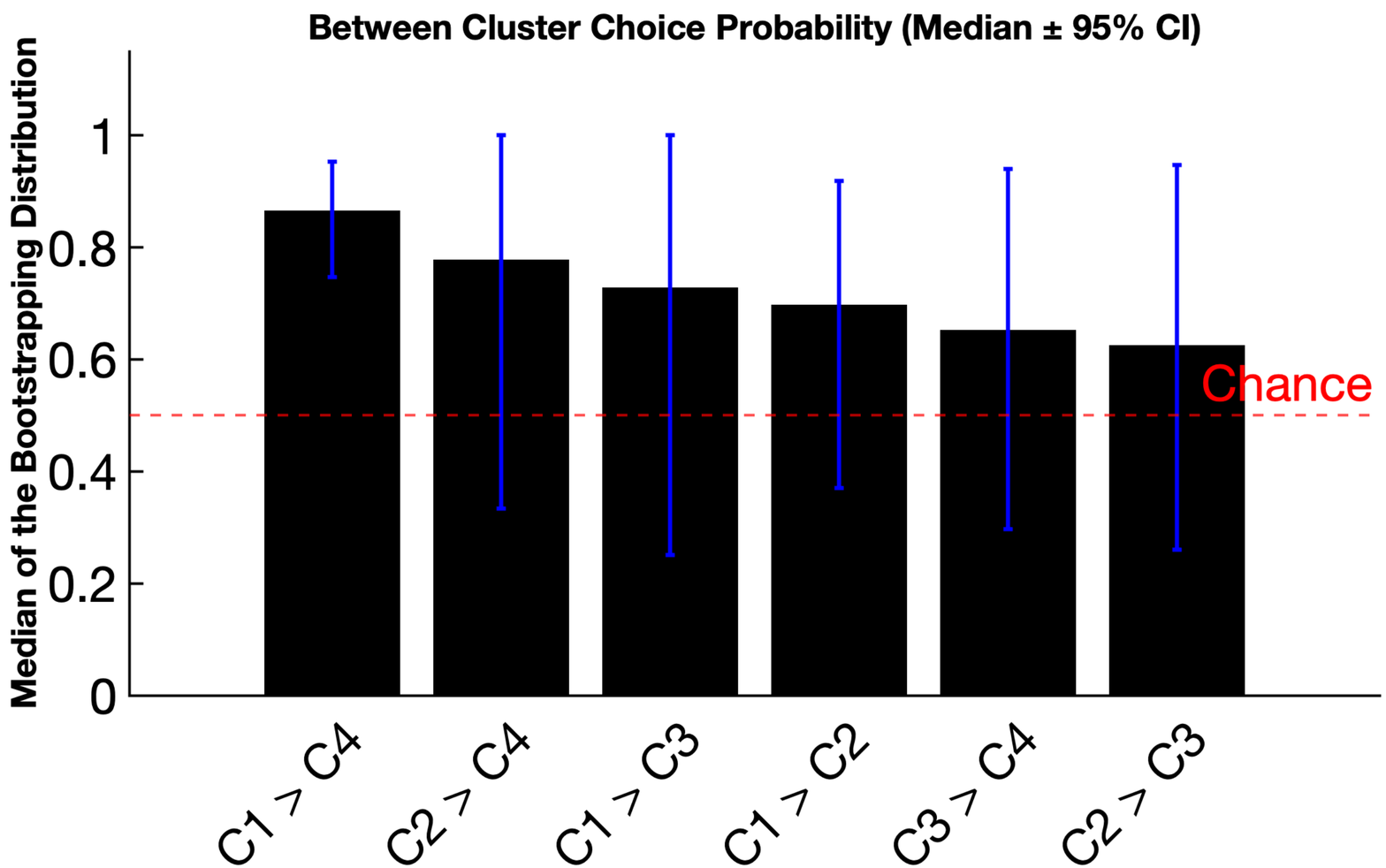
Distribution of Metric Scores in each Cluster



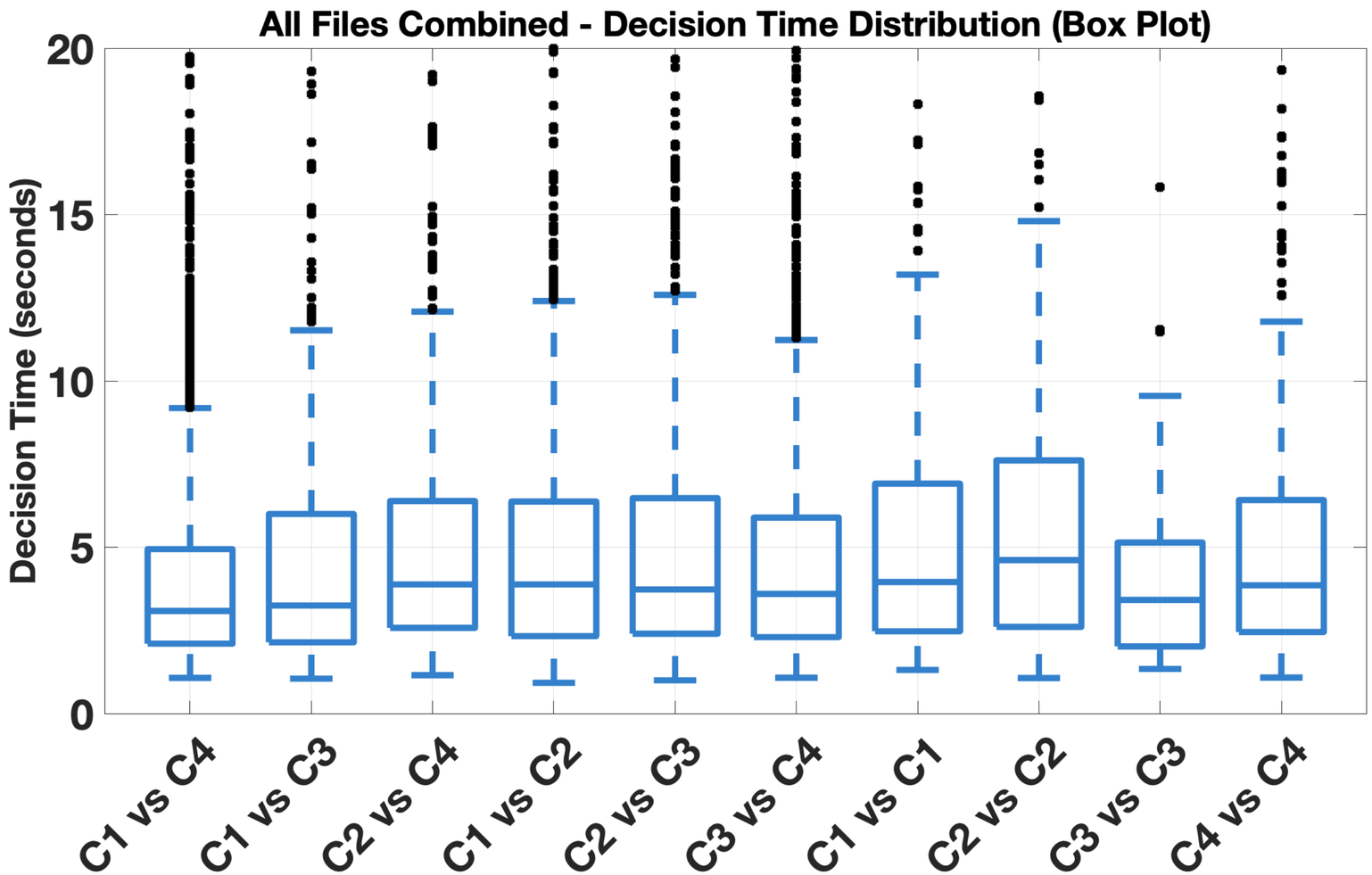
Distribution of Metric Scores in each Cluster



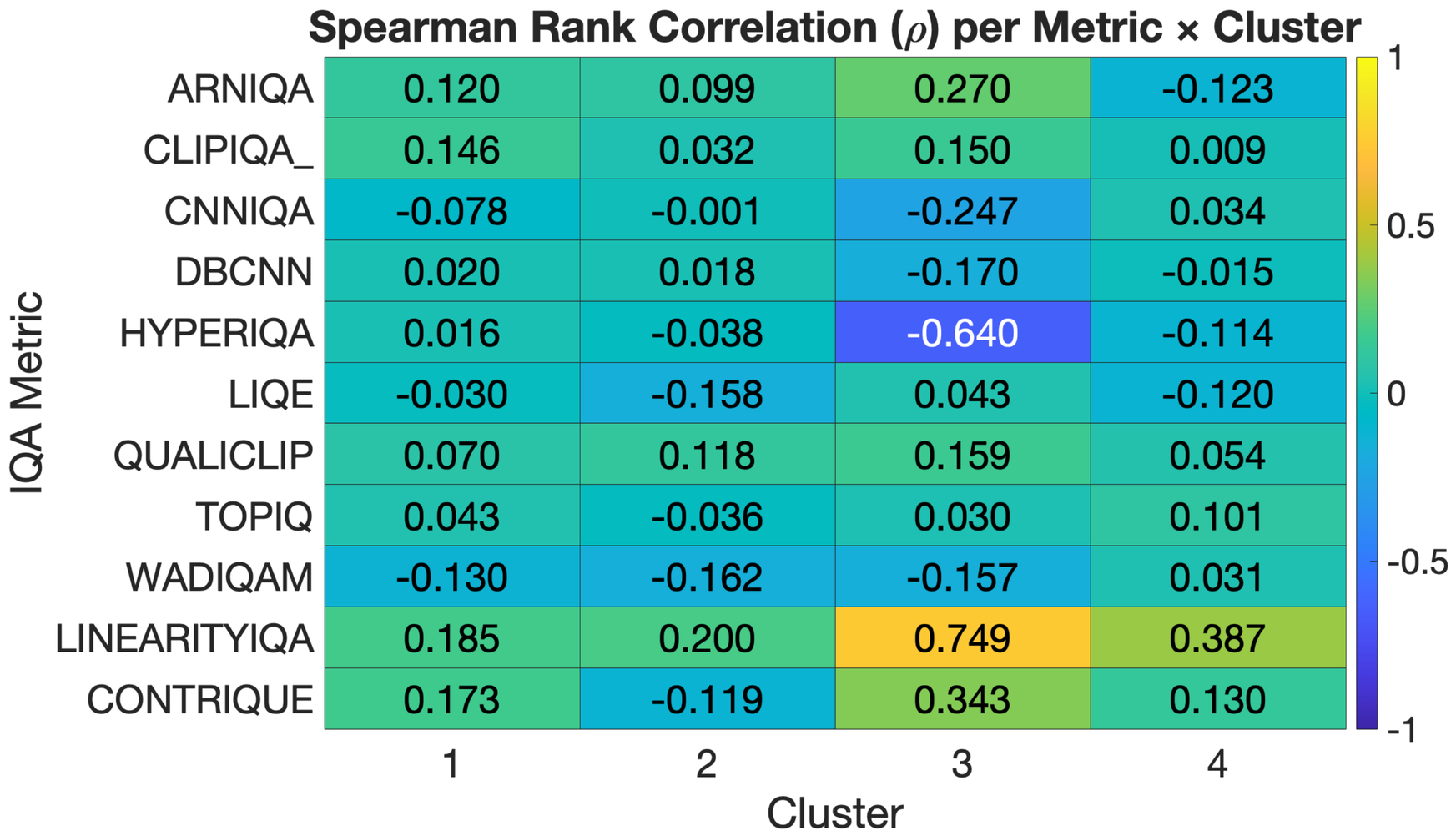
Between Clusters Choice Probabilities



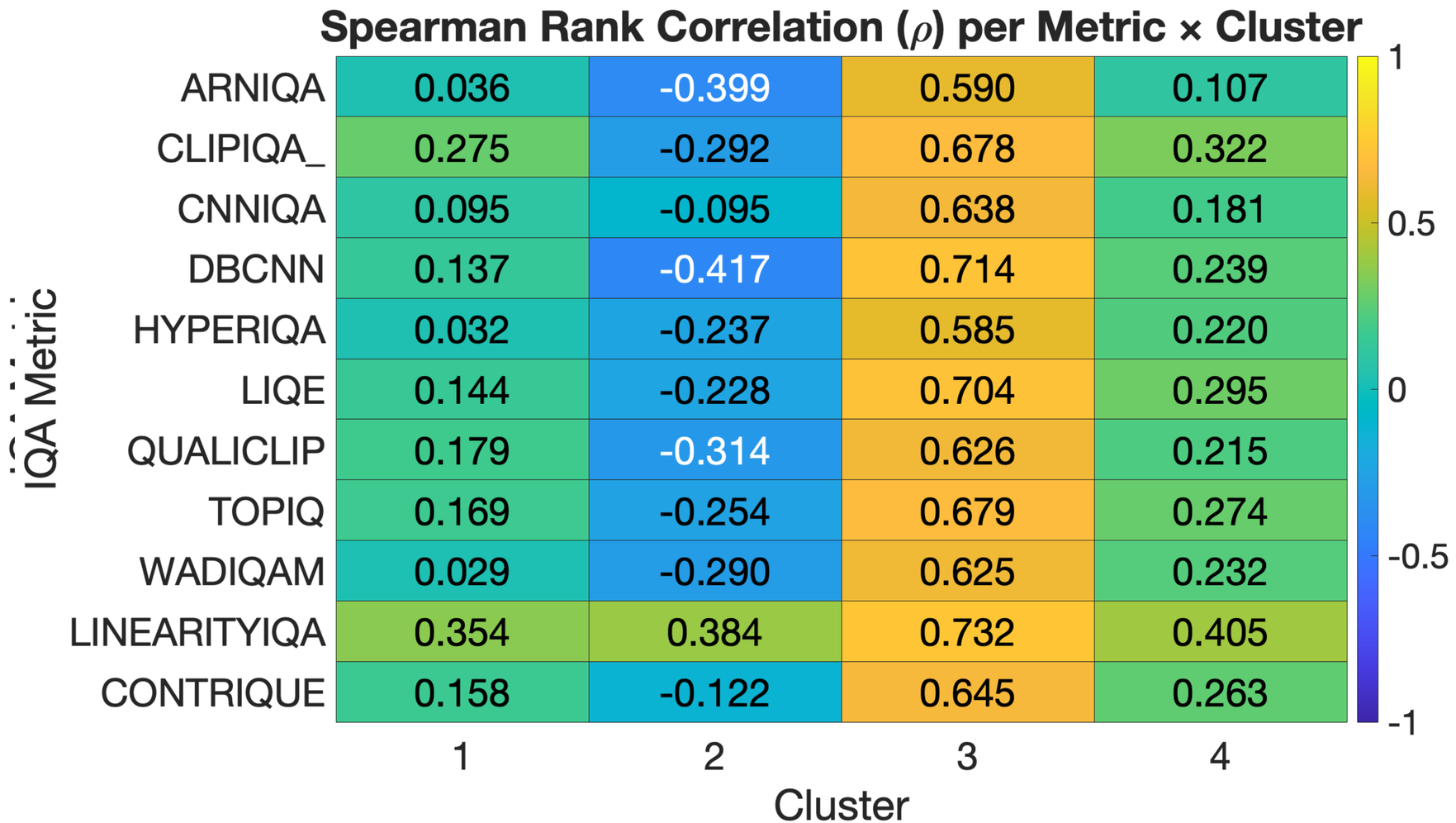
Decision Times



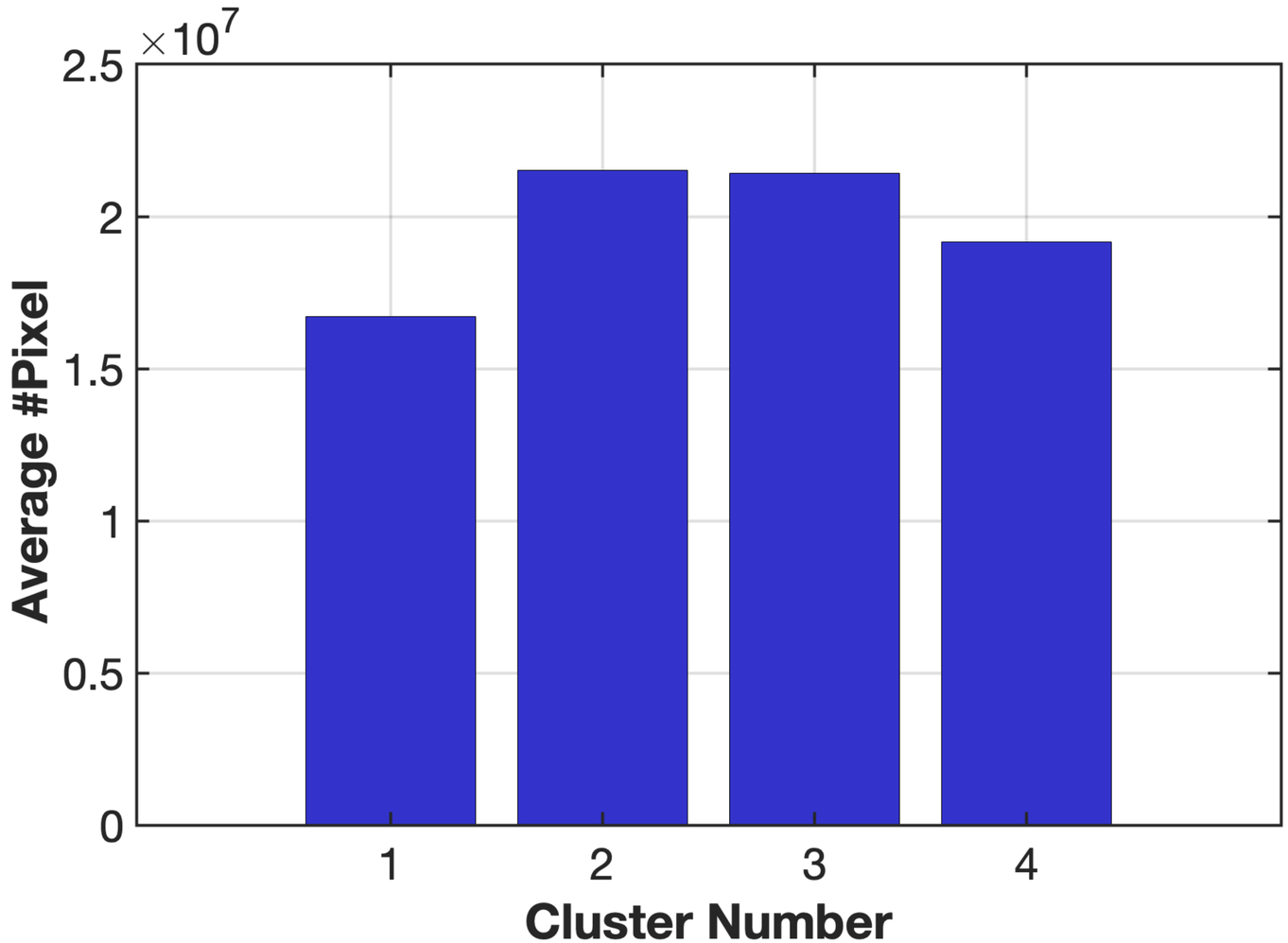
Within-Cluster Metrics Accuracy at native Resolution



Within-Cluster Metrics Accuracy at Full HD Resolution



Average Native Resolution per cluster

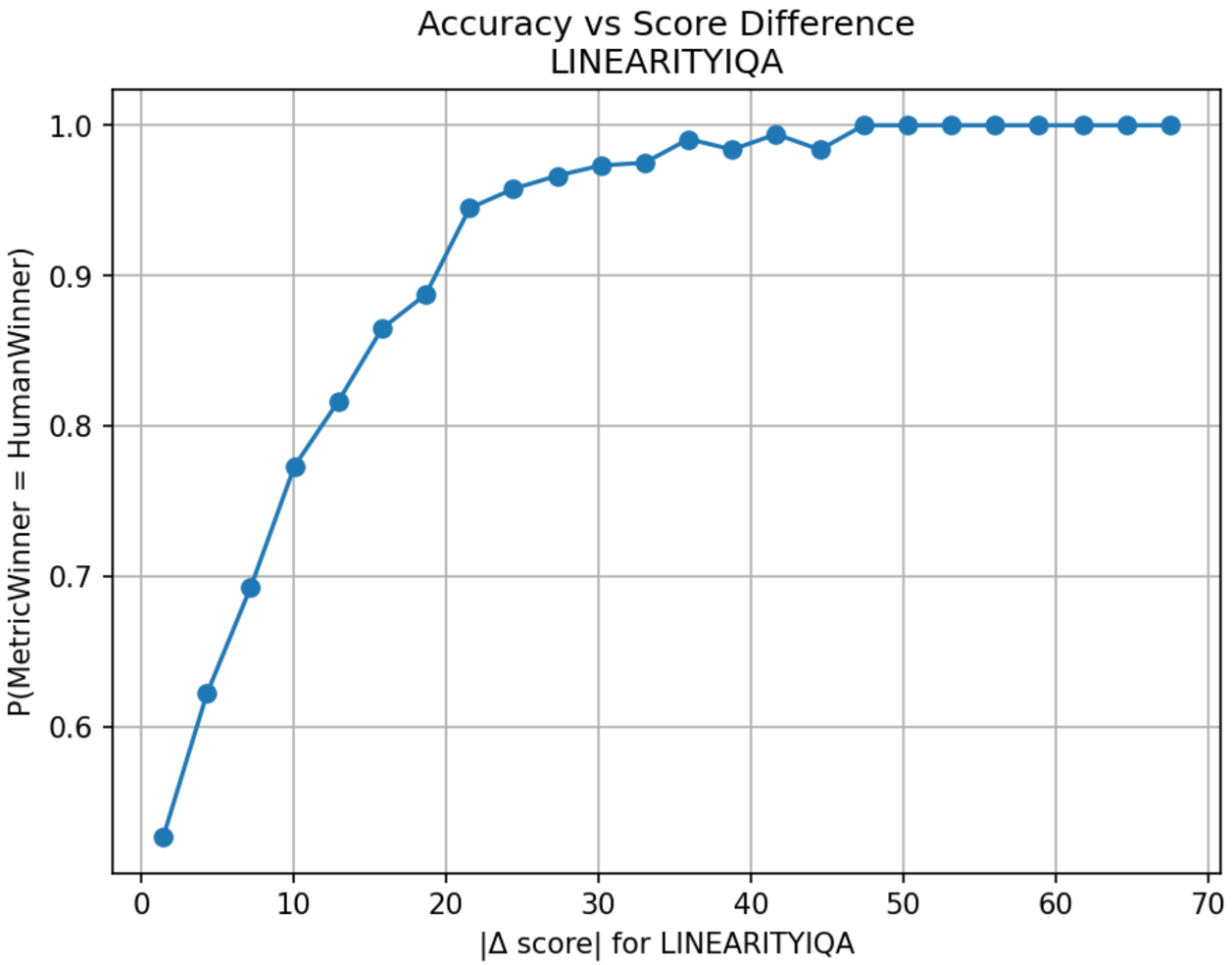


Metric Sensitivity

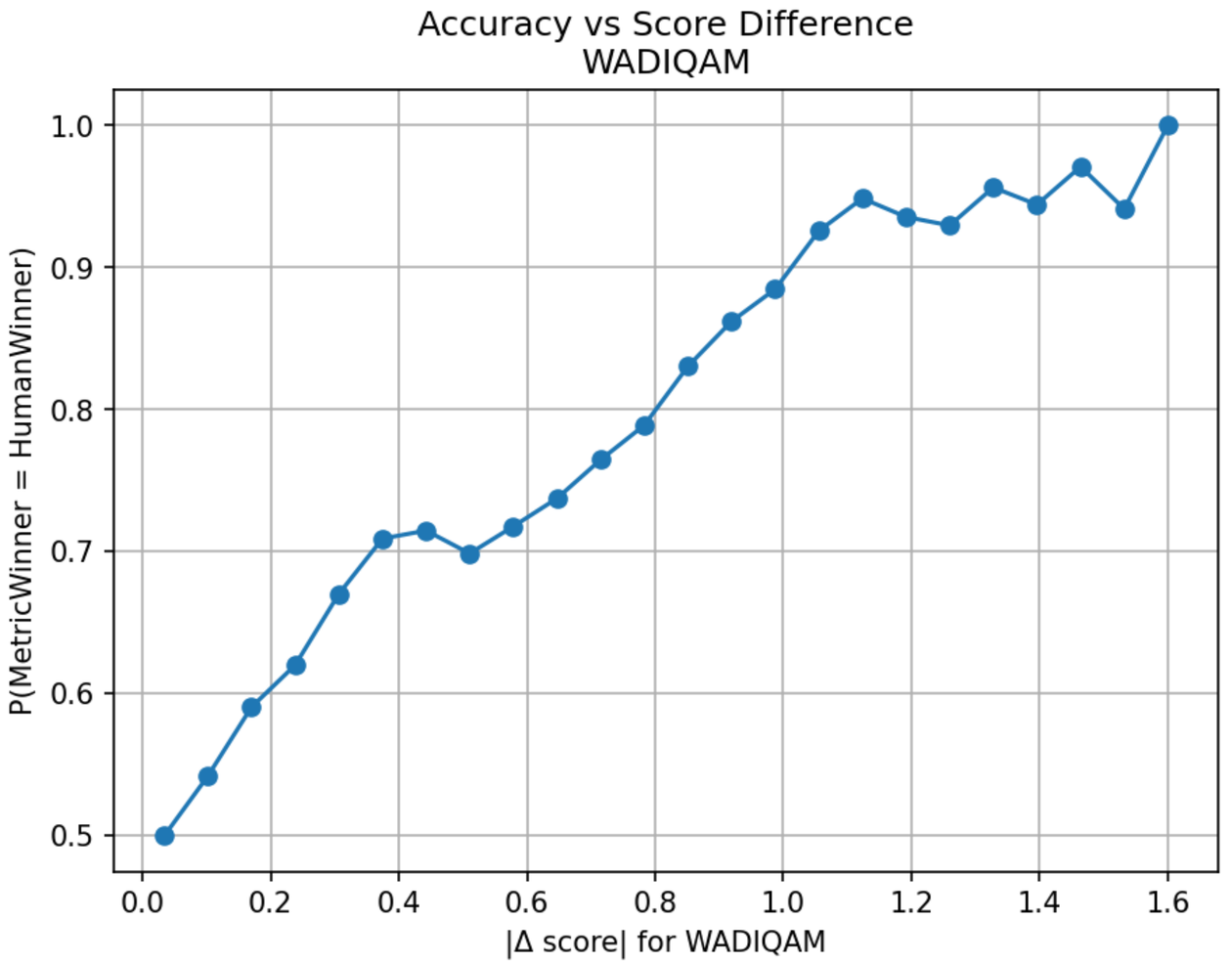
This curve shows how often LINEARITYIQA agrees with **human choices** as a function of the **absolute score difference Δ** between two images.

The accuracy is low and unstable for small Δ, indicating that LINEARITYIQA cannot reliably capture perceptual differences within clusters.

A **Δ>20 is required** for LINEARITYIQA to reach 95% alignment with human judgments. (As a comparison, **Δ is approximately 7 for VMAF in many applications**)



Metric Sensitivity



Conclusions

- **We built a new subjectively annotated dataset for assessing the quality of lightly compressed, high-resolution images viewed at Full HD resolution**
  - This type of dataset is needed in IQA literature

- **The analysis of the collected data shows that:**
  - Most metrics capture only significant quality differences (between clusters), as reflected by the clear gap between the correlations within clusters and the correlation computed on the full dataset
  - Small metric score or rank differences do not guarantee perceptual similarity
  - Display resolution seems to matter for some metrics: Full HD predictions slightly improve metrics' accuracy

- **Ongoing work:**
  - A formal statistical model to manage outliers and measure perceptual differences between and within clusters