



Fine-Grained HDR Image Quality Assessment

Mohsen Jenadeleh, Jon Sneyers, Davi Lazzarotto, Shima Mohammadi,
Dominik Keller, Atanas Boev, António Pinheiro, Thomas Richter,
Alexander Raake, Touradj Ebrahimi, João Ascenso, Dietmar Saupe



Video Quality Experts Group (VQEG)

06 May 2025

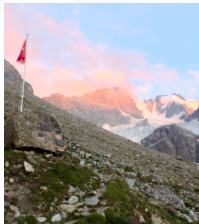
Introduction

- High Dynamic Range (HDR) imaging enhances realism by increasing contrast, brightness, and color gamut far beyond SDR limitations.
- New image compression formats such as JPEG XL, JPEG AI, AVIF, and JPEG XT support HDR, but their performance at high visual fidelity is not well understood.
- Traditional metrics (e.g., PSNR, SSIM) are not reliable for evaluating subtle impairments in HDR content.
- Existing HDR datasets are sparse, lack codec diversity, and rarely probe the near-lossless perceptual quality range.

Objective

We evaluate the JPEG AIC-3 subjective test methodology and scale reconstruction model for fine-grained quality assessment of HDR images.

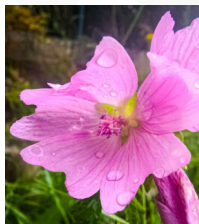
HDR Sources (Tone-mapped Previews)



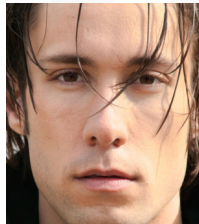
alps-sunset



building



flower



p29-crop



room1

*Tone-mapped SDR previews of the HDR source images used in the subjective experiment.
All images have a resolution of 840×944 and are encoded in Rec.2100 PQ.*

Dataset Overview

- Source images: 840×944 , Rec.2100 PQ color space, 10-bit RGB.
 - The sources were selected from 23 candidates by visual inspection by JPEG AIC group.
- Encoding recipes: (after consulting with the respective developers)
 - JPEG XL: libjxl v0.11.1, `cjxl -q Q`
 - JPEG AI: verification model v7.0, 10-bit yuv444, all coding tools enabled
 - AVIF: libavif v1.1.1, libaom 3.11.0, `avifenc --autotiling -d 10 --cicp 9/16/9 -q Q`
 - JPEG XT: libjpeg v1.70, `jpeg -h -qt 3 -oz -R 2 -q Q`
- 5 compression levels per codec per source.
 - 5 bitrates were selected based on visual inspection of the compressed images.
 - Matching bitrates for JPEG XL, JPEG AI, AVIF
 - Higher bitrates for JPEG XT (to match the visual quality range)
- Total: 100 HDR test images; represented as 16-bit PNG files with an ICC v4.4 profile with cicp tag for Rec.2100 PQ so they can be viewed in Chrome

Subjective Test Methodology

JPEG AIC-3 Methodology*:

- Two triplet comparison types:
 - **Plain Triplet Comparisons (PTC)** — side-by-side, with toggle against source
 - **Boosted Triplet Comparisons (BTC)** — 2x zoom, 10 Hz flicker with source
- Each source image: 120 same-codec + 24 (20%) cross-codec triplets per codec
 - For 5 sources for BTC test: 720 triplets
 - For 5 sources for PTC test: 720 triplets
- The triplets for each experiment was divided to 6 batches of 120 questions each.
- Each participant was allowed to complete maximum 2 batches of 120 questions.

**ISO/IEC DIS 29170-3, "Information technology — advanced image coding and evaluation — Part 3: Subjective quality assessment of high-fidelity images", 2025.*

Viewing Conditions and Display Configuration

- All labs were calibrated in compliance with:
 - **ITU-R BT.2100-3**: HDR reference viewing conditions (PQ, BT.2020)
 - **ITU-R BT.2246-8**: Ambient light, display luminance, and screen setup for critical viewing
- Ambient illumination set to **5 lux**
- Viewing distance: **$1.5\times$ – $3.1\times$ stimuli height**, depending on BTC/PTC protocol.

HDR Displays and Configuration

- **Operating system:** macOS, version 15.0 or newer
- **Three lab (Konstanz, EPFL, and IT)** used 16-inch MacBook Pro (2021 or later) with Liquid Retina XDR display:
 - Native resolution: 3456×2234 , peak brightness: 1000 cd/m^2 , SDR white: 100 cd/m^2
 - Color profile: HDR P3, refresh rate: 60 Hz
- **One lab (TU Ilmenau)** used Sony BVM-HX310 mastering monitor:
 - Resolution: 3840×2160 , connected via HDMI 2 to Mac mini
 - Refresh rate: 30 Hz (to avoid chroma subsampling), same luminance settings
- **Browser:** Google Chrome version 134.x
- **Image display:** Images shown 1:1 (image pixel to native display resolution)

Triplet Response Statistics

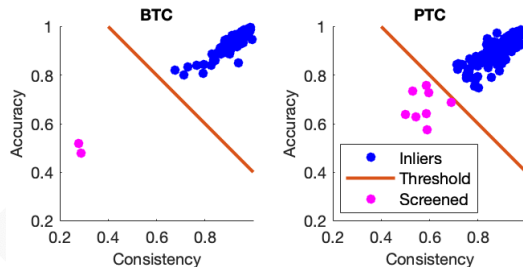
- **Participants:** 151 total, 4 labs in 3 countries.
- Each lab collected **25%** of the responses from unique participants for BTC and PTC.
- The average duration per batch was 20.1 minutes for PTC and 10.7 minutes for BTC.

Type	Measure	Distortion level difference				
		1	2	3	4	5
BTC	Avg. time (s)	5.7	4.7	4.1	3.5	3.1
	Rsp. ratio correct	0.637	0.865	0.942	0.977	0.985
	Rsp. ratio not sure	0.316	0.118	0.044	0.012	0.001
	Rsp. ratio incorrect	0.080	0.052	0.028	0.024	0.006
PTC	Avg. time (s)	10.8	9.4	8.1	6.8	5.9
	Rsp. ratio correct	0.560	0.720	0.830	0.925	0.977
	Rsp. ratio not sure	0.359	0.228	0.143	0.051	0.017
	Rsp. ratio incorrect	0.080	0.052	0.028	0.024	0.006
Response counts		Left	Not sure	Right	Skip	Total
BTC		7,838	2,405	7,019	18	17,280
PTC		6,708	3,547	6,889	136	17,280

Response accuracy increases with greater distortion level differences.

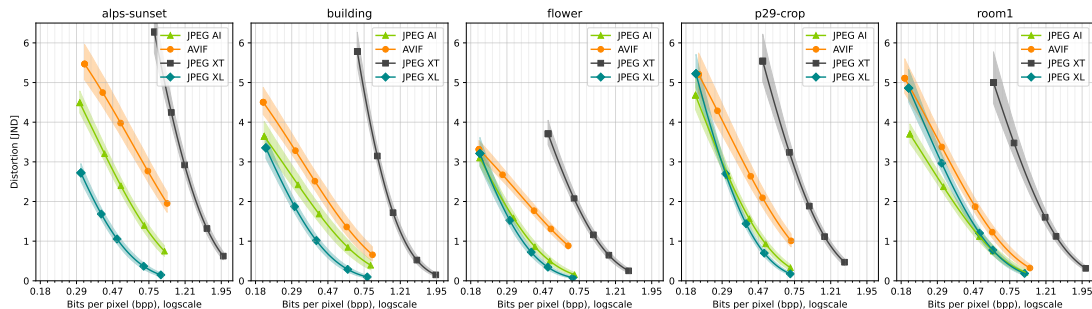
Data cleansing

- Total ratings collected: 34,560 (17,280 BTC + 17,280 PTC)
- 144 batches for each experiment type
- Each triplet received 24 responses; mirrored pairs included to measure consistency
- **Batch filtering:** low-quality batches removed based on:
 - Accuracy (based on known distortion direction)
 - Consistency (symmetry agreement in mirrored triplets)
- **Retained:** 142 BTC batches, 134 PTC batches



Bitrate–Distortion Results

- JPEG AIC-3 scale reconstruction method is used to reconstruct the quality scale in JND unites (Model details in Dietmar's talk #118).
- Maximum likelihood estimation was used to fit a unified model consisting of an exponential rate-distortion function and a quadratic boosting transform to the BTC and PTC responses simultaneously.



Perceived distortion in JND units. 95% confidence intervals via 1000× bootstrap.

Objective Metric Correlations – SDR Metrics

Metric	Overall		Per-codec		Per-source	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
PSNR-Y (PQ)	0.587	-0.596	0.741	-0.792	0.730	-0.733
PSNR-Y after TM	0.612	-0.611	0.698	-0.701	0.813	-0.816
SSIM	0.674	-0.672	0.720	-0.713	0.876	-0.889
SSIM after TM	0.544	-0.533	0.635	-0.638	0.770	-0.767
MS-SSIM (PQ)	0.757	-0.731	0.764	-0.749	<u>0.918</u>	-0.907
MS-SSIM after TM	0.679	-0.669	0.709	-0.707	0.905	-0.907
PSNR-HVS (PQ)	0.520	-0.526	0.771	-0.798	0.635	-0.639
PSNR-HVS after TM	0.731	-0.748	0.824	-0.831	0.829	-0.861
IW-PSNR (PQ)	0.658	-0.670	0.826	-0.875	0.728	-0.734
IW-PSNR after TM	0.805	-0.817	0.853	-0.859	0.900	<u>-0.914</u>
IW-SSIM (PQ)	0.647	-0.658	0.775	-0.806	0.773	-0.780
IW-SSIM after TM	0.676	-0.680	0.709	-0.704	0.885	-0.892
VMAF (PQ)	0.584	-0.585	0.740	-0.793	0.688	-0.691
VMAF after TM	0.678	-0.680	0.760	-0.784	0.801	-0.791
VMAF-neg (PQ)	0.599	-0.604	0.760	-0.816	0.714	-0.718
VMAF-neg after TM	0.713	-0.721	0.786	-0.808	0.834	-0.837

*SDR metrics include versions with and without tone-mapping (TM).
Top 5 values are underlined.*

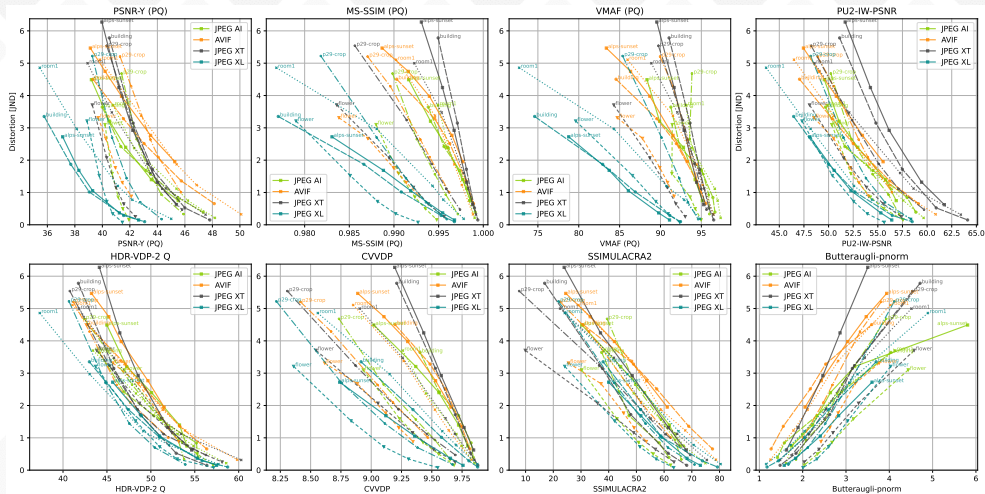
Objective Metric Correlations – HDR Metrics

Metric	Overall		Per-codec		Per-source	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
LGFM	0.462	-0.421	0.542	-0.536	0.596	-0.747
PU2-PSNR	0.622	-0.621	0.723	-0.734	0.798	-0.792
PU2-IW-PSNR	<u>0.813</u>	<u>-0.828</u>	<u>0.895</u>	<u>-0.908</u>	0.864	-0.873
PU2-SSIM	0.677	-0.676	0.720	-0.725	0.878	-0.889
PU2-MS-SSIM	0.770	-0.743	0.773	-0.758	<u>0.922</u>	<u>-0.909</u>
PU2-IW-SSIM	0.757	-0.768	0.859	-0.882	0.819	-0.837
HDR-VDP-2 (quality)	0.936	-0.946	0.963	-0.965	<u>0.955</u>	<u>-0.949</u>
HDR-VDP-3 (detect)	0.742	-0.744	0.786	-0.783	0.889	-0.899
CVVDP	0.736	-0.753	0.782	-0.784	0.883	-0.900
Butteraugli-max	<u>0.836</u>	<u>0.845</u>	<u>0.878</u>	<u>0.889</u>	0.888	0.881
Butteraugli-pnorm	<u>0.882</u>	<u>0.886</u>	<u>0.910</u>	<u>0.901</u>	<u>0.939</u>	<u>0.947</u>
SSIMULACRA 1	0.546	0.512	0.569	0.522	0.819	0.892
SSIMULACRA 2	<u>0.906</u>	<u>-0.895</u>	<u>0.907</u>	<u>-0.888</u>	0.968	-0.958

Metrics tailored to HDR or perceptually uniform spaces.

*Top 5 values are underlined; best overall in **bold**.*

Objective Metrics vs. JND (subset)



Metric predictions vs. subjective JND scores. Line styles: source images; Colors: codecs.

Statistical Comparison of Metrics

- The Meng–Rosenthal–Rubin (MRR) test evaluates whether two metrics differ significantly in their correlation with subjective scores. It tests the null hypothesis:

$$H_0 : r_{XZ} = r_{YZ}$$

where r_{XZ} and r_{YZ} are correlations of metrics X and Y with the subjective scores Z .

- The test statistic is:

$$Z = (z_1 - z_2) \left(\frac{2(1 - r_{XY})h}{n - 3} \right)^{-0.5}$$

where:

$$z_1 = \tanh^{-1}(r_{XZ}), \quad z_2 = \tanh^{-1}(r_{YZ})$$

$$h = \frac{1 - f \cdot \bar{r}^2}{1 - \bar{r}^2}, \quad f = \frac{1 - r_{XY}}{2(1 - \bar{r}^2)}, \quad \bar{r}^2 = \frac{r_{XZ}^2 + r_{YZ}^2}{2}$$

- We use a two-tailed test with significance level $\alpha = 0.05$ to determine if differences are statistically meaningful.

**Xiao-Li Meng, Robert Rosenthal, and Donald B Rubin, "Comparing correlated correlation coefficients.," Psychological Bulletin, vol. 111, no. 1, pp. 172, 1992.*

MRR Statistical Comparison Matrix

	HDR_VDP_2	SSIMULACRA2	Butteraugli_pnorm	Butteraugli_max	PU2_W_PSNR	W_PSNRAfterTM	PU2_W_SSIM	W_SSIM_PO_	HDR_VDP_3_detect_	MS_SSIM_MatlabAfterTM	PSNR_HVSAfterTM	PU2_MS_SSIM	CVVDP	VMAF_negAfterTM	W_SSIMAfterTM	MS_SSIM_Matlab_PO_	VMAFAfterTM	W_PSNR_PO_	SSIM_MatlabAfterTM	PU2_SSIM	SSIM_libvmafAfterTM	SSIM_Matlab_PO_	PSNR_YAfterTM	PU2_PSNR	SSIM_libvmaf_PO_	VMAF_neg_PO_	PSNR_Y_PO_	VMAF_PO_	MS_SSIM_libvmafAfterTM	MS_SSIM_libvmaf_PO_	SSIMULACRA1	PSNR_HVS_PO_	LQFM
HDR_VDP_2	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SSIMULACRA2	-1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Butteraugli_pnorm	-1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Butteraugli_max	-1	-1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PU2_W_PSNR	-1	-1	-1	0	0	0	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
IW_PSNRAfterTM	-1	-1	-1	0	0	0	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PU2_W_SSIM	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
IW_SSIM_PO_	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
HDR_VDP_3_detect_	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1
MS_SSIM_MatlabAfterTM	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	1	1	1	1	1	1	1	1	1
PSNR_HVSAfterTM	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1
PU2_MS_SSIM	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1
CVVDP	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1
VMAF_negAfterTM	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
IW_SSIMAfterTM	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
MS_SSIM_Matlab_PO_	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VMAFAfterTM	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
IW_PSNR_PO_	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SSIM_MatlabAfterTM	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
PU2_SSIM	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SSIM_libvmafAfterTM	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SSIM_Matlab_PO_	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PSNR_YAfterTM	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PU2_PSNR	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SSIM_libvmaf_PO_	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VMAF_neg_PO_	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PSNR_Y_PO_	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VMAF_PO_	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MS_SSIM_libvmafAfterTM	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MS_SSIM_libvmaf_PO_	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SSIMULACRA1	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PSNR_HVS_PO_	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LQFM	-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Each cell shows whether the row metric has a significantly higher (+1), significantly lower (−1), or no significant difference (0) in its correlation with subjective scores compared to the column metric. Based on MRR test.

Conclusion

- The study demonstrates that JPEG AIC-3 methodology enables accurate subjective quality assessment for HDR content even in the visually lossless range.
- The multi-lab evaluation confirms the robustness of the approach, with precise JND estimation (average 95% CI width of 0.27 at 1 JND).
- The resulting HDR dataset spans visible to imperceptible distortions and is suitable for benchmarking HDR-capable codecs and metrics.
- Evaluation of objective metrics revealed that while SDR metrics are not suitable for HDR content even with tone mapping, some HDR-specific and perceptually optimized models perform well (e.g., HDR-VDP-2). However, even the best metrics still have some source or codec bias, e.g. overestimating the distortion for the `flower` image or for JPEG XL.
- Therefore, careful subjective evaluation is still required for reliable codec evaluation and quality assessment.

Future work

- Is there a significant difference between responses from different fully controlled labs?
- We also ran the experiment in **partially** controlled environments (**Clouldinary**, **HUAWEI**) to test reproducibility outside the lab, but results aren't analyzed yet due to few participants.

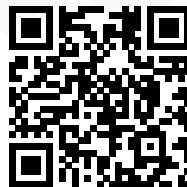
Resources and Dataset

HSR image dataset and code will be available upon paper acceptance:

<https://github.com/jpeg-aic/>

The release will include:

- Source and compressed images
- Triplet responses (plain and boosted)
- JND reconstructions and metric scores
- Statistical analysis scripts



Scan to access repository

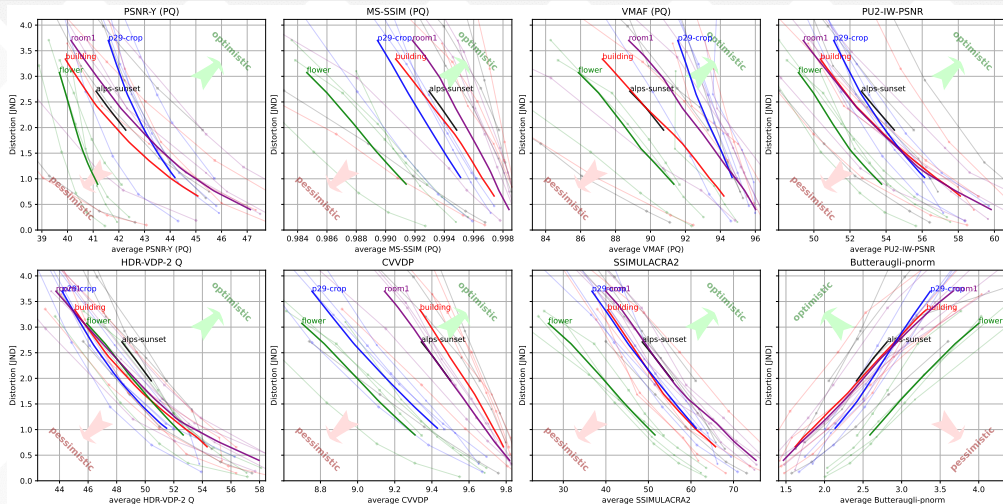
Thank you for your attention!

Target Bitrates

- Each codec uses its own method to match target bitrates.
- JPEG XL, JPEG AI, and AVIF share identical bitrate targets per source.
- JPEG XT required higher bitrates to reach comparable quality.

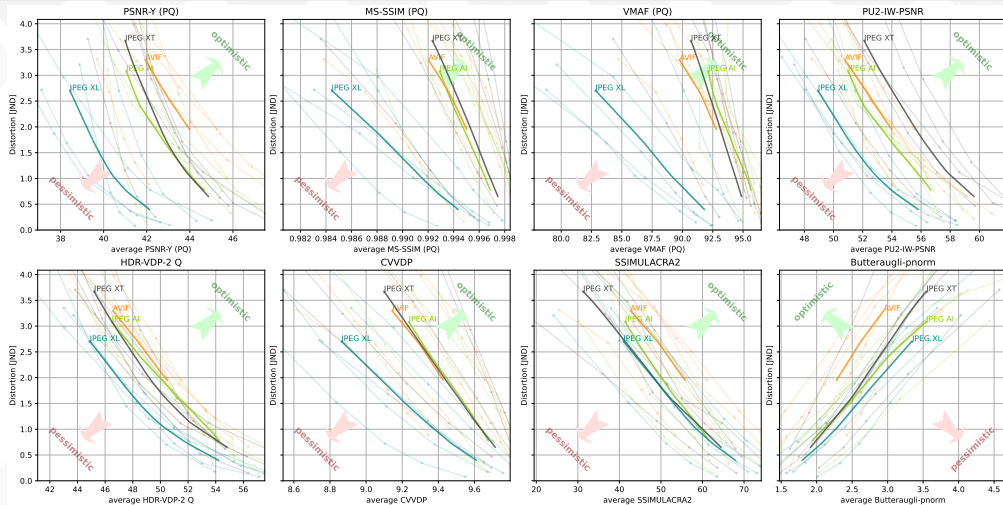
Source	AVIF / JPEG XL / JPEG AI	JPEG XT
alps-sunset	[0.3, 0.4, 0.6, 0.9, 1.1]	[0.9, 1.1, 1.5, 2.1, 2.5]
building	[0.2, 0.3, 0.5, 0.7, 1.1]	[0.7, 0.9, 1.3, 1.7, 2.5]
flower	[0.2, 0.3, 0.4, 0.6, 0.8]	[0.7, 0.9, 1.1, 1.5, 1.9]
p29-crop	[0.2, 0.3, 0.4, 0.5, 0.7]	[0.7, 0.9, 1.1, 1.3, 1.7]
room1	[0.2, 0.3, 0.5, 0.8, 1.1]	[0.7, 0.9, 1.3, 1.9, 2.5]

Source Bias: Impact of Image Content on Metric Scores



Metric predictions vs. subjective JND scores. Lines per source/codec, colored by source; per-source average highlighted in the distortion range covered by all codecs.

Codec Bias: Impact of Codec Type on Metric Scores



Predicted metrics' scores vs. subjective JND scores. Lines per source/codec, colored by codec; per-codec average highlighted in the distortion range covered by all sources.