WHAT IS CLAIMED IS:

1. A computer-implemented method, comprising:

receiving a plurality of training pairs and a subjective matrix that includes individual subjective scores provided by a plurality of subjective reviewers for the plurality of training pairs;

computing a first plurality of quality scores associated with a test encoding configuration based on a plurality of bootstrap quality models, wherein each bootstrap quality model included in the plurality of bootstrap quality models is trained based on at least one training pair included in the plurality of training pairs as well as a subset of the subjective matrix that includes individual subjective scores provided by a different subset of the plurality of subjective reviewers;

computing a second plurality of quality scores associated with a reference encoding configuration based on the plurality of bootstrap quality models; generating a distribution of bootstrap values for an encoding comparison metric based on the first plurality of quality scores and the second plurality of quality scores; and

determining an accuracy of a baseline value for the encoding comparison metric based on the distribution of bootstrap values, wherein the baseline value for the encoding comparison metric is generated by a baseline quality model that is trained based on the plurality of training pairs and the individual subjective scores provided by the plurality of subjective reviewers.

2. The computer-implemented method of claim 1, wherein the encoding comparison metric comprises a Bjontegaard rate difference (BD-rate).

3. The computer-implemented method of claim 1, wherein each quality score included in the first plurality of quality scores is a different value for a Video Multimethod Assessment Fusion (VMAF) metric.

4. The computer-implemented method of claim 1, wherein the test encoding configuration is associated with a first coder/decoder (codec), and the reference encoding configuration is associated with a second codec.

5. The computer-implemented method of claim 1, further comprising:

computing a third plurality of quality scores associated with the test encoding configuration based on the baseline quality model;

computing a fourth plurality of quality scores associated with the reference encoding configuration based on the baseline quality model; and

computing the baseline value for the encoding comparison metric based on the third plurality of quality scores and the fourth plurality of quality scores.

6. The computer-implemented method of claim 1, wherein computing the first plurality of quality scores comprises:

performing one or more encoding operations based on the test encoding configuration and a portion of video content, to generate a portion of encoded video content; and

computing a first quality score based on the portion of encoded video content and a first bootstrap quality model included in the plurality of bootstrap quality models.

7. The computer-implemented method of claim 1, wherein each quality score included in the first plurality of quality scores is associated with a different combination of a portion of encoded video content, a bitrate setting, and a bootstrap quality model included in the plurality of bootstrap quality models.

8. The computer-implemented method of claim 1, wherein the baseline quality model is trained based on the training database.

9. The computer-implemented method of claim 1, further comprising performing one or more content analysis operations based on the distribution of bootstrap values to quantify a performance of the test encoding configuration for a first type of video content.

10. The computer-implemented method of claim 1, further comprising performing one or more statistical analysis operations based on the distribution of bootstrap values and another distribution of bootstrap values associated with a different encoding configuration to compare an encoding performance of the test encoding configuration to an encoding performance of the different encoding configuration.

11. One or more non-transitory computer readable media including instructions that, when executed by one or more processors, cause the one or more processors to perform the steps of:

receiving a plurality of training pairs and a subjective matrix that includes individual subjective scores provided by a plurality of subjective reviewers for the plurality of training pairs;

computing a first plurality of quality scores associated with a test encoding configuration based on a plurality of bootstrap quality models, wherein each bootstrap quality model included in the plurality of bootstrap quality models is trained based on at least one training pair included in the plurality of training pairs as well as a subset of the subjective matrix that includes individual subjective scores provided by a different subset of the plurality of subjective reviewers;

computing a second plurality of quality scores associated with a reference encoding configuration based on the plurality of bootstrap quality models;

generating a baseline value for an encoding comparison metric based on a baseline quality model trained based on the plurality of training pairs and the individual subjective scores provided by the plurality of subjective reviewers, the test encoding configuration, and the reference encoding configuration;

generating a distribution of bootstrap values for the encoding comparison metric based on the first plurality of quality scores and the second plurality of quality scores; and

determining an accuracy of the baseline value for the encoding comparison metric based on the distribution of bootstrap values.

12. The one or more non-transitory computer readable media of claim 11, wherein the encoding comparison metric specifies a percentage bitrate change when encoding using the test encoding configuration relative to encoding using the reference encoding configuration while maintaining the same quality score.

13. The one or more non-transitory computer readable media of claim 11, wherein each quality score included in the first plurality of quality scores is a different value for a Video Multimethod Assessment Fusion (VMAF) metric.

14. The one or more non-transitory computer readable media of claim 11, wherein the test encoding configuration is associated with a first coder/decoder (codec), and the reference encoding configuration is associated with a second codec.

15. The one or more non-transitory computer readable media of claim 11, wherein computing the first plurality of quality scores comprises:

performing one or more encoding operations based on the test encoding configuration and a portion of video content, to generate a portion of encoded video content; and

computing a first quality score based on the portion of encoded video content and a first bootstrap quality model included in the plurality of bootstrap quality models.

16. The one or more non-transitory computer readable media of claim 11, wherein each quality score included in the first plurality of quality scores is associated with a different combination of a portion of encoded video content, a bitrate setting, and a bootstrap quality model included in the plurality of bootstrap quality models.

17. The one or more non-transitory computer readable media of claim 11, further comprising generating a confidence interval based on the distribution of bootstrap values and a confidence level.

18. The one or more non-transitory computer readable media of claim 11, further comprising performing one or more content analysis operations based on the distribution of bootstrap values to quantify a performance of the test encoding configuration for a first type of video content.

19. The one or more non-transitory computer readable media of claim 11, further comprising performing a Kolmogorov Smirnov test based on the distribution of bootstrap values and another distribution of bootstrap values associated with a different encoding configuration to compare an encoding performance of the test encoding configuration to an encoding performance of the different encoding configuration.

20. A system, comprising:

one or more memories storing instructions; and

one or more processors that are coupled to the one or more memories and, when executing the instructions, are configured to:

receive a plurality of training pairs and a subjective matrix that includes individual subjective scores provided by a plurality of subjective reviewers for the plurality of training pairs;

compute a first plurality of perceptual quality scores associated with a test codec based on a plurality of bootstrap quality models, wherein each bootstrap quality model included in the plurality of bootstrap quality models is trained based on at least one training pair included in the plurality of training pairs as well as a subset of the subjective matrix that includes individual subjective scores provided by a different subset of the plurality of subjective reviewers;

compute a second plurality of perceptual quality scores associated with a reference codec based on the plurality of bootstrap quality models;

generate a distribution of bootstrap values for an encoding comparison metric based on the first plurality of perceptual quality scores and the second plurality of perceptual quality scores; and

determine an accuracy of a baseline value for the encoding comparison metric based on the distribution of bootstrap values, wherein the baseline value for the encoding comparison metric is generated by a baseline quality model that is trained based on the plurality of training pairs and the individual subjective scores provided by the plurality of subjective reviewers.