Uncertainty and Assessment
Volume 1 · Number 1 · 2026

A Face Score Is an Incomplete Subject Proxy: Rank Disagreement in Two OpenS2V Leaderboards

Julian Douglas[1] · David Barker[2] · Jason Reed[3]

[1] Italy · [2] Finland · [3] Spain

Original publication: October 9, 2026 | Article 562

Abstract

Selecting a subject-preserving video model by face similarity can hide disagreement with a broader consistency measure. In the downloaded OpenS2V-Eval 1.1 tables, FaceSim and NexusScore had rank correlations of 0.471 across 15 Open-domain entries and 0.545 across 11 Single-domain entries. Individual rank differences reached nine and seven places, respectively. Regressions adjusting for motion amplitude gave positive face-score slopes, but both 95% intervals included zero. The Single-domain leaders also differed between metrics. These small, fixed leaderboards support displaying the two measures separately; they cannot establish video-level reliability or the superiority of a generation method.

Exploratory research draft using released model-level aggregates. Video generation and the Argus system were not executed; scientific review by the authors remains pending.

A Face Score Is an Incomplete Subject Proxy: Rank Disagreement in Two OpenS2V Leaderboards

Exploratory research draft using released model-level aggregates. Video generation and the Argus system were not executed; scientific review by the authors remains pending.

Abstract

Selecting a subject-preserving video model by face similarity can hide disagreement with a broader consistency measure. In the downloaded OpenS2V-Eval 1.1 tables, FaceSim and NexusScore had rank correlations of 0.471 across 15 Open-domain entries and 0.545 across 11 Single-domain entries. Individual rank differences reached nine and seven places, respectively. Regressions adjusting for motion amplitude gave positive face-score slopes, but both 95% intervals included zero. The Single-domain leaders also differed between metrics. These small, fixed leaderboards support displaying the two measures separately; they cannot establish video-level reliability or the superiority of a generation method.

1. The selection problem behind an identity number

A practitioner choosing a video generator may be offered several numbers under the heading of consistency. A score based on a face embedding looks especially convenient when a recognizable person is important. Yet a generated subject can retain a similar face while changing clothing, body appearance, or other reference details. Conversely, an object-focused task can make the face statistic a poor description of the requested subject. The practical problem is whether one convenient column preserves the ordering implied by a broader subject measure, rather than whether the columns share a favorable direction.

Argus explicitly motivates identity preservation beyond a single frontal observation and introduces a multi-view identity-memory approach (Meng et al., 2026). That distinction makes metric substitution a relevant question even when no new generator can be trained. Our study audits the released OpenS2V tables, not the numerical claims of Argus. It asks: within a fixed domain leaderboard, does FaceSim provide an adequate ordering proxy for NexusScore, and how sensitive is their adjusted association to individual model entries? An answer about these aggregates cannot establish which generated people a viewer would recognize.

The benchmark documentation distinguishes subject consistency, naturalness, textual relevance, motion, and aesthetics (Yuan et al., 2025). This separation supplies the conceptual basis of the audit. We formulate proposition P1 as a positive, imperfect association between FaceSim and NexusScore after adjusting for motion amplitude. The imperfect component matters: a positive slope could coexist with selection-relevant rank reversals. We therefore examine ordering and conditional association together, without defining a threshold at which one metric becomes an acceptable replacement for another.

2. Why related evaluation tools are not interchangeable

Video benchmarks already recognize that generation quality has several components. VBench organizes evaluation into distinct dimensions, whereas the Fréchet Video Distance framework measures discrepancies between video distributions (Huang et al., 2023; Unterthiner et al., 2018). Neither construction makes every identity statistic equivalent to overall video quality. Our analysis keeps the two released consistency columns in their original roles and avoids turning their average into a new quality index. No generated videos, reference images, or feature embeddings were downloaded for recomputation.

Face-recognition objectives explain why a face-based score has a legitimate but bounded role. FaceNet learns an embedding for face recognition and clustering, and ArcFace develops an angular-margin objective for recognition (Schroff et al., 2015; Deng et al., 2018). Their relevance is the measurement domain: a face descriptor targets facial identity, rather than every property of a video subject. An association between a face descriptor and another metric is consequently an empirical relationship among outputs, not a definitional identity. Domain composition can change that relationship even when both measurements are computed correctly.

Personalization research also uses different routes to preserve reference information. DreamBooth fine-tunes a text-to-image model for a subject, while IP-Adapter introduces image-prompt conditioning compatible with text conditioning (Ruiz et al., 2022; Ye et al., 2023). Custom Diffusion studies efficient customization, and identity-specific video work includes ID-Animator and ConsisID (Kumari et al., 2022; He et al., 2024; Yuan et al., 2024). These sources motivate attention to what is being preserved and through which evidence. The present leaderboard audit does not identify which conditioning mechanism caused a high or low metric value.

Temporal coherence adds another layer. ConsistI2V addresses consistency in image-to-video generation; AnimateDiff, VideoCrafter2, CogVideoX, and Stable Video Diffusion develop other video-generation approaches (Ren et al., 2024; Guo et al., 2023; Chen et al., 2024; Yang et al., 2024; Blattmann et al., 2023). We cite these as examples of a research setting with multiple model families and temporal objectives, without importing their performance figures. Repeated appearances of a model name across benchmark domains are particularly important: they represent linked evaluations of a system, not independent replications of a randomly sampled video population.

More general visual representations do not eliminate this distinction. CLIP and self-supervised vision transformers offer different representation-learning objectives (Radford et al., 2021; Caron et al., 2021), while structural similarity and learned perceptual similarity address image comparisons (Wang et al., 2004; Zhang et al., 2018). A diverse set of available representations increases the number of defensible measurement choices. It does not make their resulting rankings interchangeable. Our narrow audit concerns two named columns whose values were released together, with no attempt to reconstruct the benchmark's underlying feature calculations.

3. A deliberately small evidence base

We used the downloaded OpenS2V-Eval 1.1 Open-domain and Single-domain CSV files. The former contains 15 model entries and the latter 11, with 16 distinct printed model names across their union. Each retained row provides FaceSim, NexusScore, and motion amplitude. Percentage strings were converted to numerical points on the zero-to-one-hundred display scale. We checked that these fields were present and finite and that model names were unique within each file. We made no spelling-based merger of different printed model names.

The Human-domain CSV was excluded because it does not contain the released NexusScore column required for this paired comparison. Its omission is a measurement-availability decision, not evidence that the domain is unimportant. The literature on missing data distinguishes missing measurements from observed zeros (Little and Rubin, 2002). Here there is no defensible zero to insert and no auxiliary video-level information with which to impute the missing metric. A later release with that column would support a new audit, but it cannot be reconstructed from the current file.

All reported counts concern model-domain aggregates. The 26 rows are not 26 independent generators, and they are certainly not 26 individual videos. Benchmark prompts, subjects, and evaluation seeds are not identifiable in these CSV summaries. Datasheet-style provenance motivates preserving the exact downloaded tables, file hashes, domain labels, and access conditions (Gebru et al., 2021). We retain those materials locally so that the rank calculations can be repeated after a leaderboard update without confusing old and new snapshots.

Rank disagreement was calculated separately within each domain. The highest score receives rank one; equal values would receive average ranks. For each entry we subtract its NexusScore rank from its FaceSim rank, then summarize the absolute difference. Spearman correlation describes overall monotone association. It does not establish agreement, just as a high correlation between two measurement methods need not make their individual readings interchangeable (Bland and Altman, 1986). Rank positions also depend on the competing entries in the same snapshot.

4. Conditional comparison and uncertainty

The regression outcome is NexusScore in displayed percentage points. Explanatory variables are FaceSim divided by ten and motion amplitude divided by ten. Thus the first slope describes a ten-point face-score difference conditional on the recorded amplitude column. Motion amplitude is an adjustment variable because static identity retention and movement can present different evaluation demands. Its inclusion does not establish that movement is exogenous or that the regression has isolated an identity mechanism.

We fit a separate ordinary least squares model in each domain and use HC3 covariance with a Student reference distribution. Heteroskedasticity-consistent inference concerns the residual variance specification (White, 1980); it does not turn selected leaderboard entries into random experimental units. A clustered pooled analysis would introduce an additional problem because names recur across domains. Cluster-robust inference depends on having a defensible sampling and dependence structure (Cameron and Miller, 2015). We instead retain the domain boundary and acknowledge that shared architectures can still make entries dependent within it.

The resulting intervals describe an exploratory model-based comparison across entries. They do not measure uncertainty from generating videos repeatedly, rating the same subject again, or rerunning an embedding model. No such repeated observations were released here. Reliability coefficients require a specified measurement design, as emphasized in guidance on intraclass correlation (Koo and Li, 2016). We therefore report neither a video-level reliability coefficient nor a probability that one generator would beat another on an unseen identity.

To assess sensitivity to single entries, we refit each regression after deleting one model at a time. This is a descriptive influence exercise, not cross-validation of a deployed predictor. Structured-validation guidance warns against disregarding dependence, and model-selection work warns against optimistic error estimates after tuning (Roberts et al., 2017; Varma and Simon, 2006). We do not select the favorable deletion, choose covariates from the resulting significance levels, or describe leave-one-out coefficient

ranges as independent confidence intervals. Figure 1 shows the measurement relationships actually examined.

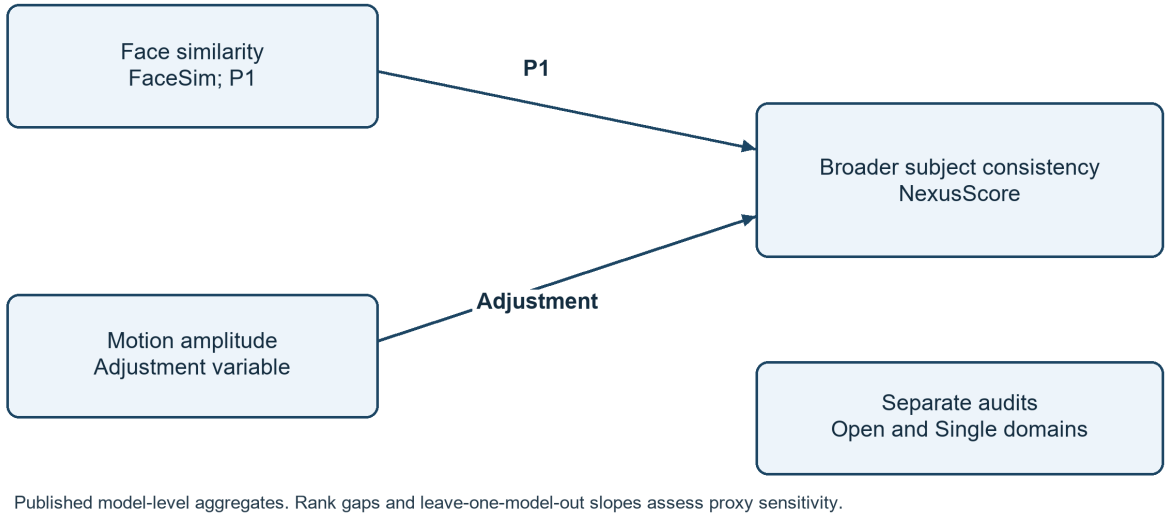

Figure 1. Research model for P1. Arrows identify regression relationships, rather than experimentally identified causal paths. Rank comparisons and deletion sensitivity accompany each domain-specific fit.

5. Results: association leaves room for different choices

In the Open domain, the Spearman correlation between FaceSim and NexusScore is 0.471. The median absolute rank difference is three places and the largest is nine. RefAlign-14B leads both columns, so the disagreement should not be exaggerated into a claim that every choice changes. The result instead shows that agreement at the top can coexist with appreciable reordering among other available entries. A researcher comparing a restricted candidate set could therefore face a different practical ordering from the one suggested by the leading entry alone.

For the Single domain, the corresponding correlation is 0.545, the median absolute gap is one place, and the maximum gap is seven. VACE-14B leads FaceSim, whereas HunyuanCustom leads NexusScore. The smaller median gap does not remove the top-choice distinction. Together these summaries show why a single correlation is insufficient: it can conceal both where rankings diverge and whether a specific decision boundary is affected. They also avoid assuming that the ordinal distances between neighboring models correspond to equal perceptual differences.

Table 1 reports the fitted relationships. A ten-point FaceSim difference is associated with 1.111 additional NexusScore points in the Open-domain model, with a 95% interval from -0.306 to 2.528. In the Single domain, the slope is 2.641 points, with an interval from -0.296 to 5.579. Both estimates are positive, but the data and residual assumptions do not rule out zero. These are inconclusive conditional associations rather than evidence of no relationship or proof that the domains have different population slopes.

Table 1. Separate OLS comparisons of released aggregate metrics; HC3 standard errors in parentheses. NexusScore is measured in displayed percentage points. Both regressions include an intercept and the two listed predictors; no domain pooling or video-level observations are used.

Model	Term	Estimate (SE)	95% CI	N
Open domain	FaceSim per 10 points	1.1111 (0.6503)	[-0.3058, 2.5280]	15
Open domain	Motion amplitude per 10 points	0.0866 (1.7257)	[-3.6733, 3.8464]	15
Open domain: fit	R-squared = 0.2054	-	-	15
Single domain	FaceSim per 10 points	2.6413 (1.2739)	[-0.2964, 5.5790]	11
Single domain	Motion amplitude per 10 points	-0.4126 (3.7913)	[-9.1553, 8.3301]	11
Single domain: fit	R-squared = 0.3049	-	-	11

Motion amplitude has a slope of 0.087 in the Open domain and -0.413 in the Single domain. Its wide intervals cover substantial positive and negative values, making a mechanistic tradeoff interpretation unwarranted. The models explain approximately 20.5% and 30.5% of the respective cross-entry outcome variation. These fitted fractions describe the same observations used for estimation. They are neither out-of-sample accuracy nor a decomposition of the benchmark's definition of subject identity.

Deleting one entry at a time gives FaceSim slopes ranging from 0.702 to 1.429 in the Open domain and from 1.783 to 3.225 in the Single domain. The positive directions persist in this limited exercise, but their magnitudes move. That stability in sign does not repair the small sample or establish calibrated intervals. The analysis also cannot reveal whether the same difficult clips drove both scores, because the underlying clip-level covariance is absent from the public aggregate tables.

6. What the proxy audit permits

The research question has a practical answer within these files: FaceSim does not preserve every ordering implied by NexusScore. This statement is directly observed and requires no claim that either measure is ground truth. Choosing the broader score instead would still leave naturalness, prompt adherence, and motion quality to assess separately. The audit supports retaining distinct columns in a selection record and naming the criterion that justifies a chosen model, especially when a narrow identity cue is being used for a broader subject requirement.

One immediate reporting improvement requires no additional computation: retain the identity of the selected metric beside the model choice. A report that says only that a candidate has the highest consistency score leaves the decision rule unrecoverable. Naming FaceSim or NexusScore, the evaluated domain, and the table version makes that rule inspectable. The same practice would allow a later evaluator to distinguish a genuinely changed score from a changed candidate set. It does not create the missing clip-level measurements, but it prevents avoidable ambiguity about the aggregate evidence actually used.

There are several competing explanations for disagreement. A metric may emphasize a different construct; a model may perform differently on faces and non-face details; or the scores may contain evaluation noise. Measurement-error analysis cautions against treating a recorded variable as an error-free latent construct (Carroll et al., 2006). Aggregate tables cannot distinguish these explanations. A stronger follow-up would obtain matched per-video scores, subject categories, repeated seeds, and human judgments, then examine whether differences remain within the same subjects and conditions.

Nor should readers search this audit for isolated significant associations among many metric combinations. False-discovery-rate methodology addresses families of statistical tests (Benjamini and Hochberg, 1995). We keep the two prespecified descriptive domain fits visible, retain their inconclusive intervals, and make no catalog of favorable metric pairs. The deletion results are influence summaries. They are not additional confirmatory tests. A future larger analysis would need its own selection rules and uncertainty design rather than treating this exploratory specification as a validated prediction rule.

7. Closing interpretation

The released leaderboards contain enough information to show a concrete proxy problem and too little information to explain it completely. FaceSim and NexusScore are positively ordered on average, yet individual positions differ and one domain has different leading models. Reporting both measures preserves information that disappears when they are collapsed into a single identity claim. The contribution is a reproducible audit of measurement substitution in a fixed snapshot, with the limits of aggregate evidence retained in the interpretation.

Data and preparation record

The analysis table preserves model names, original score strings, converted values, ranks, and domain membership. Saved result objects contain all coefficients, uncertainty estimates, and deletion ranges; the native diagram records the tested relationships. The source tables remain subject to their original distribution conditions. No new personal data, facial embeddings, generation runs, or participant judgments were collected, and no assertion of author approval or submission readiness is made.

References

1. Ren, Weiming; Yang, Huan; Zhang, Ge; Wei, Cong; Du, Xinrun; Huang, Wenhao; Chen, Wenhu. (2024). ConsistI2V: Enhancing Visual Consistency for Image-to-Video Generation. arXiv:2402.04324. https://arxiv.org/abs/2402.04324.
2. Schroff, Florian, et al. "FaceNet: A Unified Embedding for Face Recognition and Clustering" *arXiv preprint arXiv:1503.03832*, 2015. https://arxiv.org/abs/1503.03832.
3. Wang, Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing*, vol. 13, no. 4, 2004, pp. 600-612. https://doi.org/10.1109/TIP.2003.819861.
4. Unterthiner, Thomas, et al. "Towards Accurate Generative Models of Video: A New Metric & Challenges" *arXiv preprint arXiv:1812.01717*, 2018. https://arxiv.org/abs/1812.01717.
5. He, Xuanhua; Liu, Quande; Qian, Shengju; Wang, Xin; Hu, Tao; Cao, Ke; Yan, Keyu; Zhang, Jie. (2024). ID-Animator: Zero-Shot Identity-Preserving Human Video Generation. arXiv:2404.15275. https://arxiv.org/abs/2404.15275.
6. Huang, Ziqi, et al. "VBench: Comprehensive Benchmark Suite for Video Generative Models" *arXiv preprint arXiv:2311.17982*, 2023. https://arxiv.org/abs/2311.17982.
7. Caron, Mathilde, et al. "Emerging Properties in Self-Supervised Vision Transformers" *arXiv preprint arXiv:2104.14294*, 2021. https://arxiv.org/abs/2104.14294.
8. Guo, Yuwei; Yang, Ceyuan; Rao, Anyi; Liang, Zhengyang; Wang, Yaohui; Qiao, Yu; Agrawala, Maneesh; Lin, Dahua; Dai, Bo. (2023). AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. arXiv:2307.04725. https://arxiv.org/abs/2307.04725.
9. Koo, Terry K., and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine*, vol. 15, no. 2, 2016, pp. 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.
10. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
11. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.

12. Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.

13. Radford, Alec, et al. "Learning Transferable Visual Models From Natural Language Supervision" *arXiv preprint arXiv:2103.00020*, 2021. https://arxiv.org/abs/2103.00020.

14. Kumari, Nupur; Zhang, Bingliang; Zhang, Richard; Shechtman, Eli; Zhu, Jun-Yan. (2022). Multi-Concept Customization of Text-to-Image Diffusion. arXiv:2212.04488. https://arxiv.org/abs/2212.04488.

15. Yuan, Shenghai; Huang, Jinfa; He, Xianyi; Ge, Yunyuan; Shi, Yujun; Chen, Liuhan; Luo, Jiebo; Yuan, Li. (2024). Identity-Preserving Text-to-Video Generation by Frequency Decomposition. arXiv:2411.17440. https://arxiv.org/abs/2411.17440.

16. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.

17. Yang, Zhuoyi; Teng, Jiayan; Zheng, Wendi; Ding, Ming; Huang, Shiyu; Xu, Jiazheng; Yang, Yuanming; Hong, Wenyi; Zhang, Xiaohan; Feng, Guanyu; Yin, Da; Zhang, Yuxuan; Wang, Weihan; Cheng, Yean; Xu, Bin; Gu, Xiaotao; Dong, Yuxiao; Tang, Jie. (2024). CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. arXiv:2408.06072. https://arxiv.org/abs/2408.06072.

18. Deng, Jiankang, et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition" *arXiv preprint arXiv:1801.07698*, 2018. https://arxiv.org/abs/1801.07698.

19. Zhang, Richard, et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric" *arXiv preprint arXiv:1801.03924*, 2018. https://arxiv.org/abs/1801.03924.

20. Ye, Hu, et al. "IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models" *arXiv preprint arXiv:2308.06721*, 2023. https://arxiv.org/abs/2308.06721.

21. Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.

22. Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.

23. Chen, Haoxin; Zhang, Yong; Cun, Xiaodong; Xia, Menghan; Wang, Xintao; Weng, Chao; Shan, Ying. (2024). VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models. arXiv:2401.09047. https://arxiv.org/abs/2401.09047.

24. Yuan, Shenghai; He, Xianyi; Deng, Yufan; Ye, Yang; Huang, Jinfa; Lin, Bin; Luo, Jiebo; Yuan, Li. (2025). OpenS2V-Nexus: A Detailed Benchmark and Million-Scale Dataset for Subject-to-Video Generation. arXiv:2505.20292. https://arxiv.org/abs/2505.20292.

25. Blattmann, Andreas; Dockhorn, Tim; Kulal, Sumith; Mendelevitch, Daniel; Kilian, Maciej; Lorenz, Dominik; Levi, Yam; English, Zion; Voleti, Vikram; Letts, Adam; Jampani, Varun; Rombach, Robin. (2023). Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. arXiv:2311.15127. https://arxiv.org/abs/2311.15127.

26. Meng, Zijie; Liu, Jiwen; Liu, Yufei; Tong, Chengzhuo; Liu, Xiaoqiang; Zhang, Yuanxing; Xu, Yulong; Wan, Pengfei. (2026). ARGUS: Stacked Multi-View Identity Mosaic Injection for Subject-Preserving Video Generation. arXiv:2606.11670. https://arxiv.org/abs/2606.11670.

27. Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8.

28. Ruiz, Nataniel, et al. "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation" *arXiv preprint arXiv:2208.12242*, 2022. https://arxiv.org/abs/2208.12242.

29. Benjamini, Yoav and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, 1995, pp. 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

30. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.