

An Agentic Multi-Modal Evaluation Framework for Audio-Driven Generative Video Synthesis: Bridging Semantic Coherence and Quality Alignment

Shichen Wu

Data Science Quantitative Finance, University of California, Berkeley, Berkeley, CA, USA

DOI: 10.63575/CIA.2026.40205

Abstract

Evaluating audio-driven talking-face video requires distinguishing synchronization, visual quality, naturalness, and emotional consistency. This paper develops an evaluation protocol and examines its statistical behavior using a reproducible synthetic dataset. The dataset contains 600 video-level records arranged as 200 source groups with three condition labels, together with ratings from 24 simulated raters. No generated videos, human participants, or model inference outputs are used in the reported analysis. Under the specified simulation, LSE-C correlates with simulated lip-sync scores at $\rho = 0.626$ and with the holistic composite at $\rho = 0.253$. The DOVER-labeled variable reaches $\rho = 0.744$ on the holistic composite, while out-of-fold ridge predictions reach $\rho = 0.821$. Simulated reference-guided scores show higher agreement than absolute scores, and one simulated judge exhibits substantial order sensitivity. These patterns reflect assumptions imposed by the data generator, including condition effects, reduced reference-protocol noise, and position bias. They illustrate how pooled correlations, metric conflicts, and pairwise agreement can be analyzed; they do not establish the relative performance of the named systems. The resulting protocol specifies the data and validation needed for an empirical study with real videos and human ratings.

Keywords: *talking-face evaluation; method validation; metric agreement; multimodal judges*

1. Introduction

1.1. Background and Motivation

Audio-driven talking-face systems include mouth-region synthesis guided by a pretrained synchronization expert [1] and diffusion-based portrait animation [2]. Their evaluation involves distinct properties: speech–mouth alignment, image quality, motion naturalness, and audiovisual emotional consistency. A metric designed for one property need not measure the others. Using a related network for both training and evaluation also motivates testing whether improvements in its score correspond to improvements perceived by viewers.

VBench organizes visual video evaluation into separate dimensions and checks their agreement with human preferences [3]. Its original evaluation framework does not directly assess speech–mouth synchronization. Audio-capable multimodal judges offer another possible evaluation channel, but their suitability for this task requires controlled validation of input processing, scoring rubrics, and presentation order.

This paper examines how an evaluation pipeline can characterize disagreement between synchronization scores, quality scores, semantic scores, and simulated judge responses. It uses synthetic records to demonstrate the analysis and describes the additional evidence required to answer the corresponding empirical questions.

1.2. Scope and Research Questions

The proposed application is speech-driven talking-face video. Three questions guide the protocol: how automatic signals agree with dimension-specific mean opinion scores; how absolute, reference-guided, and pairwise judging differ; and how pooled agreement changes when calculated separately within conditions. The present simulation demonstrates these calculations without answering which real evaluator is most accurate.

A. Evaluators Under Study

The schema includes six per-video objective-score variables, a set-level distance calculation, a learned-evaluator variable, and two multimodal-judge configurations. CLAP-Affect is defined only for the MEAD-labeled subset. Names such as Wav2Lip, SadTalker, DOVER, and Gemini-2.5-Pro identify intended components of a future empirical protocol. In the reported results they denote synthetic variables or condition labels, not measured outputs.

B. Contributions

The contribution is a reproducible analysis example comprising dimension-specific correlations, source-group bootstrap intervals, out-of-fold metric fusion, conflict counts, and pairwise order consistency. It also makes explicit which observed patterns are imposed by the simulator. The work introduces neither a validated metric nor an empirical ranking of existing generators or judges.

2. Related Work

2.1. Audio-Driven Talking-Face Generation

A. Inpainting and Coefficient-based Methods

Wav2Lip synthesizes the mouth region using visual input and speech features, with a pretrained synchronization expert contributing supervision. SadTalker predicts expression and head-pose coefficients from audio and renders an animated portrait [4]. These different input and synthesis settings motivate separate evaluation of synchronization, visual quality, and naturalness. Their architectural differences alone do not establish a fixed ordering on those properties.

B. Diffusion-based Methods

Diffusion-based approaches such as EMO synthesize portrait motion directly from audio-conditioned features without requiring an intermediate three-dimensional face representation. Related audio-to-video work adapts text-to-video models and introduces AV-Align to assess audiovisual temporal correspondence [5]. This motivates considering both general audiovisual alignment and speech-specific synchronization, while validating each measure for the intended content.

2.2. Automatic Metrics for Generated Video

FAST-VQA uses fragment sampling for efficient no-reference video quality assessment [6]. T2VQA-DB provides subjective ratings of text-to-video outputs for evaluating learned quality predictors [7]. More directly related work, THQA-10K, studies subjective and objective assessment of generated talking heads [8]. These resources provide context for an empirical talking-face evaluation. The present work is limited to a synthetic demonstration of the statistical protocol and does not claim to be the first study of human alignment in this domain.

2.3. Learned and Multimodal-LLM-based Judges

VideoScore learns aspect-level evaluation from annotated generated videos [9]. GenAI Arena collects user preferences across visual generation tasks [10]. These studies motivate evaluating learned scores against human judgments rather than treating model outputs as ground truth. For audio-driven talking-face content, an empirical comparison should additionally verify that a judge processes the audio track and that its scores respond to controlled synchronization changes.

3. Method Validation Design and Evaluation Protocol

3.1. Synthetic Records and Intended Video Pool

HDTF is a high-resolution audiovisual resource for talking-face research [11], and MEAD supplies emotional audiovisual recordings [12]. The simulation uses their names only as subset labels. It creates 100 HDTF-labeled source identifiers with distinct speaker identifiers and 100 MEAD-labeled source identifiers representing 25 speakers across neutral, happy, angry, and sad labels. It does not select or decode recordings from either corpus. Each record has nominal metadata specifying eight seconds, 25 frames per second, and 512×512 pixels.

Each of the 200 source identifiers has three condition records labeled Real, Wav2Lip, and SadTalker, yielding 600 records. Real denotes a simulated reference condition, not an observed ground-truth recording. Table 1 gives the record counts and nominal durations; Figure 1 summarizes the intended empirical workflow. The simulation was generated with random seed 2026. Its scores depend on latent synchronization, visual-quality, naturalness, and emotion variables, with condition-specific means, noise, and shared source difficulty.

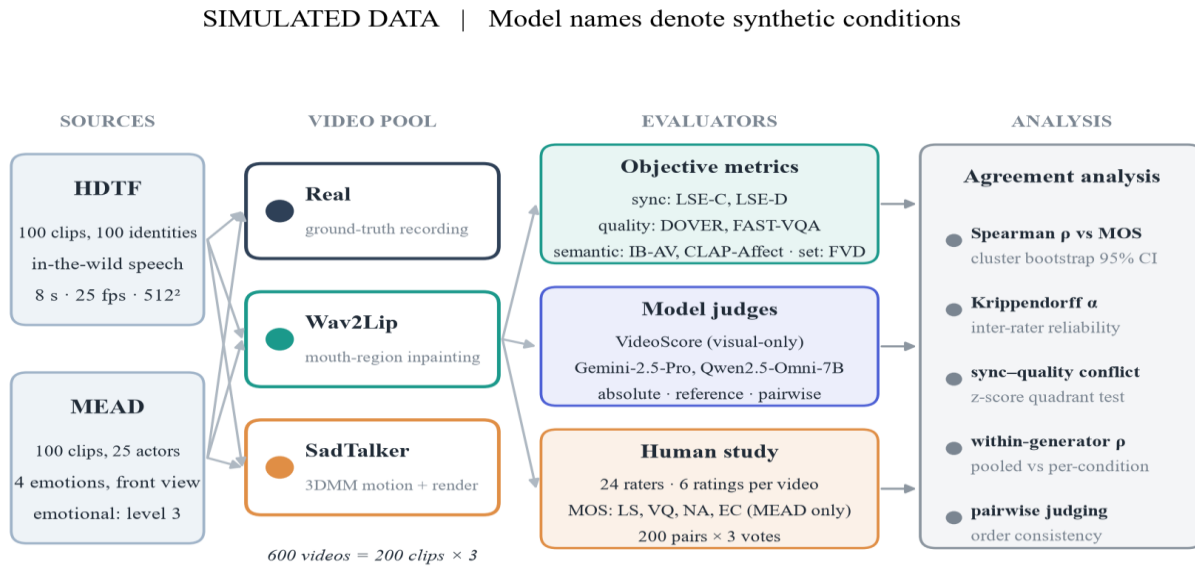


Figure 1. Intended Evaluation Design

The workflow specifies three conditions for each of 200 source clips, six ratings per video from a pool of 24 raters, and three preference votes for each of 200 generated-video pairs. The reported demonstration instantiates this structure using synthetic records only. No video generation or human data collection is documented by the supplied dataset.

Table 1. Composition of the Synthetic Record Pool

Subset label	Sources	Speakers	Emotion labels	Records	Nominal duration	Dimensions
HDTF	100	100	not labelled	300	40 min	LS, VQ, NA
MEAD	100	25	4 emotions $\times$ 25	300	40 min	LS, VQ, NA, EC
Total	200	125	—	600	80 min	—

Source: simulated manifest. Durations are nominal totals, not measured footage. LS = lip-sync; VQ = visual quality; NA = naturalness; EC = emotional consistency. Speaker counts refer to synthetic identifiers.

3.2. Automatic Metrics and Judges

A. Objective Metrics

Table 2 distinguishes the intended evaluators from their simulated representations. An empirical implementation would obtain LSE-C and LSE-D from a synchronization network [13], DOVER from aesthetic and technical quality branches [14], and FAST-VQA from video fragments. IB-AV would compare normalized ImageBind audio and aggregated vision embeddings [15]. CLAP-Affect is a proposed score combining LAION-CLAP audio–text similarities [16] and ImageBind image–text similarities over the same emotion labels. It is not an established metric validated by those papers. A real implementation must specify prompts, frame aggregation, normalization, softmax temperature, and the Jensen–Shannon divergence convention. In the simulation, these scores are generated directly from latent variables and noise; neither embeddings nor emotion distributions are computed. DOVER, FAST-VQA, and CLAP-Affect are clipped to [0, 1] by the simulator. Those bounds do not define the native output scales of the original models.

Table 2. Intended Evaluators and Synthetic Representations

Variable	Family	Intended input	Synthetic output	Better	Target
LSE-C	Synchronization	Audio + mouth crops	Confidence proxy	Higher	LS
LSE-D	Synchronization	Audio + mouth crops	Distance proxy	Lower	LS
DOVER	Perceptual quality	Video frames	Clipped score, 0–1	Higher	VQ
FAST-VQA	Perceptual quality	Video fragments	Clipped score, 0–1	Higher	VQ
IB-AV	Semantic alignment	Audio + 8 frames	Similarity proxy	Higher	Exploratory EC
CLAP-Affect	Semantic alignment	Audio + frames	Clipped proxy, 0–1	Higher	EC
FD	Distribution	Random features	Set-level distance	Lower	—
VideoScore	Learned judge	Frames only	Aspect scores, 1–4	Higher	VQ, NA
Gemini-2.5-Pro	Omni-modal judge	Audio + video	Integer, 1–5	Higher	LS, VQ, NA, EC
Qwen2.5-Omni-7B	Omni-modal judge	Audio + video	Integer, 1–5	Higher	LS, VQ, NA, EC

Source: proposed evaluation schema and simulation code. All reported score values are synthetic. MOS denotes the mean of simulated ratings. FD denotes Fréchet distance on synthetic features, not FVD measured from videos.

B. Judge Protocols

The simulation contains a VideoScore-labeled variable and judge configurations named Gemini-2.5-Pro and Qwen2.5-Omni-7B. Their scores are sampled from latent variables rather than returned by the named models. Absolute and reference-guided scores are rounded and clipped to the five-point rating scale; the reference configuration multiplies the noise standard deviation by 0.72. Pairwise responses include predefined first-position biases. A future empirical protocol would compare single-video scoring with scoring assisted by a disjoint same-speaker reference and would present each generated-video pair in both orders. It would also record model identifiers, prompts, preprocessing, decoding settings, raw responses, and query dates. The supplied simulation contains no such inference records.

3.3. Simulated Ratings and Statistical Analysis

A. Rating Structure

Twenty-four simulated rater identifiers each receive 150 distinct video records. Each of the 600 records has six distinct raters, giving 3,600 video–rater sessions. Each session contains LS, VQ, and NA scores; the 300 MEAD-labeled records also have EC scores. This produces 12,600 individual dimension scores. Ratings are sampled

from latent variables with rater-specific bias, sensitivity, and noise, then rounded and clipped to integers from 1 to 5. MOS is the mean across raters for a dimension; MOS-H is the unweighted mean of MOS-LS, MOS-VQ, and MOS-NA. It is a constructed composite, not an independently rated overall-quality score. Separately, 200 pairs receive three simulated preference votes each, yielding 600 votes. Their latent preference function uses weights of 0.30, 0.35, and 0.35 for synchronization, visual quality, and naturalness; it is therefore not identical to MOS-H.

B. Agreement Statistics

Agreement is measured using Spearman rank correlation. Percentile intervals use 1,000 bootstrap resamples of source groups, preserving the three conditions associated with each source. These intervals describe resampling variation in the synthetic dataset, not uncertainty about real-model performance. Krippendorff’s α is calculated with squared interval distances on individual scores. Ridge fusion uses LSE-C, sign-reversed LSE-D, DOVER, FAST-VQA, and IB-AV to predict MOS-H. Five outer folds are grouped by source identifier; standardization and RidgeCV fitting occur within each training fold, and correlations use only out-of-fold predictions. The reported fusion interval bootstraps those fixed predictions without refitting the model. A conflict requires oppositely signed, globally standardized LSE-C and DOVER values, each with magnitude greater than 0.5. Pairwise agreement averages 400 ordered verdicts per judge, and order consistency averages 200 pairs. Bootstrap resampling for pairwise agreement preserves both orders of each pair.

4. Method Validation Results and Analysis

4.1. Agreement with Generated Reference Ratings

A. Objective Metrics

Krippendorff’s α is 0.398 for LS, 0.434 for VQ, 0.512 for NA, and 0.478 for EC. These values indicate appreciable disagreement among simulated raters; they do not establish that six ratings are sufficient for reliable empirical evaluation. Table 3 shows an imposed rank reversal: the Wav2Lip-labeled condition has higher LSE-C than Real (8.623 versus 7.620), but lower MOS-LS (3.487 versus 3.771). The code explicitly adds a synchronization-score bonus to the Wav2Lip condition. SadTalker has higher synthetic DOVER and MOS-VQ values than Wav2Lip, whereas Wav2Lip has a higher MOS-H (3.001 versus 2.949). The Fréchet distances of 302.6 and 253.9 are computed from 64-dimensional random features and have no interpretation as measured video quality.

Table 3. Condition Means in the Synthetic Dataset

Measure	Real	Wav2Lip	SadTalker
LSE-C $\uparrow$	7.620	8.623	6.575
LSE-D $\downarrow$	7.166	6.522	7.813
DOVER $\uparrow$	0.718	0.548	0.596
FAST-VQA $\uparrow$	0.655	0.550	0.584
IB-AV $\uparrow$	0.230	0.226	0.210
FD $\downarrow$	—	302.6	253.9

Measure	Real	Wav2Lip	SadTalker
MOS-LS	3.771	3.487	2.941
MOS-VQ	3.898	2.718	3.138
MOS-NA	4.132	2.798	2.769
MOS-EC (MEAD)	3.837	3.405	2.473
MOS-H	3.934	3.001	2.949

N = 200 records per condition; MOS-EC uses 100 per condition. FD is computed against the simulated Real condition using random features. Source: recomputed simulation data.

Table 4 and Figure 2 summarize correlations within the synthetic dataset. LSE-C reaches $\rho = 0.626$ with MOS-LS; DOVER reaches $\rho = 0.782$ with MOS-VQ and $\rho = 0.744$ with MOS-H. LSE-C reaches only $\rho = 0.253$ with MOS-H. IB-AV and CLAP-Affect correlate with MOS-EC at 0.346 and 0.424, respectively. These values follow from the simulator and cannot establish that a real contrastive audio model detects facial affect or its timing; language–audio alignment [17] alone does not validate that claim. Ridge fusion reaches $\rho = 0.821$ (95% interval 0.795–0.843), exceeding the DOVER-labeled variable by 0.077 in this dataset. The use of human-aligned metric combinations has precedent [18], but the present gain is a synthetic example, not a demonstrated empirical improvement.

Table 4. Spearman Correlations with Simulated Mean Opinion Scores

Evaluator	MOS-LS	MOS-VQ	MOS-NA	MOS-EC	MOS-H [95% CI]
LSE-C	0.626	−0.032	0.095	0.337	0.253 [0.192, 0.311]
−LSE-D	0.585	−0.031	0.083	0.276	0.233 [0.169, 0.295]
DOVER	0.382	0.782	0.670	0.218	0.744 [0.708, 0.779]
FAST-VQA	0.337	0.634	0.514	0.167	0.605 [0.549, 0.654]
IB-AV	0.220	0.072	0.174	0.346	0.187 [0.115, 0.262]
CLAP-Affect	0.172	0.113	0.233	0.424	0.206 [0.113, 0.303]
VideoScore	0.345	0.640	0.544	0.267	0.616 [0.568, 0.658]
Ridge fusion (CV)	—	—	—	—	0.821 [0.795, 0.843]

N = 600 except correlations involving MOS-EC or CLAP-Affect, which use N = 300. Fusion uses five objective variables. Source: recomputed simulation data.

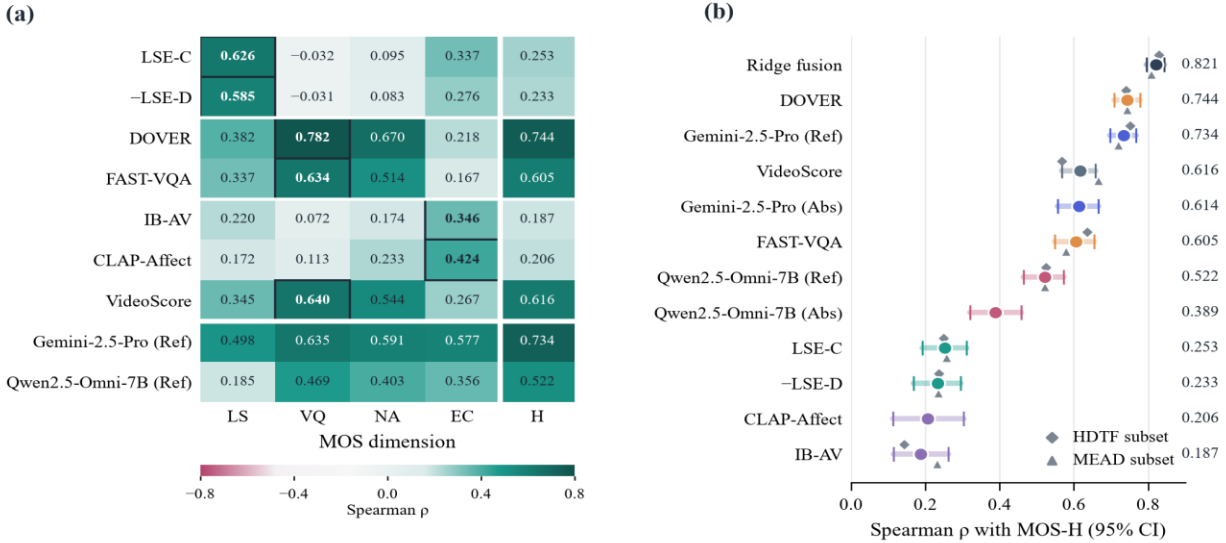


Figure 2. Agreement with Simulated Mean Opinion Scores

All plotted observations are synthetic. (a) Dimension-specific correlations; judge rows use the reference configuration. MOS-EC and CLAP-Affect use 300 MEAD-labeled records. (b) MOS-H correlations with 95% source-group bootstrap intervals and subset estimates. Fusion reaches 0.821, DOVER 0.744, and the Gemini-labeled reference configuration 0.734; IB-AV reaches 0.187. These are scenario-specific statistics, not model benchmark scores.

B. Simulated Judge Configurations

Table 5 shows higher correlations for both reference configurations on every dimension. The Gemini-labeled configuration changes from 0.614 to 0.734 on MOS-H, a difference of 0.120 (95% interval 0.063–0.179); the Qwen-labeled configuration changes from 0.389 to 0.522. This pattern is expected because reference scores were generated with lower noise. Under that configuration, MOS-LS correlations are 0.498 and 0.185. Absolute judge composites exceed MOS-H by 0.213 and 0.273 points on average. Figure 3(c) describes the association between synthetic judge scores and simulated MOS-VQ; it does not demonstrate actual model leniency or reference effectiveness.

Table 5. Spearman Correlations of Simulated Judge Scores

Judge	Protocol	MOS-LS	MOS-VQ	MOS-NA	MOS-EC	MOS-H [95% CI]
Gemini-2.5-Pro	Absolute	0.415	0.475	0.487	0.471	0.614 [0.557, 0.666]
Gemini-2.5-Pro	Reference-guided	0.498	0.635	0.591	0.577	0.734 [0.698, 0.767]
Qwen2.5-Omni-7B	Absolute	0.108	0.350	0.344	0.296	0.389 [0.320, 0.458]
Qwen2.5-Omni-7B	Reference-guided	0.185	0.469	0.403	0.356	0.522 [0.464, 0.573]

N = 600; MOS-EC uses N = 300. Model names are simulation labels. Source: recomputed simulation data.

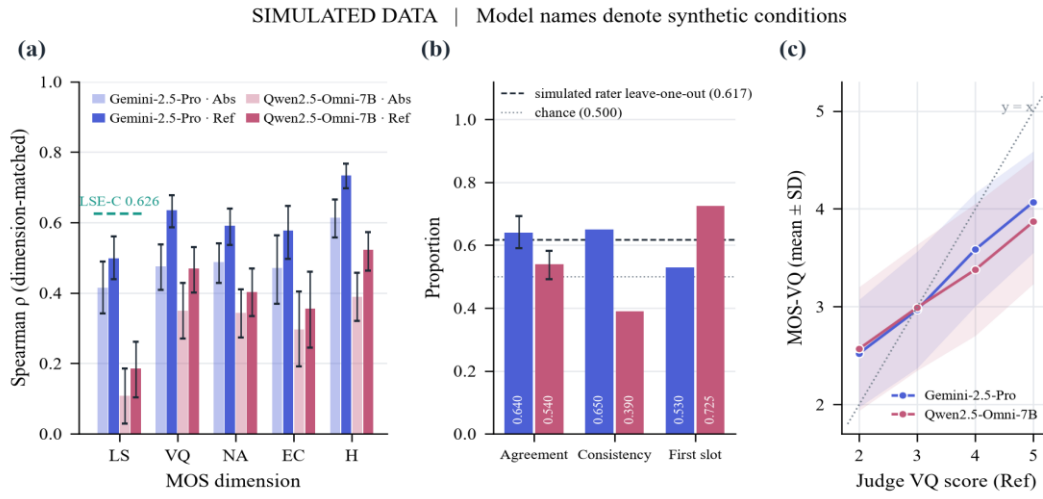


Figure 3. Simulated Judge Scores and Pairwise Responses

All panels use synthetic data. (a) Absolute and reference-configuration correlations with 95% source-group bootstrap intervals. (b) Agreement and first-position rates use 400 ordered verdicts per configuration; consistency uses 200 pairs. The simulated leave-one-out rater agreement is 0.617, with half-credit when the other two raters disagree. (c) Mean MOS-VQ $\pm$ one standard deviation for each reference-configuration VQ score. At score 5, the means are 4.071 and 3.869. Standard deviations describe dispersion, not confidence intervals.

4.2. Divergence between Synchronization and Perceptual Quality

Synthetic LSE-C and DOVER have Spearman correlation -0.048 . The conflict criterion identifies 127 of 600 records (21.2%): 73 Wav2Lip-labeled records, 36 Real records, and 18 SadTalker-labeled records. These correspond to within-condition rates of 36.5%, 18.0%, and 9.0%, respectively. Wav2Lip accounts for 73/127, or 57.5%, of conflicts. Among conflicting records, the sign of the standardized DOVER deviation matches that of MOS-H for 100/127 records (78.7%), compared with 27/127 (21.3%) for LSE-C. This is agreement about being above or below the pooled mean, not an accuracy measure against an independent binary ground truth. Figure 4 illustrates the condition rank reversal and the dependence of agreement on the grouping used.

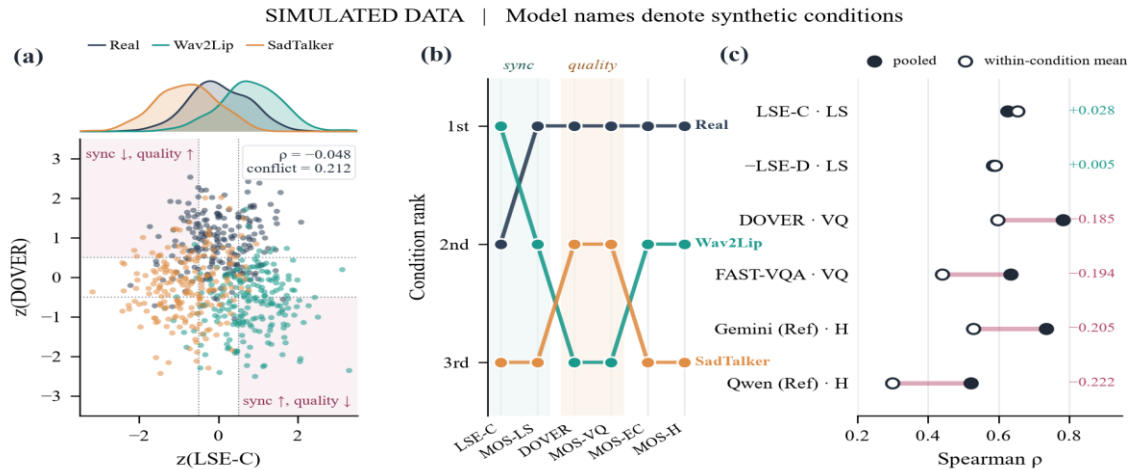


Figure 4. Synthetic Conflicts and Within Condition Agreement

(a) Globally standardized synthetic LSE-C and DOVER scores; 127 of 600 records meet the conflict criterion. (b) Condition ranks calculated from Table 3. (c) Difference between pooled correlation and the arithmetic mean of three within-condition correlations. The within-condition value changes by +0.028 for LSE-C, -0.185 for DOVER, -0.194 for FAST-VQA, -0.205 for the Gemini reference configuration, and -0.222 for the Qwen reference configuration. These differences describe the simulated conditions only.

4.3. Dependence and Order Sensitivity

A. Pooled and Within Condition Correlations

Table 6 separates pooled agreement from agreement within each synthetic condition. The LSE-C correlation with MOS-LS is 0.626 pooled and 0.654 when the three within-condition correlations are averaged. For DOVER versus MOS-VQ, the corresponding values are 0.782 and 0.597; for FAST-VQA, 0.634 and 0.440. The Gemini reference configuration changes from 0.734 to 0.529 on MOS-H, and the Qwen configuration changes from 0.522 to 0.300. These patterns show how condition separation can affect pooled correlation under the assumed latent structure. The within-condition mean is a descriptive arithmetic average, not an estimate of generalization to unseen speakers or generators. Source-group folds may contain different clips with the same MEAD speaker identifier in training and test sets.

Table 6. Pooled and Within Condition Spearman Correlations

Evaluator pair	Pooled	Real	Wav2Lip	SadTalker	Within mean
LSE-C vs MOS-LS	0.626	0.628	0.667	0.667	0.654
–LSE-D vs MOS-LS	0.585	0.569	0.597	0.603	0.590
DOVER vs MOS-VQ	0.782	0.553	0.663	0.574	0.597
FAST-VQA vs MOS-VQ	0.634	0.433	0.546	0.342	0.440
Gemini-2.5-Pro (Ref) vs MOS-H	0.734	0.464	0.595	0.528	0.529
Qwen2.5-Omni-7B (Ref) vs MOS-H	0.522	0.367	0.171	0.361	0.300

Pooled N = 600; each condition N = 200. Within mean is the arithmetic average of the three displayed condition correlations. Source: recomputed simulation data.

B. Simulated Position Bias

The Gemini-labeled configuration agrees with the simulated three-vote majority on 64.0% of its 400 ordered verdicts (95% interval 59.0%–69.2%) and is consistent across orders for 130/200 pairs (65.0%). The Qwen-labeled configuration agrees on 54.0%, is consistent for 78/200 pairs (39.0%), and chooses the first position in 290/400 verdicts (72.5%). The simulated rater leave-one-out score is 61.7%, but its two-vote comparison with half-credit for ties differs from the judge’s three-vote majority comparison; proximity does not establish human-level performance. Position bias in text judging motivates this diagnostic [19], but bias in the present audiovisual scenario is imposed in code. Agreement on consistent pairs is 93/130 (71.5%) and 47/78 (60.3%), retaining 65.0% and 39.0% of pairs. This selected-subset result does not improve performance on discarded pairs. Fréchet Video Distance is a set-level measure [20]; the random-feature distances in Table 3 are only computational stand-ins.

5. Discussion and Future Work

5.1. Interpretation of the Method Validation

The simulation demonstrates that the same analysis pipeline can detect condition-level rank reversals, differences between pooled and within-condition correlations, and order-sensitive pairwise responses. It does not show that LSE-C is the most accurate real synchronization measure, that DOVER transfers best to talking faces, or that one multimodal model is superior to another. These outcomes depend on assumed latent coefficients and noise levels. In particular, the Wav2Lip score bonus, lower noise in reference-guided scores, and judge-specific position biases explicitly favor the patterns observed in Section 4. Interpreting them as discoveries about the named systems would be circular.

For an empirical application, dimension-specific scores and the construction of any composite should be reported explicitly. MOS-H assigns equal weight to three dimensions and excludes emotional consistency; conclusions about it therefore concern that particular composite. Fusion should be evaluated on held-out groups with

preprocessing confined to training data, and its uncertainty should account for refitting where the claim concerns a learned procedure. Pairwise evaluation should report both orderings, agreement denominators, retained coverage after filtering, and the treatment of ties. Pooled and within-generator results answer different questions and should be interpreted separately. These are requirements for a transparent protocol, not recommendations validated by the synthetic model rankings.

5.2. Limitations and Empirical Validation

The primary limitation is the absence of observed videos, model inference, and human responses. Nominal durations, resolutions, corpus labels, and speaker identifiers do not establish that suitable source clips or disjoint reference clips exist. A real study would need verified media, documented sampling and preprocessing, generator checkpoints, audio-preserving judge inputs, and recorded model settings and outputs. Human participation, consent, rating conditions, exclusions, and applicable ethics review would need to be documented from actual records. Source-level bootstrap intervals and cross-validation do not account fully for recurring speakers or crossed rater effects; speaker-aware splits and dependence-sensitive uncertainty estimates should be evaluated. The current fusion intervals hold fitted predictions fixed, and no parameter sensitivity analysis establishes stability across simulation settings. The reference protocol, semantic-affect score, and pairwise selection rule each require empirical validation. A subsequent study could test controlled audio offsets, multiple generator families, varied speaking styles, and independent overall-quality ratings. Until such evidence is available, the numerical findings remain properties of this synthetic dataset.

References

- [1] Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. V. (2020). A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 484–492). ACM. <https://doi.org/10.1145/3394171.3413532>
- [2] Tian, L., Wang, Q., Zhang, B., & Bo, L. (2024). EMO: Emote portrait alive — Generating expressive portrait videos with audio2video diffusion model under weak conditions. In *Computer Vision – ECCV 2024* (pp. 244–260). Springer. https://doi.org/10.1007/978-3-031-73010-8_15
- [3] Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., & Liu, Z. (2024). VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 21807–21818). IEEE. <https://doi.org/10.1109/CVPR52733.2024.02060>
- [4] Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., & Wang, F. (2023). SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8652–8661). IEEE. https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_SadTalker_Learning_Realistic_3D_Motion_Coefficients_for_Stylized_Audio-Driven_Single_CVPR_2023_paper.html
- [5] Yariv, G., Gat, I., Benaim, S., Wolf, L., Schwartz, I., & Adi, Y. (2024). Diverse and aligned audio-to-video generation via text-to-video model adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7), 6639–6647. <https://doi.org/10.1609/aaai.v38i7.28486>
- [6] Wu, H., Chen, C., Hou, J., Liao, L., Wang, A., Sun, W., Yan, Q., & Lin, W. (2022). FAST-VQA: Efficient end-to-end video quality assessment with fragment sampling. In *Computer Vision – ECCV 2022* (pp. 538–554). Springer. https://doi.org/10.1007/978-3-031-20068-7_31
- [7] Kou, T., Liu, X., Zhang, Z., Li, C., Wu, H., Min, X., Zhai, G., & Liu, N. (2024). Subjective-aligned dataset and metric for text-to-video quality assessment. In *Proceedings of the 32nd ACM International Conference on Multimedia* (pp. 7793–7802). ACM. <https://doi.org/10.1145/3664647.3680868>
- [8] Zhou, Y., Cao, J., Zhang, Z., Wen, F., Jiang, Y., Jia, J., Liu, X., Min, X., & Zhai, G. (2025). Who is a better talker: Subjective and objective quality assessment for AI-generated talking heads. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 12201–12211).

- https://openaccess.thecvf.com/content/ICCV2025/html/Zhou_Who_is_a_Better_Talker_Subjective_and_Objective_Quality_Assessment_ICCV_2025_paper.html
- [9] He, X., Jiang, D., Zhang, G., Ku, M., Soni, A., Siu, S., Chen, H., Chandra, A., Jiang, Z., Arulraj, A., Wang, K., Do, Q. D., Ni, Y., Lyu, B., Narsupalli, Y., Fan, R., Lyu, Z., Lin, B. Y., & Chen, W. (2024). VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 2105–2123). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.127>
- [10] Jiang, D., Ku, M., Li, T., Ni, Y., Sun, S., Fan, R., & Chen, W. (2024). GenAI Arena: An open evaluation platform for generative models. In Advances in Neural Information Processing Systems: Vol. 37. Datasets and Benchmarks Track. Curran Associates. <https://arxiv.org/abs/2406.04485>
- [11] Zhang, Z., Li, L., Ding, Y., & Fan, C. (2021). Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3661–3670). IEEE. https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_Flow-Guided_One-Shot_Talking_Face_Generation_With_a_High-Resolution_Audio-Visual_Dataset_CVPR_2021_paper.html
- [12] Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., & Loy, C. C. (2020). MEAD: A large-scale audio-visual dataset for emotional talking-face generation. In Computer Vision – ECCV 2020 (Lecture Notes in Computer Science, Vol. 12366, pp. 700–717). Springer. https://doi.org/10.1007/978-3-030-58589-1_42
- [13] Chung, J. S., & Zisserman, A. (2017). Out of time: Automated lip sync in the wild. In Computer Vision – ACCV 2016 Workshops (Lecture Notes in Computer Science, Vol. 10117, pp. 251–263). Springer. https://doi.org/10.1007/978-3-319-54427-4_19
- [14] Wu, H., Zhang, E., Liao, L., Chen, C., Hou, J., Wang, A., Sun, W., Yan, Q., & Lin, W. (2023). Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 20144–20154). IEEE. https://openaccess.thecvf.com/content/ICCV2023/html/Wu_Exploring_Video_Quality_Assessment_on_User_Generated_Contents_from_Aesthetic_ICCV_2023_paper.html
- [15] Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., & Misra, I. (2023). ImageBind: One embedding space to bind them all. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15180–15190). IEEE. https://openaccess.thecvf.com/content/CVPR2023/html/Girdhar_ImageBind_One_Embedding_Space_To_Bind_Them_All_CVPR_2023_paper.html
- [16] Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., & Dubnov, S. (2023). Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 1–5). IEEE. <https://doi.org/10.1109/ICASSP49357.2023.10095969>
- [17] Elizalde, B., Deshmukh, S., Al Ismail, M., & Wang, H. (2023). CLAP: Learning audio concepts from natural language supervision. In ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 1–5). IEEE. <https://arxiv.org/abs/2206.04769>
- [18] Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., & Shan, Y. (2024). EvalCrafter: Benchmarking and evaluating large video generation models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 22139–22149). IEEE. <https://doi.org/10.1109/CVPR52733.2024.02090>
- [19] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In Advances in Neural Information Processing Systems (Vol. 36, pp. 46595–46623). Curran Associates. https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html

- [20] Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., & Gelly, S. (2018). Towards accurate generative models of video: A new metric & challenges. arXiv:1812.01717. <https://arxiv.org/abs/1812.01717>