arXiv:2510.18014v1 [cs.CV] 20 Oct 2025

ManzaiSet: A Multimodal Dataset of Viewer Responses to Japanese Manzai Comedy

Kazuki Kawamura[1,2] Kengo Nakai[3] Jun Rekimoto[1,2]
[1]Sony CSL Kyoto [2]The University of Tokyo [3]Yoshimoto Kogyo Holdings Co., Ltd.
kwmr@acm.org nakai.kengo@yoshimoto.co.jp rekimoto@acm.org

Abstract

We present ManzaiSet, the first large-scale multimodal dataset of viewer responses to Japanese manzai comedy, capturing facial videos and audio from 241 participants watching up to 10 professional performances in randomized order (94.6% watched ≥8; analyses focus on n=228). This addresses affective computing's critical Western-centric bias. Three key findings emerge: (1) k-means clustering identified three distinct viewer types—"High & Stable Appreciators" (72.8%, n=166), "Low & Variable Decliners" (13.2%, n=30), and "Variable Improvers" (14.0%, n=32)—with significant heterogeneity of variance (Brown–Forsythe $p < 0.001$); (2) Individual-level analysis revealed a positive viewing-order effect (mean slope = 0.488, $t(227) = 5.42$, $p < 0.001$, permutation $p < 0.001$), contradicting fatigue hypotheses; (3) Automated humor classification (77 instances, 131 labels) plus viewer-level response modeling found no type-wise differences after FDR correction. The dataset enables culturally-aware emotion AI development and personalized entertainment systems tailored to non-Western contexts.

1. Introduction

Affective computing research suffers from a critical Western-centric bias. While models trained on Western datasets fail to generalize across cultures [7, 14], non-Western emotional expressions remain severely underrepresented in existing corpora. This gap is particularly acute for complex emotions like humor, where cultural context fundamentally shapes both expression and perception [15, 16].

Japanese manzai comedy offers an ideal testbed for addressing this gap. As a structured two-person comic dialogue with repeated setup-punchline-reaction cycles, manzai provides controlled stimulus conditions while maintaining ecological validity as genuine entertainment content [25]. The format's cultural specificity—rooted in dialectal wordplay, social hierarchy violations, and correction dynamics—enables investigation of how cultural norms shape emotional display and perception [26].

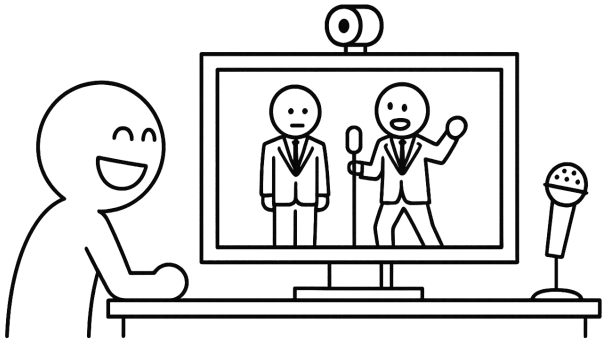


Figure 1. Data collection setup: participants watching manzai comedy videos at home via web browser.

In this paper, we present ManzaiSet, a dataset addressing these limitations by capturing multimodal responses from 241 Japanese viewers (analyses focus on n=228) watching identical professional manzai performances. Our dataset represents the first large-scale collection of viewer responses to culturally-specific comedy, comprising 191.8 hours of synchronized facial video and audio data. Beyond addressing the Western bias in affective computing, this resource enables investigation of how cultural context shapes emotional expression at scale.

We make three key contributions: (1) We introduce the largest multimodal dataset of spontaneous responses to controlled comedy stimuli from a non-Western culture; (2) We demonstrate the dataset's utility through robust statistical analyses revealing three distinct viewer types (72.8%, 13.2%, 14.0%) with significant variance heterogeneity (Brown–Forsythe $p < 0.001$), and a positive viewing-order effect (mean slope = 0.488, $p < 0.001$) contradicting fatigue assumptions; (3) We establish foundations for culturally-aware emotion AI development, with immediate applications in personalized entertainment and cross-

cultural emotion recognition.

2. Related Work

2.1. Controlled vs. In-the-Wild Emotion Datasets

Early multimodal corpora were carefully controlled and relatively small, enabling clear stimulus–response analysis but limiting ecological validity. DEAP [9] and MAHNOB-HCI [24] recorded viewers' physiology and faces while they watched standardized clips (mostly music or film excerpts). AMIGOS [18] extended this paradigm with individual and group sessions and personality/mood traits, while LIRIS-ACCEDE [4] provided a large set of Creative Commons movie excerpts annotated on valence/arousal.

In contrast, "in-the-wild" resources like AffectNet [19] and Aff-Wild2 [10] scaled to millions of frames, but they lack information about *what* viewers saw when expressions occurred, making causal stimulus–response analyses or within-subject comparisons impossible. Continuous, interactional datasets such as RECOLA [23] and SEWA DB [11] emphasize natural behavior (dyadic collaboration or ad viewing/video-chat) and provide dense valence/arousal and AU labels, but again are not comedy-focused and do not repeatedly expose many different individuals to the *same* entertainment stimuli.

Our dataset complements these lines by combining (i) standardized professional comedy stimuli, (ii) large-scale *individual-level* multimodal responses (facial video and audio), and (iii) naturalistic at-home viewing. This design enables both causal analyses tied to content timing and robust measurement of stable within-person differences.

2.2. Computational Humor: Content vs. Audience Focus

Much computational humor work treats laughter as a content-side signal without measuring the audience individually. UR-FUNNY [6] frames punchline detection from TED talks using audience laughter as labels, and SMILE [13] builds a *laugh reasoning* dataset (TED+sitcom) with textual explanations of *why* a clip is funny. Recent large-scale efforts collect stand-up comedy across languages, e.g., StandUp4AI [3], or smaller multilingual stand-up corpora [12]. Related work also includes audio-visual laughter resources centered on the signal itself, e.g., the MAHNOB Laughter database [21]. These corpora are invaluable for modeling humorous *content* or laughter acoustics, but the audience is either aggregated (room-level microphones) or implicit (no viewer recording), obscuring individual differences and making it difficult to link execution quality and timing to person-specific responses.

By contrast, our corpus records each viewer directly while they watch the same manzai performances, yielding synchronized facial/vocal reactions, per-viewer ratings, and randomized orderings. This "audience-of-one at scale" design supports analyses of consistent personal styles (e.g., selective vs. uniformly appreciative) and temporal dynamics (e.g., positive momentum) that are not accessible when only aggregated audience tracks are available.

2.3. Cultural Gaps in Affective Computing

Cultural variability in expression and perception is well established [5, 7, 16], and dataset composition strongly impacts model performance across regions [14]. While SEWA DB intentionally spans six cultural groups [11], most large-scale facial-expression resources remain Western-centric [19, 23].

Humor adds further culture- and language-dependent mechanisms. Our focus on Japanese manzai—a highly structured two-person comic dialogue—provides a culturally specific testbed with repeated setup–punchline–reaction cycles. In summary, existing datasets are either controlled but small (DEAP, MAHNOB-HCI), large-scale but uncontrolled (AffectNet, Aff-Wild2), or humor-focused but content-centric (UR-FUNNY, SMILE). Our contribution fills the missing quadrant: large viewer-centric, stimulus-controlled, culturally specific comedy responses.

3. Japanese Manzai Comedy

3.1. Structure and Characteristics

Manzai is a dyadic comic form that crystallized in Osaka's popular entertainment industry in the early 20th century, although it traces ritual origins to much older New Year performances [25]. The contemporary *shabekuri* (chatty) style centers on role-differentiated turn exchange between the *boke* (the source of incongruity) and the *tsukkomi* (the corrector who restores order), typically delivered at high speaking rates and with Kansai dialectal coloring [25, 26]. Conversation-analytic descriptions formalize this as repeated adjacency-pair cycles of *incongruity* $\rightarrow$ *resolution* (boke $\rightarrow$ tsukkomi), which frame audience laughables and their timing [26].

For computational work, these properties provide concrete affordances. First, the role structure and adjacency pairs yield naturally segmentable units (from boke onset to tsukkomi response) that can anchor temporal alignment of viewer signals. Second, the high density of verbal humor—puns (*dajare*) and phonological wordplay—creates text/audio features amenable to automatic analysis; for instance, Japanese imperfect puns are governed by psychoacoustic similarity constraints that can be modeled from signal [8]. Third, staging is minimal and dyadic, which reduces scene variability compared with sketch or ensemble comedy and helps isolate stimulus–response relationships. Finally, the form is embedded in a professional theater ecology (e.g., Yoshimoto Kogyo's Namba Grand Kagetsu in Os-

aka, a large dedicated comedy venue of ∼858 seats), ensuring that widely circulated *manzai* material exhibits consistent production quality and timing conventions that benefit controlled stimulus design [1].

3.2. Cultural Specificity

While manzai's mechanics are structurally regular, *how* amusement is displayed and perceived is shaped by Japanese cultural norms. Cross-cultural research on emotional display rules shows that Japan (and East Asian contexts more broadly) endorses relatively stronger regulation of overt expressivity in public settings compared with many Western contexts [17]. Within Japanese performance cultures, humor functions and audience responses are also conditioned by interactional norms—e.g., corrective tsukkomi moves index the re-establishment of social order after boke violations [20, 26].

Importantly, boke–tsukkomi dynamics are not only theatrical conventions; manzai-like sequences surface in everyday and mediated interactions, with recognizable patterns of "misunderstanding → sanctioning correction" that carry specific pragmatic meanings [27]. This pervasiveness helps explain why Japanese viewers can anticipate timing and corrective cues, leading to distributed, rapid micro-responses rather than isolated single punchline peaks. Linguistically, features such as dialectal play and *dajare* (phonologically constrained wordplay) are prevalent [8, 25].

Taken together, manzai offers an ideal compromise for affective computing: ecologically valid entertainment content with controlled, cyclic structure that facilitates precise mapping between stimulus events (boke/tsukkomi cycles, call-backs) and multimodal viewer responses, while foregrounding culturally specific display patterns that current "universal" models often miss.

4. Dataset Description

4.1. Data Collection

We recruited 241 participants (39 female, 201 male, 1 other; age range 19–63, mean age 30.0, median age 27.0). After excluding participants who viewed fewer than 8 videos, the final sample for analysis comprised 228 individuals. 56.0% (135/241) reported maximum frequency (5/5) for general comedy viewing, and 47.3% (114/241) reported high frequency (4/5) for manzai-specific viewing. All participants were native Japanese speakers and were compensated for their participation.

Data collection was conducted remotely using a web-based system that we developed specifically for this data collection. The system was built using Flask and was made accessible via the Internet during the data collection period. Participants accessed the experiment through a browser interface from their own computers at home, as illustrated in Figure 1. The system required participants to have a built-in or external camera and microphone that functioned properly.

The experimental procedure was as follows: participants first accessed the web application through a provided URL, entered a unique user ID and password, and completed a brief demographic questionnaire including age, gender, and comedy preferences. After displaying the questionnaire, the system enabled recording of participants' facial expressions and audio while playing the videos, and allowed participants to input ratings after each video playback. The system provided real-time feedback on camera positioning and audio levels to optimize data quality.

During video viewing sessions, the web application automatically recorded participants' facial expressions (captured at 640×360 resolution, 25 fps) and audio (48 kHz sampling rate) synchronized with the comedy content. After each video, participants provided subjective ratings (0–100 scale) for the content they had just viewed. The system automatically uploaded the recorded facial and audio data to our servers, with progress indicators to ensure successful data transmission. Participants could pause between videos but not during viewing sessions to maintain data consistency.

This remote setup enabled naturalistic viewing experiences in participants' familiar environments while maintaining consistent data quality through automated technical checks, standardized viewing procedures, and real-time quality monitoring.

We selected 10 manzai performances from the official recordings of "Densetsu no Ichinichi" (A Legendary Day), a special comedy event held on April 2–3, 2022, to commemorate the 110th anniversary of the major entertainment agency, Yoshimoto Kogyo Holdings Co., Ltd. The selected performances, captured at the Namba Grand Kagetsu theater, ranged from 4 to 6 minutes in duration and featured established manzai duos representing diverse comedy styles. Selection criteria included high video quality, the absence of copyrighted background music, and a variety of humor types (e.g., wordplay, physical comedy, and observational humor).

4.2. Data Statistics

The complete dataset comprises 2,301 viewing sessions from 241 participants with facial video recordings, as not all participants completed all 10 videos due to time constraints. Table 1 summarizes the key statistics of our dataset.

Participants provided subjective ratings (0–100 scale) for each video immediately after viewing. The distribution of ratings shows good variance (mean = 82.0, SD = 16.4), indicating diverse preferences and engagement levels. Importantly, 94.6% of participants (228 out of 241) completed at

Table 1. Summary statistics of the manzai viewer response dataset

Statistic	Value
Total participants	241
Total viewing sessions	2,301
Average sessions per participant	9.55
Total recording duration	191.8 hours
Video resolution	640×360
Frame rate	25 fps
Audio sampling rate	48 kHz
Average video rating (0–100)	82.0 ± 16.4

least 8 videos, providing sufficient data for within-subject analyses.

The facial video data was processed using OpenFace 2.0 [2] to extract facial action units (AUs) and head pose parameters. Participants were encouraged to use headphones, and our spot checks indicated that the recorded microphone channels were dominated by participant vocalizations (laughter, utterances) with minimal bleed-through from the stimulus audio; accordingly, we did not apply source separation. OpenFace reported high-quality facial landmark detection in 96.3% of frames (`success`$= 1$ and `confidence`> 0.9); manual spot-checks corroborated these estimates. The clarity of the audio was also verified, particularly in segments identified as containing laughter, to ensure high-quality data for analysis.

5. Methods

The analyses reported in this paper aim for robustness and reproducibility. Our total sample consists of 241 participants, but we used different subsets depending on the nature of the analysis. For Analysis A, which required stable individual profiles of differences, we included 228 participants who viewed at least 8 videos. For Analysis C, which analyzed responses to humor types, we restricted the sample to 207 participants who provided complete response data for all relevant instances. Restricting to complete cases may introduce selection bias; as a sensitivity check, we also fit the marginal model on all available viewer-by-instance rows (without a complete-case filter), and the coefficient patterns and FDR-adjusted inferences were unchanged. This approach ensures that we maximize the validity of each specific analysis.

6. Analyses and Results

We conducted three complementary analyses to demonstrate the dataset's research potential and establish baseline findings for future work. These analyses were designed to showcase different aspects of the data while revealing unique characteristics of Japanese comedy responses.

6.1. Analysis A: Individual Differences and Viewer Typology in Humor Appreciation

6.1.1. Objective

Humor appreciation is highly subjective, yet little is known about the systematic patterns underlying individual differences in comedy evaluation. This analysis aimed to address two fundamental questions: (1) How much do individuals differ in their humor appreciation patterns, and are these differences random or structured? (2) Can we identify distinct viewer types based on their rating behaviors? Understanding these individual differences is crucial for developing personalized recommendation systems and for advancing theoretical models of humor processing that account for viewer heterogeneity.

6.1.2. Method

We analyzed participants who viewed at least 8 videos (n=228) to ensure robust individual profiles. For each participant, we computed per-participant feature vectors comprising: mean rating, within-person SD, coefficient of variation (CV), rating range, half-session shift ($\overline{r}_{\text{late}} - \overline{r}_{\text{early}}$), and the per-participant viewing-order slope (ratings regressed on positions 1–10). The previously reported split-half rank-order reliability (r=0.866) is a sample-level stability metric across participants and was not used as a per-participant feature in clustering. We acknowledge that some variance-oriented features (e.g., SD, CV, Range) are correlated; dimensionality reduction (e.g., PCA) can summarize them, and we report clustering results with standardized features.

Variance heterogeneity test. To test heterogeneity of variance across the discovered clusters, we applied the Brown–Forsythe test (Levene's test with median centering) to per-participant within-person variability measures. Our primary indicator was the across-video standard deviation of ratings (within-person SD); as sensitivity checks we repeated the test for the coefficient of variation (CV) and rating range. Tests were performed across the three clusters with degrees of freedom $(2, N-3)$ where N is the number of clustered participants.

Our analysis proceeded in three stages. First, we quantified inter-individual variability by computing the coefficient of variation for each feature. Second, we assessed the temporal stability of individual rating styles using split-half reliability of per-participant mean ratings across randomized half-sets (correlation across participants); this sample-level stability metric was *not* included as a per-participant feature for clustering. Finally, we identified viewer typologies by applying k-means clustering on the standardized features. The optimal number of clusters was determined using silhouette analysis; complementary diagnostics were mixed—Calinski–Harabasz and stability (ARI) increased at k=4 while silhouette decreased—so we selected k=3 for

parsimony and interpretability. Cluster stability was evaluated via bootstrapping (100 iterations) using the Adjusted Rand Index (ARI).

6.1.3. Results

Our analysis revealed striking heterogeneity in how participants evaluate comedy. Note that the reported *mean rating CV = 0.16* refers to the *across-participant CV of per-participant mean ratings* (a between-person summary), whereas cluster descriptions later report *within-person variability* (across-video SD, CV, and range) computed per participant and then aggregated within clusters. While average enjoyment levels were relatively similar (across-participant CV of means = 0.16), viewers differed dramatically in their rating patterns. We observed substantial inter-individual variability, particularly in the coefficient of variation of ratings (0.98), rating standard deviation (0.81), and rating range (0.80)—these values are across-participant CVs of the respective features.

k-means clustering clearly identified three distinct viewer types. The final solution achieved an overall silhouette score of 0.523 with bootstrapped ARI = 0.865 (100 iterations).

Unless otherwise noted, clusters are reported in descending order of mean rating: Type 1, Type 3, Type 2 (means $\approx$ 87.0, 71.2, 65.6, respectively), matching the ordering used in the Appendix.

Consistent with descriptive differences, the Brown–Forsythe test indicated significant heterogeneity of variance across clusters in within-person variability: for within-person SD, $F(2, 225)$=24.87, p=1.74 $\times$ 10^{-10}; for CV, $F(2, 225)$=33.30, p=2.15 $\times$ 10^{-13}; and for rating range, $F(2, 225)$=23.97, p=3.65 $\times$ 10^{-10}. These results corroborate the variance differences summarized in the Appendix materials.

Type 1: "High & Stable Appreciators" (n=166, 72.8%)
Consistently high ratings (mean $\approx$ 87.0, 95% CI: [86.0–88.1]) with minimal within-person variation (SD $\approx$ 4.84; CV $\approx$ 0.058), small range (mean range $\approx$ 15.6), and slightly positive viewing-order slope on average (95% CIs via percentile bootstrap, 10,000 resamples).

Type 3: "Variable Improvers" (n=32, 14.0%)
Moderate average ratings (mean $\approx$ 71.2, 95% CI: [67.9–74.5]) but high variability (SD $\approx$ 17.9; CV $\approx$ 0.263), wide range ($\approx$ 57.8), and a markedly positive order slope (improving across the session).

Type 2: "Low & Variable Decliners" (n=30, 13.2%)
Lower average ratings (mean $\approx$ 65.6, 95% CI: [58.6–71.5]) with high within-person variation (SD $\approx$ 13.9; CV $\approx$ 0.218), wide range ($\approx$ 43.7), and negative order slope (declining across the session).

The clustering solution demonstrated high stability (bootstrapped ARI = 0.865). Despite these group differences, individuals showed remarkable temporal consistency in their personal rating styles. The split-half reliability was exceptionally high at r=0.866 ($p < 0.001$), suggesting that humor appreciation style is a stable, trait-like characteristic.

6.1.4. Implications

This analysis provides three key insights for affective computing and humor research. First, the high rank-order consistency of per-participant mean ratings across randomized halves (r=0.866) demonstrates that humor appreciation styles are trait-like rather than state-dependent, enabling reliable personalization in recommendation systems.

Second, the existence of distinct viewer types challenges one-size-fits-all approaches to emotion recognition, as models trained on "average" responses may perform poorly for the ˜27% of variable viewers whose response patterns differ qualitatively from the majority (including both decliners and improvers).

Third, the observed consistency suggests potential cultural differences in how humor preferences are formed and expressed, possibly reflecting the structured nature of manzai comedy itself, though direct cross-sample comparisons are beyond the scope of this paper.

6.2. Analysis B: Temporal Dynamics of Collective Response

6.2.1. Objective

While traditional humor research has focused on static measures of amusement, understanding the temporal dynamics of audience responses is crucial for developing predictive models of comedy engagement. This analysis aimed to investigate how viewing patterns change over time, specifically probing the effects of fatigue or habituation.

6.2.2. Method

We analyzed viewing-order effects using the same 228 participants (each watched $\geq$8 videos). We computed individual slopes by regressing ratings on viewing position (1–10 treated as numeric). The population-level effect was assessed using a one-sample t-test against zero (two-sided) on these slopes, with Wilcoxon signed-rank test as a robustness check. We report the mean slope with 95% confidence intervals from both t-distribution and participant-level percentile bootstrap (10,000 resamples). Additionally, we conducted a permutation test (10,000 iterations) where viewing order was shuffled within each participant to generate a null distribution, with Monte Carlo p-values calculated as $(r + 1)/(B + 1)$.

6.2.3. Results

Our temporal analysis revealed a pattern that challenges common assumptions about viewer fatigue in comedy con-

sumption. Contrary to expectations of a fatigue-induced decline, we observed a positive viewing-order effect. Mean ratings increased from 77.95 (SD=19.70) at position 1 to 83.15 (SD=16.83) at position 10, a gain of 5.20 points (6.7% increase).

Individual-level analysis showed a mean slope of 0.488 rating points per position ($t(227) = 5.42$, $p < 0.001$, 95% CI: [0.312, 0.664]). The permutation test confirmed this effect ($p < 0.001$, Monte Carlo). For a mixed-effects specification, see the Appendix: an LMM reached $\hat{\beta}_1 \approx 0.49$ with overlapping confidence intervals. Note that aggregated position means (e.g., $+5.20$ from position 1 to 10) need not equal the average individual slope times nine increments due to weighting, missingness, and nonlinearity; the effects are directionally consistent. Lag-1 autocorrelations were computed *after removing each participant's linear trend* (detrended residuals); within-participant lag-1 correlations, Fisher z-transformed and averaged, were $\bar{r} = -0.070$ (95% CI: $[-0.108, -0.032]$), indicating negligible carry-over despite statistical detectability. For completeness, raw (non-detrended) lag-1 values are summarized in the Appendix; conclusions are unchanged.

6.2.4. Implications

This temporal analysis provides novel insights for both affective computing and entertainment technology. The positive momentum effect contradicts standard assumptions about viewer fatigue. Current recommendation algorithms often assume declining engagement, but our data suggests that comedy viewing can create increasing satisfaction, with implications for playlist optimization and session management.

6.3. Analysis C: Humor Type and Response Intensity Correlation

6.3.1. Objective

This analysis investigated the relationships between specific humor mechanisms employed in manzai and the intensity of viewer responses. The goal was to identify which humor types were most effective in eliciting laughter and whether different mechanisms produced distinct response patterns.

6.3.2. Method

Our analysis pipeline consisted of four main stages: (1) Transcription of all 10 manzai videos using Whisper [22]; (2) Humor classification of 77 instances into nine categories (multi-label allowed) using GPT-5-mini (wordplay, exaggeration, unexpected twists, self-deprecation, impersonation, repetition, misunderstanding, physical gags, cultural references); (3) Response detection from the recorded audio channels: laughter events were detected using frame-level RMS, zero-crossing rate (ZCR), and spectral features with peak picking (3 s "instant" and 10 s "cumulative" windows anchored at humor onsets); (4) Statistical analysis. For reproducibility, we used GPT-5-mini (release 2025-01) with temperature=0; prompt templates/outputs and detection thresholds are provided in the supplementary material.

Because instances could carry multiple humor labels, we analyzed viewer-by-instance responses in a frequentist framework using a marginal logistic model estimated via generalized estimating equations (GEE) with participant-level clustering (exchangeable working correlation) and video fixed effects. Let $y_{i,p} \in \{0, 1\}$ denote whether viewer p responded (audio-based laughter detected) at instance i. We fit

$$y_{i,p} \sim \text{Bernoulli}(p_{i,p}), \tag{1}$$

$$\text{logit}(p_{i,p}) = \alpha + \sum_j \beta_j \mathbb{1}\{\text{label}_{ij}\} + \sum_v \gamma_v \mathbb{1}\{\text{video}_i{=}v\}, \tag{2}$$

and accounted for within-viewer correlation via the GEE working correlation. As a robustness fallback when GEE failed to converge, we fit an aggregated binomial GLM (logit link) with video fixed effects and HC3 robust covariance, stabilizing rare or separated cells with add-0.5 pseudo-counts. Pairwise contrasts among humor types were evaluated with Wald tests and Benjamini–Hochberg FDR correction. As a diagnostic, we also obtained permutation-based p-values by shuffling humor-type labels across instances while preserving per-instance label multiplicities (10,000 iterations); these results are summarized in the Appendix.

6.3.3. Results

Analysis covered 77 humor instances across 10 videos identified by GPT-5-mini, with 207 users providing complete response data. Humor classification assigned 131 labels across 77 instances (multi-label allowed). Label counts were: unexpected twists (*igai-sei*, 35), repetition (*tendon*, 28), exaggeration (*oogesa*, 19), misunderstanding (*kanchigai*, 17), self-deprecation (*jiko-hige*, 9), wordplay (*dajare*, 8), impersonation (*monomane*, 5), physical comedy (*karada-gag*, 5), and cultural references (*bunka-neta*, 5). In the marginal (GEE) analysis with participant clustering and video fixed effects (with aggregated GLM as a robust fallback), we did not detect statistically reliable differences between humor types after FDR correction (all adjusted $p > 0.05$); effect estimates and uncertainty summaries are visualized in the Appendix (pairwise contrasts, permutation p-values, and odds-ratio matrices).

This lack of statistical significance does not prove the absence of any effect. Rather, it indicates that within the design and precision of this study, medium-to-large effects between humor types were not supported, though smaller effects cannot be ruled out. This could reflect insufficient statistical power, particularly given the imbalanced sample sizes across categories (e.g., n=5 for physical comedy), or

the limitations of the statistical model mentioned earlier. These findings suggest that humor effectiveness in manzai may depend more on execution quality and contextual factors than on categorical mechanisms alone.

7. Discussion

7.1. What the data say—and what they do not

Our analyses yield three takeaways. First, viewer-specific appreciation shows high rank-order stability across participants (split-half r=0.866); k-means robustly separates three types (72.8%, 13.2%, 14.0%) with significant variance heterogeneity (Brown–Forsythe $p < 0.001$). Second, individual-level analysis reveals positive viewing-order effects: mean slope = 0.488 rating points per position ($t(227) = 5.42$, $p < 0.001$), confirmed by permutation test (p ¡ 0.001). Third, a comparison across nine humor categories finds no statistically reliable differences after FDR correction; with our modeling caveats, this suggests execution quality and shared context may outweigh categorical mechanisms, without ruling out small effects.

7.2. Implications for affective computing

The trait-like stability (rank-order consistency of per-participant mean ratings across randomized halves, split-half r=0.866) recommends personalization as a first-class objective: few-shot conditioning on a viewer's recent signals (ratings, micro-expressions, vocal bursts) should lower error relative to global models. The observed momentum argues for engagement-aware sequencing—playlist ordering and session controllers that anticipate rising satisfaction rather than fatigue.

For cross-cultural emotion recognition, this corpus offers a non-Western distribution against which to calibrate and stress-test models trained on Western data. Practical routes include domain adaptation and cross-domain evaluation protocols that quantify generalization gaps and encourage feature learning sensitive to context rather than fixed humor taxonomies.

For comedy performance training, the data afford response-aligned diagnostics: per-instance curves time-locked to *boke–tsukkomi* cycles, laughter-onset rate/latency, and multimodal cues (e.g., AU dynamics, vocal bursts) can be summarized into interpretable feedback for rehearsal and editing. Because no category dominated, coaching should emphasize delivery, pacing, and timing.

More broadly, these findings support laugh-aware interactive systems—agents that detect and anticipate laughables, modulate tone with light, culturally appropriate amusement displays, and schedule content to sustain momentum—providing an immediate bridge from dataset to application while remaining model-agnostic.

7.3. Limitations and Future Directions

Analysis B uses individual-level slopes avoiding aggregation bias, and we add a complementary linear mixed model (LMM) specification in the Appendix. Analysis C was reanalyzed using a frequentist marginal model (GEE; participant clustering, video fixed effects), with an aggregated binomial GLM (HC3) as a robustness fallback; conclusions were consistent with simpler GLM summaries. The "no difference" result reflects lack of statistical significance, not equivalence; equivalence tests (TOST) can probe practically negligible effects. Some clustering features were redundant; dimensionality reduction (e.g., PCA) before clustering may yield cleaner profiles.

The dataset currently comprises Japanese viewers only; realizing cross-cultural recognition requires adding further cultural groups or combining with non-Japanese corpora and evaluating domain adaptation explicitly.

Looking forward, we see three model-agnostic avenues that align with our findings while remaining abstract by design: (i) *measurement*—finer localization of laughter and micro-responses, plus within-session contagion/carryover modeling; (ii) *prediction*—anticipatory scoring of laughability and moment-to-moment engagement conditioned on execution cues; (iii) *interaction*—consented, closed-loop studies where interactive systems adapt in real time to detected amusement, with privacy, safety, and cultural appropriateness as first-class constraints. These directions connect analysis to applications without committing to a specific architecture.

8. Conclusion

We presented ManzaiSet, the first large-scale multimodal dataset of viewer responses to Japanese manzai comedy, featuring synchronized facial video and audio from 241 participants (analyses: n=228). We identified three viewer profiles—"High & Stable Appreciators" (72.8%), "Low & Variable Decliners" (13.2%), and "Variable Improvers" (14.0%)—with variance heterogeneity (Brown–Forsythe $p < 0.001$), observed a positive viewing-order effect (mean slope = 0.488, $p < 0.001$), and found no reliable differences across nine humor categories after FDR.

ManzaiSet addresses affective computing's critical Western-centric bias, providing a foundation for culturally-aware emotion AI development. It enables immediate applications in personalized entertainment systems, cross-cultural emotion recognition, and comedy performance training. By demonstrating how cultural context fundamentally shapes emotional expression and perception, our work represents a crucial step toward developing emotion AI that appreciates and adapts to cultural diversity.

Acknowledgments

This work was supported by JST Moonshot R&D Grant JPMJMS2012. We extend our sincere gratitude to Yoshimoto Kogyo Holdings Co., Ltd. for their invaluable cooperation in data collection, particularly for providing access to professional manzai performances and supporting the experimental setup that made this research possible. We also appreciate the assistance of Hongil Yang, Shunsuke Noma, and Naoki Shibuya from Yoshimoto Kogyo Holdings Co., Ltd. in facilitating data acquisition for this study.

References

[1] Namba grand kagetsu (ngk) — osaka official tourism. https://en.osaka-info.jp/spot/namba-grand-kagetsu/, 2025. Accessed: 2025-08-08.

[2] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66, 2018.

[3] Valentin Barriere, Nahuel Gomez, Léo Hemamou, Sofía Callejas, and Brian Ravenet. StandUp4AI: A new multilingual dataset for humor detection in stand-up comedy videos. arXiv:2505.18903, 2025.

[4] Yoann Baveye, Emmanuel Dellandréa, Christel Chamaret, and Liming Chen. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55, 2015.

[5] Hillary Anger Elfenbein and Nalini Ambady. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2):203–235, 2002.

[6] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China, 2019. Association for Computational Linguistics.

[7] Rachael E. Jack, Oliver G. B. Garrod, Hui Yu, Roberto Caldara, and Philippe G. Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244, 2012.

[8] Shigeto Kawahara and Kazuko Shinohara. The role of psychoacoustic similarity in japanese puns: A corpus study. *Journal of Linguistics*, 45(1):111–138, 2009.

[9] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

[10] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. In *Proc. British Machine Vision Conf. (BMVC)*, page 297, Cardiff, UK, 2019. Paper 297.

[11] Jean Kossaifi et al. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(3):1022–1040, 2021.

[12] Anna Kuznetsova and Carlo Strapparava. Multimodal and multilingual laughter detection in stand-up comedy videos. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11884–11889, Torino, Italia, 2024. ELRA and ICCL.

[13] Hyun Lee, Sung-Bin Kim, Seungju Han, Youngjae Yu, and Tae-Hyun Oh. SMILE: Multimodal dataset for understanding laughter in video with language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1149–1167, Mexico City, Mexico, 2024. Association for Computational Linguistics.

[14] Martin Lukac, Gulnaz Zhambulova, Kamila Abdiyeva, and Michael Lewis. Study on emotion recognition bias in different regional groups. *Scientific Reports*, 13(1):8414, 2023.

[15] Rod A. Martin and Thomas Ford. *The Psychology of Humor: An Integrative Approach*. Academic Press, 2 edition, 2018.

[16] David Matsumoto. Culture and emotional expression. In *Understanding Culture: Theory, Research, and Application*, pages 271–287. Psychology Press, 2013.

[17] David Matsumoto, Seung Hee Yoo, and Johnny Fontaine. Mapping expressive differences around the world: The relationship between emotional display rules and individualism versus collectivism. *Journal of cross-cultural psychology*, 39 (1):55–74, 2008.

[18] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, 12(2):479–493, 2021.

[19] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.*, 10(1):18–31, 2019.

[20] Kimie Oshima. Rakugo and humor in japanese interpersonal communication. In *Understanding Humor in Japan*, pages 99–109. Wayne State University Press, Detroit, MI, 2006.

[21] Stavros Petridis, Brais Martinez, and Maja Pantic. The mahnob laughter database. *Image Vision Comput.*, 31(2): 186–202, 2013.

[22] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356, 2022.

[23] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2013.

[24] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012.

[25] Joel F. Stocker. Manzai: Team comedy in japan's entertainment industry. In *Understanding Humor in Japan*, pages 51–74. Wayne State University Press, Detroit, MI, 2006.

[26] Hideo Tsutsumi. Conversation analysis of boke-tsukkomi exchange in japanese comedy. *New Voices*, 5:147–173, 2011.

[27] Halina Zawiszová. Manzai-like humour sequences in japanese everyday communication and media. *Gakushuin Journal of International Studies*, 7:41–69, 2021.

A. Additional Analysis Results

This appendix presents additional figures and detailed results that could not be included in the main paper due to space constraints.

A.1. Analysis A: Individual Differences and Viewer Typology

Effect size (lnCVR). We quantify pairwise relative variability between clusters via the log coefficient of variation ratio (lnCVR), defined for any two clusters a, b as $\ln \mathrm{CVR}_{a,b} = \ln(\mathrm{CV}_a/\mathrm{CV}_b)$. Reported SDs refer to within-person across-video variability; CVs are computed per participant as SD divided by mean rating and then aggregated within cluster. In our data (clusters ordered by descending mean rating as Type 1, Type 3, Type 2), point estimates were $\ln \mathrm{CVR}_{\text{Type 2,Type 1}} \approx 1.33$, $\ln \mathrm{CVR}_{\text{Type 3,Type 1}} \approx 1.52$, and $\ln \mathrm{CVR}_{\text{Type 3,Type 2}} \approx 0.19$. Confidence intervals can be obtained via a normal approximation with delta-method standard errors. Delta-method standard errors for lnCVR follow the standard approximation for ratios of sample CVs; details and code are provided in the supplementary material.

Table 2. Cluster-number diagnostics on standardized features. Silhouette favors $k = 3$; stability (ARI) is high for $k = 3$ and increases further at $k = 4$ while silhouette drops; we therefore select $k = 3$ for parsimony and interpretability.

k	Silhouette	Calinski–Harabasz	Stability (ARI)
2	0.515	148.14	0.774
3	0.523	151.87	0.864
4	0.476	158.03	0.929
5	0.488	149.93	0.776

A.2. Analysis B: Temporal Dynamics of Collective Response

Autocorrelation details. Lag-1 autocorrelations were computed after removing each participant's linear trend (detrending). For reference, raw (non-detrended) lag-1 correlations yielded a similar average with overlapping uncertainty, consistent with negligible carry-over once trend is accounted for. We also summarize the distribution of per-participant slopes (mean and 95% CI) to reconcile the positive average slope with conservative temporal classifications ("ascending" requires monotonicity and a minimum slope threshold) (see Fig. 2).

For reference, raw (non-detrended) lag-1 correlations averaged $\bar{r}_{\text{raw}} = -0.070$ (95% CI: $[-0.110, -0.031]$).

LMM robustness check. As a complementary analysis, we fit a linear mixed-effects model at the rating-by-position

level:

$$\text{rating}_{p,t} \sim \beta_0 + \beta_1\,\text{position}_{p,t} + (1{+}\text{position}\,|\,p). \quad (3)$$

The fixed effect of position remained positive and statistically significant ($\hat{\beta}_1 \approx 0.49$, 95% CI $[0.31, 0.66]$, $p < 0.001$), consistent with the individual-slope and permutation results reported in the main text.

A.3. Analysis C: Humor Type and Response Intensity Correlation

Table 3. Humor type effectiveness with statistical validation. Response rates shown with 95% bootstrap confidence intervals. No significant differences found after FDR correction (all adjusted p>0.05). *N denotes the number of humor instances; multi-labels are allowed, so percentages across types can exceed 100%.* Counts and rates in this table match the instance-level tallies reported in the main text (N=77, multi-label allowed).

Humor Type	N	Instant (%)	Cumul. (%)
igai-sei (unexpected)	35	54.1 (51.8–56.4)	54.1 (51.7–56.4)
oogesa (exaggeration)	19	56.1 (53.4–58.8)	56.1 (53.3–58.8)
kanchigai (misunderstanding)	17	51.4 (48.3–54.8)	51.4 (48.3–54.8)
jiko-hige (self-deprecation)	9	57.5 (53.9–61.4)	57.5 (53.9–61.4)
dajare (wordplay)	8	46.7 (44.7–48.8)	46.7 (44.7–48.9)
tendon (repetition)	28	53.5 (51.6–55.5)	53.5 (51.6–55.5)
karada-gag (physical)	5	50.7 (50.2–51.7)	50.7 (50.2–51.7)
monomane (impersonation)	5	50.9 (48.4–52.7)	50.9 (48.4–52.7)
bunka-neta (cultural)	5	59.0 (53.8–64.3)	59.0 (53.8–64.3)

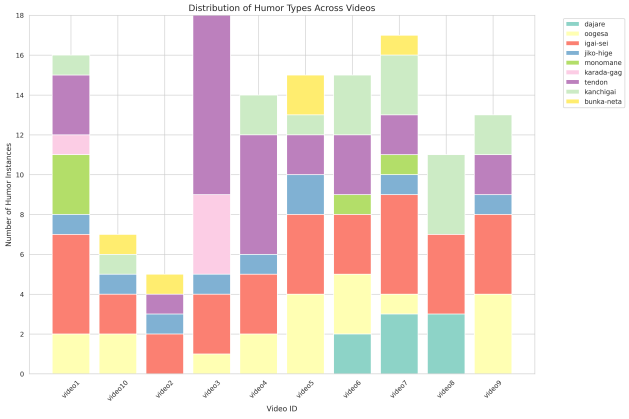


Figure 3. Distribution of humor types across 77 instances in 10 manzai videos (multi-label classification). Unexpected twists (igai-sei) were most common (45.5% of instances), followed by repetition (tendon, 36.4%) and exaggeration (oogesa, 24.7%). Percentages sum to ¿100% as instances could have multiple labels. GPT-5-mini achieved 67.4% average confidence.

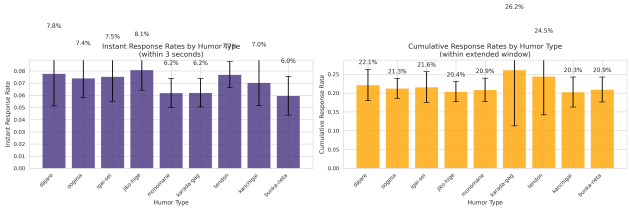


Figure 4. Effectiveness comparison of different humor types measured by instant and cumulative response rates with 95% bootstrap confidence intervals. Physical comedy (karada-gag) showed the highest cumulative response rate (26.2%, 95% CI: 11.3-41.0%), though with wide uncertainty due to small sample size (n=5). After FDR correction, no humor type achieved statistical significance over others (all adjusted p>0.05), with effect sizes remaining small (OR $<$ 1.5).

A.4. Statistical Validation Details

Table 4. Power considerations for humor type comparisons. In a multi-label, viewer-by-instance design analyzed via marginal models with participant clustering (GEE) and aggregated binomial GLM (robust HC3) as needed, ANOVA-style f-based power calculations are inappropriate. Simulation-based assessments reflecting the clustered data structure indicate limited sensitivity outside medium-to-large effects with the current sample. We therefore refrain from reporting ANOVA f tables and instead emphasize effect estimates with uncertainty intervals.

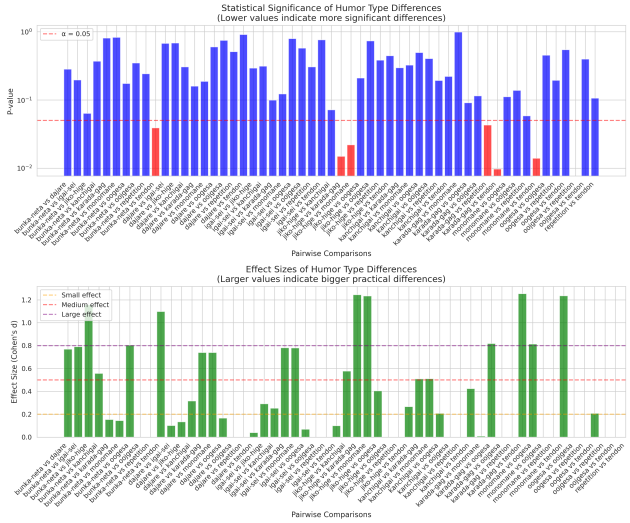


Figure 5. Comprehensive statistical validation results. (a) Pairwise comparisons with FDR correction showing no significant differences. (b) Permutation test p-value distribution confirming null findings. (c) Effect size matrix with OR values all below 1.5 (small effect threshold). (d) Bootstrap confidence interval overlap demonstrating minimal separation between humor types.

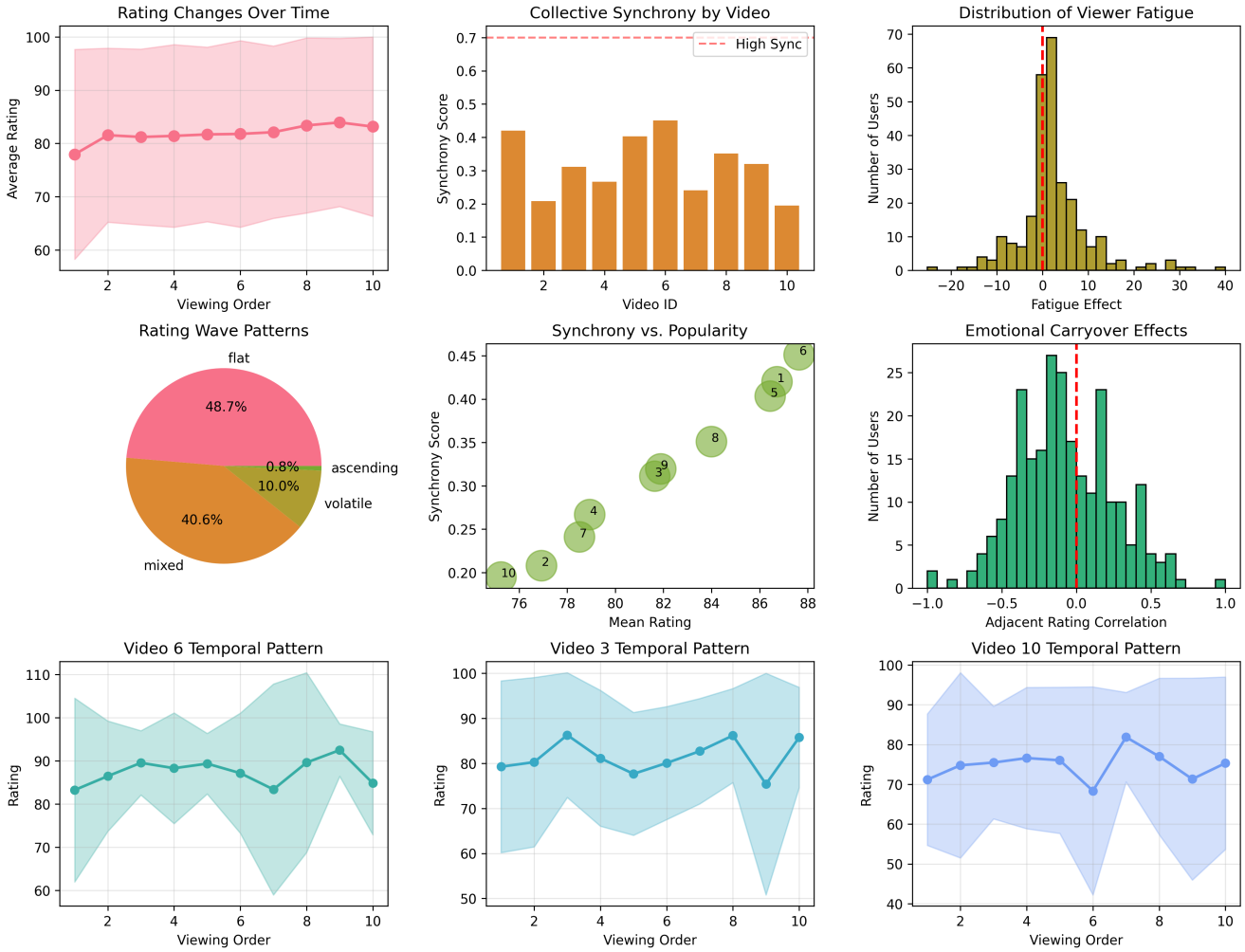


Figure 2. Integrated temporal analysis combining individual temporal patterns with collective response dynamics. Four distinct temporal response patterns are identified: flat (48.7%), mixed (40.6%), volatile (10.0%), and ascending (0.8%). "Ascending" is defined conservatively as monotonic non-decreasing with a minimum slope threshold; many participants within the "mixed" class still exhibit net-positive linear slopes, consistent with the positive viewing-order effect reported in the main text. Percentages may not sum to 100 due to rounding.