

Timing is Everything: Temporal Scaffolding of Semantic Surprise in Humor

Yuxi Ma^{1,2,3,5,6,*}, Yongqian Peng^{1,2,4,5,6,*}, Junchen Lyu^{1,4}, Chi Zhang^{3,5}✉, and Yixin Zhu^{2,1,5,6}✉

*Equal contributors Project Website: <https://mayuxi.com/research/talkshow>

¹ Institute for Artificial Intelligence, Peking University ² School of Psychological and Cognitive Sciences, Peking University

³ School of Intelligence Science and Technology, Peking University ⁴ Yuanpei College, Peking University

⁵ State Key Laboratory of General Artificial Intelligence, Peking University

⁶ Beijing Key Laboratory of Behavior and Mental Health, Peking University

Abstract

Humor is a fundamental cognitive phenomenon in which humans derive pleasure from the expectation violations and their resolution, exemplifying the brain’s dynamic capacity for predictive processing. Classical humor theories emphasize semantic incongruity as the primary driver of amusement, yet overlook temporal dynamics despite comedians’ intuition that “timing is everything.” The extent to which temporal structure contributes to humor appreciation and how it interacts with semantic content remains poorly understood. Here, we propose the Dual Prediction Violation (DPV) framework to capture the interplay between content and timing. By analyzing 828 professional Chinese stand-up performances, we show that temporal features substantially outweigh semantic incongruity in predicting audience appreciation. Specifically, we find that peak semantic violations matter more than average incongruity levels, and pauses systematically lengthen before high-surprise punchlines—a strategic coupling that distinguishes successful from unsuccessful performances. These findings reframe humor as temporally scaffolded, where timing and semantic content operate in strategic coordination rather than independently. Our DPV framework bridges humor theory with predictive processing, demonstrating that temporal structure plays a central role in naturalistic humor appreciation with implications for understanding multi-scale prediction integration in linguistic processing.

Keywords: humor cognition; predictive processing; temporal dynamics; semantic incongruity; naturalistic language

Introduction

Humor offers a unique window into how humans manage expectations and derive reward from prediction violations. Classical theories, particularly the two-stage incongruity-resolution model (Suls, 1972, 1983), propose that humor arises when audiences detect and resolve semantic violations. Recent developments in the predictive processing framework have refined this view (Van de Cruys et al., 2017), proposing that humor emerges not merely from resolving incongruity but from the *efficiency* of resolution. Critically, Van de Cruys et al. (2017) argue that pleasure depends on the *speed* of understanding—the brain rewards itself for unexpectedly rapid error correction, a metacognitive pleasure signal tied to resolution dynamics rather than content alone.

This raises a fundamental question: what determines resolution efficiency? We propose that the answer lies in temporal structure. Existing frameworks focus overwhelmingly on *what* violates expectations (semantic content), while neglecting *when* and *how* these violations unfold. Yet temporal scaffolding—pacing, pauses, rhythmic patterns—may fundamentally shape whether prediction errors resolve swiftly (producing pleasure) or stall (producing confusion). Timing does

not merely deliver semantic surprise; it may actively gate cognitive access to that content.

This theoretical gap is partly methodological. As Moran et al. (2004) noted, prior literature has relied heavily on decontextualized stimuli: written jokes, isolated punchlines, or brief text strings. While such controlled materials successfully isolate semantic incongruity mechanisms, they eliminate the rich temporal dynamics that characterize professional comedic delivery. Consequently, we lack quantitative evidence for how temporal dynamics interact with semantic surprise to drive audience engagement in ecologically valid settings.

The current study addresses this gap by analyzing naturalistic performances at scale. We created a dataset of 828 professional Chinese stand-up comedy performances from major televised competitions (2017–2025), representing 86 hours of performance time with accompanying audience appreciation scores. While the televised format introduces editing decisions, professional editors work to maintain narrative coherence and pacing that align with the audience’s cognitive rhythms rather than arbitrarily distorting the temporal structure. This analytical approach allows us to disentangle the relative contributions of content and timing without the constraints of artificial laboratory environments.

We propose a Dual Prediction Violation (DPV) framework: humor arises from the interplay between two prediction systems (see Fig. 1). *Semantic prediction violations* occur when content deviates from contextually expected continuations. *Temporal prediction violations* occur when pacing and pause patterns deviate from established rhythmic expectations, inspired by dynamic attending (Large & Jones, 1999). Critically, these systems interact hierarchically: temporal structure acts as a cognitive gatekeeper determining *when* semantic information is processed relative to attentional peaks. Extended pauses before high-surprise content create temporal windows for priming semantics and accumulating prediction errors; when the punchline arrives within this optimal window, pre-activated cognitive resources enable rapid resolution, thereby maximizing the efficiency gain that drives pleasure.

To test this framework, we analyzed the 828 Chinese stand-up comedies, quantifying temporal dynamics through pause patterns and speech rate, alongside semantic incongruity via embedding-based distance measures. Our analyses reveal three key findings. First, temporal features substantially outperform semantic features in predicting audience appreciation, suggesting that timing dynamics provide the primary mechanism for humor success. Second, semantic violations show their strongest effects at peak moments—defined by maximum

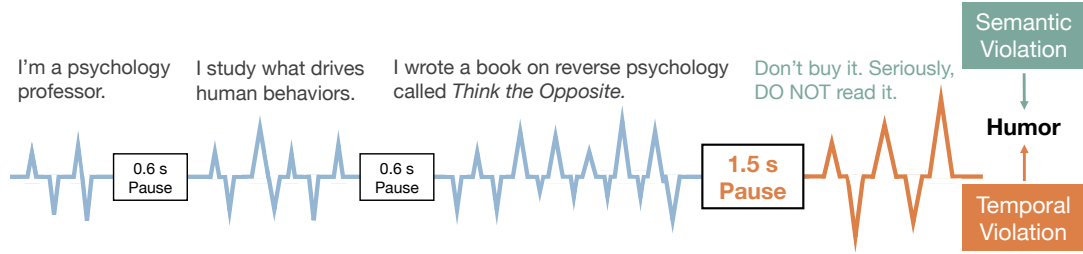


Figure 1: **Illustration of the proposed Dual Prediction Violation (DPV) mechanism.** Setup statements establish temporal and semantic expectations through regular patterns (0.6s pauses, small semantic distances between consecutive sentences). The punchline simultaneously violates both dimensions: an extended pause (1.5s) disrupts established rhythm while semantically distant content deviates sharply from contextual predictions. This strategic coupling of temporal and semantic violations creates conditions for efficient prediction error resolution, generating humor.

deviation from contextual expectations rather than average incongruity—consistent with predictive processing’s emphasis on salient prediction errors. Third, sentence-level analysis of high-performing sets reveals systematic coordination between timing and content: pauses are extended more strongly before high-surprise content compared to low-performing sets. This differential coupling pattern is consistent with the temporal scaffolding of semantic processing, repositioning timing from a delivery detail to a potential cognitive mechanism in humor appreciation.

Related Work

Humor Theories and Computational Approaches

Classical incongruity theory posits that humor arises from detecting and resolving semantic violations (Suls, 1972, 1983). Jokes succeed when audiences first encounter content violating contextual expectations, then discover alternative interpretations rendering violations meaningful. Attardo and Raskin (1991) systematized this view in their General Theory of Verbal Humor (GTVH), proposing six knowledge resources—including script opposition and logical mechanisms—that characterize joke structure. These frameworks dominated humor research for decades, positioning semantic content as the primary determinant of funniness.

Recent theoretical advances have integrated the principles of predictive processing. Van de Cruys et al. (2017) argue that pleasure emerges not from mere incongruity resolution, but from *efficient* resolution—when the brain resolves prediction errors faster than anticipated, it generates amusement as a metacognitive reward signal. This framework yields a critical prediction: any factor that modulates resolution speed should also modulate humor intensity. Neuroimaging supports dissociable neural correlates for joke comprehension *vs.* perceived funniness (Moran et al., 2004; Vrticka et al., 2013), confirming that understanding and pleasure reflect distinct processes.

Computational approaches have increasingly operationalized these theoretical constructs. Early work identified semantic violations through wordplay detection, script opposition, or lexical substitution (Cattle & Ma, 2018; Veale, 2004). Contemporary methods leverage distributional semantics to quantify semantic surprise: embedding-based approaches measure punchline deviation from setup context in semantic space,

operationalizing incongruity as vector distances in BERT or GPT (Annamoradnejad & Zoghi, 2024; Kao et al., 2016; Mihalcea et al., 2010). These computational measures effectively predict humor ratings in written text, demonstrating that quantifiable semantic features capture core aspects of funniness.

Yet these approaches, theoretical and computational alike, share a critical limitation: they treat humor primarily as a *content* problem. Timing appears, if at all, as a delivery variable amplifying inherently funny semantic structures. This reflects methodological constraints: whether analyzing written jokes or laboratory presentations with fixed timing, existing work eliminates the temporal variability characterizing naturalistic performance (Moran et al., 2004). More fundamentally, if humor pleasure derives from resolution efficiency as Van de Cruys et al. (2017) propose, *temporal dynamics* of information delivery should centrally determine when and how efficiently audiences resolve incongruity. Yet quantitative analyses examining timing-semantic interactions remain sparse.

Temporal Dynamics in Cognition

Research beyond humor demonstrates that temporal structure fundamentally shapes cognitive processing. Dynamic attending theory shows that rhythmic regularities generate anticipatory neural states, modulating when information is processed most efficiently (Large & Jones, 1999). Audiences entrain to temporal patterns, creating attentional windows at predicted moments (Jones & Boltz, 1989). Violations of these temporal expectations (unexpected pauses or accelerations) force attentional reorientation, potentially priming cognitive systems for salient information (Barnes & Jones, 2000).

Suspense research provides converging evidence that temporal scaffolding determines semantic processing (Lehne & Koelsch, 2015). Lehne et al. (2015) demonstrated that extended anticipatory periods before resolution engage predictive mechanisms, with neural activity reflecting accumulated prediction error. Critically, anticipation *duration* modulates both prediction error intensity and subsequent reward signals upon resolution (Fiorillo et al., 2003). In speech processing, prosodic boundaries marked by pauses signal upcoming information structure, enabling listeners to prepare appropriate cognitive resources (Bögels et al., 2011).

Despite these insights, temporal dynamics remain underex-

plored in humor research (Attardo, 2001). While some studies acknowledge comic timing matters (Norrick, 2001), quantitative analyses examining how pause patterns, speech rate, and rhythmic structure interact with semantic surprise are scarce.

Methods

Dataset

We compiled an initial corpus of 1102 professional Chinese stand-up comedy performances totaling 107 hours, broadcast between 2017 and 2025. The dataset comprises five seasons of *Rock & Roast* (脱口秀大会), two seasons of *Stand-up Comedy and Friends* (脱口秀和TA的朋友们), and two seasons of *Comedy King* (喜剧之王). We processed each performance using timestamped Automatic Speech Recognition (ASR) to extract temporal and semantic features.

Audience appreciation was measured through live voting during each performance. Our primary analyses focus on 828 performances (86 hours) with complete voting data. To account for heterogeneous voting systems across shows, we normalized votes by calculating the vote rate (votes received divided by total possible votes). This normalization enables comparison across shows while preserving relative performance quality within each competition. The mean performance duration was 373.06 seconds ($SD = 83.49$ seconds).

Feature Extractions

Temporal Features From timestamped ASR transcripts, we extracted temporal dynamics capturing baseline pacing and rhythmic variability. *Average pause duration* measures mean silence length between speech segments (seconds), quantifying baseline temporal spacing. *Pause variability* captures the standard deviation of pause durations (seconds), operationalizing temporal unpredictability in rhythmic structure. *Speech rate* is computed as characters (syllables) per second during speech, reflecting delivery speed; in Mandarin Chinese, each written character corresponds to a single spoken syllable. These features capture distinct aspects of temporal structure relevant to audience prediction dynamics: average pause duration and speech rate establish baseline temporal expectations, while pause variability operationalizes deviations from established rhythmic patterns.

Semantic Features We quantified semantic incongruity through embedding-based distances between consecutive sentences (Ma et al., 2025). Transcripts were automatically segmented into sentences by the ASR system. For each consecutive sentence pair, we computed embeddings using OpenAI’s text-embedding-3-small model and calculated semantic distance as $1 - \text{cosine similarity}$ from L2-normalized embeddings. Higher distances indicate greater deviation from preceding context, directly operationalizing sentence-level semantic incongruity.

For each performance, we extracted distributional statistics capturing different aspects of semantic violation. *Average distance* measures overall incongruity across the performance.

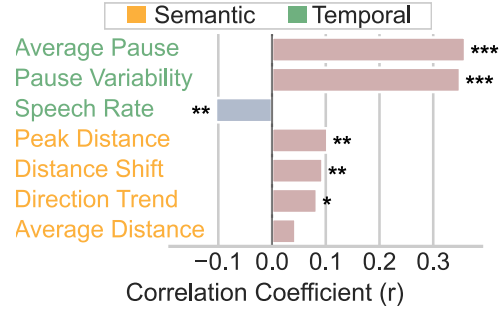


Figure 2: **Correlation between features and audience appreciation.** Partial correlation coefficients (controlling for performance duration) between temporal features (green labels) and semantic features (yellow labels) with audience vote rates. Temporal dynamics substantially outperform semantic features in predicting appreciation, with average pause duration and pause variability showing the strongest effects. Statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Peak distance captures maximum incongruity, identifying the sentence deviating most strongly from its predecessor. *Distance trend* quantifies whether semantic distances systematically change over time, computed as the slope of distance regressed on sentence position. *Distance shift* measures whether distances are higher in the second versus the first half, capturing structural changes in semantic patterning.

Results

Temporal Dynamics and Audience Appreciation

We examined relationships between features and audience appreciation using partial correlations controlling for performance duration, as duration naturally correlates with many structural aspects. Group comparisons contrasted high-performing (top 20%, $n = 171$) and low-performing (bottom 20%, $n = 169$) sets using independent samples t-tests with Cohen’s d effect sizes. The slight group size difference reflects tied vote rates at cutoff boundaries.

Temporal dynamics emerged as the strongest predictor of audience voting behavior (see Fig. 2), substantially outweighing all other features. Average pause duration demonstrated a strong positive correlation with votes (partial $r = 0.36$, $p < 0.001$), representing one of the largest effects observed. Pause variability (standard deviation of pause durations)—capturing temporal unpredictability—also predicted engagement robustly (partial $r = 0.35$, $p < 0.001$). These findings align with dynamic attending theory (Large & Jones, 1999), suggesting performers strategically manipulate temporal expectations to heighten anticipation and amplify subsequent surprise.

Speech rate showed a negative correlation with votes (partial $r = -0.10$, $p < 0.01$), indicating slower delivery facilitates engagement. This pattern suggests rushed pacing may undermine comedic impact, potentially by providing insufficient time for prediction generation. Combined with pause effects, this indicates temporal control involves both strategic silence (extended pauses) and modulated speech (slower rate), jointly expanding the temporal window for predictive processing.

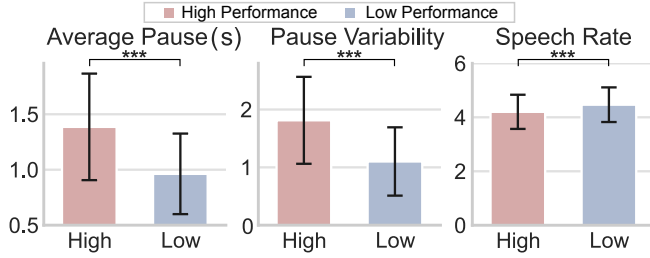


Figure 3: **Group comparisons of temporal features.** Mean values for high-performing (top 20%, red) and low-performing (bottom 20%, blue) sets across three temporal dimensions: average pause duration (seconds), pause variability (seconds), and speech rate (characters per second). High-performing comedians exhibit substantially longer and more variable pauses, alongside slower speech rates. Error bars represent standard deviation. Statistical significance: *** $p < 0.001$.

Group comparisons provided converging evidence with large effect sizes (see Fig. 3). High-performing sets exhibited substantially longer average pauses ($M_{high} = 1.39$ s, $SD = 0.48$ s vs. $M_{low} = 0.96$ s, $SD = 0.36$ s; $t(322) = 8.96$, $p < 0.001$, Cohen’s $d = 0.99$). Pause variability showed similar discrimination ($M_{high} = 1.81$ s, $SD = 0.75$ vs. $M_{low} = 1.10$ s, $SD = 0.59$; $t(322) = 9.48$, $p < 0.001$, $d = 1.05$). Speech rate differences were smaller but consistent: successful performers spoke more slowly ($M_{high} = 4.20$, $SD = 0.63$ vs. $M_{low} = 4.47$, $SD = 0.64$; $t(322) = -3.72$, $p < 0.001$, $d = -0.41$).

Semantic Violations and Audience Appreciation

Semantic distance measures supported incongruity-based mechanisms, though effect sizes were smaller than those for temporal features (see Fig. 2). Peak distance—representing maximum semantic deviation in a performance—correlated with audience appreciation (partial $r = 0.10$, $p < 0.01$). Distance shift—measuring whether distances increased from first to second half—predicted appreciation (partial $r = 0.09$, $p < 0.01$), suggesting successful performers escalate unpredictability over time rather than maintaining constant incongruity. Direction trend—quantifying systematic changes in semantic distance throughout performance—showed a positive correlation (partial $r = 0.08$, $p < 0.05$).

Group comparisons mirrored these patterns (see Fig. 4). Peak distance was higher in successful performances ($M_{high} = 0.83$, $SD = 0.04$ vs. $M_{low} = 0.81$, $SD = 0.05$; $t(322) = 4.50$, $p < 0.001$, $d = 0.50$). Average distance showed similar but weaker discrimination ($M_{high} = 0.61$, $SD = 0.03$ vs. $M_{low} = 0.60$, $SD = 0.03$; $t(322) = 2.37$, $p = 0.018$, $d = 0.26$).

Strategic Coupling: Timing and Content

Beyond main effects, we tested whether comedians strategically coordinate timing with content—specifically, whether they amplify pauses before semantically surprising sentences. This *strategic coupling hypothesis* proposes that temporal control scaffolds semantic violations: longer pauses preceding surprising content allow greater prediction generation, amplifying subsequent prediction errors.

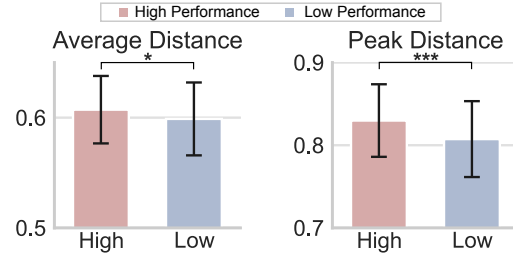


Figure 4: **Group comparisons of semantic features.** Mean values for high-performing (top 20%, red) and low-performing (bottom 20%, blue) sets across semantic dimensions: average distance (left) and peak distance (right). High-performing comedians show greater semantic incongruity, with peak distance demonstrating stronger discrimination than average distance. Error bars represent standard deviation. Statistical significance: * $p < 0.05$, *** $p < 0.001$.

Within-performance timing-content coordination We analyzed sentence-level relationships between semantic distance and pause duration within individual performances. To isolate strategic timing-content coupling, we defined surprise levels relative to each specific performance. We identified consecutive sentence pairs and classified them based on semantic distance: the top 20% as high-surprise pairs and the bottom 20% as low-surprise pairs. This within-performance approach controls for individual stylistic differences; because comedians differ in baseline semantic complexity, global classification would conflate content style with temporal strategy. By comparing high-surprise and low-surprise sentence pairs within a comedian’s own set, we tested whether performers systematically modulate timing when introducing semantic violations relative to their established patterns.

Across all performances, pauses before high-surprise pairs were significantly longer (see Fig. 5a; $M_{high} = 1.37$ s) than those preceding low-surprise pairs ($M_{low} = 1.01$ s; $t(1638) = 6.15$, $p < 0.001$). This 0.36-second difference—a 35.6% increase—demonstrates systematic timing-content coupling: pauses extend systematically before semantically surprising content rather than distributing uniformly across transitions.

Performance levels moderates timing-content coupling

To determine whether timing-content coordination varies with performance quality, we tested whether successful comedians exhibit stronger coupling than less successful performers.

A two-way ANOVA revealed significant main effects of both performance level ($F(1, 3274) = 49.05$, $p < 0.001$) and surprise level ($F(1, 3274) = 38.35$, $p < 0.001$), plus a significant interaction ($F(1, 3274) = 4.23$, $p = 0.040$). High-performing comedians demonstrated steeper timing modulation (see Fig. 5b): mean pauses increased from 1.14s (low-surprise) to 1.61s (high-surprise), a 41.2% increase. Low-performing comedians showed weaker modulation, with pauses increasing from 0.84s to 1.07s, only a 27.4% increase. This interaction demonstrates that strategic timing-content coupling is not merely a stylistic choice but may be a marker of comedic expertise. Successful comedians show **greater differentiation** in pause patterns across semantic surprise levels.

Discussions

We proposed a Dual Prediction Violation (DPV) framework that examines how temporal and semantic mechanisms jointly drive humor appreciation in naturalistic comedy. Three findings emerged: temporal dynamics substantially outperform semantic features in predicting audience engagement; semantic violations have the strongest effects at peak moments; and successful comedians systematically extend pauses before semantically surprising content. Together, these patterns suggest that temporal structure plays a more central role in humor processing than current theories acknowledge.

Temporal Dynamics as a Dominant Predictor

The DPV framework hypothesizes that humor arises from coordinated violations of temporal and semantic predictions. Our first finding, that temporal features substantially outperform semantic features, provides initial support for hierarchical organization where temporal dynamics dominate engagement. The average pause duration and the pause variability both showed large effect sizes in group comparisons, while the semantic distance measures showed medium effects. This significant difference in predictive power challenges the predominant focus on semantic content in humor theory (Attardo & Raskin, 1991; Suls, 1972). These findings do not imply that semantic content is unimportant—peak semantic distance showed consistent positive correlations with appreciation. Rather, they suggest that among measured factors, temporal structure accounts for substantially more variance in audience responses than semantic incongruity. This asymmetry raises theoretical questions about the relative contributions of “what” versus “when” in humor processing.

Several mechanisms may explain the dominance of timing. First, pauses create temporal windows for prediction generation. Extended silence allows audiences to elaborate contextual expectations before encountering violations (Lehne & Koelsch, 2015). The negative correlation with speech rate provides con-

verging evidence: slower delivery extends processing time, potentially facilitating prediction elaboration. Second, pause variability prevents audiences from settling into stable temporal expectations, maintaining cognitive flexibility through dynamic attention (Large & Jones, 1999). Third, temporal unpredictability may independently generate prediction errors that prime audiences to detect novelty, heightening sensitivity to subsequent content surprises.

However, our correlational design cannot determine whether the temporal structure directly causes appreciation or reflects other unmeasured factors. High-performing comedians may differ in confidence, stage presence, or audience rapport—qualities that manifest behaviorally as longer pauses and slower delivery.

Semantic Violations as Peak Events

The DPV framework predicts that semantic violations should matter, and our findings refine this prediction. Among semantic features, peak distance (maximum semantic deviation) emerged as the strongest predictor of appreciation. This means what matters most is not overall incongruity levels but the presence of particularly salient violations. The average semantic distance showed weaker effects, despite capturing more information about performance-wide incongruity, indicating that audiences respond more to discrete high-surprise moments than to chronic moderate incongruity. This finding aligns with predictive processing accounts that emphasize salient prediction errors (Clark, 2018). The brain prioritizes resolving large, discrete violations over processing continuous low-level uncertainty. In comedy, this suggests that a few well-placed, surprising moments may be more effective than maintaining a constant, moderate level of incongruity throughout.

The positive correlations of distance shift and direction trend with appreciation provide additional evidence for content structuring strategies. Successful performers appear to escalate semantic violations over time rather than maintain constant incongruity levels. This may create a trajectory of increasing prediction error that sustains engagement, though future research should examine whether this reflects deliberate strategy or emergent properties of joke sequencing.

Strategic Coupling Distinguishes Expertise

The most direct evidence for coordination between temporal and semantic systems came from sentence-level analyses. Across all performances, pauses were 35.6% longer before high-surprise content compared to low-surprise transitions. Critically, high-performing comedians showed stronger timing-content coupling (41.2% pause increase) than low-performing comedians (27.4% increase).

This pattern demonstrates timing-content coordination may be a marker of comedic expertise, but the causal mechanism remains ambiguous. Three interpretations are plausible. First, expert comedians may deliberately extend pauses before surprising content to scaffold audience processing—creating temporal windows for prediction elaboration before violations ar-

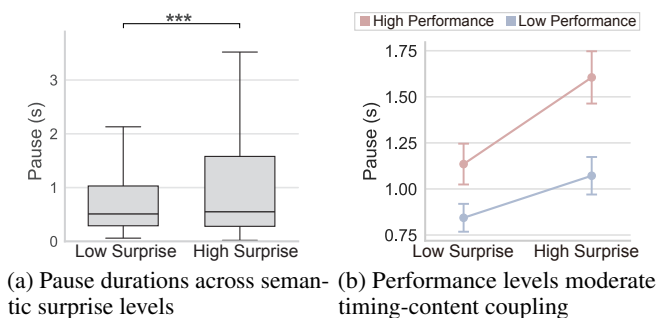


Figure 5: Strategic coupling between timing and semantic surprise. Pauses systematically extend before high-surprise content, with successful comedians showing stronger modulation. (a) Across all performances, pauses before high-surprise pairs (top 20% semantic distance) are 35.6% longer than before low-surprise pairs (bottom 20% semantic distance). (b) High-performing comedians (top 20%, red) exhibit steeper timing modulation (41.2% increase) than low-performing comedians (bottom 20%, blue; 27.4% increase). Error bars represent the standard error of the mean. Statistical significance: *** $p < 0.001$.

rive. Second, semantically complex or surprising content may require more cognitive effort to formulate and deliver, naturally eliciting longer pauses from performers. Third, expertise may manifest as sensitivity to both dimensions: skilled comedians create more surprising content and have better temporal control, with coordination emerging from dual competencies rather than causal scaffolding.

Our data cannot distinguish among these possibilities. However, the interaction effect—where high-performing comedians show *stronger* coupling than low-performing comedians—suggests coordination is not merely an artifact of content complexity. If surprising content simply required a longer delivery time, we would expect similar pause extensions across all performers. That experts show greater coordination suggests either deliberate strategic timing or that expertise involves heightened sensitivity to how content complexity affects temporal demands. Future experimental work manipulating timing while holding content constant could distinguish whether temporal structure causally affects processing or reflects performer-side demands.

The coordination finding is consistent with the temporal scaffolding of semantic processing, but alternative explanations remain viable. Longer pauses before surprising content could reflect comedians’ own processing needs, audience feedback mechanisms, or emergent properties of how complex content is constructed during performance. Establishing causal direction and mechanism requires controlled experiments beyond naturalistic observation.

Limitations and Future Directions

Several limitations warrant consideration. First, our use of professional performances in competitive settings provides ecological validity but introduces two interpretive constraints. Audience voting provides a behavioral proxy for comedic expertise, though it reflects audience preference and may not fully capture all dimensions of performer skill. Additionally, editorial decisions may modify temporal structure: while editors likely preserve narrative coherence and pacing, we cannot rule out that editing systematically alters pause patterns. Future work can examine live performances to confirm whether these patterns persist without editorial mediation.

Second, our dataset focuses on Chinese stand-up comedy, raising generalizability questions. While temporal processing mechanisms are thought domain-general and culturally universal (Large & Jones, 1999), specific pause durations and timing-content relationships may vary across cultures and languages. Chinese is a tonal, character-based language, potentially affecting how timing cues semantic boundaries. Cross-linguistic replication would clarify which findings reflect universal cognitive mechanisms versus culturally specific norms.

Third, our semantic distance measures capture local sentence-to-sentence transitions but not higher-level narrative structure, callbacks, or thematic coherence. Comedy often involves long-range dependencies where jokes build on earlier setups or subvert expectations established minutes earlier.

Future computational models should integrate hierarchical semantic representations to capture how local and global incongruity jointly shape experience, and may employ token surprisal to capture online prediction error at a finer grain.

Fourth, our analysis focused on features extractable from ASR transcripts, excluding prosodic features beyond timing (pitch, loudness, voice quality), gestural communication, audience interaction, and stage presence. The large effect sizes for pause-based features suggest timing alone captures substantial variance, but complete accounts require multimodal analysis integrating linguistic, paralinguistic, and visual cues.

Finally, the correlational design limits causal inference. We cannot determine whether temporal structure causally affects audience processing or reflects performer-side demands. Experimental studies manipulating pause durations while holding content constant are needed to establish causality.

Several directions could extend our framework. First, fine-grained audience response data (*e.g.*, continuous laughter ratings, physiological measures) would reveal how timing-content coupling affects moment-by-moment engagement. Second, testing the DPV framework across other humor formats (improvisation, sketch comedy) and related domains (music, rhetoric, storytelling) would clarify whether strategic coupling reflects humor-specific or domain-general principles. Third, examining individual differences would connect humor research with expertise and creativity: Do successful comedians consciously control timing-content coupling? Do audiences vary in sensitivity to temporal versus semantic violations? Addressing these questions would illuminate both humor cognition and predictive processing more generally.

Conclusion

By computationally analyzing 828 Chinese professional stand-ups, we establish that naturalistic humor success depends on coordinated violations of temporal and semantic predictions. Temporal control dominates overall effects, but strategic coupling—amplifying pauses before semantically surprising content—distinguishes professional expertise. These findings integrate humor theory with predictive processing frameworks, demonstrating that timing scaffolds semantic incongruity through cognitive gating mechanisms. Our DPV framework advances understanding of how experts manipulate audience expectations in naturalistic settings, with implications for humor cognition, computational humor generation, and creativity research (Ma et al., 2026; Peng et al., 2025).

Acknowledgement We thank Fengyuan Yang for creating elegant visuals. The work is supported by the National Natural Science Foundation of China (32595491, 62376009), the PKU-BingJi Joint Laboratory for Artificial Intelligence, the Wuhan Major Scientific and Technological Special Program (2025060902020304), the Hubei Embodied Intelligence Foundation Model Research and Development Program, and the National Comprehensive Experimental Base for Governance of Intelligent Society, Wuhan East Lake High-Tech Development Zone.

References

- Annamoradnejad, I., & Zoghi, G. (2024). Colbert: Using bert sentence embedding in parallel neural networks for computational humor. *Expert Systems with Applications*, 249, 123685 (cit. on p. 2).
- Attardo, S. (2001). *Humorous texts: A semantic and pragmatic analysis* (Vol. 6). Walter de Gruyter. (Cit. on p. 3).
- Attardo, S., & Raskin, V. (1991). Script theory revis (it) ed: Joke similarity and joke representation model. *Humor: International Journal of Humor*, 4(3), 293–347 (cit. on pp. 2, 5).
- Barnes, R., & Jones, M. R. (2000). Expectancy, attention, and time. *Cognitive Psychology*, 41(3), 254–311 (cit. on p. 2).
- Bögels, S., Schriefers, H., Vonk, W., & Chwilla, D. J. (2011). Prosodic breaks in sentence processing investigated by event-related potentials. *Language and Linguistics Compass*, 5(7), 424–440 (cit. on p. 2).
- Cattle, A., & Ma, X. (2018). Recognizing humour using word associations and humour anchor extraction. *International Conference on Computational Linguistics* (cit. on p. 2).
- Clark, A. (2018). A nice surprise? predictive processing and the active pursuit of novelty. *Phenomenology and the Cognitive Sciences*, 17(3), 521–534 (cit. on p. 5).
- Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 299(5614), 1898–1902 (cit. on p. 2).
- Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological Review*, 96(3), 459 (cit. on p. 2).
- Kao, J. T., Levy, R., & Goodman, N. D. (2016). A computational model of linguistic humor in puns. *Cognitive Science*, 40(5), 1270–1285 (cit. on p. 2).
- Large, E. W., & Jones, M. R. (1999). The dynamics of attending: How people track time-varying events. *Psychological Review*, 106(1), 119 (cit. on pp. 1–3, 5, 6).
- Lehne, M., Engel, P., Rohrmeier, M., Menninghaus, W., Jacobs, A. M., & Koelsch, S. (2015). Reading a suspenseful literary text activates brain areas related to social cognition and predictive inference. *PloS One*, 10(5), e0124550 (cit. on p. 2).
- Lehne, M., & Koelsch, S. (2015). Toward a general psychological model of tension and suspense. *Frontiers in Psychology*, 6, 79 (cit. on pp. 2, 5).
- Ma, Y., Peng, Y., Yang, F., Zha, S., Zhang, C., Jia, Z., Zheng, Z., & Zhu, Y. (2026). Narrativeloom: Enhancing creative storytelling through multi-persona collaborative improvisation. *ACM Conference on Human Factors in Computing Systems (CHI)* (cit. on p. 6).
- Ma, Y., Peng, Y., & Zhu, Y. (2025). Word embeddings track social group changes across 70 years in china. *Annual Meeting of the Cognitive Science Society (CogSci)* (cit. on p. 3).
- Mihalcea, R., Strapparava, C., & Pulman, S. (2010). Computational models for incongruity detection in humour. *International Conference on Intelligent Text Processing and Computational Linguistics* (cit. on p. 2).
- Moran, J. M., Wig, G. S., Adams Jr, R. B., Janata, P., & Kelley, W. M. (2004). Neural correlates of humor detection and appreciation. *Neuroimage*, 21(3), 1055–1060 (cit. on pp. 1, 2).
- Norrick, N. R. (2001). On the conversational performance of narrative jokes: Toward an account of timing. *Humor: International Journal of Humor Research*, 14(3), 255–274 (cit. on p. 3).
- Peng, Y., Ma, Y., Wang, M., Wang, Y., Wang, Y., Zhang, C., Zhu, Y., & Zheng, Z. (2025). Probing and inducing combinational creativity in vision-language models. *ACM Conference on Human Factors in Computing Systems (CHI)* (cit. on p. 6).
- Suls, J. (1972). A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. In *The psychology of humor: Theoretical perspectives and empirical issues* (pp. 81–100, Vol. 1). Academic Press. (Cit. on pp. 1, 2, 5).
- Suls, J. (1983). Cognitive processes in humor appreciation. In *Handbook of humor research: Volume 1: Basic issues* (pp. 39–57). Springer. (Cit. on pp. 1, 2).
- Van de Cruys, S., Metzinger, T. K., & Wiese, W. (2017). Affective value in the predictive mind. In *Philosophy and predictive processing* (pp. 1–21). MIND Group; Frankfurt am Main. (Cit. on pp. 1, 2).
- Veale, T. (2004). Incongruity in humor: Root cause or epiphenomenon? *Humor: International Journal of Humor Research*, 17(4), 419–428 (cit. on p. 2).
- Vrticka, P., Black, J. M., & Reiss, A. L. (2013). The neural basis of humour processing. *Nature Reviews Neuroscience*, 14(12), 860–868 (cit. on p. 2).