Similar Predictions, Different Processes: A Multi-Level Comparison of Human and Multimodal LLM Language Prediction

Shuqi Wang[1], Zhenguang G. Cai[1,2]

[1]Department of Linguistics and Modern Languages, The Chinese University of Hong Kong
[2]Brain and Mind Institute, The Chinese University of Hong Kong
shuqiwang@link.cuhk.edu.hk, zhenguangcai@cuhk.edu.hk

Abstract

Humans and large language models (LLMs) both generate predictions during language processing, but whether they integrate structural and prosodic cues similarly during visually grounded speech remains underexplored. Multimodal LLMs that jointly process speech and vision now make it possible to compare not only what humans and models predict, but also when predictions emerge. We compared Mandarin speakers and Qwen2.5-Omni-7B on Mandarin dative constructions in a visual world paradigm (VWP), asking how these cues guide predictions about upcoming referents. Experiment 1 used a cloze-in-VWP task to assess offline prediction outputs; Experiment 2 examined online processing via human eye-tracking and a model audio-to-image cross-modal attention measure. In Experiment 1, humans and the model were both sensitive to structure and prosody, consistent with partial output-level alignment, but the model showed a larger structural effect and a condition-specific atypical prosody pattern. In Experiment 2, the time courses diverged: humans showed structural effects before the contrastive connective, whereas the model's sensitivity emerged later, after connective onset. These findings indicate that output-level and process-level alignment can dissociate in this paradigm. This study contributes a methodology for multi-level human–model comparison and provides empirical constraints on claims about the cognitive plausibility of multimodal LLMs.

1 Introduction

Humans process language incrementally and predictively, generating expectations about upcoming input from contextual, syntactic, prosodic, and visual information (Kuperberg and Jaeger, 2016; Pickering and Gambi, 2018; Altmann and Kamide, 1999). Large language models (LLMs) trained on next-token prediction also generate linguistic predictions and show sensitivity to formal regularities such as syntax and morphology (Mahowald et al., 2024; Linzen and Baroni, 2021), with internal representations that partially predict human neural responses (Schrimpf et al., 2021). These parallels invite the question of whether human and model predictions align—and if so, at what level of analysis (Caucheteux and King, 2022; Goldstein et al., 2022).

Following Marr (1982), we distinguish output-level alignment in *what* is predicted from process-sensitive alignment in *when* and *how* predictions unfold. Prior human–LLM comparisons span both levels, from behavioral metrics such as accuracy, perplexity, and surprisal (Futrell et al., 2019; Wilcox et al., 2020) to internal representations, attention patterns, and neural alignment (Sood et al., 2020; Eberle et al., 2022; Gao et al., 2025; Goldstein et al., 2025). However, less work has examined spoken prediction under shared visual grounding, where listeners integrate speech, prosody, and visual information about potential referents (Tanenhaus et al., 1995; Ito and Speer, 2008; Weber et al., 2006). Multimodal LLMs that process speech and vision (Xu et al., 2025) now make it possible to test this comparison in a shared multimodal setting, examining not only *what* is predicted but also *when* and *how* predictions unfold in real time.

The present study addresses two research questions. First, how do humans and a multimodal LLM integrate syntactic structure and prosodic stress when predicting referents in visually grounded spoken language? Second, when humans and the model show similar output-level prediction patterns, are these patterns accompanied by similar temporal cue-deployment profiles at the process level?

We compare Mandarin speakers and Qwen2.5-Omni-7B on Mandarin dative constructions in the visual world paradigm (VWP), crossing double-object (DO) versus prepositional-object (PO) word

order with the presence and location of contrastive stress. Experiment 1 uses a cloze-in-VWP task to measure offline prediction outputs—which referent each system selects after the sentence preamble. Experiment 2 uses the standard VWP to compare human eye movements with the model's audio-to-image cross-modal attention over the same visual regions. This dual-paradigm design tests whether similar prediction outputs in this Mandarin task are accompanied by similar temporal cue-deployment profiles.

2 Related Work

2.1 Structural and Prosodic Cues in Referential Prediction

Prior work motivates a set of operational predictions about how syntactic structure and prosodic prominence are weighted during referential prediction. In Mandarin, sentence-final position is a default site of informational focus (Xu, 2004), and double-object (DO) versus prepositional-object (PO) datives place different arguments in this position, shaping contrast expectations (Ziegler and Snedeker, 2019). Prosodic prominence provides another cue: contrastive accent can guide anticipatory reference resolution in spoken comprehension (Dahan et al., 2002; Ito and Speer, 2008; Weber et al., 2006).

These findings motivate two dimensions of prediction. First, one cue may dominate: a structure-dominant prediction treats DO versus PO structure as the primary determinant of the anticipated alternative, whereas a prosody-dominant prediction treats overt stress as redirecting prediction toward the stressed constituent's alternative. Between these poles, the *Enduring Focus Hypothesis* predicts that default positional focus continues to shape sentence-level focus even when overt contrastive accents are present (Harris and Carlson, 2018; Harris, 2023). Second, if both cues contribute, they may combine in different ways. A cue-convergence pattern predicts stronger effects when stress and default focus select the same contrast set, consistent with evidence that pitch accent and default focus position can have additive effects in focus interpretation (Clifton and Frazier, 2016). A cue-diagnosticity pattern predicts stronger stress effects when stress marks a constituent lacking default structural focus, consistent with probabilistic accounts in which accent value depends on expected prominence (Calhoun, 2010; Yan and Calhoun, 2020; Li et al., 2018). We use these labels as task-level predictions about cue weighting in the present design.

2.2 Measuring Prediction Outputs and Online Prediction Processes

Different experimental measures capture different aspects of prediction. Cloze tasks measure prediction outputs by eliciting continuations to sentence fragments (Taylor, 1953); their similarity to autoregressive language modeling makes them useful for human–model comparison (Eisape et al., 2020; Goldstein et al., 2022). The visual world paradigm (VWP) provides a complementary process-sensitive measure: anticipatory eye movements to potential referents reveal how spoken input guides attention over time (Tanenhaus et al., 1995; Huettig et al., 2011). Pairing a cloze-in-VWP task (Corps et al., 2022) with a standard VWP therefore allows us to distinguish *what* is predicted from *when* structural and prosodic cues influence attention.

2.3 Human–LLM Alignment and Attention-Based Process Measures

Evaluating human–LLM alignment requires specifying the level at which alignment is claimed. Behavioral similarity can indicate comparable outputs, but it does not by itself establish similarity in representations or processing dynamics (Guest and Martin, 2023; Lin, 2025). Prior work has found both convergence and divergence across levels: LLMs show sensitivity to formal linguistic regularities (Linzen and Baroni, 2021; Mahowald et al., 2024), and some model representations predict human neural or behavioral responses (Schrimpf et al., 2021; Caucheteux and King, 2022; Goldstein et al., 2022). At the same time, process-level comparisons remain mixed, especially for attention-based measures.

We draw on the distinction between formal and functional linguistic competence to separate cue sensitivity from cue deployment (Mahowald et al., 2024). In our setting, syntactic structure and prosodic prominence are linguistic cues to which both humans and LLMs may be sensitive, whereas the process-level question is whether these cues are deployed jointly with visual context as the spoken sentence unfolds. This motivates two predictions for the online comparison. Under a shared-deployment prediction, similar output-level prediction patterns should be accompanied by simi-

lar temporal profiles: human fixations and model audio-to-image cross-modal attention should show structure and prosody effects in comparable windows. Under a deployment-dissociation prediction, similar output-level patterns should be accompanied by different online trajectories, with cue effects appearing at different times or in different parts of the spoken input.

Transformer attention has been compared with human gaze or attention in several settings, with mixed results (Sood et al., 2020; Eberle et al., 2022; Gao et al., 2025; Goldstein et al., 2025). In multimodal LLMs, cross-modal attention can provide a time-resolved signal of how tokens from one modality allocate weight to another, but the measure depends on choices such as modality boundaries, query–key direction, region mapping, and head/layer aggregation. We therefore treat audio-to-image cross-modal attention as an operational measure of allocation to visual regions, not as a direct equivalent of human gaze.

3 Methodology

3.1 General Design and Materials

We designed two matched experiments using matched stimuli and visual–auditory conditions, differing in task demands across experiments: Experiment 1 measures offline prediction outputs (*what* is predicted), while Experiment 2 traces online processing dynamics (*when* predictions emerge). Human and model tasks used matched materials, with input formats and instructions adapted to each system.

Spoken sentences followed a 2 × 3 factorial design (see Figure 1, left panel). The Structure factor manipulated dative construction type: double-object (DO; e.g., *He gave the girl the cake, but not the boy/flower*) versus prepositional-object (PO; e.g., *He gave the cake to the girl, but not the boy/flower*). The Stress factor manipulated prosodic prominence across three levels: Neutral, Theme stress (contrastive accent on the mentioned theme), and Recipient stress (contrastive accent on the mentioned recipient). Each visual display contained four images corresponding to four entity types: mentioned theme, mentioned recipient, alternative theme, and alternative recipient (see Figure 1, right panel).

We created 36 experimental items (from an initial pool of 50; norming details in Appendix A) and distributed them across six lists in a Latin square design. Within each list, each item appeared in exactly one of the six Structure × Stress conditions; across lists, each item appeared once in each condition. Complete stimuli and model notebooks are available in the public repository listed in the Data and Code Availability statement.

3.2 Experiment 1: Offline Prediction Output (Cloze-in-VWP)

Experiment 1 employed a cloze-in-VWP task (Corps et al., 2022; Milburn et al., 2016). Auditory stimuli were truncated after the contrastive connective *but not*. Participants heard these preambles while viewing the four-image display, then selected the image they predicted would be mentioned next and provided a text continuation.

For the human experiment, 118 native Mandarin speakers completed the task online via Gorilla (Anwyl-Irvine et al., 2020). Each participant saw one Latin-square list containing 36 experimental trials and 72 fillers.

For the model experiment, we used Qwen2.5-Omni-7B (Xu et al., 2025). For each item in each list, we sampled 20 independent generations with a fixed set of 20 random seeds. Each run used a fresh conversation context. The 20 generations were not averaged: each was coded as a trial-level response and entered the GLMM as an observation, with run-level variability modeled in the statistical analyses (Appendix C). Model responses were coded using the same response categories as the human continuations. The complete inference scripts are available in the public repository listed in the Data and Code Availability statement. Full prompting, hardware, seed, and decoding details are provided in Appendix B and the public repository.

3.3 Experiment 2: Online Processing (Standard VWP)

Experiment 2 employed the standard VWP. Human participants heard complete sentences while viewing the four-image displays, with no explicit response during the critical prediction window; intermittent comprehension questions were used to ensure attention. For humans (N = 58), eye movements were recorded using an EyeLink 1000 Plus at 1000 Hz. Fixations were assigned to the four image regions and a REST region, computed in 50 ms bins, and aligned to target-word onset (0 ms). Blinks and track losses were excluded from proportion calculations. Full preprocessing details are provided in Appendix B.

Design: 2 (Structure: DO vs. PO) x 3 (Stress: Neutral vs. Theme vs. Recipient)

- **DO structure:**

他	送	给	了	女孩	蛋糕,	而不是	男孩	/	鲜花。
3SG	*give*	*DAT*	*PFV*	*girl*	*cake*	*but.not*	*boy*		*flower*
				mentioned recipient	mentioned theme		alternative recipient		alternative theme

He gave the girl the cake, but not the boy/flower.

- **PO structure:**

他	送	了	蛋糕	给	女孩,	而不是	男孩	/	鲜花。
3SG	*give*	*PFV*	*cake*	*DAT*	*girl*	*but.not*	*boy*		*flower*
			mentioned theme		mentioned recipient		alternative recipient		alternative theme

He gave the cake to the girl, but not the boy/flower.

	DO structure	PO structure
Neutral	He gave the girl the cake, but not the boy/flower.	He gave the cake to the girl, but not the boy/flower.
Theme	He gave the girl the **cake**, but not the flower.	He gave the **cake** to the girl, but not the flower.
Recipient	He gave the **girl** the cake, but not the boy.	He gave the cake to the **girl**, but not the boy.

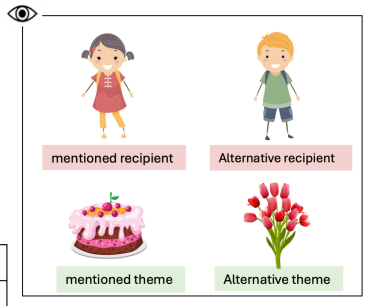

Figure 1: Experimental design and sample stimuli. The left panel illustrates the 2×3 factorial design crossing Structure (DO vs. PO) with Stress (Neutral, Theme, Recipient). Mandarin example sentences are shown with morpheme-by-morpheme English glosses following the Leipzig Glossing Rules, with free English translations below. Bold underline in the condition table indicates contrastive stress placement. The right panel shows an example four-image visual display; entity role labels are shown for illustration only and were not visible to participants.

For the model-side time-course analysis, we extracted audio-to-image cross-modal attention from Qwen2.5-Omni-7B. Because Qwen2.5-Omni processes text, audio, and image tokens in a shared self-attention sequence rather than through a separate cross-attention module, we operationalized cross-modal attention as attention from audio-query tokens to image-key tokens. For the reported analyses, model audio input was truncated before the target noun, matching the human prediction window.

For each item and each of the model's 28 layers, we extracted the audio-query/image-key attention submatrix, averaged across heads within layer, and mapped image tokens to patch-aligned visual regions. Attention was averaged within each region and normalized across the four image regions plus REST, yielding AOI-style regional proportions. Audio token indices were aligned to target-word onset at 40 ms resolution (Qwen2.5-Omni's 25 Hz audio token rate), comparable to human fixation proportions at 50 ms resolution. Layer-specific regional time courses were retained for inferential analysis, with Layer treated as a grouping factor in the statistical models; visualizations report layer-aggregated averages. Implementation details, including token-boundary detection, image grid mapping, and region coordinates, are provided in Appendix B.

4 Analysis and Results

Full model specifications, exclusion details, and complete statistical tables are provided in Appendix C; analysis code is available in the public repository listed in the Data and Code Availability statement.

4.1 Experiment 1: Offline Prediction Output

Analysis procedure. Responses were coded from the text continuation as alternative theme (1) or alternative recipient (0). Responses outside this contrast space, referring to entities not depicted, or referring to already-mentioned referents were excluded (full criteria and counts in Appendix B.1.3). After exclusions, 4,015 of 4,246 human responses (94.6%) and 2,881 of 4,320 model generations (66.7%) entered the binary analysis. Generalized linear mixed-effects models (GLMM) tested effects of Structure and Stress; Structure was sum-coded (DO $= -0.5$, PO $= 0.5$), Stress was treatment-coded with Neutral as the reference, and Participant Type was treatment-coded with humans as the reference in the combined analysis. Random effects were determined by forward model comparison (Appendix C).

Human and model results. Figure 2 (left panel) displays human responses across conditions. A significant positive intercept ($\beta = 2.51$, $z = 14.02$, $p < .001$) indicated an overall theme bias. A significant Structure effect emerged ($\beta = -1.56$, $z = -6.76$, $p < .001$), with greater theme bias in DO than PO constructions, consistent with sentence-final default focus. Both Stress conditions showed significant effects: Theme stress increased theme bias ($\beta = 0.66$, $z = 4.64$, $p < .001$), whereas Recipient stress decreased it ($\beta = -1.33$, $z =$

-11.37, $p < .001$). Significant Structure × Stress interactions also emerged (Theme: $\beta = 0.67$, $p = .020$; Recipient: $\beta = 1.00$, $p < .001$), indicating that stress effects were larger when prosody marked the argument that lacked default structural focus. These results are consistent with the *Enduring Focus Hypothesis* and *cue-diagnosticity pattern* introduced in Section 2.1, with both cues contributing and prosody exerting its strongest influence in non-default positions.

Figure 2 (right panel) displays model results. The model showed the same broad cue sensitivity: significant Structure, Stress, and Structure × Stress effects, including a larger Structure effect ($\beta = -4.69$, $p < .001$). At the response-pattern level, this indicates partial output-level alignment, with a stronger structure-driven pattern in the model.

Human–model comparison. To directly compare human and model behavior, we analyzed the combined dataset with Participant Type as an additional factor. The main effect of Participant Type was not significant ($\beta = -0.13$, $z = -0.34$, $p = .737$), indicating comparable baseline response rates—consistent with partial output-level alignment. However, two interactions revealed systematic differences in cue weighting. First, a Participant Type × Structure interaction ($\beta = -2.78$, $z = -6.49$, $p < .001$) indicated that the model's structure effect was approximately three times the human magnitude. Second, a three-way interaction (Participant Type × Structure × Recipient stress: $\beta = 1.46$, $z = 3.41$, $p < .001$) revealed a condition-specific divergence in prosodic integration. Full model specifications are in Appendix C, Table 4.

Thus, Experiment 1 showed partial output-level alignment, but also stronger structure weighting and condition-specific prosody deviations in the model. The higher model exclusion rate is discussed as a task-grounding difference in Section 5.2.

4.2 Experiment 2: Online Processing Dynamics

Analysis procedure. Time-course data were aligned to target-word onset (time = 0 ms). The prediction window spanned from first clause offset (-1300 ms) through target-word onset. Within this window, we distinguished two sub-windows: a *Pause window* (-1300 to -600 ms), from first clause offset to connective onset, and a *Connective window* (-600 to 0 ms), from connective onset to target-word onset. This division separates early prediction driven by structural and prosodic context alone (the Pause window) from later prediction in which the explicit contrastive connective *but not* also becomes available (the Connective window).

The dependent measure was the log ratio of allocation to the alternative-theme versus the alternative-recipient, computed from fixation proportions for humans and from area-of-interest (AOI)-style audio-to-image cross-modal attention proportions for the model. Positive values indicate theme bias; negative values indicate recipient bias. We applied two complementary analyses: window-based linear mixed-effects models (LME) within the Pause and Connective windows, and time-course generalized additive mixed models (GAMM) over the full prediction window. GAMM significant windows indicate intervals where the pointwise 95% confidence interval for the smooth difference excludes zero; they are interpreted as reliable condition-difference intervals, not exact cognitive onsets.

Human results. Figure 3 (upper panel) shows human fixation time courses. In the Pause window, a significant Structure effect emerged ($\beta = -0.81$, $t = -2.82$, $p = .005$), with no reliable Stress effects. In the Connective window, the Structure effect persisted ($\beta = -1.14$, $t = -3.57$, $p < .001$), and Recipient stress emerged ($\beta = -0.85$, $t = -3.40$, $p = .001$). Thus, Structural effects emerged early—before the connective *but not* was heard—while prosodic effects emerged later.

GAMMs supported the same pattern. The Structure contrast was significant across the full prediction window when collapsed across stress conditions (-1300 to 0 ms); in the Neutral condition, it was significant from -854 to 0 ms. Stress effects emerged later: Recipient stress differed from Neutral from -446 to 0 ms, whereas Theme stress differed from Neutral only from -39 to 0 ms. Within-structure analyses showed the cue-diagnosticity pattern (Section 2.1): in DO structures, Recipient stress differed from Neutral from -643 to 0 ms, whereas Theme stress did not differ reliably; in PO structures, Theme stress differed from Neutral from -249 to 0 ms, whereas Recipient stress did not differ reliably. Prosodic stress thus had its clearest effect when it marked the argument lacking default structural focus.

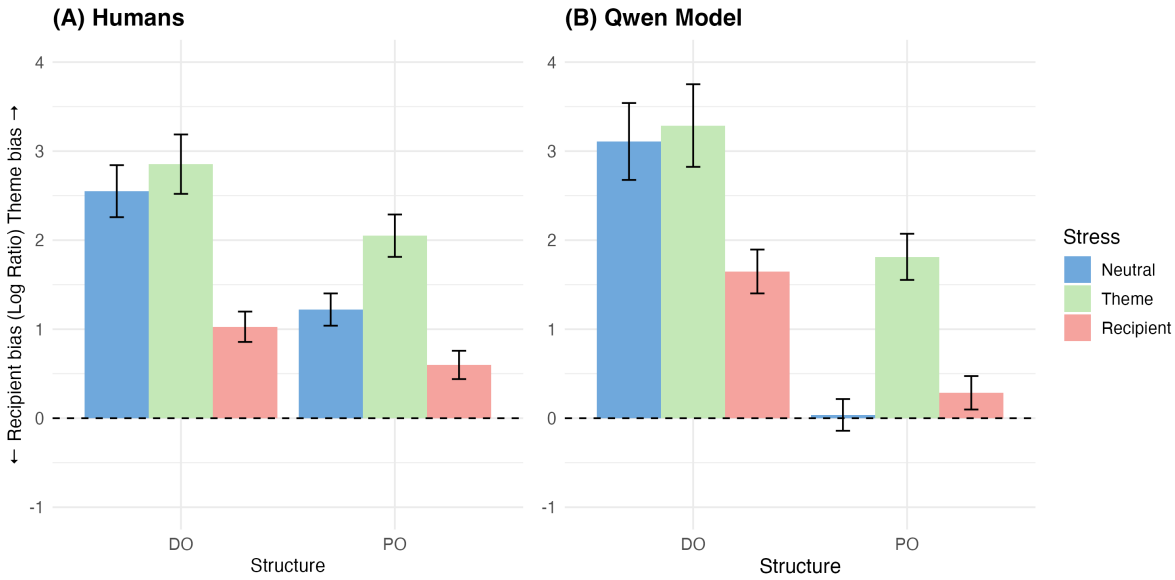

Figure 2: Offline prediction bias in Experiment 1 (cloze-in-VWP). Left panel: human results; right panel: model results. The y-axis displays the log-ratio bias [log(Alt. Theme/Alt. Recipient)]. Positive values indicate a bias toward the theme; negative values indicate a bias toward the recipient. Error bars represent 95% confidence intervals.

Model results. Figure 3 (lower panel) shows the model's audio-to-image attention profile. In the Pause window, no main effects emerged, though a Structure × Theme stress interaction appeared ($p = .025$). In the Connective window, a Structure effect emerged ($p = .005$), along with Theme stress ($p = .029$) and marginal Recipient stress ($p = .096$). Thus, unlike humans, the model did not show a reliable main Structure effect before connective onset.

The GAMMs likewise showed later model attention effects. The Structure contrast collapsed across stress was significant only near target-word onset (-13 to 0 ms). In the Neutral condition, the Structure contrast was significant from -194 to 0 ms. Within structures, however, the model showed the same cue-diagnosticity pattern as humans but later: in DO structures, Recipient stress differed from Neutral from -220 to 0 ms, whereas Theme stress did not differ reliably; in PO structures, Theme stress differed from Neutral from -181 to 0 ms, whereas Recipient stress did not differ reliably. Full results are provided in Appendix C, Tables 7–8 and Table 9.

Human–model comparison. Experiment 2 therefore supports the deployment-dissociation prediction (Section 2.3). Both systems showed cue-diagnosticity at the level of condition structure, but their measured temporal profiles differed: Human structural sensitivity emerged from -854 ms—before the contrastive connective *but not* was heard, consistent with anticipatory effects based on the pre-connective sentence context. Model structural sensitivity emerged only from -194 ms, after connective onset: a $\sim$650 ms difference in the estimated timing of measurable structure effects. Prosodic effects showed a parallel pattern: in humans, Recipient stress effects emerged from -643 ms within DO structures; in the model, from -220 ms.

5 Discussion

5.1 Cue Integration in Human Predictive Language Processing

Our human results clarify how syntactic structure and prosodic prominence jointly shape contrastive prediction in Mandarin. Two patterns are noteworthy. First, both cues contributed: prosodic stress did not eliminate the influence of sentence-final default focus, consistent with the *Enduring Focus Hypothesis* (Harris and Carlson, 2018; Harris, 2023); at the same time, structural information alone did not capture the observed stress effects. Second, when both cues were available, prosodic stress exerted its clearest effect on the constituent lacking default structural focus—the cue-diagnosticity pattern introduced in Section 2.1. In DO structures, Recipient stress differed reliably from Neu-

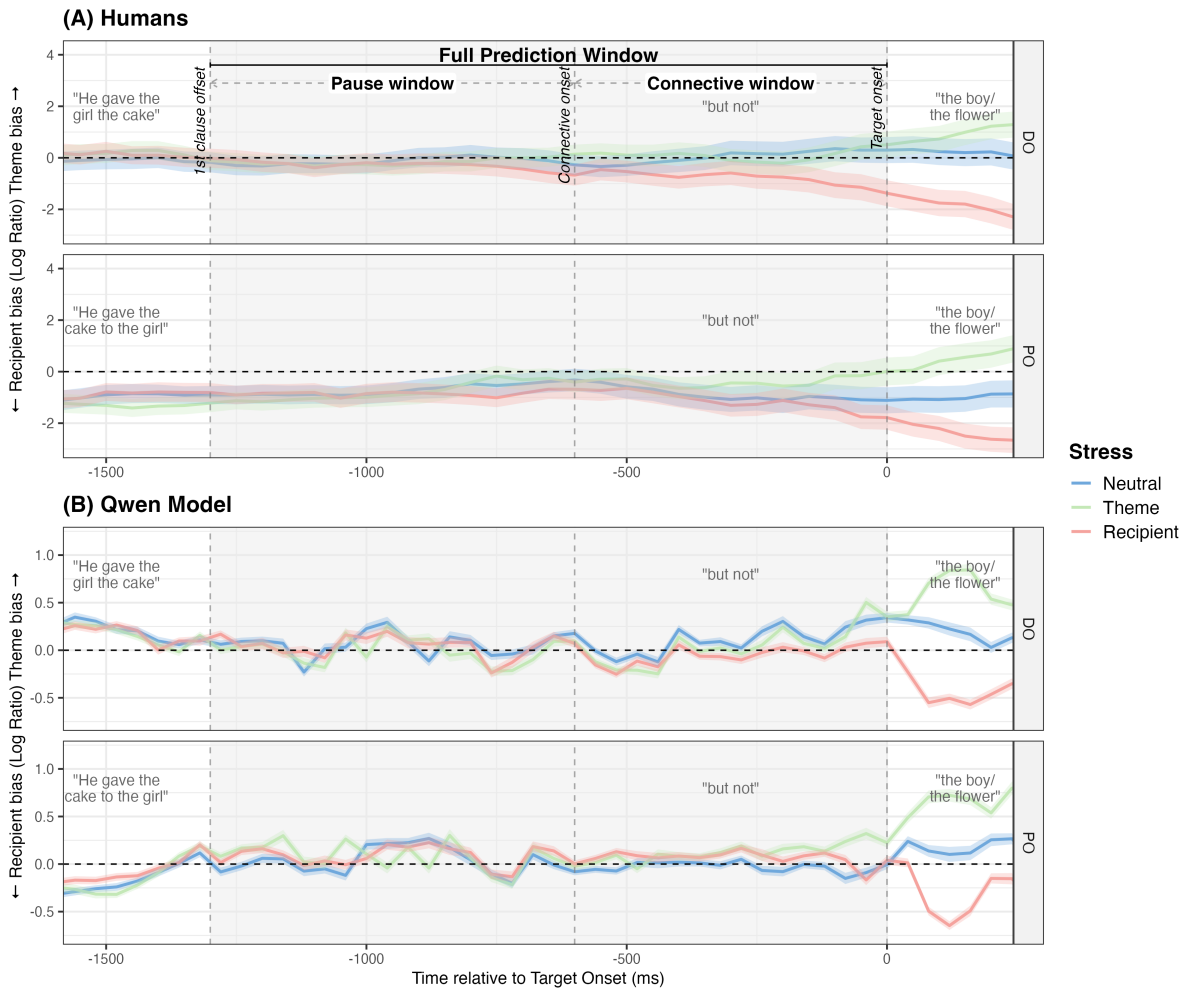

Figure 3: Time-course of online predictive bias in Experiment 2 (standard VWP). Upper panel: human eye-tracking results. Lower panel: model attention results. Data are aligned to target-word onset (0 ms). The y-axis represents the log-ratio bias $[\log(\text{Alt. Theme}/\text{Alt. Recipient})]$, derived from fixation proportions (humans) and audio-to-image cross-modal attention (model). Positive values indicate a bias toward the theme; negative values indicate a bias toward the recipient. The grey shaded region (-1300 to 0 ms) marks the prediction window (from first clause offset to target-word onset). Shaded bands denote 95% confidence intervals.

tral, whereas Theme stress did not; in PO structures, the reverse held. And this pattern appeared both at the offline output level (Experiment 1) and in the online time course (Experiment 2).

These findings extend prior work on focus interpretation and cue interaction (Clifton and Frazier, 2016; Harris and Carlson, 2018; Yan and Calhoun, 2020) to a spoken, visually grounded prediction task. They also show why separating output and time-course measures is useful. The cloze-in-VWP task identified the referential alternatives listeners selected after the preamble, whereas the standard VWP showed that structural information influenced fixation patterns earlier than prosodic stress. The human results therefore support a graded cue-integration account: default focus provides an early structural bias, while prosodic prominence modulates that bias as the sentence unfolds.

5.2 Human–Model Alignment Across Output and Process Levels

Comparing Mandarin speakers and Qwen2.5-Omni-7B reveals partial output-level alignment together with systematic differences in task grounding, cue weighting, and measured temporal profiles.

Partial output-level alignment. At the level of offline prediction outputs, the model reproduced several main human patterns within the subset of fully on-task alternative continuations: an overall theme bias, a Structure effect, Stress effects, and

a cue-diagnosticity interaction. This partial alignment is consistent with evidence that LLMs are sensitive to formal linguistic regularities (Linzen and Baroni, 2021; Mahowald et al., 2024) and with the use of cloze-style prediction as a basis for human–model comparison (Eisape et al., 2020; Goldstein et al., 2022). However, this alignment is conditional on responses that remained within the displayed alternative theme–recipient space, and should not be interpreted as full behavioral equivalence.

Divergence in cue weighting and temporal deployment. Output-level similarity coexisted with two systematic divergences. First, the model's structure effect was approximately three times the human magnitude. In the PO-Recipient condition, where prosodic and structural cues converged on the same constituent, the model's predictions moved in the opposite direction from human predictions. Second, the measured time courses of cue deployment differed: human structural sensitivity was reliable before the contrastive connective, whereas the model's attention-derived sensitivity was reliable later, after connective onset. Prosodic effects showed a similar difference in the within-structure analyses. These findings support the deployment-dissociation prediction introduced in Section 2.3: similar output-level patterns can be accompanied by different temporal cue-deployment profiles in this paradigm.

This pattern can be described using the distinction between formal and functional linguistic competence (Mahowald et al., 2024), with the qualifications developed in Section 2.3. The model showed sensitivity to syntactic structure and prosodic prominence, yet its weighting and situated deployment of these cues differed from human comprehenders in the present visually grounded, time-sensitive task. More generally, behavioral or representational similarity does not by itself license inferences about shared cognitive mechanisms (Guest and Martin, 2023; Lin, 2025). The relevant conclusion here is therefore not that the systems have fundamentally different mechanisms, but that partial output-level alignment was not accompanied by similar temporal cue-deployment profiles.

On-task prediction and visual grounding. Beyond the contrasts captured by the primary GLMM, the higher exclusion rate for the model (33.3% vs. 5.4% for humans; Appendix B.1.3) indicates another locus of divergence. The primary GLMM compares humans and the model only within the displayed alternative theme–recipient space; outside that subset, Qwen2.5-Omni-7B was less consistently constrained by the task-defined alternatives and, in some cases, by the visual display. Thus, the partial output-level alignment reported above holds for the subset of fully on-task alternative continuations; outside that subset, the model already shows a broader task-grounding difference. Together with the time-course results, this suggests that human–model differences arise not only in the timing of cue deployment, but also in the reliability with which visually grounded, task-appropriate predictions are produced.

Implications and future directions. These findings support multi-level human–model comparison: output choices, task-grounding behavior, and process-sensitive time courses can dissociate. For evaluating multimodal LLMs, the findings underline the value of process-sensitive measures that complement output comparisons: output similarity in a single paradigm should not be read as sufficient evidence for shared processing dynamics. Methodologically, the cross-modal attention measure used here is one of several possible operational measures, and the design choices involved (Section 3) leave room for further refinement, for instance through gradient-weighted attention measures (Azarkhalili and Libbrecht, 2025) or layer-resolved analyses (Goldstein et al., 2025). Future work should test whether the same patterns hold across additional multimodal models, languages, and prediction paradigms, and should further examine layer- or head-specific attention patterns and intervention-based measures of model cue use.

6 Conclusion

This study compared Mandarin speakers and Qwen2.5-Omni-7B on the integration of syntactic structure and prosodic prominence during visually grounded spoken language prediction. Across two matched experiments, humans and the model showed partial output-level alignment: both were sensitive to structure and prosody, and both showed a cue-diagnosticity pattern in which prosody had its clearest effect on constituents lacking default structural focus. At the same time, the model showed stronger structure weighting, condition-specific prosody deviations, a higher rate of responses outside the task-defined alterna-

tive space, and later attention-derived cue effects. These findings indicate that output-level alignment in this paradigm can dissociate from task-grounding behavior and from process-sensitive temporal profiles. More broadly, the study shows how controlled psycholinguistic paradigms can support multi-level human–model comparison by jointly assessing what systems predict, how reliably those predictions remain grounded in the task context, and when relevant cues influence attention to potential referents.

Limitations

Several limitations warrant acknowledgment. First, the model results are based on a single multimodal LLM (Qwen2.5-Omni-7B) tested on one Mandarin dative prediction task. The observed pattern—partial output-level alignment, higher model exclusion rates, and later attention-derived cue effects—may reflect properties of this architecture, its training data, the prompt format, or the specific structure–prosody manipulation. Testing additional speech–vision models, model sizes, prompting strategies, and decoding settings is necessary before drawing conclusions about multimodal LLMs more broadly.

Second, the human and model tasks were closely matched but not identical: human participants perceived spoken sentences and visual displays in an experimental interface, whereas the model received audio and image inputs through its processor and a task prompt, with visual displays adapted to the model's image-tokenization constraints. These adaptations were necessary for a controlled comparison, but the two systems were not exposed to identical perceptual environments.

Third, the model-side online measure is an operational proxy. Audio-to-image cross-modal attention provides a time-resolved signal of how the model allocates weight from spoken input to visual regions, but it is not a direct equivalent of human gaze, and its construction involves design choices (Section 3). Although layer-specific time courses were retained for statistical modeling, the present analyses do not establish which heads or layers are causally responsible for the observed effects; intervention-based methods or alternative attribution measures (Azarkhalili and Libbrecht, 2025) would be needed to strengthen mechanistic interpretation.

Finally, the human results come from Mandarin speakers and one construction type. Extending the paradigm to other languages, constructions, and prosodic systems would clarify which aspects of the cue-diagnosticity pattern and the human–model deployment dissociation generalize beyond this task.

Ethics Statement

All human experiments were conducted with informed consent under procedures approved by The Chinese University of Hong Kong. Participants were recruited as described in Appendix B, compensated for their time, and informed that their data would be used for research purposes. All released human data are anonymized and contain no personally identifying information. Visual stimuli were sourced from Clipart.com, a commercial image repository; redistribution of clipart-derived materials follows the applicable licensing restrictions. The study evaluates a publicly described model for scientific purposes and does not involve deployment in sensitive decision-making contexts.

Data and Code Availability

Analysis scripts, model prompting code, the attention-extraction pipeline, model outputs, anonymized analysis-ready human data, stimulus metadata, audio files, and display metadata are available at `https://github.com/wang-shuqi/conll2026-vwp-qwen-prediction`. Clipart-derived visual materials are shared only to the extent permitted by the applicable license.

Acknowledgments

We thank Chi Fong Wong for helpful suggestions on the code implementation, and Wu Hanlin for recording the auditory stimuli. We are also grateful to the participants of the norming studies and main experiments for their time, and to the anonymous CoNLL reviewers for their constructive feedback.

References

Gerry TM Altmann and Yuki Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

Alexander L. Anwyl-Irvine, Jessica Massonnié, Adam Flitton, Natasha Kirkham, and Jo K. Evershed. 2020. Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, 52(1):388–407.

Manabu Arai, Roger P. Van Gompel, and Christoph Scheepers. 2007. Priming ditransitive structures in comprehension. *Cognitive Psychology*, 54(3):218–250.

Behrooz Azarkhalili and Maxwell Libbrecht. 2025. Generalized attention flow: Feature attribution for transformer models via maximum flow. ArXiv:2502.15765.

Dale J. Barr. 2008. Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, 59(4):457–474.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Zhenguang G. Cai, Martin J. Pickering, and Patrick Sturt. 2013. Processing verb-phrase ellipsis in Mandarin Chinese: Evidence against the syntactic account. *Language and Cognitive Processes*, 28(6):810–828.

Zhenguang G. Cai, Nan Zhao, and Martin J. Pickering. 2022. How do people interpret implausible sentences? *Cognition*, 225:105101.

Sasha Calhoun. 2010. The centrality of metrical structure in signaling information structure: A probabilistic perspective. *Language*, 86(1):1–42.

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134.

Charles Clifton and Lyn Frazier. 2016. Focus in corrective exchanges: Effects of pitch accent and syntactic form. *Language and Speech*, 59(4):544–561.

Ruth E. Corps, Charlotte Brooke, and Martin J. Pickering. 2022. Prediction involves two stages: Evidence from visual-world eye-tracking. *Journal of Memory and Language*, 122:104298.

Ruth E. Corps, Meijian Liao, and Martin J. Pickering. 2023. Evidence for two stages of prediction in non-native speakers: A visual-world eye-tracking study. *Bilingualism: Language and Cognition*, 26(1):231–243.

Delphine Dahan, Michael K. Tanenhaus, and Craig G. Chambers. 2002. Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, 47(2):292–314.

Oliver Eberle, Stephanie Brandl, Jonas Pilot, and Anders Søgaard. 2022. Do transformer models show similar attention patterns to task-specific human gaze? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4295–4309, Dublin, Ireland. Association for Computational Linguistics.

Tiwalayo Eisape, Noga Zaslavsky, and Roger Levy. 2020. Cloze distillation: Improving neural language models with human next-word prediction. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 401–411.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. ArXiv:1903.03260.

Changjiang Gao, Zhengwu Ma, Jiajun Chen, Ping Li, Shujian Huang, and Jixing Li. 2025. Increasing alignment of large language models with language processing in the human brain. *Nature Computational Science*, 5(11):1080–1090.

Ariel Goldstein, Eric Ham, Mariano Schain, Samuel A. Nastase, Bobbi Aubrey, Zaid Zada, Avigail Grinstein-Dabush, Harshvardhan Gazula, Amir Feder, Werner Doyle, Sasha Devore, Patricia Dugan, Daniel Friedman, Michael Brenner, Avinatan Hassidim, Yossi Matias, Orrin Devinsky, Noam Siegelman, Adeen Flinker, and 1 others. 2025. Temporal structure of natural language processing in the human brain corresponds to layered hierarchy of large language models. *Nature Communications*, 16(1):10529.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, and 1 others. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380.

Olivia Guest and Andrea E. Martin. 2023. On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, 6(2):213–227.

Jesse A. Harris. 2023. The enduring effects of default focus in *Let Alone* ellipsis: Evidence from pupillometry. In *Experiments in Linguistic Meaning*, volume 2, pages 117–128.

Jesse A. Harris and Katy Carlson. 2018. Information structure preferences in focus-sensitive ellipsis: How defaults persist. *Language and Speech*, 61(3):480–512.

Falk Huettig, Joost Rommers, and Antje S. Meyer. 2011. Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, 137(2):151–171.

Aine Ito and Pia Knoeferle. 2022. Using the visual world paradigm to study sentence comprehension in typologically diverse languages. In *Language, Cognition, and Mind*. Springer.

Aine Ito, Martin J. Pickering, and Martin Corley. 2018. Investigating the time-course of phonological prediction in native and non-native speakers of English: A visual world eye-tracking study. *Journal of Memory and Language*, 98:1–11.

Kiwako Ito and Shari R. Speer. 2008. Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language*, 58(2):541–573.

Gina R. Kuperberg and T. Florian Jaeger. 2016. What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31(1):32–59.

Weijun Li, Nianlong Deng, Yufang Yang, and Lin Wang. 2018. Process focus and accentuation at different positions in dialogues: An ERP study. *Language, Cognition and Neuroscience*, 33(2):255–274.

Zhicheng Lin. 2025. Six fallacies in substituting large language models for human participants. *Advances in Methods and Practices in Psychological Science*, 8(3):25152459251357566.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(1):195–212.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press, Cambridge, MA.

Evelyn Milburn, Tessa Warren, and Michael Walsh Dickey. 2016. World knowledge affects prediction as quickly as selectional restrictions: Evidence from the visual world paradigm. *Language, Cognition and Neuroscience*, 31(4):536–548.

Martin J. Pickering and Chiara Gambi. 2018. Predicting while comprehending language: A theory and review. *Psychological Bulletin*, 144(10):1002–1044.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020. Interpreting attention models with human visual attention in machine reading comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 12–25.

Michael K. Tanenhaus, Michael J. Spivey-Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Wilson L. Taylor. 1953. "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433.

Jacolien van Rij, Petra Hendriks, Hedderik van Rijn, R. Harald Baayen, and Simon N. Wood. 2019. Analyzing the time course of pupillometric data. *Trends in Hearing*, 23:1–22.

Jacolien van Rij, Martijn Wieling, R. Harald Baayen, and Hedderik van Rijn. 2015. *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs*. R package.

Andrea Weber, Martine Grice, and Matthew W. Crocker. 2006. The role of prosody in the interpretation of structural ambiguities: A study of anticipatory eye movements. *Cognition*, 99(2):B63–B72.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. ArXiv:2006.01912.

Simon N. Wood. 2017. *Generalized Additive Models: An Introduction with R*, 2nd edition. Chapman and Hall/CRC.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. ArXiv:2503.20215.

Liejiong Xu. 2004. Manifestation of informational focus. *Lingua*, 114(3):277–299.

Mengzhu Yan and Sasha Calhoun. 2020. Rejecting false alternatives in Chinese and English: The interaction of prosody, clefting, and default focus position. *Laboratory Phonology*, 11(1).

Jayden Ziegler and Jesse Snedeker. 2019. The use of syntax and information structure during language comprehension: Evidence from structural priming. *Language, Cognition and Neuroscience*, 34(3):365–384.

A Stimuli Construction and Norming

This appendix details the construction, recording, and norming procedures for all experimental stimuli.

A.1 Sentence Construction

The complete set of 216 sentence stimuli (36 items $\times$ 6 conditions in the 2 $\times$ 3 factorial design crossing Structure with Stress), together with stimulus metadata, audio files, display metadata, and

sharable materials, is available in the public repository listed in the Data and Code Availability statement. Each item appeared in all six conditions, distributed across six experimental lists in a Latin square design. Individual visual display files are not redistributed due to commercial licensing of the clipart source; the assembly script and entity labels are provided so that displays can be regenerated.

Experimental sentences were selected and adapted from two sources: 22 items from Cai et al. (2013) and 28 items from Cai et al. (2022), yielding an initial pool of 50 items. Each item was constructed in both the DO and PO dative structures, followed by the contrastive conjunction 而不是 ('but not…').

Adaptations included converting items from Cai et al. (2013) to the perfective aspect (adding 了 *le*) and removing article–classifier modifiers, and translating items from Cai et al. (2022) (originally in English) into Mandarin. Items were then screened for syllable matching between recipient and theme nouns, entity uniqueness across items, imageability for the visual world paradigm, and visual distinctiveness of referent pairs. After applying these criteria, the initial pool was reduced to 42 items.

Cloze norming. Each experimental sentence ends with the contrastive conjunction 'but not…', after which a target word—either an alternative theme or an alternative recipient—is expected. To determine the most natural and plausible target for each item, a sentence completion (cloze) task was conducted. A total of 159 native Mandarin speakers participated via Qualtrics. Participants were distributed across six lists in a Latin square design (26–28 per list), yielding 6,678 valid data points. The task began with four practice trials, after which participants were reminded to (a) provide a natural continuation rather than transcribe the sentence, and (b) keep their responses concise and plausible. On each trial, participants listened to an audio recording of a sentence fragment ending with 'but not…' and typed a continuation. The highest-probability alternative was selected as the target for each condition, subject to two constraints: (a) the same entity could not serve as the target in more than one item, and (b) the entity had to be imageable for the visual world paradigm. When the highest-probability alternative violated these constraints, the next most probable alternative was selected instead. Two additional items were removed due to insufficient visual distinctiveness or semantic implausibility of entity combinations, yielding 40 items for audio recording and naturalness norming. Four further items were excluded after naturalness norming and visual-display preparation (see Appendix A.2), yielding the final set of 36 experimental items reported in Section 3.

A.2 Audio Recording, Editing, and Norming

Recording procedure. All sentences were recorded by a male native Mandarin speaker from Beijing with training in phonetics in a sound-proof recording booth. Recordings were made using a Shure SM58 dynamic microphone connected to a Focusrite Scarlett Solo audio interface, captured in Adobe Audition at a sampling rate of 44.1 kHz (mono). The speaker was instructed to produce each sentence as naturally as possible. For the Theme stress and Recipient stress conditions, the mentioned theme or mentioned recipient was highlighted in red in the recording script; the speaker was asked to produce contrastive stress on these words to naturally elicit a contrastive focus that would lead listeners to anticipate the upcoming alternative. For the Neutral condition, the speaker produced the sentence with even, unmarked prosody.

Post-recording editing. Individual sentence clips were extracted from the recording sessions using Adobe Audition. The preamble portion of each sentence (i.e., everything before the contrastive conjunction) was kept as naturally produced. To ensure temporal consistency across conditions and items, two regions were normalized: (a) the inter-clause silence between the preamble and the conjunction 'but not' was set to 700 ms (based on the grand mean of 697 ms across all recordings), and the silent region was replaced with silence; (b) the duration of the conjunction 'but not' was adjusted to 600 ms (based on the grand mean of 549 ms) using Audition's time-and-pitch stretching function, which preserves the original F0 contour. Leading silence before the subject and trailing silence after the target word offset were removed.

Segment boundaries were annotated in Adobe Audition for use in the eye-tracking analysis. Markers were placed at the onset of each major constituent, including the subject, verb, dative marker, aspect marker, recipient, theme, inter-clause si-

lence, conjunction, and each syllable of the target word. Together, the temporal normalization and segment annotation ensured that stimuli were temporally aligned across conditions and items, providing a consistent time-locking basis for the subsequent time-course analyses of eye-tracking data. All 216 audio files (MP3 format) are provided in the public repository listed in the Data and Code Availability statement.

Naturalness norming. A norming study was conducted to evaluate the naturalness of the recorded sentences. A total of 128 native Mandarin Chinese speakers (none of whom participated in the main experiment) were recruited via Qualtrics. The 40 items were distributed across 8 lists in a Latin square design, such that each participant heard each item in only one of the six conditions (2 structures $\times$ 3 stress patterns). Each participant rated 40 target sentences and 72 filler sentences. Participants listened to each sentence and rated its naturalness on a 5-point Likert scale (1 = *very unnatural*, 5 = *very natural*).

Based on the norming results, 4 items were excluded. Two items received low naturalness ratings (M = 2.41 and M = 2.80). The other two were excluded primarily for residual target overlap and imageability concerns that became apparent during visual display preparation, despite having acceptable naturalness ratings (M = 3.19 and M = 3.57). For the final set of 36 items, the overall mean naturalness rating was 3.22 (SD = 1.32). Across conditions, mean ratings ranged from 2.89 (DO, Recipient stress) to 3.48 (DO, Theme stress). No condition was excluded wholesale; the variation across conditions reflects natural differences in how contrastive stress interacts with argument structure in Mandarin.

A.3 Visual Display Construction

Image sourcing. All images were sourced from Clipart.com,[1] a commercial clipart library. This source was chosen to ensure a consistent illustrative style across all referent images. As a subscription-based service whose individual image files require licensing for use, Clipart.com images are less likely than freely available web images to appear in standard web-scraped pretraining corpora; this choice was made to reduce the likelihood that the specific visual stimuli were memorized during model pretraining, though we cannot directly verify whether they appear in Qwen2.5-Omni's training data. Each experimental item required four images: the mentioned recipient, the mentioned theme, the alternative recipient, and the alternative theme. For each entity, two candidate images were selected via keyword search based on the criteria of imageability and stylistic fit, yielding two image lists for subsequent norming.

Image preparation. Each image was processed in Adobe Photoshop following a standardized procedure: (a) the image was placed on a 2400 $\times$ 2400 pixel transparent canvas, (b) the original background was removed, (c) the referent was centered with consistent margins on all sides, and (d) the image was exported as a PNG file. For consistency, human referents were depicted in matching perspectives (e.g., if the mentioned recipient was shown in full-body view, the alternative recipient was also shown in full-body view).

Name agreement norming. A name agreement norming study was conducted to select the best image for each referent. For each entity, two candidate clipart images were selected, yielding two image lists. A total of 45 native Mandarin speakers (19 female, 26 male; mean age = 21.7 years) participated via Qualtrics, with approximately half assigned to each list (22 to List 1, 23 to List 2). On each trial, participants saw a noun label alongside one of the candidate images and rated how well the image matched the label on a 5-point Likert scale (1 = *very inconsistent*, 5 = *very consistent*). Images were displayed at 300 $\times$ 300 pixels.

Image selection criteria. The final image for each entity was selected based on the following considerations, in order of priority:

1. **Name agreement score**: The image with the higher mean rating was generally preferred. However, when the two candidates did not differ significantly, additional criteria were applied.
2. **Stylistic consistency**: Images within the same item were matched for artistic style (e.g., both cartoon or both realistic), body perspective (full-body vs. half-body for human referents), and color palette.
3. **Visual distinctiveness**: The four images in each visual display needed to be clearly distinguishable from one another at the experimental display size.

[1] https://clipart.com

Images that required replacement during the screening process (e.g., due to entity substitutions) underwent an additional round of norming with the same procedure.

Visual display assembly. The four images for each item (mentioned recipient, mentioned theme, alternative recipient, alternative theme) were combined into a single visual display. Each referent appeared equally often in each of the four screen positions (top-left, top-right, bottom-left, bottom-right) across items, ensuring that position did not systematically cue any entity role.

For the human eye-tracking experiment, individual images were scaled to 300 × 300 pixels and placed on a 1024 × 768 pixel canvas, following the display resolution commonly used in visual world studies (e.g., Corps et al., 2023; Ito et al., 2018). The four images were centered at the following coordinates: top-left (256, 192), top-right (768, 192), bottom-left (256, 576), and bottom-right (768, 576).

For the multimodal LLM experiment, the display dimensions were adjusted to accommodate Qwen's image tokenization, which operates in multiples of 28 pixels. Each image was scaled to 308 × 308 pixels and placed on a 1008 × 756 pixel canvas with spacing between images to preserve the spatial layout: top-left [84, 28, 392, 336], top-right [616, 28, 924, 336], bottom-left [84, 420, 392, 728], and bottom-right [616, 420, 924, 728]. The displays were generated programmatically using Python. Because the individual clipart images are subject to commercial licensing restrictions, the repository provides the display assembly scripts, entity labels, and display metadata; redistribution of the original image files is limited to what is permitted by the applicable license.

A.4 Filler Items

In addition to the 36 experimental items, 72 filler sentences were included. Fillers used seven different connective types—但是 ('but'), 而不是 ('but not'), 然后 ('then'), 并且 ('and'), 所以 ('so'), 因为 ('because'), and 只好 ('had to')—to prevent participants from developing expectations specific to any single construction. These fillers varied in whether the post-conjunction target referent was a person or a non-person entity (36 each) and in whether the sentence contained a prosodically stressed word (30 stressed, 42 neutral). The stressed fillers varied in which constituent carried prominence, including adverbs, adjectives, nouns, and verbs, so that participants could not use the presence of stress alone to predict which word class would be highlighted.

The filler design served several purposes: (a) the variety of connective types ensured that participants processed each sentence for meaning rather than developing a strategy specific to the 而不是 ('but not') construction used in experimental items; (b) the balanced distribution of person and non-person targets prevented participants from anticipating the semantic category of the upcoming referent; and (c) the varied stress patterns prevented participants from associating prosodic prominence exclusively with the contrastive stress patterns used in experimental items. Visual displays for filler items were prepared following the same procedure as for experimental items.

B Data Collection and Experiment Procedure

This appendix provides detailed procedural information for all four sub-experiments, supplementing the overview in Section 3.

B.1 Experiment 1: Cloze-in-VWP

B.1.1 Human Experiment

Participants. 118 native Mandarin-speaking university students (62 female; mean age = 21.64 years) were recruited from universities in mainland China. All reported normal or corrected-to-normal vision and normal hearing. None had participated in the norming studies described in Appendix A. Participants were randomly assigned to one of six Latin-square lists (19–20 per list) and received 15 RMB compensation for the approximately 30-minute session.

Platform. The experiment was conducted online using Gorilla Experiment Builder (`https://gorilla.sc`; Anwyl-Irvine et al., 2020). Participants completed the experiment on personal computers with headphones in a quiet environment.

Trial procedure. Each trial consisted of the following sequence:

1. **Fixation cross** (500 ms): A central fixation cross appeared on screen.
2. **Visual preview** (2000 ms): The four-image display appeared without audio.
3. **Audio presentation**: The sentence preamble was played, truncated after the contrastive

connective *but not* (而不是). The visual display remained on screen throughout.

4. **Response phase**: Participants clicked on the image they predicted would be mentioned next, then typed their predicted continuation. Both responses were required before advancing.

Design and list structure. The 36 experimental items were distributed across six lists in a Latin-square design, ensuring each participant encountered each item in exactly one of the six conditions. Each list contained 36 experimental items and 72 filler items (108 trials total), presented in pseudo-randomized order with no more than two consecutive experimental items. Two practice trials preceded the main experiment.

B.1.2 Model Experiment

Model and hardware. We used Qwen2.5-Omni-7B (Xu et al., 2025), a multimodal LLM capable of processing text, audio, and visual inputs. The model was run on Google Colab with a single NVIDIA A100 40 GB GPU, using the Hugging-Face Transformers library.

Prompt design. The model received a Chinese-language prompt specifying the task context. The following is an English translation of the instruction; the exact prompt template is provided in the public repository listed in the Data and Code Availability statement:

> **Instruction**: You are a participant in a psycholinguistic experiment. You will first see four images and then hear a sentence related to the images (the sentence will mention some of the images). Your task is: look at the images carefully and listen to the sentence to understand its meaning. Note: at the end of some trials, there will be a question about the trial. Below are the images and the sentence: `<image><audio>`

The `<image>` and `<audio>` tokens were replaced with the assembled visual display (1008 × 756 pixels) and the audio stimulus, respectively. The model was then prompted to select a quadrant and provide a text continuation.

Response collection. For each of the 36 items in each of the six lists, the model completed 20 independent generations, yielding 4,320 responses. Decoding used `do_sample=True`, `temperature=1.0`, `top_p=0.9`, and `max_new_tokens=64`; all unspecified generation parameters followed the model and Transformers defaults. We used the same fixed set of 20 random seeds for all items: 11, 23, 37, 41, 53, 67, 79, 83, 97, 101, 113, 127, 131, 149, 163, 173, 181, 197, 211, and 223. Each generation used a fresh conversation context. The 20 generations were not averaged: each was coded as a trial-level response and entered the GLMM as an observation, with run-level variability modeled in the statistical analyses. The model prompting notebook and outputs are available in the public repository listed in the Data and Code Availability statement.

B.1.3 Response Coding and Exclusions

Responses in Experiment 1 were coded from the text continuation rather than from the raw click or selected quadrant alone. Each continuation was classified as referring to the alternative theme, the alternative recipient, or other. We also flagged whether the referred entity was depicted in the display and whether it was one of the two alternative referents. When the selected image or quadrant conflicted with an unambiguous text continuation, the continuation determined the final response category. The same continuation-based coding rule was applied to human and model outputs.

The primary binary GLMM included only continuations that unambiguously referred to the displayed alternative theme or displayed alternative recipient. We excluded responses coded as other, ambiguous, action-based, or off-task; responses referring to entities not depicted in the display; and responses referring to the already mentioned theme or recipient rather than to one of the two alternative referents. Table 1 summarizes the resulting exclusion counts.

Dataset	Total	Excluded	GLMM N
Human	4,246	231	4,015
Qwen2.5-Omni-7B	4,320	1,439	2,881

Table 1: Experiment 1 response exclusions for the binary theme-versus-recipient GLMM. Excluded responses were coded as other, ambiguous, action-based, or off-task; referred to entities not depicted in the visual display; or referred to the already-mentioned theme or recipient rather than to one of the two alternative referents.

B.2 Experiment 2: Standard VWP

B.2.1 Human Experiment

Participants. 58 native Mandarin speakers (30 female; mean age = 22.3 years) were recruited from The Chinese University of Hong Kong and

local participant pools. All reported normal or corrected-to-normal vision and normal hearing. None had participated in Experiment 1 or the norming studies. Participants received 100 HKD compensation for the approximately 40-minute session.

Apparatus. Eye movements were recorded using an EyeLink 1000 Plus desktop-mounted eye-tracker (SR Research, Ontario, Canada) at 1000 Hz sampling rate, tracking the right eye. Stimuli were displayed on a 24-inch monitor (1024 × 768 pixels) at approximately 60 cm viewing distance. Audio was presented through headphones. A chin rest and forehead rest stabilized head position. A 9-point calibration and validation procedure was performed at the start, with recalibration between blocks as needed.

Trial procedure. Each trial consisted of the following sequence:

1. **Drift correction**: A central fixation point appeared; the experimenter initiated the trial once stable fixation was achieved.
2. **Visual preview** (2000 ms): The four-image display appeared without audio.
3. **Audio presentation**: The complete sentence (including the target word after *but not*) was played. The visual display remained on screen.
4. **Comprehension question**: On one-third of trials (randomly selected), a yes/no question appeared to ensure attention.

Design and list structure. The same Latin-square design as Experiment 1 was used. Each list contained 36 experimental and 72 filler items, presented in four blocks of 27 trials with breaks between blocks.

Eye-tracking data preprocessing. Fixations were assigned to five areas of interest: four image quadrants (TL, TR, BL, BR) and the remaining screen area (REST). Fixation proportions were computed for each 50 ms time bin, following the standard bin size in VWP research (e.g., Altmann and Kamide, 1999; Ito et al., 2018; Corps et al., 2023), aligned to target word onset (time = 0 ms). Blinks and track losses were excluded from proportion calculations.

B.2.2 Model Experiment

Attention extraction procedure. Cross-modal attention weights were extracted from Qwen2.5-Omni-7B using the following procedure. Qwen2.5-Omni concatenates all modality tokens (text, audio, image) into a single sequence processed through standard Transformer self-attention—there is no separate cross-modal attention module; self-attention *is* cross-modal attention when queries and keys span different modalities.

1. **Modality boundary identification**: Audio and image token boundaries were identified via special tokens (`<|audio_start|>`, `<|audio_end|>`, `<|vision_start|>`, `<|vision_end|>`).
2. **Attention submatrix extraction**: From each of the 28 layers' attention matrices (heads × Q × K), the submatrix $\mathbf{A}$[:, audio, image] was extracted, capturing how each audio token attends to each image patch.
3. **Head aggregation**: Attention weights were averaged across all attention heads within each layer.
4. **Region mapping**: The visual display was encoded by the Vision Transformer into a 27 × 36 patch grid. Each display quadrant maps to a set of patches corresponding to the bounding box of one image. The four quadrant regions in pixel coordinates were: TL [84, 28, 392, 336], TR [616, 28, 924, 336], BL [84, 420, 392, 728], BR [616, 420, 924, 728]. All coordinates are multiples of 28 pixels (= patch size × merge factor = 14×2), ensuring clean mapping between pixels and patches.
5. **Aggregation and normalization**: Attention was averaged within each quadrant and normalized across four quadrants plus a REST category (patches outside any quadrant), yielding AOI-style regional proportions that parallel fixation proportions for analysis.

Temporal alignment. Qwen2.5-Omni processes audio at 25 Hz (one token per 40 ms). Audio token indices were converted to millisecond timestamps and aligned to target word onset, producing time-course data at 40 ms resolution comparable to human fixation proportions at 50 ms resolution. Sentences were truncated before target nouns, matching the human prediction window. We retained layer-specific regional time courses for inferential analyses, with Layer treated as a grouping factor in the statistical models; visualizations report layer-aggregated averages. The resulting audio-to-image attention measure is an

operational index of model allocation to visual regions during spoken input, not a direct equivalent of human gaze. The attention-extraction pipeline and layer-specific attention data are available in the public repository listed in the Data and Code Availability statement.

C Statistical Model Specifications

This appendix provides full specifications for all statistical models reported in the paper. Analyses were conducted in R (version 4.3+) using `lme4` (Bates et al., 2015) for generalized linear mixed-effects models (GLMM) and linear mixed-effects models (LME), and `mgcv` (Wood, 2017) for generalized additive mixed models (GAMM). GAMM models were fit using the `bam()` function for computational efficiency with large time-course datasets. All GAMM models included AR1 autocorrelation correction to account for temporal dependencies in adjacent time bins; the AR1 parameter (ρ) was estimated from a base model without the correction using `start_value_rho()` from the `itsadug` package (van Rij et al., 2015). Model comparison via AIC confirmed that AR1 models provided substantially better fit in all cases.

C.1 Experiment 1: Offline Prediction Output

C.1.1 Dependent Variable and Coding

The dependent variable for Experiment 1 was continuation type, coded as 1 for alternative-theme continuations and 0 for alternative-recipient continuations. Positive coefficients therefore indicate increased theme bias. Responses were included only if the text continuation unambiguously referred to the displayed alternative theme or displayed alternative recipient. Responses classified as other, ambiguous, action-based, or off-task, responses referring to entities not depicted in the display, and responses referring to already-mentioned entities were excluded from the binary analyses (Appendix B.1.3).

Fixed effects were coded as follows:

- **Structure**: sum-coded (DO = -0.5, PO = 0.5). With the dependent variable coded as alternative theme = 1, negative Structure coefficients indicate greater theme bias in DO than PO.
- **Stress**: treatment-coded with Neutral as reference level (Theme vs. Neutral; Recipient vs. Neutral).
- **Participant Type** (combined analysis only): treatment-coded with Human as reference level.

C.1.2 Model Selection

Random effects structures were determined by forward model comparison: starting from a model with only random intercepts for Subject (or Run, for the model) and Item, additional random slopes were added incrementally, and models were compared using likelihood ratio tests (`anova()`). The most complex model that converged and significantly improved fit was retained.

All GLMMs were fit with the `bobyqa` optimizer and a maximum of 10^6 function evaluations to ensure convergence.

C.1.3 Human Model

```
glmer(T_con ~ cStructure * Stress
  + (1 + Structure | SubjectID)
  + (1 | Item),
  family = binomial)
```

N = 4,015 observations; 118 participants; 36 items.

Table 2: GLMM results for human responses (Experiment 1).

Fixed effect	β	SE	z	p
Intercept	2.51	0.18	14.02	$< .001$
Structure	-1.56	0.23	-6.76	$< .001$
Stress: Thm	0.66	0.14	4.64	$< .001$
Stress: Rec	-1.33	0.12	-11.37	$< .001$
Str $\times$ Thm	0.67	0.29	2.32	.020
Str $\times$ Rec	1.00	0.23	4.27	$< .001$

C.1.4 Model (Qwen2.5-Omni-7B)

```
glmer(T_con ~ cStructure * Stress
  + (1 + cStructure | Run)
  + (1 + cStructure
      + cStructure:Stress | Item),
  family = binomial)
```

N = 2,881 observations; 20 runs; 36 items. Each of the 36 items in each of 6 lists was presented to the model 20 times before response exclusions. Run-level variability was modeled via the Run grouping factor.

C.1.5 Combined Human–Model Analysis

```
glmer(T_con ~
  cStructure * Stress * Participant
  + (1 + cStructure + Participant
     | SubjectID)
  + (1 + Stress + cStructure
      + Participant | Item),
  family = binomial)
```

Table 3: GLMM results for model responses (Experiment 1).

Fixed effect	β	SE	z	p
Intercept	2.65	0.39	6.84	$< .001$
Structure	-4.69	0.59	-7.93	$< .001$
Stress: Thm	1.28	0.27	4.66	$< .001$
Stress: Rec	-0.94	0.22	-4.28	$< .001$
Str $\times$ Thm	2.70	0.89	3.02	.003
Str $\times$ Rec	2.75	0.66	4.16	$< .001$

N = 6,896 observations (4,015 human + 2,881 model). SubjectID indexes individual human participants and model runs (Omni_01–Omni_20). Participant Type is treatment-coded with Human as reference; thus, interaction terms reflect how model behavior *deviates* from the human pattern.

C.2 Experiment 2: Online Processing Dynamics

C.2.1 Dependent Variable

The dependent variable was the log ratio of attention to the alternative theme location versus the alternative recipient location:

$$\mathrm{DV} = \log\left(\frac{\mathrm{Prop(Alt.\ Theme)}}{\mathrm{Prop(Alt.\ Recip.)}}\right) \tag{1}$$

Positive values indicate a theme bias; negative values indicate a recipient bias. The log ratio is a standard dependent variable in VWP research for comparing fixations between two co-present regions of interest without violating the independence assumption (Arai et al., 2007; Barr, 2008). A small constant (0.001) was added to both numerator and denominator to avoid undefined values when either proportion was zero (Ito and Knoeferle, 2022).

C.2.2 Time Alignment and Windows

All time-course data were aligned to target word onset (Time = 0 ms). Two theoretically motivated time windows were defined:

- **Pause window** (-1300 to -600ms): from first clause offset to connective onset, corresponding to the inter-clause silence period.
- **Connective window** (-600 to 0 ms): from connective onset to target-word onset, corresponding to presentation of *but not* (而不是).

C.2.3 Window-Based Analysis: LME Models

Fixed effects were coded identically to Experiment 1. Random effects structures were determined by forward model comparison.

Human.

```
# Pause window
lmer(log_ratio_alt ~
  Stress * cStructure
  + (1 | Subject)
  + (1 + Stress | Item))

# Connective window
lmer(log_ratio_alt ~
  Stress * cStructure
  + (1 + Stress | Subject)
  + (1 | Item))
```

Model (Qwen2.5-Omni-7B). For model data, Layer (L1–L28) served as the grouping factor for layer-level variability. This grouping factor captures variation across Transformer layers for statistical modeling; it is not intended as a cognitive analogue of human participant variability.

```
# Pause window
lmer(log_ratio_alt ~
  Stress * cStructure
  + (1 + cStructure | Layer)
  + (1 + Stress * cStructure
     | Item))

# Connective window
lmer(log_ratio_alt ~
  Stress * cStructure
  + (1 + cStructure | Layer)
  + (1 + Stress * cStructure
     | Item))
```

C.2.4 Time-Course Analysis: GAMM Models

All GAMM models followed the same general specification:

```
bam(log_ratio_alt ~ Condition
  + s(Time, by = Condition)
  + s(Time, GroupingFactor,
      by = Condition,
      bs = "fs", m = 1)
  + s(Time, Item,
      by = Condition,
      bs = "fs", m = 1),
  rho = rho_value,
  AR.start = Is_start)
```

where `Condition` is either Structure (DO vs. PO) or Stress (e.g., Recipient vs. Neutral), and `GroupingFactor` is Subject (for human data) or Layer (for model data). The `bs = "fs"` specification creates factor smooth interactions, allowing each level of the grouping factor to have its own smooth over time—the recommended approach for modeling individual differences in nonlinear time-course data (van Rij et al., 2019). The `m = 1` setting applies a first-derivative penalty for computational efficiency.

Significant time windows were identified using the `plot_diff()` function from `itsadug`, which computes pointwise confidence intervals for the

Table 4: GLMM results for the combined human–model analysis in Experiment 1 after excluding continuations outside the displayed alternative theme–recipient space. Participant: Qwen indicates how the model deviates from the human reference level. The dependent variable was coded as 1 for alternative-theme continuations and 0 for alternative-recipient continuations; positive coefficients indicate increased theme bias.

Fixed effect	β	SE	z	p
Intercept	2.61	0.20	12.91	$< .001$
Structure	-1.67	0.24	-6.82	$< .001$
Stress: Theme	0.55	0.17	3.18	.001
Stress: Recipient	-1.41	0.15	-9.72	$< .001$
Participant: Qwen	-0.13	0.38	-0.34	.737
Structure × Theme	0.77	0.29	2.63	.008
Structure × Recipient	1.09	0.24	4.55	$< .001$
Structure × Qwen	-2.78	0.43	-6.49	$< .001$
Theme × Qwen	0.45	0.27	1.69	.091
Recipient × Qwen	0.53	0.22	2.45	.014
Str × Thm × Qwen	1.91	0.51	3.73	$< .001$
Str × Rec × Qwen	1.46	0.43	3.41	$< .001$

Table 5: LME results for human eye-tracking data (Exp. 2), Pause window (-1300 to -600ms).

Fixed effect	β	SE	t	p
Intercept	-0.54	0.24	-2.24	.031
Str: Thm	-0.07	0.25	-0.27	.792
Str: Rec	-0.30	0.27	-1.09	.283
Structure	-0.81	0.29	-2.82	.005
Thm×Str	-0.35	0.40	-0.86	.389
Rec×Str	-0.07	0.40	-0.18	.861

Table 6: LME results for human eye-tracking data (Exp. 2), Connective window (-600 to 0 ms).

Fixed effect	β	SE	t	p
Intercept	-0.39	0.19	-1.99	.051
Str: Thm	0.43	0.28	1.57	.123
Str: Rec	-0.85	0.25	-3.40	.001
Structure	-1.14	0.32	-3.57	$< .001$
Thm×Str	0.41	0.45	0.90	.367
Rec×Str	0.65	0.45	1.44	.150

Table 7: LME results for model attention data (Exp. 2), Pause window (-1300 to -600ms).

Fixed effect	β	SE	t	p
Intercept	-0.01	0.05	-0.28	.778
Str: Thm	-0.003	0.02	-0.18	.860
Str: Rec	0.02	0.02	1.02	.315
Structure	-0.03	0.04	-0.63	.534
Thm×Str	0.10	0.04	2.34	.025
Rec×Str	0.06	0.04	1.30	.204

Table 8: LME results for model attention data (Exp. 2), Connective window (-600 to 0 ms).

Fixed effect	β	SE	t	p
Intercept	0.04	0.05	0.88	.383
Str: Thm	0.05	0.02	2.28	.029
Str: Rec	-0.04	0.02	-1.71	.096
Structure	-0.17	0.06	-2.93	.005
Thm×Str	0.23	0.05	4.55	$< .001$
Rec×Str	0.25	0.07	3.56	.001

difference between two smooths; time regions where the 95% confidence interval excludes zero are reported as significant.

Table 9 summarizes all GAMM analyses, reporting the comparison, data subset, significant time window(s) from `plot_diff()`, and whether the parametric coefficient for the condition difference reached significance.

Interpreting GAMM results. In GAMM analyses, the parametric coefficient (β) tests the *overall* (intercept) difference between conditions across the entire time window, while the smooth terms capture *how* the effect changes over time. A non-significant parametric term combined with a significant smooth difference (identified via `plot_diff()`) indicates that the two conditions diverge during a specific temporal sub-window but do not differ on average across the full analysis period. This is expected in predictive processing data, where effects typically emerge late in the time window. The significant time windows from `plot_diff()` are the primary inferential results for the time-course analysis.

Table 9: Summary of GAMM time-course analyses (Experiment 2). All models were fit over the prediction window (-1300 to 0 ms). "Sig. window" reports the estimated time interval in which the pointwise difference between smooths was reliable (95% CI excluding zero). "n.s." = difference not significant at any time point.

System	Comparison	Data subset	Sig. window (ms)	β	p
Structure effects (PO vs. DO)					
Human	PO vs. DO	All Stress	-1300 to 0	-0.21	$< .001$
Human	PO vs. DO	Neutral only	-854 to 0	-0.26	.002
Human	PO vs. DO	Theme only	-1300 to -552	-0.20	.011
Human	PO vs. DO	Recip. only	-1300 to -893	-0.16	.033
Model	PO vs. DO	All Stress	-13 to 0	-0.02	.749
Model	PO vs. DO	Neutral only	-194 to 0	0.08	.253
Model	PO vs. DO	Theme only	n.s.	-0.08	.326
Model	PO vs. DO	Recip. only	-556 to -414	-0.07	.374
Stress effects (collapsed across Structure)					
Human	Rec. vs. Neut.	All Str.	-446 to 0	-0.13	.093
Human	Thm. vs. Neut.	All Str.	-39 to 0	0.03	.623
Model	Rec. vs. Neut.	All Str.	n.s.	-0.01	.908
Model	Thm. vs. Neut.	All Str.	n.s.	0.01	.892
Stress effects within Structure (cue diagnosticity)					
Human	Rec. vs. Neut.	DO only	-643 to 0	-0.17	.024
Human	Thm. vs. Neut.	DO only	n.s.	0.0003	.997
Human	Rec. vs. Neut.	PO only	n.s.	-0.08	.348
Human	Thm. vs. Neut.	PO only	-249 to 0	0.06	.503
Model	Rec. vs. Neut.	DO only	-220 to 0	0.08	.259
Model	Thm. vs. Neut.	DO only	n.s.	0.07	.326
Model	Rec. vs. Neut.	PO only	n.s.	-0.07	.365
Model	Thm. vs. Neut.	PO only	-181 to 0	-0.09	.261