Modeling Bottom-up Information Quality during Language Processing

Cui Ding
University of Zürich
cui.ding@uzh.ch

Yanning Yin
University of Basel
yanning.yin@unibas.ch

Lena A. Jäger
University of Zürich
lenaann.jaeger@uzh.ch

Ethan Gotlieb Wilcox
Georgetown University
ethan.wilcox@georgetown.edu

Abstract

Contemporary theories model language processing as integrating both top-down expectations and bottom-up inputs. One major prediction of such models is that the quality of the bottom-up inputs modulates ease of processing—noisy inputs should lead to difficult and effortful comprehension. We test this prediction in the domain of reading. First, we propose an information-theoretic operationalization for the "quality" of bottom-up information as the mutual information (MI) between visual information and word identity. We formalize this prediction in a mathematical model of reading as a Bayesian update. Second, we test our operationalization by comparing participants' reading times in conditions where words' information quality has been reduced, either by occluding their top or bottom half, with full words. We collect data in English and Chinese. We then use multimodal language models to estimate the mutual information between visual inputs and words. We use these data to estimate the specific effect of reduced information quality on reading times. Finally, we compare how information is distributed across visual forms. In English and Chinese, the upper half contains more information about word identity than the lower half. However, the asymmetry is more pronounced in English, a pattern which is reflected in the reading times.

https://github.com/DiLi-Lab/Bottom-Up-Information.git

1 Introduction

During reading, individuals actively expend cognitive effort to extract information. Many contemporary theories of language comprehension in general, and reading in particular, model this process as a rational integration of bottom-up and top-down information (Legge et al., 1997; Norris, 2006; Bicknell and Levy, 2010; Gibson et al., 2013; Gauthier and Levy, 2023). Bottom-up information refers to the perceptual input (e.g., visual forms of words), while top-down information includes the prior beliefs and expectations about what messages or word-forms are likely to be encountered, and is guided by the reader's linguistic and contextual knowledge. A central prediction of such models is that the ease of reading should be influenced by the quality of the bottom-up information. In the modality of visual reading, visual signals that effectively convey information about the intended message are expected to facilitate fast and effortless comprehension (Balota, 1994). Conversely, degraded visual signals—caused by factors such as lighting, occlusion, or visual interference—are likely to increase processing effort and raise the likelihood of errorful reading.

This prediction fits well within noisy channel models of reading. In a noisy channel model (Shannon, 1948), a message is encoded and sent over a channel, where it is potentially corrupted. A receiver, at the other end of the channel, must decode the most probable intended message given the received inputs. Previous work has looked at the role of noise during reading, demonstrating how noise over uncertain inputs can lead to non-veridical interpretations (Levy, 2008b; Gibson et al., 2013).

While intuitive, to the best of our knowledge, the impact of noisy inputs on reading effort has not been quantified within a formal computational model of reading. That is, although many theories of reading assume that poorer sensory input leads to more effortful processing, and classic experimental work has shown that reduced visual signals increase processing difficulty and interact with other lexical properties (Rumelhart and Siple, 1974), they have not derived or tested this relationship quantitatively. In this paper, we aim to fill this gap by providing an information-theoretically grounded, quantitative account of how bottom-up input quality affects processing effort. Our central proposal is that input quality can be formalized as the mutual information (MI) between (visual) input

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 11721–11733
November 4-9, 2025

and word identity. From an information-theoretic perspective, a signal is informative to the extent that it reduces uncertainty about a target variable—in this case, the identity of a word. We assume that greater processing effort manifests in longer reading times, and therefore predict that reductions in mutual information should lead to systematic slowdowns in reading.

This paper makes three contributions: First, we instantiate the above operationalization of visual input quality in reading under a formal model of reading as a Bayesian update. Second, we provide a quantitative estimate of the cost of reduced input quality on processing effort. To do so, we use multimodal language models to estimate mutual information over a dataset of partially masked word images. We then collect human reading times on the same stimuli, using the MoTR paradigm (Wilcox et al., 2024), which simulates eye-tracking, and can be used to collect data over the web. We use these data to estimate the relationship as a specific slowdown in terms of nats of information gain (the pointwise variant of mutual information) per millisecond of processing time. Our data suggest that the cost of reduced information is not linear—small losses in informational quality can lead to disproportionately large increases in reading time, particularly in the upper regions of a signal's informational range.

Our third contribution is to compare how information is distributed across visual forms of words in two typologically distinct languages. To that end, we collect data in both English and Chinese, representing alphabetic and logographic scripts, respectively. We find that, in both languages, the upper half of a word contains more information about word identity than the lower half. However, the asymmetry is more pronounced in English than in Chinese, a pattern that is reflected in the reading times.

2 Formal Model

2.1 Reading as Bayesian Update

Following an extensive prior literature (Norris, 2006; Bicknell and Levy, 2010; Gauthier and Levy, 2023), we model word recognition as a Bayesian update process. Readers incrementally process a word w drawn from a vocabulary $\mathcal{W}$, where $w \in \mathcal{W}$ denotes a realization of a random variable W taking values in $\mathcal{W}$. We refer to a word at a particular timestep, t, as w_t and the corresponding random variable at this timestep as W_t. We assume that readers intake individual samples of input e, where $e \in \mathbb{R}$ denotes a realization of a random variable E ranging over the samples[1]. Input samples could be either a patch of visual input for visual reading or a haptic percept in the case of braille. Following previous work (e.g., Bicknell and Levy, 2010), we model the process of reading as one of sequential word identification given input e and a previous context of words $\mathbf{w}_{<t}$. In such models, readers are assumed to rationally integrate their prior expectations about a word, $P(w_t \mid \mathbf{w}_{<t})$, with the likelihood of the observed input e, $P(e \mid w_t, \mathbf{w}_{<t})$. Instead of a single sample, we assume that readers integrate evidence over k samples, $\mathbf{e}_{1:k}$. The rational update process we use to model reading is therefore:

$$P(w_t \mid \mathbf{e}_{1:k}, \mathbf{w}_{<t}) \propto \quad (1)$$
$$P(w_t \mid \mathbf{w}_{<t}) \times \prod_{i=1}^{k} P(e_i \mid w_t, \mathbf{w}_{<t})$$

This tells us how readers update beliefs about a word given inputs and priors. But reading is a dynamic process. How do readers choose when to move on to the next word? Previous work models this by proposing that readers draw samples until the uncertainty about the current word reaches a threshold, ϕ, at which point they move on (e.g., Li and Futrell, 2024). We quantify uncertainty as the entropy of the posterior distribution. That is, sampling continues until:

$$H(P(w_t \mid \mathbf{e}_{1:k}, \mathbf{w}_{<t})) \leq \phi \quad (2)$$

However, given a particular actual input w^* we cannot be certain how many samples a reader draws or what information each sample contains. To account for this uncertainty, we therefore make the prediction that readers will move on when the *expected* entropy falls below this threshold, where the expectation is taken over uncertain inputs:

$$\mathbb{E}_{\mathbf{E}_{1:k}}[H(W_t \mid \mathbf{E}_{1:k}, \mathbf{w}_{<t})] \leq \phi \quad (3)$$

Although we assume that reading does take place given a context, for the rest of this section, we will drop the word-context term, $\mathbf{w}_{<t}$. We note that it would be easy to add this term back into the subsequent equations as a conditioning variable without changing the overall model.

[1]For simplicity, we model inputs as continuous and univariate. However, we acknowledge that inputs may be more aptly modeled as multivariate and see this as an easy extension of the formal presentation given here.

2.2 Quality of Bottom-Up Evidence

We model the quality of the inputs as the mutual information between the inputs and the word identity, i.e., as $I(W; E)$. High-quality inputs do a better job of reducing uncertainty over words. For a given word-identification step, we can write the mutual information between a word and the total number of samples drawn as $I(W; \mathbf{E}_{1:k})$. Using the chain rule of mutual information (Cover, 1999) and assuming that there is *conditional independence* between samples, given W, we can derive the following inequality:[2]

$$I(W; \mathbf{E}_{1:k}) = \sum_{i=1}^{k} I(W; E_i \mid \mathbf{E}_{1:i-1}) \quad (4a)$$

$$\leq \sum_{i=1}^{k} I(W; E_i) \quad (4b)$$

assuming cond. independence

$$\leq k \times I(W; E); \quad (4c)$$

How is the mutual information between inputs and words related to the reading process, as described above? We assume that taking samples and processing these samples takes cognitive effort. Following previous work, we also assume a link between effort and time (Levy, 2008a; Hale, 2001). Therefore, the more samples, k, a reader needs to take in order to reduce uncertainty, the longer it will take them to read a given word.

We can now link the quality of inputs to our reading process through the definition of mutual information:

$$I(W; \mathbf{E}_{1:k}) = H(W) - H(W \mid \mathbf{E}_{1:k}) \quad (5)$$

Plugging in the inequality from 4c, and the definition of conditional entropy,[3] we rearrange the terms:

$$\mathbb{E}_{\mathbf{E}_{1:k}}[H(W \mid \mathbf{E}_{1:k})] \geq H(W) - k \times I(W; E) \quad (6)$$

That is, the expected entropy of the posterior distribution, given uncertain inputs, is greater than the entropy over words minus the number of samples taken times the mutual information between the samples and the words.

For our model of reading, we are interested in when the entropy of the posterior distribution is approximately ϕ. In particular, we are interested in how many samples must be drawn to reach this threshold, as this determines the effort (and therefore the time) required to reduce uncertainty enough to move on to the subsequent word. Substituting in our threshold parameter in and rearranging the terms, we have:

$$k \geq \frac{H(W) - \phi}{I(W; E)} \quad (7)$$

The minimum number of samples required to reach the threshold grows with the entropy of the distribution over W. Likewise, it decreases with the mutual information between W and E. Because we assume a link among the number of samples, effort and time, this leads us to the following two predictions:

Prediction 1 ***Top-Down Processing & Entropy:*** *As the entropy of a word-position W increases, average reading time increases.*

Prediction 2 ***Bottom-up Processing & Mutual Information:*** *As the mutual information between words W and their visual representations E decreases, average reading time increases.*

In fact, Prediction 1 has already been investigated by Pimentel et al. (2023), whose results confirm our prediction. Pimentel et al. refer to the entropy over the next word, given a set of previous words $H(W_t \mid \mathbf{w}_{<t})$ as a word's *contextual entropy*. They find that as word-level contextual entropy increases, so too does reading time. For the rest of this paper, therefore, we are interested in testing Prediction 2, namely whether the quality of bottom-up evidence, modeled as mutual information between words and visual information, affects word-by-word reading times.

3 Methods

3.1 Materials

We use a portion of the advanced OneStopQA dataset (Berzak et al., 2020). This dataset contains Guardian news articles, along with carefully constructed reading comprehension questions, which are linked to individual spans in the text. For our study, we selected three articles: "101-Year-Old Bottle Message", "Inky the Octopus Escapes from Aquarium", and "Japan Calls Time on Long Hours Work Culture". A team member with experience in English-Chinese translation hand-translated these texts and their questions into Mandarin. This small

[2] For more discussion of these assumptions, see Section A.

[3] That is: $H(X \mid Y) = \mathbb{E}_Y[H(X \mid Y)]$.

Figure 1: Example showing a screen from a MoTR trial with our three different reading conditions.

translated corpus, which we term the **Chinese OneStopQA**, is released along with the publication of this article (see code repository).

The English subset contains 1,793 words (mean word length $= 4.6$, $SD = 2.53$), while the Chinese subset contains 3,182 characters. In terms of experimental presentation, one Chinese character occupies roughly 1.46 times the pixel space of an English letter, making an average English word about 3.2 times longer than a Chinese character. The average Zipf frequency is slightly higher in English ($M = 5.77$, $SD = 1.45$) than in Chinese ($M = 4.84$, $SD = 1.90$), largely due to the low frequency of transliterated Western named entities in the Chinese translations.

Creating Noisy Words To create noised reading conditions, we occluded (i.e., masked with white) either the upper or lower half of every word in the dataset. There are potentially many ways to add noise to the texts. Other options would be to occlude the first half or the second half of words, as well as Gaussian noise. Previous work has shown that the beginnings of words tend to carry more disambiguatory information than their endings. For example, Pimentel et al. (2021) demonstrates this cross-linguistically using information-theoretic measures, while Alhama et al. (2019) presents a perceptually constrained connectionist model explaining fixation biases toward word onsets. These findings are consistent with psycholinguistic evidence of the optimal viewing position effect in visual word recognition(Brysbaert and Nazir, 2005).

However, these studies focus on the linear distribution of information across letter positions, which applies naturally to alphabetic scripts such as English but not to logographic systems like Chinese, where characters are two-dimensional and not arranged linearly. We were also concerned that completely removing some letters or characters would make reading too difficult or frustrating for participants, and that the removal of letters or characters demands very careful handling to avoid confounds (Rayner, 1998; Rayner et al., 2006). Masking the upper or lower half retains some information about each character, which presents a paradigm that is relatively readable for participants, especially in the degraded conditions. In addition, unlike simply adding Gaussian noise, upper and lower half occlusion allows us to investigate *where*, in vertical space, information is localized in English and Chinese orthographic systems. Our strategy leads to two additional research questions:

Sub Research Question 1 *Is information split up differentially between the upper and lower halves of orthographic words?*

Sub Research Question 2 *Does the relative informativeness of upper vs. lower halves differ across languages?*

3.2 Data Collection

Mouse Tracking for Reading (MoTR) To test our main predictions, we need a way of measuring (average) human reading times in our different conditions. To do so, we use Mouse Tracking for Reading (MoTR; Wilcox et al., 2024). In a MoTR trial, a blurred text is presented on a screen. A small region around the tip of a user's mouse brings the text into focus. Participants move the mouse to incrementally *reveal* and read the text, while their mouse location is recorded and used as a proxy for gaze location. The time-stamped x/y coordinates are then turned into incremental word-by-word reading times, similar to word-level reading times in an eye-tracking-while-reading experiment. As in eye-tracking, there are several ways to compute reading times. For our main analysis, we use the **first-pass reveal time (FPRT)**, defined as the total amount of time a participant spends revealing a word during their first pass reading. Conveniently, the same acronym (FPRT) is used in eye-tracking for **first-pass reading time**, but we use "reveal" to emphasize that it is computed from mouse movements rather than eye fixations.

Wilcox et al. (2024) show that MoTR reading measures are strongly correlated with eye-tracking and self-paced-reading measures. MoTR has been used to collect data in English and Russian (Oğuz et al., 2025), but not in Chinese.

Participants We recruited 54 English and 57 Chinese speakers on Prolific, requiring a minimum approval rate of 98% and the corresponding language to be their first and native language. Participants were compensated 3.75 GBP for a median reading time of 25 minutes.

Procedure Each participant read the article paragraphs presented screen by screen, with each screen randomly assigned to one of three conditions: upper-half occluded (i.e., lower-half visible), lower-half occluded (i.e., upper-half visible), or unoccluded (see Figure 1). In addition to reading texts and answering comprehension questions, we ask participants to rate the ease of reading after finishing all the trials using a multiple-choice question: "Which do you find easier to read: text showing only the top half or only the bottom half?" The options are "Top half only," "Bottom half only," and "About the same." The actual experiments are available online for Chinese[4] and English[5].

3.3 Mutual Information Estimation

In Section 2, our model concerns the mutual information between words, W, and (visual) evidence sampled by the reader, $\mathbf{E}$. However, we do not have direct access to this evidence. Instead, as a proxy for our visual evidence, we estimate the mutual information between words W and their orthographic representations $\mathbf{o} \in \mathbb{R}^d$, where $\mathbf{o}$ is a realization of a random variable $\mathbf{O}$ that ranges over representations of different words. Following Pimentel et al. (2020), we decompose the mutual information as

$$I(W; \mathbf{O}) = H(W) - H(W \mid \mathbf{O}) \tag{8a}$$
$$\approx H_\theta(W) - H_\theta(W \mid \mathbf{O}) \tag{8b}$$

where θ denotes the parameters of the models employed for entropy estimation. We estimate each of these two terms separately.

We estimate **unconditional entropy** $H_\theta(W)$ with a maximum likelihood estimation of the unigram distribution of Chinese characters and English words. We take the $9,933$ unique Chinese characters included in the modern Chinese character database[6], and the $60,384$ English words in the SUBTLEXus database (Brysbaert and New, 2024), and look up their frequencies using the Python library *wordfreq* (Speer, 2022) that supports both languages and aggregates data from multiple domains, including subtitles, Wikipedia, news, fiction, and web content. Normalizing the frequencies, we obtain the empirical distribution $p_\theta(w)$ and from it we can directly compute the entropy $H_\theta(W)$. The empirical entropies are 5.59 and 7.12 nats for Chinese characters and English words.

We estimate the **conditional entropy** $H_\theta(W \mid \mathbf{O})$ in two stages. First, we compute the word entropy conditioned on a specific orthographic representation, $H_\theta(W \mid \mathbf{O} = \mathbf{o})$. We refer to this as **pointwise conditional entropy**. We compute this value by taking the expectation of the information content, or **surprisal** of all words given the orthographic representation $\iota_\theta(w \mid \mathbf{o})$, where $\iota_\theta(\cdot) = -\log p_\theta(\cdot)$. Given a model with parameters θ that can produce our probability distribution of interest, that is, $p_\theta(w \mid \mathbf{o})$, the pointwise conditional entropy is calculated as:

$$H_\theta(W \mid \mathbf{o}) \approx \sum_{w \in \mathcal{W}} p_\theta(w \mid \mathbf{o}) \iota_\theta(w \mid \mathbf{o}) \tag{9}$$

We then estimate conditional entropy as the expectation of the pointwise conditional entropy with respect to $\mathbf{O}$, following the identity $H(W \mid \mathbf{O}) = \mathbb{E}_{\mathbf{O}}[H(W \mid \mathbf{O} = \mathbf{o})]$. We take the expectation over a set of held-out test samples:

$$H_\theta(W \mid \mathbf{O}) \approx \frac{1}{N} \sum_{n=1}^{N} H_\theta(W \mid \mathbf{o}^n) \tag{10}$$

where $\mathbf{o}^n$ is the n^{th} orthographic representation in the test set.

We note that using these methods, we can estimate not only the mutual information $I(W; \mathbf{O})$, but also its pointwise variant, also called the **information gain (IG)**, for a particular orthographic representation, where $IG(W; \mathbf{o}) = H(W) - H(W \mid \mathbf{o})$. While our formal prediction is made in terms of mutual information, in Section 4.3, we use IG to investigate the relationship between information contained in individual visual inputs and their respective reading times.

In recent work, similar methods have been used to study the relationship between words (as represented by text) and prosody, or the melody of speech (Wolf et al., 2023; Regev et al., 2025; Wilcox et al., 2025). However, these previous works learn distributions over real-valued variables that represent pitch. We wish to learn distributions over discrete w-valued variables $p_\theta(w \mid \mathbf{o})$. To obtain this distribution, we use multimodal language

[4] https://cuierd.github.io/Re-Veil/multilingual_motr/zh/
[5] https://cuierd.github.io/Re-Veil/multilingual_motr/en/
[6] https://lingua.mtsu.edu/chinese-computing/

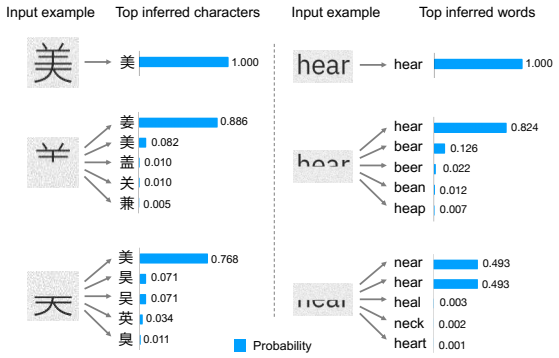

Figure 2: Results of fine-tuned Qwen2.5 model for the Chinese character 美 ("beautiful") and the English word *hear*. The preference for *hear* over *heal* in upper half occlusion likely reflects pre-training frequency bias, which we control for by training TransOCR from scratch.

models, which we fine-tune to produce conditionalized distributions over words, given visual inputs. We do so with the following methods:

Data Generation We adapt the Python library *TRDG*[7] to generate images of Chinese characters and English words from text as their orthographic representations, applying upper-, lower-half occlusion to create our different experimental conditions. For each character or word under each condition, we generated six images with different fonts and font weights[8] to enhance visual variability, and added a small amount of Gaussian noise to the image backgrounds (Li et al., 2025). We generate $16,800$ Chinese character images and $44,800$ English word images under each of the three occlusion conditions as training data. For test data, we generate images of all Chinese characters in the Chinese OneStopQA dataset and all English words in the selected OneStopQA subset, again under each occlusion condition.

A Bayesian Baseline Model As a simple reference, we implement a Bayesian baseline to estimate the pointwise conditional entropy $H(W \mid \mathbf{O} = \mathbf{o})$. In this model, lexical frequencies provide priors $p(w)$, and structural similarity (SSIM; Wang et al., 2004), computed with *scikit-image* (Van der Walt et al., 2014), serves as a likelihood $p(\mathbf{o} \mid w)$. Posterior probabilities $p(w \mid \mathbf{o})$ are obtained by normalizing the product of priors and likelihoods across all candidate images, with the denominator corresponding to the marginal likelihood $p(\mathbf{o})$. We then compute $H(W \mid \mathbf{O} = \mathbf{o})$ from these posteriors for each input image and average across the dataset to obtain $H(W \mid \mathbf{O})$. While straightforward, this baseline has some limitations. First, its computational cost scales quadratically with dataset size ($O(N^2)$). Second, SSIM captures only low-level pixel similarity, often producing clustered scores for orthographically distinct characters or words. Third, its estimates are largely driven by lexical frequency. For these reasons, we turn to multimodal models based on artificial neural networks (ANNs) for more reliable and scalable estimation.

ANN-based Predictive Multimodal Models We use three different multimodal model settings: First, we evaluate the pre-trained Qwen2.5-VL-7B-Instruct[9] in a zero-shot setting. Qwen2.5-VL-7B is an open-source vision-language model developed by Alibaba, designed for high-accuracy multimodal analysis with enhanced visual understanding and text-image alignment (Wang et al., 2024; Bai et al., 2025). As upper- and lower-half occluded words are likely out-of-distribution with respect to the model's training data, we do not expect the mutual information estimate to be tight in this setting. For a better estimate, we then fine-tune Qwen2.5-VL-7B on our task-specific data to improve its performance. To complement the estimate from the pre-trained model, we also train a separate transformer-based OCR model (TransOCR; Yu et al., 2023), from scratch, to perform the same prediction task. The model combines a ResNet encoder with a Transformer decoder for character recognition. Full training configurations and prompt designs for the Qwen and TransOCR models are provided in Section B and Section C, respectively.

4 Results

4.1 Human Reading Results

We show reading times in Figure 3(a). In both languages, reading full words resulted in the shortest average FPRT, as predicted. Interestingly, both languages follow a *Full* < *Upper* < *Lower* pattern, with lower-half visibility leading to the longest FPRT. To quantify these effects, we fit linear mixed-effects models with visibility condition as a predictor, using sliding contrasts to compare *Upper* vs. *Full* and

[7] https://github.com/Belval/TextRecognitionDataGenerator

[8] For Chinese, the fonts are XinYiJiXiangSong, FZHei-B01, FZKai-Z03, NotoSansSC-Regular, NotoSerifSC-SemiBold, and SourceHanSans. For English, they are Droid-Sans, Lato-Bold, NotoSans, PTSerif, Raleway, and Sansation.

[9] https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct

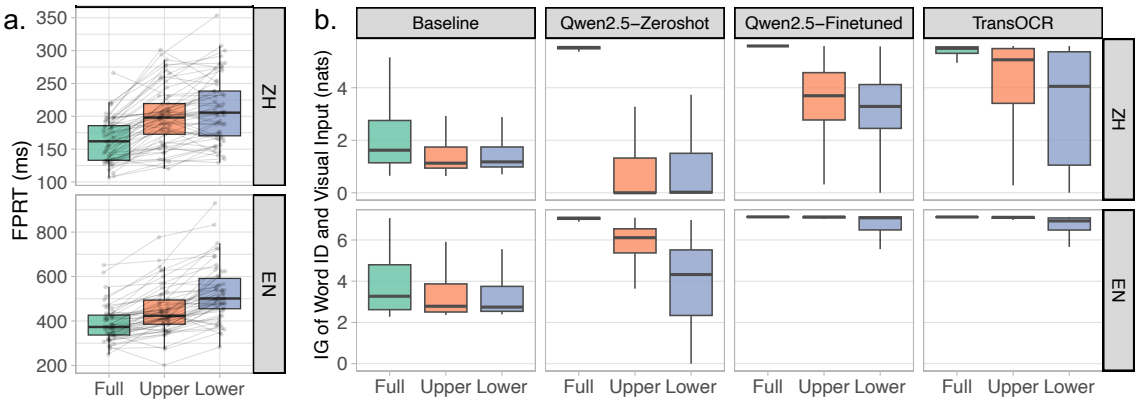

Figure 3: **(a)** Reading times (FPRT) measured under three visibility conditions. Boxes represent the interquartile range (middle 50%), center lines indicate the median, and whiskers show the overall data spread. Grey lines trace each participant's mean across conditions. EN: English; ZH: Simplified Chinese **(b)** Information gain (IG) between word identity and visual form under the three conditions, obtained with each of our estimation techniques.

Lower vs. *Upper*. Word length (EN only), lexical frequency, surprisal, and contextual entropy are included as control variables, with random intercepts for subjects and items. In Chinese, both contrasts are significant: $\beta = 37.79$ ms ($p < 2e{-}16$) and $\beta = 12.64$ ms ($p < 2e{-}16$). In English, the effects are larger: $\beta = 69.93$ ms ($p < 2e{-}16$) and $\beta = 90.09$ ms ($p < 2e{-}16$)[10].

These results can be interpreted as implying a visual asymmetry in both languages between ease of processing with respect to just upper and lower halves of words. The asymmetry is stronger in English, where the lower half leads to greater slowdowns. Participants' subjective ratings confirm this asymmetric pattern and further show that English lower halves are perceived as harder to read than Chinese ones (Appendix D). In addition, we summarize comprehension question performance in Appendix E. Accuracy declines with occlusion, but remains well above chance (25%), indicating that reading is effective.

4.2 Mutual Information Results

To give a visual sense of how our models perform, in Figure 2, we show sample images in the three experimental conditions, along with the predictions from the fine-tuned Qwen2.5 model. As a performance check, we also report the model accuracies (Acc) in Table 1. Baseline accuracies are not reported because they only compare each test image against other images in the test set, rather than predicting over the full vocabulary. This makes its accuracy values not directly comparable to the other models. Accuracies in the unoccluded Full condition are uniformly high (>97%). Under occlusion, zero-shot Qwen2.5 drops dramatically in Chinese (<6%) but is still strong in English (52–85%). Fine-tuning improves performance in both languages, while TransOCR achieves the most robust accuracy overall.

		Acc (%)			MI (nat)		
Language	Model	Full	Upper	Lower	Full	Upper	Lower
ZH	Baseline	-	-	-	4.85	4.42	4.41
	Qwen2.5-zs	97.9	5.2	3.2	5.42	0.27	0.32
	Qwen2.5-ft	99.0	48.6	44.3	5.57	3.62	3.27
	TransOCR	97.9	78.8	65.7	5.26	4.09	3.17
EN	Baseline	-	-	-	6.43	6.12	5.99
	Qwen2.5-zs	98.8	84.1	51.5	6.99	5.74	3.86
	Qwen2.5-ft	99.7	93.1	68.1	7.11	7.01	6.66
	TransOCR	98.8	95.7	65.8	7.07	7.00	6.68

Table 1: Model accuracy (%) and mutual information estimates, $I(W;\mathbf{O})$, for Chinese (ZH) and English (EN) with different models.

Our main analyses focus on mutual information, $I(W;\mathbf{O})$ in Table 1 and information gain, $IG(W;\mathbf{o})$ in Figure 3(b), across conditions and models. Figure 3(b) shows IG between word identity and visual input under three visibility conditions, estimated with the Bayesian baseline, Qwen2.5-VL-7B-Instruct (zero-shot and fine-tuned), and TransOCR. Table 1 reports MI as the IG averaged across all inputs. MI decreases systematically from *Full* to *Upper* to *Lower* in both languages.

For a statistical test of our observed trends, we fit linear mixed-effects models for each language–model pair, with visibility as the main predic-

[10]FPRT is calculated for Chinese *characters* and English *words*, which may explain the generally longer reading times in English.

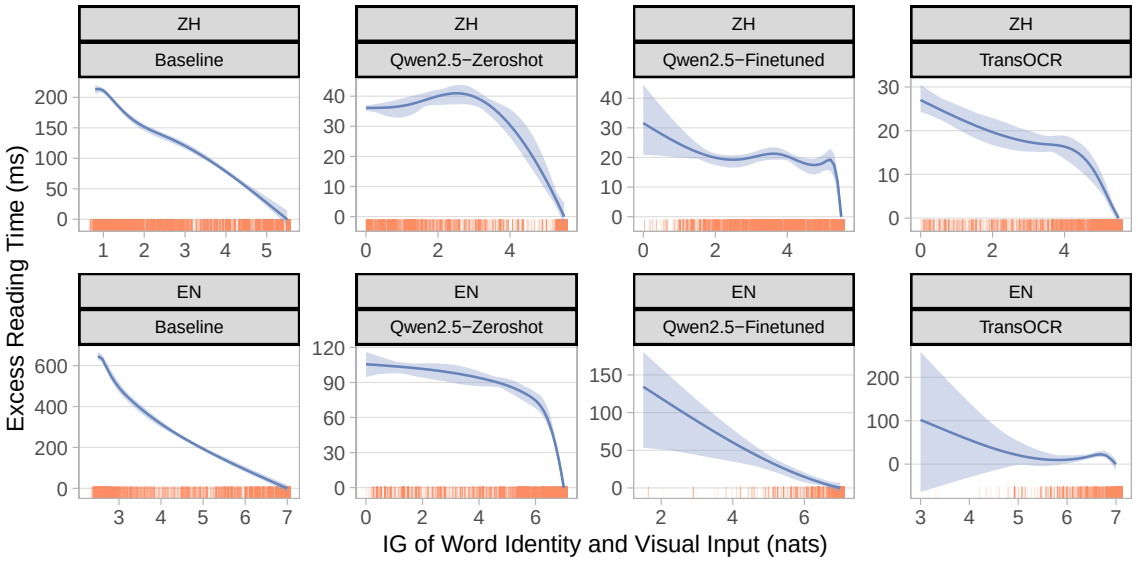

Figure 4: Relationship between informational quality of individual words (information gain; IG) and excess reading time. Solid blue lines are smoothed GAM fits; shaded regions show bootstrapped 95% confidence intervals. Red tick marks along the bottom (rug plots) indicate the distribution of IG data points. Reading times are aligned to end at zero at the highest MI end to emphasize the relative excess reading time when information quality decreases.

tor. As in the reading time analysis, we use the same sliding contrasts. Word length (EN only), lexical frequency, surprisal, and contextual entropy are included as controls, with random intercepts for items.[11] In Chinese, all models show significant IG reductions when only the upper half is visible (Baseline: $\beta = -.58$; Qwen2.5-Zeroshot: $\beta = -4.55$; Qwen2.5-Finetuned: $\beta = -1.85$; TransOCR: $\beta = -.99$ nats), and IG from fine-tuned models dropped further when visibility changed from Upper to Lower (Qwen2.5-Finetuned: $\beta = -.37$; TransOCR: $\beta = -1.01$ nats). In English, the zero-shot model showed the largest overall drop (Upper vs. Full: $\beta = -1.43$; Lower vs. Upper: $\beta = -2.09$ nats), while the baseline and fine-tuned models show smaller but consistent reductions (Baseline: $\beta = -.36, -.07$; Qwen2.5-Finetuned: $\beta = -.11, -.47$; TransOCR: $\beta = -.08, -.35$ nats). All effects were statistically significant at $p < .0001$. Panels (a) and (b) of Figure 3, taken together, reveal a clear pattern: as visual input degrades from *Full* to *Upper* to *Lower*, as measured by IG, reading times increase.

4.3 Word-Level Relationship

In this section, we test the relationship between reading time and informational quality at the *word level*. To do so, we fit linear mixed-effects models with reading time of an orthographic representation as the dependent variable and its IG as a fixed effect. We also included frequency, surprisal, contextual entropy, and (in EN) word length as additional fixed effects, as well as by-subject and by-item random intercepts.

We find a significant effect of IG on reading time across all models and measures, with a consistent negative effect: the higher the informational quality of the input, the faster it is read. In Chinese, all four IG estimates were significant predictors of FPRT: $\beta = -14.94$ ms (Baseline), $\beta = -7.53$ ms (Qwen2.5-Zeroshot), $\beta = -10.19$ ms (Qwen2.5-Finetuned), and $\beta = -4.97$ ms (TransOCR). In English, the effects were even larger: $\beta = -15.36$ ms (Baseline), $\beta = -23.67$ ms (Qwen2.5-Zeroshot), $\beta = -51.48$ ms (Qwen2.5-Finetuned), and $\beta = -66.42$ ms (TransOCR). All effects were statistically significant at $p < .0001$.

4.4 Nonlinear Relationship Between Information Quality and Reading Time

While our linear regression models show that informational quality affects reading time, it makes strong assumptions about the functional form of this relationship. In order to get a better sense of

[11] Empirically, IG shows little correlation with these controls ($-.03$ to $.05$ in Chinese; $-.22$ to $.15$ in English). An exception is the Bayesian baseline, where IG correlates strongly with frequency (>0.9) in both languages.

how these two variables are related, we visualize them together in Figure 4. We used generalized additive models (GAMs). GAMs are models that allow for non-linear relationships between predictor and response variables. We fit GAMs to predict reading times with smooth terms for IG, controlling for frequency, surprisal, contextual entropy, and (for English) word length. [12] We applied bootstrap smoothing over 20 resamples and computed confidence intervals for the estimated effects. We observe a consistent trend across both languages and all multimodal models: reading time remains relatively stable at lower IG estimates but decreases rapidly as IG increases in the upper end of its range. The Bayesian baseline does not show this pattern, as its IG values largely reflect lexical frequency.

5 Discussion

Turning back to our main prediction, we argue that our results provide converging evidence that visual quality, as measured by mutual information or information gain, impacts ease of processing. First, we find a consistent ordering, both in terms of reading times and mutual information, across our three experimental conditions. Second, we find a significant effect of the pointwise mutual information, or information gain, of individual words on reading times. While intuitive, the idea that bottom-up informational quality impacts ease of reading has not been quantified within a formal framework of reading. Our methods and experiments provide a specific estimate for the relationship between visual informational quality and reading times, which in English is between 25–66 ms/nat and in Chinese 5–10 ms/nat. However, these numbers should be taken only as rough estimates, as the exact functional form may not be linear.

Turning now to our two sub research questions outlined in section 3.1: Interestingly, we find that information is not distributed evenly between the top and bottom halves of words. Both English and Chinese place more information about word identity in the top half of their orthographic systems, a feature which we argue is reflected in the quicker reading times for our *Upper* condition. This asymmetry may connect to more general visual biases, such as the top-heavy bias in object recognition (Viola Macchi et al., 2004) or the upper visual field advantage (Previc, 1990), as well as the way writing systems are shaped by reading and writing direction.

Interestingly, Pimentel et al. (2021) find similar informational asymmetries between the beginnings and ends of words, using an even wider set of languages. Exploring whether their asymmetry in reading times and extending our results to more languages is an important direction for future research. Finally, we find some suggestive evidence that this asymmetry is stronger in English, reflected in the larger effect sizes for the *Upper* vs. *Lower* contrast in our reading data. Future work should investigate such differences in greater detail.

Limitations

There are several limitations with the present work. In our formal model, we made several assumptions: that visual samples of a given word E are drawn i.i.d. during reading, and that visual inputs are conditionally independent from each other given W. While these assumptions are strong, they are compatible with a "simple but fast" approach to reading. We discuss them in more detail in Section A. Moreover, our model assumes that the entropy threshold ϕ (Eq. 2) is always eventually reached, whereas in reality, readers may sometimes move on without reaching this threshold, especially under poor input quality. Besides, our model treats reading as a linear process, abstracting away from regressions, skips, and parafoveal preview. By focusing on FPRT as an index of early lexical processing, we capture word-level difficulty while necessarily ignoring the full dynamics of scanpaths. These simplifications keep the model straightforward but also mark directions for future work.

Another limitation concerns our approach to estimating mutual information between word identity and orthographic representation in Chinese. We used characters, rather than lexical words, as the unit of analysis. This choice was motivated by two considerations: first, the average word length in our OneStopQA Chinese dataset is approximately 1.4 characters; second, Chinese characters, unlike English letters, carry substantial visual and semantic complexity. As such, characters may serve as a more suitable unit for modeling bottom-up visual processing in Chinese, analogous to words in English. Nonetheless, using lexical words might produce slightly different estimates of mutual information. Future work could examine whether

[12] For example, for EN data, the GAM was specified as: `FPRT ~ s(IG, bs='cr', k=6) + te(freq, len, bs='cr') + te(surp, len, bs='cr') + te(entropy, len, bs='cr')`.

similar patterns hold when words are used instead of characters.

Another limitation of the present work is that we did not extensively explore how top-down (contextual) processing can be integrated into the investigation of bottom-up processing. While much current research in psycholinguistics emphasizes the role of top-down expectations, our study is intended as a contribution specifically to the modeling of bottom-up processing. As described in Section 2.1, our formal Bayesian model is capable of incorporating contextual terms in a straightforward manner. However, for the scope of this analysis, we opted to exclude this component, leaving its exploration for future work.

Moreover, our study is limited to English and Chinese, as they are chosen to represent alphabetic and logographic systems, given the available corpora and our expertise. While this provides meaningful cross-linguistic insight, extending to more languages is important to generalize the conclusions of this work. In particular, the omission of Semitic languages is a notable gap, as their nonconcatenative morphology and distinct reading behaviors provide a critical test of generalizability (see e.g., Alhama et al., 2019 and Lerner et al., 2014). Including such languages remains an important direction for future work.

Acknowledgments

We thank the reviewers from ACL Rolling Review for their valuable feedback. We are also grateful to Kirill Semenov, Junlin Li, and our colleagues in the Computational Linguistics department at the University of Zürich for their helpful discussions and feedback. CD was supported by the MeRID grant (212276, PI: LJ).

Ethics Statement

We do not foresee any specific ethical concerns deriving from this work. Experimental participants engaged in standard reading tasks, and were compensated for their time. We used AI models for analysis. As with all AI systems, it is important to acknowledge a general risk of bias or misuse.

References

Raquel G Alhama, Noam Siegelman, Ram Frost, and Blair C Armstrong. 2019. The role of information in visual word recognition: A perceptually-constrained connectionist account. In *the 41st Annual Meeting of the Cognitive Science Society (CogSci 2019)*, pages 83–89. Cognitive Science Society.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

David A. Balota. 1994. Visual word recognition. In *Handbook of Psycholinguistics*, pages 303–358. Academic Press, San Diego, CA.

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. STARC: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5726–5735, Online. Association for Computational Linguistics.

Klinton Bicknell and Roger Levy. 2010. A rational model of eye movement control in reading. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1168–1178.

Marc Brysbaert and Tatjana Nazir. 2005. Visual constraints in written word recognition: evidence from the optimal viewing-position effect. *Journal of Research in Reading*, 28(3):216–228.

Marc Brysbaert and Boris New. 2024. The subtlex word frequency norms. In *Reference Module in Social Sciences*. Elsevier.

Thomas M. Cover. 1999. *Elements of Information Theory*. John Wiley & Sons, New York, NY, USA.

Jon Gauthier and Roger Levy. 2023. The neural dynamics of word recognition and integration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 980–995, Singapore. Association for Computational Linguistics.

Edward Gibson, Leon Bergen, and Steven T. Piantadosi. 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Gordon E Legge, Timothy S Klitz, and Bosco S Tjan. 1997. Mr. Chips: An ideal-observer model of reading. *Psychological Review*, 104(3):524.

Itamar Lerner, Blair C Armstrong, and Ram Frost. 2014. What can we learn from learning models about sensitivity to letter-order in visual word recognition? *Journal of Memory and Language*, 77:40–58.

Roger Levy. 2008a. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Roger Levy. 2008b. A noisy-channel model of human sentence comprehension under uncertain input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 234–243, Honolulu, Hawaii. Association for Computational Linguistics.

Jiaxuan Li and Richard Futrell. 2024. An information-theoretic model of shallow and deep language comprehension. In *Proceedings of the 46th Annual Meeting of the Cognitive Science Society (CogSci 2024)*, page to appear. Cognitive Science Society.

Zhecheng Li, Guoxian Song, Yujun Cai, Zhen Xiong, Junsong Yuan, and Yiwei Wang. 2025. Texture or semantics? Vision-language models get lost in font recognition. *arXiv preprint arXiv:2503.23768*.

Dennis Norris. 2006. The Bayesian reader: Explaining word recognition as an optimal bayesian decision process. *Psychological Review*, 113(2):327.

Metehan Oğuz, Cui Ding, Ethan Gotlieb Wilcox, and Zuzanna Fuchs. 2025. Using MoTR to probe gender agreement in Russian. *OSF*.

Tiago Pimentel, Ryan Cotterell, and Brian Roark. 2021. Disambiguatory signals are stronger in word-initial positions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 31–41, Online. Association for Computational Linguistics.

Tiago Pimentel, Clara Meister, Ethan G. Wilcox, Roger P. Levy, and Ryan Cotterell. 2023. On the effect of anticipation on reading times. *Transactions of the Association for Computational Linguistics*, 11:1624–1642.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Fred H Previc. 1990. Functional specialization in the lower and upper visual fields in humans: Its ecological origins and neurophysiological implications. *Behavioral and Brain Sciences*, 13(3):519–542.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372.

Keith Rayner, Sarah J White, Rebecca L Johnson, and Simon P Liversedge. 2006. Raeding wrods with jubmled lettres: There is a cost. *Psychological Science*, 17(3):192–193.

Tamar I Regev, Chiebuka Ohams, Shaylee Xie, Lukas Wolf, Evelina Fedorenko, Alex Warstadt, Ethan Wilcox, and Tiago Pimentel. 2025. The time scale of redundancy between prosody and linguistic context. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30476–30488, Vienna, Austria. Association for Computational Linguistics.

David E Rumelhart and Patricia Siple. 1974. Process of recognizing tachistoscopically presented words. *Psychological Review*, 81(2):99–118.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. 2014. scikit-image: Image processing in Python. *PeerJ*, 2:e453.

Cassia Viola Macchi, Chiara Turati, and Francesca Simion. 2004. Can a nonspecific bias toward top-heavy patterns explain newborns' face preference? *Psychological Science*, 15(6):379–383.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

Ethan Wilcox, Cui Ding, Giovanni Acampa, Tiago Pimentel, Alex Warstadt, and Tamar I Regev. 2025. Using information theory to characterize prosodic typology: The case of tone, pitch-accent and stress-accent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24439–24451, Vienna, Austria. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Cui Ding, Mrinmaya Sachan, and Lena Ann Jäger. 2024. Mouse Tracking for Reading (MoTR): A new naturalistic incremental processing measurement tool. *Journal of Memory and Language*, 138:104534.

Lukas Wolf, Tiago Pimentel, Evelina Fedorenko, Ryan Cotterell, Alex Warstadt, Ethan Wilcox, and Tamar Regev. 2023. Quantifying the redundancy between prosody and text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9765–9784, Singapore. Association for Computational Linguistics.

Haiyang Yu, Xiaocong Wang, Ke Niu, Bin Li, and Xiangyang Xue. 2023. Scene text segmentation with text-focused transformers. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 2898–2907, New York, NY, USA. Association for Computing Machinery.

A Assumptions of Formal Model

In this appendix, we discuss the assumption(s) of our formal model, namely that our samples $\mathbf{E}$ are conditionally independent of each other, given W. First, we walk through the step from 4a to 4b. We have by the definition of mutual information:

$$I(W; \mathbf{E}_{1:i}) \tag{11}$$

$$= \sum_{i=1}^{k} I(W; E_i \mid \mathbf{E}_{1:i-1}) \tag{12}$$

$$= \sum_{i=1}^{k} H(E_i \mid \mathbf{E}_{1:i-1}) - H(E_i \mid W, \mathbf{E}_{1:i-1}) \tag{13}$$

For the first term in this sum we can use the inequality $H(E \mid \mathbf{E}_{1:i-1}) \leq H(E)$. This is because adding information can only decrease entropy. Furthermore, assuming conditional independence between the samples, given W, we have that $H(E \mid W, \mathbf{E}_{1:i-1}) = H(E \mid W)$. Therefore, we can rewrite as:

$$\leq \sum_{i=1}^{k} H(E_i) - H(E_i \mid W) \tag{14}$$

$$\leq \sum_{i=1}^{k} I(E_i; W) \tag{15}$$

which, given the symmetry of mutual information, is what we have in 4b.[13]

Regarding the assumption of conditional independence: This means that if the reader knows the word's identity, then previous samples do not necessarily predict what will be sampled next. We believe that this assumption is somewhat strong. However, it may be compatible with the view that readers adopt a simple, but fast, sampling strategy, in which prior evidence or even incremental knowledge about the word's identity from samples does not determine future sampling behavior. Given that reading happens at a very quick timescale, where word identification takes potentially only tens of milliseconds, such a "simple but fast" approach is not unreasonable.

[13] Thank you to Tiago Pimentel for pointing out an earlier issue with our derivation, which has been corrected in this version of the paper. Any remaining mistakes are, of course, the fault of the authors.

B Qwen2.5-VL-7B-Instruct Fine-Tuning Details

We fine-tune Qwen2.5-VL-7B-Instruct using QLoRA with 4-bit quantization and LoRA adapters applied to attention projection layers with rank 8, $\alpha = 16$, and dropout 0.05. The model is trained for up to 100 epochs. Early stopping is applied based on validation loss. The training will terminate if no improvement for three consecutive epochs. AdamW (learning rate 2e-4), batch size 4, gradient accumulation of 8, and gradient clipping of 1.0. Training data consists of system and user prompts with bottom-half character images; the model predicts a single Chinese character. We format inputs using Qwen's chat template and pass them, along with images, through the model. The LM head outputs token-level logits, which are converted to probabilities. We compute cross-entropy loss directly on the gold assistant tokens, i.e., on the predictive distribution of the LM head, rather than on sampled outputs. Image inputs are processed using the Qwen processor. Training is conducted on a single GPU (RTX 3090 Ti). Each training sample consists of a fixed system prompt and a task-specific user prompt. For example, for the lower-half recognition task, the templates used are as follows:

Chinese prompt

> **<system prompt>** 你是一个善于识别汉字的智能助手。图片只展示了一个汉字的下半部分，请你根据下半部分准确识别该汉字，只回答一个汉字。
>
> **<user prompt>** 这张图片显示的是一个汉字的下半部分，上半部分被遮挡住了。请根据可见部分判断这是什么汉字，只回答一个汉字，不要包含其他内容。这个汉字是：

English prompt

> **<system prompt>** You are a helpful assistant that can identify English words in images. The image will show only the lower half of an English word, with the upper half masked. Identify the word accurately based on the visible portion. Please answer with a single word, and do not include any other text.
>
> **<user prompt>** The image contains the lower half of an English word. The upper

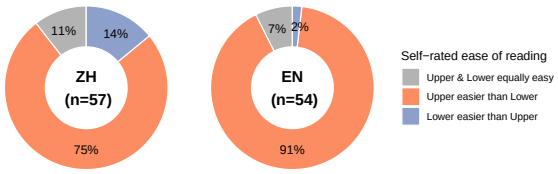

Figure 5: Self-rated ease of reading across visibility conditions. Participants were asked to judge whether the upper or lower half of words was easier to read.

half is masked. What is the word in the image? Please answer with a single word, and do not include any other text. The word is:

C TransOCR Training Details

We trained the Transformer-based OCR model (TransOCR) for character recognition using the PyTorch framework. The model takes grayscale images resized to 32×256 pixels as input and is trained to predict character sequences in an autoregressive manner. Training was conducted using the Adadelta optimizer (ρ = 0.9, weight decay = 1e-4) with an initial learning rate of 1.0 and a batch size of 16. The loss function was standard cross-entropy over predicted character classes. We applied early stopping with a patience of 5 epochs based on validation accuracy.

All models were trained on two NVIDIA GPUs (RTX 3090 Ti) with multi-GPU support (DataParallel), and model checkpoints were saved at each epoch. The best-performing model was selected based on validation accuracy.

During inference, character predictions were generated step-by-step. At each step, the model outputs a probability distribution over the character vocabulary via a softmax layer. We denote this distribution given the image input $\mathbf{o}$ as $p_\theta(\mathbf{c} \mid \mathbf{o})$, where $\mathbf{c}$ is a realization of a random variable $\mathbf{C}$ ranging over characters. We first compute the character-level conditional entropy $H_\theta(\mathbf{C} \mid \mathbf{o})$ at each step using $H_\theta(\mathbf{C} \mid \mathbf{o}) = -\sum_{\mathbf{c}} p_\theta(\mathbf{c} \mid \mathbf{o}) \log p_\theta(\mathbf{c} \mid \mathbf{o})$, and then sum up the entropies of all steps to obtain the word-level conditional entropy $H_\theta(W \mid \mathbf{o}) = -\sum_{\mathbf{C}} H_\theta(\mathbf{C} \mid \mathbf{o})$.

D Self-Rated Ease of Reading

As shown in Figure 5, in both Chinese and English, participants overwhelmingly rated the upper half of words as easier to read. This asymmetry was more pronounced in English, where 91% of participants preferred the upper half, compared to 75% in Chinese.

E Human Performance on Comprehension Questions

Language		Full	Upper	Lower
Chinese	Acc	66%	60%	56%
English	Acc	81%	77%	60%

Table 2: Comprehension question accuracy (Acc) for Chinese and English participants.

We did not include these results from Table 2 in the main text due to limitations in our experimental design. In the Chinese experiment, some screens lacked questions due to paragraph splitting, leading to mismatches between question accuracy and occlusion condition, i.e., answers for a given question could appear on the previous screen under a different occlusion condition. This likely explains the lower overall accuracy in Chinese. In the English experiment, we corrected this issue by including questions for all screens, though they were not generated with the original OneStopQA procedure (Berzak et al., 2020) due to resource constraints.