Full, Keyword, or Glossed? Exploring How Captioning Modes Influence EFL Learners' Listening Comprehension and Caption Reading Patterns: An Eye-Tracking Study

Musa Nushi (Corresponding Author)

Shahid Beheshti University, Tehran, Iran

m_nushi@sbu.ac.ir

https://orcid.org/0000-0003-1917-5372

Kimia Askarian

Friedrich-Schiller-Universität Jena, Weimar, Germany

Kimia.askarian@uni-jena.de

https://orcid.org/ 0009-0007-2240-0506

Mehran Hosseinkhani

Shahid Beheshti University, Tehran, Iran

me.hosseinkhani@mail.sbu.ac.ir

https://orcid.org/0009-0002-4018-2235

Abstract

This study investigated the effects of different captioning modes on intermediate English as a Foreign Language (EFL) learners' listening comprehension and how caption reading patterns influence comprehension of audio-visual materials. Using a two-phase sequential quasi-experimental design, 76 Iranian intermediate EFL learners were assigned to three groups receiving full, keyword, or glossed captioned videos during a six-session instructional program. Listening comprehension was measured through IELTS-based tests, and a subsequent eye-tracking phase with 36 participants explored learners' visual attention and caption reading behaviors. One-way ANOVA results indicated a statistically significant effect of captioning mode on listening comprehension, with full captioning leading to a significantly higher performance than keyword captioning, while no significant differences were found between full and glossed captioning or between keyword and glossed captioning. Eye-tracking data analysis revealed differences in caption reading patterns between high- and low-performing learners: high-performing learners demonstrated more selective attention, focusing primarily on verbs and spending less time on captions overall, whereas low-performing learners exhibited longer and more distributed attention across caption elements. The findings suggest that captioning mode and learners' attention allocation play important roles in listening comprehension and that efficient caption use is associated with better learning outcomes. The study provides pedagogical implications for the use of captioned audio-visual materials and contributes to understanding cognitive processing in multimedia-assisted language learning.

Keywords: audio-visual materials, captioning mode, eye-tracking, multimedia-assisted language learning, quasi-experimental design, visual attention

Introduction

Listening is commonly regarded as a challenging skill for second language (L2) learners and instructors (Nushi and Orouji 2020). To address these difficulties, various pedagogical approaches have been proposed, including the use of authentic audio-visual materials, which expose learners to real-life language and provide visual and contextual cues that may support comprehension (Rost 2024). However, such materials are often difficult to understand without assistance because many learners lack sufficient linguistic knowledge to process unmodified input (Alamri 2025). Consequently, captions and subtitles have been introduced to facilitate understanding by enabling dual processing of auditory and textual information and by creating a supportive target language learning environment (Field 2019).

A survey of the extant literature on captioning, however, reveals rather contradictory views regarding the effect of captioning on learners' listening comprehension. While some researchers (e.g., Mahalingappa et al. 2024; Montero Perez et al. 2015; Winke 2013) argue in favor of the merits of captions on learners' listening comprehension; others have reported captioning as deleterious or ineffective (e.g., Aldera and Mohsen 2013; Li 2025 Montero Perez et al. 2014). For instance, Aldera and Mohsen (2013) explain that captioning may divide learners' attention across visual, textual, and auditory channels—imposing excessive cognitive load and thereby hindering listening comprehension. This inconsistency suggests that the effects of captioning are not uniform and may be mediated by learner-related and task-related variables, such as proficiency level, attentional capacity, input complexity, and the captioning type. Moreover, methodological differences across studies may partially account for these inconsistencies. For instance, earlier studies often lack precise control over input difficulty or fail to distinguish between comprehension of linguistic versus visual information (Dong et al. 2015).

As a result, it remains unclear not only whether captions are beneficial, but under what conditions and through which cognitive mechanisms they exert their effects.

Researchers have utilized varying methodologies to investigate listening, one example of which is eye-tracking. Eye-tracking research uses unique hardware and software to measure the movement of the eyes, record and store that data, and provide visual outputs (Conklin et al. 2019). Although eye-tracking has been a popular approach for exploring reading (Conklin et al. 2020), there have been several significant studies of eye movements in listening literature as well (e.g., Batty 2021; Liu and Aryadoust 2026). Altmann (2011) argues that listeners focus their visual attention on objects as they appear or are anticipated to appear in the linguistic stream; moreover, Dolgunsöz (2015) believes that eye movements are reliable proxies of attention which is directed to written words during listening. Hence eye-tracking methodology can be influential in illuminating certain aspects of listening processing.

Although previous research has examined the effects of captioning on L2 listening comprehension, most studies have either compared a limited range of caption types or focused solely on learning outcomes without examining the underlying cognitive processes. Furthermore, relatively few studies have used eye-tracking to investigate how learners allocate visual attention to different linguistic elements within captions. As a result, little is known about how different captioning modes influence both listening comprehension outcomes and learners' attention to specific parts of speech during captioned viewing. The present study addresses this gap by simultaneously comparing three captioning modes (full, keyword, and glossed) while integrating listening performance measures with eye-tracking data to examine learners' attention allocation across parts of speech. By combining outcome-based and process-based evidence within a single

design, the study provides a more comprehensive understanding of how captioning supports listening comprehension.

Literature Review

Captioning

According to Garza (1991), captioning is the exact and simultaneous transcription of what is being heard. Three modes of captioning that are commonly used in the classroom environment (Teng 2022) include: keyword captioning (KC), which consists of words that are important for the meaning of the sentence or paragraph; glossed captioning (GC), which provides access to meaning through corresponding L1 context-bound translation; and full captioning (FC), which is the simultaneous transcription of what is being heard (Montero Perez et al. 2015).

Captioning has been a popular aid among other classroom supplementary materials and is widely used by teachers (Weingärtner et al. 2024). This popularity has invited many researchers to closely investigate its effects on learning and comprehension resulting in contradictory results. For example, Guillory (1998) reported no significant differences between FC and KC, though they were both better than non-captioning condition, and Dong et al. (2015) similarly found limited differences across captioning modes. However, these findings contrast with those of Montero Perez et al. (2015), who demonstrated clear advantages for FC, suggesting that the effectiveness of captioning may depend on factors such as task design and input difficulty. Although Guillory (1998) provided early evidence supporting captioning, the study suffers from methodological limitations, including the lack of clarity regarding input difficulty and the nature of comprehension being measured. Such limitations reduce the interpretability and generalizability of the findings, highlighting the need for more controlled investigations.

Utilization of Captions in Comprehension of Audio-Visual Materials

This study is grounded in Cognitive Theory of Multimedia Learning (Mayer 2020), which posits that learning occurs through the integration of information from dual channels—auditory and visual—within the constraints of limited working memory. According to this theory, meaningful learning depends on the learner's ability to select, organize, and integrate relevant information across these channels. A key implication of this framework is that instructional materials should be designed to optimize cognitive processing while minimizing unnecessary cognitive load. In multimedia environments such as captioned videos, learners must process spoken input, written captions, and visual context simultaneously. This creates the potential for both facilitation and overload, depending on how information is presented.

Utilization of captions during video-based L2 listening activities has been studied extensively (Winke et al. 2013; Gass et al. 2019). The majority of studies have investigated whether captioned videos are better than non-captioned ones for language learning, with the conclusion that captioning leads to better performance on listening comprehension and aid vocabulary acquisition. However, according to Danan (2016), captions "might become a distraction or a crutch" (14) to learners if they direct learners to focus too much on the words on the screen rather than on the listening process. Hence, captions may be considered redundant and destructive to learning process, segment or parse speech and not necessary for comprehension. Researchers have pointed out that some learners may concentrate on the aural input and close their eyes to eliminate the visuals, (the written input stream) which they find distracting (Winke et al. 2010). They might also ignore the captions to decrease the amount of incoming information (Taylor 2005).

In the present study, distraction was operationalized as inefficient, disproportionate, or unstable allocation of visual attention across multimodal input. This conceptualization is grounded in the eye–mind assumption proposed by Marcel Just and Patricia Carpenter (1980), which posits that eye movements closely reflect ongoing cognitive processing. It is further supported by research in eye-tracking methodology (e.g., Maran et al. 2021), which demonstrates that gaze behavior provides reliable indicators of attentional distribution and processing dynamics.

Specifically, distraction was inferred from a combination of eye-tracking indicators that have been associated with increased cognitive load and less efficient processing in prior research. First, fixation duration (or looking time) has been widely linked to processing difficulty and cognitive effort, with longer fixations typically indicating greater effort in lexical or semantic integration (Inhoff et al. 2008; Rayner 1998). However, in the present study, fixation duration alone was not interpreted as evidence of distraction; rather, it was considered in conjunction with other measures to avoid oversimplification. Second, saccade frequency was included as an index of attentional stability, as higher saccadic activity has been associated with less stable gaze patterns and difficulty integrating information across input streams (Duchowski 2002).

In addition, ratio-based measures of gaze behavior were employed to capture the proportional allocation of attention to captions relative to their on-screen duration. Greater ratios of gaze time to caption duration were interpreted as reflecting potential over-reliance on textual input, which may indicate inefficient processing and reduced integration of auditory and visual information. This interpretation aligns with theoretical accounts of split-attention effects in multimedia learning (e.g., Mayer 2020; Sweller 1999), where learners divide their attention between multiple sources of information that need to be mentally integrated, thereby increasing cognitive load and

reducing processing efficiency. Finally, combined gaze patterns were analyzed to provide a more comprehensive understanding of attentional processes. Following prior eye-tracking research in second language acquisition (e.g., Winke et al. 2013; Conklin et al. 2019), inefficient processing was inferred from patterns that included longer gaze durations, increased saccadic activity, and disproportionate allocation of attention.

This multi-indicator approach is consistent with methodological recommendations in eye-tracking research, which emphasize that individual measures should not be interpreted in isolation but rather as part of a broader pattern of cognitive processing. Taken together, this operationalization is further supported by cognitive load theory (Sweller 1988), which posits that learning is hindered when attention is divided across multiple sources of information, leading to inefficient processing. It is also consistent with the cognitive theory of multimedia learning (Mayer 2020), which highlights the limitations of working memory and the potential for cognitive overload when verbal and visual streams compete for attentional resources.

While the general utility of captions for L2 learning has largely been established (Teng 2024), many pedagogical and theoretical questions about captioning have remained unanswered. For instance, Vanderplank (2016 as cited in Gass et al. 2019, 246) notes that most of what applied linguists know about captions "is still largely anecdotal," highlighting the need for further investigation into how learners actually use captioned video to build their language abilities. More recent meta-analyses and experimental studies (e.g., Teng 2024) suggest that captions generally facilitate listening and vocabulary learning, but their impact is strongly mediated by factors such as caption format, learner proficiency, and cognitive resources. Researchers are yet to learn how learners use the language made available in captioned videos.

Utilization of Captioning as an Educational Tool in EFL Context: Eye Tracking-Based Studies

Eye-movement recording is regarded as one of the most comprehensive indicators of cognitive processes as individuals interpret written text (Aryadoust and Ang 2021). As a diagnostic tool, eye-tracking can provide quantitative evidence of the user's visual attentional processes (Mashat 2024). The hypothesis is that eye movements and cognition are linked (Just and Carpenter 1980). Winke et al. (2013) point out that not many foreign language acquisition studies have utilized eye-tracking methods to investigate learners' use of captions. Most of those studies have investigated the processing of viewers' attention to standard subtitles (video with the sound in L2 and the text in the L1) versus reversed subtitles (video with the sound in L1 and the text in L2).

Eye-tracking studies have shown that caption use varies depending on factors such as the relationship between learners' native and target languages, content familiarity, and individual differences (e.g., d'Ydewalle and De Bruycker 2007; Muñoz 2017). Research has demonstrated differences in time spent on captions across language groups and subtitle types, suggesting that cognitive load, split attention, and language characteristics influence how learners process captioned input (d'Ydewalle and De Bruycker, 2007; Winke et al. 2013). In this context, according to Gass et al. (2019), factors affecting L2 caption use should be studied thoroughly so it could be profitably utilized in designing instructional materials and learning a language.

Researchers still hesitate regarding what parts of captions learners focus on, or even whether they skip some parts of captions entirely (Winke 2013). Neither is there a consensus on how learners balance the simultaneous intake of audio, video, and text. Several studies attempted to address these questions via interviews (Sydorenko 2010; Taylor 2005; Winke et al. 2010) or verbal reports (Gruba 2006). For example, Li (2025) as well as Yeldham (2018) suggest that

when watching captioned videos, learners may prioritize reading over listening, which may prevent them from processing the aural stream and benefiting from the information provided by the aural stream and contextual clues. In this context, findings on the effectiveness of captioning across proficiency levels are also inconsistent, with some studies showing greater benefits for lower-proficiency learners (e.g., Wu et al. 2022), while others report stronger advantages for intermediate and advanced learners compared to elementary learners (e.g., Alabsi 2020; Montero Perez et al. 2014; Taylor 2005).

Taken together, the literature reveals three major limitations. First, findings on the effectiveness of different captioning modes remain inconsistent. Second, most studies have focused on outcome measures without examining underlying cognitive processes. Third, relatively few studies have explored fine-grained attention allocation to linguistic elements within captions, particularly using eye-tracking methods in EFL contexts. Therefore, a comprehensive investigation that integrates both performance outcomes and real-time processing data across multiple captioning modes is needed. Accordingly, the current study takes a detailed examination of three types of captioning (i.e., FC, KC and GC) to find the most effective one for L2 learners. It also investigates students' caption reading patterns using an eye-tracking device to provide useful solutions and implications for L2 classroom use and instructional materials design. Thus, the present study is guided by the following research questions:

1. Do different modes of captioning (i.e., FC, KC, and GC) affect intermediate EFL learners' listening comprehension?

2. Are there any significant differences in how the high-performing learners use each mode of captions compared with their low-performing counterparts?

3. To what extent and which type of captioning mode lead to inefficient attention allocation (i.e., potential distraction), as indicated by eye-tracking metrics?

Method

Research Design

This study employed a two-phase sequential quasi-experimental design to investigate the effects of different captioning modes on intermediate EFL learners' listening comprehension and to examine learners' caption reading patterns. The study combined quantitative experimental procedures with eye-tracking measures to provide both performance-based and process-oriented evidence. In the first phase, a quasi-experimental comparison group design was used to examine the impact of three captioning modes (i.e., FC, KC, and GC) on learners' listening comprehension. Intact classes were randomly assigned to one of the three treatment conditions. Participants completed a six-session instructional program over three weeks, during which they watched captioned instructional videos and completed listening comprehension tests.

The second phase employed an eye-tracking design to investigate learners' visual attention and caption reading patterns while watching captioned videos. Participants were categorized into high-performing and low-performing groups based on their post-test scores from the first phase. Eye-movement data, including fixation duration, saccades, pupil dilation, and looking time, were collected to examine differences in attention allocation across captioning modes and parts of speech. This phase aimed to provide process-level evidence on how learners interacted with captions and how reading behaviors may be associated with comprehension outcomes.

Participants

The study's first phase started with 76 Persian-speaking female EFL learners ranging in age from 16 to 45 years (M = 19.5). Participants were selected via convenience sampling from among those studying at the Iran Language Institute (ILI) in Tehran. The participants were all at intermediate level of English language proficiency, as determined by their scores on the standardized proficiency (or placement) test employed by the institute. The sample consisted of seven intact classes randomly assigned to three groups: FC (N=24, two classes), KC (N=28, three classes), and GC (N=24, two classes).

For the second phase (the eye-tracking session), 36 intermediate EFL learners between 18 and 26 years old (M = 21.7) were chosen via voluntary sampling from the same institute. They were divided into two groups in line with their performance on the test – the high group and the low group. Participant recruitment for the second phase was determined by methodological and practical requirements associated with the eye-tracking procedure. Unlike the classroom-based first phase with 76 learners, the eye-tracking phase involved individual testing, specialized equipment, and controlled laboratory conditions, which necessitated a smaller sample to ensure reliable data collection. Different participants were selected to avoid practice or carryover effects resulting from prior exposure to the videos and tests, which could have influenced attention patterns and cognitive processing. Such familiarity might have threatened the validity of the findings. All participants in both phases provided informed consent prior to participation to ensure voluntary involvement.

Instruments

Listening Comprehension Tests

The researchers designed six listening tests (see supplementary materials), modeled on the IELTS listening section, to measure students' progress throughout the six-session program and examine the effects of different captioning modes on listening comprehension. The tests were carefully developed to ensure validity and suitability. Items followed IELTS formats/task types and were matched to the instructional videos and the learners' intermediate proficiency level. To ensure content validity, two experienced English language teaching (ELT) specialists reviewed the tests for clarity, relevance, and appropriateness, and revisions were made accordingly. A pilot study with a similar group assessed reliability and difficulty, resulting in further refinements to strengthen the instruments.

Videos

Six short video clips were chosen to be presented during the six-session program of the research. These clips were four minutes long maximum and were retrieved from the Cambridge University Press website[1]. The materials in the website are classified based on learners' language proficiency levels and the researchers selected level-appropriate videos for the participants in this study (i.e., intermediate). To further ensure the suitability of the selected materials, two

[1] https://drupal-s3fs-prod.s3.eu-west-1.amazonaws.com/files/6915/9301/0079/Interchange_4_Level_2_Unit_1_Video.mp4
https://drupal-s3fs-prod.s3.eu-west-1.amazonaws.com/files/1315/9301/0711/Interchange_4_Level_2_Unit_2_Video.mp4
https://drupal-s3fs-prod.s3.eu-west-1.amazonaws.com/files/6315/9301/2184/Interchange_4_Level_2_Unit_6_Video.mp4
https://drupal-s3fs-prod.s3.eu-west-1.amazonaws.com/files/3115/9301/2334/Interchange_4_Level_2_Unit_7_Video.mp4
https://drupal-s3fs-prod.s3.eu-west-1.amazonaws.com/files/8615/9301/2659/Interchange_4_Level_2_Unit_9_Video.mp4
https://drupal-s3fs-prod.s3.eu-west-1.amazonaws.com/files/6215/9301/3341/Interchange_4_Level_2_Unit_12_Video.mp4

experienced ELT specialists reviewed the videos for linguistic appropriateness, clarity of content, and alignment with the learners' intermediate proficiency level. Although objective linguistic indices such as lexical density or speech rate were not calculated, the expert review process and the use of level-classified instructional materials were intended to minimize variability in input difficulty across videos. The content of these videos varies, yet they all share one feature in common, that is, they are all made for instructional purposes.

Keyword Determination

The procedure for keyword determination was adopted from similar previous research (e.g., Guillory 1998; Park 2004; Prez et al. 2018). Five ELT experts, with at least 10 years of teaching experience, were asked to watch the clips and underline the words they consider significant in rendering the overall content of the video clips. Those words on which the teachers had more than 50 percent agreement were selected to be used for KC and GC. The teachers chose 152 words in total which were then classified into three groups of nouns, (n =65), verbs (n=66), and adjectives-adverbs (n =21).

Data Collection

The Pilot Study

Prior to the actual experiment, the researchers conducted a pilot study to detect the possible problems with the data collection procedure. The participants in this phase were 24 intermediate EFL learners (not included in the main treatment) from the same institute from which the main study participants were chosen. Since the tests contained different items and were not intended as parallel forms, the internal consistency of each test was examined separately using Cronbach's alpha (see supplementary materials). The piloting was conducted over two sessions and the

results revealed that the content of the tests was somewhat beyond the students' proficiency level and the reliability of the tests was not at an acceptable range (i.e., from 0.32 to 0.61). Hence, the tests were modified for the main study. Moreover, the piloting helped the researchers make slight adjustments to the treatment procedure, especially in terms of time budgeting and presentation of the materials. As a result, reliability coefficients for the six tests ranged from 0.76 to 0.89, indicating acceptable internal consistency for the main study.

The Six-Session Program

Seven intact classes were assigned to three groups: FC, KC, and GC. They watched the selected videos during a six-session program which lasted three weeks. In each session, they first watched the videos, and then took the tests. The students of the FC group were presented with the full-captioned videos and the students of the KC group were provided with the keyword-captioned videos; however, the GC group could see the KC along with their Persian equivalents on the screen. The students were given the tests before the videos started and could read the questions while watching the videos.

The Eye-Tracking Sessions

Thirty-six participants in the eye-tracking sessions were asked to watch all six videos – the same ones used in the first phase of the study – in a row, with a small break between every two videos; the whole process took about 25 minutes. The eye-tracking device was RED model manufactured by Senso-Motoric Instruments. The procedure required the captioning to be presented in a fixed sequence – keyword, glossed, and full – for every subject since the post-test was on the FC videos, so every subject would get the same sequence of modes of captioning. If

the sequence were to be of a different mode for each subject, the results would have been greatly affected by the participants' exhaustion, which could have led to different eye movements.

The post-test in the eye-tracking phase was administered using FC videos based on the results of the first phase of the study. Since the findings of the initial phase indicated that FC led to the highest listening comprehension performance, FC was selected to provide a consistent and effective condition for examining learners' visual attention and caption reading patterns. This decision enabled the researchers to support the findings from the first phase and further investigate the cognitive processing underlying the different captioning modes used in the experimental phase.

It is worth mentioning that the participants were only told there was going to be a post-test of the set of videos; none of them were aware that they would be tested on the content of the last two videos with FC on the post-test – this was due to limiting the participants to view the videos with different attention allocation. Furthermore, the participants each received a different sequence of videos so that the video contents would not affect their caption reading. Each mode of captioning was given two different videos to ensure that if the participants did not comprehend the captioning process on the first video, they would be familiarized with it on the second video to not affect the results. The post-test was given to the participants along with refreshments, and they each had about 15 minutes to complete the test until the next subject was ready to start.

Results

Different Modes of Captioning and Listening Comprehension

The primary aim of the current study is to find out whether different modes of captioning (FC, KC, GC) affect learners' listening comprehension. To this end, a one-way ANOVA was used.

Before proceeding to data analysis, we first checked normality of distribution and homogeneity of variances. In order to evaluate the normality of the data, skewness and kurtosis statistics were run. The value for skewness was between −.147 and .276 and the value for kurtosis was .296. Thus, they were far less than the cutoff values of ± 2.0 for skewness and kurtosis showing the univariate normality of the response (Loewen and Plonsky 2017). This was further supported by the non-significant results of Kolmogorov-Smirnov and Shapiro-Wilk statistics (p = .636). Regarding homogeneity of variances, Levene's Test of Error Variances was examined. Results showed the test was not significant (p =.158), indicating that the variance in scores did not violate this assumption. Descriptive statistics for three types of captioning are presented in Table 1, and the mean scores for FC, KC, and GC were 31.87, 27.17, and 28.29, respectively.

Table 1: Descriptive Statistic for the Three Types of Captioning Total Score

					95% Confidence Interval for Mean	
	N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound
full captioning	24	31.8750	6.91761	1.41205	28.9540	34.7960
keyword captioning	28	27.1786	7.82708	1.47918	24.1435	30.2136
glossed captioning	24	28.2917	5.50477	1.12366	25.9672	30.6161

In order to examine whether there were statistically significant differences in participants' listening comprehension scores across multiple testing sessions and captioning conditions over time, a multivariate test of repeated measures ANOVA was conducted Prior to conducting the repeated measures ANOVA, it was gathered that based on Mauchly's Test, the Chi-square test did not turn out significant (p = 0.12), thus showing the Sphericity assumption was not violated.

According to Table 2, the repeated measures ANOVA indicated a significant main effect of time (i.e., session), suggesting that participants' listening comprehension improved across the six sessions, regardless of captioning condition.

Table 2: Multivariate Tests of Repeated Measures ANOVA

GROUP	Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared (ηp^2)
Full captioning	.339	7.417	5.00	19.00	.001	.661
Keyword captioning	.269	12.470	5.00	23.00	.000	.731
Glossed captioning	.358	6.802	5.00	19.00	.001	.642

To test whether the differences of means between groups were statistically significant, a one-way between-groups ANOVA was run (see Table 3). The analysis was meant to determine the impact of different modes of captioning on listening comprehension, which showed there was a statistically significant difference between the three groups (F (2,73) = 3.207, p=.046, η^2=.080). Hence, it could be concluded that the scores obtained by the three groups show that different modes of captioning do affect listening comprehension.

Table 3: One-Way Between Groups ANOVA

Total score	Sum of Squares	df	Mean Square	F	η^2	Sig.
Between Groups	303.296	2	151.648	3.207	.080	.046
Within Groups	3451.690	73	47.283			
Total	3754.987	75				

Following the one-way between groups ANOVA, a Tukey post hoc test (see Table 4) revealed that the mean difference between FC and KC (M= 4.969) was significant (p = .043, d=.63). However, there was no statically significant difference between FC and GC as well as KC and GC (p = .175, p = .830).

Table 4: Multiple Comparisons Tukey HSD

		Mean			95% Confidence Interval	
		Difference	Std.		Lower	Upper
(I) GROUP	(J) GROUP	(I-J)	Error	Sig.	Bound	Bound
Full captioning	Keyword captioning	4.69643	1.91281	.043	.1202	9.2727
	Glossed captioning	3.58333	1.98502	.175	-1.1657	8.3324
Keyword captioning	Full captioning	-4.69643	1.91281	.043	-9.2727	-.1202
	Glossed captioning	-1.11310	1.91281	.830	-5.6894	3.4632
Glossed captioning	Full captioning	-3.58333	1.98502	.175	-8.3324	1.1657
	Keyword captioning	1.11310	1.91281	.830	-3.4632	5.6894

Eye-tracking Results

To address the second and third research questions, an eye-tracking experiment was carried out to determine whether learners differed significantly in their caption-reading behaviors. Thirty-six participants took part in this phase and were categorized into high- and low-performing groups based on their scores on the listening comprehension test administered during the eye-tracking session. Table 5 summarizes the participants' age and test scores for both groups. The results indicate that the classification was appropriate, as the high-performing group achieved a mean score of 17.2, while the low-performing group obtained a mean of 9.61. The age distribution was also balanced, with similar averages in both groups.

Table 5: Participants' Age and Grade Descriptive Indicators Based on Their Groupings

Group	Indicators	Minimum	Maximum	Mean	Standard Deviation
Low	**Age**	18	25	18.94	2.6
	Grade	4	13	9.61	2.8
High	**Age**	18	26	19.8	2.2
	Grade	14	20	17.2	2

The most important indicators of the validity of measurement of the eye-tracking device include the deviation from the X axis and the Y axis, and the tracking ratio calculated for both groups (see supplementary materials). It is observed that the deviation indicators from the X and Y axes are both under 1 degree, which shows the device's measurement error in both groups is at an acceptable rate. Likewise, the tracking ratio indicator shows a low percentage of lost time in the eye-tracking of both groups. Therefore, the validity of the eye-tracking device is confirmed.

To address the second and third research questions, a mixed ANOVA was employed because the design included a between-subjects variable (high- vs. low-performing groups) and a within-subjects variable (parts of speech), along with several dependent variables derived from eye-tracking measures. The analyses revealed significant main and interaction effects for both group and within-subject factors, indicating consistent differences in learners' visual attention and processing patterns (detailed statistical results for each analysis are reported in the supplementary materials). Only content words (verbs, nouns, and adjectives/adverbs) were examined, as function words received minimal fixations. Before conducting the analyses, Mauchly's tests confirmed that the sphericity assumption was met, with non-significant chi-square results (see supplementary materials for details of each test).

The Average Duration of Looking Time at Full Captioning Based on Parts of Speech

Based on Figure 1, results indicate that learners with lower performance levels tended to spend more time looking at captioned words across all content word categories compared to higher-performing learners. This suggests that low-performing learners may rely more heavily on captions to support their comprehension and may require more processing time to understand lexical information. In contrast, high-performing learners showed shorter fixation times, which may reflect more efficient processing, greater automaticity in language comprehension, or less dependence on textual support.

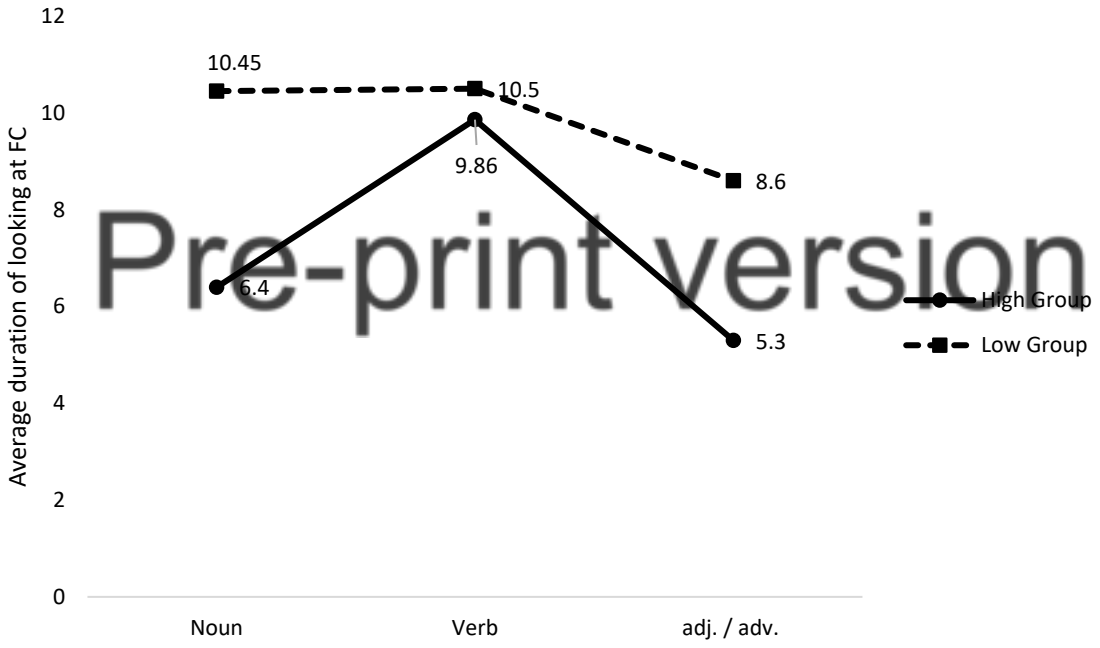

Figure 1: The visual representation of the interaction effect and parts of speech on the average duration of looking at FC

The Average Saccades on Full Captioning Based on Parts of Speech

The average saccades on FC based on different parts of speech and participating groups are presented in Figure 2. The higher number of saccades in the low-performing group for nouns

and adjectives/adverbs indicates that these learners shifted their gaze more frequently between captioned words, which may reflect less stable attention, increased processing effort, or difficulty in integrating lexical information. In contrast, the high-performing group showed fewer saccades for these word categories, suggesting more focused and efficient visual processing. However, the high-performing group produced more saccades for verbs, which may indicate greater sensitivity to action-related information that is crucial for understanding meaning in context.

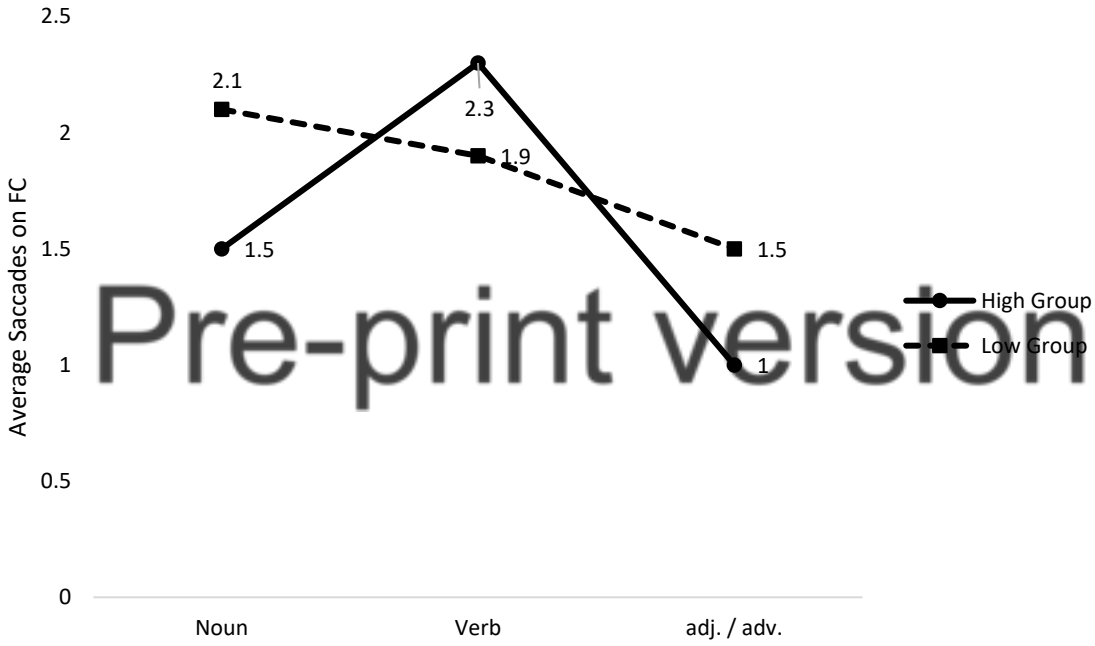

Figure 2: The visual representation of the interaction effect and parts of speech on the average saccade on FC

The Average Fixation on Full Captioning Based on Parts of Speech

The average fixation time on FC based on different parts of speech and participating groups is presented in Figure 3. In essence, the high-performance group showed longer fixation times on nouns, which may indicate deeper processing of key lexical items that carry essential meaning in

the input. In contrast, the low-performance group demonstrated longer fixation times on verbs and adjectives/adverbs, suggesting that they required longer durations to process these word types, possibly due to greater cognitive effort or difficulty in understanding the information.

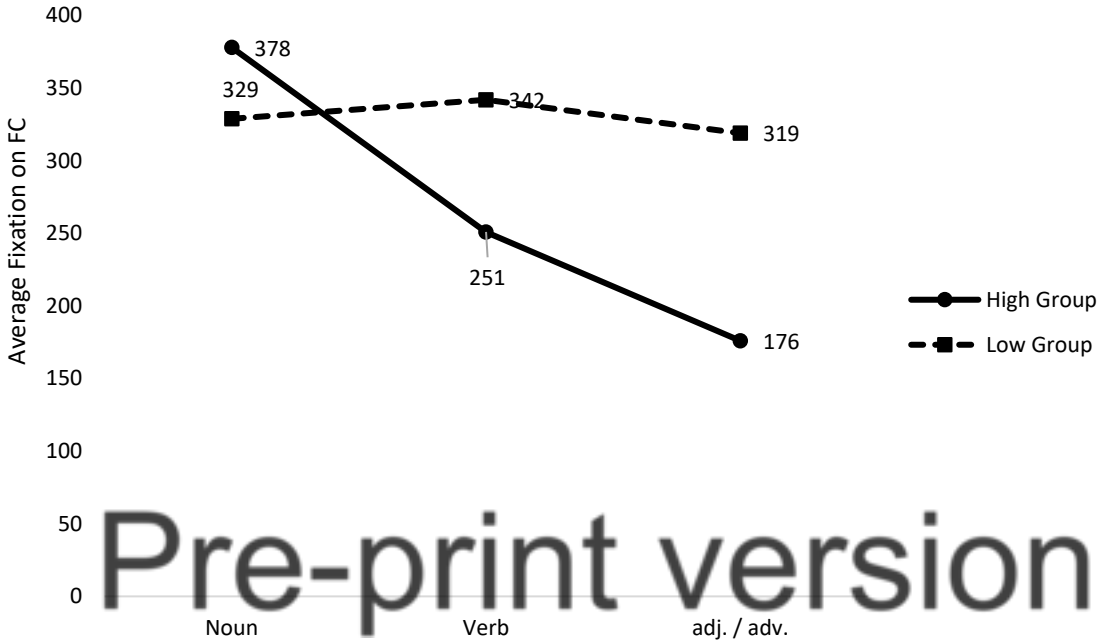

Figure 3: The visual representation of the interaction effect and parts of speech on the average fixation on FC

The Average Pupil Diameter on Full Captioning Based on Parts of Speech

The average pupil diameter on FC based on different parts of speech and participating groups is presented in Figure 4. The high-performance group showed an average pupil diameter of 4.1 millimeters for nouns, 4.8 millimeters for verbs, and 3.7 millimeters for adjectives and adverbs. However, in the low-performing group's average pupil diameter for nouns, verbs, and adjectives and adverbs are 3.6, 3.5, 3.9 millimeters respectively.

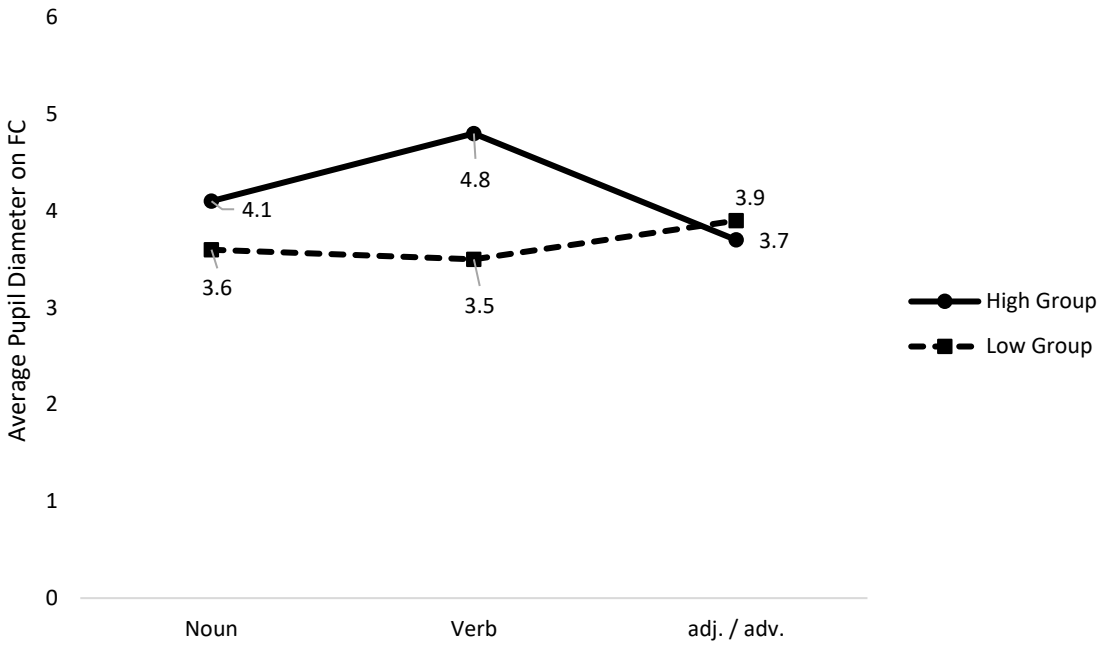

Figure 4: The visual representation of the interaction effect and parts of speech on the average pupil diameter on FC

The Ratio of Looking time at the Captioning to the Duration of the Captioning Based on Types of Captioning

The average ratio of fixation on the captioning to the whole captioning duration based on different types of captioning and participating groups is presented in Figure 5. Based on the graph, it is observed, in the low group the calculated ratio for KC and GC has increased significantly, while the same is not seen in the high group. The results indicate that low-performing learners, especially in GC and KC, relied more heavily on captions, as reflected by their higher fixation-to-caption duration ratios. This suggests greater dependence on textual input and less efficient processing of audiovisual information. In contrast, high-performing learners showed lower ratios, especially in FC, indicating more selective and efficient use of captions.

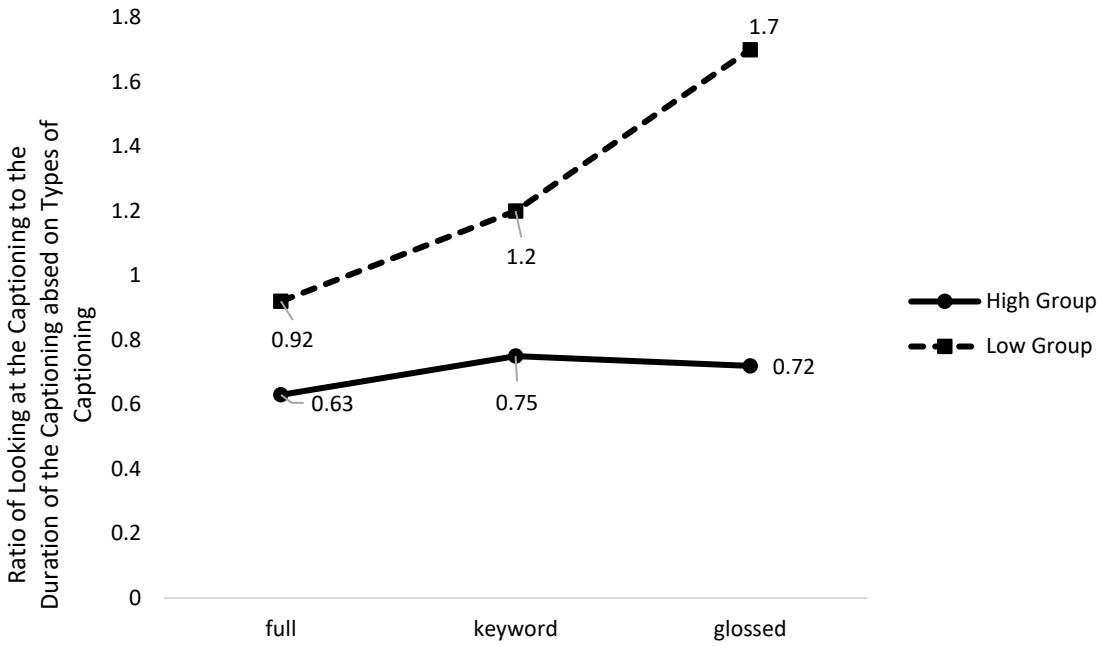

Figure 5: The visual representation of the interaction effect and parts of speech on the ratio of looking at the captioning to the duration of the captioning based on types of captioning

The Ratio of Saccades on Captioning to Captioning Duration Based on Types of Captioning

The average ratio of saccades on captioning to the whole captioning duration based on different modes of captioning and participating groups is presented in Figure 6. Based on the graph, it is observed, the calculated ratio in low-performing participants is significantly higher in GC and KC compared to FC, while this difference is not seen in the high-performing participants. Low-performing learners showed higher ratios across all captioning modes—FC, KC, and GC—with particularly high values in KC and GC, suggesting greater cognitive effort, less stable attention, and more difficulty integrating caption information. In contrast, high-performing learners showed lower saccade ratios across FC, KC, and GC, indicating more stable visual attention and more efficient processing of captioned input.

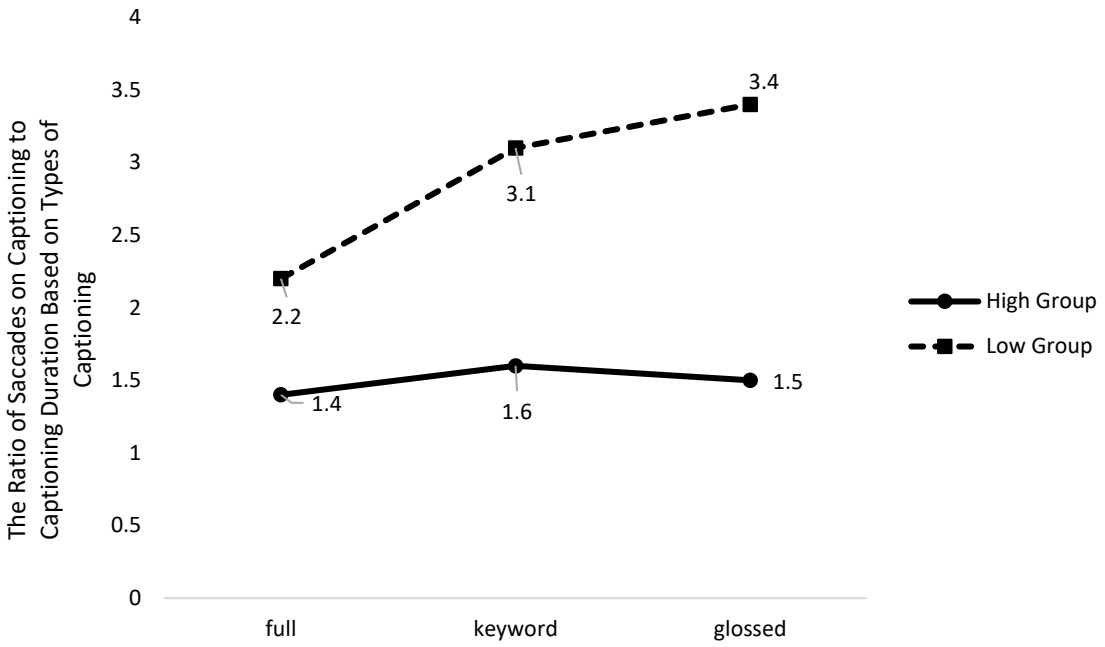

Figure 6: The visual representation of the interaction effect and parts of speech on the ratio of saccades on captioning to captioning duration based on types of captioning

The Ratio of Fixation on Captioning to Captioning Duration Based on Types of Captioning

The average ratio of fixation on captioning to the whole captioning duration based on different types of captioning and participating groups is presented in Figure 7. Based on the graph, it is observed, only low-performing participants were faced with the difference in the calculated ratio based on different types of captioning, while high-performing participants were not distracted. High-performing learners showed slightly higher fixation ratios across FC, KC, and GC, indicating greater attention to captions and more consistent engagement with the textual input. In contrast, low-performing learners demonstrated lower fixation ratios, particularly in KC and GC, suggesting reduced attention to captions and less sustained processing of captioned information.

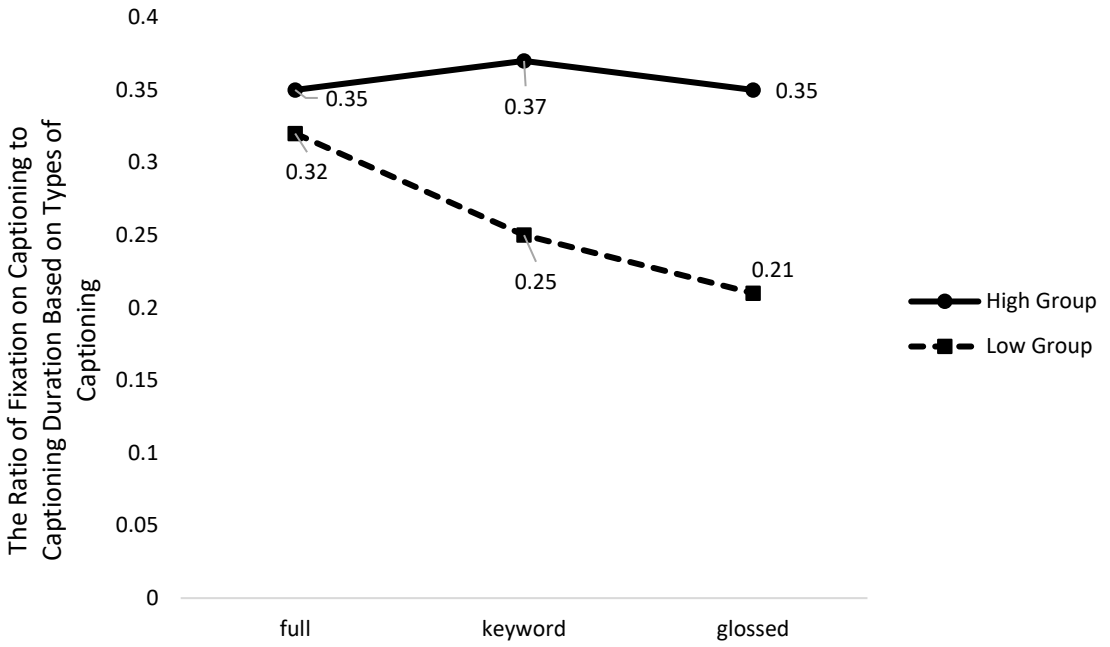

Figure 7: The visual representation of the interaction effect and parts of speech on the ratio of fixation on captioning to captioning duration based on types of captioning

The Ratio of Pupil Diameter on Captioning to Captioning Duration Based on Types of Captioning

The average ratio of pupil diameter on captioning to the whole captioning duration based on different types of captioning and participating groups is presented in Figure 8. The ratio of pupil diameter to caption duration provides insight into learners' cognitive load and processing efficiency across captioning modes. The results indicate that high-performing learners showed relatively stable ratios across FC, KC, and GC, suggesting consistent cognitive engagement regardless of caption type. In contrast, low-performing learners exhibited a noticeable decrease in this ratio, particularly in GC, which may reflect reduced cognitive engagement or difficulties in effectively processing the combined textual and translated input. This pattern suggests that

glossed captions may impose additional processing demands on lower-performing learners, potentially leading to less efficient allocation of cognitive resources.

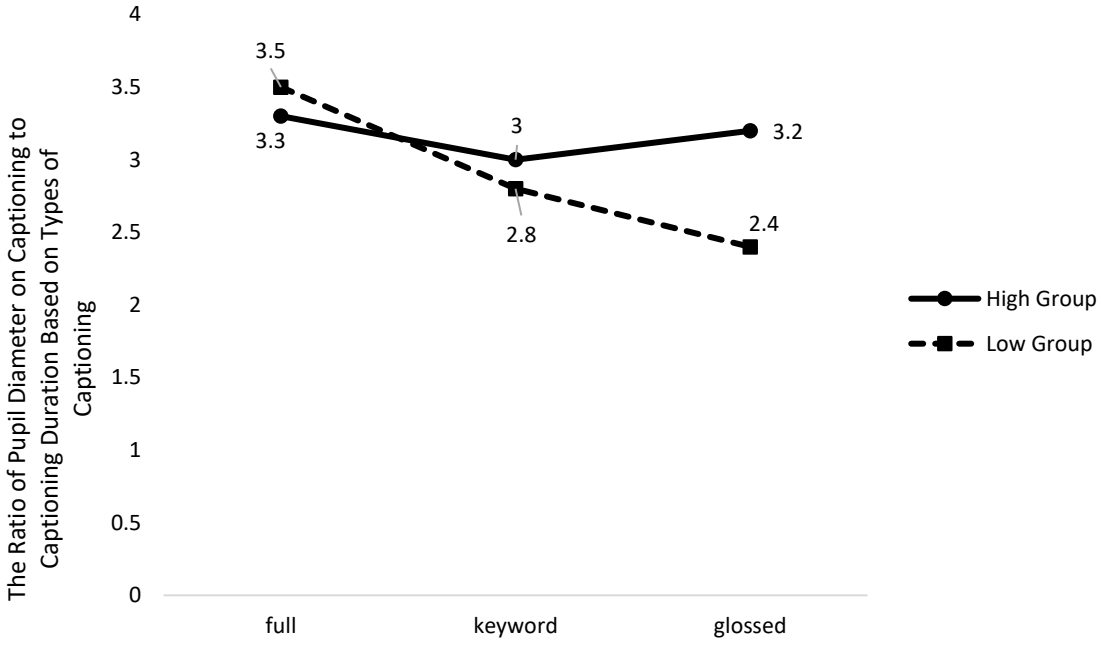

Figure 8: The visual representation of the interaction effect and parts of speech on the ratio of pupil diameter on captioning to captioning duration based on types of captioning

According to the results of the analyses, KC and GC were associated with increased indicators of attentional dispersion compared to FC, particularly among low-performing learners.

Discussion

To answer the first research question of the current study, a three-week program was implemented to investigate how the three groups of participants perform on the given post-tests. Data analysis revealed that there was a significant difference between the three groups, especially between the FC and KC groups, and the FC group outperformed the other two groups on listening comprehension tests conducted in this phase of the study.

Beyond statistical significance, the magnitude of the observed effects provides important insight into the practical relevance of the findings. The repeated-measures analyses yielded large partial eta squared values across captioning conditions (ηp^2 ranging from .642 to .731), indicating that a substantial proportion of the variance in listening comprehension was associated with instructional exposure over time. According to conventional benchmarks (i.e., small = .01, medium = .06, large = .14), these values represent large effects, suggesting that sustained engagement with captioned materials meaningfully enhanced learners' listening comprehension.

For the one-way ANOVA comparing captioning modes, eta squared was calculated using the sums of squares values, yielding $\eta^2 = .080$. This value represents a medium effect size, indicating that captioning mode accounted for approximately 8% of the variance in listening comprehension scores. Although modest in magnitude, this proportion of explained variance is educationally meaningful in classroom settings.

For the significant pairwise comparison between FC and KC, Cohen's d was calculated to assess the magnitude of the mean difference. The resulting value ($d = 0.63$) indicates a medium effect size, suggesting the advantage of FC over KC. Although the effect size between FC and KC was modest, the mean difference of approximately 4.7 points represents a pedagogically meaningful difference in listening outcomes.

The superior performance observed in the FC condition can be interpreted in light of multimedia learning theory (Mayer 2020). Full captions may have facilitated the integration of auditory and textual information, thereby supporting dual-channel processing. Although the redundancy principle suggests that excessive textual input may hinder learning (Mayer 2020), the findings indicate that, for intermediate learners, the additional support provided by full

captions outweighed potential cognitive costs. These findings are in accordance with those of the previous studies confirming the undeniable benefits of captioning for ESL learners in vocabulary development and listening comprehension (e.g., Alotaibi et al. 2023). However, these findings are different from what Guillory (1998) and Dong et al. (2015) reported in their studies, which showed there was no significant difference between KC and GC.

The next two questions of the current study were answered using an eye-tracking device to examine learners' attention allocation during captioned viewing. As mentioned before, to see whether there is a significant pattern to students' caption reading, the researchers divided the participants into two groups of high performing and low performing. Data analysis revealed that the low-performing group spent more time looking at the captions overall. In contrast, the high-performing group focused more on verbs than on nouns, adjectives, and adverbs, whereas the low-performing group distributed their attention more evenly across these word types. Moreover, the high group's saccades on verbs were significantly more frequent than other parts of speech, while the low-performing participants paid more attention to nouns. The high group's fixation was mostly on nouns, and least of all on adjectives and adverbs, with a medium fixation on verbs; the low group was not so greatly different in this case. As stated in previous studies, fixations happen to be longer when readers come across long, low frequency, or contextually implausible words or phrases (Inhoff et al. 2008; Joseph et al. 2008). Another study by Duchowski (2002) also reported that when reading becomes more advanced or conceptually more complex, eye-fixation duration increases and saccade length decreases.

Both high and low groups' pupil diameter were quite similar in looking at nouns, adjectives and adverbs, but the high-performing participants' pupil diameter was slightly wider while looking at verbs, which indicates the depth of processing parts of speech. These results are

consistent with prior research that showed reductions in maximum pupil size can demonstrate learning across trials while completing a task (Coyne et al. 2019).

To answer the last question, the ratio of different eye movements to the overall time of different modes of captioning on screen was calculated. Based on the mixed ANOVA results, the calculated ratio of looking time to caption duration was significantly higher for KC and GC in the low-performing group. Similarly, the ratio of saccades relative to caption duration was significantly higher for KC and GC than for FC in the low-performing group, whereas this pattern was not observed in the high-performing group. Additionally, only the low-performing participants faced the difference in the calculated ratio based on the different types of captioning, while the high-performing participants were not distracted. The calculated ratios for FC and GC do not have a significant difference in either group; on the contrary, for GC the ratio has decreased significantly for the low-performing participants.

It can be concluded that KC and GC are associated with inefficient attention allocation in comparison with the FC, although this pattern was only seen in low-performing participants. The increased indicators of attentional dispersion observed in the KC and GC conditions may reflect split-attention effects (Mayer 2020; Sweller 1999), as predicted by multimedia learning theory. The reduced or fragmented textual input may have required learners to allocate cognitive resources across multiple sources of information, resulting in less efficient processing. To sum up, these results are in accordance with what was shown in previous studies where learners highlighted the usefulness of captions for different linguistic and learner-related reasons and that the keyword captions were generally deemed as distracting for learners' listening experience (Montero Perez et al. 2014).

The findings also carry several concrete implications for classroom practice and multimedia design. First, the superior performance associated with FC suggests that instructors working with intermediate learners should initially provide full textual support rather than reduced formats such as keyword captions. This approach appears particularly beneficial during early exposure to authentic audio-visual input, where learners may require full lexical scaffolding to build comprehension confidence.

Second, the eye-tracking results indicate that low-performing learners demonstrated longer and more dispersed attention patterns across caption elements, suggesting inefficient allocation of cognitive resources. This finding highlights the importance of incorporating explicit strategy instruction into listening courses. Teachers may consider training learners to focus selectively on meaning-bearing elements such as verbs and key nouns, pause videos strategically, and alternate attention between auditory and textual streams. Such guided training may help learners develop more efficient caption-reading behaviors similar to those observed among high-performing learners.

Third, the findings suggest implications for digital learning environments and educational technology design. Developers of language-learning platforms and captioning systems may consider implementing adaptive captioning features that allow learners to switch between caption modes based on proficiency level or task complexity. For example, learners could begin with full captions and gradually transition to keyword captions as their listening skills improve. This graduated approach may reduce cognitive overload while maintaining meaningful language exposure.

A key contribution of the present study lies in its integrative methodological design, which simultaneously examined learning outcomes and real-time processing behaviors across

multiple captioning modes. While previous research has often compared caption formats or investigated eye-movement patterns separately, relatively few studies have combined experimental listening measures with fine-grained eye-tracking analyses within a unified framework. By linking performance-based results with visual attention indicators, the present study advances understanding of how captioning influences not only what learners achieve but also how they cognitively process multimodal input. Furthermore, the study extends the existing literature by providing empirical evidence on attention allocation to specific parts of speech during captioned listening. This fine-grained analysis offers new insight into the linguistic elements that learners prioritize during comprehension, thereby contributing to a more process-oriented account of multimedia-assisted language learning.

Conclusion

This study examined the effects of three captioning modes—FC, KC, and GC—on intermediate EFL learners' listening comprehension and explored learners' caption reading patterns through eye-tracking measures. The findings demonstrated that captioning mode significantly influenced listening comprehension, with FC leading to significantly higher performance than KC, while no significant differences were found between FC and GC or between KC and GC. Eye-tracking results further revealed clear differences in attention allocation between high- and low-performing learners. High-performing learners demonstrated more selective and efficient processing, focusing particularly on verbs and spending less time on captions overall, whereas low-performing learners showed longer and more distributed attention across caption elements and relied more heavily on textual input. Moreover, KC and GC appeared to be associated with greater distraction for low-performing learners, suggesting that reduced or segmented textual support may increase cognitive load for learners with less efficient processing strategies. Overall,

the findings highlight that both captioning mode and learners' attention allocation patterns play a crucial role in listening comprehension and multimedia processing.

From a pedagogical perspective, the findings suggest that FC may provide more effective support for intermediate EFL learners' listening development, particularly in instructional contexts involving authentic audio-visual materials. Teachers and instructional designers are therefore encouraged to integrate full-captioned videos into listening instruction, especially for learners who lack efficient caption-reading strategies. Additionally, the results underscore the importance of explicitly training learners in effective caption use, such as selective attention to key linguistic elements and balanced processing of auditory and textual input. Curriculum designers may also consider learners' proficiency levels and cognitive demands when selecting captioning formats for multimedia materials.

Despite its contributions, the study has several limitations. First, the participants were limited to Persian-speaking intermediate female learners from a single language institute, which restricts the generalizability of the findings to other proficiency levels, genders, or learning contexts. Second, the relatively small sample size in the eye-tracking phase and the fixed sequence of caption presentation may have influenced learners' attention patterns. Third, the treatment duration was relatively short, and the study did not examine long-term effects of caption use or strategy development. Another limitation concerns the lack of objective linguistic indices of video difficulty, such as lexical density, speech rate, or syntactic complexity. Although videos were selected based on proficiency level and expert review, future research should incorporate quantitative measures of input difficulty to enhance methodological precision and replicability. Finally, the use of convenience sampling may have introduced sampling bias, as participants were selected based on accessibility rather than random selection. This limits the

representativeness of the sample and reduces the generalizability of the findings to broader EFL populations.

Future research should investigate the effectiveness of captioning modes across different proficiency levels, linguistic backgrounds, and learning settings to enhance generalizability. Longitudinal studies could explore how caption use and attention allocation develop over time and whether strategy training improves learners' processing efficiency. Further research may also examine additional variables such as working memory, cognitive load, playback speed, and learner training in caption-reading strategies. Expanding the use of eye-tracking and other process-oriented measures may provide deeper insights into the cognitive mechanisms underlying multimedia-assisted language learning and contribute to more effective instructional design.

Funding

The authors declare that the research was carried out without receiving any external funding.

Authors' Contribution

The first author contributed to supervision, conceptualization of the study, and data collection. The second author was responsible for data collection, data analysis, and implementation of the research procedures. The third author contributed to data analysis and manuscript editing.

Disclosure of interest

The authors confirm that there are no conflicts of interest related to this study.

Acknowledgements

The authors express their sincere gratitude to the participants for their time and involvement. They also extend their appreciation to the editor and the anonymous reviewers for their insightful and constructive comments.

References

Alabsi, Thuraya. 2020. “Effects of Adding Subtitles to Video via Apps on Developing EFL Students’ Listening Comprehension.” *Theory and Practice in Language Studies* 10 (10): 1191–1199. https://doi.org/10.17507/tpls.1010.02

Alamri, Wafa. 2025. “Evaluating the Benefits and Challenges of Using Authentic Materials in EFL Context for Listening Purpose.” *Frontiers in Education* 10: 1611308. https://doi.org/10.3389/feduc.2025.1611308

Aldera, Abdullah S., and Mohammed Ali Mohsen. 2013. “Annotations in Captioned Animation: Effects on Vocabulary Learning and Listening Skills.” *Computers & Education* 68: 60–75. https://doi.org/10.1016/j.compedu.2013.04.018

Alotaibi, Hind M., Hassan Saleh Mahdi, and Deema Alwathnani. 2023. "Effectiveness of Subtitles in L2 Classrooms: A Meta-Analysis Study" *Education Sciences* 13, no. 3: 274. https://doi.org/10.3390/educsci13030274

Aryadoust, Vahid, and Bee Hoon Ang. 2021. “Exploring the Frontiers of Eye Tracking Research in Language Studies: A Novel Co-Citation Scientometric Review.” *Computer Assisted Language Learning* 34 (7): 898–933. https://doi.org/10.1080/09588221.2019.1647251

Batty, Aaron Olaf. 2021. “An Eye-Tracking Study of Attention to Visual Cues in L2 Listening Tests.” *Language Testing* 38 (4): 511–535. https://doi.org/10.1177/0265532220951504

Conklin, Kathy, Ana Pellicer-Sánchez, and Gareth Carrol. 2019. *Eye-Tracking: A Guide for Applied Linguistics Research*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108233279

Conklin, Kathy, Sara Alotaibi, Ana Pellicer-Sánchez, and Laura Vilkaitė-Lozdienė. 2020. "What Eye-Tracking Tells Us about Reading-Only and Reading-While-Listening in a First and Second Language." *Second Language Research* 36 (3): 257–276. https://doi.org/10.1177/0267658320921496

Coyne, Joseph T., Noelle Brown, Cyrus K. Foroughi, and Ciara M. Sibley. 2019. "Improving Pupil Diameter Measurement Accuracy in a Remote Eye Tracking System." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 63 (1): 49–53. https://doi.org/10.1177/1071181319631176

Danan, Martine. 2016. "Enhancing Listening with Captions and Transcripts." *Applied Language Learning* 26: 1–24.

d'Ydewalle, Géry, and Wim De Bruycker. 2007. "Eye Movements of Children and Adults While Reading Television Subtitles." *European Psychologist* 12 (3): 196–205. https://doi.org/10.1027/1016-9040.12.3.196

Dolgunsöz, Emrah. 2015. "Measuring Attention in Second Language Reading Using Eye-Tracking: The Case of the Noticing Hypothesis." *Journal of Eye Movement Research* 8 (5). https://doi.org/10.16910/jemr.8.5.4

Dong, Jiangqiao, Yajing Zhou, and Guiru Liu. 2015. "The Effect of Caption Modes on EFL Students' Video Comprehension." *Journal of Language Teaching and Research* 6 (2): 397–404. http://dx.doi.org/10.17507/jltr.0602.21

Duchowski, Andrew T. 2002. “A Breadth-First Survey of Eye-Tracking Applications.” *Behavior Research Methods, Instruments, & Computers* 34: 455–470. https://doi.org/10.3758/BF03195475

Field, John. 2019. “Second Language Listening: Current Ideas, Current Issues.” In *The Cambridge Handbook of Language Learning*, edited by John W. Schwieter and Alessandro Benati, 283–319. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108333603.013

Garza, Thomas J. 1991. “Evaluating the Use of Captioned Video Materials in Advanced Foreign Language Learning.” *Foreign Language Annals* 24 (3): 239–258. https://doi.org/10.1111/j.1944-9720.1991.tb00469.x

Gass, Susan, Paula Winke, Daniel R. Isbell, and Jieun Ahn. 2019. “How Captions Help People Learn Languages: A Working-Memory, Eye-Tracking Study.” *Language Learning & Technology* 23 (2): 84–104. https://doi.org/10.10125/44684

Gruba, Paul. 2006. “Playing the Videotext: A Media Literacy Perspective on Video-Mediated L2 Listening.” *Language Learning & Technology* 10 (2): 77–92. https://doi.org/10.64152/10125/44062

Guillory, Helen Gant. 1998. “The Effects of Keyword Captions to Authentic French Video in Foreign Language Instruction.” *CALICO Journal* 15: 89–108.

Inhoff, Albrecht W., Matthew S. Starr, Matthew Solomon, and Lars Placke. 2008. “Eye Movements during the Reading of Compound Words and the Influence of Lexeme Meaning.” *Memory & Cognition* 36 (3): 675–687. https://doi.org/10.3758/mc.36.3.675

Just, Marcel, and Patricia A. Carpenter. 1980. *A Theory of Reading: From Eye Fixations to Comprehension*. Pittsburgh: Carnegie Mellon University. https://doi.org/10.1184/R1/6613262.v1

Joseph, Holly S. S. L., Simon P. Liversedge, Hazel I. Blythe, Sarah J. White, Susan E. Gathercole, and Keith Rayner. 2008. "Children's and Adults' Processing of Anomaly and Implausibility during Reading: Evidence from Eye Movements." *The Quarterly Journal of Experimental Psychology* 61 (5): 708–723. https://doi.org/10.1080/17470210701400657

Li, Yan. 2025. "Listen or Read? The Impact of Proficiency and Visual Complexity on Learners' Reliance on Captions" *Behavioral Sciences* 15, no. 4: 542. https://doi.org/10.3390/bs15040542

Liu, Tingting, and Vahid Aryadoust. "An Eye-Tracking Study of the Impact of Item Presentation and Item Format on Cognitive Processing in L2 Listening Assessment." *Studies in Second Language Acquisition*, 2026, 1–28. https://doi.org/10.1017/S0272263126101648.

Loewen, Shawn, and Luke Plonsky. 2017. *An A–Z of Applied Linguistics Research Methods*. London: Bloomsbury Academic. https://doi.org/10.1007/978-1-137-40322-3_1

Mahalingappa, Laura, Jiaxuan Zong, and Nihat Polat. 2024. "The Impact of Captioning and Playback Speed on Listening Comprehension of Multilingual English Learners at Varying Proficiency Levels." *System* 120: 103192. https://doi.org/10.1016/j.system.2023.103192

Maran, Thomas, Marco Furtner, Simon Liegl, Theo Ravet-Brown, Lucas Haraped, and Pierre Sachse. 2021. "Visual Attention in Real-World Conversation: Gaze Patterns Are

Modulated by Communication and Group Size." *Applied Psychology* 70: 1602–1627. https://doi.org/10.1111/apps.12291

Mashat, Arwa. 2024. "Eye Physiology, Movement, and the Role of Eye Tracking in Instructional Design." In *Navigating the World of Multimedia: Innovation and Applications*. IntechOpen. https://doi.org/10.5772/intechopen.1008301

Mayer, Richard E. 2020. *Multimedia Learning* (3rd ed.). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316941355

Montero Perez, Maribel, Elke Peters, and Piet Desmet. 2015. "Enhancing Vocabulary Learning through Captioned Video: An Eye-Tracking Study." *The Modern Language Journal* 99: 308–328. https://doi.org/10.1111/modl.12215

Montero Perez, Maribel, Elke Peters, and Piet Desmet. 2014. "Is Less More? Effectiveness and Perceived Usefulness of Keyword and Full Captioned Video for L2 Comprehension." *ReCALL* 26 (1): 21–43. https://doi.org/10.1017/S0958344013000256

Muñoz, Carmen. 2017. "The Role of Age and Proficiency in Subtitle Reading: An Eye-Tracking Study." *System* 67: 77–86. https://doi.org/10.1016/j.system.2017.04.015

Nushi, Musa, and Fereshte Orouji. 2020. "Investigating EFL Teachers' Views on Listening Difficulties among Their Learners: The Case of Iranian Context." *SAGE Open* 10 (2). https://doi.org/10.1177/2158244020917393

Park, Myongsu. 2004. "The Effects of Partial Captions on Korean EFL Learners' Listening Comprehension." PhD diss., *Dissertation Abstracts International, A: The Humanities and Social Sciences* 65 (4): 1287-A. (UMI Order No. DA3128868).

Rayner, Keith. 1998. “Eye Movements in Reading and Information Processing: 20 Years of Research.” *Psychological Bulletin* 124 (3): 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Rost, Michael. 2024. *Teaching and Researching Listening* (4th ed.). New York: Routledge. https://doi.org/10.4324/9781003390794

Sweller, John. 1988. “Cognitive Load during Problem Solving: Effects on Learning.” *Cognitive Science* 12 (2): 257–285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, John. 1999. *Instructional design in technical areas.* Camberwell, Australia: ACER Press.

Taylor, Gregory. 2005. “Perceived Processing Strategies of Students Watching Captioned Video.” *Foreign Language Annals* 38 (3): 422–427. https://doi.org/10.1111/j.1944-9720.2005.tb02228.x

Teng, Mark Feng. 2024. “Captioned Viewing for Language Learning: A Cognitive and Affective Model.” In *Innovations in Technologies for Language Teaching and Learning*, edited by Hung Phu Bui and Ehsan Namaziandost, Studies in Computational Intelligence 1159. Cham: Springer. https://doi.org/10.1007/978-3-031-63447-5_1

Teng, Mark Feng. 2022. “Vocabulary Learning through Videos: Captions, Advance-Organizer Strategy, and Their Combination.” *Computer Assisted Language Learning* 35 (3): 518–550. https://doi.org/10.1080/09588221.2020.1720253

Vanderplank, Robert. 2016. “‘Effects of’ and ‘Effects with’ Captions: How Exactly Does Watching a TV Programme with Same-Language Subtitles Make a Difference to Language Learners?” *Language Teaching* 49 (2): 235–250. https://doi.org/10.1017/S0261444813000207

Weingärtner, Henrike, Maximiliane Windl, Lewis L. Chuang, and Fiona Draxler. 2024. "Useful but Distracting: Viewer Experience with Keyword Highlights and Time-Synchronization in Captions for Language Learning." In *Proceedings of MUM 2024: The 23rd International Conference on Mobile and Ubiquitous Multimedia*, December 1–4, Stockholm, Sweden, 235–248. https://doi.org/10.1145/3701571.3701574

Winke, Paula M. 2013. "The Effects of Input Enhancement on Grammar Learning and Comprehension: A Modified Replication of Lee (2007) with Eye-Movement Data." *Studies in Second Language Acquisition* 35 (2): 323–352. https://doi.org/10.1017/S0272263112000903

Winke, Paula, Susan Gass, and Tetyana Sydorenko. 2013. "Factors Influencing the Use of Captions by Foreign Language Learners: An Eye-Tracking Study." *The Modern Language Journal* 97 (1): 254–275. https://doi.org/10.1111/j.1540-4781.2013.01432.x

Winke, Paula, Susan Gass, and Tetyana Sydorenko. 2010. "The Effects of Captioning Videos Used for Foreign Language Listening Activities." *Language Learning & Technology* 14: 66–87.

Wu, Huizhen, Ping Yu, Shenshen Yang, and Xuanyuan Chen. 2022. "Video Captioning Effects on EFL Listening Comprehension and Vocabulary Learning: Help or Hurdle?" *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)* 12 (2): 1–16. https://doi.org/10.4018/IJCALLT.291534

Yeldham, Michael. 2018. "Viewing L2 Captioned Videos: What's in It for the Listener?" *Computer Assisted Language Learning* 31 (4): 367–89. https://doi.org/10.1080/09588221.2017.1406956