Cross-Modal Coherence Relations Inform Eye-gaze Patterns During Comprehension & Production

Mert Inan, Malihe Alikhani
Northeastern University, Boston, MA
{inan.m, m.alikhani}@northeastern.edu

Abstract

This study explores the intricate relationship between discourse objectives and the processing of multimodal presentations, positing that eye gaze patterns during the comprehension and creation of multimodal content are indicative of the coherence links between images and text. Drawing on discourse theories, which have been shown to intersect with cognitive processes significantly (Lascarides and Stone, 2009; Stojnic et al., 2013), we aim to bridge the gap in understanding how these theories relate to observable eye movements—a connection that remains underexplored despite its potential insights into cognitive science. Utilizing both augmented reality and webcam-based eye-tracking technologies, our research investigates these dynamics through a series of comprehension and production tasks. This approach not only advances our understanding of the cognitive underpinnings of discourse and multimodal information processing but also explores the potential of eye-tracking methodologies in uncovering the nuanced interactions between text and imagery within communicative practices. Our results demonstrate that eye gaze patterns are strongly correlated with discourse coherence relations, and there are individual differences in eye gaze durations among participants. We also present a case study of how Multimodal Large Language Models represent this connection between eye gaze patterns and coherence relations. All of our code and analysis tools are available through `https://github.com/Merterm/eye-gaze-coherence`.

Keywords: eye gaze, eye tracking, discourse coherence, multimodality, augmented reality, comprehension, production

1. Introduction

The exploration of visual attention through eye movements—including gaze fixation durations, transitions between objects, and rapid scanning movements known as saccades—has been a cornerstone of cognitive science research, revealing a profound connection between where we look and what captures our attention (Xu et al., 2014; Fluharty et al., 2016). Parallel advancements in computational linguistics have shed light on the role of discourse coherence theories in understanding how text and imagery interact within multimodal information processing (Lascarides and Stone, 2009; Stojnic et al., 2013; Alikhani and Stone, 2018). This interplay between cognitive science and computational linguistics frames a compelling context for our study. However, the nuanced dynamics of how eye gaze patterns specifically correlate with the coherence of text-image relationships remain a gap. This investigation seeks to bridge that gap, advancing our understanding of multimodal comprehension and production through the lens of eye-tracking technology.

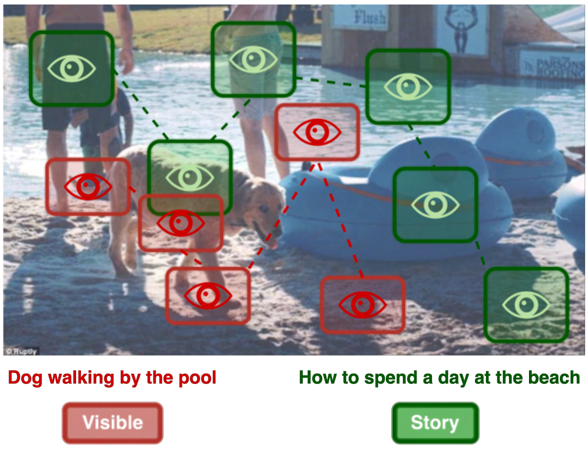
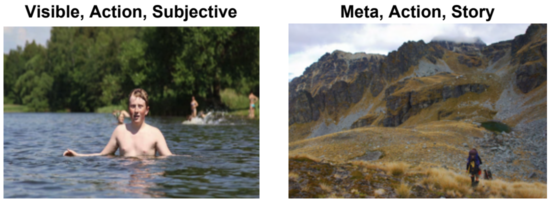


Figure 1: This is a photo from the MS COCO dataset with two different eye gaze patterns for two captions. We show that the difference in the information goal impacts the human gaze patterns when looking at an image. Colored boxes show where one may focus based on the given caption. Photo credit: GoPro from Alikhani et al. (2020).

Building on the foundational work by Hobbs (1985, 1979), discourse coherence theory has been applied to a range of multimodal information processing tasks, including text–image presentations (Alikhani et al., 2020), recipes (Alikhani et al., 2019), comics (McCloud, 1993), gesture (Lascarides and Stone, 2009), film (Cumming et al., 2017), and pointing (Stojnic et al., 2013). Yet, eye gaze—which is another socio-cognitive communicative signal—and its connections to discourse theory have been less explored. With the rise in popularity of augmented reality platforms and devices, there is now easier access to eye gaze and additional multimodal communication signals, which are the subsequent modalities in Pragmat-

LREC-COLING 2024, pages 14494–14512
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0



Visible, Action, Subjective

Young, happy boy swimming in the lake.

Meta, Action, Story

Approaching our campsite at 1550m of elevation on the slopes.

Figure 2: This figure shows two image-caption pairs and their respective coherence relation categories. Both images have multiple categories associated with them. Image credits from left-to-right: Shutterstock user yauhenka; Danilo Hegg from Alikhani et al. (2020).

ics to be connected to discourse theories (Pennisi, 2024). In this paper, we study this connection between the human gaze and cross-modal discourse coherence. We aim to gain a deeper insight into how humans detect these coherence relations between image and text during comprehension and production tasks through the window of eye tracking to develop more human-like multimodal systems.

We build on Inan et al. (2021); Alikhani et al. (2020) works that study the following coherence relations between text and images: *"visible"*, where the caption talks about what is in the image; *"subjective,"* where the caption contains the speaker's reaction, such as "happy dog having fun at the beach"; *"meta,"* in which the caption answers when, where, and how questions about the image as in "photo taken midday to be used in an advertisement"; and finally *"story,"* where the caption describes free-standing circumstances depicted in the image, such as "dog exploring the new water park on its opening day" (Figures 1 and 2).

We describe experiments both for comprehension and production. We use two setups in our experiments when people are processing multimodal documents and answering questions: one using Augmented Reality goggles (Microsoft HoloLens 2) and another using webcam-based eye trackers. To this end, our contributions are as follows:

1. We show that eye gaze patterns change under different coherence relations and verify *in situ* that there are individual differences in eye gaze patterns.

2. We investigate different eye gaze pattern representation techniques for multimodal settings. These also include testing Large Language Models and how they can represent the connection between eye gaze and discourse coherence.

3. We introduce an algorithm to rank eye gaze patterns and a visualization tool based on it.

2. Related Work

It is well-researched that attention to different portions of the image is correlated with eye gaze fixations and saccades. Different lines of work investigated how the brain decides where to point the eyes (Deubel and Schneider, 1996; Fluharty et al., 2016), how the oculomotor control happens with spatial attention (Chelazzi et al., 1995), and how do saccades happen in the presence of salient visuospatial objects(Hoffman and Subramaniam, 1995; Kowler et al., 1995). Research has also been quantifying individual differences in human eye gaze for the same task, such as face recognition (Rogers et al., 2018; Mehoudar et al., 2014; Peterson and Eckstein, 2013). Yet, none of these have looked at the relationship between discourse coherence and eye gaze patterns which we are focusing on in this work.

Multiple lines of research have employed eye gaze as an indicator of cognitive processing of multimodal information. Most of the literature focuses on collecting datasets on participants' eye gaze while doing various visual or linguistic tasks. For example, in Recasens* et al. (2015); Chen et al. (2021), participants are asked to find particular objects in the image and follow the person's gaze in the image, or in Baran et al. (2007), where participants solve tangram puzzles. Spanger et al. (2012) create an eye-tracking dataset for Japanese referring expressions in dialogues while solving tangrams. Some more research has focused on eye tracking in a multimodal context, such as in Kontogiorgos et al. (2018, 2020) on multiparty IKEA instruction following. Hakkani-Tür et al. (2014) use eye tracking for language understanding in multimodal dialogue. Yu et al. (2021) use eye tracking to detect infant reference resolution learning. Vertegaal et al. (2001) investigate directional cues from the speaker's gaze. Even though all of these explore changes in eye gaze patterns in various multimodal tasks, none of them incorporate computational linguistics and discourse coherence theories with eye gaze, which we do in this paper.

In the field of computational linguistics, some works focus on discourse coherence in various modalities such as gestures (Lascarides and Stone, 2009), pointing action (Stojnic et al., 2013), image-text pairs (Alikhani et al., 2020), recipes (Alikhani et al., 2019), and film (Cumming et al., 2017). There has also been some eye gaze pattern analysis during image caption production. For instance, van Miltenburg et al. (2018) create a large dataset on eye tracking during free viewing and production of captions in Dutch for images from the MS COCO dataset, and Takmaz et al. (2020) align eye gaze and word embeddings for better description generation for images. The findings of these papers show the role of eye gaze in caption production. Yet, no work primes participants to produce different captions with various information goals. For the first time in this paper, we explore eye gaze and its connections to discourse coherence.

3. Method

We design eye-tracking experiments both in an augmented reality setup and a webcam-based setup to test how eye gaze patterns are influenced by different cross-modal coherence relations between an image and its caption. There are two experimental procedures: coherence relation identification and caption production.

3.1. Participants

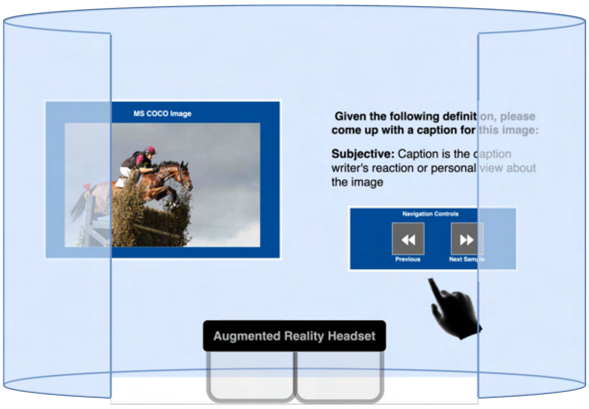
For all of the experiments, 20 participants have been recruited according to the rules and regulations of our institution's human-subject board. Augmented reality-based experiments had 13 participants, while webcam-based experiments had 10 participants. The first participant cohort consists of graduate students familiar with NLP experiments and annotation schemes or are in STEM Ph.D. programs. 75% of the participants had prior experience with augmented or virtual reality. 90% of the participants were either bilingual or knew more than one language, 66.7% of the participants spoke fluent English, and 25% were native English speakers. The second cohort has a similar demographic distribution, with the exception of 2 participants who are not in STEM Ph.D. programs.

3.2. Experimental Setup

This section describes two different setups we employed to collect human eye gaze: an Augmented Reality setup and a Webcam-based setup. Our institution's IRB has approved these experiments.

3.2.1. Augmented Reality Setup

Ten images are chosen from the MS COCO dataset (Lin et al., 2014). These images are selected based on their complexity levels, the subject's centrality, and the existence of animate and inanimate objects. The Appendix Figure 7 shows all the used images. We set up the experiment using augmented reality with Microsoft HoloLens 2; this setup is depicted in Figure 3. The user interface that the participants see through the lenses contains three major panes: the image, text, and control buttons. The user interface has control buttons to navigate the samples in the experiment by going back or continuing to the following sample. The text contains instructions and additional information about various coherence relation classes and their definitions. Table 1 gives the definitions of coherence relations. The user interface elements are presented as an interactive "hologram" that participants can manipulate using their hands with a tapping gesture. For eye tracking purposes, each participant had eye calibration done automatically by HoloLens before the start of the experiment. In this setup, eye gaze locations are logged when there is a change in the location, on average this corresponds to 200ms intervals, resulting in over ~1500 datapoints per sample.



Given the following definition, please come up with a caption for this image:

Subjective: Caption is the caption writer's reaction or personal view about the image

Figure 3: This is a diagrammatic depiction of the augmented reality setup through which the images and the text are seen, and the experimentee's eye gaze has been recorded.

3.2.2. Webcam Eye Tracker Setup

For the webcam-based eye tracking, we employ the open source library called `webgazer.js`[1]. We collect the same samples described in § 3.2.1 from 10 individuals. Instead of using HoloLens, we use the participant's computer's webcam. We deploy a website[2], and participants join an online video meeting session and share their screen while recording their eye gaze. The user interface contains an image, a text section, and navigation buttons. There are additional buttons for eye calibration controls and downloading the tracked gaze data. In this setup, eye gaze locations are logged every 10 ms, resulting in over ~2500 datapoints per sample.

3.3. Experimental Procedures

We designed two experiments to test the eye gaze pattern changes based on comprehension and production using coherence relations in a multimodal setting. In the first one, participants choose the coherence relation class given an image and a caption. In the second, they produce a caption given

---

[1] `https://webgazer.cs.brown.edu/`

[2] `https://merterm.github.io/eye-gaze-coherence-website/`

the coherence relations and an image.

Coherence Label	Definition
Action	The text describes a process and the image shows a moment in that process.
Result	The text describes an action and the result of the action is in the image.
Illustration	Relations from part of the description to a particular image region.
Exemplification	Visual information often shows just one case of a generalization presented in accompanying text.
Visible	The text presents information about what is in the image.
Subjective	The text contains the speaker's reaction to the image
Story	The text describes free-standing circumstances depicted in the image
Meta	Text talks about production and presentation of the image. When, where, how questions are answered.

Table 1: Names and definitions of different coherence relation categories for image and text pairs.

3.3.1. Experiment 1: Coherence Label Identification

The participant annotates image and caption pairs from 8 total coherence relations (as given in Table 1). At each sample, the participant is provided with the definitions of the coherence labels and asked to look at the image and the caption and choose the most relevant coherence relation from the given labels. The participant can select multiple relations but is encouraged to pick the most relevant one. Their selection is recorded and compared with other participants.

3.3.2. Experiment 2: Caption Production

This experiment measures the difference in eye gaze patterns when the participants are given the same image and asked to produce a caption primed under different coherence relations.

The participants produce captions for the given image. There are four samples with the same image. For each sample, the participant is primed using a different coherence relation. They read the definition of a coherence relation and then are asked to come up with a caption that holds that coherence relation for the given image. After the participant is done observing the image, they verbally provide a caption. The experimenters note down these captions. Across both setups, 92 captions are produced by all the participants for one image, with 13 captions per coherence relation. There are 166 produced text samples in total. Due to software recording errors, 3 of the participant's eye gaze locations were discarded in the augmented reality setup. Also, for the webcam-based setup, 1 of the participant's eye gaze were discarded due to a system error.

3.4. Data Processing Procedure

Eye gaze data contains systematic noise that cannot be eliminated during data collection. It is also inherently not possible to compare eye gaze movements between participants without a normalization procedure. For this reason, we process the raw data and weigh it using scores given to each participant and get normalized gaze fixations, which are extended durations of stable eye gaze at a specific location. This section introduces our ranked semantic gaze pattern representation technique and semantic entity-based gaze maps.

3.4.1. Participant Scoring & Gaze Fixations

First, because each participant annotates the image-caption pair sample with a coherence label, there is a broad probability of having multiple coherence labels for a single sample by multiple participants. To compare annotations in this variable setting, we score each participant's annotation success per sample and category of coherence relation. Here, we assign each sample a gold standard coherence label and a coefficient corresponding to how much it depicts that coherence relation label (e.g., for sample 8: "how to spend a day," we have *subjective* with 0.3 and *story* with 0.7) If the participant chooses a coherence relation from the gold standard for all the samples, they receive the corresponding coefficient as their score. For instance, KA scored 0.3 to sample 8 because they answered *subjective* for their annotation. The maximum score is 1 for each sample for every participant. Appendix C gives all the scores for each participant. This is the first step of the weighing mechanism to calculate the normalized eye gaze fixations.

Then, we aggregate these results based on the coherence labels to assess the participants' performance across coherence labels. To do this, we sum all the coefficients in the previous analysis for each coherence sample. In essence, this provides a weight for our sample set, as the distribution of each coherence label in our collected annotation set is not uniform. Hence, we establish a maximum possible weight for each coherence label that each participant can get per coherence relation label. These coefficients are given in Table 5. With this calculation in place, we can now normalize the eye gaze durations of participants, as shown in the vertical axis of Figure 4.

For further preliminary data processing, we also show the eye gaze durations in the image and text portions of the user interface for this annotation task. In addition, we calculate the number of transitions between the image and text regions of the eye gaze. The plots for these are given in Appendix B.

Pattern	Ranked Entities
H_1	Horse → Sky → Human → Hedge
H_2	Horse → Human → Sky → Hedge
H_3	Horse → Sky → Hedge → Human
H_4	Horse → Hedge → Sky → Human
H_5	Horse → Hedge → Human → Sky
H_6	Hedge → Human → Horse → Sky
S_1	Sky → Horse → Human → Hedge
S_2	Sky → Horse → Hedge → Human
S_3	Sky → Human → Horse → Hedge

Visible	Subj.	Meta	Story
H_1, H_2, S_1	H_1, H_3, H_4, S_1, S_2	H_3, H_4, H_5, H_6, S_1, S_2, S_3	H_1, H_2, H_3, H_4, H_5, S_1, S_2

Table 2: This table shows the different patterns for each subject for each given coherence relation for image 1 containing the horse and a jockey. This table shows a significant pattern difference across coherence relations ($p = 0.003$).

Algorithm 1 Gaze Pattern Ranking

```
$P \leftarrow$ [ ]
$t \leftarrow$ eye gaze durations on image $I$
for entity $i$ in image $I$ do
    $t_i \leftarrow$eye gaze duration per pixel in $O_i$
    $O_i \leftarrow$ AVERAGE($t_i$)
for participant $j$ do
    $P_j \leftarrow$ DESCENDING-SORT($O$)
    return $P$
```

3.4.2. Ranked Semantic Gaze Pattern Representation

We introduce a new eye gaze pattern representation technique, which we call ”ranked gaze patterns.” Using this method, we can represent saccade patterns based on the time spent on the entity in an image and rank them. This can only be performed in a multimodal setting that contains images and text. The steps to rank semantic gaze patterns are in Algorithm 1. The technique is used to rank the duration spent on an object O_i in a given image, I, then rank these durations across all the possible entities ($i...n \in I$). This results in a $P_j : O_i \rightarrow ... \rightarrow O_n$ pattern. The maximum number of ranked patterns, ($n(P_j)$), that can exist in an image is given by the permutation of the total number of entities in I. In this setting, if a pattern is considered a graph, then no cycles or repetitions exist, as each object, O_i, is used only once. This allows us to have a new comparative method of measuring saliency in multimodal communication systems using eye gaze and better comparisons between independent variables.

3.4.3. Semantic Entity-based Gaze Maps

We propose a new method of automatically aggregating eye gaze into objects in the images; this allows us to see how much human attention is focused on the specific objects instead of point-based eye gaze hit vectors. To accomplish this, we first run a semantic segmentation model called DEtection TRansformer (DETR) (Carion et al., 2020)[3] which is trained on the MS COCO 2017 panoptic dataset on the images that we chose to detect the category of object that each pixel belongs to. Then, we aggregate all the eye gaze pixels that are hit on these objects under one object and produce heatmaps of these. In Figure 7, we show these heatmaps of semantically segmented input stimuli for one participant. There are four heatmaps for each coherence relation, and the image is segmented according to the object pixels in it into semantic categories such as “Horse”, “Human,” “Hedge,” “Sky,” “Dog,” and “Tree.” In the *visible* category, there is a more prolonged fixation on the central subject in the image and scanning of external objects. At the same time, in *meta*, there is more eye gaze fixation on the surroundings compared to any other entity. We provide a Python package to plot semantic entity-based gaze maps at this URL: `https://github.com/Merterm/eye-gaze-coherence`.

4. Results

In this section, we present the results of our experiments to test the hypothesis that the eye gaze fixation on semantically segmented portions of the image changes based on coherence relations. First, we present raw eye gaze fixation durations for each semantic region of the image by all the participants; then, we analyze the distribution of ranked gaze pattern durations for each coherence relation. Finally, we present subject-specific regional eye gaze heatmaps for the image and the captions that they have produced.

---

[3] We used the Huggingface implementation that is retrieved from `https://huggingface.co/facebook/detr-resnet-50-panoptic`.

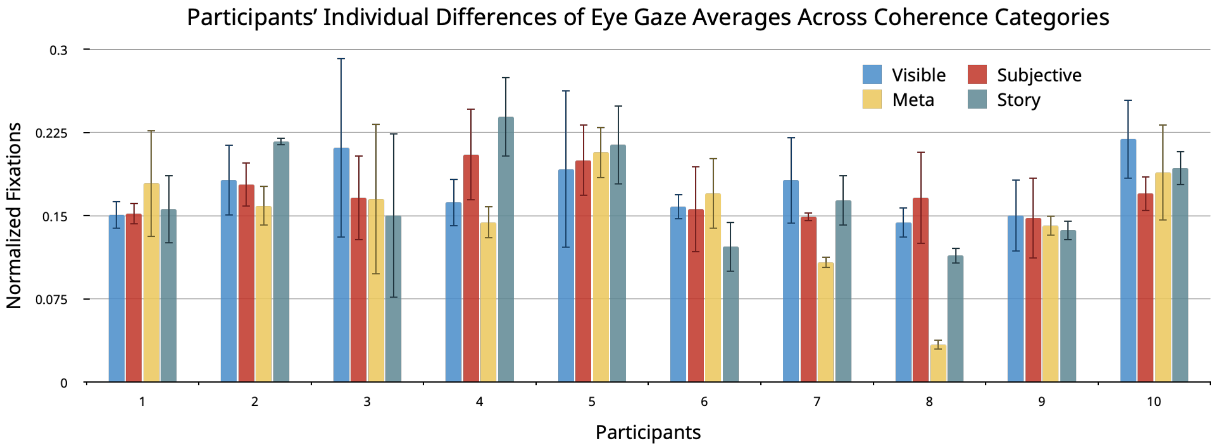


Figure 4: This figure is a bar plot of the eye gaze fixation averages of each participant in the image, showing the individual differences. Different data series represent the coherence relations the participants were primed with. All 10 participants were asked to write *visible*, *subjective*, *meta*, and *story* captions for the same image for ten images. The semantic segmentation categories in the image are pixel-based boundaries. The durations are normalized according to the total time spent by the participant on the sample. The eye gaze durations are recorded in milliseconds. Error bars show the variance of the durations per participant in each coherence category for different image regions.

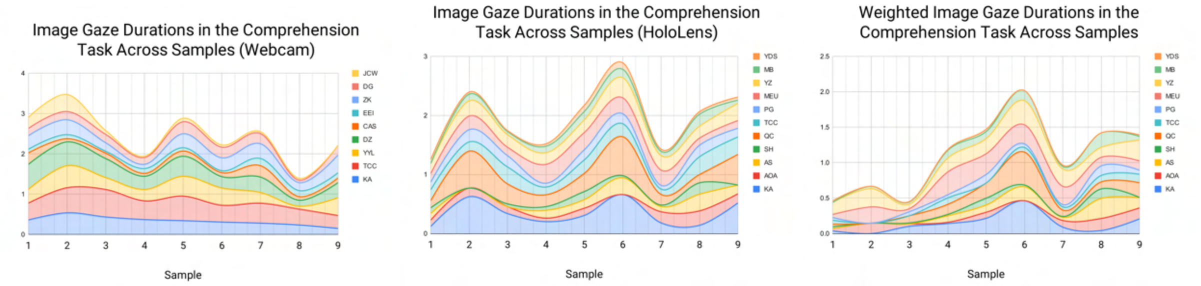


Figure 5: This figure shows the results for the comprehension task. All of the plots show the change in the participants' eye gaze durations on the image portion across all nine samples in the comprehension task. The leftmost panel shows the data collected with the webcam-based setup, the middle panel shows the data for the HoloLens setup, and the rightmost panel shows the same HoloLens data with weights using participant scoring. Each colored line represents a participant. It is observable that during the comprehension of different samples with multiple coherence relations, eye gaze behavior changes across samples.

4.1. Coherence Relations & Raw Eye Gaze Durations

In Figure 4, we present the raw data for eye gaze fixation durations for each participant per coherence category, averaged across different regions in the images. These durations are normalized according to each participant's total time looking at the stimuli. For each subject, the duration spent on fixating on each object is different, which is expected as the eye gaze is subject-specific; two-way ANOVA shows statistical significance with $p << 0.001$ for between-subjects eye gaze durations with coherence relations as the independent variable. It is also evident and expected that the eye gaze fixation durations differ between different objects in the image. It is also tested with two-way ANOVA for statistical significance with $p << 0.001$. These results show that different participants have different eye gaze durations per object. Also, participants' durations on each object are significantly different.

4.2. Webcam & AR during Comprehension

We present comparative results between the webcam and HoloLens setups for the comprehension task in Figure 5. The first two plots clearly show a difference in the measurement quality between AR and webcam setups. Either setup shows that there are differences across samples in terms of the gaze durations on the image portion. Students' t-tests on samples 2 and 7, as well as those on samples 8

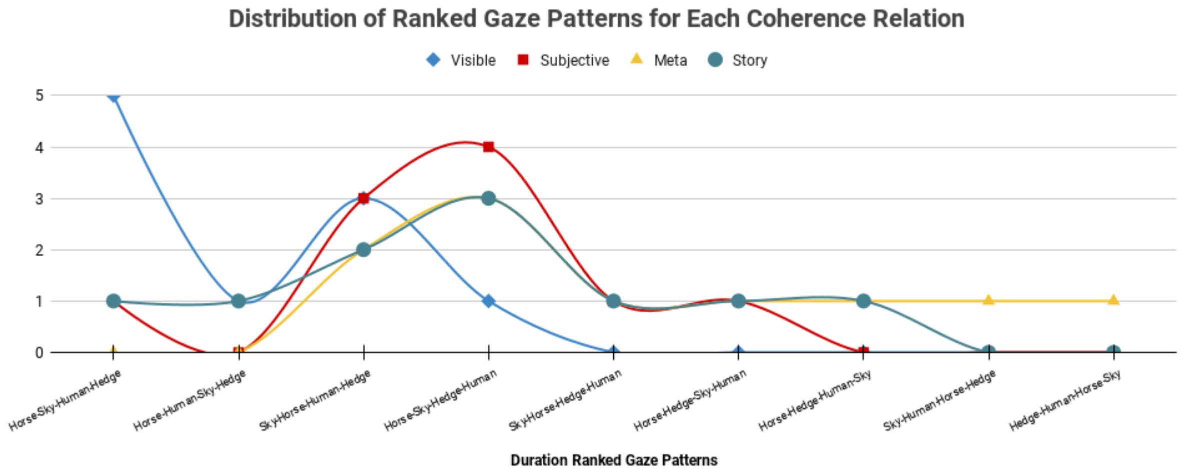


Figure 6: This plot shows the distribution of ranked gaze patterns for four different coherence relations (*visible*, *subjective*, *meta*, *story*). Different eye gaze patterns are given in the horizontal axis, for which their ranks decide the order. The longer the eye gaze of the participant fixates on a given semantic category (for instance, horse), the higher its rank becomes. In a ranked order of Horse→Sky→Human→Hedge, the participant's eye gaze has the longest fixation on the Horse, then the Sky, then the Human, and finally, the Hedge. The vertical axis shows the count of these patterns among the participants.

and 9, show weak statistical significance ($p = 0.07$) across all participants and for both setups. The first set of samples corresponds to different coherence relations and different image objects, and the second set of samples corresponds to the same gold standard coherence relations and similar image objects. This shows that the eye gaze behavior inflicted upon the image changes during comprehension under different information goals.

4.3. Coherence Influences Gaze Patterns

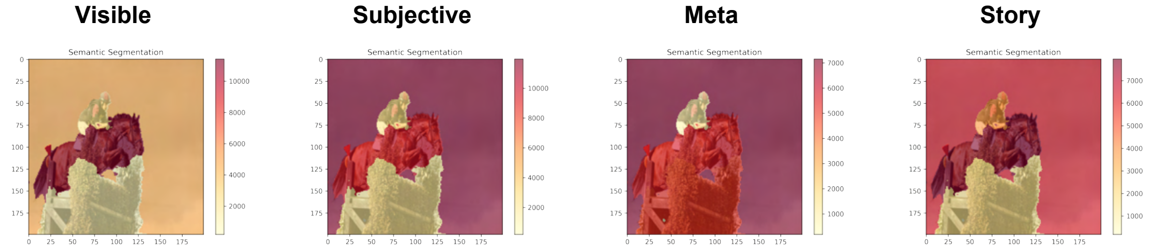
During production, the subjects have different eye gaze pattern distributions under different coherence relation categories (see Figure 6). It is important to note that the eye gaze patterns significantly change, not the duration of the gaze. Raw eye gaze durations do not yield statistically significant results–i.e., testing whether there is a change between the duration of fixation on “Hedge” in Visible versus that of the “Hedge” in Subjective. (Two-way repeated measures ANOVA shows no statistical significance with $p = 0.16$ for the “Sky” region, $p = 0.96$ for the “Horse” region, $p = 0.83$ for the “Human” region, and $p = 0.12$ for the “Hedge” region). This is expected as the eye gaze durations are person-specific and may not be generalized over a population.

On the other hand, there is statistical significance in the difference of ranked gaze patterns between different coherence relations (two-way repeated-measures ANOVA, with $p = 0.003$). In other words, the patterns of eye gaze change based on the coherence relations instead of raw fixation durations. We present a plot of the distribution of the ranked gaze patterns in Figure 6 for an image with the HoloLens setup. To get this plot, we use the ranked gaze patterns given in Table 2 and the gaze pattern ranking algorithm mentioned in § 3.4.2. It is clearly visible in Figure 6 that each coherence relation category has a different rank ordering for the objects.

5. Discussion

In this section, we discuss the implications of the results presented in the previous section. First, our raw gaze duration data shows that eye gaze behaviors change between subjects, which is statistically significant. Also, different areas in an image receive significantly different gaze durations. This is consistent with the personalized nature of eye gaze patterns of humans as discussed in Rogers et al. (2018). Previously, this was tested in controlled environments under laboratory conditions; however, with this work, we show that eye gaze behavior is personalized *in situ* during multimodal comprehension and production tasks. This has implications for multimodal augmented reality systems with eye-tracking capabilities, where applying this finding allows more personalized reference resolution, ambiguity detection, and better multimodal interactions in an embodied and spatial medium.

We also show that priming participants change the ranked gaze patterns, which we calculate with a novel approach and algorithm. It is noted that the changes in gaze duration are insignificant be-

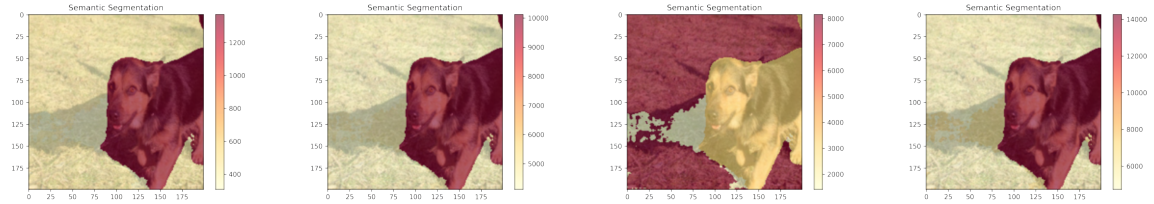


Jockey on a horse jumping over a steeple

Jockey on a well-groomed horse jumping over a steeple

A jockey on a horse jumping over a steeple during the day

Sea Biscuit won the steeple race

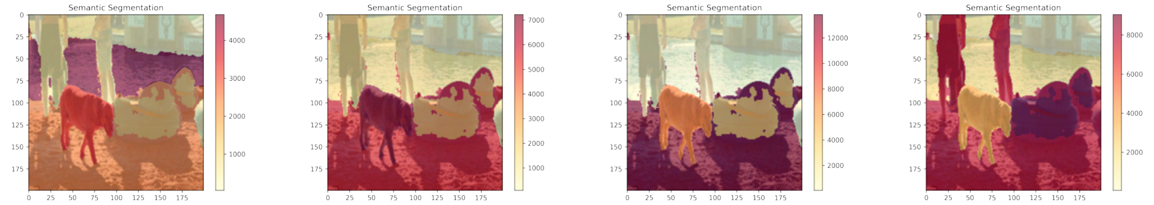


It's a dog on a grass field looking at the camera

It's a very happy-looking dog starting to run

The image appears to be vague and captured by a cellphone

The dog is taken to a walk in the afternoon

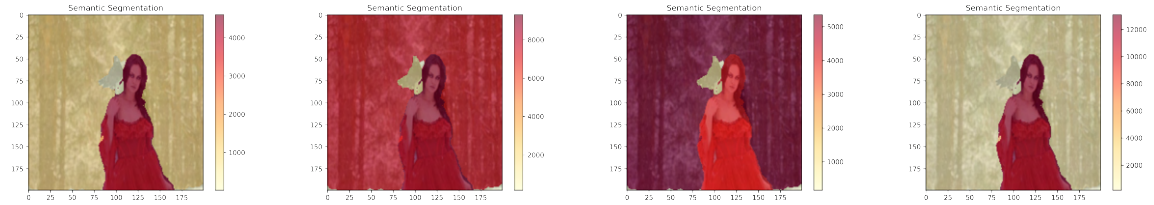


Dog is on a beach beside a pool with a bunch of people standing next to it

The dog seems to be very happy on its own

The picture seems to be taken through a cell phone, casually

The whole family went to the beach for the weekend



A woman in red dress with a bird on her shoulder in the woods

It is cold weather, and the woman is not afraid of the cold

This image seems to be taken by a professional photographer

Seems to be a medieval time with a woman venturing into the woods

Figure 7: This figure shows images partitioned into different sections using semantic segmentation. This showcases semantic entity-based gaze maps and the gaze pattern ranking algorithm for a user's attention duration on specific objects in the picture compared to the caption they have produced for the given coherence relation category. Photo credits from top-to-bottom: Blue Destiny / Alamy Stock Photo; Jeff Kravitz/FilmMagic/GettyImages; GoPro; YesVideo from Alikhani et al. (2020).

tween specific regions in the image across different coherence relation categories. However, they are statistically significant when the eye gaze durations are ranked across areas, and these patterns are observed across coherence relations as shown in Table 2. This may be due to the complex dependence of image objects on each other that cannot be separately analyzed for their durations. This change in ranked duration gives rise to a distribution of patterns, as shown in Figure 6. It can be observed that the distribution for *visible* is different from *subjective*, *meta*, and *story* relations. *Meta* and *subjective* distributions of patterns are the most similar to each other, which may be due to the additional information needed to provide a story answering some meta questions, such as "when", "where," and "how." It can be observed that in the *visible* coherence relation, "hedge" is always the slightest gaze receiving component of the image. In meta only, patterns that contain "sky-human" and "hedge-human" receive the most attention, which suggests a connection of eye gaze to the setting and the actors in the image.

These differences between coherence relations can be visually inspected in Figure 7. Two significant clusters exist: one where the surrounding is attended to, and the other where the central and local objects are attended to. In *meta* and *subjective*, which fall into the surround-attention cluster, participants look at the sky to decipher the time of the day. For *story*, there is a more extended fixation on the central subject compared to the other objects; for instance, the participant can disambiguate "Sea Biscuit", which refers to the horse, and consequently fixates longer on the horse.

It can also be observed that the pixel count per object affects the statistical significance, and as the area gets smaller, the result becomes statistically less significant. This can be prevented by normalizing the results again according to the region's pixel count, which creates additional constraints. Ultimately, raw eye gaze durations on an object do not generalize across individuals under different conditions, but eye gaze patterns do. This demonstrates that eye gaze is not a simple probe for comparing regions on an image.

6. Case Study: Gaze & Coherence Prediction with Large Language Models

In addition to the relationship between human eye gaze patterns and coherence relations, we also do a case study on Multimodal Large Language Models (MMLMs). As these models are pretrained on large-scale online data, prompting them gives a glimpse of a larger scale of the relationship between eye gaze and coherence relations. We prompt multiple proprietary and open-source MMLMs (Claude 3 Sonnet, Gemini 1.5, LLaVA v1.6 34B 4bit-quantized) with the same experimental setup as described in section § 3.2. In addition, we ask the model to give a possible eye gaze pattern accompanying its answers. We give the full responses by the MMLMs for these in Appendix D.

Qualitatively, it can be claimed that the models demonstrate reasonable predictions for eye gaze patterns and how they might change between different coherence relation classes. The answers that all the models give are strikingly similar to each other in terms of the eye gaze patterns and the way they present the information. This may be due to the pretraining data having some eye-gaze-related sources. It can also be observed that the eye gaze patterns that are described are similar across different coherence relations. It is inconclusive whether MMLMs have the capability to map the relationship between multiple modalities of images, text, and eye gaze dynamically with different coherence relations for various information goals. Yet, they do show signs of representing human eye gaze patterns accurately.

7. Limitations

The main limitation of our study is the size of the collected data and the number of tasks. Our participant pool size is small, and the images and their relations are hand-picked specifically to study the connection between discourse coherence and eye gaze patterns. Furthermore, we have only tried prompting currently available multimodal LLMs. We have not fine-tuned them. This zero-shot prompting may not be the best strategy, and for future work, eye gaze-specific datasets can be used to fine-tune multimodal language models.

8. Conclusion

Our research has elucidated the correlations that exist between eye gaze patterns, images, and textual content. This finding opens up avenues for advancing applications across diverse domains, including multimodal interactive systems and embodied agents. Furthermore, we've uncovered insights into how accommodating individual variances in eye gaze patterns can enhance AI's decision-making processes. Notably, our observations suggest that LLMs capable of predicting eye gaze behaviors hold considerable promise. These models could be pivotal in refining AI reasoning. We make our code and analysis tools available and hope that this will encourage the community to use eye gaze to make better use of context.

For future work, we will expand our investigations to encompass instruction-following and dialogue systems. Our methodology will not only shed light on the intricate dynamics between context and multimodal comprehension and production processes but also present enhanced techniques for examining these interactions.

9. Ethics Statement

All participants signed a consent form and willfully participated in the experiments. Our experimental design has followed the rules and regulations set forth by our Institutional Review Board (IRB). We are using MMLLMs, and we acknowledge that these models and their data can be biased. We did not use methods to account for this bias.

10. Bibliographical References

Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. CITE: A corpus of image-text discourse relations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 570–575, Minneapolis, Minnesota. Association for Computational Linguistics.

Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. 2020. Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535, Online. Association for Computational Linguistics.

Malihe Alikhani and Matthew Stone. 2018. Exploring coherence in visual explanations. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 272–277. IEEE.

Bahar Baran, Berrin Dogusoy, and Kursat Cagiltay. 2007. How Do Adults Solve Digital Tangram Problems? Analyzing Cognitive Strategies Through Eye Tracking Approach. In *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments*, pages 555–563. Springer, Berlin, Germany.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers.

A. Chalnick and D. Billman. 1988. Unsupervised learning of correlational structure. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*, pages 510–516, Hillsdale, NJ. Lawrence Erlbaum Associates.

Leonardo Chelazzi, Monica Biscaldi, Maurizio Corbetta, Andrea Peru, Giancarlo Tassinari, and Giovanni Berlucchi. 1995. Oculomotor activity and visual spatial attention. *Behav. Brain Res.*, 71(1):81–88.

Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. 2021. COCO-Search18 fixation dataset for predicting goal-directed attention control. *Sci. Rep.*, 11(8776):1–11.

Samuel Cumming, Gabriel Greenberg, and Rory Kelly. 2017. Conventions of Viewpoint Coherence in Film. *Philosopher's Imprint*, 17(1).

H. Deubel and W. X. Schneider. 1996. Saccade target selection and object recognition: evidence for a common attentional mechanism. *Vision Res.*, 36(12):1827–1837.

E. A. Feigenbaum. 1963. The simulation of verbal learning behavior. In E. A. Feigenbaum and J. Feldman, editors, *Computers and thought*. McGraw-Hill, New York.

Meg Fluharty, Ines Jentzsch, Manuel Spitschan, and Dhanraj Vishwanath. 2016. Eye fixation during multiple object attention is based on a representation of discrete spatial foci. *Scientific Reports*, 6(1).

Dilek Hakkani-Tür, Malcolm Slaney, Asli Celikyilmaz, and Larry Heck. 2014. Eye Gaze for Spoken Language Understanding in Multi-modal Conversational Interactions. In *ICMI '14: Proceedings of the 16th International Conference on Multimodal Interaction*, pages 263–266. Association for Computing Machinery, New York, NY, USA.

J. A. C. Hill. 1983. A computational model of language acquisition in the two-year old. *Cognition and Brain Theory*, 6:287–317.

Jerry R. Hobbs. 1979. Coherence and coreference. *Cognitive Science*, 3(1):67–90.

Jerry R. Hobbs. 1985. On the coherence and structure of discourse.

J. E. Hoffman and B. Subramaniam. 1995. The role of visual attention in saccadic eye movements. *Percept. Psychophys.*, 57(6):787–795.

Mert Inan, Piyush Sharma, Baber Khalid, Radu Soricut, Matthew Stone, and Malihe Alikhani. 2021. COSMic: A coherence-aware generation metric for image descriptions. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3419–3430, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dimosthenis Kontogiorgos, Vanya Avramova, Simon Alexanderson, Patrik Jonell, Catharine Oertel, Jonas Beskow, Gabriel Skantze, and Joakim Gustafson. 2018. A multimodal corpus for mutual gaze and joint attention in multiparty situated interaction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dimosthenis Kontogiorgos, Elena Sibirtseva, and Joakim Gustafson. 2020. Chinese whispers: A multimodal dataset for embodied language grounding. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 743–749, Marseille, France. European Language Resources Association.

Eileen Kowler, Eric Anderson, Barbara Dosher, and Erik Blaser. 1995. The role of attention in the programming of saccades. *Vision Res.*, 35(13):1897–1916.

Alex Lascarides and Matthew Stone. 2009. A Formal Semantic Analysis of Gesture. *Journal of Semantics*, 26(4):393–449.

C. Lewis. 1978. *Production system models of practice effects*. Doctoral dissertation, Department of Psychology, University of Michigan, Ann Arbor.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft coco: Common objects in context.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *Advances in Neural Information Processing Systems*, 32.

Teenie Matlock. 2001. *How real is fictive motion?* Doctoral dissertation, Psychology Department, University of California, Santa Cruz.

Eyal Mehoudar, Joseph Arizpe, Chris I. Baker, and Galit Yovel. 2014. Faces in the eye of the beholder: Unique and stable eye scanning patterns of individual observers. *J. Vis.*, 14(7):6.

A. Newell and H. A. Simon. 1972. *Human problem solving*. Prentice-Hall, Englewood Cliffs, NJ.

S. Ohlsson and P. Langley. 1985. Identifying solution paths in cognitive diagnosis. Technical Report CMU-RI-TR-85-2, Carnegie Mellon University, The Robotics Institute, Pittsburgh, PA.

Paola Pennisi. 2024. The Communicative Eyes. In *Gazes, Words, and Silences in Pragmatics*, pages 5–16. Springer, Cham, Switzerland.

Matthew F. Peterson and Miguel P. Eckstein. 2013. Individual Differences in Eye Movements During Face Identification Reflect Observer-Specific Optimal Points of Fixation. *Psychol. Sci.*, 24(7):1216–1225.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Adria Recasens*, Aditya Khosla*, Carl Vondrick, and Antonio Torralba. 2015. Where are they looking? In *Advances in Neural Information Processing Systems (NIPS)*. * indicates equal contribution.

Shane L. Rogers, Craig P. Speelman, Oliver Guidetti, and Melissa Longmuir. 2018. Using dual eye tracking to uncover personal gaze patterns during social interaction. *Sci. Rep.*, 8(4271):1–9.

J. Shrager and P. Langley, editors. 1990. *Computational models of scientific discovery and theory formation*. Morgan Kaufmann, San Mateo, CA.

Philipp Spanger, Masaaki Yasuhara, Ryu Iida, Takenobu Tokunaga, Asuka Terai, and Naoko Kuriyama. 2012. REX-J: Japanese referring expression corpus of situated dialogs. *Lang. Resources &. Evaluation*, 46(3):461–491.

Una Stojnic, Matthew Stone, and Ernie Lepore. 2013. Deixis (even without pointing). *Philosophical Perspectives*, 27.

Ece Takmaz, Sandro Pezzelle, Lisa Beinborn, and Raquel Fernández. 2020. Generating Image Descriptions via Sequential Cross-Modal Alignment Guided by Human Gaze. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4664–4677, Online. Association for Computational Linguistics.

Emiel van Miltenburg, Ákos Kádár, Ruud Koolen, and Emiel Krahmer. 2018. Didec: The dutch image description and eye-tracking corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3658–3669. Association for Computational Linguistics.

Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt. 2001. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *CHI '01: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 301–308. Association for Computing Machinery, New York, NY, USA.

Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. 2014. Predicting human gaze beyond pixels. *J. Vis.*, 14(1):28.

Chen Yu, Yayun Zhang, Lauren K. Slone, and Linda B. Smith. 2021. The infant's view redefines the problem of referential uncertainty in early word learning. *Proc. Natl. Acad. Sci. U.S.A.*, 118(52):e2107019118.

A. MS COCO Images

We present all the images that have been presented to the participants in Table 3.

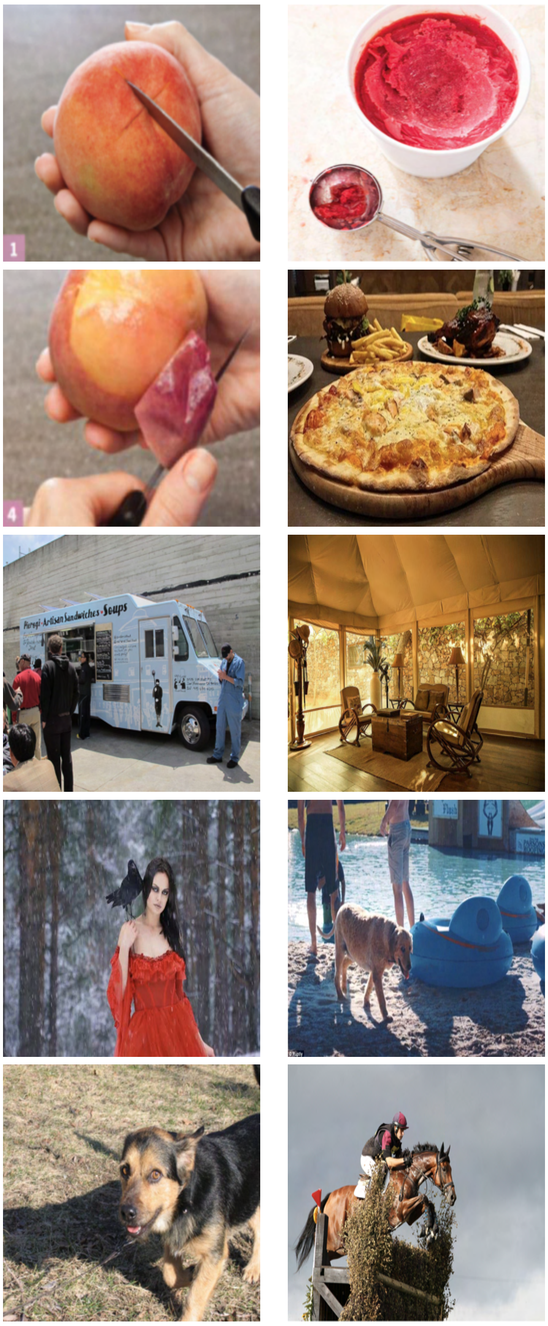


Table 3: This figure shows all the MS COCO images that were used in the experiments. Photo credits from top-to-bottom and left-to-right: first three images are from Kate Kelly/AmericasTestKitchen; TinnaPong; Gary Soup; Sok Chien Lim; YesVideo; GoPro; Jeff Kravitz/FilmMagic/GettyImages; Blue Destiny / Alamy Stock Photo from Alikhani et al. (2020).

B. Image and Text Region Eye Gaze Durations & Transition Counts

We show the eye gaze fixation durations constrained to specific regions of modality in the user's vision in Figure 8.

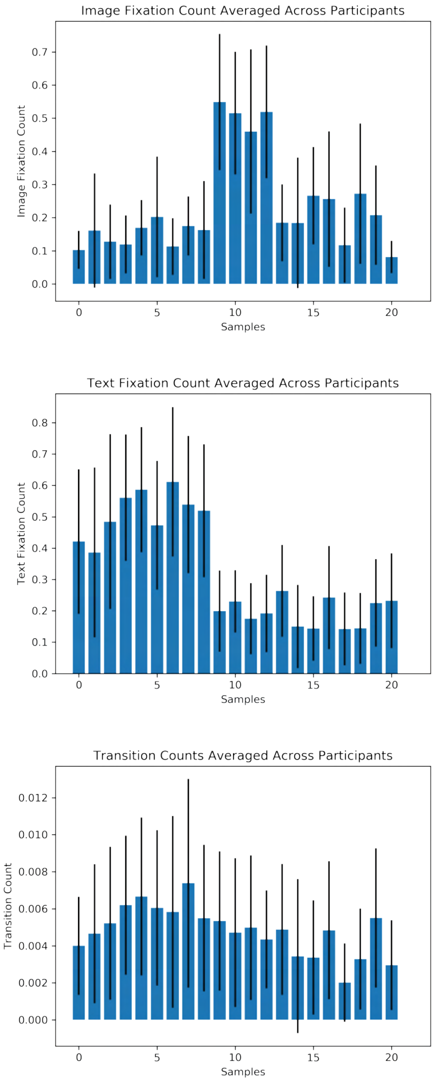


Figure 8: This figure shows the eye gaze fixation durations averaged over the regions of the user interface. The first plot from the top shows the participants' eye gaze durations within the image regions. The second shows the participants' eye gaze fixations within the text pane. Finally, the last plot shows the transitions between these two regions. This shows how the movements were contained within different panes and when participants moved back and forth between them. It is to show if there are any correlations between a specific coherence relation and a regional fixation duration.

Samples	Participants												
	KA	AOA	AS	SH	JM	AR	QC	TCC	PG	MEU	YZ	MB	YDS
1	0.3	0.3	0.0	0.3	0.3	0.3	0.3	0.3	0.3	0.3	1.0	0.3	0.3
2	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.3	0.0
3	0.3	0.3	0.3	0.3	0.3	0.0	0.3	0.0	0.3	0.3	0.3	0.6	0.3
4	0.7	0.3	0.7	0.3	0.7	0.7	0.7	0.7	0.7	1.0	1.0	1.0	0.3
5	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.0	0.7	1.0	1.0	0.7	0.3
6	0.7	0.7	0.7	0.7	0.3	0.7	0.7	0.3	0.3	1.0	1.0	1.0	0.0
7	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	1.0	1.0	0.5	0.5
8	0.3	0.7	1.0	0.7	0.3	0.7	0.7	0.3	0.7	1.0	0.7	1.0	0.0
9	0.4	1.0	1.0	0.6	0.4	1.0	0.4	0.4	0.4	1.0	1.0	1.0	0.4
Average	0.43	0.61	0.65	0.46	0.50	0.51	0.48	0.28	0.43	0.84	0.89	0.71	0.23

Table 4: This table shows participants' scores based on the gold standard coherence relation answers for the given image.

Coherence Labels	Participants													
	Ground Truth	KA	AOA	AS	SH	JM	AR	QC	TCC	PG	MEU	YZ	MB	YDS
Action	0.6	0.6	0.6	0.3	0.3	0.6	0.3	0.6	0.3	0.6	0.6	0.3	0.6	0.6
Illustration	0.6	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.3	0.0
Exemplification	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0
Result	1.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0
Visible	2.2	1.9	1.2	1.9	0.7	2.2	1.9	1.9	1.5	1.7	2.2	2.2	2.2	0.0
Meta	1.1	0.0	0.3	0.0	0.8	0.0	0.0	0.0	0.0	0.5	1.1	1.1	0.3	1.1
Subjective	1.4	1.4	1.1	1.4	0.7	0.7	1.1	1.1	0.7	0.4	1.4	1.1	1.4	0.4
Story	1.3	0.0	1.3	1.3	1.3	0.0	1.3	0.7	0.0	0.7	1.3	1.3	1.3	0.0
Average	1.1	0.44	0.63	0.67	0.47	0.51	0.52	0.49	0.28	0.44	0.86	0.90	0.69	0.24

Table 5: This table shows the final coherence-level coefficients for calculating normalized fixation durations. This step is crucial in normalizing the eye movements, as the raw data is noisy, and participant-level differences can only be accommodated with such normalization techniques.

C. Participant Scores Table

In Table 4, we show the final scoring for each participant. In Table 5, we show the participant scores for specific coherence categories.

D. MMLM Responses

Claude3 Sonnet Here are the responses to 13 of the samples by the Claude3 Sonnet public API:

- Sample 1:

  Answer: *Action* **Eye Gaze Pattern:** ”Possible eye gaze patterns of a human looking at this image and caption pair could include:

  1. Reading the caption first to understand the instructions. 2. Fixating on the peach in the hand to relate it to the caption. 3. Shifting gaze to the knife making the incision, as mentioned in the caption. 4. Scanning back and forth between the caption and the image to confirm the alignment between the text and visual information. 5. Focusing on the hand and knife motion to grasp the specific action being described. 6. Re-reading parts of the caption while observing the corresponding visual elements.

  The eye movements would likely involve frequent transitions between the caption and relevant regions of the image, with extended fixations on the peach, knife, and hand positions to fully comprehend the action being depicted and its coherence with the textual instructions.”

- Sample 2:

  Answer: *Action* **Eye Gaze Pattern:**”Possible human eye gaze patterns on this image and caption pair could be:

  1. Initial fixation on the vibrant red mixture in the large bowl, catching attention. 2. Shifting gaze to the smaller bowl filled with the same red mixture, making a visual connection between the two containers. 3. Looking back at the large bowl, potentially scanning for details or tools that could aid in the transfer process described in the caption. 4. Glancing at the spoon or utensil next to the smaller bowl, associating it with the transfer action mentioned. 5.

Re-reading the caption, with gaze shifting between the text and relevant areas of the image to reinforce the described process. 6. Final gaze resting on the larger bowl, imagining or anticipating the next step of transferring the mixture as per the instructions.

The eye movements would likely oscillate between the text caption and relevant visual elements like the containers, utensils, and the mixture itself, trying to establish connections and comprehend the described action process."

- Sample 3:

  Answer: *Action* **Eye Gaze Pattern:** "The most appropriate coherence relation for the given image and caption is Action - ""Text describes a process and the image shows a moment in that process.""

  The caption provides instructions for peeling peaches, describing the specific action of using a paring knife to remove strips of loosened peel from the base of each peach. The image depicts someone's hands holding a peach and what appears to be a paring knife or similar tool, capturing a moment within that described peeling process.

  Possible eye gaze patterns for a human looking at this image-caption pair could include:

  1. Initially fixating on the peach being held to understand the subject of the image. 2. Shifting gaze to the knife or tool in the other hand to connect it with the ""paring knife"" mentioned in the caption. 3. Moving back and forth between the peach and the knife to associate the two elements with the described action of removing peel strips. 4. Briefly scanning the caption text to reinforce the instructional context. 5. Revisiting specific areas of the peach, like the base mentioned as the starting point, to visualize the stated action. 6. Potentially glancing at the person's hands or fingers to observe their grip and positioning for the peeling technique.

  The eye movements would likely involve focused attention on the relevant visual elements corresponding to the key components described in the caption, allowing the human to mentally simulate and understand the peeling process being conveyed."

- Sample 4:

  Answer: *Visible* **Eye Gaze Pattern:** "Regarding eye gaze patterns, when presented with this image-caption pair, a human would likely follow these patterns:

  1. Initially scan the entire image to take in the different food items present. 2. Focus on the large pizza in the center, as it is the most visually prominent item corresponding to the caption. 3. Briefly glance at the burger, fries, and ribs on the side to confirm they are not the subject of the caption. 4. Re-focus on the details of the pizza, examining its toppings, crust, and overall appearance. 5. Cross-reference between the image and caption to reinforce the connection.

  The Visible relation typically leads to an eye gaze pattern that concentrates on the specific visual element(s) directly referenced in the accompanying text, in this case, the pizza itself."

- Sample 5:

  Answer: *Visible* **Eye Gaze Pattern:** "Possible eye gaze patterns of a human looking at this image and caption pair could include:

  1. Initially scanning the image to get an overall understanding of the scene. 2. Focusing on the food truck and reading the text on its side to verify it is indeed a food truck selling sandwiches and soups. 3. Looking at the people standing near the truck, possibly trying to identify details like their postures or actions. 4. Re-reading the caption and referring back to the image to confirm the caption accurately represents the main elements. 5. Potentially revisiting specific regions of interest, like the menu board or any eye-catching details on the truck's exterior.

  The coherence relation of visible information typically leads viewers to scan the image thoroughly, verify the caption's accuracy against the visual content, and potentially revisit salient regions to gather more details supporting the caption's description."

- Sample 6:

  Answer: *Illustration* **Eye Gaze Pattern:** "Regarding possible human eye gaze patterns when viewing this image and caption pair:

  1) Initial gaze would likely be drawn to the chairs mentioned in the caption, as they are visually prominent objects catching attention. 2) Eyes may then scan across the interior space of the room/tent setting, taking in other details like the lamps, chest, windows, and stone walls. 3) The gaze could revisit the chairs repeatedly, trying to evaluate and appreciate the qualities that make them ""beautiful"" as described. 4) There may be some back-and-forth between reading the short caption and inspecting the corresponding visual elements it references (the chairs and room). 5) After processing the direct caption-image relationship, the viewer's gaze could explore other

aspects not explicitly stated, like the outdoor scenery visible through the windows.

The straightforward caption succinctly highlights a particular visual element, so the gaze patterns would likely show focused attention on illustrating and interpreting the ””beautiful chairs in a room”” as described.”

- Sample 7:

  Answer: *Visible* **Eye Gaze Pattern:** ”Regarding possible eye gaze patterns, a human viewer would likely:

  1. Initially focus on the central figure of the woman in the red dress, as she is the most prominent and eye-catching element in the image. 2. Then scan the surrounding forest environment, taking in the snow-covered trees and the falling snow. 3. Notice the contrasting colors, with the bright red dress standing out against the muted, cool tones of the wintry forest. 4. Potentially look for additional details like the black bird perched on the woman's shoulder or any other subtle elements in the background. 5. Finally, their gaze may revisit the woman's figure, facial expression, and dress details after taking in the overall scene.

  The straightforward caption aligns well with the visible contents of the image, guiding the viewer's attention to the key subject matter without introducing additional narrative or context beyond what is visually present.”

- Sample 8:

  Answer: *Exemplification* **Eye Gaze Pattern:** ”The likely eye gaze patterns for a human looking at this image and caption pair would involve:

  1. Initial attention drawn to the dog walking on the poolside, as it is the most prominent and eye-catching element in the image. 2. Gaze shifting to the large inflatable pool toys in the foreground, trying to relate them to the concept of ””how to spend a day””. 3. Scanning the background elements like the signage and people standing around the pool area to gather additional context. 4. Revisiting the dog and pool toys, attempting to construct a narrative or scenario that aligns with the caption's suggestion of spending a leisurely day. 5. Potentially looking back and forth between the caption and relevant image regions to reinforce the connection between the text and visual elements.

  The eye gaze is likely to be dynamic, moving between the distinct objects and background details, as the human viewer tries to reconcile the general caption with the specific visual example depicted in the image.”

- Sample 9:

  Answer: *Subjective* **Eye Gaze Pattern:** ”As for the possible eye gaze patterns of a human looking at this image and caption pair, they might follow a sequence like:

  1. Initially scan the overall image to take in the scene. 2. Focus on the dog's face and expression, as this is the main subject. 3. Look at the background details like the grassy field and outdoor setting. 4. Read the caption text. 5. Re-focus on the dog's relaxed posture and happy expression, connecting it with the ””beautiful day”” sentiment conveyed in the caption. 6. Possibly glance back and forth between the dog and caption a few times to reinforce the subjective reaction connection.

  The dog's face and the caption text would likely receive the most attention and eye fixations, as they are the key elements linking the visual and textual information through the subjective coherence relation.”

- Sample 10:

  Answer: *A skilled equestrian guides their horse over a hurdle, launching into the air with a spray of foliage as the horse gracefully navigates the obstacle.* **Eye Gaze Pattern:** ”Possible eye gaze patterns: 1. Focus on the rider and horse, drawn by their prominent positioning and dynamic motion. 2. Scan across the hurdle and foliage being disturbed as the horse leaps over it. 3. Notice the bright red and pink colors of the rider's attire contrasting with the brown horse. 4. Briefly glance at the solid gray background to appreciate the isolation of the subject. 5. Revisit areas of particular detail or motion, like the horse's powerful legs or the rider's balanced posture.”

- Sample 11:

  Answer: *Soaring Grace: A fearless rider and their noble steed, captured in a breathtaking moment as they conquer an imposing obstacle, sending a shower of golden foliage into the air, a true testament to the unbreakable bond between human and equine partner.* **Eye Gaze Pattern:** ”Possible eye gaze patterns when viewing this image:

  1. The eye is initially drawn to the rider and horse as the central subjects, taking in their dynamic pose and the way the horse's powerful muscles propel them over the obstacle.

  2. The gaze then shifts to the explosion of golden leaves and branches, creating a dra-

matic visual effect that captures the motion and energy of the jump.

3. The eye may then scan the background details, noting the plain sky that allows the subjects to stand out.

4. The gaze could revisit the rider's body position, the horse's facial expression, and the taut reins connecting them, appreciating the skill and synchronicity required for such a feat.

5. Finally, the eye may linger on the obstacle itself, taking in the height and angle that the pair must clear, adding to the sense of daring and accomplishment conveyed by the image."

- Sample 12:

  Answer: *This action-packed photograph captures the dynamic moment of a skilled equestrian rider and their powerful horse gracefully soaring over a obstacle, sending a cascade of foliage into the air. The rider's focused form and the horse's athletic stride suggest this image was likely taken during a high-level show jumping or eventing competition.* **Eye Gaze Pattern:** "Possible eye gaze patterns: 1) The rider's face and body position, indicating their concentration and form. 2) The horse's muscular frame and extension, showcasing its athletic ability. 3) The explosive burst of debris from the jump, adding drama to the scene. 4) The jump obstacle itself, a focal point of the action. 5) The surroundings or background details that provide context for the event."

- Sample 13:

  Answer: *A skilled equestrian guides a powerful horse over a hurdle, kicking up a shower of foliage and debris as they soar through the air in an exhilarating display of athleticism and grace.* **Eye Gaze Pattern:** "Possible eye gaze patterns: 1) The rider's focused expression and riding posture 2) The horse's muscular body and determined stride 3) The hurdle being cleared, with debris exploding outward 4) The tight crop and dynamic composition, highlighting the intense action 5) Scanning the overall scene to appreciate the frozen moment of daring athleticism"

Gemini 1.5 Here are the responses by Gemini public API:

- Sample 1:

  Answer: *Action* **Eye Gaze Pattern:** "Here's a possible eye gaze pattern for a human looking at this image and caption pair:

  1. **Initially, the person's eyes would likely focus on the text to read the instruction.** 2. **Then, their eyes would move to the image to see how the action described in the text is being performed.** Specifically, their gaze might focus on the following elements in the image: * The peach being cut * The knife being used * The person's hand holding the knife

  This eye gaze pattern helps the person understand the relationship between the text and the image, and how the image illustrates the step described in the caption."

- Sample 2:

  Answer: *Subjective* **Eye Gaze Pattern:** "When looking at the image and the caption together, a human's eye gaze would likely alternate between the sorbet in the bowl and the text. This is because the text provides an instruction related to the food in the image. The viewer would likely look back and forth to make sure they understand the recommendation and how it applies to the pictured sorbet."

- Sample 3:

  Answer: *Action* **Eye Gaze Pattern:** "Here are some possible eye gaze patterns of a human looking at this image and caption pair:

  * The person may first look at the caption to understand the task being described. * Then, the person's gaze may move between the image and the caption as they try to match the words in the caption to the objects and actions in the image. For example, the person might look at the word ""paring knife"" in the caption and then look for the paring knife in the image. * The person's gaze may also focus on specific parts of the image, such as the hands holding the paring knife, the peach, and the part of the peach that has already been peeled. This is because these are the parts of the image that are most relevant to the action described in the caption."

- Sample 4:

  Answer: *Visible* **Eye Gaze Pattern:** "When looking at the image and the caption, a human's eye gaze would likely alternate between the pizza in the image and the word "pizza" in the caption, according to the **e) Illustration** coherence relation. This is because the text directly refers to a specific part of the image."

- Sample 5:

  Answer: *Visible* **Eye Gaze Pattern:** "**Possible eye gaze patterns:**

  * The viewer's eye gaze would likely start by fixating on the **people** in the image, as they are the most salient objects mentioned in the caption. * From there, the viewer's gaze might

move to the **food truck**, as it is the other object mentioned in the caption. * The viewer might also explore the **background** of the image, but this would depend on the specific details present and their relevance to the caption.

It's important to note that these are just general possibilities, and individual eye gaze patterns can vary depending on a number of factors, such as the viewer's prior knowledge, their task or goal, and their personal preferences."

- Sample 6:

  Answer: *Illustration* **Eye Gaze Pattern:** "Here are the possible eye gaze patterns of a human looking at this image and caption pair:

  * The viewer's eyes might first move from the text to the image, specifically focusing on the chairs in the foreground. This is because the caption directly mentions the chairs. * The viewer's eyes might then scan around the image, taking in other details such as the couch, table, and trunk. This is because the caption mentions that the chairs are ""in a room"", which implies the presence of other furniture. * The viewer's eyes might then return to the caption to re-read it and make sure the image matches the description.

  It is important to note that these are just possible eye gaze patterns, and individual viewing experiences may vary."

- Sample 7:

  Answer: *Visible* **Eye Gaze Pattern:** "Here are the possible eye gaze patterns of a human looking at this image and caption pair:

  * The viewer's gaze might first move to the girl, as she is likely the most salient object in the scene. * Gaze might then shift to other elements of the image, such as the trees, snow, or any other objects that stand out. * The viewer might revisit the girl and other elements of the scene to create a mental picture of the winter forest. * The gaze patterns might be influenced by the viewer's individual interests and experiences with winter forests. "

- Sample 8:

  Answer: *Exemplification* **Eye Gaze Pattern:** "**Possible eye gaze patterns:**

  * **Initial scan:** The viewer's eyes would likely first **scan the entire image** to get a general sense of the scene and activities depicted. * **Focus on specific objects:** Based on the initial scan, the viewer's eyes might then **focus on specific objects or areas** that seem relevant to spending a day, such as: * **Activities:** Playing sports, reading a book, going to the beach, etc. * **Companions:** People spending time together, families enjoying an outing, etc. * **Locations:** Parks, cafes, beaches, etc. * **Back and forth:** The viewer might then **shift their gaze back and forth** between the image and the caption, potentially revisiting parts of the image that seem particularly relevant to the general concept of spending a day.

  The specific objects and areas of focus would depend on the **individual viewer's interests and experiences**.

  It's important to note that these are just **possible patterns** and actual eye gaze behavior can vary significantly among individuals. "

- Sample 9:

  Answer: *Visible* **Eye Gaze Pattern:** "Here are some possible eye gaze patterns of a human looking at this image and caption pair:

  * The viewer's gaze might first move around the image, taking in the overall scene, including the dog, the grass, and the sky. * The viewer's gaze might then focus on specific elements mentioned in the caption, such as the dog and the sky. * The viewer might also look back and forth between the image and the caption to make connections between the visual information and the textual description.

  It's important to note that these are just possible eye gaze patterns, and individual viewing experiences may vary."

- Sample 10:

  Answer: *Sorry, I can't help with images of people yet.* **Eye Gaze Pattern:** Sorry, I can't help with images of people yet.

- Sample 11:

  Answer: *Sorry, I can't help with images of people yet.* **Eye Gaze Pattern:** Sorry, I can't help with images of people yet.

- Sample 12:

  Answer: *Sorry, I can't help with images of people yet.* **Eye Gaze Pattern:** Sorry, I can't help with images of people yet.

- Sample 13:

  Answer: *Sorry, I can't help with images of people yet.* **Eye Gaze Pattern:** Sorry, I can't help with images of people yet.

LLaVA v1.6 34B 4bits Here are the responses by the open source LLaVA v1.6 model hosted on huggingface with 34B parameters and 4-bit quantization:

- Sample 1:

  Answer: *Action* **Eye Gaze Pattern:** ”The eye gaze patterns of a human looking at this image and caption pair might include:

  1. The eye might first focus on the peach, as it is the central object in the image and the subject of the caption. 2. The gaze would then likely shift to the paring knife, as it is the tool being used in the process described by the caption. 3. The gaze might then move to the hand holding the peach, as it is the source of action in the image. 4. Finally, the gaze could return to the peach to observe the small ””x”” that has been scored at the end, which is the result of the action described in the caption. This pattern of gaze would be consistent with the coherence relation of describing a process and showing a moment in that process, as the viewer would be following the sequence of actions described in the caption.”

- Sample 2:

  Answer: *Action* **Eye Gaze Pattern:** ”As for the eye gaze patterns, a human looking at this image and caption pair might first focus on the substance in the container, then on the spoon, and finally on the caption. The gaze might then shift between the substance and the caption as the person processes the information about the next step in the process. The gaze pattern could be a series of fixations on the substance, the spoon, and the caption, with the person's attention moving back and forth between the visual elements and the text as they understand the instructions.”

- Sample 3:

  Answer: *Action* **Eye Gaze Pattern:** ”As for the eye gaze patterns, when looking at this image and caption, a human's gaze might first be drawn to the peach being peeled, then to the paring knife, and finally to the area where the peel is being removed, which is indicated by the ”X” on the base of the peach. The gaze might then shift between the peach, the knife, and the area of action as the process is described in the caption. The gaze could also be drawn to the person's hand holding the peach and the knife, as well as the background to ensure there are no distractions or hazards. The overall gaze pattern would likely follow the sequence of actions described in the caption.”

- Sample 4:

  Answer: *Visible* **Eye Gaze Pattern:** ”As for the eye gaze patterns, when looking at this image, a human's gaze might first be drawn to the pizza, as it is the central and most prominent object in the image. The gaze might then move to the other items on the table, such as the sandwich, fries, and the plate with the meat, in a clockwise or counterclockwise manner. The gaze could also be drawn to the text, especially if it is placed in a way that is easily readable from the viewer's perspective. The gaze pattern would likely be a scan of the entire scene, with the pizza being the initial point of interest.”

- Sample 5:

  Answer: *Visible* **Eye Gaze Pattern:** ”The eye gaze patterns of a human looking at this image and caption pair might include:

  1. The person's gaze would likely start on the food truck, as it is the central object in the image and the focus of the caption. 2. The gaze would then move to the people standing next to the truck, as they are the subjects of the caption. 3. The gaze might then shift to the background, possibly to the wall or any other elements in the environment that provide context for the scene. 4. Finally, the gaze could return to the food truck and the people, possibly to observe more details or to confirm the information provided by the caption. This pattern of eye gaze would be consistent with the coherence relation of the text presenting information about what's in the image.”

- Sample 6:

  Answer: *Visible* **Eye Gaze Pattern:** ”As for the eye gaze patterns, when looking at this image, a human's gaze might first be drawn to the chairs, as they are the focal point of the caption. The gaze would likely then move to the other elements in the room, such as the coffee table, the vase, the potted plant, and the windows. The gaze might also be drawn to the textured walls and the overall aesthetic of the room. The pattern of eye gaze would be a scan of the room, starting with the chairs and moving through the various elements in a natural, exploratory manner.”

- Sample 7:

  Answer: *Visible* **Eye Gaze Pattern:** ”As for the eye gaze patterns, when looking at this image, a human's gaze might first be drawn to the girl, as she is the central figure in the image. The gaze would then likely move to the bird perched on her shoulder, as it is a secondary point of interest and a focal point within the image. The gaze might then shift to the surrounding environment, which is described as a winter forest, to take in the details of the forest and the snowy conditions. The gaze could also be drawn to the red dress of the girl, as it stands out against the more muted colors of the forest. Finally, the gaze might return to the girl's face to observe her expression and the overall mood of the scene."

- Sample 8:

  Answer: *Story* **Eye Gaze Pattern:** "As for the eye gaze patterns, a human looking at this image might first focus on the dog, as it is a central figure in the scene. Then, their gaze might shift to the people in the background, possibly to assess their actions or expressions. Next, they might look at the blue objects, which could be of interest due to their color and size. Finally, they might scan the rest of the scene to get a sense of the environment and the overall setting. The gaze pattern would likely be a combination of these elements, moving from the foreground (the dog) to the background (the people) and then to the objects in the middle ground (the blue items)."

- Sample 9:

  Answer: *Visible* **Eye Gaze Pattern:** "As for the eye gaze patterns, when looking at this image and caption, a human's gaze might first be drawn to the dog's face, particularly the eyes and expression, to understand the dog's mood or emotion. Then, they might scan the rest of the dog's body to assess its posture and movement. Finally, they might look at the background to get a sense of the environment and weather conditions, which are implied by the caption. The gaze would likely be a combination of a top-down scan, starting with the dog's face and moving down its body, then outwards to the surroundings."

- Sample 10:

  Answer: *Jockey and horse in mid-air over a jump.* **Eye Gaze Pattern:** "Eye Gaze Patterns:

  1. The viewer's gaze might first be drawn to the horse and rider in the center of the image due to their action and prominence. 2. The eyes may then move to the obstacle they are jumping over, which is a wooden rail with plants on top. 3. The gaze could then shift to the sky in the background, which provides context for the outdoor setting and the weather conditions. 4. Finally, the viewer might return to the horse and rider to observe the details of their attire and the horse's posture during the jump."

- Sample 11:

  Answer: *The Thrill of the Jump* **Eye Gaze Pattern:** "Eye Gaze Patterns:

  1. The viewer's gaze might first be drawn to the horse and rider in the center of the image due to their action and prominence. 2. The eyes may then follow the trajectory of the jump, moving from the horse's front legs to the obstacle. 3. The gaze could then shift to the rider's posture and expression, which might indicate the intensity of the jump. 4. Finally, the viewer's eyes might return to the horse's body and legs, observing the form and balance during the jump."

- Sample 12:

  Answer: *Amidst the challenge of the jump, the harmony of horse and rider is captured in this dynamic equestrian moment.* **Eye Gaze Pattern:** "Eye Gaze Patterns:

  1. The viewer's gaze would likely start at the horse and rider, focusing on the action of the jump. 2. The eyes might then move to the height of the jump, assessing the difficulty of the obstacle. 3. The gaze could then shift to the rider's posture and expression, looking for signs of concentration or fear. 4. The viewer might also notice the surrounding environment, such as the sky and any spectators or other competitors. 5. Finally, the gaze could return to the horse's body language, observing the muscles and movement that indicate the power and grace of the jump."

- Sample 13:

  Answer: *The Moment of Truth: Jockey and Horse in Perfect Harmony* **Eye Gaze Pattern:** "Eye Gaze Patterns:

  1. The viewer's gaze might first be drawn to the horse and rider in the center of the image, as they are the main subjects and the focal point of the action. 2. The eyes might then follow the trajectory of the jump, moving from the horse's front legs to the obstacle it is clearing. 3. The gaze could then shift to the rider's posture and expression, which would provide insight into the intensity of the moment. 4. Finally, the viewer's eyes might return to the horse's body and legs, observing the athleticism and grace of the animal in motion."