# An Experimental Investigation of Relational Categorization and Visual Relational Processing

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy
in the Graduate School of The Ohio State University

By

Yuhui Du, M.A.

Graduate Program in Psychology

The Ohio State University

2023

Dissertation Committee

Alexander A. Petrov, Advisor

Julie D. Golomb

John E. Hummel

Andrew B. Leber

# Abstract

Relational processing plays a central role in human cognition. An important theoretical goal of cognitive psychologists is to understand the learning mechanisms and representations that support explicit relational processing in human minds. With this overarching goal, I conduced four behavioral experiments to shed light on two specific questions: (i) the representational format of relation-defined categories, and (ii) whether the visual system (as contrasted with central cognition) constructs and manipulates explicit relational representations. Experiments 1 and 2, reported in Chapter 2, were designed to investigate the first question. Experiments 3 and 4, reported in Chapters 3 and 4, were designed to investigate the second question. Chapter 2 reports two category learning experiments in which participants were trained on a fictional medical diagnosis task involving relation- and feature-defined categories. A post-training reconstruction task was also included to probe the resulting category representations. Both correlational and causal evidence was found supportive of the *extreme-value hypothesis*: only participants who learned relation-defined categories exaggerate values away from the trained means as reconstructed members. The finding contributes to the growing literature that, unlike feature-defined categories, relation-defined categories are represented by their extreme members. In Chapter 3, I used a continuous carryover presentation design with visual stimuli depicting comparative relations. An inhibitory effect of response time upon the switch (versus repetition) of relational roles was observed across consecutive trials. In Chapter 4, I used a priming paradigm with visual stimuli depicting eventive relations.

The response times showed various priming effects as a function of role switching and/or role repetitions relative to a control condition. The results of Experiments 3 and 4 suggest the human visual system is sensitive to abstract relational roles. Taken together, the Thesis provides empirical evidence that sheds light on the mechanisms for relational processing in both vision and central cognition.

# Table of Contents

# Chapter 4 Seeing Eventive Relations

## Introduction

While Experiment 3 explored perceiving *comparative* relations between pairs of objects, Experiment 4 focused on *eventive* relations between agents and patients. Published studies provided evidence that the human visual system not only encodes relations (e.g., Hafri et al., 2013; Stankiewicz & Hummel, 2002), but encodes abstract relational roles (e.g., Hafri et al., 2018). Nevertheless, these authors also agreed that their abstract relational roles might be correlated with certain typical visuo-spatial featural properties, such as head and body orientation, or whether the limbs of one human were outstretched in the direction of the other human. In the experiment of Hafri et al. (2013), the accuracy of identifying either agent or patient was indeed lower when patients also had typical agent-like features. This concern may also challenge the finding of abstract relational roles in vision reported by Hafri et al. (2018). Although these authors interpreted the observed role switch cost in terms of processing abstract roles in vision, the cost might be explained at least in part in terms of relatively low-level visual features instead of abstract roles. To extend these published studies, we conducted an experiment that complements the design of Hafri et al. (2018) in the following sense: In the published study, the human figures that played the Patient role in each interacting pair looked similar across all event types (e.g., punching, pulling, and tickling). It was the Agent who assumed characteristic poses with outstretched limbs. By contrast, in our Experiment 4, the animals that play the Agent role look the same across event types (e.g., breaking or deforming), whereas the Patients undergo characteristic deformations such as breaking in half or bending their hind legs.

More importantly, Experiment 4 aimed to explore whether visual system is sensitive to *abstract* relational roles – that is, roles that are encoded in some uniform representational format that is shared across situations that have minimal overlap of these lower-level properties. In other words, we set out to explore whether the difference between response time in trials with repeated and switched roles, referred as *role switch cost* by Hafri et al. (2018), would persist when no featural properties were perfectly correlated with abstract roles.

Three eventive relations were used in Experiment 4: *breaking*, *deforming*, and *launching*. For *deform* events, there were a few patient-like features that can be extracted from a still image, although they were not as obvious as the features in *break* events. For *launch* events, however, no patient-like features were available, and this made it impossible to differentiate the roles of such an event from a still image. By contrast, short video clips allowed easy identification of both relational roles (Agent and Patient) for all event types. One key idea in Experiment 4 was to induce priming effects by means of short videos instead of still images. This is one of the biggest differences of Experiment 4 from Experiment 3 and published experiments.

Experiment 4 also aimed to examine the role of selective attention in perceiving relational roles. Although the published literature emphasized that spatial attention directing towards to the objects was necessary to explicitly encode relations among parts within the objects, it was not clear if selective attention towards the particular relation-relevant dimension was necessary. Previous experiments by Hafri and colleagues, as well as Experiment 3, suggested that the selective attention might not be necessary. Clevenger and Hummel (2014) also proposed that multiple relations should be “stacked” onto pairs of objects, as their empirical evidence supported the model with fixed proposition (e.g., above-and-larger (2, 1)) better than another with separate propositions (e.g., above (2, 1) and larger (2, 1)). Therefore, selective attention might not be necessary when attention was directed to a separate dimension. Nevertheless, previous experiments either focused on more concrete comparative relations or their results could be explained in terms of lower-level featural properties. Experiment 4 might shed light on the requirement of selective attention in perceiving abstract relational roles.

Furthermore, in order to distinguish visual perception of relational roles from post-perceptual processing, Experiment 4 induced priming effects by means of short sentences in a subsequent complementary task. In this task the relational roles were primed in linguistic format, which mainly relied on post-perceptual processing. Psycholinguistic studies have confirmed that thematic roles (e.g., *agent*, *goal*, or *patient*) can prime subsequent utterances (e.g., Chang et al., 2003; Hare & Goldberg, 1999). For example, Chang and colleagues (2003) found that participants were more likely to incorrectly repeat the sentence “the farmer heaped straw onto the wagon” (*theme-location*) as “the farmer heaped the wagon with straw” (*location-theme*) after being primed with another *location-theme* sentence (e.g., “the maid rubbed the table with polish”). Despite its effect on following utterances, however, thematic roles conveyed by sentences were not expected to prime visual relational roles. Therefore, we only predicted the role switch cost in the *primary task* that used videos as primes, but no role switch cost in the *complementary task* that used sentences as primes.

To summarize, Experiment 4 extended and complemented the study by Hafri et al. (2018). In comparison to previous experiments, Experiment 4 was different in three important ways. First, we changed the format of primes from still images to videos so that neither typical agent- or patient-like featural properties nor temporal information (e.g., who moved first) were fully correlated with the abstract roles. Therefore, if we could replicate the role switch cost of RTs, it would provide a strong support to the claim that vision encodes abstract relational roles. Second, we included neutral conditions in which no eventive relations were available. By comparing the RTs of role repeated and switched trials to these neutral trials, we could distinguish role switch cost from role repeat benefit. Third, we included a subsequent isomorphic task in which abstract roles were primed by sentences. By comparing the results of tasks that only differed in the format of primes, we could further explore whether the potential role switch cost was predominately due to the visual system instead of post-perceptual processes.

Overall, we predicted (i) longer RTs in both role repeated and switched trials as compared to the neutral trials where eventive relational roles were neither repeated nor switched; (ii) longer RTs in role switched trials as compared to role repeated trials, as referred to as the role switch cost; and (iii) any effects produced by repeated and/or switched relational roles only persisted in the primary task where abstract roles were primed by videos, while no differences were expected in the subsequent task where abstract roles were primed by sentences.

Methods

*Participants.* Printed flyers were posted on bulletin boards throughout the Psychology Department and an electronic announcement was posted on the Research Experience

Program website of the Ohio State University (OSU). 23 members of the OSU community participated in the experiment in person and received $15 per hour for their participation. The experiment was divided into two sessions, one session per day, for participants to keep focused. All participants signed an informed consent form at the beginning of their first session. In almost all cases, the second session was scheduled within one week from the first session. On average, the first session took 60 minutes, and the second session took 80 minutes. The study was approved by the Ohio State University Behavioral and Social Sciences Institutional Review Board.

Before recruiting participants for Experiment 4, we had conducted a pilot experiment with 4 participants. We ran a power analysis based on the effect observed in pilot data using R simr package (Green & MacLeod, 2016). Power analysis results suggested a durable 80% replication rate of the effects (as recommended in Brysbaert & Stevens, 2018) could be achieved with 4 times of the number of participants included in the pilot study. Therefore, we planned to recruit 20 participants in Experiment 4 and aimed to have 16 good participants who passed all sanity checks. As described below, we did recruit 23 participants and had 21 of them completed all tasks. However, only 15 of them passed the inclusion criteria and we were unable to collect more data due to disruption and noise caused by construction work on the roof above the lab. Therefore, we acknowledge that the results in Experiment 4 are preliminary as they come from an incomplete sample with fewer participants than the number we planned for based on our power analysis.

Among the 23 participants who completed the first session, two did not show up for the second session and were dropped from the study. Thus, 21 participants completed both sessions. The *foreground color* for the visual localization task was manipulated between subjects – 12 of them were instructed to search for the red and 9 for the blue animal in each pair. Also, the experimental design used six alternative *templates (or conditions)* for counterbalancing stimuli across trials, as described in Table 4.1 below. These templates were randomly assigned between subjects -- with 3, 4, 6, 1, 1, and 6 participants in conditions 1 through 6, respectively.

We applied a series of pre-determined exclusion criteria that eliminated 6 participants from the sample. Concretely, according to the first pre-determined criterion, participants were included in the main analysis only if they achieved at least 85% accuracy in both sessions. One participant was eliminated for having chance-level (49.3%) accuracy in the first session, and 2 participants were eliminated for having chance-level (49.1% and 49.3%) accuracy in the second session. The second exclusion criterion was based on the accuracy on a series of catch questions that were included in the trial sequence to examine whether participants encoded events correctly. Specifically, we dropped three more participants: one had chance-level accuracy (26.4%) on catch questions, whereas two others had relatively low accuracy on both catch questions (61.1% and 47.2%) and main trials (87.2% and 87.4%). Overall, 15 valid participants remained in our main analysis. Eight of them searched for red and 7 for blue targets, and the six stimulus counterbalancing templates were assigned to 3, 2, 6, 0, 1, and 3 participants, respectively.

Table 4.1 Templates for Counterbalancing Stimulus Materials Across Participants in Experiment 4
There are six templates defined by the six blocks of the table below. Each participant was assigned randomly to one of these templates. The letters M, H, L, and G stand for "mouse", "horse", "lion", and "goat", respectively. Each scene contains two animals from the same species. The rows specify the animals in the priming video (in the primary task) or sentence (in the complementary task); the columns specify the animals in the subsequent still image. Each template samples trials among the cell marked by an X and ignores the other combinations.

| | 1 | Still image | | | | 2 | Still image | | | | 3 | Still image | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | M | H | L | G | | M | H | L | G | | M | H | L | G |
| Prime | M | X | | | | M | X | | | | M | X | | | |
| | H | | X | | | H | | | | X | H | | | X | |
| | L | | | | X | L | | | X | | L | | X | | |
| | G | | | X | | G | | X | | | G | | | | X |
| | 4 | Still image | | | | 5 | Still image | | | | 6 | Still image | | | |
| | | M | H | L | G | | M | H | L | G | | M | H | L | G |
| Prime | M | | | | X | M | | | X | | M | | X | | |
| | H | | X | | | H | | X | | | H | X | | | |
| | L | | | X | | L | X | | | | L | | | X | |
| | G | X | | | | G | | | | X | G | | | | X |

*Stimuli and Procedure of the Threshold Estimation and Primary Task.* As described above, Experiment 4 was conducted in two sessions. The first session included the threshold estimation task and the primary visual-priming task. The stimuli for these two tasks were 64 short videos (~1400 ms each) depicting scenes with one of three eventive relations (*breaking, deforming,* and *launching*) or a perceptual control (*dancing*). The videos were generated in an open-source animation software called Blender (Blender Foundation, 2022). Each scene involved a pair of twin animals in different colors: *mouse-mouse*, *horse-horse*, *lion-lion*, and *goat-goat*. For each unique scene (16 in total = 4 events x 4 actor pairs), the color (red or blue) and the position (left or right, see Fig. 4.1) of agent were fully counterbalanced, resulting in 64 videos in total. The two animals always faced each other. In videos depicting eventive relations, one of the animals moved from either the left or the right side of the screen to the center. The agent – that is, the animal that moves first – *launched*, *deformed,* or *broke* the other animal (the patient). The fourth scene type served as a perceptual control. In it, the two animals were not involved in any well-defined relations, they simply moved (or "danced") around their assigned positions at the center of the screen (see Fig. 4.1). As discussed earlier, the use of videos made it possible to explore abstract relations. However, even "short" videos left plenty of time for post-perceptual processing. The aim of this study was to use rapid, masked visual presentation to isolate perceptual from post-perceptual processing. Therefore, instead of implementing a video version of the carryover sequence used in Experiment 3, we only used videos as primes in the primary task. The primary task – and the associated RT data – was performed on static *target images* presented after the priming-inducing videos. On most trials, the target image could have been the final frame of some video sequence (as shown in the second column in Fig. 4.1). A static *visual mask* was constructed by overlaying fragments of images involving animals (e.g., cat, dog) that are not among the main stimuli. Furthermore, the duration of the still image within each trial was calibrated individually for each participant in the threshold estimation task right before the primary task.

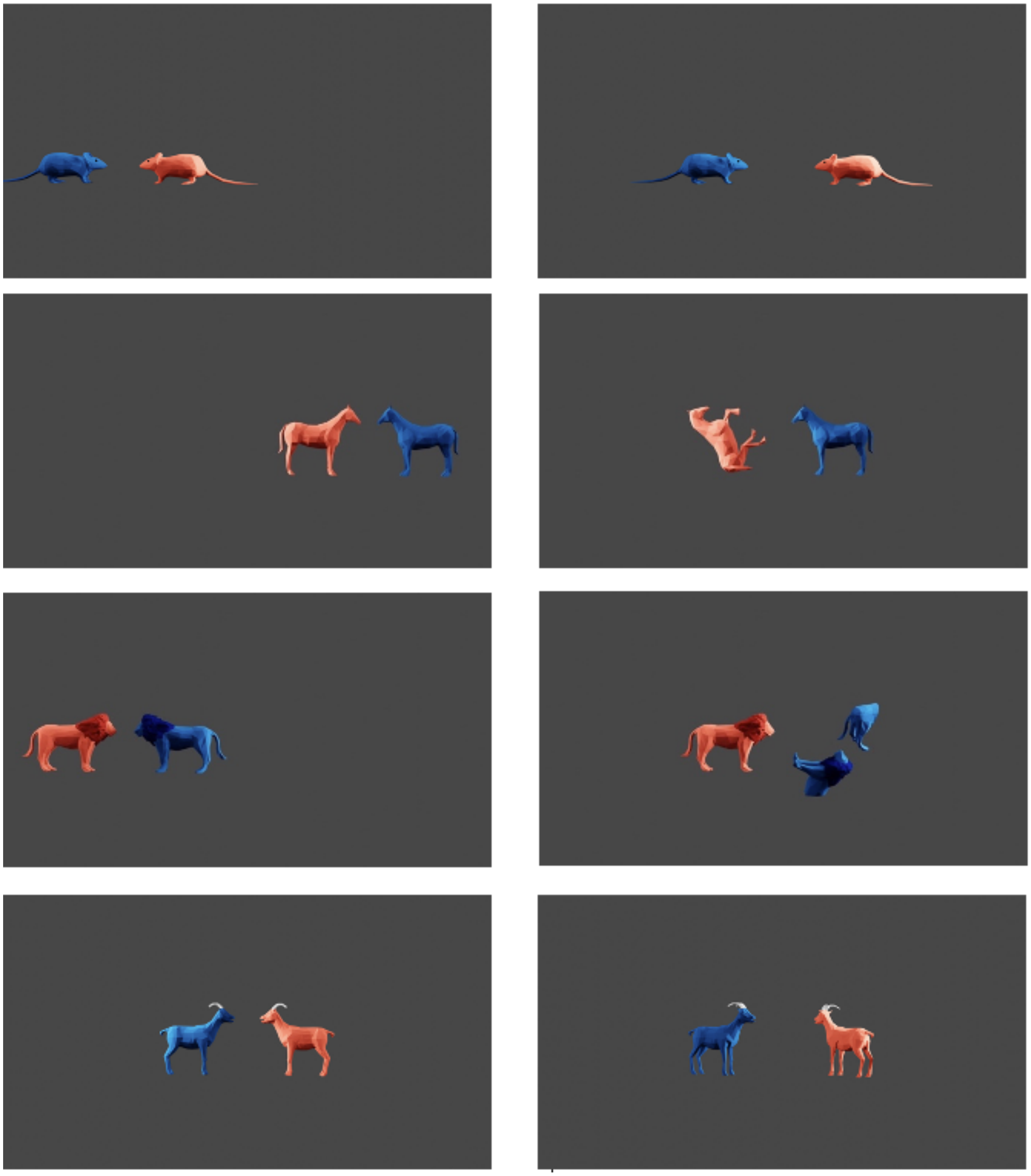

Figure 4.1 Examples of Stimuli in Experiment 4
There are four species of animals – mouse, horse, lion, and goat. The two animals in a scene are always from the same species. The first three rows represent three eventive relations – *launch, deform*, and *break* – while the fourth row represents the perceptual control ("*dance*"). For eventive relations, animals start from the left (e.g., the first and second row) or the right (e.g., the third row) of the screen (as the left column shows) and finish the relation in the middle of the screen (as the first three rows within the second column shows; the final frame of control scene is identical to its initial frame, so the picture depicts a middle frame to illustrate the movements). Agents are defined as blue mouse, blue horse, red lion in the three eventive relations respectively, with no changes on their shapes.

At the very beginning of Experiment 4, participants were instructed with examples that there was a pair of animals in each trial, where one animal might "break", "knock down", or "push away" the other animal. The instructions pointed out that there is a well-defined agent and a well-defined patient for these three relation types. In addition, the participants were instructed that animal might not interact with each other in certain trials, which suggested that there was no well-defined agent or patient.

The threshold estimation task began immediately after the instructions outlined above. It asked the participants to indicate the spatial location of the agent (with fewer perceptual cues in our stimuli as compared to patient) in still image in each trial. The key sentence in the instruction was: "As quickly and accurately as possible, press A/a if the agent appears on the left and L/l if it appears on the right." Participants were also informed that only breaking and deforming events, which included well-defined agent and patient in still images, would be presented during the threshold estimation task. The purpose of this task was to estimate the presentation duration that yields at least 90% accuracy for the given individual participant. We used a version of the staircase method to collect data adaptively (Leek, 2001). Concretely, we used two interleaved staircases – one 3-up-1-down and the other 4-up-1-down. That is, if 3 (or 4, respectively) consecutive responses were correct, the presentation duration became more difficult (i.e., shorter) by one step, whereas a single incorrect response triggered an increase of the duration (i.e., made it easier to identify the role). Taking the 60 Hz refresh rate of our monitor into account, the possible presentation durations range from 16.67 ms to 300 ms, and the size of each step is 16.67 ms. There is a total of 100 threshold estimation trials (50 trials per staircase chain). 32 unique still images had equal chance of being presented on each threshold estimation trial. After 100 responses and their corresponding durations were collected, a maximum likelihood estimation in Tensorflow.js (Smilkov et al., 2019) was used to estimate the underlying psychometric function. Our model uses a transformed logistic function, as recommended by Shen and Richards (2012) to estimate the presentation duration that leads to 90% accuracy for each participant.

As a sanity check of the estimation procedure, we conducted a parameter recovery simulation. Figure 4.2 summarizes a random subset of the simulated participants. The true psychometric functions are plotted as blue lines, the simulated responses as blue dots, and the recovered psychometric functions as red lines. Most of the recovered curves are within an acceptable range with the simulated curves, except for “s07” where the simulated threshold is 1 frame. The bad fit is not surprising in this case, though, because the intensities of all simulated stimuli are above the threshold. Nevertheless, no real participants will be expected to achieve 90% accuracy under such a short exposure either way. Overall, the reasonable recovered parameters validated the procedure of estimating individual shortest presentation durations for each participant.

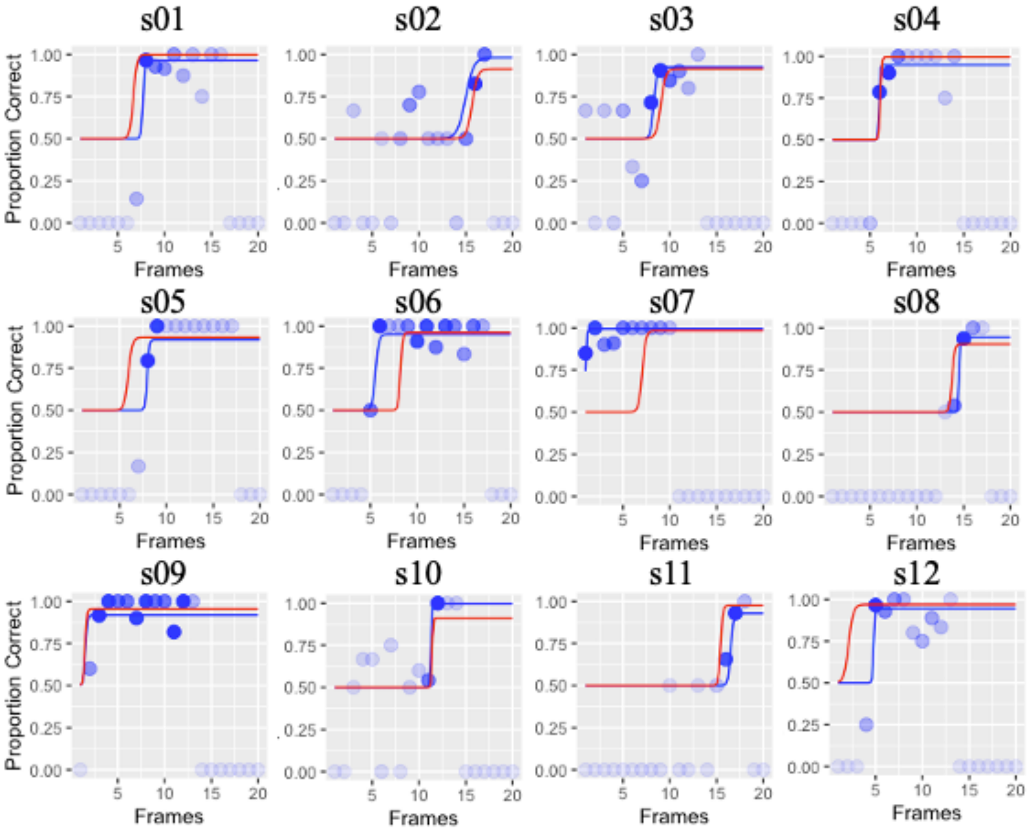

Figure 4.2 Parameters Recovery for Simulated Participants
Random subset of 12 simulated participants. For each simulated participant, a blue curve depicts the underlying psychometric function, in which threshold, slope, and lapse were randomly selected within reasonable ranges. Blue dots depict the simulated responses generated for stimuli of various intensities (i.e., number of frames for exposure), and opacity suggests the frequency of the corresponding stimulus intensity. Red curve depicts the recovered psychometric function based on the simulated responses using the same estimation software that is going to be used in the behavioral experiment.

After the threshold estimation task, the participants moved on to the primary task of the experiment. It began with a few practice trials to help them transition from identifying agent to identifying an assigned foreground color (same as Experiment 3). Then, they moved on to the main task, in which they were required to indicate the spatial location of a foreground color (red or blue). The key sentence in the instruction for a participant with blue assignment was: “As quickly and accurately as possible, press A/a if the blue animal appears on the left and L/l if it appears on the right.” The presentation sequence within each trial is depicted schematically in Figure 4.3: a priming video for 1400 ms, followed by its own final frame for extended 100 ms, a mask with a fixation crosshair for 100 ms, a static image for duration individualized in the threshold estimation task, and another mask with a fixation crosshair. The computer then waited for the participant's response. If the response time was greater than 2000 ms, a reminder with "Slow Response" would be shown. Trials were divided by intertrial intervals of 500 ms.

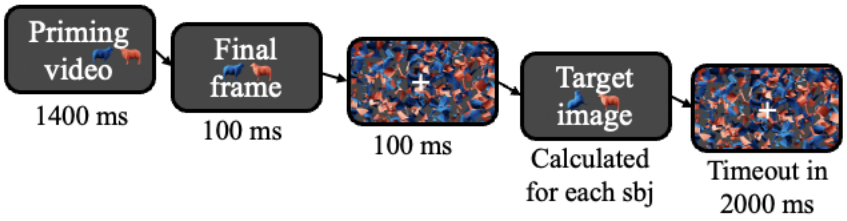

Figure 4.3 Presentation Sequence of One Trial in Primary Task
Presentation sequence of one experimental trial of the primary task in Experiment 4. The durations of priming video, the final frame (of the priming video), and mask are listed, while that of target image was calculated in the threshold estimation task for each participant.

A total of 768 trials were presented in 12 blocks in the primary task. The trial pool was generated according to one of the six templates in Table 4.1. Each template consists of 4 cells marked with Xs in Table 4.1. Each cell features prime videos with a definite animal species combined with target images with a definite species. For example, the first cell of the first template combines mouse videos with mouse targets (M-M), the second cell combines horse videos with horse targets (H-H), the third combines lion videos with goat targets (L-G), and the fourth and final cell combines goat videos with lion targets (G-L). Each cell generated 192 trials – 16 videos fully crossed with 12 target images. The videos were generated by crossing the 4 event types, 2 color assignments (the agent is red vs blue), and 2 position assignments (the agent is on the left vs right). The 12 target images were generated by crossing the 3 relation types (excluding the perceptual control, which in a still image looks identical to *launching*), 2 color assignments, and 2 position assignments. An overall presentation sequence was constructed by collecting all trials across the 4 cells of the template, and then randomizing the presentation order, under a constraint that each block consisted of all 64 unique videos.

In order to motivate the participants to pay attention to the videos, 72 trials chosen at random from the main presentation sequence included a "catch" question at the end of the trial, after recording the behavioral response on the primary task. Each catch question was presented in a pop-up box at the center of the screen, asking "What happened in the short video?" The participants then chose from four options – "break", "Knock down", "Push away" or "Dance/NA".

*Stimuli and Procedure of the Complementary Task*. As scheduled separately with participants, they completed the complementary task in a follow-up session. The complementary task had the same logical design as the primary task, except that the primes were linguistic instead of visual. Concretely, in addition to the static images used in the primary task, the complementary task included 64 English sentences. Sentences were designed to be as analogous as possible to the videos in the primary task. Sixty-four sentences were generated as the full factorial combination of 4 animal species, 4 event types, 2 colors of the agent, and 2 locations (left vs. right). Table 4.2 illustrates the sentences. The linguistic materials were presented in two modalities simultaneously – as text written on the screen and as a male voice that played through the computer sound system. The synthetic audio files were generated using an online text-to-speech tool (https://www.readthewords.com/Try.aspx).

Table 4.2 Examples of Sentences in Complementary Task
Illustration of the sentences used as primes in the complementary task. All sentences involving goats are listed. The sentences for the other three animal species are generated analogously.

| Relation type | Agent Color | Agent Location | Sentence |
|---|---|---|---|
| launching ("pushing away") | red | Left | The red goat on the left pushed away the blue goat on the right. |
| | red | Right | The red goat on the right pushed away the blue goat on the left. |
| | blue | Left | The blue goat on the left pushed away the red goat on the right. |
| | blue | Right | The blue goat on the right pushed away the red goat on the left. |
| deforming ("knocking down") | red | Left | The red goat on the left knocked down the blue goat on the right. |
| | red | Right | The red goat on the right knocked down the blue goat on the left. |
| | blue | Left | The blue goat on the left knocked down the red goat on the right. |
| | blue | Right | The blue goat on the right knocked down the red goat on the left. |
| breaking | red | Left | The red goat on the left broke the blue goat on the right. |
| | red | Right | The red goat on the right broke the blue goat on the left. |
| | blue | Left | The blue goat on the left broke the red goat on the right. |
| | blue | Right | The blue goat on the right broke the red goat on the left. |
| control | | | There was a red goat on the left and a blue goat on the right. |
| | | | There was a red goat on the right and a blue goat on the left. |
| | | | There was a blue goat on the left and a red goat on the right. |
| | | | There was a blue goat on the right and a red goat on the left. |

Similar to the primary task, the complementary task also began with a few practice trials to help participants prepare for the subsequent task. Then, they entered the main task, in which they were required to indicate the spatial location of a target color (red or blue, the same as the primary task) with the same instruction as the primary task (i.e., "As quickly and accurately as possible, press A/a if the blue animal appears on the left and L/l if it appears on the right." for a participant with blue assignment). The presentation sequence within each trial is depicted schematically in Figure 4.4: a priming sentence presented both visually and audibly for *approximately*[6] 3000 ms, a mask with a

[6] Sentences and the corresponding audios varied in length. In order to make sure the durations from the prime offset to the target onset are consistent across trials, we set up a 3000 ms window to present the text and audio in each trial. That is, the durations from the behavioral responses to the prime offset were consistent (i.e,, 3000 ms), and the durations from the prime offset to the target onset were also consistent (100 ms), whereas the durations from response to the prime onset varied across the trials.

fixation crosshair for 100 ms, a static image for individualized duration, and another mask with a fixation crosshair. Again, the computer then waited for the participant's response, sent a reminder if the response took more than 2000 ms, and moved on to the intertrial interval of 500 ms.

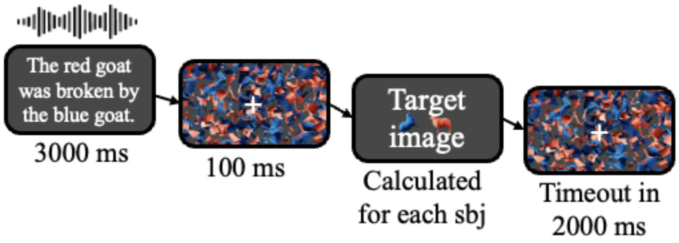

Figure 4.4 Presentation Sequence of One Trial in Complementary Task
Presentation sequence of one experimental trial of the complementary task in Experiment 4. The trial starts with a 3000 ms presentation window with text and audio being presented simultaneously. Duration of the target image is the same as the primary task for each participant, as calculated in the threshold estimation task.

The block and presentation-sequence design in the complementary task was also very similar to those in the primary task. A total of 768 trials were presented in 12 blocks. The trial pool was generated according to the same template used in the primary task, where each cell generated 192 trials with 16 sentences fully crossed with 12 target images. The sentences were generated by crossing the 4 event types, 2 color assignments of the agent (red or blue), and 2 locations of the agent (left or right). The 12 target images were generated in the same way as the primary task – crossing 3 relations, 2 color assignments, and 2 location assignments. Sequence randomization manipulation and catch questions were also identical to those in the primary task, except the catch questions asked about the sentences rather than the short videos.

Results

As described earlier, we dropped 6 participants according to various predetermined exclusion criteria. All results reported below are based on the data from the remaining 15 participants. As described in Methods, we could not collect as many participants who pass the exclusion criteria as we originally planned based on the power analysis. Therefore, the results in Experiment 4 should be considered carefully as they were observed in an incomplete sample. On average, it took 193.40 ms (SD = 88.13) to indicate the location of agent in the threshold estimation task. As for the main color-localization task, participants had an average accuracy of 96.74% (*SD* = 2.68%) in the video session and an average of accuracy of 97.95% (*SD* = 1.39%) in the linguistic session. The difference in accuracy across the two sessions is marginally statistically significant (*t* (14) = -1.94, *p* = .07). Recall that approximately 10% of the trials included a *catch question* after the color-localization task. The average accuracy on these catch questions was 82.31% (*SD* = 10.10%) in the video session and 98.89% (*SD* = 1.31%) in the linguistic session. This difference is highly statistically significant (*t (*14) = -6.13, *p* < .001). Table 4.3 contrasts the types of relational events in prime with the corresponding answers to the catch questions. Briefly, events were more likely to be memorized as *deform* and *launch* than *break* and *control*.

Table 4.3 Response Pattern for Catch Questions
Frequency of the randomly chosen prime events that followed by catch questions, separated by the primary and complementary tasks, and the corresponding answers to the catch questions.

| Prime Relational Events | | break | deform | launch | control |
|---|---|---|---|---|---|
| Primary task (Visual prime) | Break | 270 | 39 | 37 | 12 |
| | Deform | 17 | 292 | 53 | 9 |
| | Launch | 27 | 57 | 297 | 10 |
| | control | 13 | 31 | 26 | 322 |
| Complementary task (linguistic prime) | Break | 388 | 6 | 4 | 1 |
| | Deform | 6 | 351 | 2 | 1 |
| | Launch | 2 | 3 | 376 | 0 |
| | control | 7 | 12 | 8 | 344 |

The next data-processing step was to apply the trial-level exclusion criteria of the pre-determined experimental plan. Whereas the participant-level criteria excluded all data for certain individuals, the trial-level criteria exclude some trials from the data analysis. Specifically, we dropped trials with incorrect responses and trials whose responses were too fast (<200 ms) or too slow (>2000 ms). Finally, we also excluded trials whose RTs were more than 2.5 standard deviations away from the mean RT for a given participant, separately for the video and linguistic sessions. A total of 5.30% of the trials were excluded due to all criteria combined. At the end of the exclusion process, the remaining data set had a total of 21,820 trials across 15 participants. All results reported below are based on this “clean” data set. To emphasize once again, the analyses are based exclusively on data from trials with correct responses.

The means of RT for these “clean” trials were 492.20 ms (*SD* = 122.90 ms) for the primary task with visual primes and 385.85 ms (*SD* = 65.89 ms) for the complementary task with linguistic primes. Figure 4.5 contrasts the average RT of each participant in the complementary task to those in the primary task. All 15 participants showed a consistent pattern – slower responses in the primary task compared to the complementary task. There were also substantial individual differences in the speed of responding across tasks.

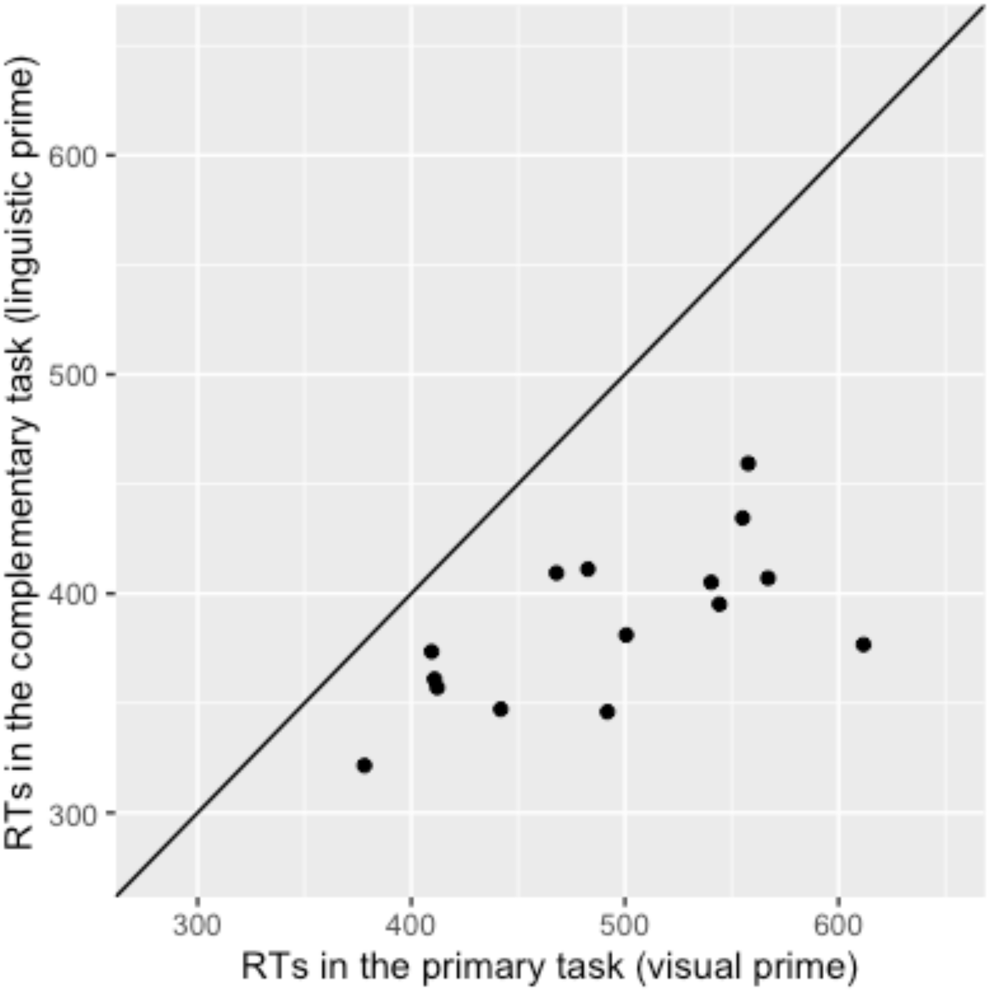

Figure 4.5 Contrast between RTs in Complementary Task and Primary Task
Contrast between the mean RTs of each participant in the complementary task and the corresponding mean RTs in the primary task. Each participant contributes one dot in this plot. Diagonal line illustrates identical mean RTs between the two tasks. All dots are below the diagonal line.

To eliminate these effects from the main results of interest – an inhibitory effect of switched roles, we used the linear mixed-effect models. The outcome variable of the model was response time (RT), the same as Experiment 3, and was manipulated to be inverse RT (-1000/RT) to improve normality of its distribution. The original full model included the following fixed effects: Session (visual vs. linguistic – format of priming in the primary and complementary tasks), Role (repeated vs. switched vs. ambiguous –

whether the animal in foreground color played the same or different role between prime and target events, or the role was not well-defined in at least one event), Relation (repeated vs. switched – whether the event types presented in the prime and target were the same), Location (repeated vs. switched – whether the animal in foreground color appeared in the same location between prime and target events or not), Animal (repeated vs. switched – whether the animals were of the same species between prime and target events), Prime Relation (break vs. deform vs. launch vs. control – the relation event in prime), Target Relation (break vs. deform vs. control – the relation – the relation type in target), Prime Location (left vs. right – the location of foreground animal appeared in prime), Target Location (left vs. right - the location of foreground animal appeared in target), Prime Animal (mouse vs. horse vs. lion vs. goat – the species of the foreground animal in prime), Target Animal (mouse vs. horse vs. lion vs. goat – the species of the foreground animal in target), Target Topology (broken vs. intact – whether the foreground animal was broken or intact), as well as the interactions between each repeated factors, and their interactions with Session. Trial number was also included as a fixed effect to account for temporal dependencies. To account for individual differences, we included each pair of repeated factors and their interactions as the random effects. The model converged with all fixed effects and two random effects – Side and Session. Based on the full model, a backward stepwise selection was applied to produce the best-fitting model based on AIC (Kuznetsova et al., 2017; Matuschek et al., 2017).

The best-fitting model includes predictors for the temporal effect, Session, properties of the prime and target events, as well as varying repeated effects and their interactions with Session, as summarized in Table 4.4. Among those, the Role factor and its interaction with Session are of most interest to us. The main effect of Role was significant across sessions. Specifically, responses were slower (*t* (21,780) = 4.16, *p* < .001) when the role repeated between prime and target events (*M* = 441.78, *SD* = 114.99) than when role was ambiguous in at least one of prime and target events (*M* = 436.04, *SD* = 109.29). Responses were also slower (*t* (21,780) = 3.02, *p* = .003) when the role switched (*M* = 440.32, *SD* = 113.41) than it was ambiguous; while no significant differences were found between trials with repeated and switched roles (*t* (21,780) = 1.08, *p* = .28). In terms of the interaction effect between Role and Session, no significant differences were found when the roles were primed as sentences in the linguistic session. In visual session, however, the 12.58 ms difference between trials with repeated roles (*M* = 499.28, *SD* = 126.41) and trials with ambiguous role in prime and/or target events (*M* = 486.70, *SD* = 119.23) reached the significant level of .01 (*t* (21,780) = 2.67, *p* = .008). Although the difference in RTs between trials with switched roles (*M* = 496.21, *SD* = 126.11) was 9.51 ms, it did not reach significant level of .05 (*t* (21,780) = 1.18, *p* = .237). In addition, there was a 3.07 ms role repetition cost in visual session, while it did not reach significant level of .05 as well (*t* (21,780) = 1.30, *p* = .193). Similarly, the best-fitting model with the interaction of Role and Session fitted the data better than a model without this interaction factor ((*χ2* (2) = 7.18, *p* < .05), and the model without the interaction factor also fitted the data better than a model without the interaction factor and the main factor of Role ((*χ2* (2) = 12.38, *p* < .01).

Table 4.4 Fixed Factors in the Best-Fitting Model of Experiment 4
Mean RTs across participant between levels of fixed factors included in the best-fitting model of Experiment 4. The order of the repeated factors is Role, Location, Relation, and Animal. The corresponding effects in prime, target, or interaction with session are listed next to the main repeated factors if significant. The fixed effect of trial number is also included in the best-fitting model, but it is not listed here as no clear contrast involved.

| Factor | Contrast | difference (ms) | t values in best-fitting model |
|---|---|---|---|
| Session | Visual 492.20 (122.90) – linguistic 385.85 (65.89) | 106.35 | 10.04*** |
| Role | Switched 440.32 (113.41) - ambiguous 436.04 (109.29) | 4.28 | 3.02** |
| | Repeated 441.78 (114.99) - ambiguous 436.04 (109.29) | 5.64 | 4.16*** |
| | Switched 440.32 (113.41) - Repeated 441.78 (114.99) | -1.36 | 1.08 |
| Role, visual session | Switched 496.21 (126.11) - ambiguous 486.70 (119.23) | 9.51 | 1.18 |
| Role, linguistic session | Switched 386.22 (62.99) - ambiguous 385.84 (68.04) | 0.38 | |
| Role, visual session | Repeated 499.28 (126.41) - ambiguous 486.70 (119.23) | 12.58 | 2.67** |
| Role, linguistic session | Repeated 385.49 (64.37) - ambiguous 385.84 (68.04) | -0.35 | |
| Role, visual session | Switched 496.21 (126.11) - Repeated 499.28 (126.41) | -3.07 | -1.30 |
| Role, linguistic session | Switched 386.22 (62.99) - Repeated 385.49 (64.37) | 0.73 | |
| Location | Switched 436.01 (110.20) - Repeated 441.04 (113.32) | -5.03 | -4.85*** |
| Location, visual session | Switched 487.11 (121.71) - Repeated 497.30 (122.87) | -10.19 | -4.43*** |
| Location, linguistic session | Switched 385.73 (66.48) - Repeated 385.97 (65.29) | -0.24 | |
| Target Location | Left 436.82 (110.20) – right 440.24 (113.36) | -3.42 | -2.73** |
| Relation | Switched 439.93 (111.31) - Repeated 435.70 (112.73) | 4.23 | 2.62** |
| Prime Relation | break 435.39 (110.24) - deform 440.37 (115.42) | -4.98 | -2.76** |
| | break 435.39 (110.24) - launch 439.44 (112.49) | -4.05 | -2.41* |
| | break 435.39 (110.24) - deform 438.88 (108.92) | -3.49 | -4.88*** |

Continued

Table 4.4 Continued

| | | | |
|---|---|---|---|
| Target Animal | Lion 433.67 (110.07) – goat 439.96 (113.09) | -6.29 | -4.91*** |
| | Lion 433.67 (110.07) – mouse 444.07 (112.21) | -10.4 | -8.87*** |
| | horse 436.27 (111.57) – goat 439.96 (113.09) | -3.69 | -2.90** |
| | horse 436.27 (111.57) – mouse 444.07 (112.21) | -7.80 | -6.62*** |
| | goat 439.96 (113.09) – mouse 444.07 (112.21) | -4.11 | 3.66*** |
| Target Topology | Broken 453.88 (118.86) – intact 435.47 (110.09) | 18.41 | 10.90*** |
| Target Topology, visual session | Broken 516.55 (127.88) – intact 487.41 (121.32) | 29.14 | 4.10*** |
| Target Topology, linguistic session | Broken 393.33 (118.86) – intact 384.35 (65.51) | 8.98 | |

These patterns are illustrated in Figure 4.6: (i) RTs in visual session are about 100 ms longer on average than linguistic session; (ii) RTs of different role conditions are about the same in linguistic session; (iii) both RTs of trials with repeated role and switched role are longer than RTs of trials with ambiguous roles (in prime and/or target events), while only the difference between repeated and ambiguous is significant.

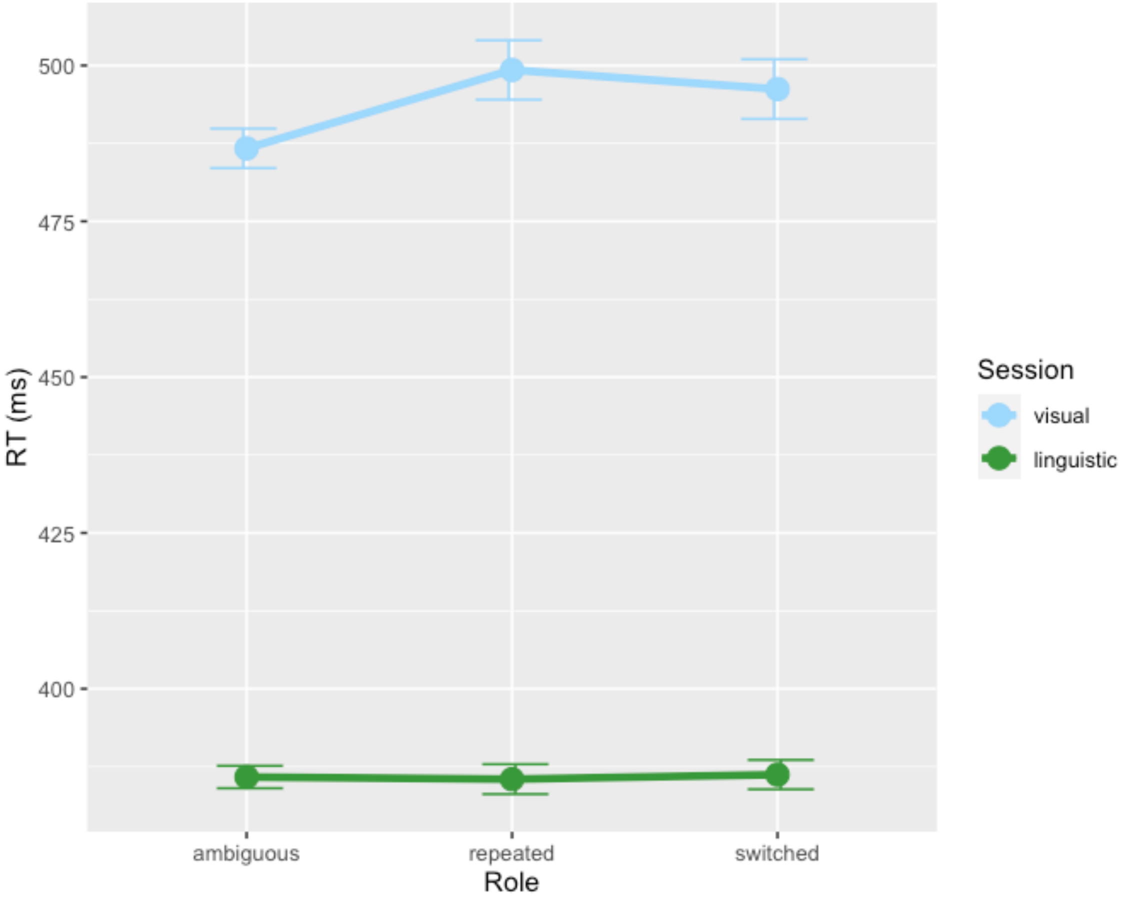

Figure 4.6 RT as a Function of Role and Session
Response time (RT) as a function of trial conditions where the roles of foreground animals repeated, switched, or were ambiguous from prime to target events, grouped by the session with visual or linguistic primes. Error bars indicate the 95% confidence intervals.

The best-fitting model also revealed effects other than the ones regarding roles (as shown in Table 4.4). We report them in the order of decreasing effect size. First of all, the model revealed a strong learning effect in the data (*t* (21,780) = -37.18, *p* < .001). As more trials were completed, the responses became faster. We also found a strong effect regarding the topological properties of the foreground animal in target (*t* (21,780) = 10.90, *p* < .001), as indicating the location of a broken animal (*M* = 453.88, *SD* = 118.86) took 18.41 ms longer than intact animal (*M* = 435.47, *SD* = 110.09). This effect did not interact with Role factor but was specific to the visual session (*t* (21,780) = 4.10, *p* < .001). Response time was also influence by the animal species in the target event. On average, it took 433.67 ms (*SD* = 110.07) to indicate a foreground lion, 436.27 ms (*SD* = 111.57) to indicate a foreground horse, 439.96 ms (*SD* = 113.09) to indicate a foreground goat, and 444.07 ms (*SD* = 112.21) to indicate a foreground mouse. Except for the comparison between lion and horse, the rest pairwise comparison all reached significant level of .01 (as listed in Table 4.4). The main factor of repeated animals and animal species in prime events, however, were not included in the best-fitting model. In terms of Relation, the best-fitting model included both main factor of repeated relations and the relation types in prime events. Unlike Role factor, responses in trials with repeated relations (*M* = 435.70, *SD* = 112.73) were faster than trials with switched relations (*M* = 439.93, *SD* = 111.31) in a significant level of .01 (*t* (21,780) = -2.76, *p* = .009). Responses were also faster when the prime events were *break* as compared to the rest three events, as shown in Table 4.4. However, the main effect of Relation did not interact with the main effect of Role, as illustrated in Figure 4.7.

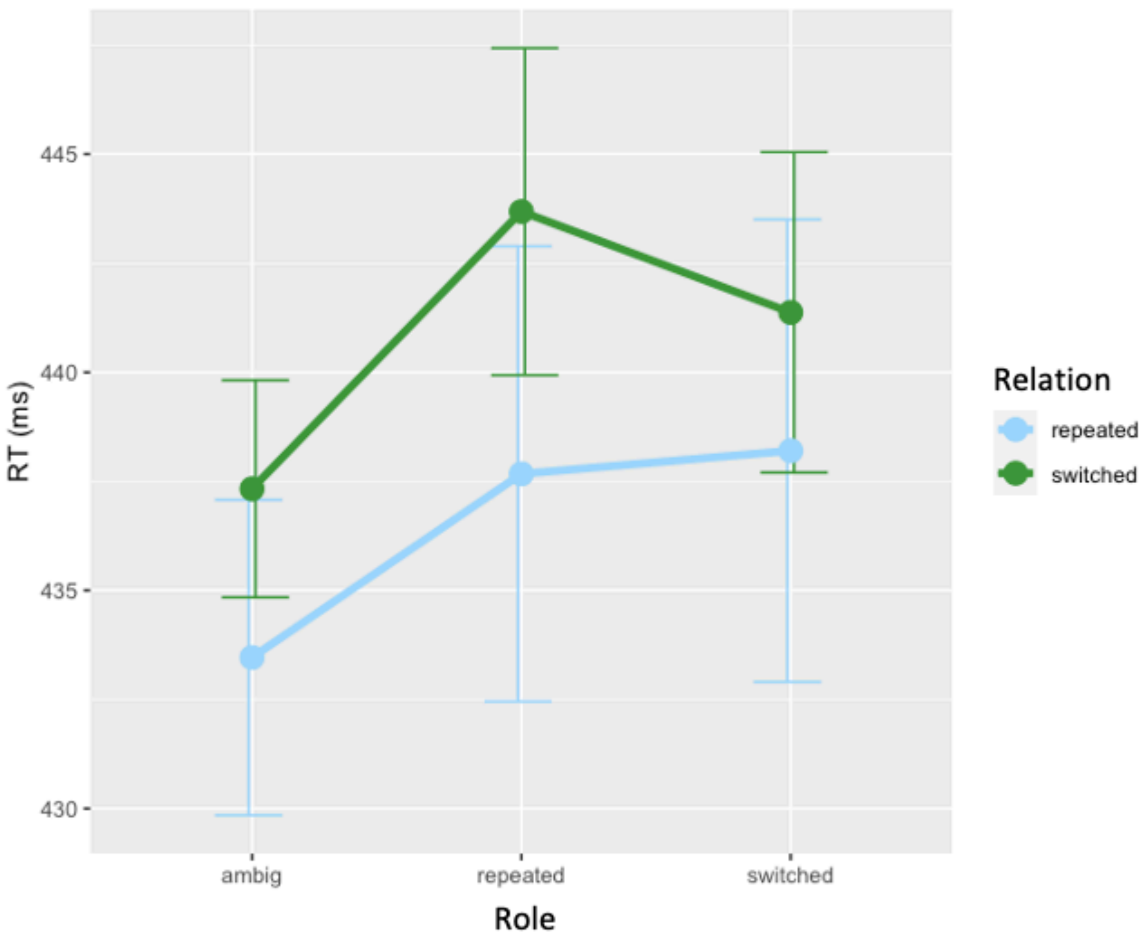

Figure 4.7 Interaction between Role and Relation
Response time (RT) as a function of trial conditions where the roles of foreground animals repeated, switched, or were ambiguous from prime to target events, grouped by trial conditions where target events repeated or switched from prime events. Error bars indicate the 95% confidence intervals.

Last but not least, we replicated the location switch benefit (or location repeat cost) in Hafri et al., (2018) and Experiment 3. Responses were 5.03 ms slower on average in trials when foreground animal appeared in repeated location (*M* = 441.04, *SD* = 113.32) than switched location (*M* = 436.01, *SD* = 110.20), which reached the significant level of .001 (*t* (21,780) = 4.85, *p* < .001). This effect was also specific to the visual session, as its interaction with Session reached the significant level of .001 as well (*t* (21,780) = 4.43, *p* < .001).

Discussion

To extend and complement previous experiments using continuous carryover presentation of trials, we modified the paradigm into separate prime and target and achieved these 3 goals: (i) priming relational roles by videos so they were not fully correlated with typical agent- or patient-like features in still images; (ii) including neutral trials, in which relational role were neither repeated or switched from the preceding primes, so role-repeated and role-switched trials can be compared to neutral trials to distinguish role switch cost from role repeat benefit (or distinguish role switch benefit from role repeat cost); and (iii) introducing a subsequent isomorphic task that only differed from the primary task in the priming format, so visual relational role in target could only be observed to repeat or switch from relational role in visual primes but not linguistic primes. On the basis of the modified paradigm, we had 3 hypotheses: (i) slower responses were expected on trials with repeated and switched roles than on neutral trials; (ii) slower responses were expected on trials with switched roles than on trials with repeated roles; and (iii) the effects of roles between consecutive relational events were expected to persist only for video primes, whereas no differences were expected for linguistic primes. The results of Experiment 4 supported the first and third hypotheses but discouraged the second hypothesis by showing an opposite trend.

As the best-fitting linear mixed-effect model showed a statistically significant main effect for the Role factor. This suggests that the main task – indicating the location of a foreground animal – was influenced by whether the role played by the foreground animal repeated or switched from the role it played in the prime. The model also showed a statistically significant main effect of the Session factor. This indicates that the main task was completed faster after linguistic primes compared to video primes. The Role-by-Session interaction was also statistically significant. The qualitative pattern of this interaction is illustrated in Figure 4.6 – the main task took longer time when both prime and target events included well-defined relational roles, as compared to when the roles were not well-defined in prime and/or target events. All these results are consistent with our first hypothesis. They are also consistent with Experiment 3, in the sense that there is an effect of repeated or switched roles. More precisely, it is better referred as role switch (or repeat cost), as the existence of roles in consecutive relational events slows down the overall process.

The third hypothesis was also confirmed as the RT differences among trials with different roles only persisted in the primary task that used videos as primes, but not the complementary task that used sentences as primes. However, instead of the role switch cost observed in Hafri et al. (2018), we found a role repetition cost of 3.07 ms. In our data, indicating the location of a foreground animal took longer time when its role repeated the role in priming video than when its role switched from the role in priming video. Two explanations might account for the inconsistency between our results and Hafri et al. (2018). First, since we changed the prime from still images to videos, the durations of prime events were prolonged in comparison to those of trials in the experiments of Hafri and colleagues. Therefore, our stimuli were more likely to initiate visual adaptation, which might have inhibited the subsequent encoding of the same relational roles in target. Although visual adaptation was believed to appear only after extensive repetition, recent studies found its timescale can be much shorter than expected (Webster, 2015). This explanation was also supported by the main effect of Session.

Regardless of the abstract roles, response was slower when it followed another visual event of 1500 ms than a conceptual event of 3000 ms. This is consistent with the hypothesis that the increase in RT in the primary task might stem from visual adaptation rather than fatigue in general.

Alternatively, the switch from a role switch cost to a role repetition cost might be a result of using computer generated low-resolution images of animals rather than photographs of human actors in Hafri et al. (2018). While people have extensive prior experience of seeing a person chase, push, or kick another person, it might not be natural for them to see a crudely rendered animal break, knock down, or push away another crudely rendered animal. The contrast between facilitatory effect of familiar events with human actors versus inhibitory effect of unfamiliar events with low-resolution animal actors was consistent with the functional group hypothesis by Green and Hummel (2006). That is, while object identification sensitivity increased when the object was presented in proximity to a distractor that both related and interacted with the object, it decreased when the object was presented in proximity to a distractor that interacted but lacked the semantic relation to the object. This suggests that when humans perform these eventive relations, the whole scenario might be treated as functional group that facilitate recognition, whereas the scenarios with artificial animals are hard to be treated as functional groups.

The two explanations outlined above are not mutually exclusive, and there might be other possibilities. The current results of our experiment cannot differentiate among these alternative explanations. Nevertheless, our results did suggest a role of vision in encoding abstract relational roles. Further research needed to elucidate how the human visual system represents and processes abstract relational roles.

In addition to the effects of Role and Session, the response time was also found to be influenced by many other factors, such as if the to-be-searched foreground animal appeared on the same location between prime and target events, the types of relations presented in the prime events, and if the animal was intact or broken in the target event. These findings are informative on their own. More important, they confirmed that our experiment had sufficient statistical power to detect RT differences on the order of 5 ms. Therefore, it is unlikely that we failed to detect an effect of such magnitude because we lacked statistical power.

Among those significant effects, two factors are worth further discussion. As described earlier, we aimed to manipulate the lower-level featural properties so they were not fully correlated with the abstract relational roles. Not surprising, we did find the effects of topological properties in RT – it took longer time to respond if the foreground animal was broken in the target image compared to when it was intact. Nevertheless, the main effect of topological properties of the foreground animal in target did not eliminate the effect of foreground animal's roles between prime and target events. Moreover, if such topological properties can account for most variance in abstract roles, we should have observed an interaction effect between repeated roles and repeated relations, as repeated topological properties manifested repeated roles in repeated breaking events. It was not supported by our results. Therefore, the results of Experiment 4 supported the claim that vision can encode abstract relational roles, and this could not be fully explained by featural properties that were correlated with the abstract roles.

Last but not least, we replicated the location repetition cost in Experiment 3 and Hafri et al. (2018; who referred to it as *side switch benefit*). Once again, the basic observation is that indicating the location of a foreground animal (or object, person) takes longer time when the animal appeared on the same location two (or more) times in a row. Hafri and colleagues provided a brief explanation that participants might expect a side switch when they observed as a switch in actors. Nevertheless, this speculative (in Hafri and colleagues' words) account is a rather post-perceptual account that may be at odds with the claim that the entire process was exclusively in vision. Further research is needed to investigate this consistent location effect. In particular, the examination of the inhibitory effect of location repetition may shed light on the potential inhibitory or facilitatory effect of abstract role repetition.

## Limitations and Suggestions for Future Work

First of all, one limitation of Experiment 4 is the number of valid participants who passed all exclusion criteria. Although power analysis based on pilot results suggest at least 16 participants would be good enough for a desired replicate rate of 80%, we only had 15 valid participants in the current sample. Therefore, more participants will be recruited in future experiments.

In addition, the current experimental paradigm can be modified and extended in several ways. First, as described earlier, we included a threshold estimation task to calculate appropriate target durations for individual participant. The criterion we chose for Experiment 4 ensured that the stimulus presentation duration was long enough for each participant to achieve 90% accuracy in indicating the location of agents. As results, there was a ceiling effect on the main task of indicating the location of a foreground color. Future experiments could make the main task more difficult by decreasing the presentation duration by setting a lower accuracy in the threshold estimation task. This would produce more errors and there might be informative patterns in these errors, and/or in the combination of correct and incorrect responses with the corresponding RTs.

In addition, the current experiment relied on catch questions to ensure that participants processed priming videos and sentences. Since the catch question only asked about the general event type rather than agent or patient, we claimed that they did not contaminate the main results of abstract roles. Nevertheless, more task-irrelevant control (e.g., eye-tracking) can be used to eliminate this concern.

To summarize, Experiment 4 used a modified prime-target paradigm and found evidence that the visual system encodes abstract relational roles. This finding cannot be fully explained by alternative accounts, such as lower-level patient-like properties or post-perceptual processing. Although Experiment 4 confirmed the role of vision in perceiving abstract relational roles, the current results posed new questions about the specific mechanism of this process. More experiments are needed to answer how vision encodes relations, in addition to whether it encodes relations.