Common Ground Building through Generative Cognitive Modules: Examining the Roles of Initial Perception, Imaging and Captioning

Ryunosuke Baba[1], Junya Morita[1], Takeru Amaya[1], Ryuichiro Higashinaka[2], Yugo Takeuchi[1]

[1] Shizuoka University, 3-5-1 Johoku, Chuo-ku, Hamamatsu, 432-8011, Japan

[2] Nippon Telegraph and Telephone Corporation, 1-1 Hikarinooka, Yokosuka 239-0847, Japan

baba.ryunosuken.20@shizuoka.ac.jp, j-morita@inf.shizuoka.ac.jp amaya.takeru.19@shizuoka.ac.jp, ryuichiro.higashinaka@ntt.com, takeuchi@inf.shizuoka.ac.jp

Abstract

To advance our understanding of referential communication and common ground formation, this study presents a novel generative cognitive model that integrates deep neural networks for visual perception, image generation, and language captioning. Using the Tangram Naming Task (TNT), we simulate the sender–receiver interaction with modular processes replicating holistic cognitive strategies. Through controlled simulation experiments, we reveal that language generation plays a more crucial role than visual perception in establishing common ground, while intermediate image generation enhances linguistic diversity—a key aspect of natural communication. Our results bridge cognitive modeling and large generative models, demonstrating how internal cognitive dynamics can be visualized and quantitatively evaluated. This study contributes to the growing field of cognitive-inspired human–AI communication and provides a blueprint for grounding-rich simulations in collaborative tasks.

Keywords: common ground, cognitive modeling, referential communication, generative AI, image captioning, Tangram Naming Task, simulation-based cognitive science

Introduction

Common ground is cognitive constructs related to beliefs, knowledge, and frames, which participants share to establish communication (Stalnaker, 2002). To form this, the sender (director) of a message explores shared beliefs/knowledge with the receiver (matcher) while searching for appropriate expressions (Clark & Schaefer, 1989). Furthermore, the building of common ground is not a one-way process from sender to receiver; feedback from the receiver is also necessary (Heller, Gorman, & Tanenhaus, 2012). Through this process, an implicit understanding of the meaning of the expression is formed among the participants, and the conversation proceeds smoothly.

Studies examining the building of common ground often use experimental tasks in which participants share the names of abstract images. In such tasks, participants metaphorically represent the images as concrete objects (Wu & Keysar, 2007). During this process, although communication appears to be superficially established, in reality, there may be differences in mutual understanding.

In order to understand the factors that contribute to the building of the common ground, it is necessary to develop a cognitive model that can visualize its internal processes. In the field of cognitive science, numerous models have been proposed concerning the emergence of symbols. Many of these models represent experimental situations by substituting real-world phenomena with abstract symbols in a computer system. However, relatively few studies have addressed the formation of analog internal representations (e.g., images) or the generation of natural language, which contains rich meanings and contextual dependencies.

For example, Reitter and Lebiere (2011) developed a computational model for a task, originally used in a pychological experiment couducted by Fay, Garrod, Roberts, and Swoboda (2010), where a director draws a picture related to a target concept, and a matcher is tasked with identifying an intended concept. In their study, the director model conveys symbolic representations of concepts, drawn from external sources, rather than describing analog images directly. The matcher then searches for the corresponding concept embedded in its network.

Contrary to the above previous study, the current study integrates several deep learning-based modules to generate language and images people usually use, as components of a communication model. Through this model, we aim to clarify the internal cognitive processes and functions involved in common ground building through simulations. The following sections provide previous research, the process targeted in this study, and the model of common ground representing this process. As applications of the model, several simulation studies are presented to reveal concrete process of common ground building.

Related Studies

In the field of cognitive science, researchers have extensively examined how two participants interact to achieve successful communication, particularly focusing on the process of building common ground through shared information (Chandu, Bisk, & Black, 2021). This line of inquiry aligns with the domain of "experimental semiotics," which investigates how novel communication systems emerge in contexts where a shared linguistic background or common assumptions are absent (Galantucci & Garrod, 2011).

Within this research framework, studies on common ground formation generally fall into two main categories: referential communication tasks, where participants use natural language to refer to objects or concepts (Knutsen, Bangerter, & Mayor, 2019), and pictorial tasks, where meaning is conveyed through visual representations such as drawings or images (Garrod, Fay, Lee, Oberlander, & MacLeod, 2007).

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

Referential communication tasks have been treated by researchers including Schober and Clark (1989); Metzing and Brennan (2003). In there studies, participants are divided into a director (sender) and a matcher (receiver). The director provides information about a specific figure, and the matcher selects the corresponding figure based on that information. Specifically, the information is sent based on a diagram that only the director can observe, and the matcher is required to identify the correct shape from several candidates. In such studies, the quality and quantity of the provided information have been analyzed to examine how they influence the process of forming common ground between the two participants.

Regarding referential communication, the influence of visual context on language generation has also been investigated. Tourtouri, Delogu, Sikos, and Crocker (2019) presented that redundant information, obtained from visual information in referential communication, reduces the cognitive load of the listener and makes target identification more efficient. The results indicate the importance of visual images in the building of common ground formation.

The tangram naming task has long been used as one of the tasks to examine the process of common ground building involving visual images (Clark & Wilkes-Gibbs, 1986). In the Tangram Naming Task (TNT), two participants, unable to see each other's figures, use verbal communication to infer which of a tangram the other participant is referring to. Tangrams are interpreted as silhouettes of objects, resulting in a variety of perceptions. This interpretation depends on the cognitive framework of the perceivers, and differences in the angle and visibility of the tangrams affect the building of common ground. Experiments on this task have shown that a common ground is constructed in the process of matching each other's perceptions through speech.

Several researchers have modeled the process involved in the Tangram Naming Task (TNT). For instance, Ji et al. (2022) prepared a dataset annotating tangrams based on their overall and partial features, and used it to develop deep learning models for performing the TNT. Their results demonstrated that models trained on partial features outperformed human participants even without fine-tuning. However, while this study successfully optimized task performance using deep learning, it differs from our approach in that it does not seek to replicate the underlying cognitive mechanisms of communication.

In contrast, cognitive models explicitly aim to simulate the mental processes and interactions that give rise to communicative behavior. These models are particularly effective for investigating how task success emerges from internal cognitive functions (Kotseruba & Tsotsos, 2020). For example, Reitter, Keller, and Moore (2011) proposed that the mechanisms of language production can be explained in terms of general memory and learning processes, implemented within cognitive architectures such as ACT-R (Adaptive Control of Thought-Rational; Anderson, 2007).

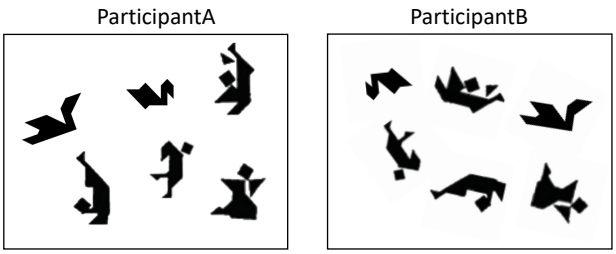

Figure 1: An example of placement in the tangram naming task.

Table 1: Example dialogue in the tangram naming task (dotted underlines: analytic, solid underlines: holistic).

A	And, you know, kicking, like, kicking a ball or something.
B	I can't see.
A	There is some kind of ball behind the head, and the feet are shaped like the guy is kicking a ball.
B	You know, the one with the separate squares?
A	Aha, yes, yes, yes.
B	Like Hokkaido?
A	Hokkaido.
B	Like a map of Japan
A	Oh, no, no, no, no, not that.
A	It's kind of a 90-degree kink.
B	Ah.
A	Foot-like, ball-kicking kind of thing.
B	Yes, yes, yes, like a little cross-legend thing?
A	Oh, yes, and that one with the little square behind it.
B	Yeah, I kind of get it.

Target Process

Among several studies employing the TNT, we focus on observations obtained by Sudo, Asano, Mitsuda, Higashinaka, and Takeuchi (2022). Figure 1 shows an example of a tangram set observed by a participant in their task. Both participants are shown the same tangram set, but the arrangement and angles of them differ. The participants are required to assign a shared name to each tangram, using only verbal communication.

Sudo et al. (2022) analyzed the utterances obtained in TNT from the perspectives of "analytic expressions" and "holistic expressions." Examples are shown in Table 1. Holistic expressions are those in which the shape of the tangram is compared to a concrete object (e.g., like Hokkaido, ball kicking), while analytic expressions are those in which the tangram is broken down into geometric figures (e.g., a square and a triangle on each side). In Sudo et al.'s data, the frequency of holistic expressions surpassed that of analytic utterances throughout the experiment, and the gap between the two widened as the sessions progressed. Notably, even in the early sessions, when common ground had not yet been established, the proportion of holistic expressions remained high. This suggests that the naming of tangrams shared between participants is not necessarily formed in a bottom-up manner through analytic expressions.

Model of Common Ground building

To explore factors involving common ground building, this study utilizes a model proposed by Morita, Yui, Amaya, Higashinaka, and Takeuchi (2023). This model represents a process of generating holistic expressions by senders and their interpretation by receivers in the TNT. This process is described as a sequential procedure involving multiple deep learning models. Figure 2 shows the process with actual examples of deep learning models. These deep learning models are assumed to function as sub-modules of a cognitive architecture as follows:

Sender process

1. **Perception**: At the beginning of the process, the model makes a brief impression of each tangram shape. An example is shown in "Perception" of Figure 2. This process is assumed to be instantiated by a Convolutional Neural Network (CNN) for general object recognition. However, for CNN models of general object recognition, the existence of a texture bias (i.e., recognition influenced by the texture of the image surface rather than the holistic shape) has been reported (Geirhos et al., 2018). To avoid this bias, Morita et al. (2023) utilizes the ImageNet Sketch dataset (Wang, Ge, Lipton, & Xing, 2019), which consists of black-and-white images that do not introduce texture bias. The 1000 labels in the dataset are mapped to the output layer of the CNN model, which was pretrained (loss = 4.85, Accuracy = 15.78%)[1]. By inputting a tangram image into the network, a 1000-dimensional probability vector (output of softmax) is obtained. From that vector, the model selects the label as a first impression of the tangram shape, which will be the input for the next image generation.

2. **Image generation**: To simulate the holistic expression shown in Table 1, the labels from ImageNet are not enough. To make a detailed expression of the tangram shapes, we use the img2img function of Stable Diffusion (Rombach, Blattmann, Lorenz, Esser, & Ommer, 2022)[2]. This function takes a language prompt and an initial image as input and generates a new image. In this study, the input tangrams and output labels from the previous step were used as input data. An actual example is shown in the "Image generation" of Figure 2.

3. **Text captioning**: Based on the image generated by the previous step, a detailed linguistic representation of the tangram is generated. This process utilizes a pre-trained Vision Encoder-Decoder model from HuggingFace[3], developed based on ViT (vision transformer; Dosovitskiy, 2020) and GPT (generative pretrained transformer; Radford, 2018). An example of the output of this step (the output of the sender) is shown in "Text captioning" of Figure 2.

Receiver process

1. **Image generation**: The receiver generates an image from the output of the sender. This process is instantiated by txt2img of Stable Diffusion based on the caption produced by the sender[4]. An example is shown in Figure 2.

2. **Tangram identification**: From the image generated in the previous step, the receiver attempts to identify the tangram image. This process is accomplished by computing the similarity between the generated image and the observable tangram. While various methods can be considered for image similarity calculation. Morita et al. (2023) assumed that the sender and receiver share basic cognitive modules, and the cosine similarity of the output layer of a CNN with a structure similar to that of the sender's perception is adopted.

Leaning Figure 3 represents the learning process proposed by Morita et al. (2023). The first column of Figure 3 shows the entire sequence of Figure 2 as one episode (transformed it into a vertical order). Each episode is applied to a tangram placed at a specific angle that the participant is observing in Sudo et al. (2022)'s experiment. The pairs of the network state and input obtained as a result are stored in an experience buffer, with labels indicating success (the receiver correctly identifies the sender's tangram) or failure. Through back-propagation using the success cases stored in the buffer, the weights of each module in the model are adjusted (Lapan, 2018). This series of procedures (collecting episodes, extracting success cases, and back-propagation) consists of a single trial and is repeated multiple times.

Experiment

As shown in the previous section, there is a model integrating multiple deep learning modules to represent a sequential communication process. However, the previous study did not prove that the above learning actually works toward common grounding. Therefore, in the current experiments, we examine the learnability of the model by manipulating each step of this process independently. Especially, this study focuses on the initial and the final steps of the process: the perception process, which captures the visual information of the tangrams, and the speech process, which does not change the view but changes the way of speaking. Applying the learning trials depicted in Figure 3 independently to each step, we examine their impacts on building common ground between the sender and the receiver. This will allow us to investigate

[1]Details about the Network architecture: Conv2D (32 filters, 3x3, ReLU), Conv2D (32 filters, 3x3, ReLU), MaxPooling2D (2x2), Dropout (0.1), Conv2D (64 filters, 3x3, ReLU), Conv2D (64 filters, 3x3, ReLU), MaxPooling2D (2x2), Dropout (0.25), Flatten, Dense (512 units, ReLU), Dropout (0.45), Dense (1000 units, softmax)

[2]v1-5-pruned-emaonly.safetensors of Stable Diffusion 1.6.0 was used.

[3]https://huggingface.co/nlpconnect/vit-gpt2-image-captioning

[4]The model of Stable Diffusion is the same as that of the sender.

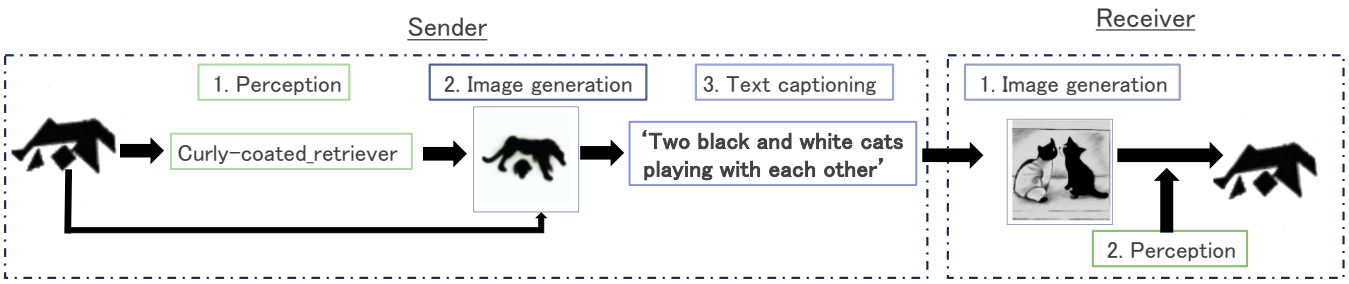

Figure 2: Example of the model

Table 2: Overall of Experiment

Process	Sender			Receiver		Max accuracy	
	Perception*	Image generation	captioning*	Image generation	Perception		
Model	CNN	SD‡(img2img)	ViT&GPT-2	SD‡(txt2img)	CNN	High†	Low†
Sim 1	-	random	-	random	-	0.35	0.06
Sim 2	5, Positive numbers	High/Low	-	High/Low	-	0.38	0.10
Sim 3	-	High/Low	5, Positive numbers	High/Low	-	0.45	0.41
Sim 4	None	None	5, Positive numbers	High/Low	-	0.16	0.43

* Epochs, batches when training the process
† The two conditions obtained in simulation 1, high and low seed values for the percentage of correct answers
‡ Stable Diffusion

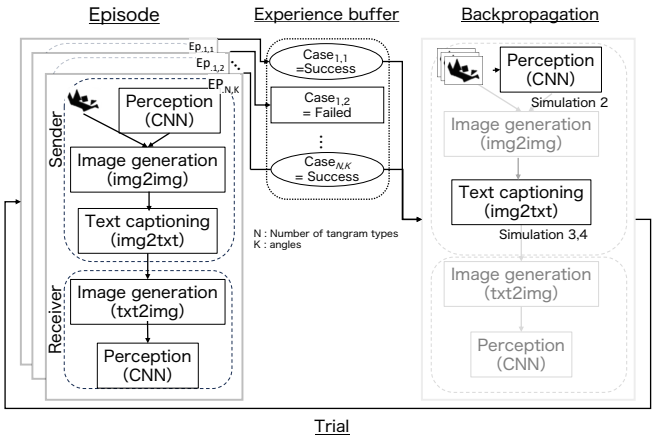

Figure 3: Overview of the model

the necessity of deep conceptual changes to build common ground in this communication task. Concerning this point, it has been discussed that common grounding is achieved only by superficial alignment by participants (Kabbach & Herbelot, 2021).

The structure of the simulation experiments conducted in this study is shown in Table 2, which consists of the processes manipulated in each simulation (columns 2-6) and an overview of the results (columns 7 and 8). In the cells related to the operation modules, a blank (-) indicates that the parameters of that module were not modified in the simulation. Additionally, the settings for the number of epochs and batches (e.g., 5, Positive numbers) indicate that the module was trained using back-propagation with those settings. Variables connected by slash (e.g., high/low) indicate that multiple models with that variable were set and compared. Cells labeled "random" indicate that variability in the deep learning seed values was investigated for that simulation.

Table 2 shows the four simulations conducted in this study. Simulation 1 explored the seed for the image generation to be set in the later simulations. Simulations 2 and 3 examined changes in the internal processes of the sender that contribute to the establishment of common ground. In these simulations, the sender's starting (perception) and the final process (language production) were independently trained using back-propagation. Simulation 4 is conducted to resolve a question about the significance of image generation that arose from Simulation 3. In all simulations, six types of tangrams were used, with eight variations of rotation angles (at 45-degree intervals) set for each. Therefore, each trial consisted of 48 episodes. Below, the objectives, procedures, and detailed results of each simulation are presented.

Simulation 1 In Stable Diffusion models, the seed value determines the initial noise pattern from which an image is generated. The model transforms this noise pattern based on the given prompt to produce a human-recognizable image. By fixing the seed value for each model, it serves to reveal a consistent generation style across different prompts specific to that model (Xu, Zhang, & Shi, 2025). As a preparatory step for the subsequent simulation, we examined how such differences in style influence the task of common grounding.

In this simulation, the seed values used for image generation by both the sender and receiver were randomly sampled. After the sampling, a trial (comprising 48 episodes, which include 6 types of tangrams and 8 angles as shown in Figure

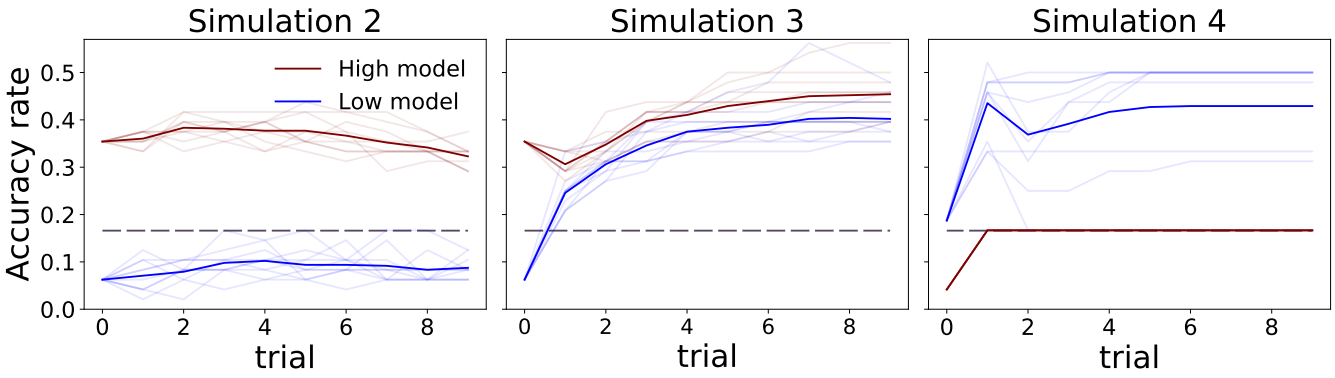

Figure 4: Simulation results (The dotted black lines are chance level))

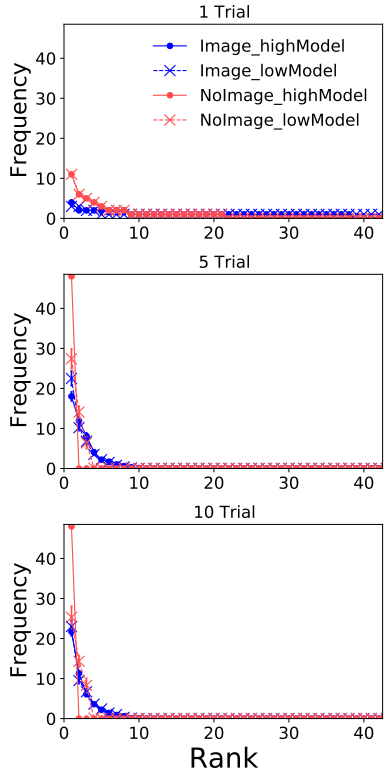

Figure 5: Rank frequency graph of Simulations 3 ,4

3) was executed once and repeated independently 100 times ($n = 100$).

The average accuracy of the 100 repetitions was resulted in 0.18, with a standard deviation of 0.05. The maximum accuracy rate was 0.35 and the minimum was 0.06. Although the mean was around the chance level (0.16 = 8/48), the percentage of correct answers varied. Taken into account this random variation caused by the initial settings, subsequent simulations were conducted using two models (hereafter referred to as "the low model" and "the high model"), with seed values corresponding to the minimum and maximum accuracy rates respectively.

Simulation 2 For the low and high models, a learning trial targeting the "Perception (CNN)" in Figure 3 was repeated ten independent runs of ten continuous trials. The results are plotted in Simulation 2 of Figure 4. The dark lines represent the average of the ten runs, while the light lines show the results of each individual run. The back-propagation algorithm was applied for each trial using each success case in the 48 episodes as a batch and repeated five epochs. In each batch, using a pair of a tangram image (x) and a 1000-dimensional output vector (y) observed in the success case, the network parameter of CNN was modulate to reduce the loss calculation based on categorical cross entropy.

From the figure, for the high model, no clear learning trend was observed even as the trials progressed. In contrast, for the low model, there was a slight increase in accuracy as the trials progressed, although it did not reach the chance level.

Simulation 3 For both the low and high models, ten learning trials targeting on the "Captioning (img2txt)" from Figure 3 was repeated 10 times. In each training trial, five epochs of batches corresponding to success cases obtained in the trial (48 episodes) were repeated. In simulation 3, each case was comprised of the image generated by the img2img (x) and the caption obtained from the txt2img (y). The network parameters of GPT were modulated to reduce CrossEntropy loss, and AdamW (lr=5e-5) was employed for optimization.

The results were plotted in Simulation 3 of Figure 4. Learning trends were observed in both models, with particularly noticeable improvement in the low model, which saw an increase from a low 6.2% to around 41%. From the result, it is suggested that common grounding is achieved only modulating the final output of the sender's process. Furthermore, it leads a question regarding the need for internal process especially image generation of the models.

Simulation 4 To investigate the impact of "Image generation" of Figure 3, an experiment was conducted under a condition without image generation. In this no-image-generation condition, the processes of the sender's perception and image

Table 3: Example Captions (...indicates abbreviation)

	Image_lowModel (Simulation 3)	NoImage_lowModel (Simulation 4)
Trial 1	*a bird is flying through the air* (3), *a black and white photo of a black and white cat* (3), *a black and white photo of a black and white animal* (2), *a black bird is flying through the air* (2), *a woman flying through the air while holding a skateboard* (1), *a black and white photo of a black and white photo* (1), *a colorful kite is suspended in the air* (1), *a black and white bird flying in the sky* (1), *a black bird flying through the air* (1), *a white microwave oven sitting on top of a table* (1) . . . Others(33)	*a black and white photo of a street sign* (11), *a black and white photo of an airplane flying in the sky* (6), *a black and white photo of a black and white bird flying* (5), *a black and white photo of a bird flying through the air* (4), *a black and white photo of two planes flying in the sky* (3), *an aerial view of an airplane flying in the sky* (2), *a black and white photo of a bird flying in the air* (2), *a black and white photo of an airplane flying* (2), *a black and white photo of two street signs* (1). . . Others(12)
Trial 10	*a black and white photo of a black and white cat* (25), *a black and white photo of a computer*(7), *a black and white photo of a* (4), *a black and white photo of a street sign* (4), *a black and white photo of a cat* (3), *two black and white cats playing with each other* (2), *a pair of scissors sticking out of the top of a woman's head* (2), *a black and white photo of a black and white photo of a black and* (1)	*a black and white photo of a black and white photo* (22), *a black and white photo of a* (16), *a black and white photo of an airplane flying* (10)

generation were removed, and captioning was performed directly from the tangram images. The experiment was carried out under conditions similar to those in Simulation 3 (img2txt learning). The success cases for this experiment consisted of pairs of the original tangram images (x) and the captions generated by those tangrams (y). In Simulation 4, both the low and high models were also set based on Simulation 1. However, the seed settings in these models were applied only to the receiver's image generation process.

The results were plotted in Simulation 4 of Figure 4. the high model did not exceed the chance level, but the low model showed an increase that surpassed the chance level. The results indicate that our model improves performance on the tangram naming task even without imaging[5].

However, when we investigated the contents of the generated captions, we found differences of variety between models with images and without images. This impression was quantified in Figure 5, which compares the frequency of captioning generated in Simulation 3 (with imaging) and Simulation 4 (without imaging). In the first trial, the captions generated in both conditions exhibited diversity. However, the diversity decreases as the trial progresses (trials 5 and 10). More importantly, the model without image generation (the red lines) showed a more pronounced convergence of captions as the trial progressed, compared to the model with image generation (the blue lines). In particular, the high model in Simulation 4 resulted in only one caption *'a black and white photo of an airplane flying'* from the fifth trial, regardless of the input tangram. To illustrate the generated captions more concretely, the captions from trial 1 and trial 10 of the low model condition in Simulations 3 and 4 are provided in Table 3.

[5]One thing that needs to be excused is the reversal of results between the high and low models. This is attributed to a change in model architecture. The previous simulations applied random seeds to image generations in the sender and the receiver, while only the receiver used the seeds in Simulation 4. Therefore, the assumed initial performance (high/low) cannot be applied in this simulation.

Discussion and conclusion

This study conducted simulations of a cognitive model that represents internal communication processes using generative models. The learnability of the model was shown in Simulation 3. This simulation also presented the importance of the surface level of the communication to build common ground with the receiver consistent with the previous discussion (Kabbach & Herbelot, 2021).

The results of Simulation 3 simultaneously raise a question about the necessity of the sender's internal processes (initial perception and image generation). This question is addressed by Simulation 4 excluding image generation. The results showed that image generation is not necessary if we only focus on the accuracy of the communication. However, we found that the role of image in generating diverse expressions in communication. This suggests that image generation is necessary to achieve human-like diversity (creativity) in language production (Chomsky, 2006).

Summarizing the study, we found advantages of the model including rich generative modules showing the concrete images and languages. The finding related to the image generation is difficult to obtain only from classic cognitive modeling approach (Reitter & Lebiere, 2011). However, our study indicates the need of more comprehensive investigations into the model used in this study. This study focused solely on the sender's module. In actual communication, however, the receiver also undergoes a learning process. It is necessary to further investigate such processes of mutual adaptation.

Furthermore, the model examined in this study achieved a maximum average accuracy of 0.45. For this study, which aims to simulate human cognitive processes, it is not necessary to aim for perfect agreement. However, in prior experiments conducted with human participants (Sudo et al., 2022), there were no instances of mismatched naming after the task. Considering this, the current accuracy rate is insufficient. To advance our understanding of the common ground building process, it will be necessary to explore more comprehensive models and examine methods that are more cognitively valid and improve the learning rate.

References

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press.

Chandu, K. R., Bisk, Y., & Black, A. W. (2021, August). Grounding 'grounding' in NLP. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the association for computational linguistics: Acl-ijcnlp 2021* (pp. 4283–4305). Association for Computational Linguistics.

Chomsky, N. (2006). *Language and mind*. Cambridge University Press.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, *13*(2), 259–294.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, *34*(3), 351–386.

Galantucci, B., & Garrod, S. (2011). Experimental semiotics: a review. *Frontiers in human neuroscience*, *5*, 11.

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: where might graphical symbol systems come from? *Cognitive Science*, *31*(6), 961–987.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Heller, D., Gorman, K. S., & Tanenhaus, M. K. (2012). To name or to describe: Shared knowledge affects referential form. *Topics in Cognitive Science*, *4*(2), 290–305.

Ji, A., Kojima, N., Rush, N., Suhr, A., Vong, W. K., Hawkins, R., & Artzi, Y. (2022). Abstract visual reasoning with tangram shapes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 582–601).

Kabbach, A., & Herbelot, A. (2021). Avoiding conflict: When speaker coordination does not require conceptual agreement. *Frontiers in Artificial Intelligence*, *3*, 523920.

Knutsen, D., Bangerter, A., & Mayor, E. (2019). Procedural coordination in the matching task. *Collabra: Psychology*, *5*(1), 3.

Kotseruba, I., & Tsotsos, J. K. (2020). 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, *53*(1), 17–94.

Lapan, M. (2018). *Deep Reinforcement Learning Hands-On: Apply modern RL methods, with deep Q-networks, value iteration, policy gradients, TRPO, AlphaGo Zero and more*. Packt Publishing.

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*(2), 201–213.

Morita, J., Yui, T., Amaya, T., Higashinaka, R., & Takeuchi, Y. (2023). Cognitive architecture toward common ground sharing among humans and generative ais: Trial modeling on model-model interaction in tangram naming task. In *Proceedings of the 2023 AAAI fall symposium on integrating cognitive architectures and generative models*. AAAI Press.

Radford, A. (2018). Improving language understanding by generative pre-training.

Reitter, D., Keller, F., & Moore, J. D. (2011). A computational cognitive model of syntactic priming. *Cognitive Science*, *35*(4), 587–637.

Reitter, D., & Lebiere, C. (2011). How groups develop a specialized domain vocabulary: A cognitive multi-agent model. *Cognitive Systems Research*, *12*(2), 175–185.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684–10695).

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive psychology*, *21*(2), 211–232.

Stalnaker, R. (2002). Common ground. *Linguistics and philosophy*, *25*(5/6), 701–721.

Sudo, S., Asano, K., Mitsuda, K., Higashinaka, R., & Takeuchi, Y. (2022). A speculative and tentative common ground handling for efficient composition of uncertain dialogue. In *Proceedings of the 13th Language Resources and Evaluation Conference* (pp. 3150–3157).

Tourtouri, E. N., Delogu, F., Sikos, L., & Crocker, M. W. (2019). Rational over-specification in visually-situated comprehension and production. *Journal of Cultural Cognitive Science*, *3*(2), 175–202.

Wang, H., Ge, S., Lipton, Z., & Xing, E. P. (2019). Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, *32*.

Wu, S., & Keysar, B. (2007). The effect of information overlap on communication effectiveness. *Cognitive Science*, *31*(1), 169–181.

Xu, K., Zhang, L., & Shi, J. (2025). Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages=3024–3034.*