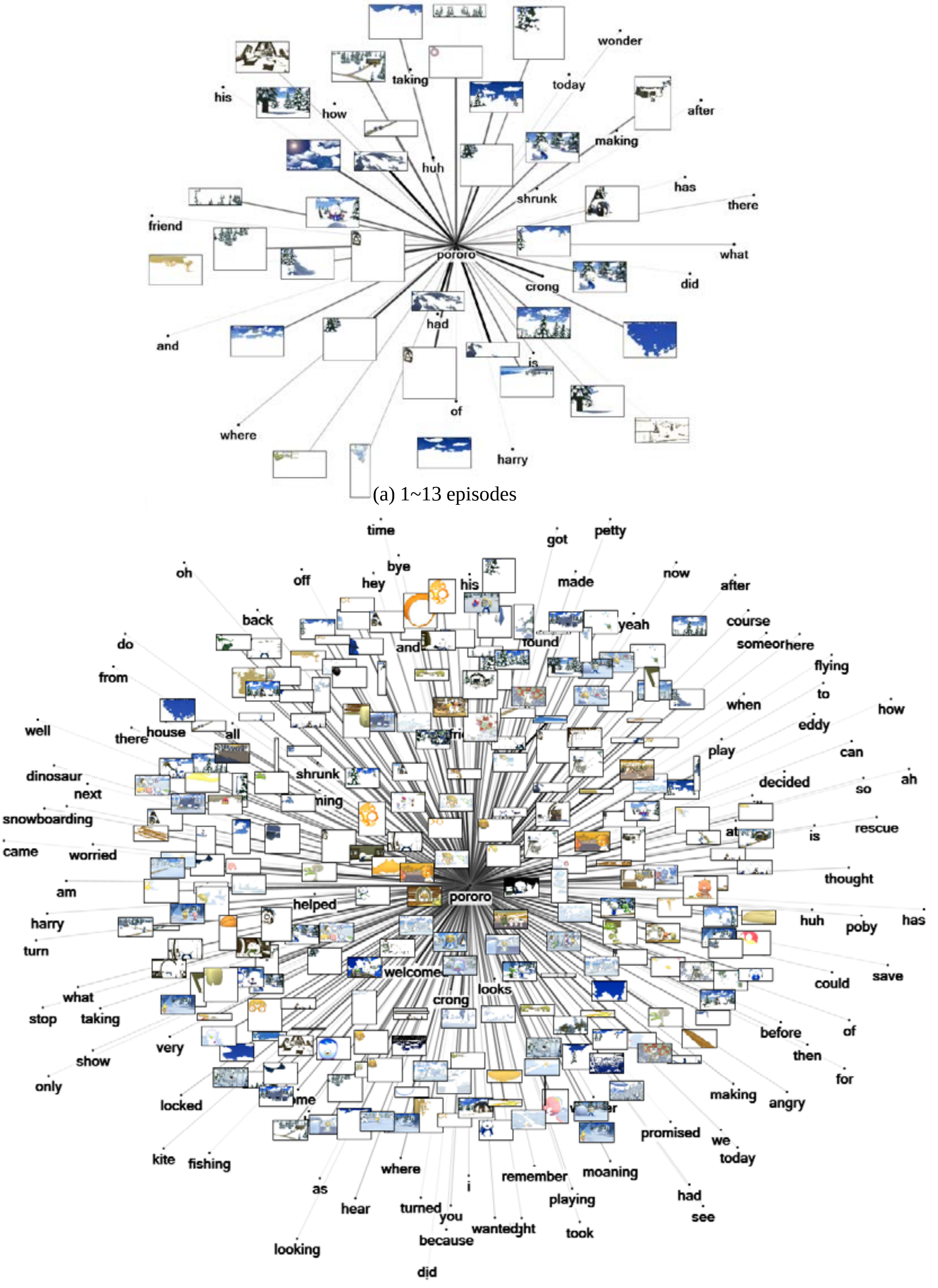
Supplementary Material

- Figure S1: Representation of concept development using visual-linguistic concept maps

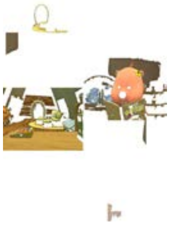
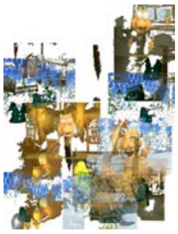
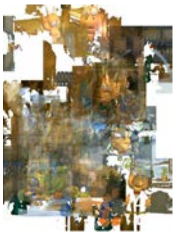
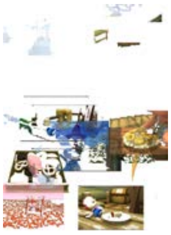
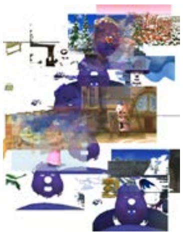
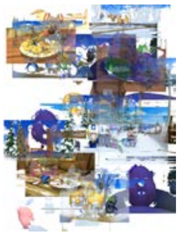


(a) 1~13 episodes

(b) 1~78 episodes

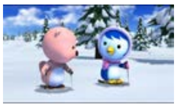
Concepts learned from videos using a deep concept hierarchy (DCH) model can be represented as a concept map, using words and patches as its nodes. Both maps are concept maps of Pororo. (a) and (b) are constructed after observing 13 and 78 episodes, respectively. The maps are built by converting the hyperedges in the clusters strongly connected to *Pororo* **c**^{2}-node into cliques and by summing the hyperedge weights. The distances between "*Pororo*" and other nodes are inversely proportional to the weights. As the number of observed videos increase, the concept on a character becomes more complex and is represented with more diverse words and images.

- Figure S2: Sentence-to-scene generation

Query sentences	**1~52 episodes (1 season)**	**1~104 episodes (2 seasons)**	**1~183 episodes (all seasons)**
• Tongtong, please change this book using magic. • Kurikuri, Kurikuri-tongtong!			
• I like cookies. • It looks delicious • Thank you, loopy			

Scene image generation from given query sentences. The generated scenes are synthesized by the weighted overlapping of image patches associated with the words in the sentences based on the constructed knowledge. This mechanism is inspired by the cross-modal reconstruction of mental imagery upon stimuli in human brains. When a child hears dialogue sentences, that is, he recalls the scenes or images related to the sentences. As the number of observed videos increase, the images become more complex and diverse. Comparing two query sentences, the first query sentences are related to *Tongtong*, a dragon magician and the second sentences are associated with *Loopy*, a chatter girl who likes cooking. Note that *Tongtong* does not appear until episode 56 and he casts "*Kurikuri*" for spell. Therefore, the images generated by the model learning from episode 1 to episode 52 seem to be unrelated to the first query sentences. However, once the concepts on *Tongtong* were constructed by observing *Tongtong*-related episodes, various images related to *Tongtong* recalled from the query sentences. Dissimilar to *Tongtong*, *Loopy* continuously comes on since episode 1 even if she less frequently appears than *Pororo* and *Crong*. In addition, she likes to make cookies. From the fact that the recalled images by the second query mainly mostly contain *Loopy*, cookies, and diverse objects related to *Loopy*, we indicate that the concept knowledge constructed by learning of DCHs enables the translation to be story-aware.

- Figure S3: Scene-to-sentence generation

Scenes	Episodes 1~13		Episodes 1~36	
	Original	And petty taught loopy how to ski.		
	UGMC	- did you ask me how to swim. - the end how grateful I think she is coming.	UGMC	- Wow petty that how that is not always so loopy taught if i can do fly it well. - How did you have to be that I could ski just.
	PRGMC	- end how was so happy - the end how did you I would	PRGMC	- How did you pororo you. - How about now you can I do not worry.
	FGMC	- To show how big you found - The end how grateful I am petty nice to lose careful	FGMC	- Harry realized that how that is it is dangerous - I thought that how that I could ski just
	SPC	- But how do someone stop. - The end how was it.	SPC	- How about now you can you give me that how that is great. - I will see let see how big.
	Original	Wow poby, you caught so many already.		
	UGMC	- Has been caught	UGMC	- Come out if you go in to hear you guys you have got a lot of fish I caught. - You have caught a lot today did you see you later.
	PRGMC	- Has been caught	PRGMC	- Everyone has caught a fish for dinner. - You have caught a lot today did you ask me how.
	FGMC	- What are you guys you have caught a lot. - What happened to ten everyone has caught a lot.	FGMC	- Poby caught a boat a secret that all the wind is so big. - You have caught a fish for the art diving.
	SPC	- Pororo no pororo has caught - She caught the first place	SPC	- You come with his new friend has caught a very interesting book recently - What about pororo has caught a lot of fish

Story-aware subtitle generation for given scene images. The subtitles are generated by aligning the words associated with the image patches contained in the given scenes. SPC denotes a model with no concept layer. The first image is an observed scene and the second is not observed by the models. Similar to Figure S2, for both the images, the model which observes more videos generates not only more complex but also descriptive sentences with diverse words. Furthermore, the model learned by FGMC provides more accurate and descriptive sentences, compared to those by PRGMC.