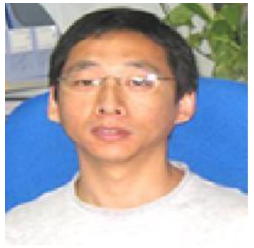
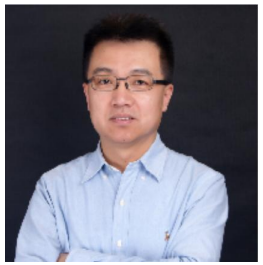
Neural Divide-and-Conquer Reasoning Framework

A Neural Divide-and-Conquer Reasoning Framework for Image Retrieval from Linguistically Complex Text

Yunxin Li, Baotian Hu, Yuxin Ding, Lin Ma, Min Zhang





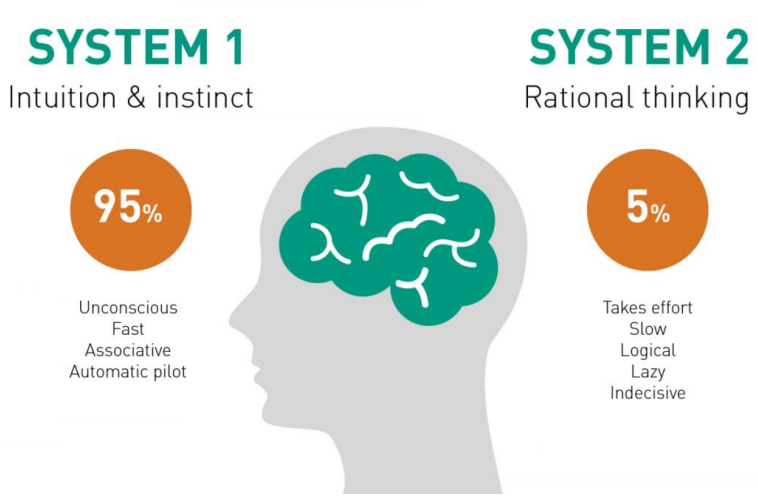


Yunxin Li, Baotian Hu, et al. A Neural Divide-and-Conquer Reasoning Framework for Image Retrieval from Linguistically Complex Text. *ACL 2023*.

Neural Divide-and-Conquer Reasoning Framework

Task: Image Retrieval from Linguistically Complex Text

- Divide-and-Conquer is a strategy of solving a large problem by breaking the problem into smaller sub-problems, solving the sub-problems, and combining them to get the desired output.
- Dual-Process Theory for human thinking: human brains contain two thinking systems: System 1 performs analogical reasoning well, System 2 is capable of abstract logical reasoning, well-suitable for complex reasoning problems.



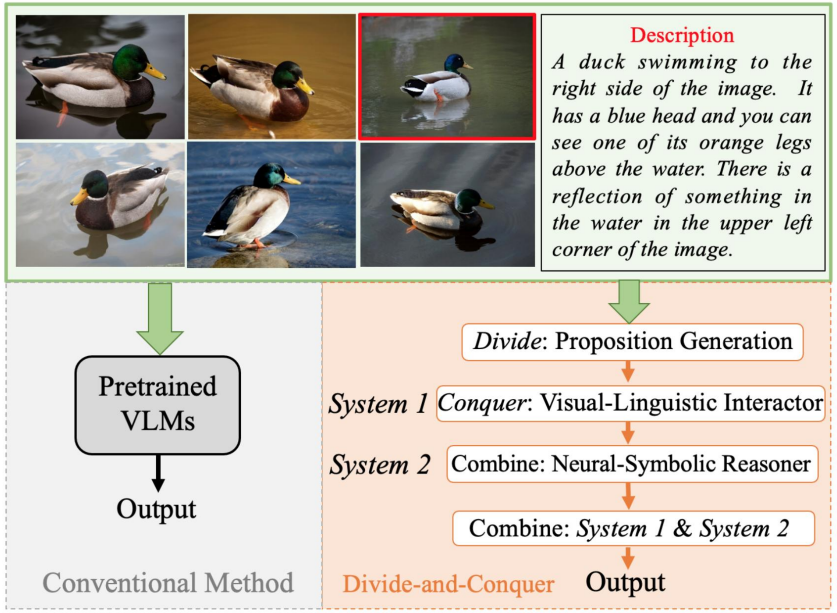
> *The automatic operations of System 1 generate surprisingly complex patterns of ideas, but only the slower System 2 can construct thoughts in an orderly series of steps.*
>
> – **Daniel Kahneman in** *Thinking, Fast and Slow*

System 1 and System 2

Yunxin Li, Baotian Hu, et al. A Neural Divide-and-Conquer Reasoning Framework for Image Retrieval from Linguistically Complex Text. *ACL 2023*.

Neural Divide-and-Conquer Reasoning Framework

1. Pretrained VLMs foucs on analogical reasoning as Sysem 1 based on the analysis of deep learning networks (Bengio, 2017, 2019), performing well on simple text-image retrieval.
2. When confronted with complex text, VLMs performance drops drastically.
3. We may need a logical reasoning System 2 to perform this complex retrieval task via logical operation.
4. Combining the advantage of System 1 and System 2 may be a significant way for complex reasoning.
5. System 1 and System 2 can be integrated with the Divide-and-Conquer Strategy.

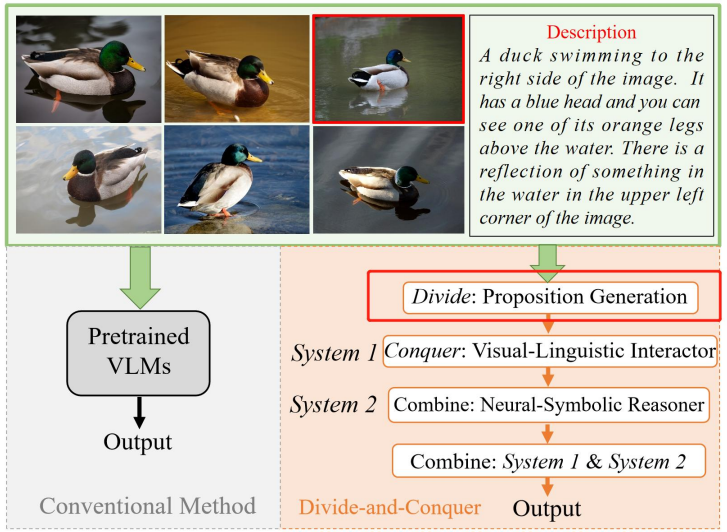


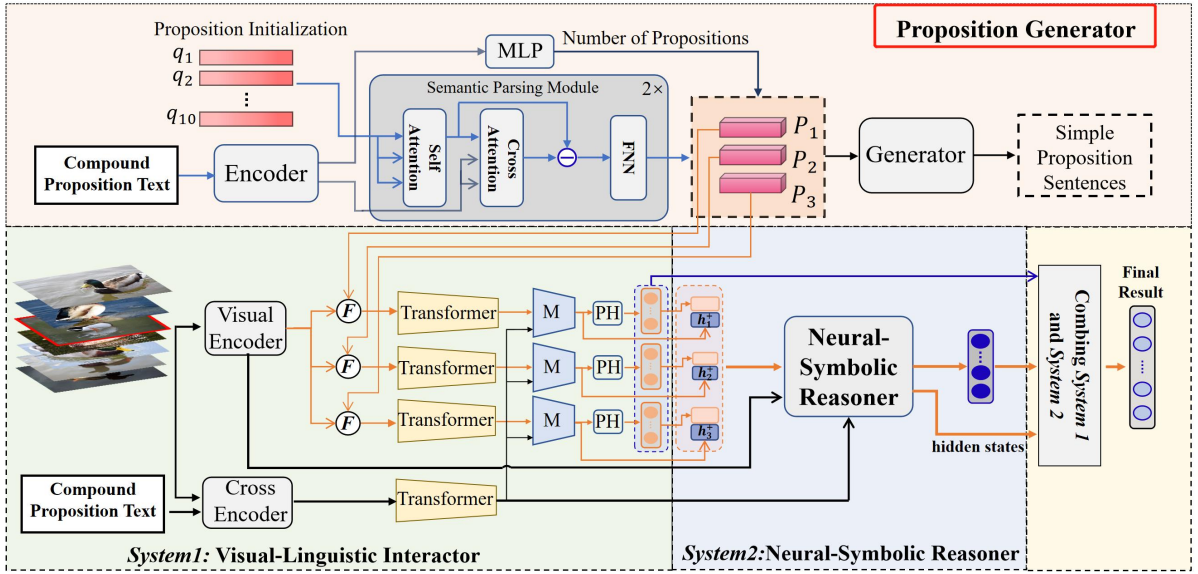
Yunxin Li, Baotian Hu, et al. A Neural Divide-and-Conquer Reasoning Framework for Image Retrieval from Linguistically Complex Text. *ACL 2023*.

Neural Divide-and-Conquer Reasoning Framework

Proposition Generator

1. The proposition generator is a sequence-to-sequence model based on the pretrained language model BART.
2. It aims to decompose the complex proposition text into representations of simple proposition sentences.
3. For explainig what simple propositions represent, we use the decoder of BART to generate the corresponding sentences according to the encoding representations.



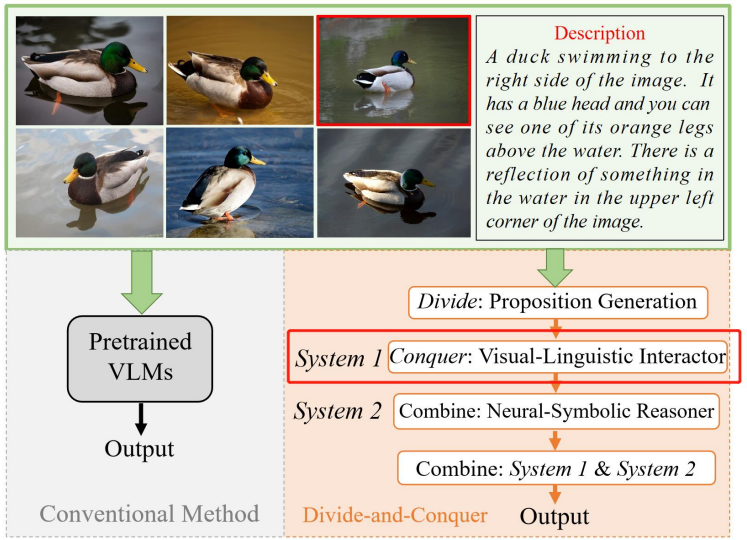


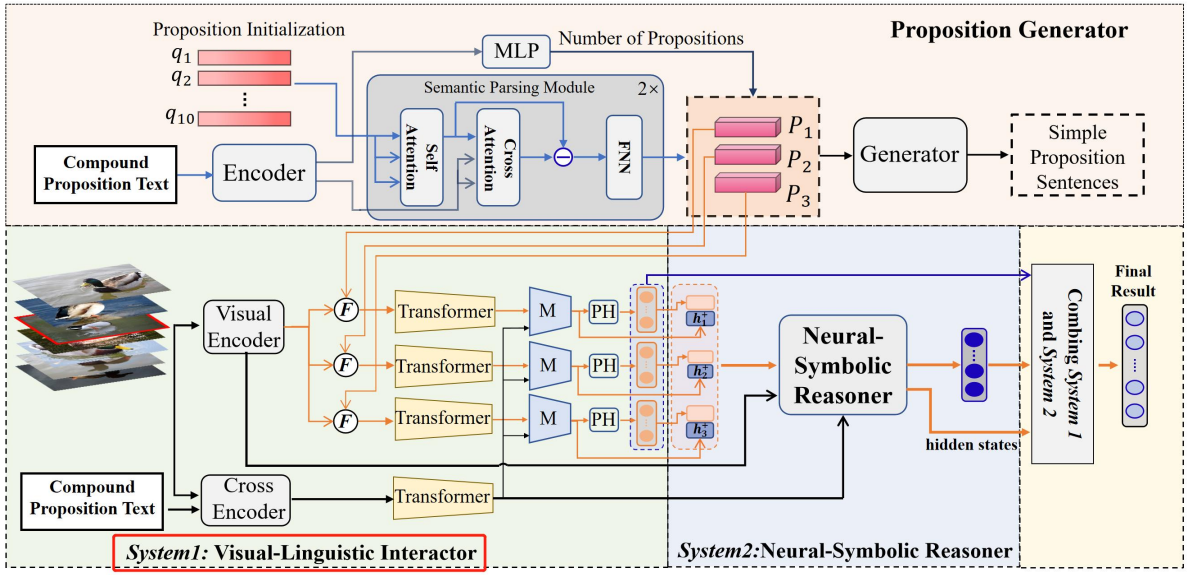
Yunxin Li, Baotian Hu, et al. A Neural Divide-and-Conquer Reasoning Framework for Image Retrieval from Linguistically Complex Text. *ACL 2023*.

Neural Divide-and-Conquer Reasoning Framework

System 1: Visual-Linguistic Interactor

1. This module aims to perform the visual-propositions information interaction, resembles the System 1.
2. This module is based on OFA, a VLM capable of multimodal information interaction.
3. The outputs of this module are matching scores of propositions-images and the reasoning states of propositions-on-images.





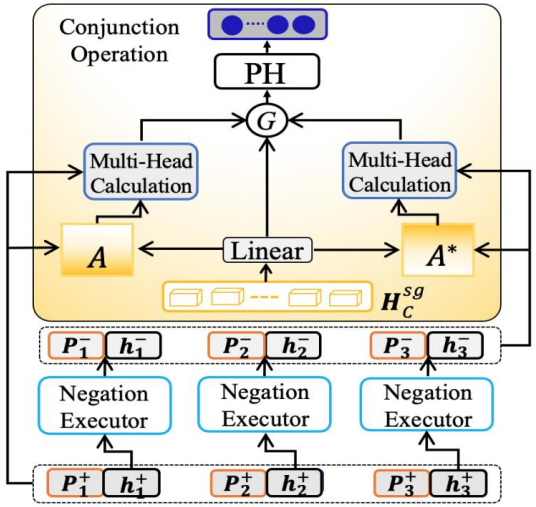
Yunxin Li, Baotian Hu, et al. A Neural Divide-and-Conquer Reasoning Framework for Image Retrieval from Linguistically Complex Text. *ACL 2023*.

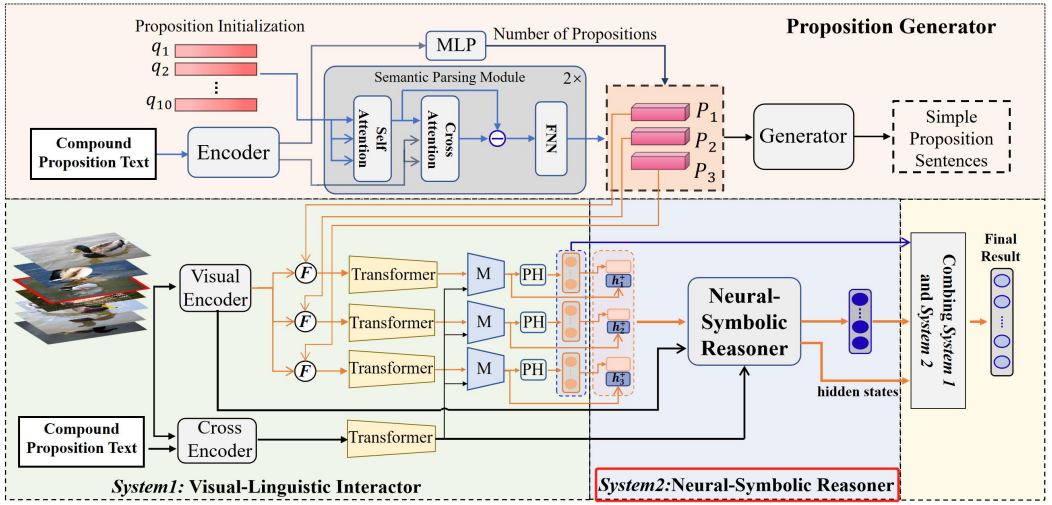
Neural Divide-and-Conquer Reasoning Framework

System 2: Neural-Symbolic Reasoner

1. This module is responsible for integrating the reasoning states and results of simple propositions to obtain the final solution of complex proposition on images.

2. It consists of the neagtion executor and the conjunction operation. Negation executor aims to gain the negational reasoning state of positive reasoning state. Conjunction operation is responsible for obtaining the inference results based on joint positive and negational reasoning states.



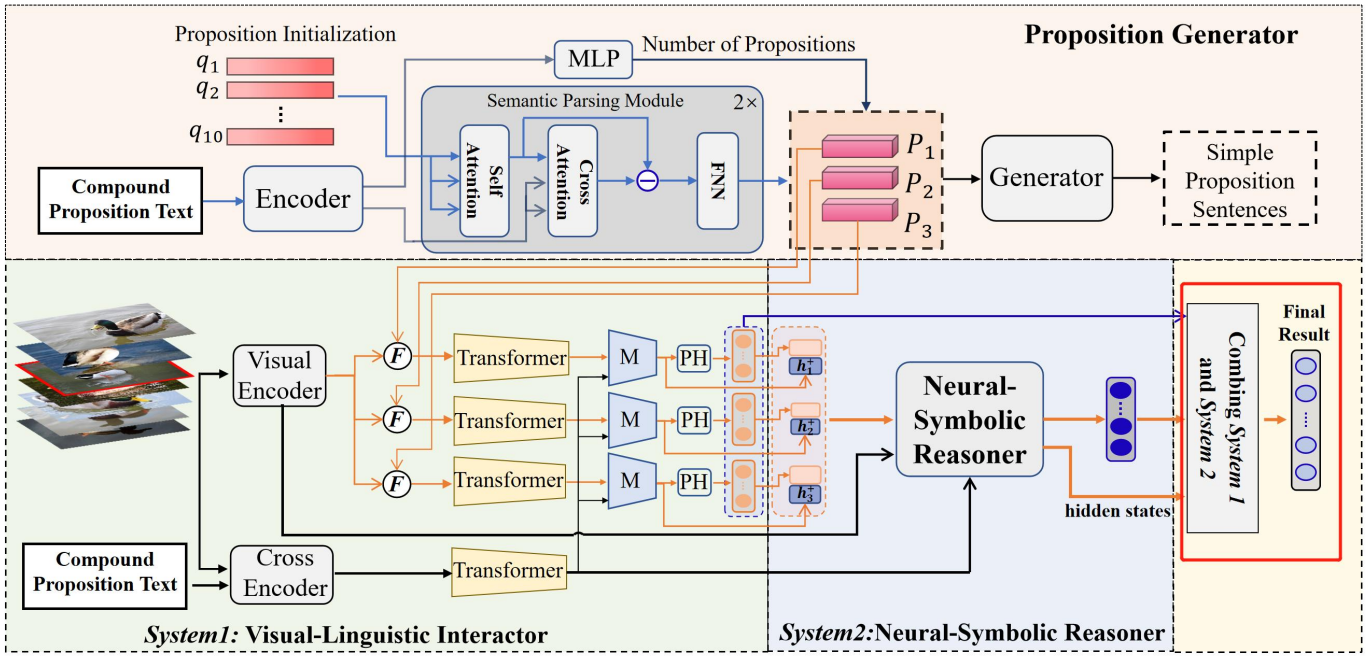


Yunxin Li, Baotian Hu, et al. A Neural Divide-and-Conquer Reasoning Framework for Image Retrieval from Linguistically Complex Text. *ACL 2023*.

Neural Divide-and-Conquer Reasoning Framework

Combining System 1 and System 2

This process is responsible for integrating the inference results of System 1 and System 2 as the final solution. The output of System 1 contains the perceptual calculation results of simple propositions on images. The output of System 2 is the logical inference result for the overall description. By doing so, the whole system utilizes the advantages of analogical inferring System 1 and logical reasoning System 2.



Yunxin Li, Baotian Hu, et al. A Neural Divide-and-Conquer Reasoning Framework for Image Retrieval from Linguistically Complex Text. *ACL 2023*.

Neural Divide-and-Conquer Reasoning Framework

Experimental Results

Method ↓ Type →	All	Video	Static
CLIP (Radford et al., 2021)	28.4	20.0	<u>60.0</u>
CLIP† (Krojer et al., 2022)	<u>29.9</u>	<u>22.0</u>	59.8
UNITER (Chen et al., 2020)	24.8	17.4	52.8
UNITER† (Krojer et al., 2022)	25.7	19.1	50.5
ViLBERT (Lu et al., 2019)	20.9	15.0	42.7
ViLBERT† (Krojer et al., 2022)	24.5	18.0	49.3
NDCR (ours)	**34.1**	**26.1**	**64.3**

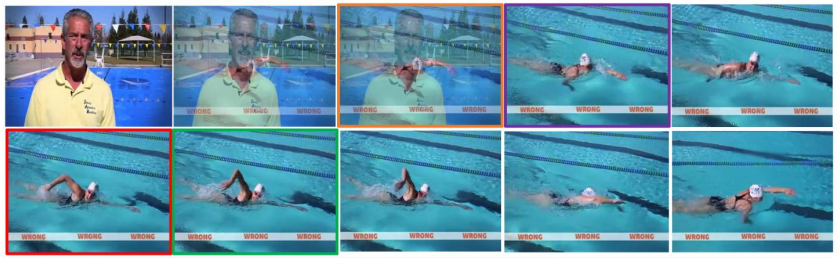
Table 1: Model performance (accuracy) on **original testing set**. The results of CLIP, UNITER, ViLBERT, and their variants(†) are reported by Krojer et al. (2022). The underscore and bold indicate the second highest value and best performance (same as following tables). We report results for all examples and two disjoint subsets: video frames and static images.

Method ↓ Type →	All	Video	Static
OFA (Wang et al., 2022)	29.0	22.1	54.8
OFA†	<u>30.0</u>	<u>23.6</u>	54.6
CLIP (Radford et al., 2021)	27.4	19.7	<u>56.5</u>
CLIP† (Krojer et al., 2022)	27.6	20.8	53.2
NDCR (ours)	**32.8**	**25.7**	59.2
System 2	32.4	25.3	**59.3**
System 2 w/o Negation	32.0	25.3	57.3
System 1	31.6	24.5	58.3
System 1 w/o Modifier	19.3	16.4	30.3

Table 2: Ablation experiments on the **testing*** set, where we manually label the testing set to conduct ablation studies. 'Negation' and 'Modifier' indicate the negation executor and modifier. We adopt the mean pooling method to aggregate the predicted results of simple proposition in System 1 and w/o Modifier.

Neural Divide-and-Conquer Reasoning Framework

Case Analysis



Compound Propositin Text: The person in the water is looking at the camera. Most of their right arm is out of the water and there is a splash covering up some of their right hand.

Simple Propositin Sentences: 1. The person in the water is looking at the camera. *2.* Most of the man's right arm is out of the water. *3.* There is a splash covering up some of right hand.

P_1^+ [0.0078, **0.3845, 0.4014**, 0.0066, 0.0589, 0.1189, 0.0121, 0.0015, 0.0017, 0.0066]
P_2^+ [0.0027, 0.0151, 0.0864, 0.0147, **0.2427**, **0.582**, 0.032, 0.0044, 0.0018, 0.018]
P_3^+ [0.0075, 0.2218, **0.3601**, 0.0124, 0.1173, **0.2399**, 0.0237, 0.0031, 0.0026, 0.0117]

Golden Label: 5, NDCR: 5, System 2: 5, System 1: 2, OFA†: 6, CLIP†: 3

Figure 4: A case from the test set, where different colors correspond to the predicted result of models. $P^+_{1,2,3}$ represent the inferred confidence scores of simple proposition sentences in System 1 and are used to obtain the results in System 2 and final combination process.

Compound Propositin Text: There is no text. The knife is near the center.

Simple Propositin Sentences: 1. There is no text. *2.* The knife is near the center.

P_1^+ [0.0594, 0.0828, 0.1096, 0.0818, 0.0818, 0.068, **0.1527**, 0.0596, **0.1522, 0.1522**]
P_2^+ [0.0497, 0.0574, 0.1179, 0.1207, 0.0801, 0.0842, **0.1238**, 0.0813, **0.2438**, 0.0413]

Golden Label: 8, NDCR: 8, System 2: 8, System 1: 6, OFA†: 7, CLIP†: 9

Figure 5: Another case from the test set, where it contains two simple proposition sentences.

Take Home Message

- Neural symbolic calculation may be a worthwhile approach to improve the compositional reasoning and planning capability of large language models.
- Divide-and-Conquer is similar to the self-asking chain-of-the-thought, aiming to decompose complex reasoning into simple problems and construct the reasoning path. Both are effective for solving complex problems.
- Dual-Process Theory could be integrated with the Divide-and-Conquer.