Multimodal Chain-of-Thought Reasoning: A Comprehensive Survey

Yaoting Wang[1], **Shengqiong Wu**[1], **Yuechen Zhang**[2], **Shuicheng Yan**[1], **Ziwei Liu**[3], **Jiebo Luo**[4], **Hao Fei**[1*]
[1]NUS, [2]CUHK, [3]NTU, [4]UR

Survey Project: https://github.com/yaotingwangofficial/Awesome-MCoT

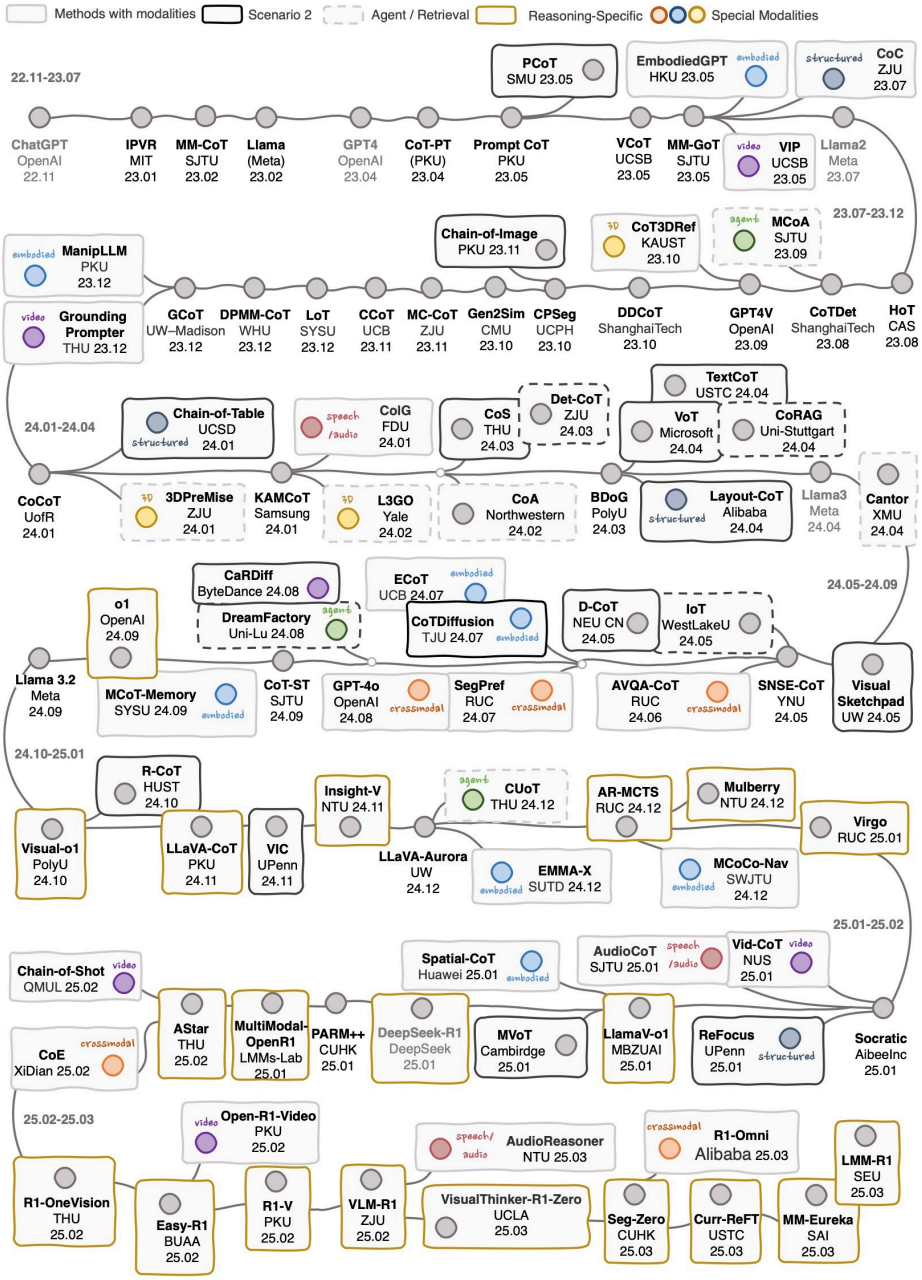

Figure 1: Developing timeline of Multimodal Chain-of-Thought (MCoT) reasoning. Models with names in gray are text-only LLMs. For clarity, the models in the figure are assumed to include the image modality by default, unless specified with special modalities indicated by colored circles.

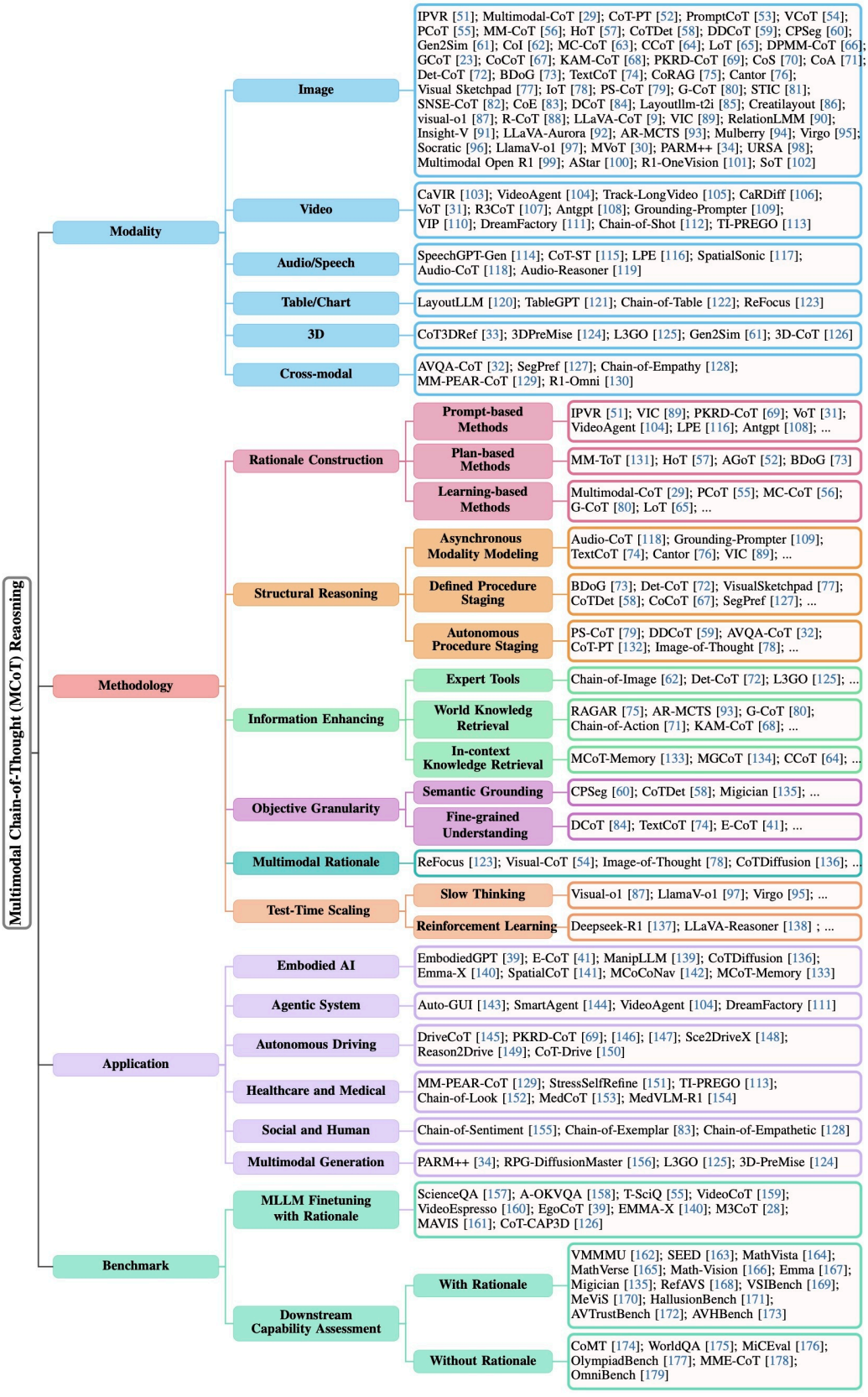

Figure 2: Taxonomy of MCoT reasoning.

Terms	Abbrev.	Description
In-context Learning	ICL	Prompting LLMs with task-specific examples without additional explicit training.
Chain-of-Thought	CoT	Prompting LLMs to reason step-by-step or breaks complex problems into logical steps.
Multimodal CoT	MCoT	Extends CoT to reason with multimodalities, *e.g.*, audio, image.
Cross-modal CoT		Reasoning with two or more multimodalities, *e.g.*, audio-visual.
Thought		A single reasoning step in CoT.
Rationale		Built upon multiple thoughts to support the final answer.

Table 1: Interpretation of MCoT-related terms.

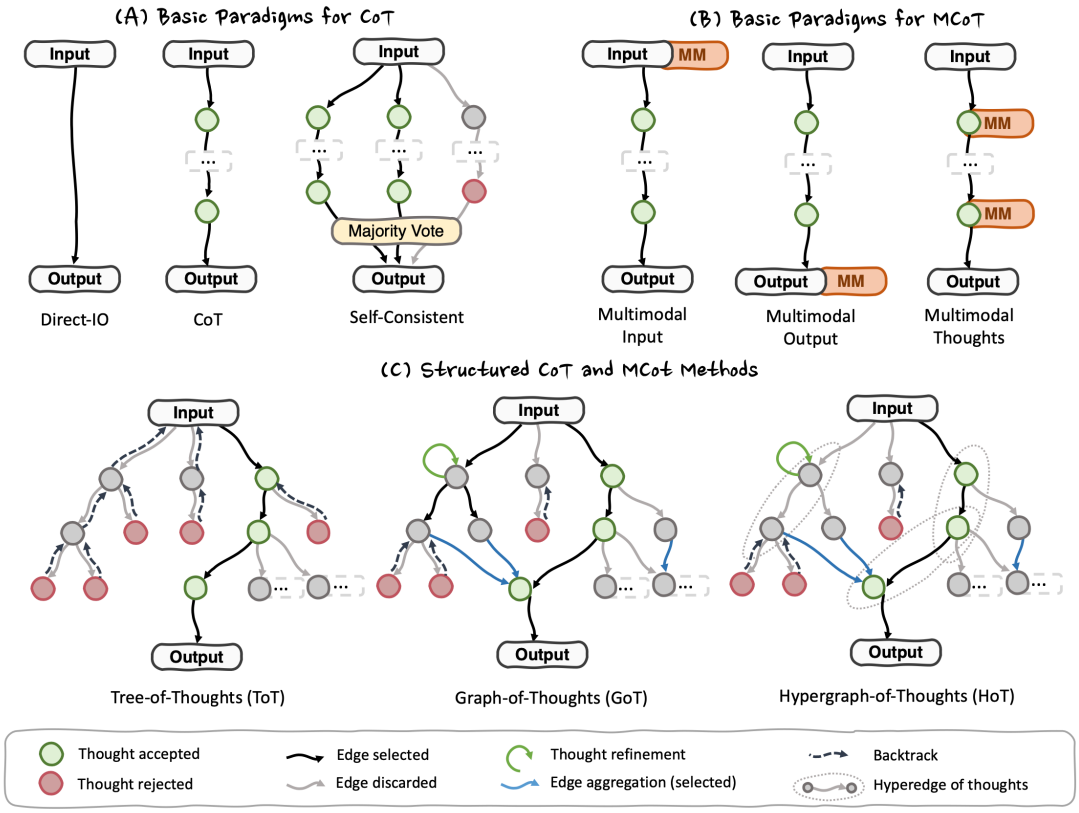

Figure 3: Different thought paradigms of CoT and MCoT.

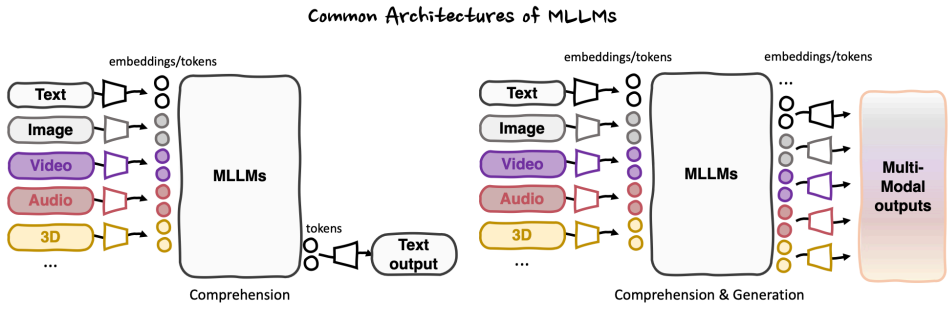

Figure 4: Common architectures for comprehension-only and comprehension-generation MLLMs.

Audio QA

Input: Based on the given audio, identify the source of the honk?

Rationale: The audio contains sounds related to a vehicle honking. Car honking is commonly heard on roads and streets when a driver uses the horn to alert other vehicles or pedestrians.

Answer: the most likely source of the honk is a car.

Visual QA

Input: Which nutrients are mainly provided by the foods in the image?
(A) Vitamins (B) Protein (C) Fats

Rationale: Oranges are in the image and they are commonly known to be a good source of Vitamin C. Therefore, the nutrient mainly provided by the foods in the image is Vitamins.

Answer: (A) Vitamins

Image Grounding

Input: The object can open bottle of beer.

Rationale: sharp blade with a pointed end to insert into the bottle cap.

Answer:

3D Synthesis

Input: A bench with five legs.

Rationale: Part specs generation… spacial specs generator… coordinate calculator… run command… spatial critic… completion critic….

Answer:

Video QA

Input: Why is the person running?

Rationale: First, a group of people is standing in front of an exercise machine and running on the track. Then, the essential function of the exercise machine is to provide resistance for the legs during exercise. This can help improve muscle strength and endurance.

Answer: They might participate in a fitness event.

AVQA

Input: Where is the first sounding instrument?

Rationale: What musical instruments appear in the video? flute, piano. What is the first instrument that sounds in the video? flute. Where is the flute? right

Answer: Right.

Generation

Input: A photo of three sports balls.

Rationale: <IMG A> The number of sport balls in the image does not match the prompt. While the lighting effect is decent, some areas have unnatural reflections and shadowing, which affects the overall realism.

Answer: <IMG B>

Math Reasoning

Input: Find the missing value in this math puzzle.

Rationale:

(5-4)^3=1, (7-3)^3=64,
(8-2)^3=216, (11-8)^3=27.

Answer: the missing value is 27.

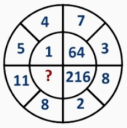

Figure 5: Examples of MCoT applications in various modalities and tasks.

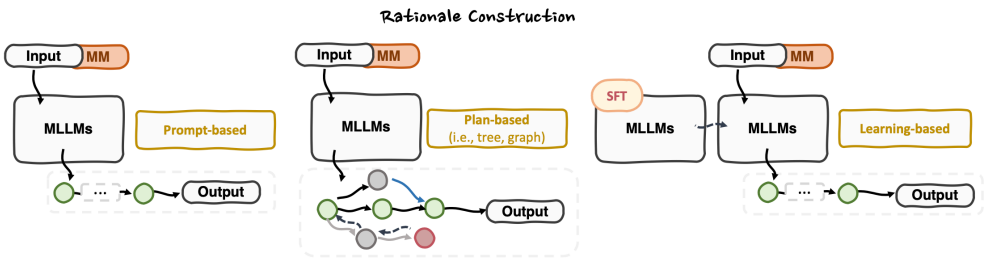

Figure 6: MCoT reasoning methods under different ***rationale construction*** perspectives.

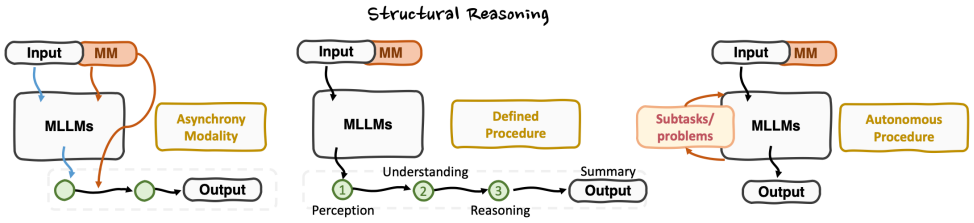

Figure 7: MCoT methods under different ***structural reasoning*** perspectives.

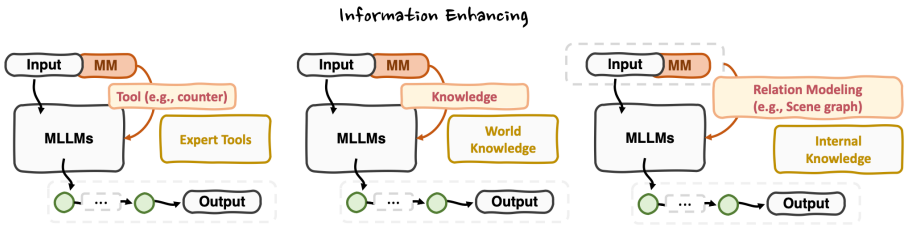

Figure 8: MCoT reasoning under perspectives with ***information enhancing***.

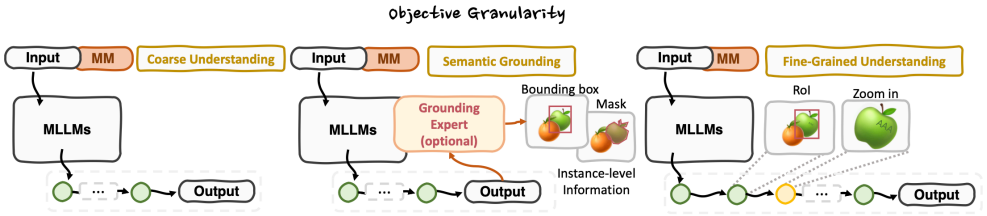

Figure 9: MCoT reasoning under the perspectives of various ***objective granularities***.

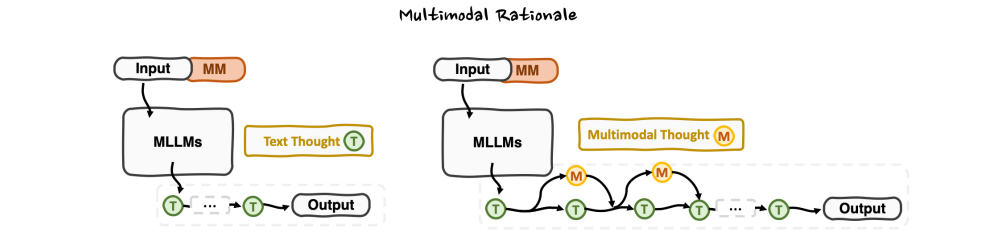

Figure 10: MCoT reasoning with ***multimodal rationale***.

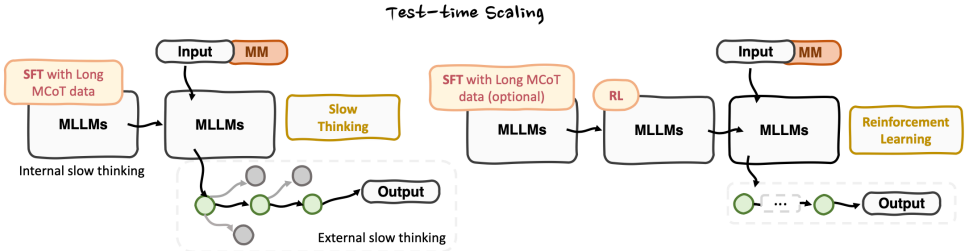

Figure 11: MCoT reasoning with ***test-time scaling*** strategies. RL can help improve reasoning quality, or active long-CoT reasoning ability without annotated long-CoT training data. SFT is optional.