MAMMQA

Rethinking Information Synthesis in Multimodal Question Answering A Multi-Agent Perspective

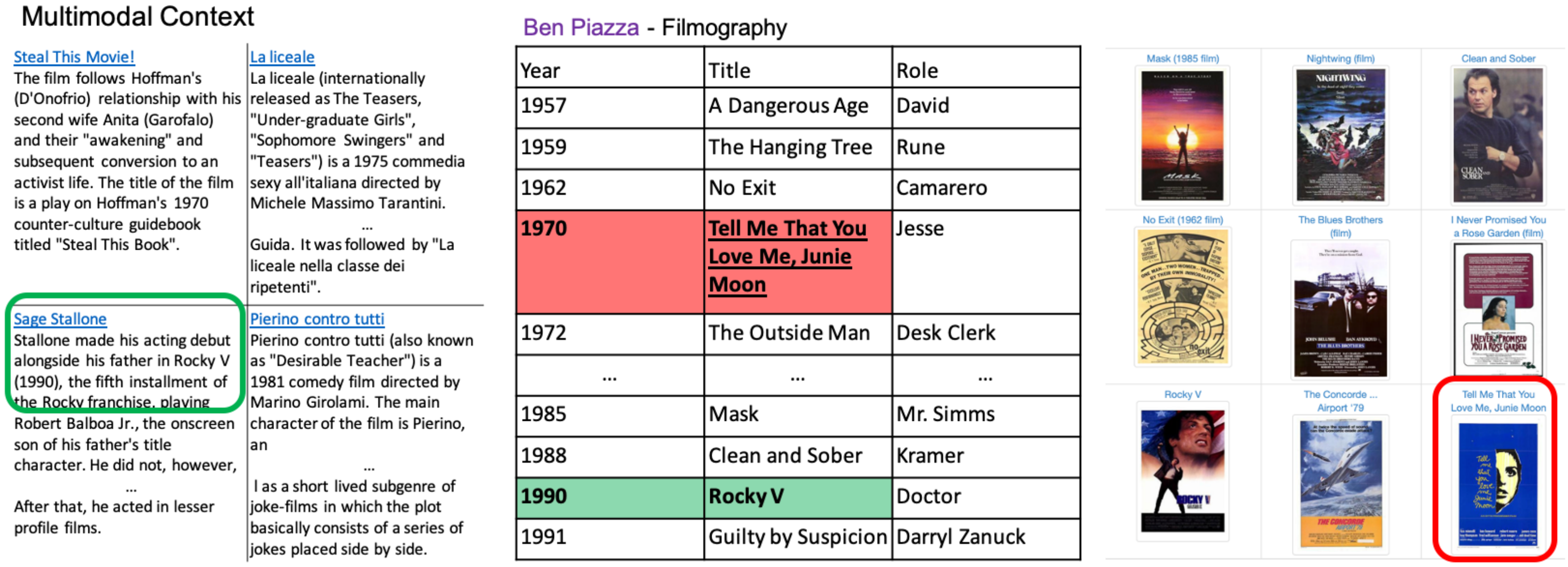
*Tejas Anvekar , *Krishna Singh Rajput, Chitta Baral, Vivek Gupta

Arizona State University

AFNLP IJCNLP AACL 2025 December 20-24 Mumbai, India

* Equal Contribution

Multi Modal QA—Matter?



Multimodal Context

Steal This Movie!
The film follows Hoffman's (D'Onofrio) relationship with his second wife Anita (Garofalo) and their "awakening" and subsequent conversion to an activist life. The title of the film is a play on Hoffman's 1970 counter-culture guidebook titled "Steal This Book".

La liceale
La liceale (internationally released as The Teasers, "Under-graduate Girls", "Sophomore Swingers" and "Teasers") is a 1975 commedia sexy all'italiana directed by Michele Massimo Tarantini.
...
Guida. It was followed by "La liceale nella classe dei ripetenti".

Sage Stallone
Stallone made his acting debut alongside his father in Rocky V (1990), the fifth installment of the Rocky franchise, playing Robert Balboa Jr., the onscreen son of his father's title character. He did not, however,
...
After that, he acted in lesser profile films.

Pierino contro tutti
Pierino contro tutti (also known as "Desirable Teacher") is a 1981 comedy film directed by Marino Girolami. The main character of the film is Pierino, an
...
l as a short lived subgenre of joke-films in which the plot basically consists of a series of jokes placed side by side.

Ben Piazza - Filmography

Year	Title	Role
1957	A Dangerous Age	David
1959	The Hanging Tree	Rune
1962	No Exit	Camarero
1970	**Tell Me That You Love Me, Junie Moon**	Jesse
1972	The Outside Man	Desk Clerk
...	...	...
1985	Mask	Mr. Simms
1988	Clean and Sober	Kramer
1990	Rocky V	Doctor
1991	Guilty by Suspicion	Darryl Zanuck

Mask (1985 film)
Nightwing (film)
Clean and Sober
No Exit (1962 film)
The Blues Brothers (film)
I Never Promised You a Rose Garden (film)
Rocky V
The Concorde ... Airport '79
Tell Me That You Love Me, Junie Moon

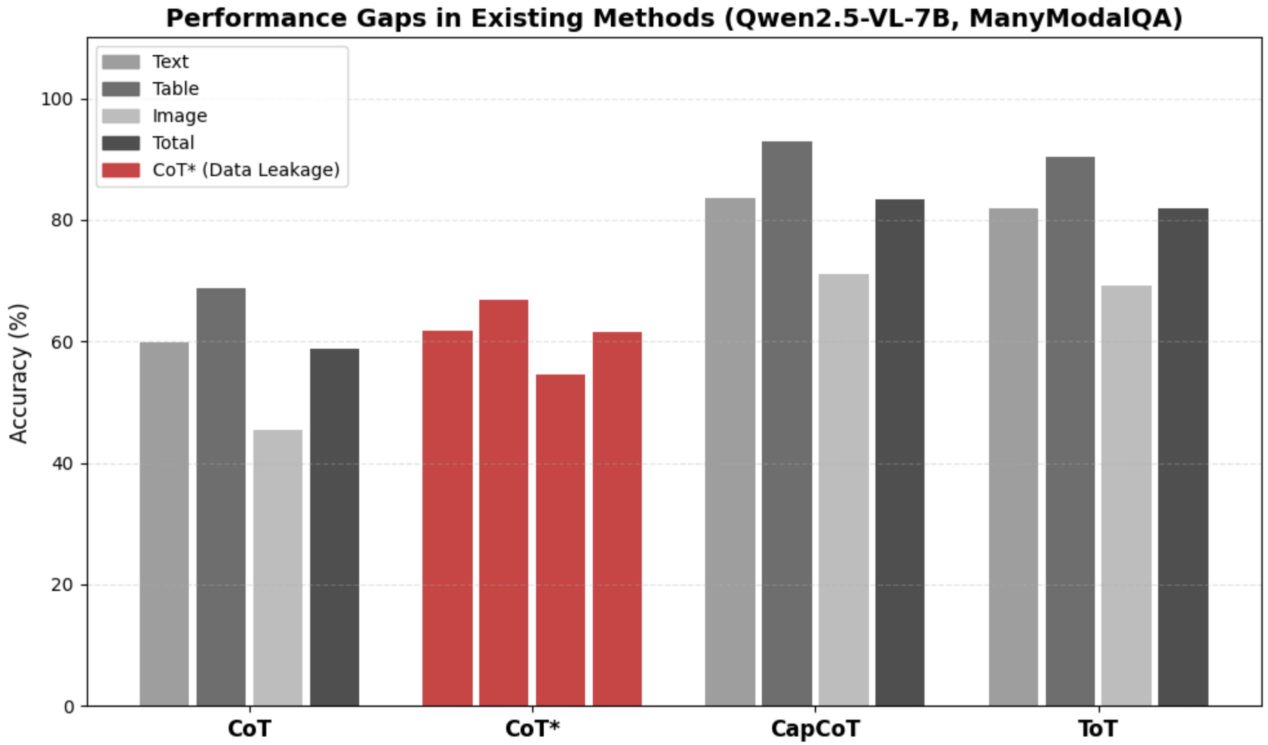
Q: Which B. Piazza title came earlier: the movie S. Stallone's son starred in or the movie with half of a lady's face on the poster?
A: Tell Me That You Love Me, Junie Moon

Real-world information is multimodal; real data spans **tables, text, charts,** and **images** not just plain text.

Enables true cross-modal, multi-hop reasoning Many questions require **combining evidence across formats**; multimodal QA links and synthesizes these signals.

Improves **accuracy, coverage** & **robustness** Multiple modalities provide **complementary evidence**, reducing errors critical for finance, science, analytics, and audits.

Gap in Current Methods

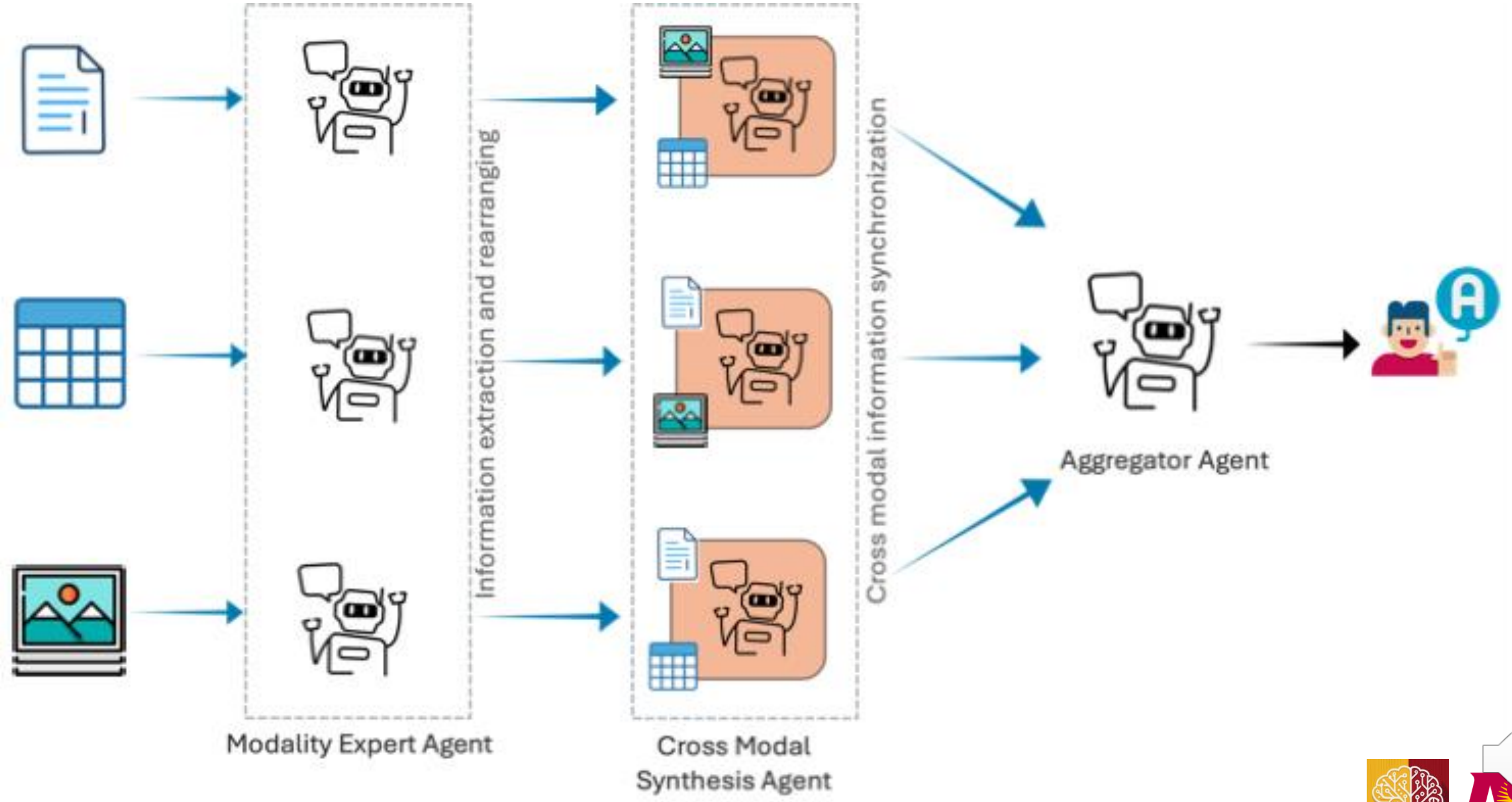


Single Model: high cognitive Burden

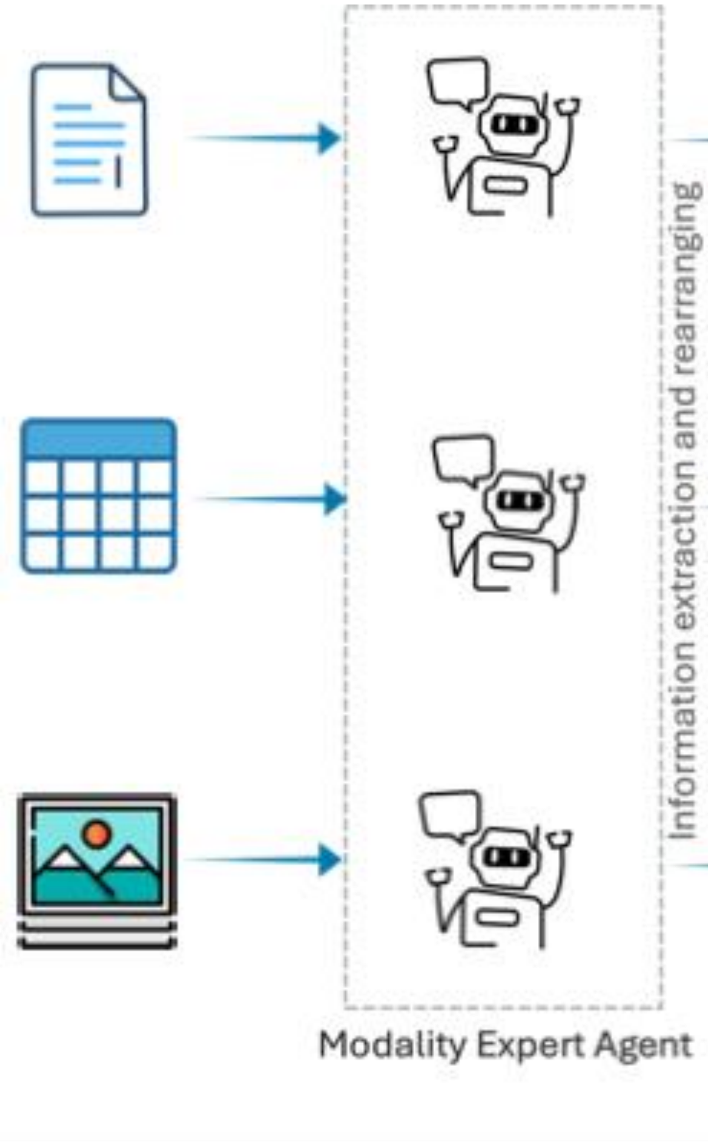
Data leakage & memorized shortcuts

Dynamic agent methods are computationally heavy

MAMMQA

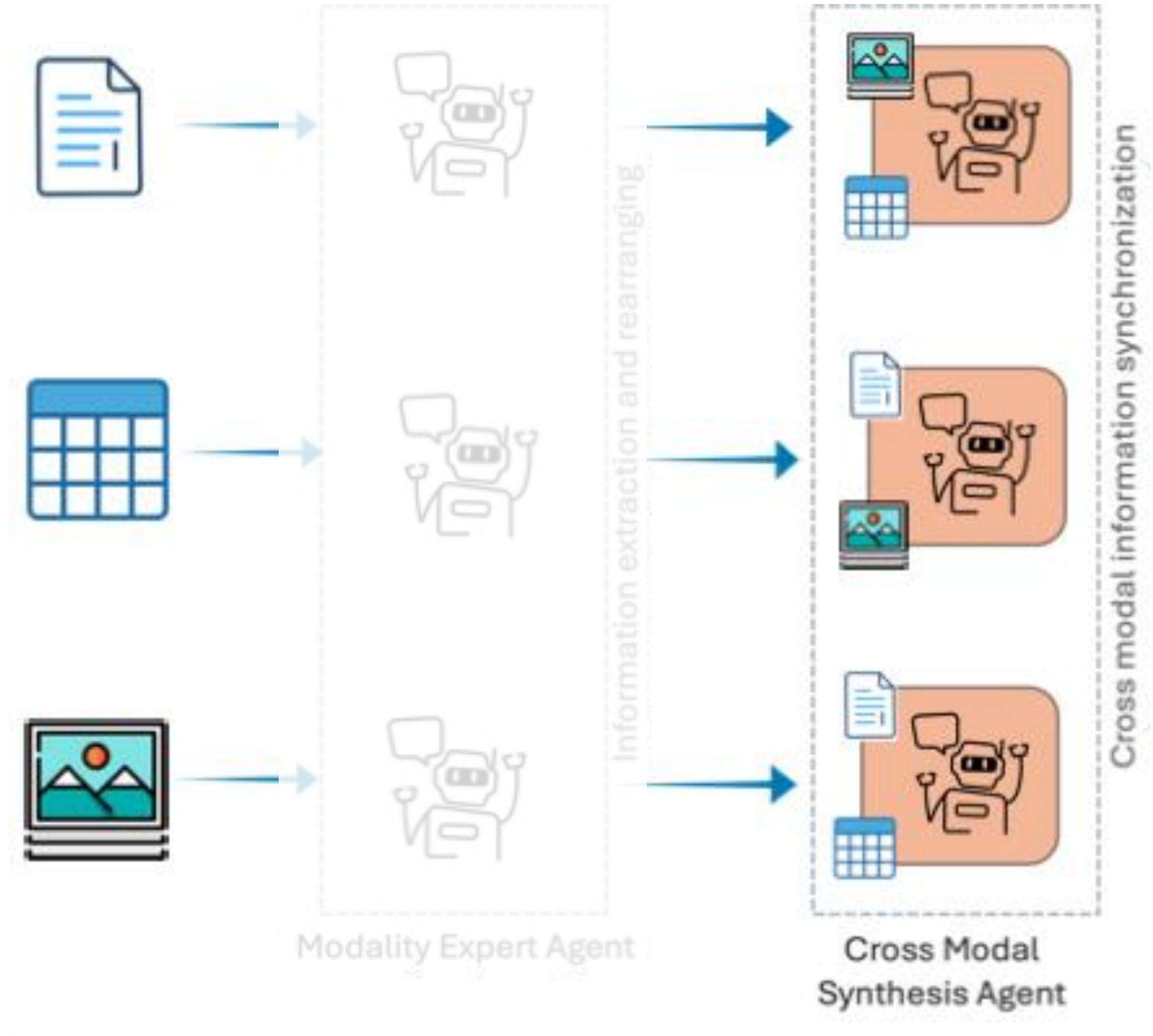

Information extraction and rearranging
Cross modal information synchronization
Aggregator Agent
Modality Expert Agent
Cross Modal
Synthesis Agent

MAMMQA Phase 1 Modality Specific Agents



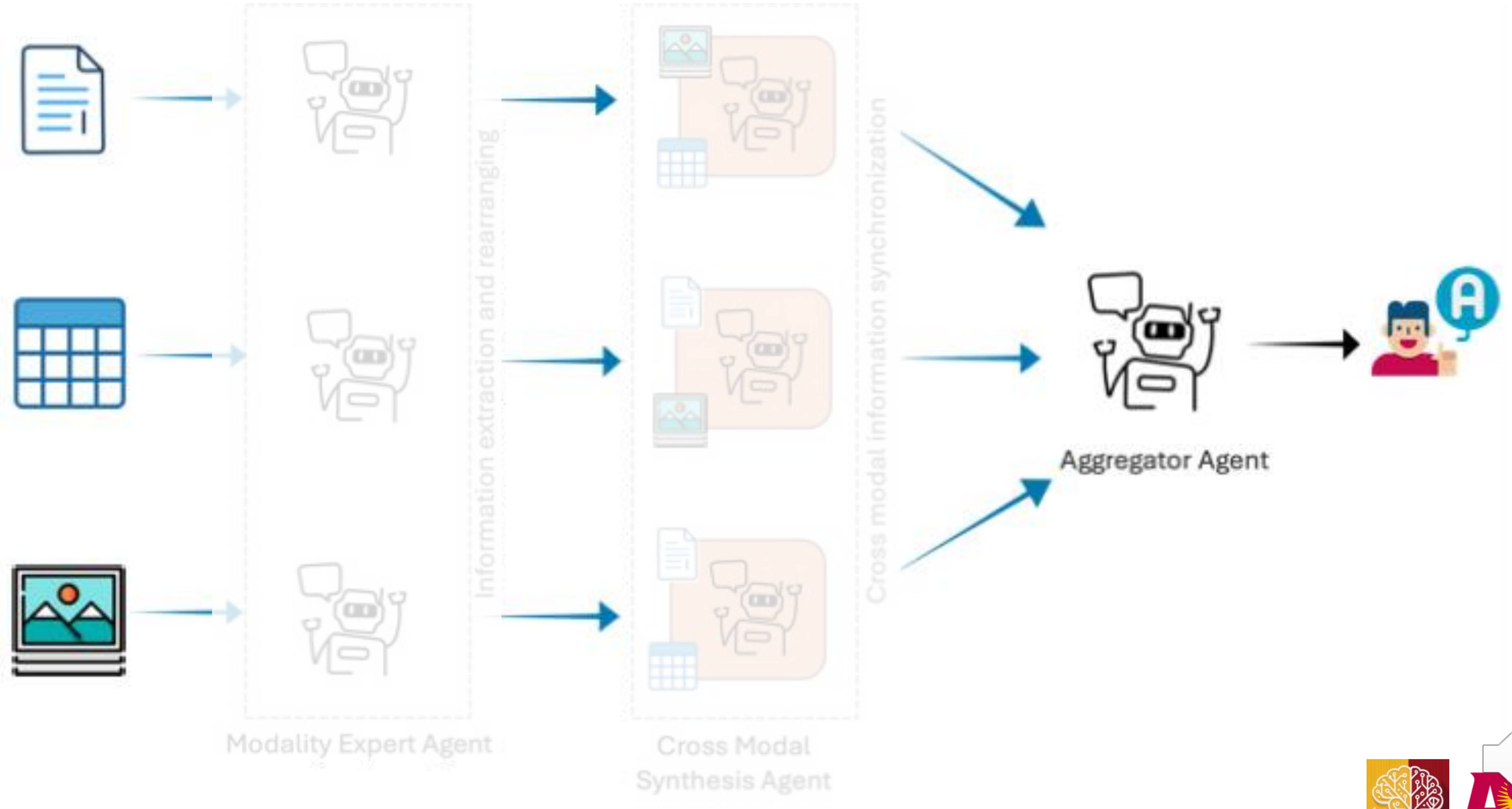
- Modality Specific Agent, disentangles query specific to modality
- Three Experts: **Text, Table, Image**
- **Less** Cognitive load per expert, +**Interpretability**

MAMMQA Phase 2 Cross Modal Agents


Information extraction and rearranging
Cross modal information synchronization
Modality Expert Agent
Cross Modal Synthesis Agent

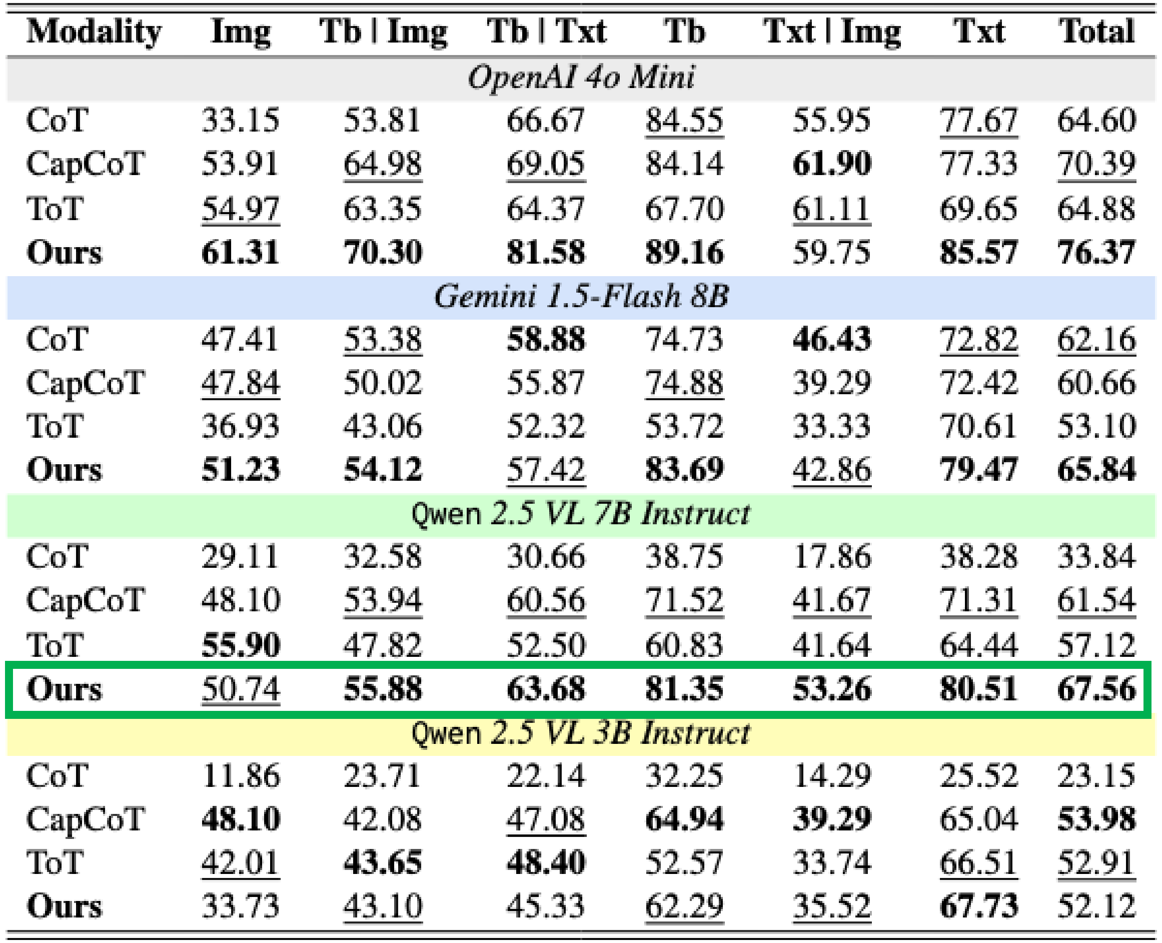
- Cross Info Experts resolves ambiguity between 2 given modality
- Less Hallucination
- Trackable Reasoning

MAMMQA Phase 3 Information Aggregation


Information extraction and rearranging
Cross modal information synchronization
Aggregator Agent
Modality Expert Agent
Cross Modal Synthesis Agent

MAMMQA

Results on MultiModalQA



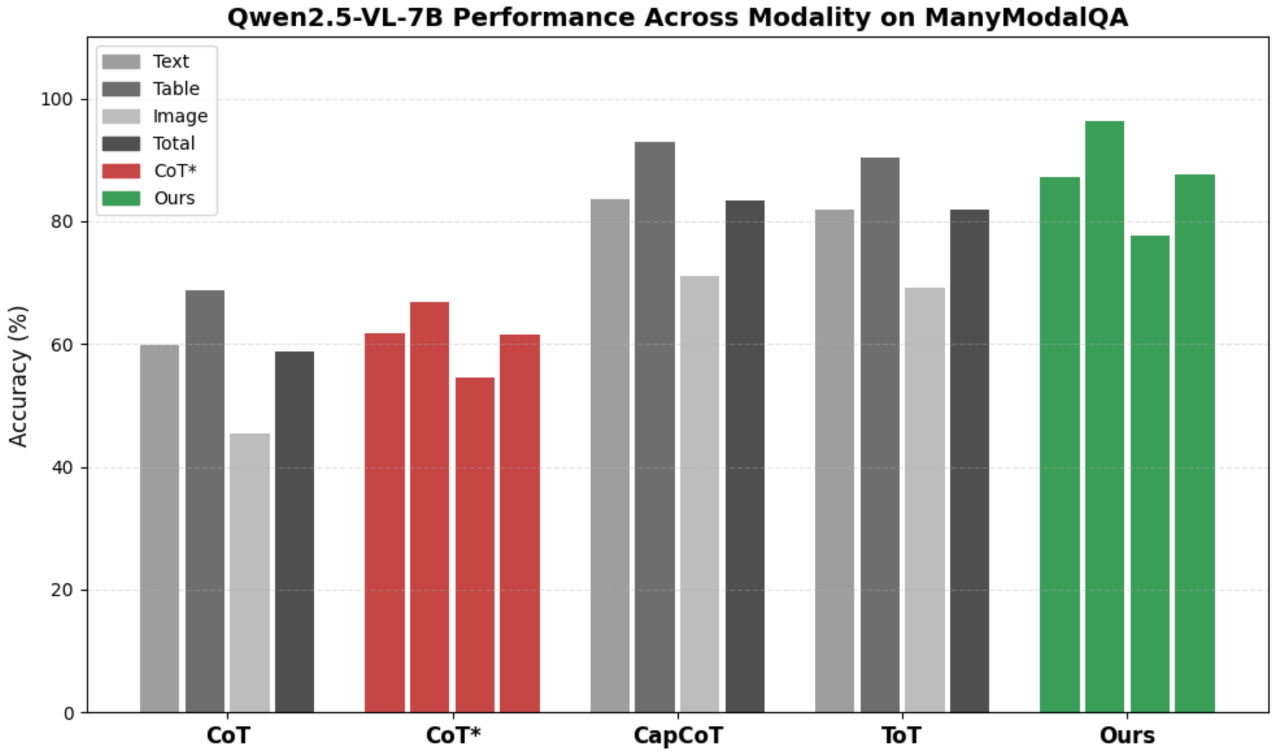
Modality	Img	Tb | Img	Tb | Txt	Tb	Txt | Img	Txt	Total
OpenAI 4o Mini							
CoT	33.15	53.81	66.67	84.55	55.95	77.67	64.60
CapCoT	53.91	64.98	69.05	84.14	**61.90**	77.33	70.39
ToT	54.97	63.35	64.37	67.70	61.11	69.65	64.88
Ours	**61.31**	**70.30**	**81.58**	**89.16**	59.75	**85.57**	**76.37**
Gemini 1.5-Flash 8B							
CoT	47.41	53.38	**58.88**	74.73	**46.43**	72.82	62.16
CapCoT	47.84	50.02	55.87	74.88	39.29	72.42	60.66
ToT	36.93	43.06	52.32	53.72	33.33	70.61	53.10
Ours	**51.23**	**54.12**	57.42	**83.69**	42.86	**79.47**	**65.84**
Qwen 2.5 VL 7B Instruct							
CoT	29.11	32.58	30.66	38.75	17.86	38.28	33.84
CapCoT	48.10	53.94	60.56	71.52	41.67	71.31	61.54
ToT	**55.90**	47.82	52.50	60.83	41.64	64.44	57.12
Ours	50.74	**55.88**	**63.68**	**81.35**	**53.26**	**80.51**	**67.56**
Qwen 2.5 VL 3B Instruct							
CoT	11.86	23.71	22.14	32.25	14.29	25.52	23.15
CapCoT	**48.10**	42.08	47.08	**64.94**	**39.29**	65.04	**53.98**
ToT	42.01	**43.65**	**48.40**	52.57	33.74	66.51	52.91
Ours	33.73	43.10	45.33	62.29	35.52	**67.73**	52.12

Table 1: Quantitative Analysis on MULTIMODALQA dataset

- **MAMMQA +20–30 pp** over CoT→ for small MLLMs like Qwen
- **CapCoT**: Plateau despite captions extracted from enterprise models like gemini.
- **ToT**– fails to scale with huge computational load

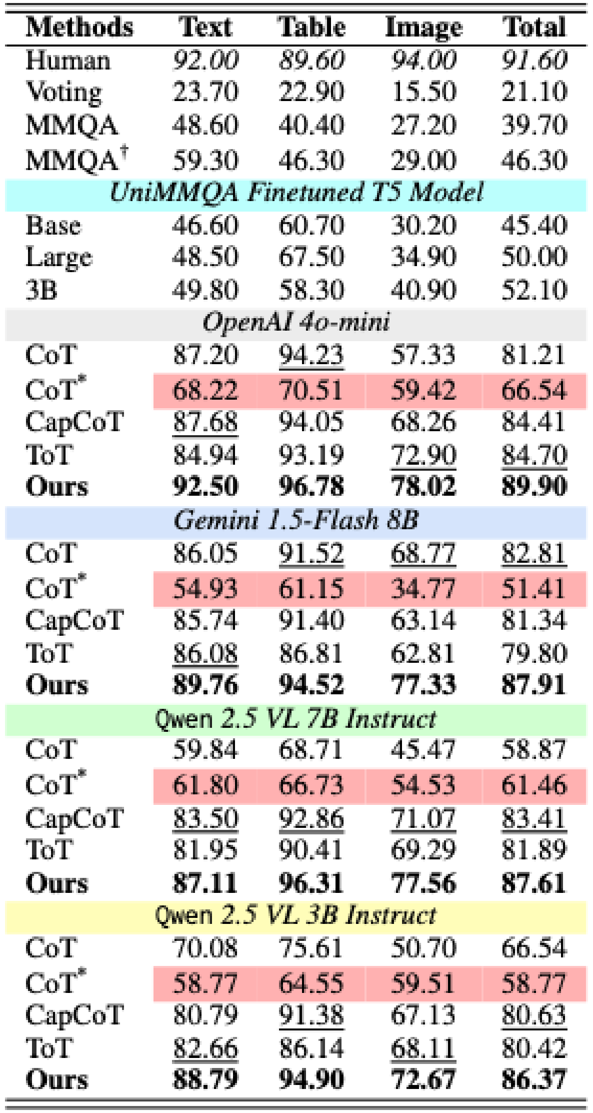
MAMMQA

Results on ManyModalQA



- MAMMQA Outperforms almost all baselines

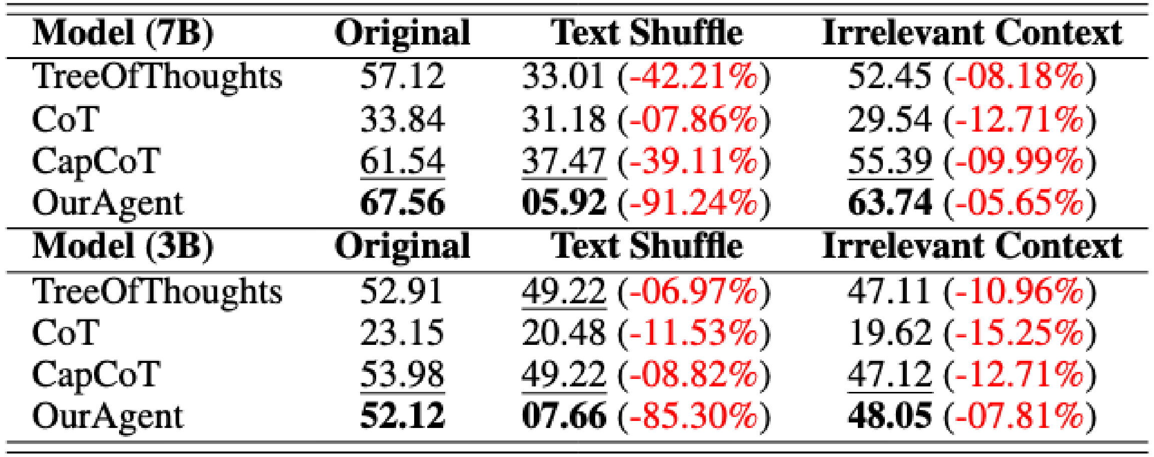
- CoT* shows Data Leakage



Methods	Text	Table	Image	Total
Human	*92.00*	*89.60*	*94.00*	*91.60*
Voting	23.70	22.90	15.50	21.10
MMQA	48.60	40.40	27.20	39.70
MMQA†	59.30	46.30	29.00	46.30
UniMMQA Finetuned T5 Model				
Base	46.60	60.70	30.20	45.40
Large	48.50	67.50	34.90	50.00
3B	49.80	58.30	40.90	52.10
OpenAI 4o-mini				
CoT	87.20	94.23	57.33	81.21
CoT*	68.22	70.51	59.42	66.54
CapCoT	87.68	94.05	68.26	84.41
ToT	84.94	93.19	72.90	84.70
Ours	**92.50**	**96.78**	**78.02**	**89.90**
Gemini 1.5-Flash 8B				
CoT	86.05	91.52	68.77	82.81
CoT*	54.93	61.15	34.77	51.41
CapCoT	85.74	91.40	63.14	81.34
ToT	86.08	86.81	62.81	79.80
Ours	**89.76**	**94.52**	**77.33**	**87.91**
Qwen 2.5 VL 7B Instruct				
CoT	59.84	68.71	45.47	58.87
CoT*	61.80	66.73	54.53	61.46
CapCoT	83.50	92.86	71.07	83.41
ToT	81.95	90.41	69.29	81.89
Ours	**87.11**	**96.31**	**77.56**	**87.61**
Qwen 2.5 VL 3B Instruct				
CoT	70.08	75.61	50.70	66.54
CoT*	58.77	64.55	59.51	58.77
CapCoT	80.79	91.38	67.13	80.63
ToT	82.66	86.14	68.11	80.42
Ours	**88.79**	**94.90**	**72.67**	**86.37**

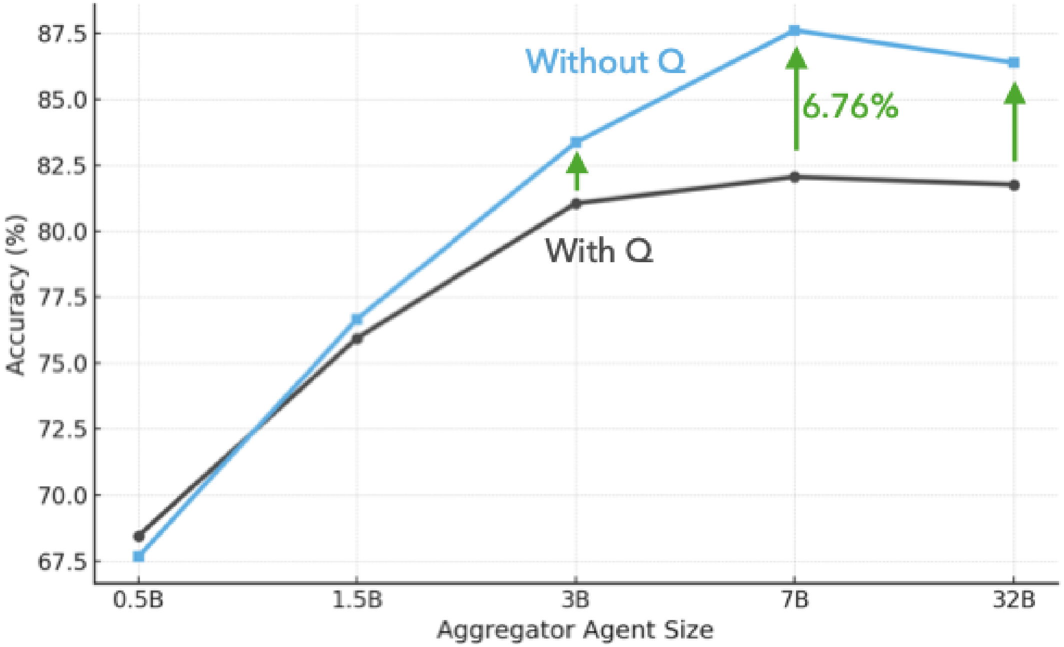
MAMMQA

Robustness Calibration



Model (7B)	Original	Text Shuffle	Irrelevant Context
TreeOfThoughts	57.12	33.01 (-42.21%)	52.45 (-08.18%)
CoT	33.84	31.18 (-07.86%)	29.54 (-12.71%)
CapCoT	61.54	37.47 (-39.11%)	55.39 (-09.99%)
OurAgent	**67.56**	**05.92** (-91.24%)	**63.74** (-05.65%)
Model (3B)	**Original**	**Text Shuffle**	**Irrelevant Context**
TreeOfThoughts	52.91	49.22 (-06.97%)	47.11 (-10.96%)
CoT	23.15	20.48 (-11.53%)	19.62 (-15.25%)
CapCoT	53.98	49.22 (-08.82%)	47.12 (-12.71%)
OurAgent	**52.12**	**07.66** (-85.30%)	**48.05** (-07.81%)

Table 5: Robustness of different reasoning strategies under perturbations across model sizes.



MAMMQA better handles Perturbation like Text Shuffle and Irreverent Context, depicting strong evidence based grounding

MAMMQA shows consistent performance boost when Aggregating information without query

Closing Notes

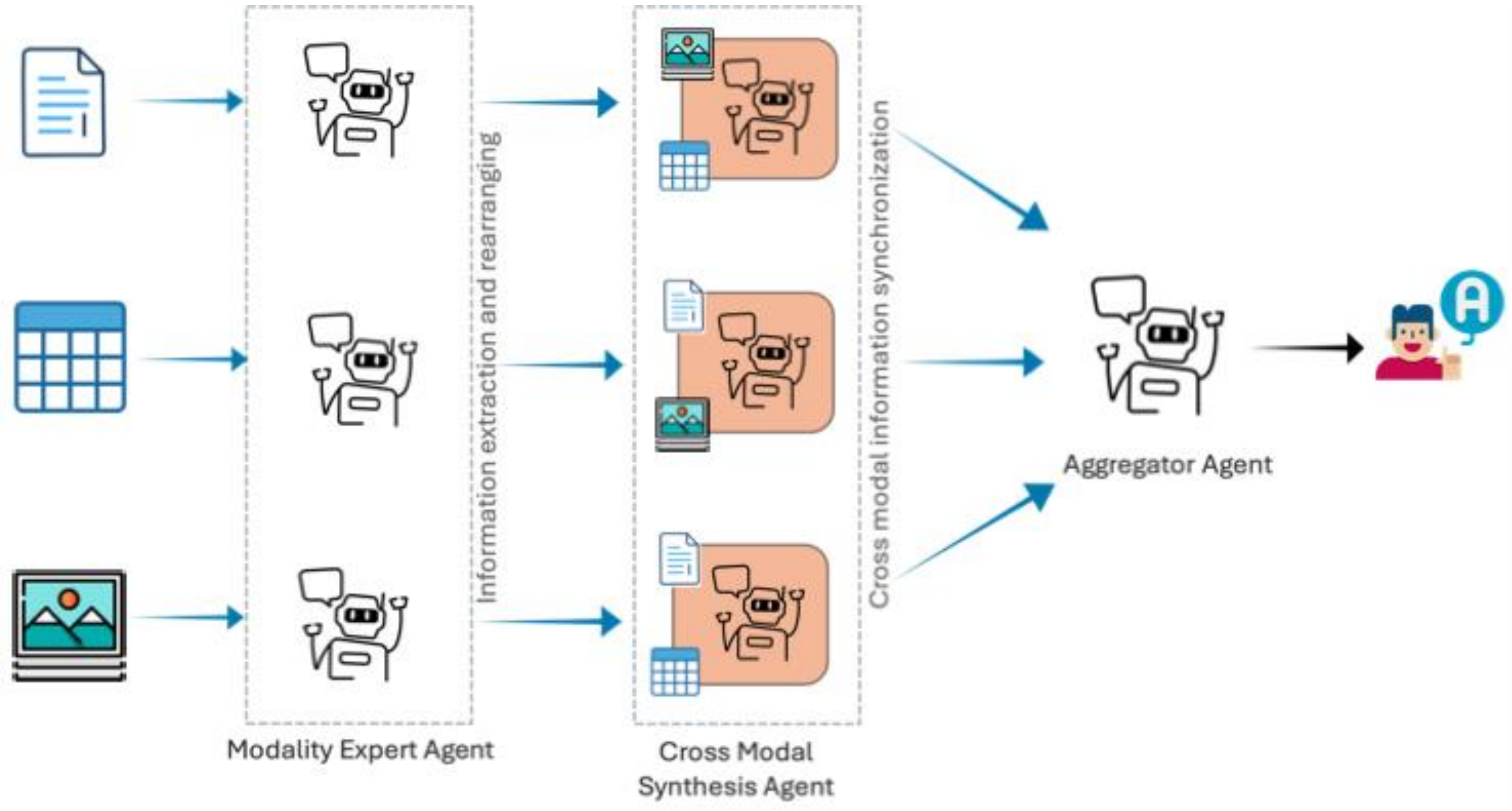
Closing Notes

- **Agentic Method** → reduces cognitive load, traceable reasoning

Closing Notes

- **Agentic Method** → reduces cognitive load, traceable reasoning

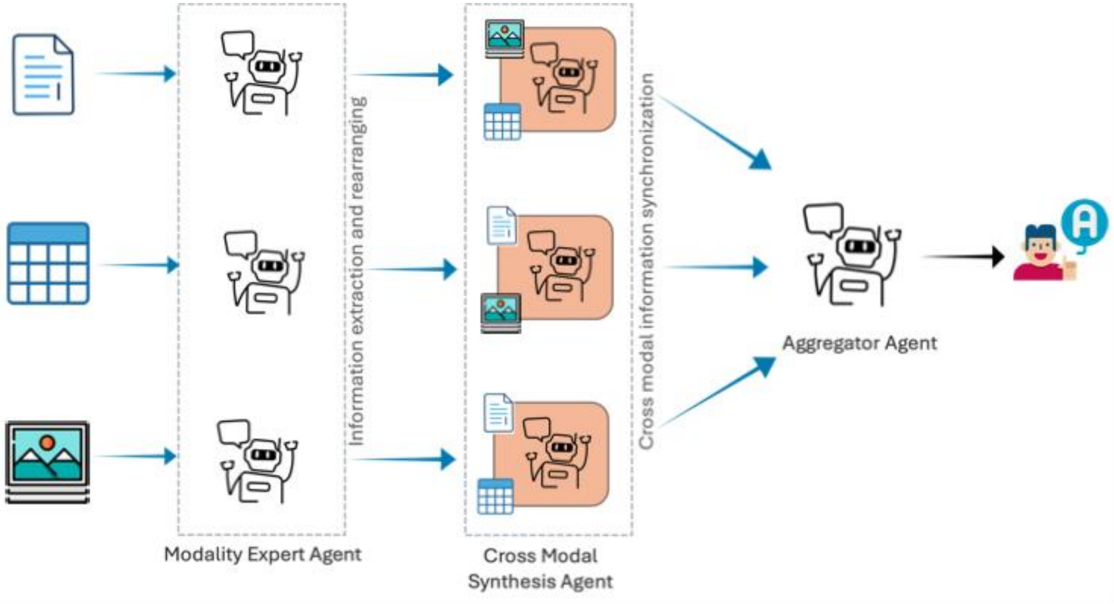
- **MAMMQA** → 3-stage MLLM agent (Extract ➔ Compare ➔ Answer) catching fine-grained ambiguity



Information extraction and rearranging
Cross modal information synchronization
Modality Expert Agent
Cross Modal Synthesis Agent
Aggregator Agent

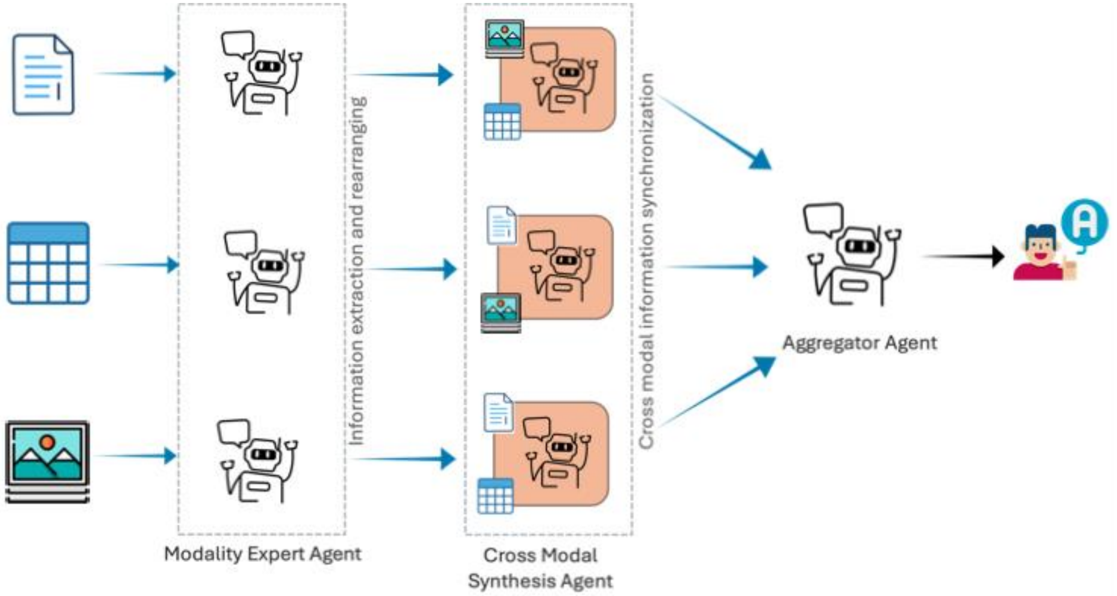
Closing Notes

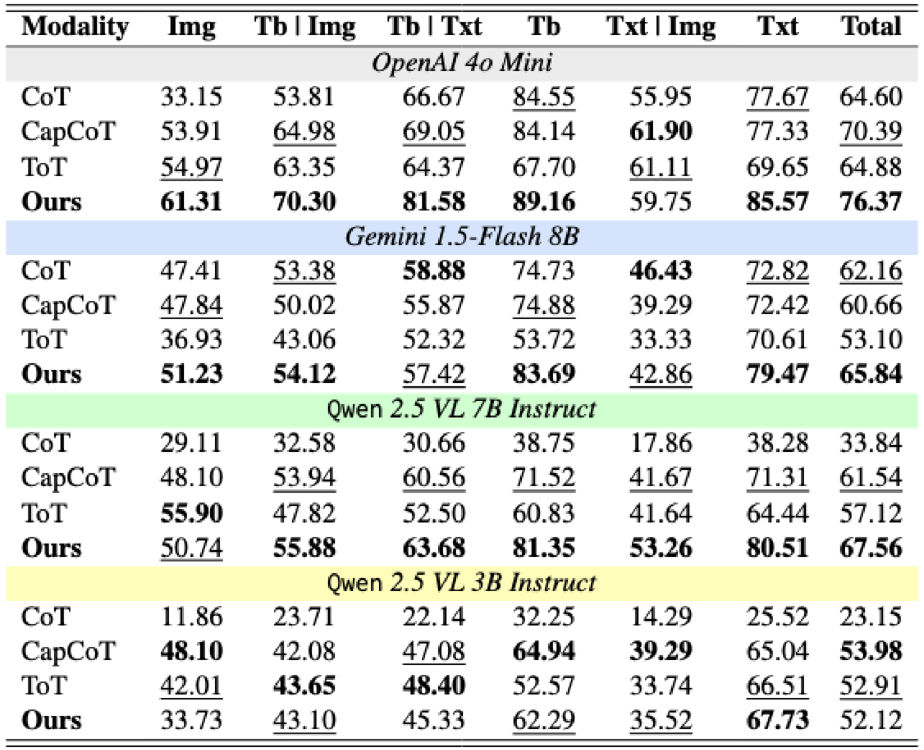
- **Agentic Method** → reduces cognitive load, traceable reasoning
- **MAMMQA** → 3-stage MLLM agent (Extract ➔ Compare ➔ Answer) catching fine-grained ambiguity


Information extraction and rearranging
Cross modal information synchronization
Modality Expert Agent
Cross Modal Synthesis Agent
Aggregator Agent
A

Closing Notes

- **Agentic Method** → reduces cognitive load, traceable reasoning
- **MAMMQA** → 3-stage MLLM agent (Extract ➜ Compare ➜ Answer) catching fine-grained ambiguity
- **Efficient** → Even our 3B Agent performs comparable to enterprise models


Information extraction and rearranging
Cross modal information synchronization
Modality Expert Agent
Cross Modal Synthesis Agent
Aggregator Agent
A

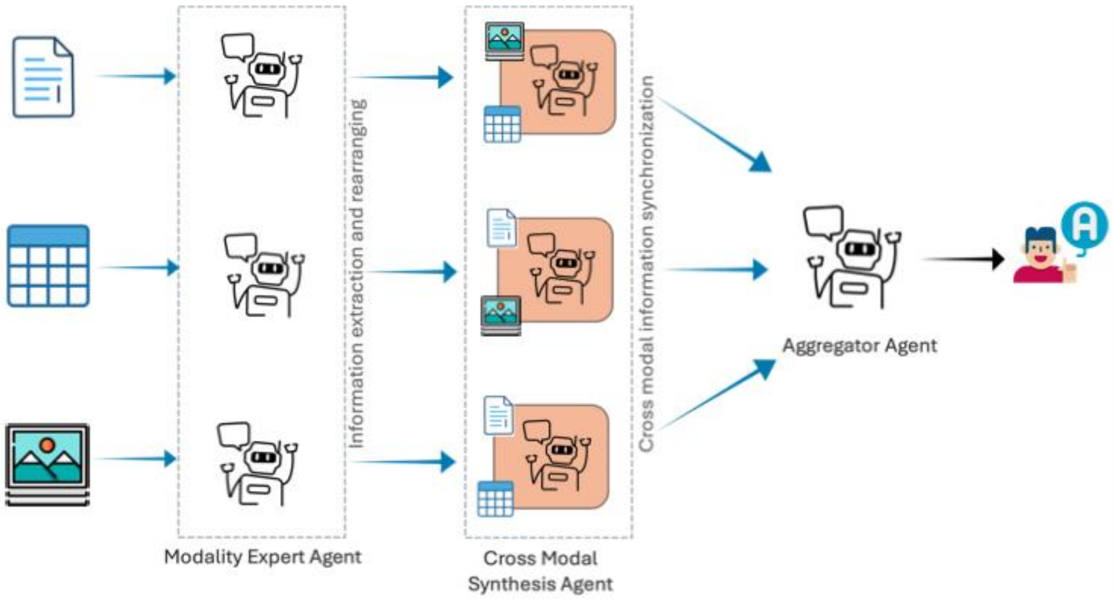


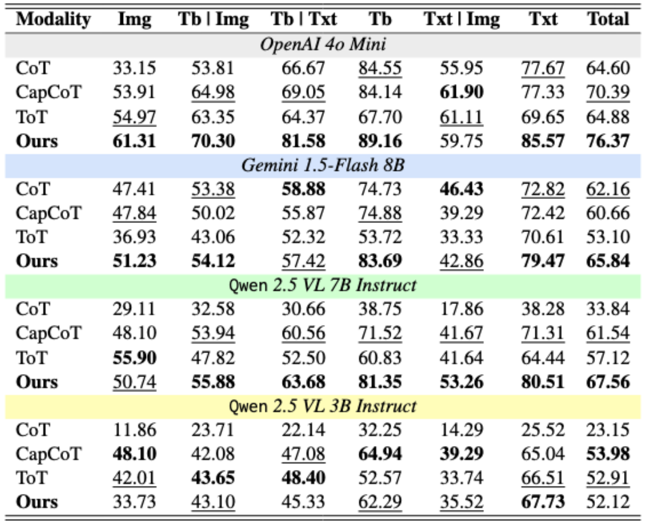
Modality	Img	Tb | Img	Tb | Txt	Tb	Txt | Img	Txt	Total
			OpenAI 4o Mini				
CoT	33.15	53.81	66.67	84.55	55.95	77.67	64.60
CapCoT	53.91	64.98	69.05	84.14	**61.90**	77.33	70.39
ToT	54.97	63.35	64.37	67.70	61.11	69.65	64.88
Ours	**61.31**	**70.30**	**81.58**	**89.16**	59.75	**85.57**	**76.37**
			Gemini 1.5-Flash 8B				
CoT	47.41	53.38	**58.88**	74.73	**46.43**	72.82	62.16
CapCoT	47.84	50.02	55.87	74.88	39.29	72.42	60.66
ToT	36.93	43.06	52.32	53.72	33.33	70.61	53.10
Ours	**51.23**	**54.12**	57.42	**83.69**	42.86	**79.47**	**65.84**
			Qwen 2.5 VL 7B Instruct				
CoT	29.11	32.58	30.66	38.75	17.86	38.28	33.84
CapCoT	48.10	53.94	60.56	71.52	41.67	71.31	61.54
ToT	**55.90**	47.82	52.50	60.83	41.64	64.44	57.12
Ours	50.74	**55.88**	**63.68**	**81.35**	**53.26**	**80.51**	**67.56**
			Qwen 2.5 VL 3B Instruct				
CoT	11.86	23.71	22.14	32.25	14.29	25.52	23.15
CapCoT	**48.10**	42.08	47.08	**64.94**	**39.29**	65.04	**53.98**
ToT	42.01	**43.65**	**48.40**	52.57	33.74	66.51	52.91
Ours	33.73	43.10	45.33	62.29	35.52	**67.73**	52.12

Table 1: Quantitative Analysis on MULTIMODALQA dataset

Closing Notes

- **Agentic Method** → reduces cognitive load, traceable reasoning
- **MAMMQA** → 3-stage MLLM agent (Extract ➜ Compare ➜ Answer) catching fine-grained ambiguity
- **Efficient** → Even our 3B Agent performs comparable to enterprise models


Information extraction and rearranging
Cross modal information synchronization
Modality Expert Agent
Cross Modal Synthesis Agent
Aggregator Agent

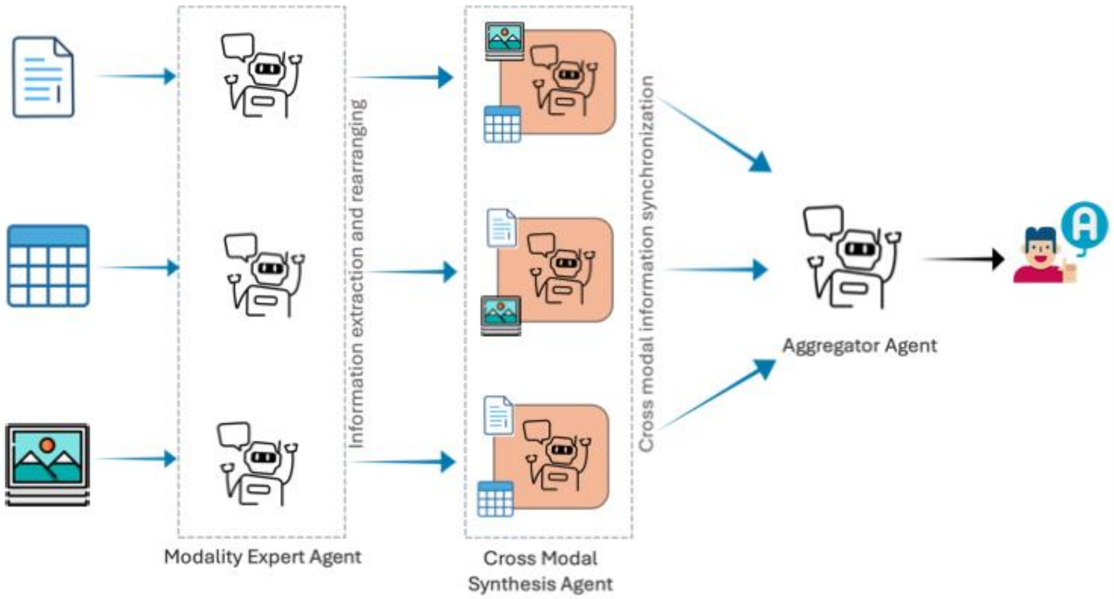


Modality	Img	Tb | Img	Tb | Txt	Tb	Txt | Img	Txt	Total
OpenAI 4o Mini							
CoT	33.15	53.81	66.67	84.55	55.95	77.67	64.60
CapCoT	53.91	64.98	69.05	84.14	**61.90**	77.33	70.39
ToT	54.97	63.35	64.37	67.70	61.11	69.65	64.88
Ours	**61.31**	**70.30**	**81.58**	**89.16**	59.75	**85.57**	**76.37**
Gemini 1.5-Flash 8B							
CoT	47.41	53.38	**58.88**	74.73	**46.43**	72.82	62.16
CapCoT	47.84	50.02	55.87	74.88	39.29	72.42	60.66
ToT	36.93	43.06	52.32	53.72	33.33	70.61	53.10
Ours	**51.23**	**54.12**	57.42	**83.69**	42.86	**79.47**	**65.84**
Qwen *2.5 VL 7B Instruct*							
CoT	29.11	32.58	30.66	38.75	17.86	38.28	33.84
CapCoT	48.10	53.94	60.56	71.52	41.67	71.31	61.54
ToT	**55.90**	47.82	52.50	60.83	41.64	64.44	57.12
Ours	50.74	**55.88**	**63.68**	**81.35**	**53.26**	**80.51**	**67.56**
Qwen *2.5 VL 3B Instruct*							
CoT	11.86	23.71	22.14	32.25	14.29	25.52	23.15
CapCoT	**48.10**	42.08	47.08	**64.94**	**39.29**	65.04	**53.98**
ToT	42.01	**43.65**	**48.40**	52.57	33.74	66.51	52.91
Ours	33.73	43.10	45.33	62.29	35.52	**67.73**	52.12

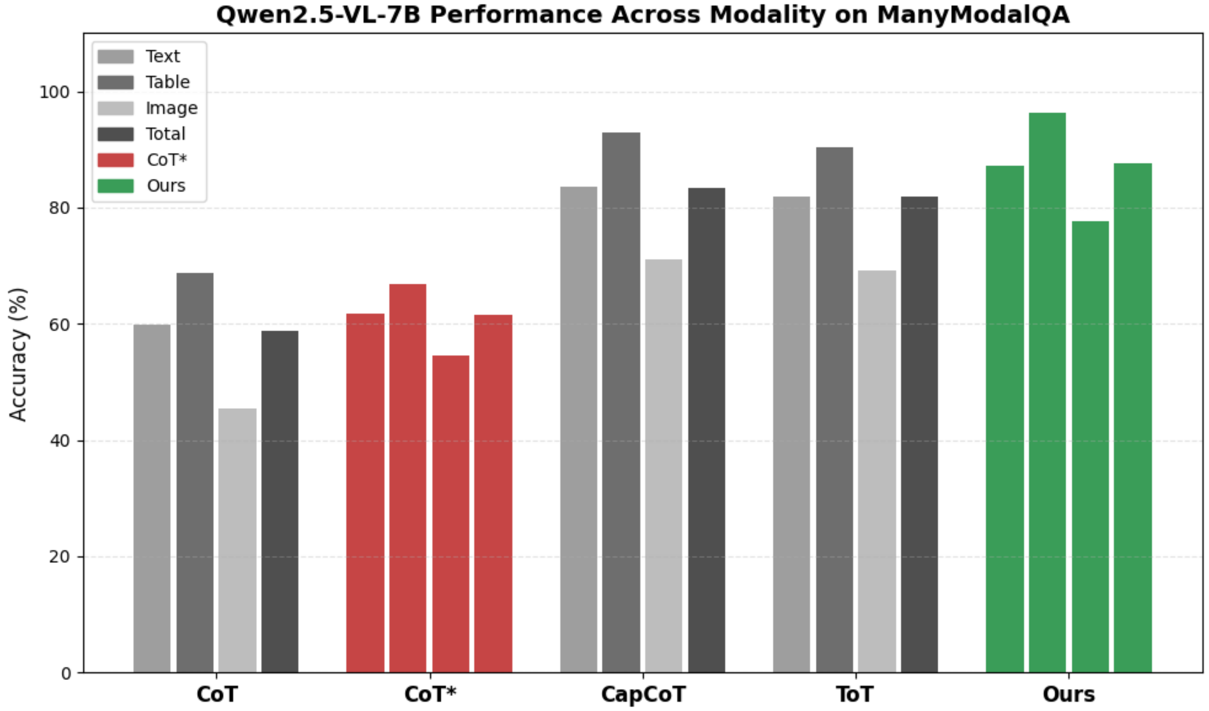
Table 1: Quantitative Analysis on MULTIMODALQA dataset

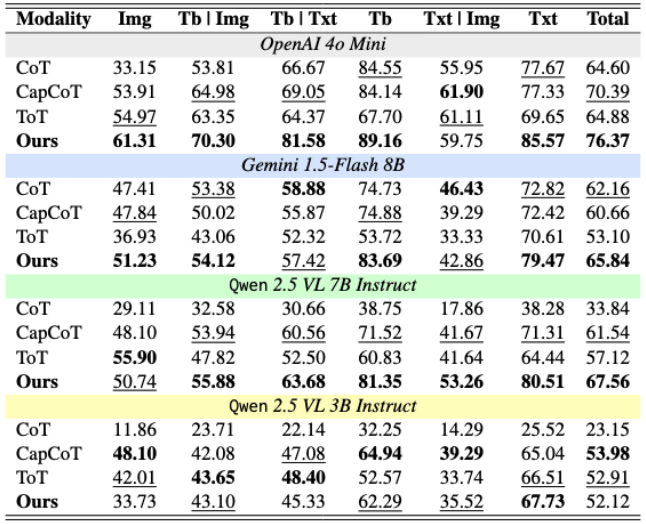
Closing Notes

- **Agentic Method** → reduces cognitive load, traceable reasoning
- **MAMMQA** → 3-stage MLLM agent (Extract ➜ Compare ➜ Answer) catching fine-grained ambiguity
- **Efficient** → Even our 3B Agent performs comparable to enterprise models
- **Interpretable** → Evidence based grounding



Information extraction and rearranging
Cross modal information synchronization
Modality Expert Agent
Cross Modal Synthesis Agent
Aggregator Agent



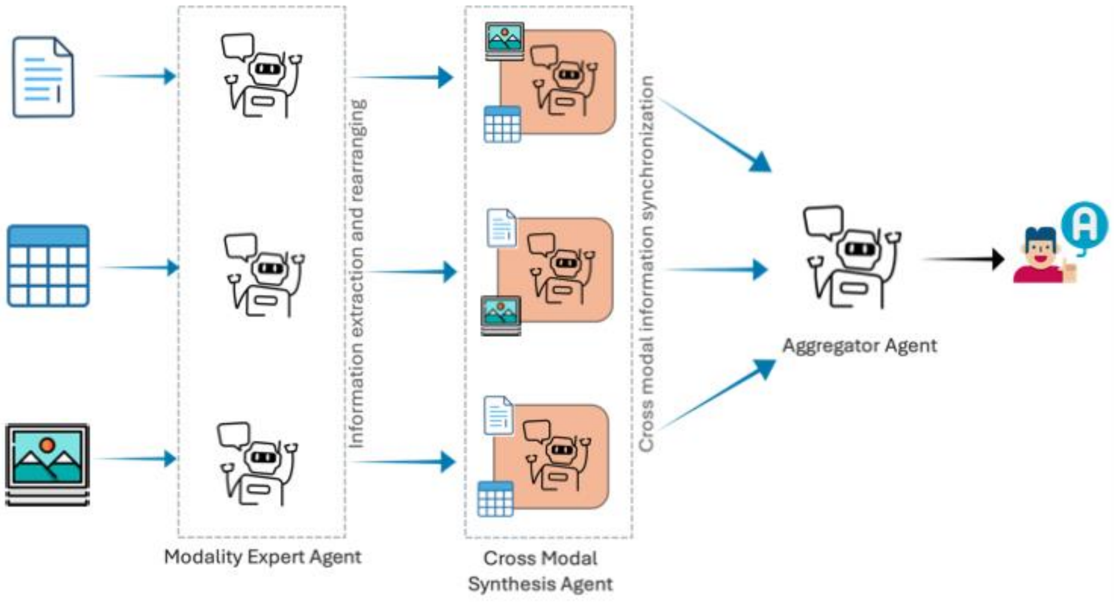


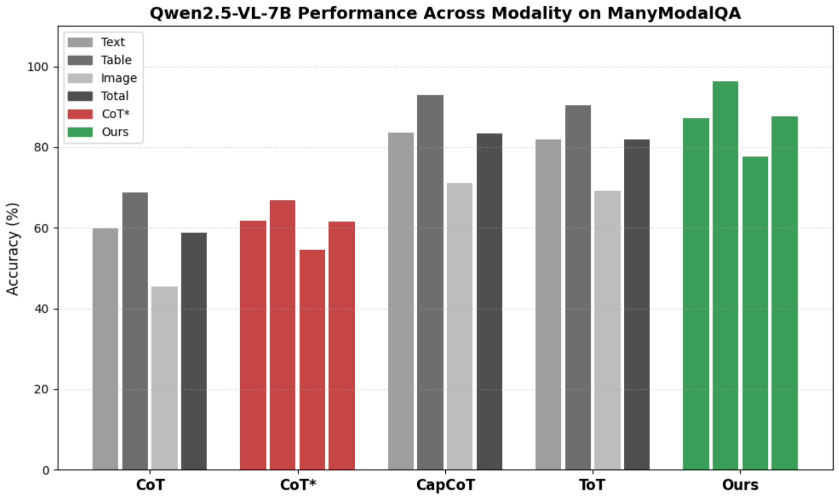
Modality	Img	Tb | Img	Tb | Txt	Tb	Txt | Img	Txt	Total
			OpenAI 4o Mini				
CoT	33.15	53.81	66.67	84.55	55.95	77.67	64.60
CapCoT	53.91	64.98	69.05	84.14	**61.90**	77.33	70.39
ToT	54.97	63.35	64.37	67.70	61.11	69.65	64.88
Ours	**61.31**	**70.30**	**81.58**	**89.16**	59.75	**85.57**	**76.37**
			Gemini 1.5-Flash 8B				
CoT	47.41	53.38	**58.88**	74.73	**46.43**	72.82	62.16
CapCoT	47.84	50.02	55.87	74.88	39.29	72.42	60.66
ToT	36.93	43.06	52.32	53.72	33.33	70.61	53.10
Ours	**51.23**	**54.12**	57.42	**83.69**	42.86	**79.47**	**65.84**
			Qwen *2.5 VL 7B Instruct*				
CoT	29.11	32.58	30.66	38.75	17.86	38.28	33.84
CapCoT	48.10	53.94	60.56	71.52	41.67	71.31	61.54
ToT	**55.90**	47.82	52.50	60.83	41.64	64.44	57.12
Ours	50.74	**55.88**	**63.68**	**81.35**	**53.26**	**80.51**	**67.56**
			Qwen *2.5 VL 3B Instruct*				
CoT	11.86	23.71	22.14	32.25	14.29	25.52	23.15
CapCoT	**48.10**	42.08	47.08	**64.94**	**39.29**	65.04	**53.98**
ToT	42.01	**43.65**	**48.40**	52.57	33.74	66.51	52.91
Ours	33.73	43.10	45.33	62.29	35.52	**67.73**	52.12

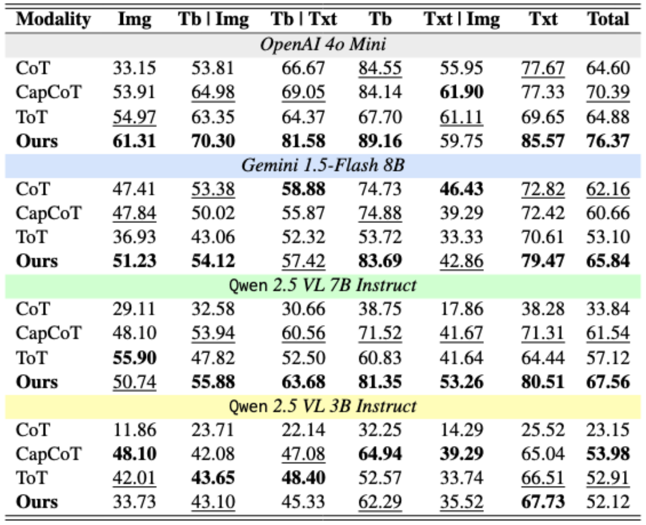
Table 1: Quantitative Analysis on MULTIMODALQA dataset

Closing Notes

- **Agentic Method** → reduces cognitive load, traceable reasoning
- **MAMMQA** → 3-stage MLLM agent (Extract ➔ Compare ➔ Answer) catching fine-grained ambiguity
- **Efficient** → Even our 3B Agent performs comparable to enterprise models
- **Interpretable** → Evidence based grounding


Information extraction and rearranging
Cross modal information synchronization
Aggregator Agent
A
Modality Expert Agent
Cross Modal Synthesis Agent





Modality	Img	Tb ǀ Img	Tb ǀ Txt	Tb	Txt ǀ Img	Txt	Total
OpenAI 4o Mini							
CoT	33.15	53.81	66.67	84.55	55.95	77.67	64.60
CapCoT	53.91	64.98	69.05	84.14	**61.90**	77.33	70.39
ToT	54.97	63.35	64.37	67.70	61.11	69.65	64.88
Ours	**61.31**	**70.30**	**81.58**	**89.16**	59.75	**85.57**	**76.37**
Gemini 1.5-Flash 8B							
CoT	47.41	53.38	**58.88**	74.73	**46.43**	72.82	62.16
CapCoT	47.84	50.02	55.87	74.88	39.29	72.42	60.66
ToT	36.93	43.06	52.32	53.72	33.33	70.61	53.10
Ours	**51.23**	**54.12**	57.42	**83.69**	42.86	**79.47**	**65.84**
Qwen *2.5 VL 7B Instruct*							
CoT	29.11	32.58	30.66	38.75	17.86	38.28	33.84
CapCoT	48.10	53.94	60.56	71.52	41.67	71.31	61.54
ToT	**55.90**	47.82	52.50	60.83	41.64	64.44	57.12
Ours	50.74	**55.88**	**63.68**	**81.35**	**53.26**	**80.51**	**67.56**
Qwen *2.5 VL 3B Instruct*							
CoT	11.86	23.71	22.14	32.25	14.29	25.52	23.15
CapCoT	**48.10**	42.08	47.08	**64.94**	**39.29**	65.04	**53.98**
ToT	42.01	**43.65**	**48.40**	52.57	33.74	66.51	52.91
Ours	33.73	43.10	45.33	62.29	35.52	**67.73**	52.12

Table 1: Quantitative Analysis on MULTIMODALQA dataset

Try MAMMQA Yourself!

https://coral-lab-asu.github.io/MAMMQA/

AFNLP

IJCNLP AACL 2025 December 20-24 Mumbai, India

Thank You

ASU Arizona State University