UniM: A Unified Any-to-Any Interleaved Multimodal Benchmark

Yanlin Li[1], Minghui Guo[1], Kaiwen Zhang[1], Shize Zhang[1], Yiran Zhao[1],
Haodong Li[2], Congyue Zhou[2], Weijie Zheng[3], Yushen Yan[2],
Shengqiong Wu[1], Wei Ji[4], Lei Cui[5], Furu Wei[5], Hao Fei[1,], Mong-Li Lee[1], Wynne Hsu[1]*

[1]National University of Singapore, [2]South China University of Technology, [3]Nanyang Technological University,
[4]Nanjing University, [5]Microsoft Research

Paper: https://arxiv.org/abs/2603.05075

Homepage: https://any2any-mllm.github.io/unim/

Content

1 Motivations

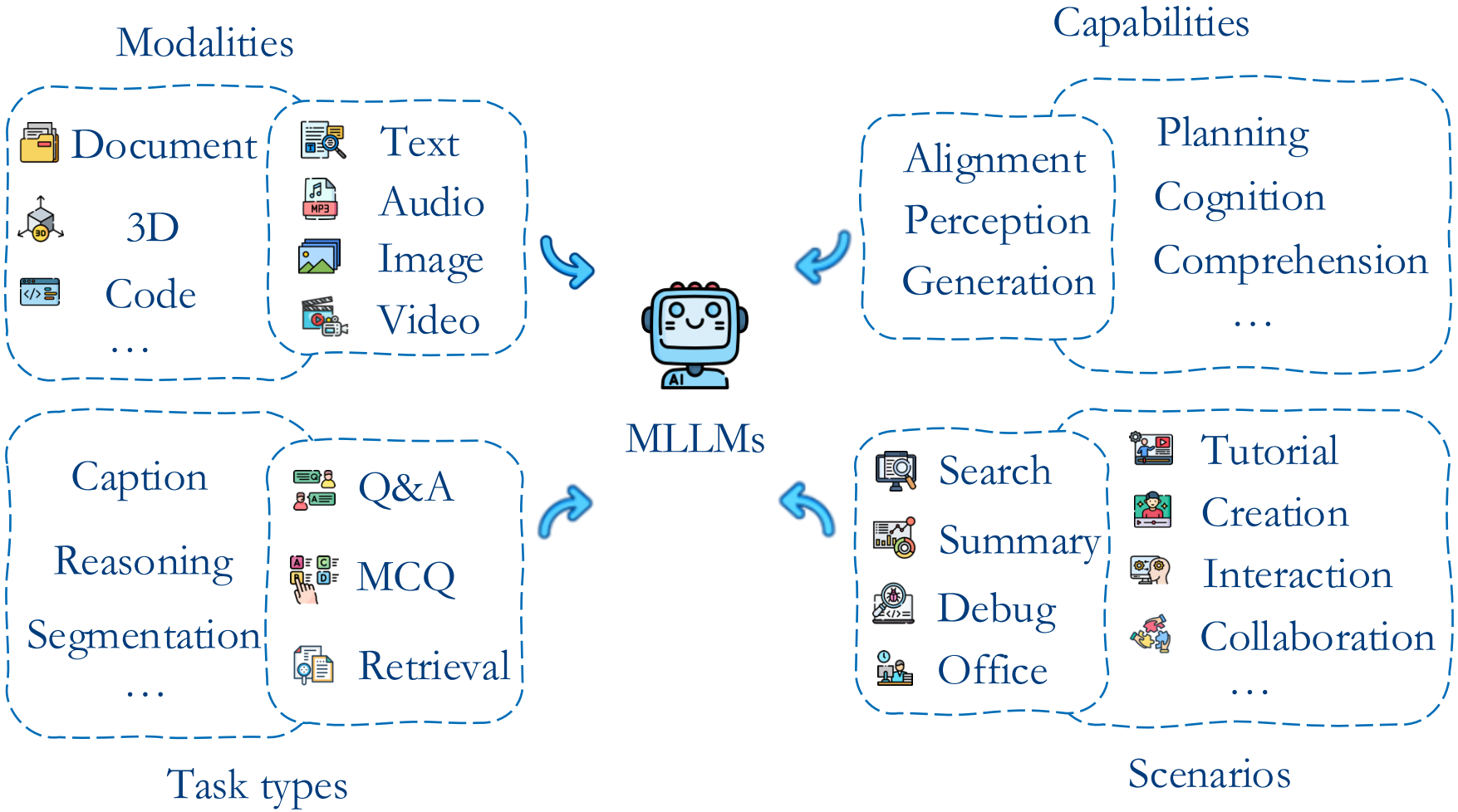
Trends in foundation MLLMs

More modalities

More task types

Richer capabilities

More realistic scenarios



1 Motivations

Any-to-Any Interleaved: Realistic Application Scenario

1 Motivations

Research Gaps

1) Limitation in modality coverage and paradigm. (only text and image fall short of capturing the realistic essential any-to-any interleaved multimodal paradigm.)

2) Lack of unified any-to-any evaluation. (current any-to-any evaluation only focuses on isolated single-modality assessment or purely automatic evaluation)

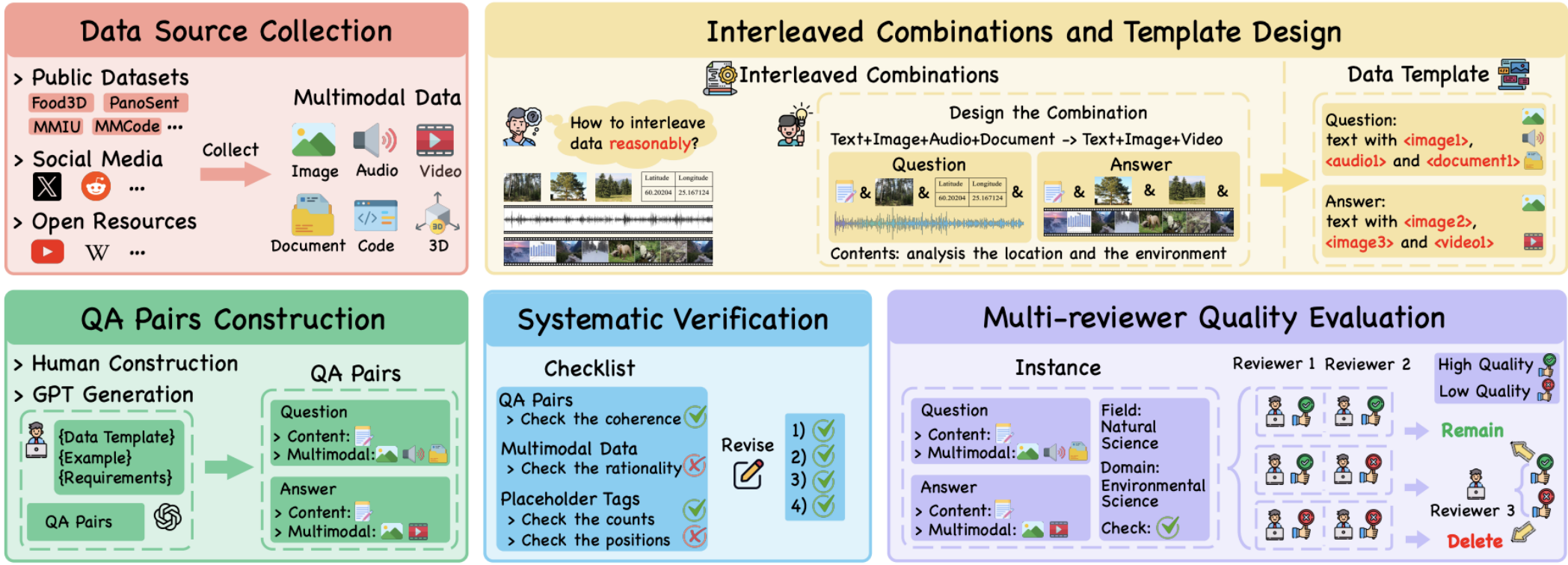
3) Single-capability task design and insufficient domain diversity. (simple tasks and general-domain scenarios, failing to reflect real-world applications.)

Benchmarks	Domains	Num.	Inter. Comb.	Cap. per Instance	Eval. Metric	Difficulty Tax.	Any-to	to-Any	Modalities
ITLVD-BENCH [34]	10	815	2	Single	5	✗	✗	✗	
OpenING [79]	8	5,400	4	Single	7	✗	✗	✗	
ISG-Bench [6]	8	1,150	3	Single	4	✗	✗	✗	
CoMM [8]	3	/	4	Single	3	✗	✗	✗	
MMIE [67]	10	20,103	3	Single	7	✗	✗	✗	
UNIM (Ours)	**30**	**31,026**	**41**	**Multiple**	**13**	✓	✓	✓	

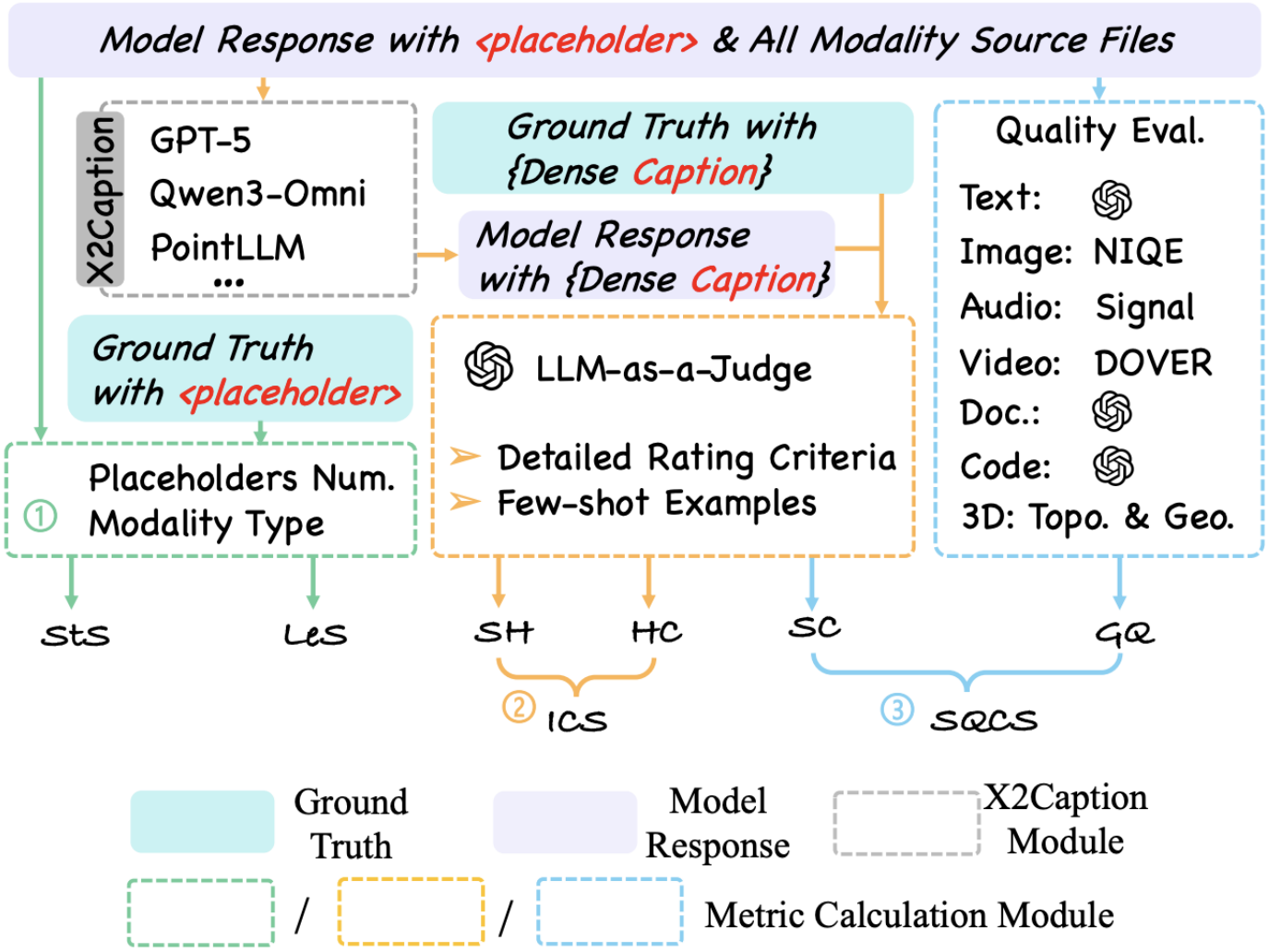
Comparison with existing interleaved multimodal benchmarks. Inter. Comb.: Interleaved combinations of modalities. Cap. per Instance: Capability per instance. Difficulty Tax.: Difficulty taxonomy.

2 UniM: Unified Any-to-Any Interleaved Benchmark

1) Any-to-Any Interleaved Modalities.
2) Universal and Diverse Capabilities.
3) Multi-domain Coverage.
4)Multiple Tasks per Instance.
5) Progressive Difficulty. Large Scale and High Quality.



3 UniM Evaluation Suite



3 dimensions, 13 metrics

(1) Semantic Correctness & Generation Quality: *SC*, *GQ*, $SQCS^{abs}$, $SQCS^{rel}$.

(2) Response Structure Integrity: StS^{abs}, StS^{rel}, LeS^{abs}, LeS^{rel}.

(3) Interleaved Coherence: *HC*, *SH*, ICS^{abs}, ICS^{rel}.

Supporting Rate: τ.

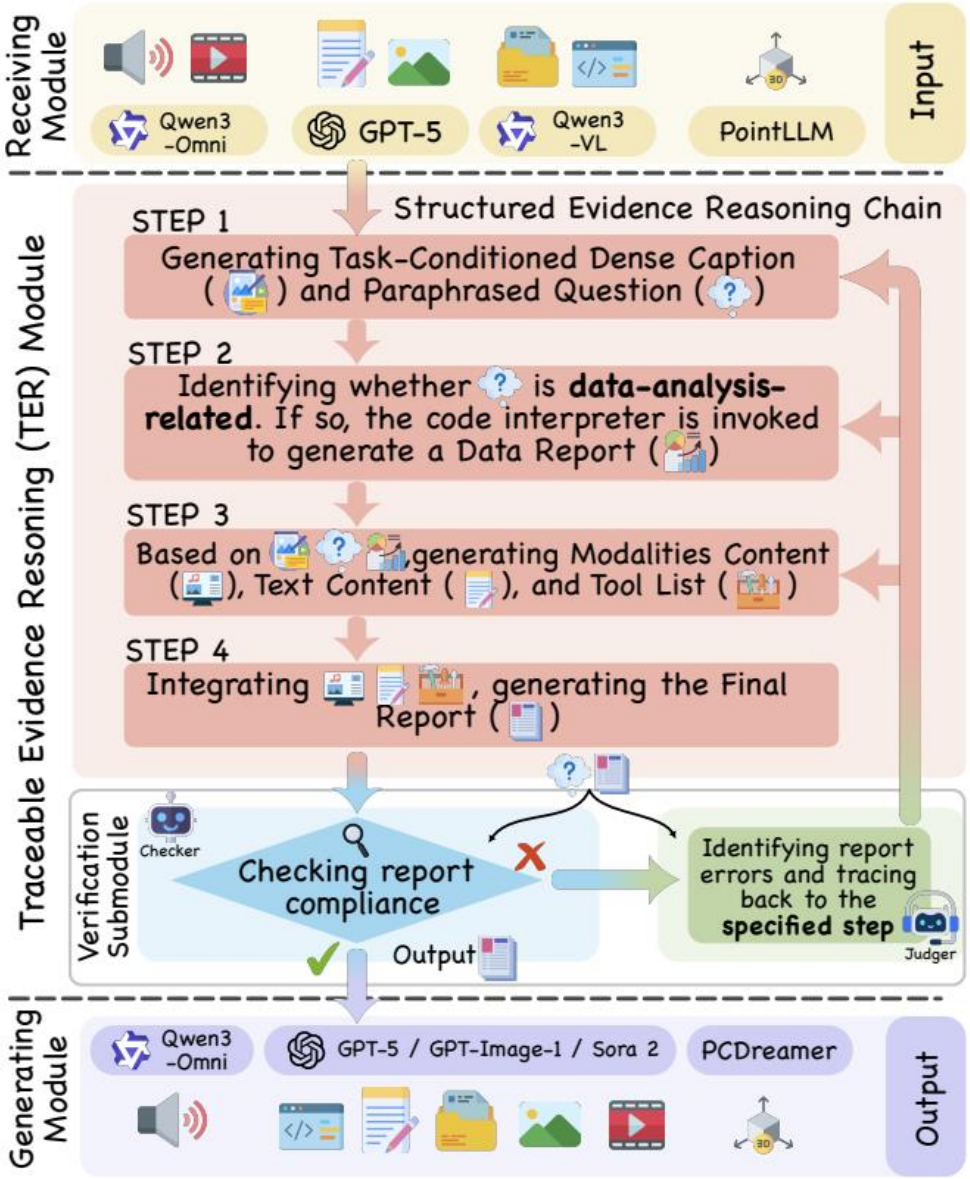
4 UniMA: A Unified Any-to-Any Agentic Model

3 Modules: Receiving Module, Traceable Evidence Reasoning Module, Generating Module.

Receiving Module converts non-text modalities into text descriptions. (all modalities to text)

Traceable Evidence Reasoning Module generates and refines evidence to produce a consistent, verifiable report for guiding generation. (core reasoning)

Generating Module produces interleaved multimodal outputs, completing a loop from understanding to generation based on the report. (modalities generation)



5 Experiments Results

Baseline models show very low StS scores (below 5%). Although slightly better on ICS, their overall performance remains low (generally below 50%), with SQCS under 20%, reflecting significant semantic misalignment. It indicates UniM is challenging for existing models. Our UniMA shows clear superior performance on UniM.

Model	Natural Science				Social Science				General Area			
	StS^{abs}	LeS^{abs}	StS^{rel}	LeS^{rel}	StS^{abs}	LeS^{abs}	StS^{rel}	LeS^{rel}	StS^{abs}	LeS^{abs}	StS^{rel}	LeS^{rel}
AnyGPT [76]	12.9	27.8	12.2	21.4	14.9	16.6	14.5	16.2	12.5	16.4	9.8	13.6
NExT-GPT [64]	2.0	2.2	1.2	1.3	1.3	1.7	1.2	1.5	2.2	2.5	1.4	1.7
MIO [59]	1.3	1.9	0.9	1.3	4.1	5.2	4.0	5.1	3.3	3.8	2.4	2.9
UNIMA	50.8	62.2	50.8	62.2	58.1	72.9	58.1	72.9	71.3	84.3	71.3	84.3

Field	Models	SC	GQ	$SQCS^{abs}$	τ	$SQCS^{rel}$
Natural Science	AnyGPT [76]	13.7	37.9	11.1	90.4	10.7
	NExT-GPT [64]	8.4	23.4	6.2	62.0	2.9
	MIO [59]	19.7	29.1	15.9	59.2	10.0
	UNIMA	59.8	79.7	57.3	100	57.3
Social Science	AnyGPT [76]	18.0	23.8	15.5	94.7	14.7
	NExT-GPT [64]	16.8	31.9	13.3	89.0	10.8
	MIO [59]	25.2	32.8	21.4	80.8	16.1
	UNIMA	76.2	81.0	72.7	100	72.7
General Area	AnyGPT [76]	19.0	30.1	17.9	90.3	17.2
	NExT-GPT [64]	5.4	30.0	4.4	76.0	3.4
	MIO [59]	24.8	37.5	21.2	71.7	15.2
	UNIMA	64.7	83.6	62.2	100	62.2

Field	Models	HC	SH	ICS^{abs}	ICS^{rel}
Natural Science	AnyGPT [76]	39.9	46.3	41.8	38.5
	NExT-GPT [64]	23.5	26.1	24.9	16.3
	MIO [59]	49.4	63.7	52.1	31.8
	UNIMA	68.4	71.9	69.1	69.1
Social Science	AnyGPT [76]	31.3	35.3	32.1	29.2
	NExT-GPT [64]	24.5	27.1	21.4	19.0
	MIO [59]	46.3	55.0	51.6	42.0
	UNIMA	73.1	76.5	73.8	73.8
General Area	AnyGPT [76]	36.5	41.9	43.6	31.3
	NExT-GPT [64]	27.9	31.1	28.1	20.0
	MIO [59]	68.3	77.7	60.0	45.7
	UNIMA	68.7	74.3	69.8	69.8

6 Conclusion

1) We introduce **UniM**, the **first Unified Any-to-Any Interleaved Multimodal** benchmark, including **31K** high-quality instances, spanning **30** diverse domains, covering **7** representative modalities, text, image, audio, video, document, code, and 3D.

2) We further develop the **UniM Evaluation Suite**, which assesses models along **3 dimensions**: Semantic Correctness & Generation Quality, Response Structure Integrity, and Interleaved Coherence, for better evaluate MLLMs under the Any-to-Any interleaved setting.

3) We propose **UniMA**, a **Unified Any-to-Any Interleaved Multimodal Agentic model** capable of supporting arbitrary interleaved combinations of 7 modalities.

Thank You !

Yanlin Li (yanlin.li@u.nus.edu)

Paper: https://arxiv.org/abs/2603.05075

Homepage: https://any2any-mllm.github.io/unim/