

### Previous MLLM trajectory

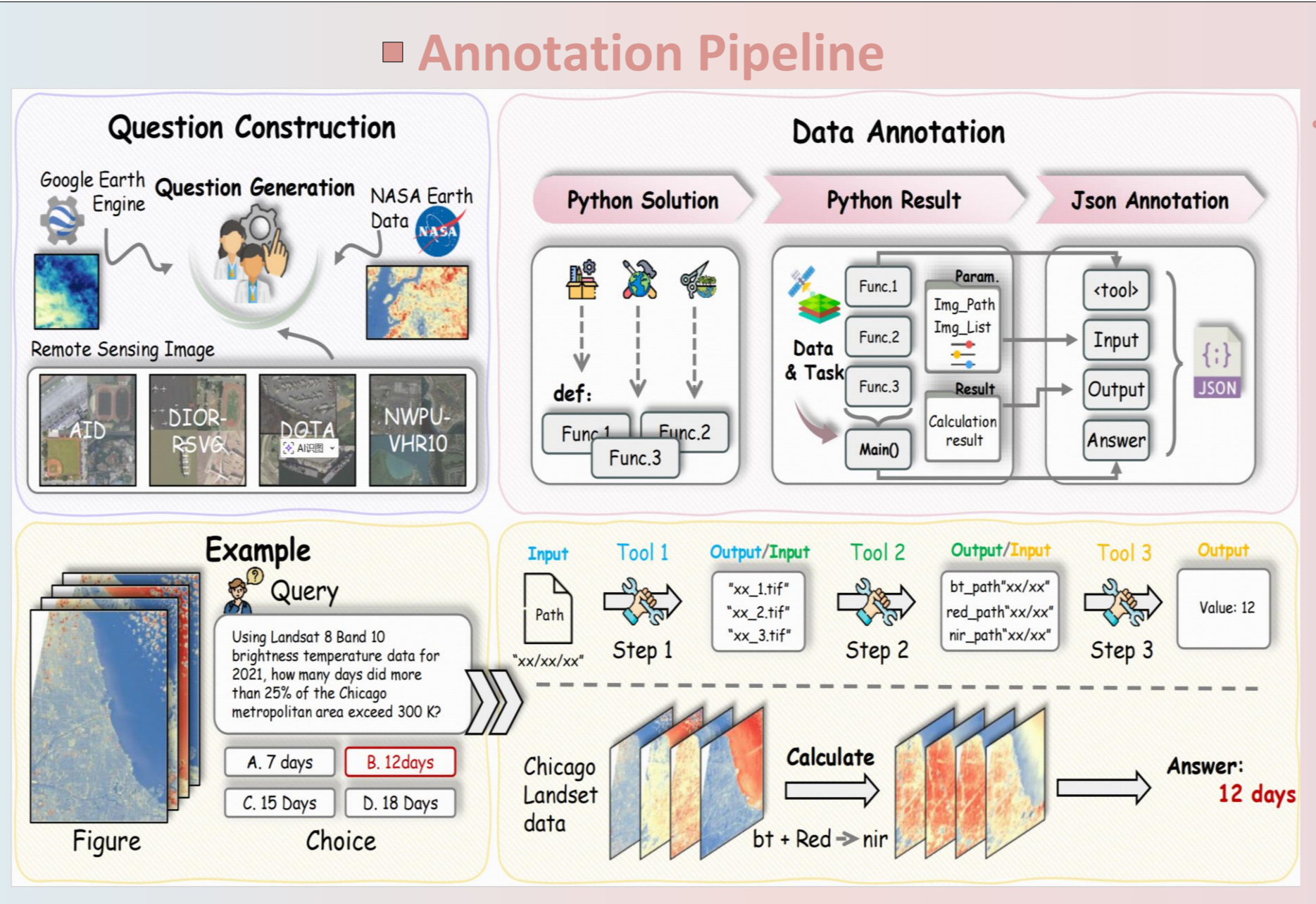
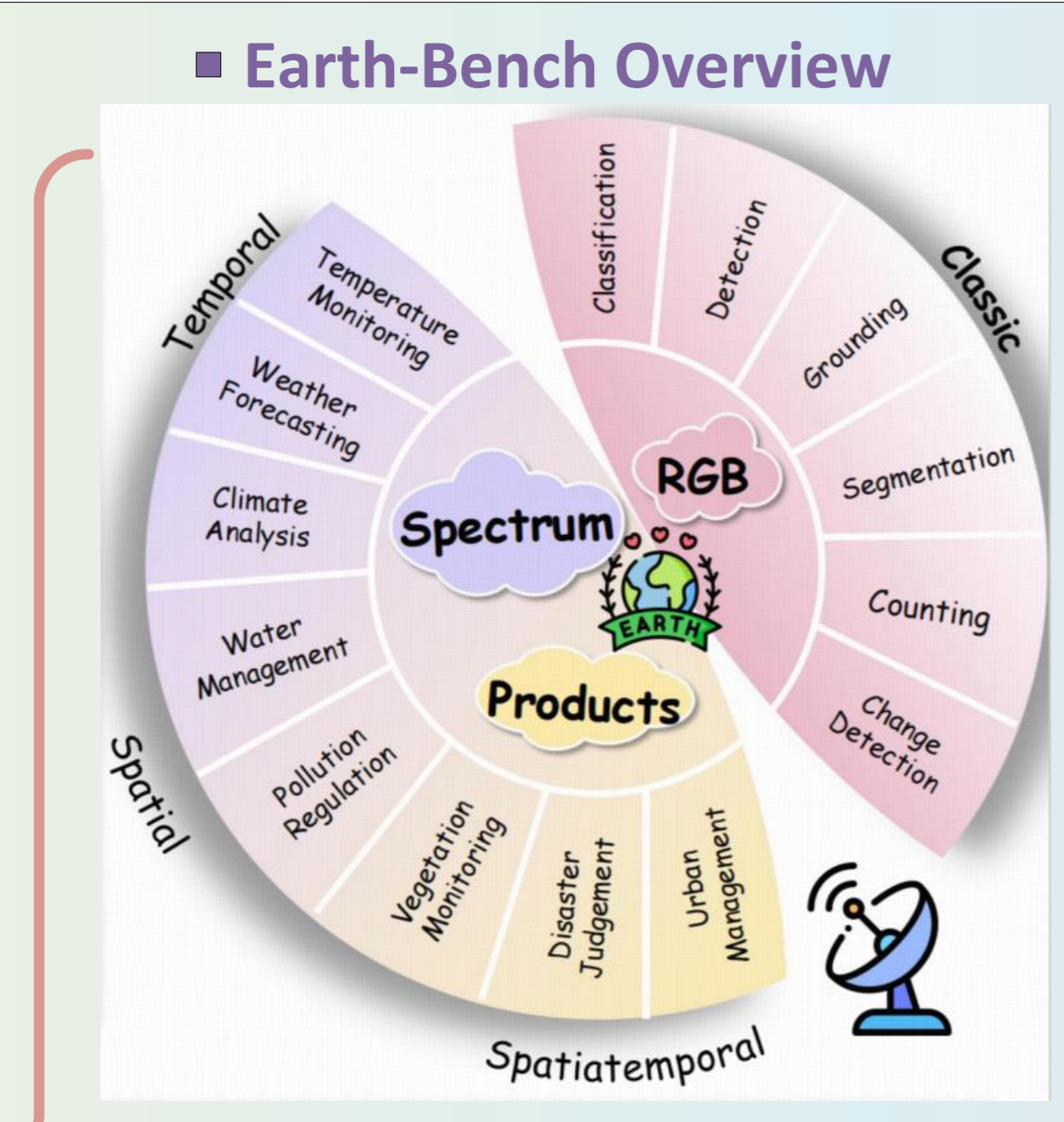
#### MLLM-Based EO Research

Describe/Classify this image<image>

#### Drawbacks

1. Only support processing RGB images
2. Only process one/ a few images at a time
3. Limited to single-step and shallow reasoning
4. Can't integrate external tools or expert models

Model performance relies more on data quality



### A new trajectory: connect EO with Agent!

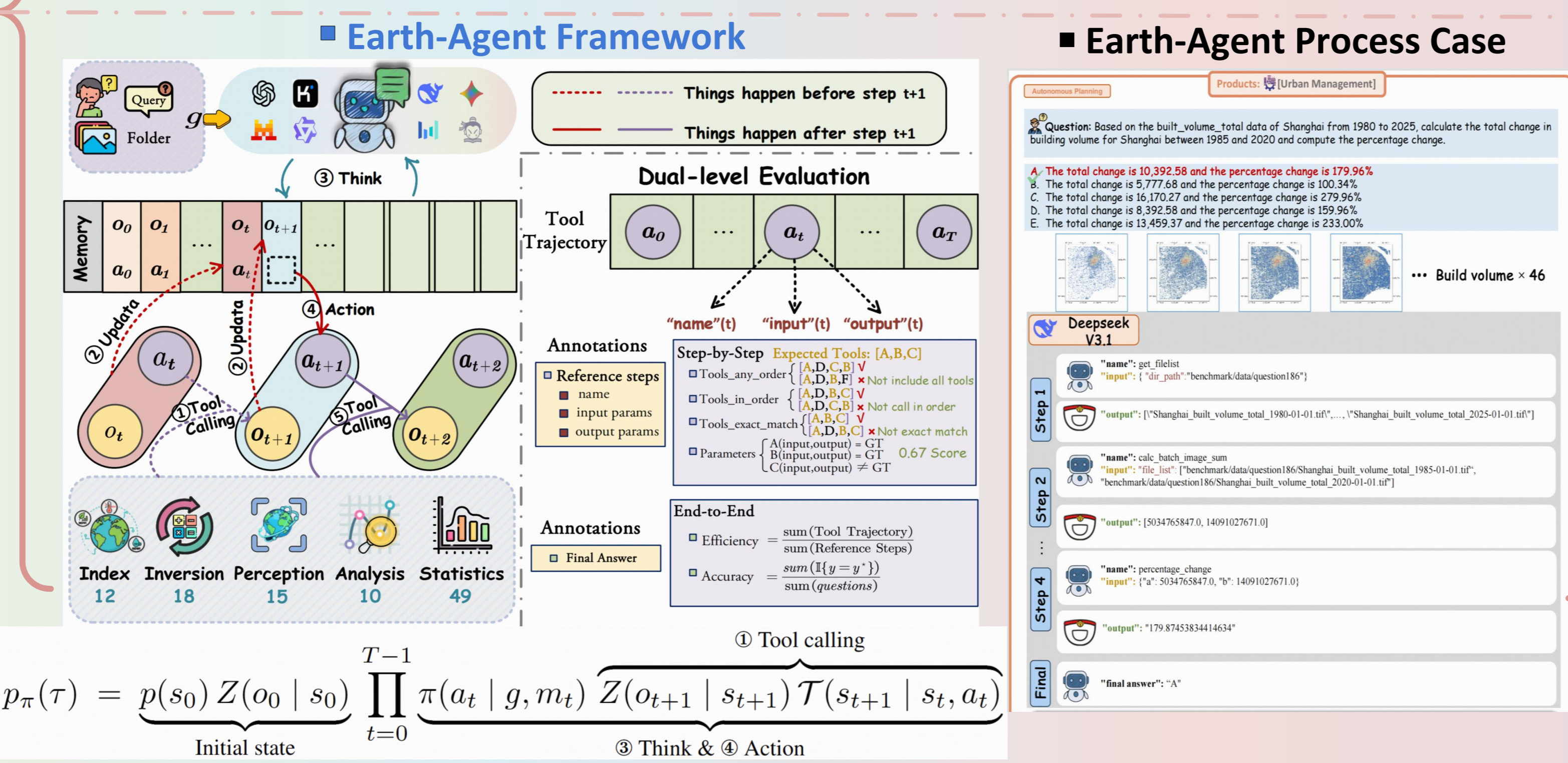
#### Earth-Agent (Ours)

Calculate the trend of building volume

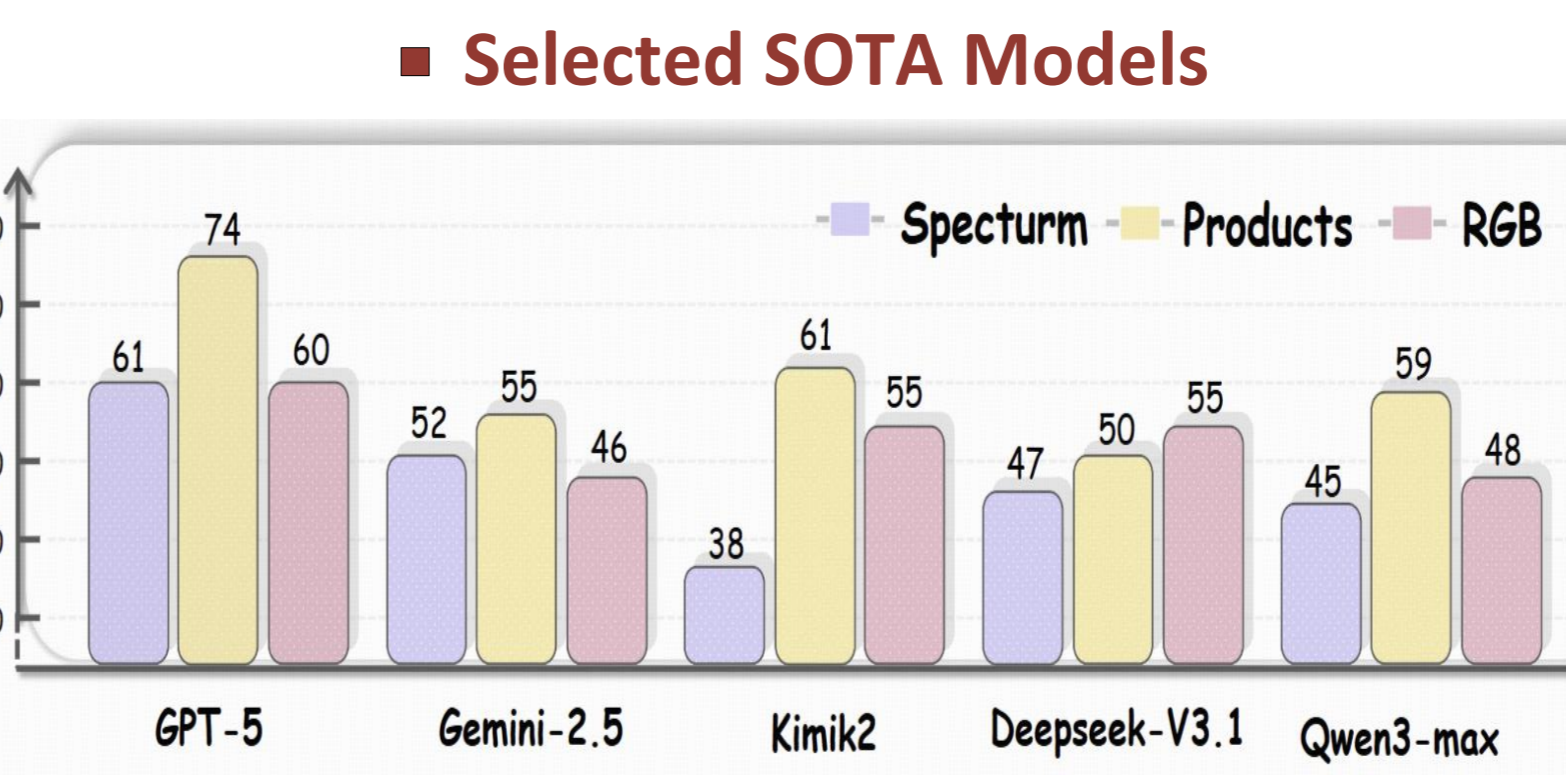
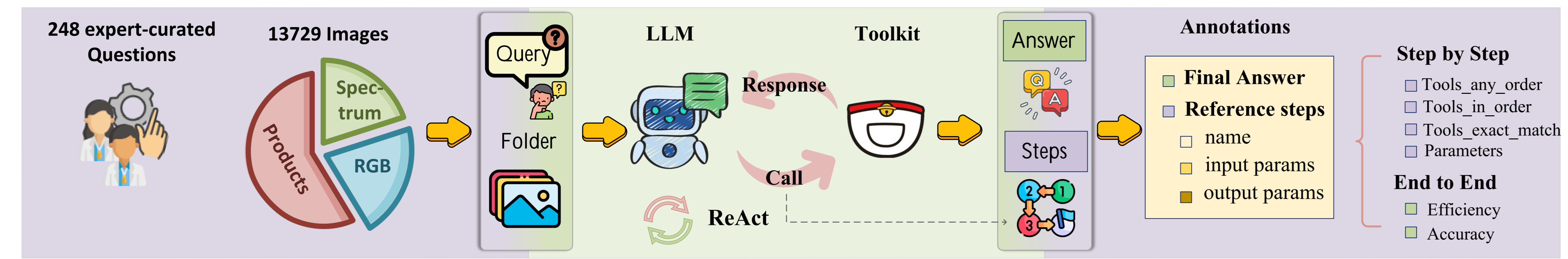
#### Advantages

1. Able to process Multi-Spectral and RGB images
2. Able to process numerous images
3. Complex & multi-step interactive reasoning
4. Interact with external tools/models extensibly
5. Systemic evaluation: trajectory & result

Model performance relies more on tool use



$$p_{\pi}(\tau) = \underbrace{p(s_0) Z(o_0 | s_0)}_{\text{Initial state}} \prod_{t=0}^{T-1} \underbrace{\pi(a_t | g, m_t) Z(o_{t+1} | s_{t+1}) \mathcal{T}(s_{t+1} | s_t, a_t)}_{\text{Think \& Action}}$$



### Earth-Agent with Different LLMs

Model	Accuracy		Efficiency		Tool-Any-Order		Tool-In-Order		Tool-Exact-Match		Parameters	
	AP	IF	AP	IF	AP	IF	AP	IF	AP	IF	AP	IF
GPT-5	65.99	62.35	2.3560	2.9093	68.74	71.41	57.71	61.06	44.97	46.01	26.11	25.91
Gemini-2.5	52.23	53.04	2.9958	2.4595	58.04	61.63	45.31	50.72	31.32	41.04	17.26	23.43
GPT-4o	43.72	44.94	2.1211	2.6007	65.65	67.02	50.76	53.04	46.26	47.41	26.55	27.92
Kimik2	50.61	56.68	1.8542	2.1793	71.03	78.86	57.57	68.83	42.11	51.15	25.90	30.45
DeepSeek-V3.1	51.42	52.23	2.6116	2.6303	78.31	78.66	62.73	64.50	48.54	49.58	30.81	31.36
Qwen3-Max	50.20	47.37	1.8810	1.9511	69.56	70.14	53.28	56.02	37.02	42.74	21.83	26.27
Seed-1.6	52.48	59.51	1.3110	1.3408	55.43	59.44	40.67	46.79	28.39	35.47	18.32	23.13
LLaMA-4	44.94	38.46	0.2886	0.3211	16.51	22.41	2.45	12.05	1.70	9.05	1.30	6.46
Qwen-Plus	42.51	38.46	1.5119	1.5854	52.04	55.96	30.75	39.77	11.69	25.51	9.12	16.95
GLM-4.5v	32.86	35.25	1.7123	2.0129	42.48	46.69	28.57	35.24	14.12	19.95	11.02	15.37
Mistral	29.96	22.67	0.9552	0.8802	27.73	29.64	11.78	20.90	9.13	18.13	7.24	11.66
Qwen3-32B	20.65	24.80	2.7274	1.9010	39.76	42.39	21.56	33.79	9.51	26.10	8.17	17.73
InternVL-3.5	26.72	26.72	0.1206	0.1750	8.83	16.62	3.87	10.59	2.02	9.32	1.46	5.32

AP: without step instruction in query IF: with step instruction in query

### Compare with General Agents

Earth-Agent performs better than other general Agents

Method	Spectrum	Products	RGB	Avg.
GPT-Agent	45.00	31.60	45.26	40.42
MGX	40.00	15.80	0.00	18.60
Manus	15.00	15.80	47.62	26.14
Coze	35.00	10.50	0.00	15.30
Earth-Agent(GPT5)	65.00	36.84	65.71	55.83
Earth-Agent(Deepseek-V3.1)	50.00	42.11	51.43	47.84

### Compare with MLLM-based Methods

Earth-Agent performs better than other MLLM-based Methods

Model	Classification		Detection		Grounding
	AID	WHU-RS19	DOTA	HRSC2016	DIOR-RSVG
MiniGPT-v2 (Chen et al., 2023)	32.96	64.80	14.8	24.8	29.892
LLaVA-1.5 (Liu et al., 2024d)	51.00	74.52	17.5	22.1	12.085
Sphinx (Lin et al., 2023)	58.20	-	15.1	25.7	0.939
Geochat (Kuckreja et al., 2024)	72.03	86.47	16.5	24.0	10.024
VHM (Pang et al., 2025)	91.70	95.80	-	-	-
LHRS-Bot (Muhtar et al., 2024)	91.26	93.17	17.1	24.4	11.826
Earth-Agent (ours)	93.42	96.12	60.88	65.60	60.46

If you like our project, don't forget to give us a star ⭐!