40th Conference on Neural Information Processing Systems

Sydney (Main)
Dec 6 – Dec 12, 2026

NeurIPS 2026 Poster

MMGraph-Agent: Agentic Multimodal RAG via Cache-Inspired Multimodal Knowledge HyperGraphs

Haoran Luo[1†], Ziyue Zhu[2*], Shangyang Wu[2], Linhao Luo[3], Yikai Guo[4], Qika Lin[5], Fangzhi Xu[6], Jiapu Wang[8], Xiaobao Wu[9], Yifan Zhu[2], Luu Anh Tuan[1,7*]

[1] Nanyang Technological University [2] Beijing University of Posts and Telecommunications
[3] Monash University [4] Beijing Institute of Computer Technology and Application
[5] National University of Singapore [6] Xi'an Jiaotong University [7] VinUniversity
[8] Nanjing University of Science and Technology [9] Shanghai Jiao Tong University
† *Project Lead* * *Corresponding Authors*

haoran.luo@ntu.edu.sg, zyzhu@bupt.edu.cn, anhtuan.luu@ntu.edu.sg

CONTENTS

MMGraph-Agent: Agentic Multimodal RAG via Cache-Inspired Multimodal Knowledge HyperGraphs



NEURAL INFORMATION PROCESSING SYSTEMS

1 Background of Study

2 Contributions

3 Methodology

4 Experimental Results

1

PART

Background of Study

NEURAL INFORMATION PROCESSING SYSTEMS

MMGraph-Agent: Multimodal RAG Meets Hypergraph Memory

Driven by complex real-world applications, **RAG** has evolved from textual pipelines to **multimodal settings** spanning text, images, tables and diagrams.

It has two complementary dimensions: **knowledge organization** (GraphRAG, HyperGraphRAG, MMGraphRAG, RAG-Anything) and **retrieval enhancement** (Search-R1, Graph-R1, MMSearch-R1) — advanced separately so far.

Existing methods still face **high construction cost, context-window limits, and no unified training** across offline and online knowledge sources.

MMGraph-Agent is a unified multimodal RAG framework built on cache-inspired multimodal knowledge hypergraphs.

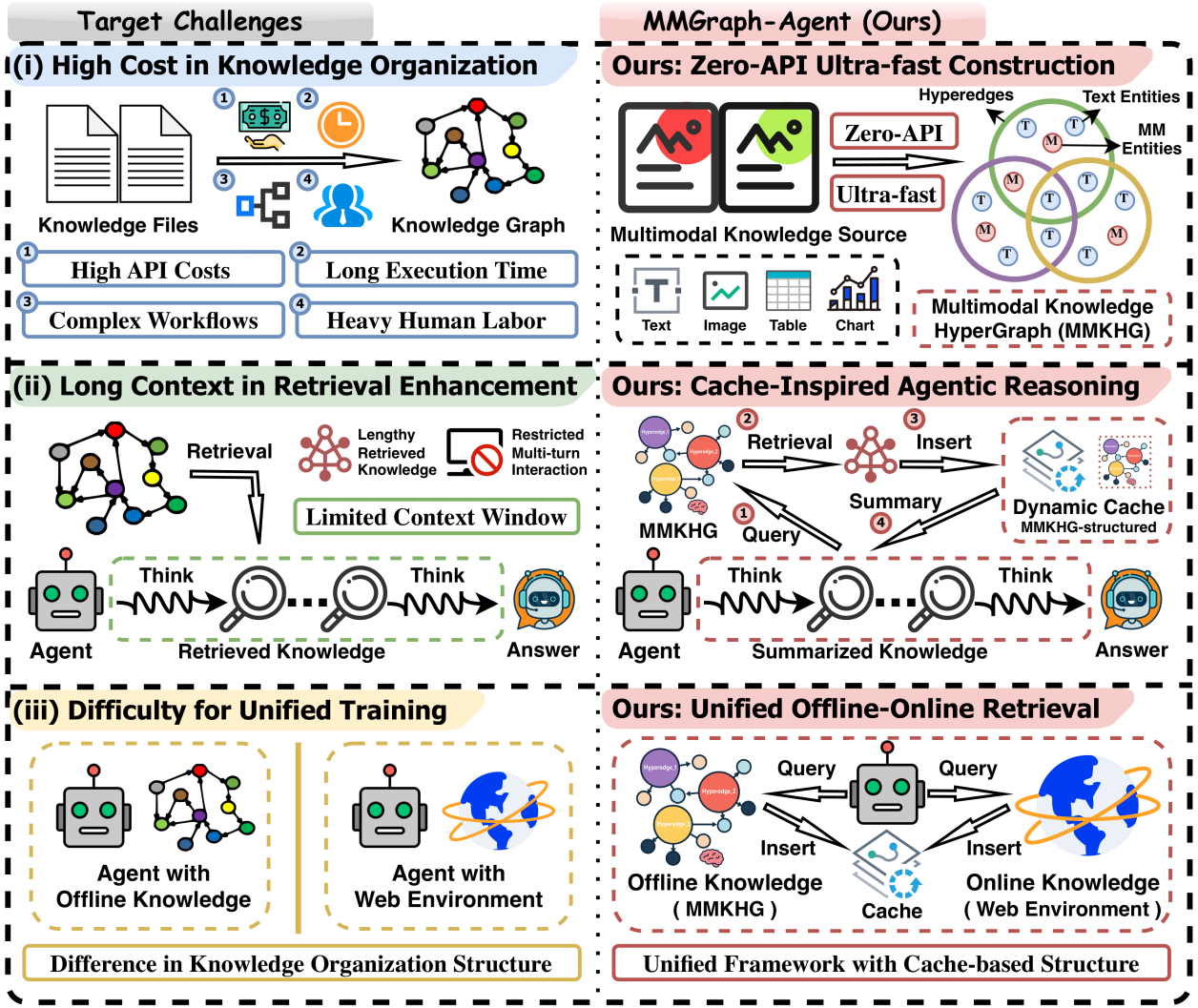


Figure 1: Target challenges (i)–(iii) and how MMGraph-Agent addresses each of them.

Task Definition

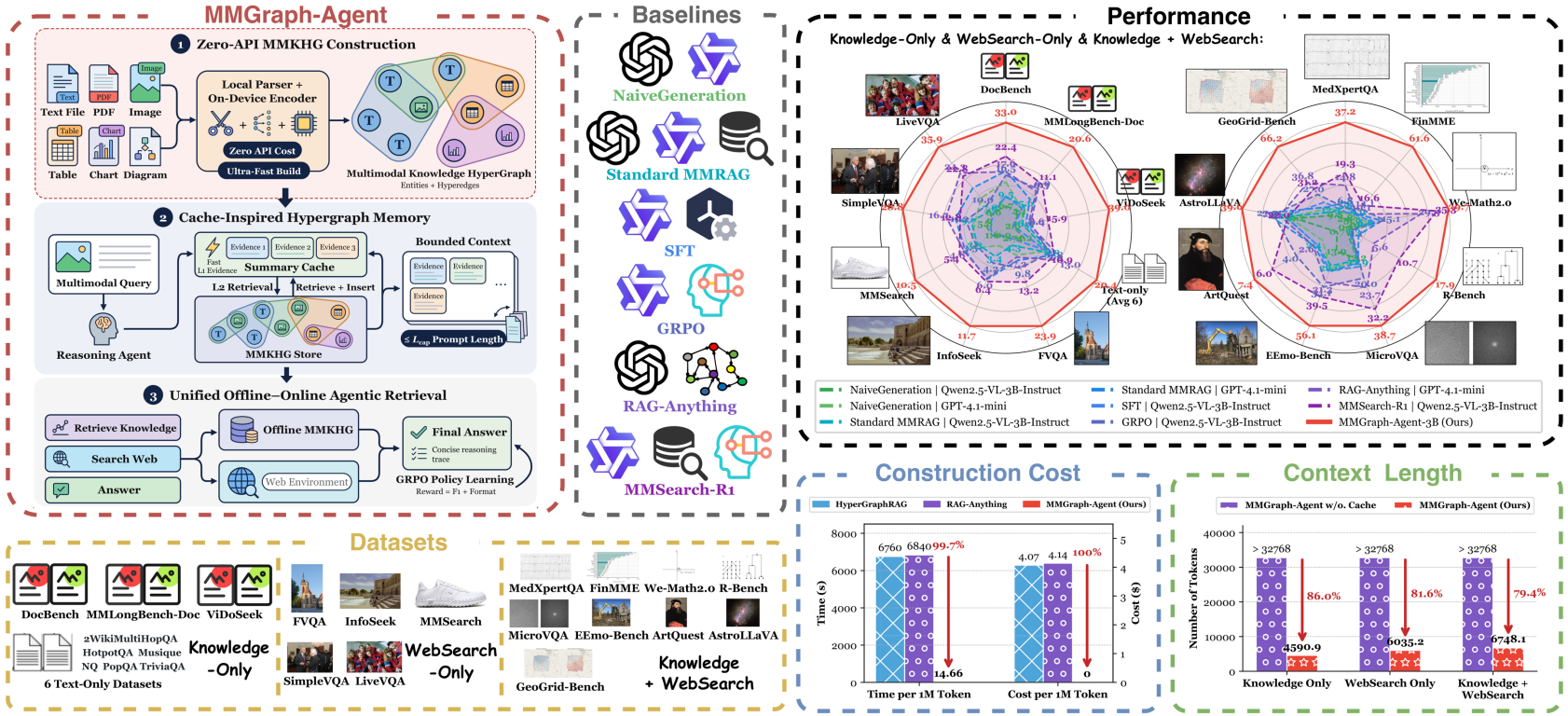


Figure 2: Overview of MMGraph-Agent: a multimodal RAG framework built on cache-inspired knowledge hypergraphs.

Let K be a static corpus of multimodal documents (text, image, table, diagram) and W a dynamic web corpus reachable only through a search interface S.

Given a query (q, v_q) with text and an optional image, learn a **policy π_θ** that emits a trajectory τ scheduling **offline retrieval** on an indexed view $\varphi(K)$, **online retrieval** on W, and the **final answer** y_τ.

φ is the offline organizer; the reward is token-level F1 against the gold answer y*.

$$\tau \sim \pi_\theta\big(\cdot \mid q, v_q; \phi(\mathcal{K}), \mathcal{S}(\mathcal{W})\big), \qquad \mathcal{J}(\pi_\theta) = \mathbb{E}_{(q,v_q,y^*)\sim\mathcal{D}}\big[R_{F_1}(y_\tau, y^*)\big], \tag{1}$$

The three modules address the **cost**, **context** and **unification** challenges in turn.

Three Main Challenges

However, existing multimodal RAG methods still face three key challenges:

i **High cost in knowledge organization**
Building multimodal knowledge graphs relies on manual annotation or expensive API calls, which limits scalability.

ii **Long context in retrieval enhancement**
Multi-round retrieval is constrained by model context windows: as retrieved evidence accumulates, context overflow becomes increasingly likely and degrades reasoning quality.

iii **Lack of a unified framework**
Structural differences make it hard to unify local private knowledge and online web knowledge in one framework (and to train across both).

2

PART

Contributions





Contributions

1 **Zero-API-cost hypergraph construction**

A lightweight multimodal knowledge hypergraph construction with **zero API cost**, substantially reducing construction cost and time while preserving knowledge completeness.

2 **Cache-inspired hypergraph memory**

A two-tier L1/L2 memory that dynamically stores, updates and serves retrieved multimodal evidence during iterative agentic reasoning, alleviating context-window limits and improving multi-round retrieval efficiency.

3 **Unified offline–online agentic RL**

Multimodal RL coordinates retrieval from **offline corpora and the online web** in a single framework; across **23 datasets** and three regimes it beats generation, RAG-based and training-based baselines.

3

PART

Methodology

NEURAL INFORMATION PROCESSING SYSTEMS

Method: MMGraph-Agent

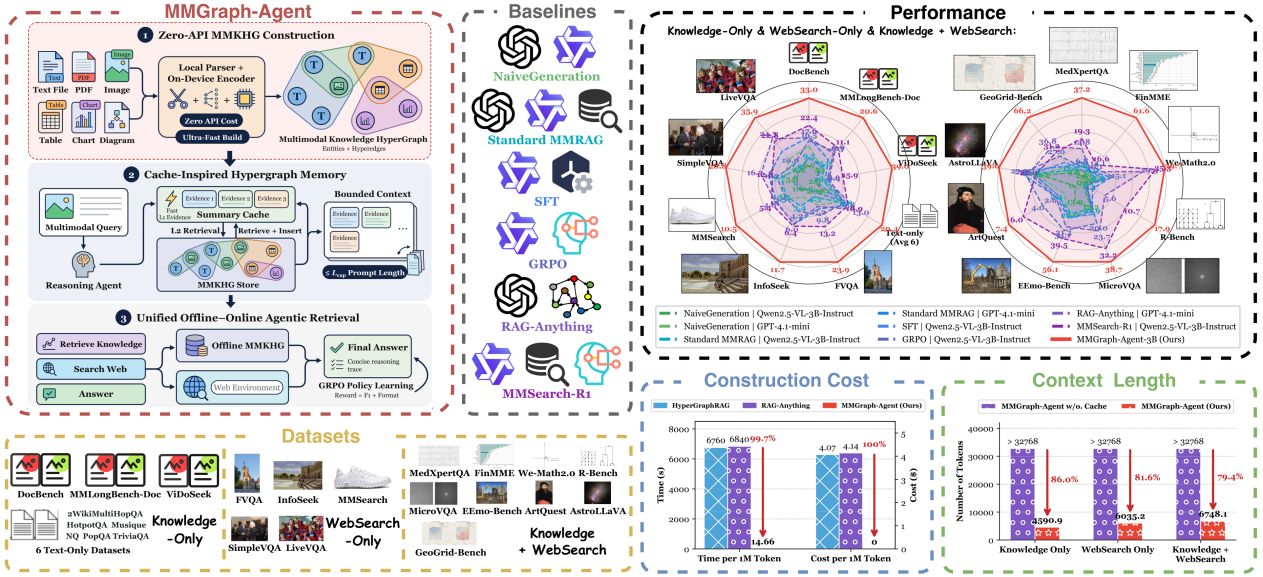


Figure 2: Overview of MMGraph-Agent built on cache-inspired multimodal knowledge hypergraphs.

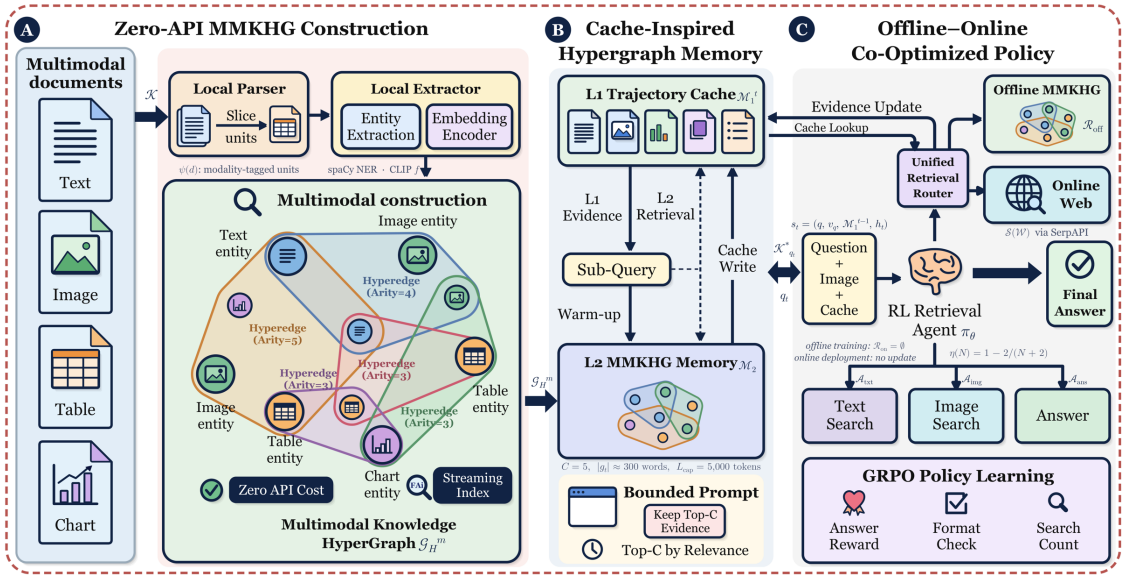


Figure 3: Methodology: zero-API MMKHG, two-tier cache, offline-trained policy.

$$\mathcal{E}_H = \{(\pi_{\text{ext}}(u), u) : u \in \psi_{\text{txt}}(d),\ d \in \mathcal{K},\ |\pi_{\text{ext}}(u)| \geq 1\},\quad \mathbf{E}_{\mathcal{E}_H} = \{f(e_H)\}_{e_H \in \mathcal{E}_H},\quad \mathcal{G}_H^m = (\mathcal{V}^m, \mathcal{E}_H), \tag{3}$$

$$\mathcal{R}_2(q_t) = \underset{x \in \mathcal{M}_2}{\text{Top-C}}\langle f(q_t), f(x)\rangle,\quad \mathcal{M}_1^t = \mathcal{M}_1^{t-1} \cup \mathcal{R}_2(q_t) \cup \mathcal{R}_{\text{on}}(q_t),\quad \mathcal{K}_{q_t}^* = \underset{x \in \mathcal{M}_1^t}{\text{Top-C}}\langle f(q_t), f(x)\rangle, \tag{4}$$

$$\text{prompt}(s_{t+1}) = \text{prompt}(s_t) \oplus g_t,\quad |g_t| \leq L_{\text{dig}},\quad |\text{prompt}(s_t)| \leq \min\big(|\text{prompt}(s_0)| + t \cdot L_{\text{dig}},\ L_{\text{cap}}\big), \tag{5}$$

$$\hat{A}_i = \frac{r(\tau_i) - \mu_G}{\sigma_G + \delta},\quad \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\tau_i}\left[\sum_{i,t} \frac{1}{G|\tau_i|} \min\big(\rho_t \hat{A}_i,\ \text{clip}(\rho_t, 1 \pm \epsilon)\hat{A}_i\big)\right] - \kappa\, \text{KL}(\pi_\theta \,\|\, \pi_{\text{ref}}), \tag{7}$$

Proposition 1. *$\mathcal{G}_H^m$ attains entity coverage comparable to an LLM extractor at zero API cost.*

Proposition 2. *The two-tier memory bounds the prompt uniformly over T with $\mathcal{G}_H^m$ fully reachable.*

Proposition 3. *The source-agnostic interface deploys the offline-trained π_θ online without updates.*

① Zero-API MMKHG Construction

A local parser, **spaCy NER** and a frozen **CLIP** encoder build the hypergraph: entities sharing a sentence form **n-ary hyperedges** at $0 API cost.

② Cache-Inspired Hypergraph Memory

Like CPU caches: an **L2 backend** stores the full hypergraph, a per-trajectory **L1 cache** accumulates evidence and serves the Top-C (Eq. 4); digests keep the **prompt bounded** ($C = 5$, $L_{\text{cap}} = 5{,}000$ tokens).

③ Unified Offline–Online Agentic Retrieval

One action space (text / image query, answer) over a **source-agnostic cache**; GRPO with a **hit-aware reward** (Eq. 7) is trained offline only and deployed online **without updates**.

4

PART

Experimental Results

NEURAL INFORMATION PROCESSING SYSTEMS

Main Results

23 multimodal QA datasets in three regimes: Knowledge-Only, WebSearch-Only and Knowledge+WebSearch; † marks out-of-distribution datasets. Metric: token-level F1.

Backbone **Qwen2.5-VL-3B**; NaiveGen, StdMMRAG and RAG-Any also have **GPT-4.1-mini** variants.

MMGraph-Agent-3B gains **+11.6 F1** on average over the second-best baseline across all 23 datasets; largest gains FinMME +44.93, GeoGrid-Bench +29.47, ViDoSeek +23.06.

+13.0 F1 over MMSearch-R1 on the 8 OOD datasets.

Datasets	NaiveGen		StdMMRAG		SFT	GRPO	RAG-Any	MMS-R1	MMGraph-Agent-3B
	Qwen2.5-3B	GPT-4.1-mini	Qwen2.5-3B	GPT-4.1-mini	Qwen2.5-3B	Qwen2.5-3B	GPT-4.1-mini	Qwen2.5-3B	(Ours)
DocBench	6.74 ±0.3	14.43 ±0.8	9.54 ±0.4	17.79 ±0.3	9.88 ±0.5	16.50 ±0.7	17.57 ±0.9	22.39 ±1.4	**33.05** ±4.4 (+10.66)
MMLongBench-Doc	4.14 ±0.2	4.78 ±0.3	5.63 ±0.2	7.33 ±1.8	4.95 ±0.3	9.91 ±1.1	9.29 ±1.5	11.08 ±1.1	**20.64** ±2.3 (+9.56)
ViDoSeek†	2.64 ±0.1	6.49 ±0.4	6.32 ±0.5	7.58 ±0.8	3.08 ±0.6	9.37 ±0.6	8.58 ±0.4	15.89 ±0.8	**38.95** ±3.5 (+23.06)
2Wiki	4.11 ±0.3	15.89 ±2.0	12.51 ±1.3	15.94 ±1.5	12.42 ±0.7	14.42 ±0.7	15.30 ±1.1	15.36 ±1.1	**17.68** ±5.3 (+1.74)
HotpotQA	7.29 ±1.3	12.79 ±1.2	11.40 ±1.3	13.85 ±1.2	15.16 ±1.0	15.51 ±0.6	13.32 ±1.0	14.72 ±1.3	**20.70** ±4.2 (+5.19)
Musique	0.45 ±0.2	0.59 ±0.3	0.65 ±0.3	0.66 ±1.1	0.27 ±0.9	0.98 ±1.0	1.22 ±0.2	0.78 ±0.7	**5.67** ±1.3 (+4.45)
NQ	3.10 ±0.2	6.82 ±1.9	9.36 ±1.2	8.83 ±1.0	7.67 ±1.3	7.79 ±0.7	10.99 ±1.1	10.90 ±0.8	**24.42** ±2.9 (+13.43)
PopQA†	7.09 ±1.2	10.90 ±1.3	14.34 ±1.5	7.29 ±1.6	7.23 ±1.2	9.21 ±1.7	14.74 ±1.2	11.58 ±1.1	**16.80** ±2.7 (+2.06)
TriviaQA†	3.11 ±0.4	8.41 ±0.3	10.82 ±1.1	19.08 ±1.8	10.05 ±1.7	11.15 ±1.3	22.30 ±0.9	12.13 ±1.3	**37.37** ±5.5 (+15.07)
FVQA†	2.66 ±0.4	6.73 ±0.4	4.34 ±0.1	7.95 ±0.5	2.53 ±0.6	9.83 ±0.8	7.17 ±0.3	13.23 ±0.8	**23.87** ±3.9 (+10.64)
InfoSeek†	0.94 ±0.1	2.16 ±0.3	3.42 ±0.2	6.14 ±0.2	1.00 ±0.2	4.20 ±0.2	6.01 ±0.2	6.39 ±0.5	**11.66** ±3.5 (+5.27)
MMSearch†	1.61 ±0.1	4.16 ±0.4	3.31 ±0.3	4.69 ±0.4	1.60 ±0.3	3.46 ±0.2	4.82 ±0.1	5.37 ±0.2	**10.51** ±2.9 (+5.14)
SimpleVQA†	5.83 ±0.1	8.58 ±0.3	9.71 ±0.6	14.07 ±0.7	4.34 ±0.5	11.22 ±0.4	16.70 ±0.4	12.77 ±1.2	**28.82** ±4.0 (+12.12)
LiveVQA†	6.59 ±0.9	13.21 ±0.3	7.63 ±0.8	17.64 ±0.2	6.48 ±0.6	10.01 ±0.8	24.32 ±0.7	22.84 ±0.6	**35.89** ±2.7 (+11.57)
MedXpertQA	2.84 ±0.2	4.93 ±0.1	7.32 ±0.2	8.95 ±0.1	5.39 ±0.4	6.81 ±0.3	14.83 ±0.2	19.27 ±0.8	**37.17** ±4.5 (+17.90)
FinMME	5.27 ±0.1	8.14 ±0.5	9.91 ±0.9	10.18 ±0.4	9.53 ±0.8	14.12 ±0.9	10.08 ±0.5	16.63 ±0.9	**61.56** ±4.1 (+44.93)
We-Math2.0	3.95 ±1.9	9.19 ±2.1	6.79 ±0.6	14.17 ±1.8	10.92 ±1.8	15.11 ±1.5	29.32 ±1.8	35.29 ±3.2	**39.74** ±4.6 (+4.45)
R-Bench	0.07 ±0.0	0.61 ±0.1	1.26 ±0.4	2.62 ±0.1	0.76 ±0.1	1.14 ±0.1	5.56 ±0.1	10.66 ±0.1	**17.85** ±2.6 (+7.19)
MicroVQA	12.18 ±0.9	13.30 ±0.2	12.90 ±0.8	13.87 ±0.7	15.56 ±0.5	19.98 ±1.0	23.72 ±0.7	32.20 ±0.6	**38.72** ±5.1 (+6.52)
EEmo-Bench	13.87 ±0.9	23.64 ±0.2	21.97 ±0.2	25.48 ±0.1	24.46 ±0.2	34.08 ±1.0	31.31 ±0.4	39.50 ±1.8	**56.11** ±5.1 (+16.61)
ArtQuest	0.67 ±0.1	1.86 ±0.1	2.38 ±0.3	2.56 ±0.2	2.36 ±0.1	3.99 ±0.3	2.61 ±0.2	5.95 ±0.2	**7.38** ±2.3 (+1.43)
AstroLLaVA	19.25 ±0.6	21.78 ±0.7	22.81 ±0.7	30.46 ±1.0	23.26 ±0.8	27.52 ±0.8	25.08 ±0.6	21.97 ±1.7	**39.57** ±1.5 (+9.11)
GeoGrid-Bench	9.85 ±0.3	15.60 ±0.6	12.49 ±0.4	16.66 ±0.6	20.62 ±0.7	27.00 ±0.2	36.76 ±0.5	31.17 ±0.6	**66.23** ±4.7 (+29.47)

Ablation & Analysis

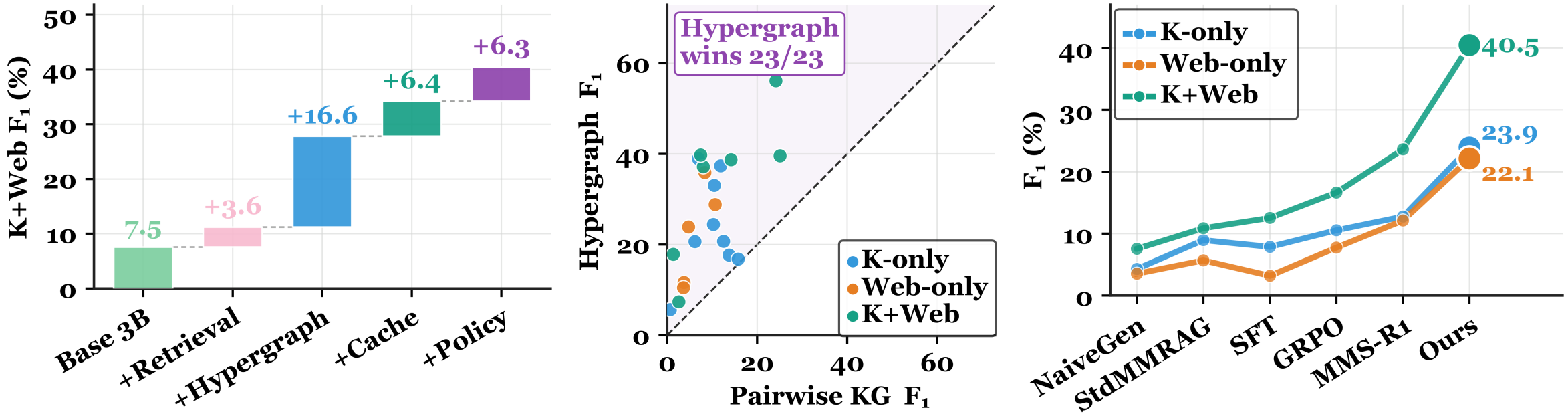
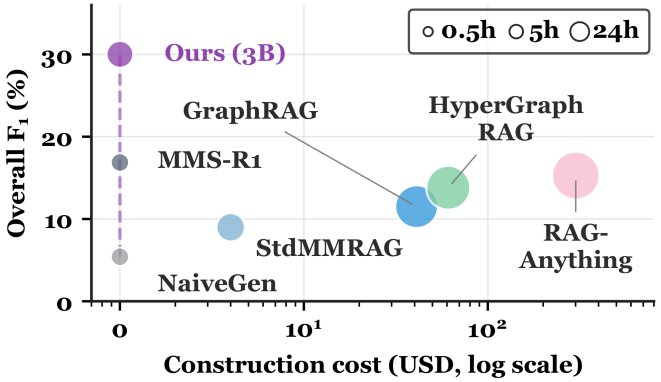
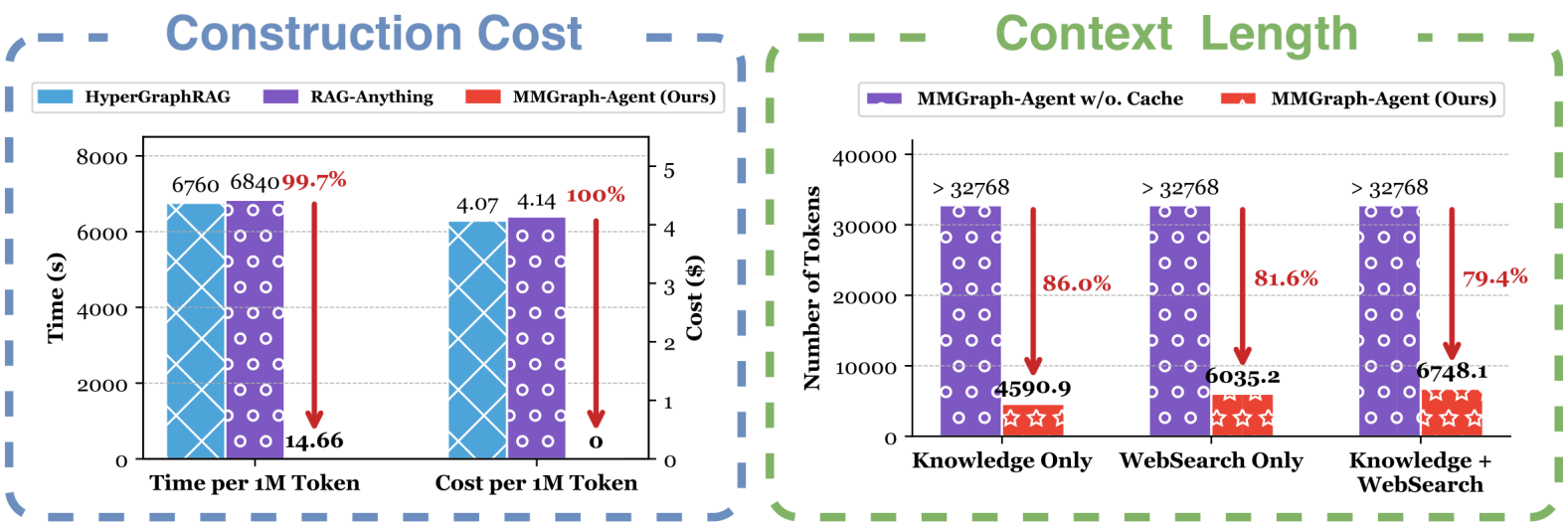


Figure 4: (a) cumulative component waterfall on K+Web; (b) multimodal hypergraph vs. a pairwise KG, one point per dataset; (c) per-regime F1 for the main baseline families and Ours.



- **Components:** K+Web F1 rises from 7.5 (base) to 40.5 via +3.6 retrieval, **+16.6 hypergraph**, +6.4 cache memory and +6.3 policy.
- **Representation:** the multimodal hypergraph beats a pairwise KG on **23/23** datasets.
- **Regimes:** 23.9 / 22.1 / 40.5 F1 (K-only / Web-only / K+Web), +11.2 / +10.0 / +16.9 over MMSearch-R1.
- **Cost–performance (Fig. 5):** overall F1 30.0 at zero API cost.

Efficiency: Construction Cost & Context Length



−99.7% build time per 1M tokens
14.66 s vs. 6760 s (HyperGraphRAG) / 6840 s (RAG-Anything)

−100% API cost per 1M tokens
$0 vs. $4.07 / $4.14

−86.0 / −81.6 / −79.4% context vs. w/o cache
K-only / Web-only / K+Web (w/o cache > 32,768 tokens)

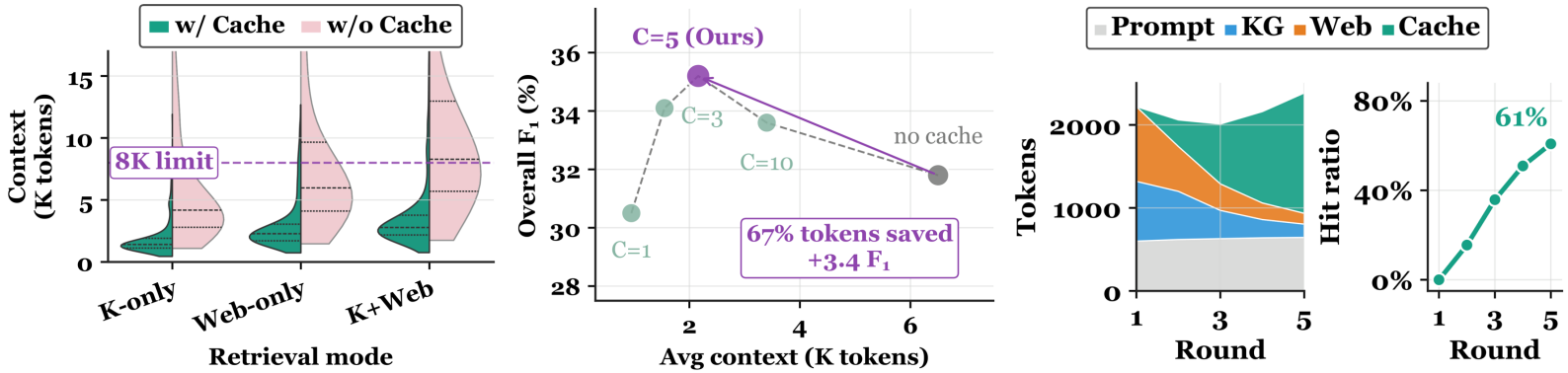


Figure 6: (a) context with/without L1 cache; (b) cache-capacity Pareto, C ∈ {1, 3, 5, 10, ∞}; (c) per-round tokens and cache-hit ratio.

- With the L1 cache, per-query context stays mostly around **1K–3K tokens**; without it, many K+Web trajectories exceed the 8K window.
- Default **C = 5** is on the Pareto frontier: **67% tokens saved, +3.4 F1**.
- Cache-hit share grows from 0% (round 1) to **~61%** by round 5.

40th Conference on Neural Information Processing Systems

Sydney (Main)
Dec 6 – Dec 12, 2026

NeurIPS 2026 Poster

Thanks for Listening

Code & Data https://github.com/LHRLAB/MMGraph-Agent

Contact haoran.luo@ntu.edu.sg