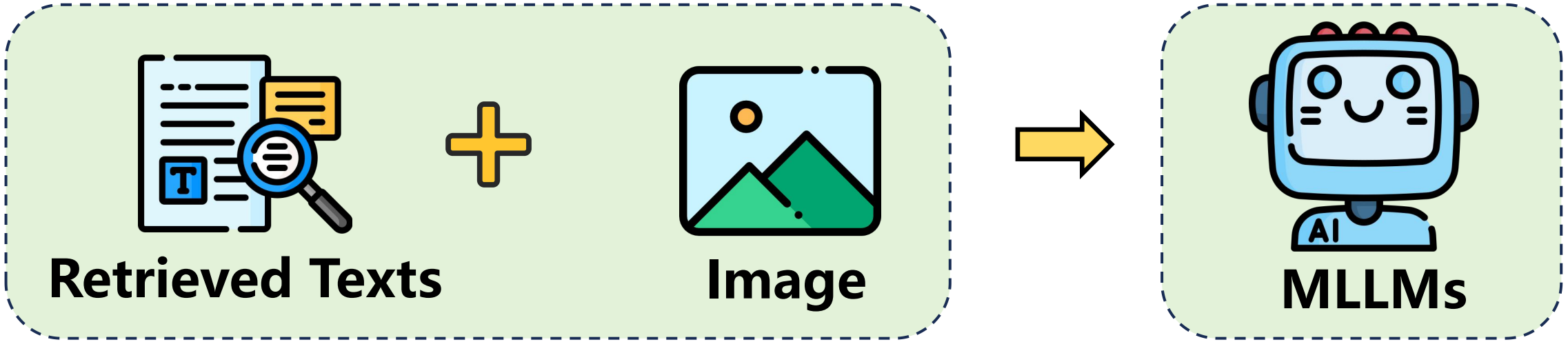
REAL: Resolving Knowledge Conflicts in Knowledge-Intensive Visual Question Answering via Reasoning-Pivot Alignment

Kai Ye, Xianwei Mao, Sheng Zhou, Zirui Shao, Ye Mo, Liangliang Liu, Haikuan Huang, Bin Li, Jiajun Bu

Speaker: Kai Ye

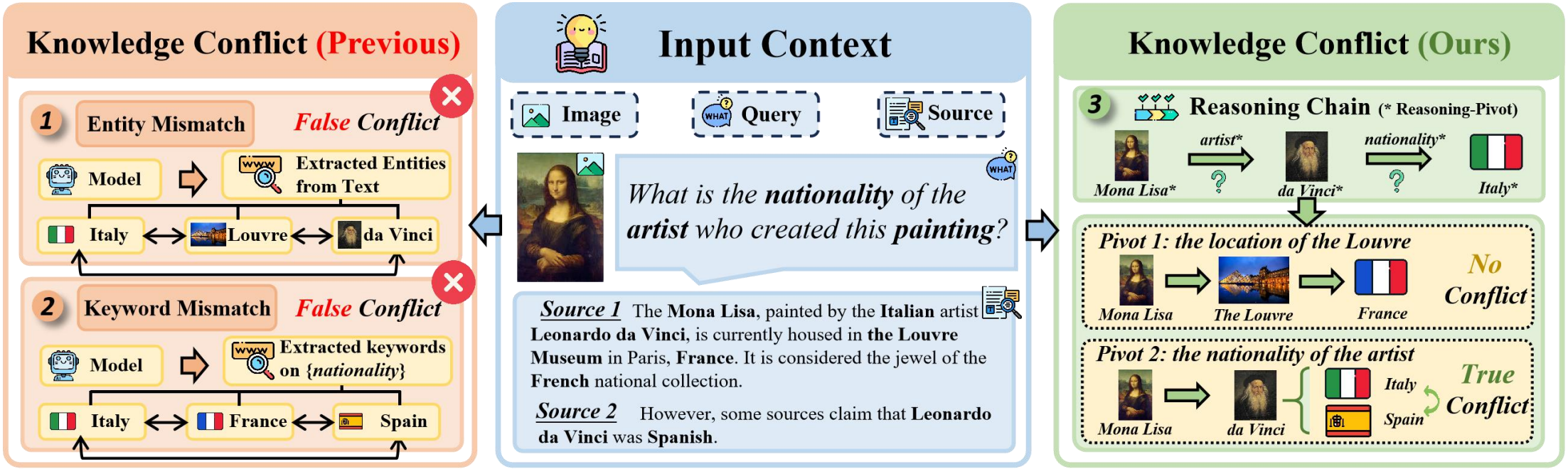
Motivation

The Paradox of Open-Domain Retrieval



- **Goal**: Enhance reasoning accuracy via external open-domain knowledge retrieval.
- **Reality***:* Inevitable retrieval noise introduces contradictions, collapsing logical chains.

Rethinking Conflicts

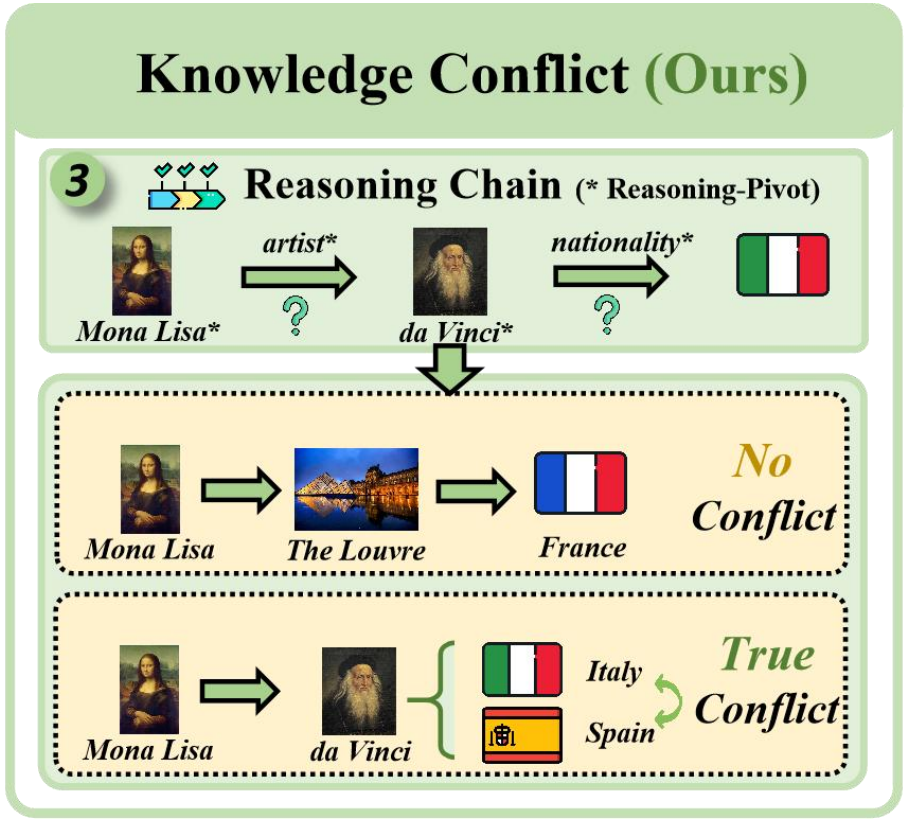


- **Conventional Methods**: Incorrectly flag valid multi-hop entity variations as false conflicts.

- **Our Reasoning-Pivot**: Accurately isolates true logical contradictions strictly within the essential reasoning node.

Definition

Defining the Reasoning-Pivot

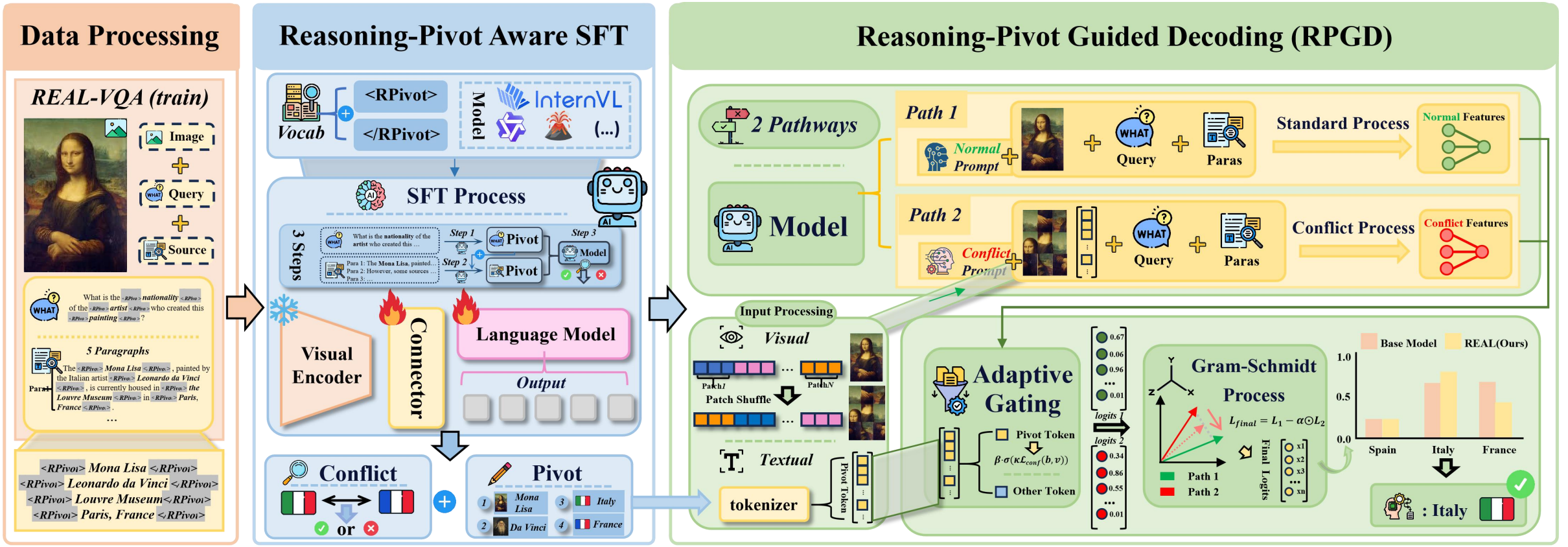


$$e_{img} \xrightarrow{p_1} e_2 \xrightarrow{p_2} \cdots \xrightarrow{p_n} y$$

An **indispensable atomic unit** in the logic chain.

True Conflict: Mutually exclusive assertions regarding the exact same reasoning-pivot

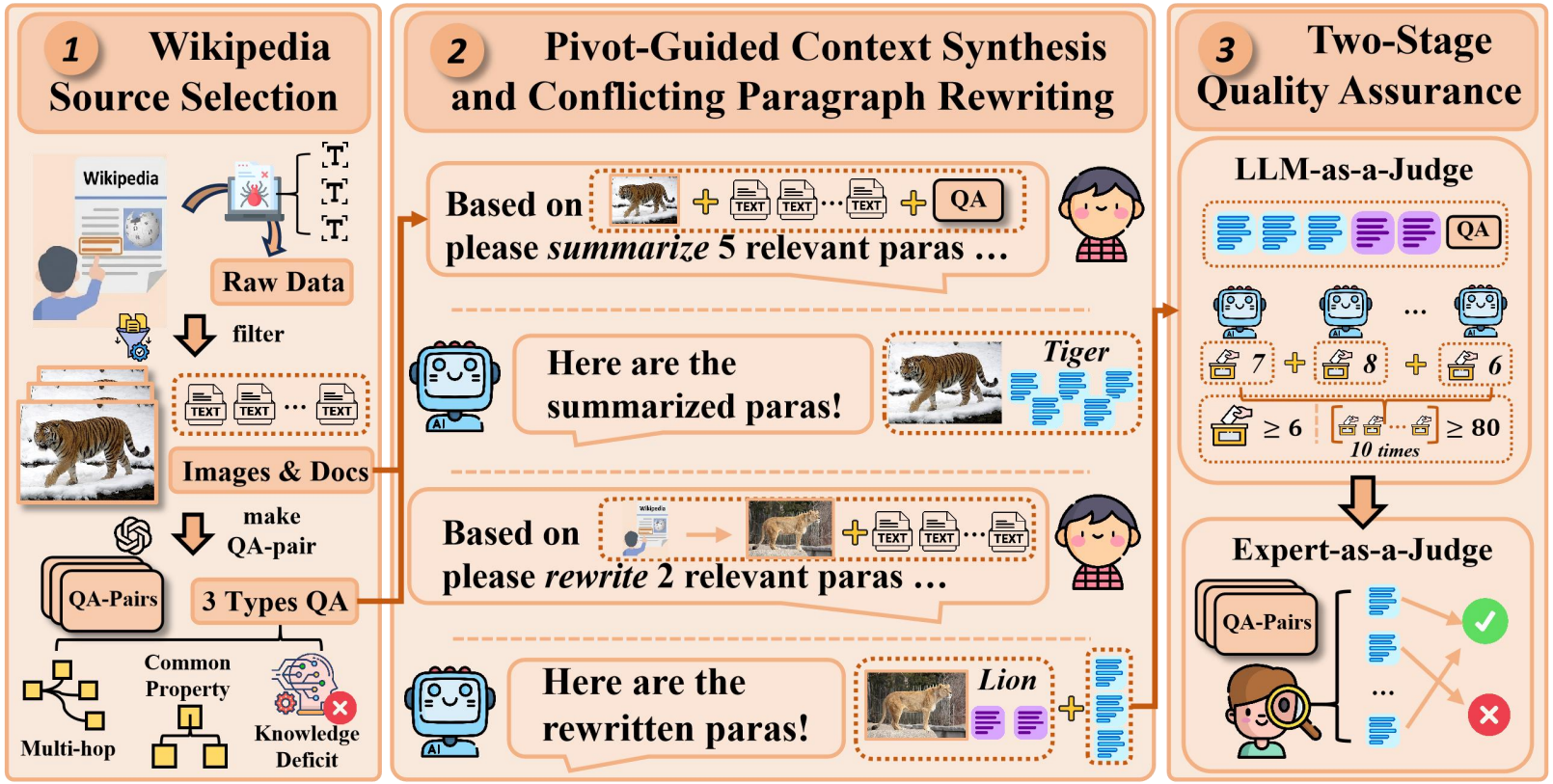
Framework



- **Stage 1 (RPA-SFT):** Aligns query and paragraph pivots via special explicit tokens.

- **Stage 2 (RPGD):** A training-free contrastive decoding strategy to structurally mitigate isolated conflicts.

Dataset



- **Counterfactual Selection**: Targets logical nodes rather than random entity swaps.
- **Context Synthesis**: Uses GPT-4o for hallucination-free passage rewriting.
- **Quality Assurance**: Strict mechanism eliminates hybrid hallucinations.

Result

Table 1. VQA accuracy scores on the E-VQA test set and the InfoSeek validation set, where **bold** values indicate the best performance and underlined values indicate the second-best performance, and † denotes results obtained with the same retriever and knowledge base, making them directly comparable.

Method	Retriever	Model	InfoSeek			E-VQA	
			Un-Q	Un-E	All	Single-Hop	All
Model Based Methods							
PromptCap† (Hu et al., 2023b)	EVA-CLIP-8B	Flan-T5-11B	4.3	3.8	4.1	10.8	11.4
Prophet++† (Shao et al., 2023)	EVA-CLIP-8B	Qwen3-VL-8B	13.2	11.6	12.3	10.8	11.4
NoteMR† (Fang et al., 2025)	EVA-CLIP-8B	Qwen3-VL-8B	28.7	29.8	29.2	25.6	23.6
VKC-MIR† (Ye et al., 2025)	OPT-66B	mPLUG-Owl3-8B	–	–	25.1	–	–
Retrieval Augmented Methods							
EchoSight† (Yan & Xie, 2024)	EVA-CLIP-8B	LLaVA-1.5-7B	27.3	26.3	26.8	31.1	28.5
Wiki-LLAVA (Caffagni et al., 2024)	CLIP-ViT-L	LLaVA-1.5-7B	30.1	27.8	28.9	25.7	27.1
RORA-VLM (Qi et al., 2024)	CLIP+Google Search	LLaVA-1.5-7B	25.1	27.3	26.2	–	20.3
ReflectiVA (Cocchi et al., 2025)	EVA-CLIP-8B	LLaMA3.1-8B	40.4	39.8	40.2	35.5	35.5
mKG-RAG (Yuan et al., 2025)	CLIP-ViT-L	LLaMA3-8B	41.4	39.6	40.5	38.4	36.3
VLM-PRF (Hong et al., 2025)	EVA-CLIP-8B	LLaMA3.1-8B	41.3	40.6	40.8	36.3	35.5
VLM-PRF (Hong et al., 2025)	EVA-CLIP-8B	InternVL3-8B	43.5	42.1	42.5	40.1	39.2
Knowledge Conflict Based Method							
REAL(Ours)†	EVA-CLIP-8B	LLaVA-1.5-7B	30.6	33.0	31.8	33.5	30.9
REAL(Ours)†	EVA-CLIP-8B	InternVL3.5-8B	**43.8**	43.7	43.8	43.9	39.2
REAL(Ours)†	EVA-CLIP-8B	Qwen3-VL-2B	33.2	33.6	33.4	40.0	35.4
REAL(Ours)†	EVA-CLIP-8B	Qwen3-VL-8B	43.1	**45.1**	**44.1**	**45.5**	**41.4**

Table 2. VQA accuracy scores on the A-OKVQA benchmark, where **MC** denotes multiple-choice and **DA** denotes direct-answer accuracy, and **bold** values indicate the best performance.

Method	Model	MC	DA
ClipCap (Mokady et al., 2021)	–	44.0	18.1
KRISP (Marino et al., 2021)	–	51.9	33.7
GPV-2 (Kamath et al., 2022)	–	60.3	48.6
PromptCap (Hu et al., 2023b)	GPT-3	73.2	56.3
Prophet (Shao et al., 2023)	GPT-3	76.4	58.2
ASB (Xenos et al., 2023)	LLaMA-2-13B	-	58.6
SKP (Wang et al., 2024)	LLaVA-1.5-7B	-	65.3
QACap (Yang et al., 2025)	Claude 3.5	76.7	66.3
REAL(Ours)	LLaVA-1.5-7B	**80.3**	**68.3**

- Evaluated on **E-VQA, Infoseek, A-OKVQA**.
- Achieves **outstanding performance** across all benchmarks.

Result

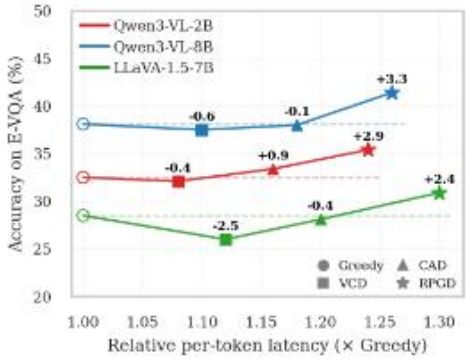
Model	Method	REAL-VQA		E–VQA		ScienceQA		MMKC	
		MCC	F1	MCC	F1	MCC	F1	MCC	F1
LLaVA-1.5-7B	Zero-shot	2.9	52.3	5.5	50.9	3.6	49.5	0.7	47.0
	Few-shot CoT	-3.7	47.4	5.1	49.6	3.5	48.2	0.0	45.5
	SFT	68.1	84.6	**38.4**	69.8	34.6	72.2	19.6	66.3
	RPA-SFT(Ours)	**89.4**	**89.6**	37.7	**70.0**	**58.0**	**77.0**	**45.5**	**72.3**
InternVL3.5-8B	Zero-shot	9.3	70.6	52.2	82.6	37.3	39.2	61.9	85.6
	Few-shot CoT	11.5	70.8	72.5	89.5	50.8	58.5	70.5	82.4
	SFT	88.4	96.1	77.7	89.1	79.4	88.5	73.1	**85.8**
	RPA-SFT(Ours)	**95.6**	**98.5**	**78.6**	**90.5**	**85.8**	**90.8**	**73.7**	**85.8**
Qwen3-VL-8B	Zero-shot	19.0	69.9	85.4	94.5	64.5	74.8	23.4	66.9
	Few-shot CoT	19.4	72.3	86.9	95.3	67.4	77.1	42.4	71.4
	SFT	89.4	96.8	82.6	91.4	87.0	93.1	38.2	73.2
	RPA-SFT(Ours)	**98.1**	**99.1**	**93.4**	**95.5**	**87.9**	**95.4**	**52.9**	**74.8**

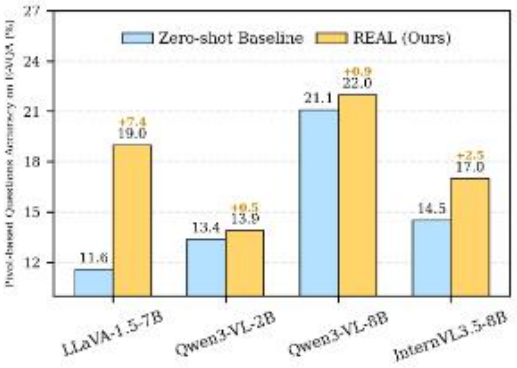
- **Robust Generalization**
- **Substantial F1 discrimination gain** across diverse model scales.

Result

High Efficiency & Component Synergy

Model	Method	InfoSeek			E-VQA	
		Un-Q	Un-E	All	Single-Hop	All
LLaVA-1.5-7B	Zero-shot	8.1	7.3	7.7	11.5	11.5
	REAL(w/o RPGD)	27.3	26.3	26.8	31.1	28.5
	REAL(Ours)	**30.6**	**33.0**	**31.8**	**33.5**	**30.9**
InternVL3.5-8B	Zero-shot	10.3	8.5	9.3	14.3	14.3
	REAL(w/o RPGD)	36.6	36.1	36.2	39.0	35.0
	REAL(Ours)	**43.8**	**43.7**	**43.8**	**43.9**	**39.2**
Qwen3-VL-8B	Zero-shot	20.4	18.2	19.2	20.1	20.3
	REAL(w/o RPGD)	38.0	39.4	38.7	42.4	38.1
	REAL(Ours)	**43.1**	**45.1**	**44.1**	**45.5**	**41.4**





RPGD Components			E-VQA	
Patch Shuffle	Adaptive Gating	Gram Schmidt	Single-Hop	All
✗	✗	✗	42.4 ↓3.1	38.1 ↓3.3
✗	✓	✓	43.9 ↓1.6	39.2 ↓2.2
✓	✗	✓	44.1 ↓1.4	39.5 ↓1.9
✓	✓	✗	43.5 ↓2.0	38.9 ↓2.5
✓	✓	✓	45.5	41.4

- **All components** (Patch Shuffle, Gating, Orthogonalization) are essential.
- Maintains inference latency within **1.3x** of standard greedy decoding.

Thank you!!!

Paper

Personal Homepage

Feel free to contact us: **mercury0926@zju.edu.cn ; zhousheng_zju@zju.edu.cn**

浙江大学 ZHEJIANG UNIVERSITY

Alibaba Group 阿里巴巴集团

ICML International Conference On Machine Learning