Wiki-R1: Incentivizing Multimodal Reasoning for Knowledge-based VQA via Data and Sampling Curriculum

Shan Ning, Longtian Qiu, Xuming He

ShanghaiTech University,Shanghai Engineering Research Center of Intelligent Vision and Imaging, Lingang Laboratory

Paper & Code

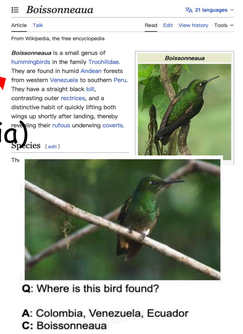
Contribution

- We propose Wiki-R1, a data-generation-based curriculum RL framework that incentivizes the reasoning ability of MLLMs on KB-VQA.
- Wiki-R1 employs a data curriculum strategy to constructs a training distributions, and introduces a curriculum sampling method to select valuable data.
- Experimental results show that Wiki-R1 consistently outperforms prior SOTAs on challenging KB-VQA benchmarks, with pronounced improvements in unseen settings.

Task Definition

Knowledge-based VQA requires:

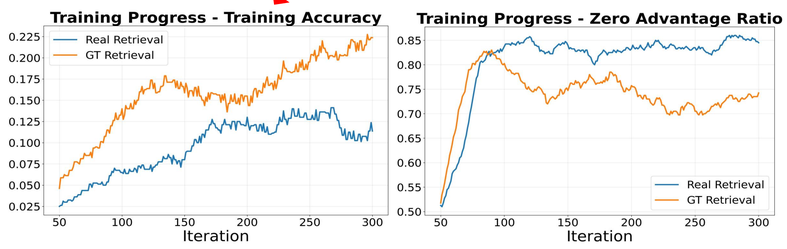
- Outside knowledge base(e.g. Wikipedia)
- Fine-grained visual perception
- Encyclopedic text understanding



Motivation

We observe two empirical phenomena in RL post-training for KB-VQA.

1. Training samples yield zero advantage, accompanied by reasoning collapse
2. Training accuracy remains low.



Where RL Might Struggle for KB-VQA?

- **Noisy Retrieval system increase overall task difficulty.**
  - Retrieval imperfection leads to noisy candidates, confusing the model in choosing valid context.
- **MLLM inherent distributions diverge from the target distribution of RAG-driven KB-VQA.**

Methodology

Our insights: Construct a curriculum learnable distribution for RL post-training

- Empower the policy model to learn valuable and informative signals.
- Boost training efficiency and stabilize post-training optimization.

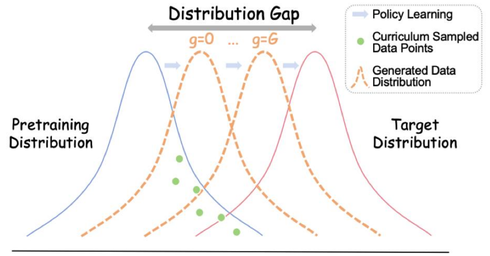
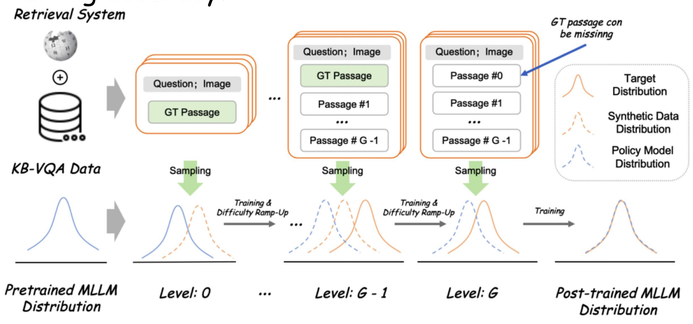


Illustration of Wiki-R1's Motivation

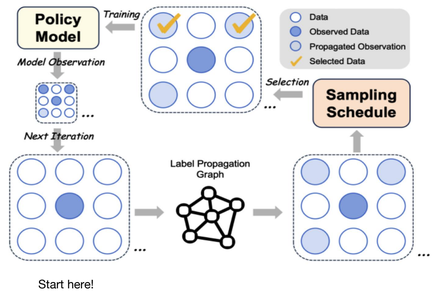
Controllable curriculum data generation

- We manipulate the retriever to generate training samples with gradually increasing difficulty.
- The level is dynamically adjusted based on observed training accuracy.



Curriculum sampling with observation propagation

- High-value samples with valid training signals yield an expected accuracy 0.5.
- We build relation graphs based on the source articles of VQA samples.
- Observations are propagated to unobserved samples via relation graphs.



Experiments

Table 1: **Performance comparison on Encyclopedic VQA and InfoSeek.** All results of retrieva augmented generation methods are reported without applying any re-ranking stage to reorder re trieved documents. *Retrieval Mode* spans two columns: the first specifies the retrieval model, whil the second indicates the type of knowledge source utilized. The *V.* and *T.* indicate the visual an textual retrieval mode. The *Con.* and *Col.* indicate textual retrieval model, Contriver (Izacard et al 2021) and Colbert V2 (Santhanam et al., 2021) respectively.

Method	Retrieval Mode		EVQA		InfoSeek			Avg.
			Single-hop	All	Unseen-Q	Unseen-E	All	
Zero-shot MLLMs								
BLIP-2	-	-	12.6	12.4	12.7	12.3	12.5	12.5
InstructBLIP	-	-	11.9	12.0	8.9	7.4	8.1	10.1
LLaVA-1.5 7B	-	-	16.0	16.9	8.3	8.9	7.8	12.4
Qwen-2.5-VL 3B	-	-	18.6	18.8	26.3	16.1	19.6	19.2
Qwen-2.5-VL 7B	-	-	26.6	26.3	25.3	17.2	19.9	23.1
GPT-4V	-	-	26.9	28.1	15.0	14.3	14.6	21.4
Retrieval-Augmented Generation								
DPR_{V+T}	CLIP ViT-B/32	V. + T.	29.1	-	-	-	12.4	-
RORA-VLM	CLIP+Google Search	V. + T.	-	20.3	25.1	27.3	-	-
Wiki-LLaVA	CLIP ViT-L/14+Con.	T.	18.3	19.6	28.6	25.7	27.1	23.4
EchoSight	EVA-CLIP-8B	T.	22.4	21.7	30.0	30.7	30.4	26.1
EchoSight	EVA-CLIP-8B	V.	26.4	24.9	18.0	19.8	18.8	21.9
ReflectiVA	CLIP ViT-L/14	T.	24.9	26.7	34.5	32.9	33.7	30.2
ReflectiVA	EVA-CLIP-8B	T.	28.0	29.2	40.4	39.8	40.1	34.7
ReflectiVA	EVA-CLIP-8B	V.	35.5	35.5	28.6	28.1	28.3	31.9
Wiki-R1 3B	EVA-CLIP-8B + Col.	V.+ T.	40.4	35.9	46.0	40.3	42.2	39.1
Wiki-R1 7B	EVA-CLIP-8B + Col.	V.+ T.	**41.0**	**37.1**	**47.8**	**42.3**	**44.1**	**40.6**

Table 4: **Ablation study of framework design on Encyclopedic VQA and InfoSeek.** We conduct experiments on Qwen-2.5-VL 3B model. Each row progressively adds components, and we mark enabled modules with ✓. The *Samp. Cur.*, *Data Cur.*, *Obs. Prop.* indicate the sampling curriculum, data curriculum generation, and observation propagation strategies.

Method	Modules			EVQA		InfoSeek		
	Data Cur.	Samp. Cur.	Obs. Prop.	Single-hop	Overall	Unseen-Q	Unseen-E	Overall
Zero-shot	-	-	-	18.6	18.8	26.3	16.1	19.6
SFT	-	-	-	21.6	25.1	38.7	24.9	29.5
SFT	✓	-	-	-	34.4	-	-	32.1
DAPO	×	×	×	35.9	31.4	44.9	39.8	41.5
	✓	×	×	39.4	34.5	**46.9**	**41.1**	**43.0**
	✓	✓	×	36.4	32.1	45.2	37.3	40.0
	✓	✓	✓	**40.4**	**35.9**	46.0	40.3	42.2

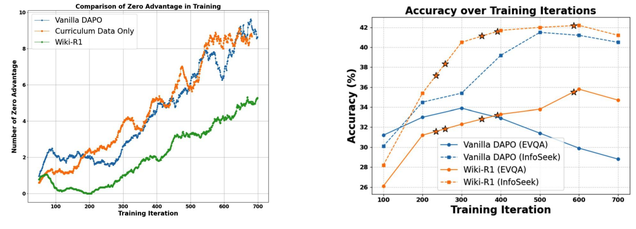


Figure 3: **Left: Number of ignored trajectories.** Trajectories are ignored when they provide zero advantage and no training signal; a larger number indicates lower training efficiency. **Right: Accuracy over training iterations.** Performance is reported on the EVQA test set and the InfoSeek validation set. The *star* denotes an increase in curriculum difficulty during Wiki-R1 training.