FrameOracle: Learning What to See and How Much to See in Videos

Chaoyu Li[1,2], Tianzhi Li[2,3], Fei Tao[1], Zhenyu Zhao[1], Ziqian Wu[1], Maozheng Zhao[1], Juntong Song[1], Cheng Niu[1], Pooyan Fazli[2]

[1]NewsBreak [2]Arizona State University [3]Carnegie Mellon University

ICML International Conference On Machine Learning

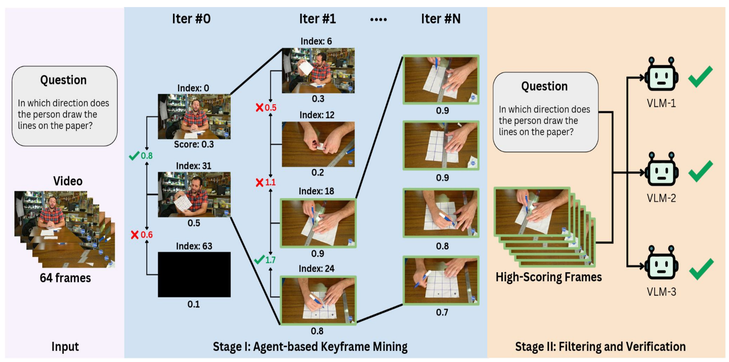
Motivation

- Most VLMs can only process limited number of frames, making performance highly dependent on whether the selected frames contain the right visual evidence.
- Common strategies such as uniform sampling or fixed-budget keyframe selection are not adaptive:
  - – Long videos may miss salient moments.
  - – Short videos may include redundant frames.
  - – Different questions over the same video require different number of frames.

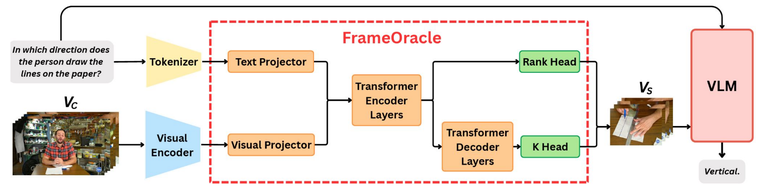
Contributions

- FrameOracle: a lightweight, **plug-and-play**, backbone-agnostic frame selector that jointly predicts:
  - – **what to see**: which frames are most relevant to the query;
  - – **how much to see**: how many frames are sufficient.
- FrameOracle-41K: the first large-scale VideoQA dataset with question -conditioned keyframe annotations, specifying the minimal sufficient frames needed to answer each question.

FrameOracle-41K: Pipeline

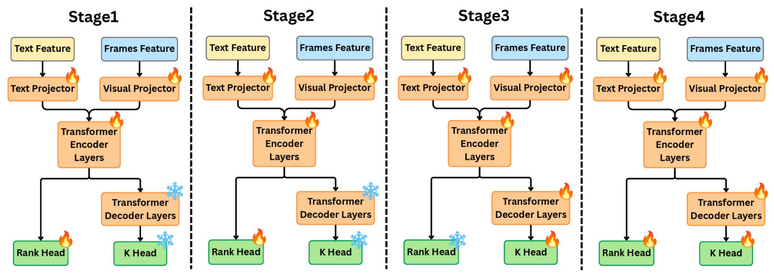


FrameOracle



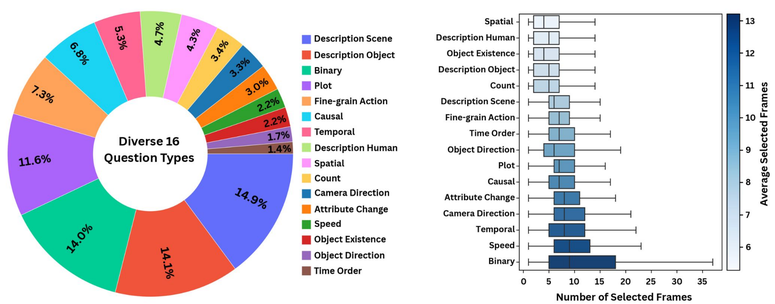
- **Rank Head**: Scores each candidate frame by question relevance.
- **K Head**: Predicts the adaptive number of frames instead of using a fixed budget.

➢ Training Curriculum



FrameOracle-41K: Statistics

- FrameOracle-41K contains **40,992** samples.
- Each sample provides the 1) **key frames** and 2) **minimal sufficient number of frames** needed to answer the question.



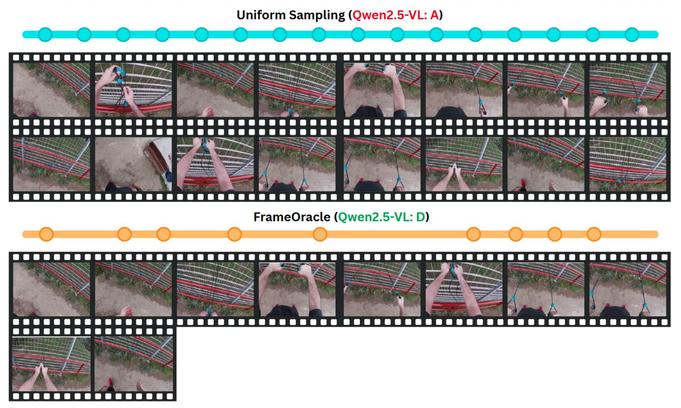
Experiments

Model	Frames	NExTQA			Perception	LVB	Video-MME	EgoSchema	MLVU	Avg.
		OE_val	OE_test	MC						
Qwen2.5-VL-3B (Bai et al., 2025b)	32	25.1	29.6	75.4	65.9	54.1	58.4	53.4	59.4	52.7
+ **FrameOracle**	32→20.9	25.6	30.5	74.8	66.7	54.3	58.5	53.8	58.4	52.8
+ **FrameOracle**	128→27.8	**26.0**	**31.7**	**76.1**	**67.8**	**54.8**	**59.7**	**54.5**	**61.6**	**54.0**
LLaVA-OneVision-7B (Li et al., 2025a)	16	14.6	16.7	78.2	56.4	55.0	56.1	60.8	60.9	49.8
+ **FrameOracle**	16→10.4	16.1	17.8	77.6	56.5	55.5	56.0	62.4	60.2	50.3
+ **FrameOracle**	64→13.9	**16.5**	**19.0**	**78.5**§	**56.9**	**56.5**	**58.1**	**63.4**	**63.7**	**51.6**
LLaVA-Video-7B (Zhang et al., 2025e)	16	27.3	32.4	81.0	64.3	55.8	59.8	54.2	61.7	54.6
+ **FrameOracle**	16→10.4	27.8	33.0	80.4	64.7	56.3	59.6	54.6	60.8	54.7
+ **FrameOracle**	64→13.9	**28.8**	**33.9**	**81.6**	**65.1**	**57.8**	**61.6**	**55.2**	**64.3**	**56.0**
VideoLLaMA3-7B (Zhang et al., 2025a)	16	27.8	32.3	**82.3**	72.3	56.1	61.2	61.4	50.9	55.5
+ **FrameOracle**	16→10.4	28.3	32.9	81.2	72.0	56.0	61.4	61.8	52.8	55.8
+ **FrameOracle**	64→13.9	**28.9**	**33.6**	82.0§	**72.8**	**56.9**	**61.8**	**62.4**	**54.1**	**56.6**
Qwen3-VL-8B (Bai et al., 2025a)	32	26.0	31.1	76.6	67.5	63.3	66.9	70.8	63.6	58.2
+ **FrameOracle**	32→20.9	26.6	32.3	76.1	68.2	64.0	67.3	71.4	62.9	58.6
+ **FrameOracle**	128→27.8	**28.1**	**33.8**	**77.3**	**69.0**	**65.2**	**69.1**	**72.3**	**66.3**	**60.1**

- With 16 frames input: FrameOracle reduces to **10.4 frames** on average while preserving accuracy.
- With 64 frames intput: FrameOracle improves average accuracy by +1.5% using only **13.9 frames**.
- These gains are validated across **5 VLM backbones** and **6 video benchmarks**.

Qualitative Example

Question: What seems to be the main purpose of the video? What actions did c perform to achieve this purpose?
A: The main objective of this instructional video is to effectively demonstrate how to easily tie your hair back.
B: The main purpose of the video is to show how to open a jar.
C: The primary objective of the video presentation is to demonstrate the most effective methods for properly cleaning your windows.
D: The main purpose of the video is to show how to use a resistance band to exercise your arms and upper body.
E: The primary objective of this video presentation is to effectively demonstrate the proper way to engage in a fun tug-of-war match with your canine companion.



arXiv

Website