# ReFoCUS: Reinforcement-guided Frame Optimization for Contextual Understanding

Hosu Lee[1]* · Junho Kim[2]* · Hyunjun Kim[1] · Yong Man Ro[1]
[1] Korea Advanced Institute of Science & Technology (KAIST) [2] University of Illinois Urbana-Champaign (* equal contribution)

KAIST · UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN · CVPR JUNE 3-7, 2026 DENVER COLORADO

arXiv | Project | HuggingFace | Hosu Lee | Junho Kim

## Frame selection via *learned policy*, not a heuristic.

Model wasn't reasoning wrong. It just **never saw the evidence**.

**Q.** Which time is displayed on the clock in the video?


Hidden-clock video questions are conventionally depicted at 10:10, `so the answer is 10:10.` ✗

Locating the clock in the third through fifth frames, the minute hand points straight up at 12 while the hour hand rests on the 5, `so the time reads 5:00.` ✓

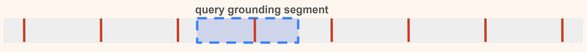
**Take away**

1. RL aligned what models say. **We're the first to align what they see.**
2. No frame labels. No retriever. **Any VLM's answer-confidence is the reward.**
3. From sub-1B to GPT-4o and Gemini, **one selector lifts them all.**

## 1 Problem — Why frame selection is the real bottleneck.

### Uniform sampling is blind to the query

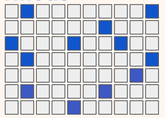
Most video-LLMs treat a video as uniformly sampled static frames. Under a tight frame budget the **decisive evidence is usually missed**
→ Thus, accuracy is capped before reasoning even begins.




### And the search space itself is also enormous

Selecting 32 of 512 candidate frames:

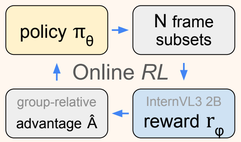
$$\binom{512}{32} \approx 7 \times 10^{50}$$



Hand-labeling "BEST" frame subsets is infeasible
→ ***ReFoCUS*** **learns the selection** with reinforcement learning instead.

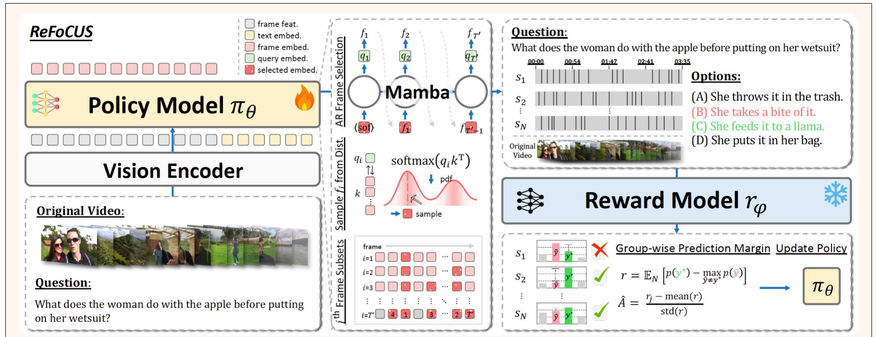
### Our approach: ReFoCUS optimizes the visual evidence a model reasons over ~~not the answer it produces.~~ We brought RL from text to vision.

| | |
|---|---|
| **Query-blind** ignores what the question asks for | **Query-conditioned** attends to the question, then picks |
| **7×10⁵⁰ subsets** far too many to search or label | **Learnable AR search** builds a subset frame-by-frame |
| **Apart from reasoning** heuristics ≠ what the model needs | **reward = VLM's confidence** aligned to downstream answering |



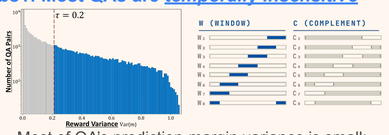

**ReFoCUS** learns to assemble the evidence frames a model reasons over with online RL.

## 2 Method — Aligning visual evidence selection with model-internal utility.




$\pi_\theta$ proposes N frame subsets → frozen $r_\varphi$ scores each by its answer margin → group-relative advantage updates $\pi_\theta$

### Obs1: Most QAs are *temporally insensitive*



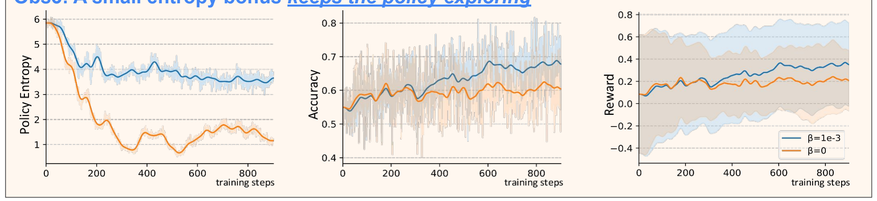

Most of QA's prediction-margin variance is small; we keep τ > 0.2. 962K → 393K .

### Obs2: search space is *too vast to explore*

| SEARCH SPACE | SETTING | Δ ON VIDEO-MME · 32-FRAME EVAL |
|---|---|---|
| $7\times10^{50}$ | 32 NO CURRICULUM | +0.6 |
| $2.8\times10^{9}$ | 4 | +1.6 |
| $5.4\times10^{16}$ | 4→8 | +2.6 |
| $5.7\times10^{50}$ | 4→8→16 | +3.0 |
| $7\times10^{50}$ ▼ | 4→8→16→32 | +3.2 |

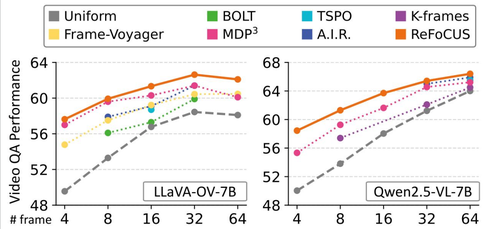
Expand the budget in stages: 4 → 8 → 16 → 32. Curriculum +3.2 vs. +0.6 from the full space.

### Obs3: A small entropy bonus *keeps the policy exploring*

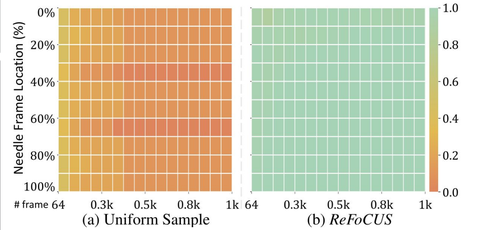



## 3 Results — One selector lifts every backbone and the selected frames are indispensable.

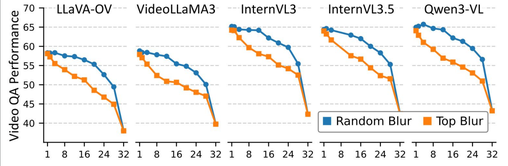
### vs. SoTA Baselines




### Needle-in-a-Haystack: Pinpoints Sparse Evidence




### Evidence Suppression: Selected frames are indispensable




### Model-Agnostic: Consistent Gains Across Backbones

| Model | LLM | Video-MME (w/o sub) short | medium | long | overall | LVB acc. (val) | MLVU m-avg | VMMMU overall | NExT-QA wups (val) |
|---|---|---|---|---|---|---|---|---|---|
| *Closed Source* | | | | | | | | | |
| Gemini 2.5 Flash [10] | - | 77.6 | 63.7 | 56.8 | 66.0 | 47.9 | 52.8 | 40.6 | 11.7 |
| + **ReFoCUS** | - | 79.3 ↑1.7 | 68.3 ↑4.6 | 60.8 ↑4.0 | 69.5 ↑3.5 | 50.9 ↑3.0 | 58.0 ↑5.2 | 45.6 ↑5.0 | 11.9 ↑0.2 |
| GPT-4o [37] | - | 68.0 | 55.0 | 53.3 | 58.8 | 49.5 | 58.7 | 62.9 | 8.5 |
| + **ReFoCUS** | - | 68.2 ↑0.2 | 60.1 ↑5.1 | 54.0 ↑0.7 | 60.8 ↑2.0 | 52.9 ↑3.4 | 65.1 ↑6.4 | 62.1 ↓-0.8 | 9.1 ↑0.6 |
| *Open Source* | | | | | | | | | |
| LLaVA-OneVision [22] | 0.5B | 53.7 | 39.9 | 37.0 | 43.5 | 44.7 | 44.8 | 17.3 | 18.1 |
| + **ReFoCUS** | 0.5B | 58.3 ↑4.6 | 44.6 ↑4.7 | 38.3 ↑1.3 | 47.1 ↑3.6 | 48.7 ↑4.0 | 50.3 ↑5.5 | 19.4 ↑2.1 | 18.9 ↑0.8 |
| InternVL3 [73] | 1B | 63.1 | 46.9 | 39.9 | 50.0 | 47.6 | 54.0 | 27.7 | 20.0 |
| + **ReFoCUS** | 1B | 66.4 ↑3.3 | 51.8 ↑4.9 | 42.6 ↑2.7 | 53.6 ↑3.6 | 50.6 ↑3.0 | 58.9 ↑4.9 | 29.3 ↑1.6 | 20.4 ↑0.4 |
| VideoLLaMA 3 [67] | 2B | 55.2 | 38.8 | 35.2 | 43.1 | 48.8 | 46.8 | 28.7 | 18.9 |
| + **ReFoCUS** | 2B | 58.9 ↑3.7 | 44.1 ↑5.3 | 38.3 ↑3.1 | 47.1 ↑4.0 | 53.7 ↑4.9 | 50.2 ↑3.4 | 29.2 ↑0.5 | 20.7 ↑1.8 |
| InternVL3 [73] | 2B | 71.0 | 56.4 | 47.8 | 58.4 | 50.9 | 62.7 | 38.3 | 24.4 |
| + **ReFoCUS** | 2B | 72.2 ↑1.2 | 60.2 ↑3.8 | 49.7 ↑1.9 | 60.7 ↑2.3 | 54.9 ↑4.0 | 68.0 ↑5.3 | 39.3 ↑1.0 | 25.0 ↑0.6 |
| InternVL3.5 [52] | 4B | 76.4 | 60.3 | 51.3 | 62.7 | 57.7 | 66.6 | 52.0 | 22.1 |
| + **ReFoCUS** | 4B | 78.0 ↑1.6 | 62.3 ↑2.0 | 57.4 ↑6.1 | 65.9 ↑3.2 | 62.6 ↑4.9 | 71.5 ↑4.9 | 53.3 ↑1.3 | 22.9 ↑0.8 |
| Qwen3-VL [59] | 4B | 74.1 | 61.0 | 51.3 | 62.1 | 57.4 | 63.1 | 54.0 | 23.8 |
| + **ReFoCUS** | 4B | 76.7 ↑2.6 | 65.7 ↑4.7 | 57.0 ↑5.7 | 66.4 ↑4.3 | 61.9 ↑4.5 | 71.9 ↑8.8 | 56.4 ↑2.4 | 24.1 ↑0.3 |
| VideoLLaMA 3 [67] | 7B | 70.4 | 57.7 | 48.9 | 59.0 | 54.8 | 52.9 | 32.8 | 25.8 |
| + **ReFoCUS** | 7B | 72.2 ↑1.8 | 60.1 ↑2.4 | 54.3 ↑5.4 | 62.2 ↑3.2 | 57.0 ↑2.2 | 59.8 ↑6.9 | 34.4 ↑1.6 | 26.5 ↑0.7 |
| LLaVA-OneVision [22] | 7B | 70.9 | 55.7 | 48.8 | 58.4 | 55.0 | 63.7 | 34.1 | 16.2 |
| + **ReFoCUS** | 7B | 72.8 ↑1.9 | 61.7 ↑6.0 | 53.4 ↑4.6 | 62.6 ↑4.2 | 61.0 ↑6.0 | 68.5 ↑4.8 | 35.7 ↑1.6 | 16.4 ↑0.2 |
| InternVL3 [73] | 8B | 75.1 | 64.4 | 53.4 | 64.3 | 57.8 | 68.1 | 49.3 | 26.6 |
| + **ReFoCUS** | 8B | 75.8 ↑0.7 | 66.8 ↑2.4 | 58.3 ↑4.9 | 67.0 ↑2.7 | 62.0 ↑4.2 | 72.7 ↑4.6 | 50.6 ↑1.3 | 26.8 ↑0.2 |
| InternVL3.5 [52] | 8B | 77.4 | 62.4 | 53.2 | 64.4 | 59.7 | 67.3 | 50.0 | 24.3 |
| + **ReFoCUS** | 8B | 76.2 ↓-1.2 | 64.9 ↑2.5 | 58.9 ↑5.7 | 66.7 ↑2.3 | 64.1 ↑4.4 | 70.6 ↑3.3 | 53.2 ↑3.2 | 24.7 ↑0.4 |
| Qwen3-VL [59] | 8B | 75.1 | 64.6 | 55.3 | 65.0 | 56.6 | 63.0 | 59.1 | 25.3 |
| + **ReFoCUS** | 8B | 79.6 ↑4.5 | 67.0 ↑2.4 | 58.9 ↑3.6 | 68.5 ↑3.5 | 63.3 ↑6.7 | 72.5 ↑9.5 | 61.1 ↑2.0 | 25.7 ↑0.4 |

### Robust to Policy Scale & Reward Choice

| Train Configuration $\pi_\theta$ | $r_\varphi$ | VMME | LVB | MLVU | VMMMU |
|---|---|---|---|---|---|
| *Effect of Policy Model Scale* | | | | | |
| 0.3B | InternVL3 2B | 64.8 | 59.2 | 66.3 | 45.9 |
| 1.3B | InternVL3 2B | 65.4 | 61.4 | 68.8 | 46.5 |
| 2.7B | InternVL3 2B | 64.5 | 60.9 | 70.1 | 46.7 |
| *Effect of Reward Model Choice* | | | | | |
| 1.3B | InternVL3 2B | 65.4 | 61.4 | 68.8 | 46.5 |
| 1.3B | Qwen3-VL 4B | 65.0 | 61.2 | 70.9 | 46.2 |
| 1.3B | LLaVA-OV 7B | 64.0 | 61.0 | 68.6 | 45.1 |

### Efficiency vs. baselines

| Model | Size | TFLOPs | Time(s) | Mem(GB) |
|---|---|---|---|---|
| *Downstream Baselines* | | | | |
| LLaVA-OV [22] | 7B | 1780 | 62 | 77.3 |
| InternVL3 [73] | 8B | 2099 | 46 | 44.6 |
| *Frame Selection Methods (added to baselines)* | | | | |
| + VideoAgent [53] | GPT-4 | 691 | - | - |
| + Frame-Voyager [65] | - | - | - | 23.0 |
| + T* [61] | 8B | 393 | 35 | 19.1 |
| + MDP³ [46] | 0.9B | 309 | 10 | 17.7 |
| + **ReFoCUS** | 1.8B | 428 | 9 | 5.3 |