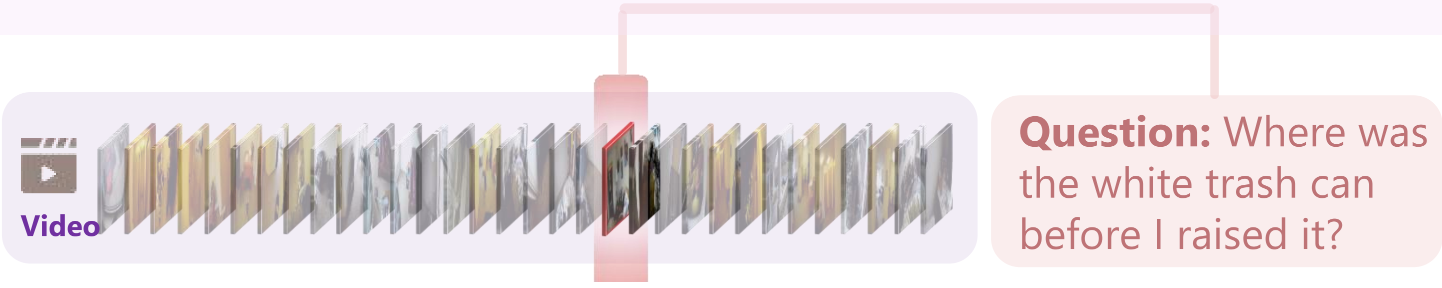
# T*: Re-thinking Temporal Search for Long-Form Video Understanding

Jinhui Ye*, Zihan Wang*, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk
Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, Jiajun Wu, Manling Li

https://mll-lab-nu.github.io/lvhaystack

Code, paper, demo, dataset

## Needle in the Long Video Haystack

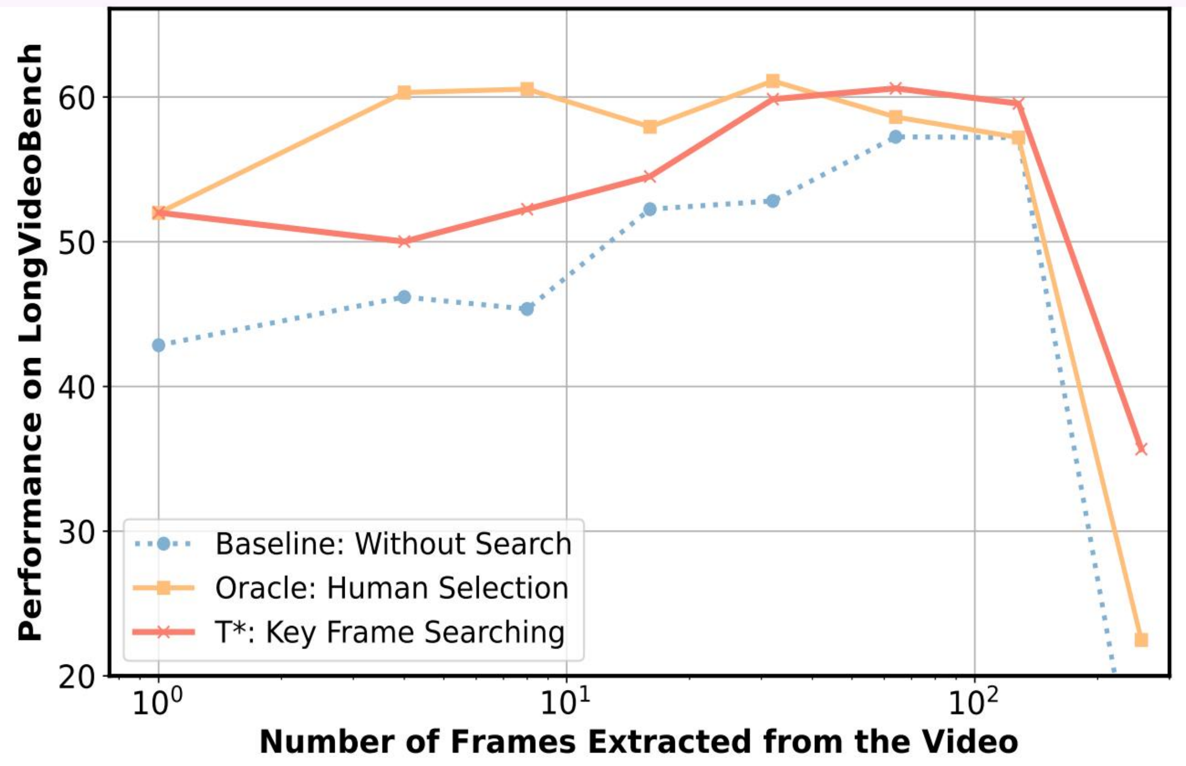



Datasets: LVHaystack / LongVideoHaystack

| Haystack-Ego4D | Haystack-LVBench |
|---|---|
| 988 videos | 246 videos |
| 432 hours | 57.7 hours |
| 15,092 QA pairs | 602 QA pairs |
| 23,800 frames | 1,070 frames |

- Objective: Select few keyframes to answer questions.
- Keyframe set must be complete and minimal.

## Current VLMs Fall Short in Long Videos



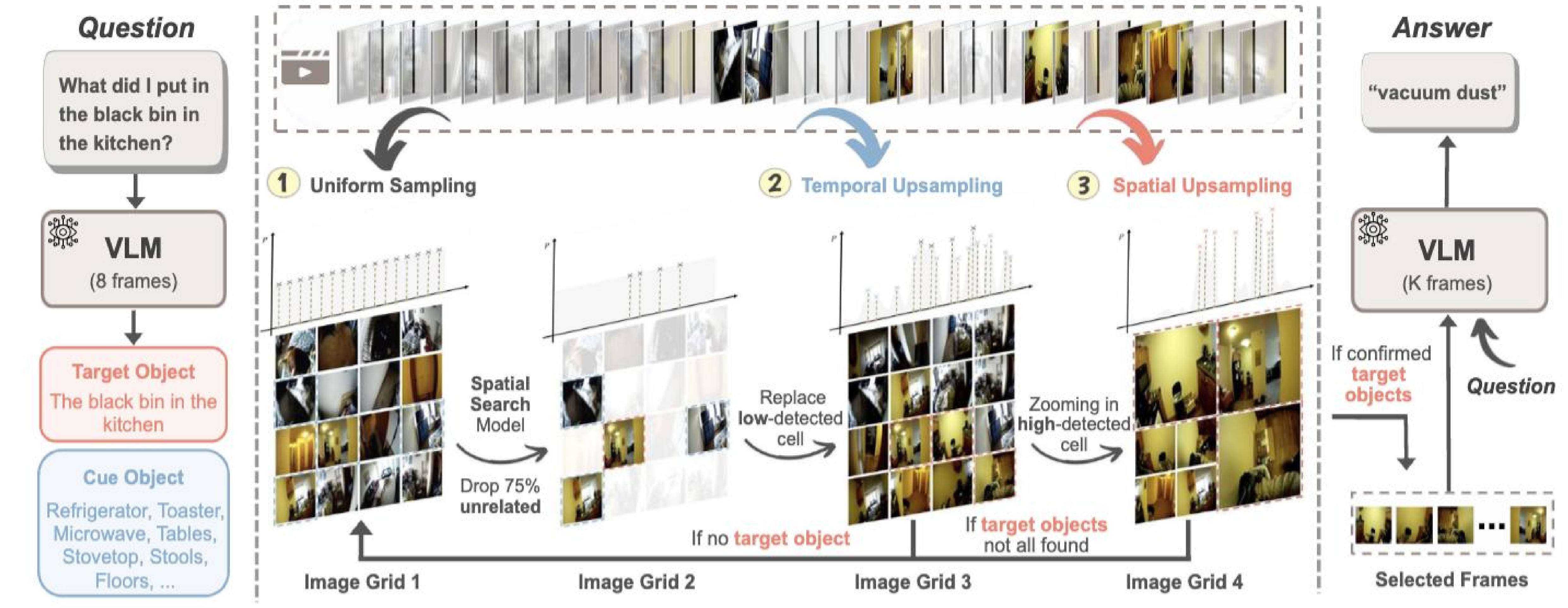

Medium-sized models (7b) **struggle** to handle more than 100 frames.

Large Models face **diminishing marginal returns** (5x more frames only brings 1.0 points)

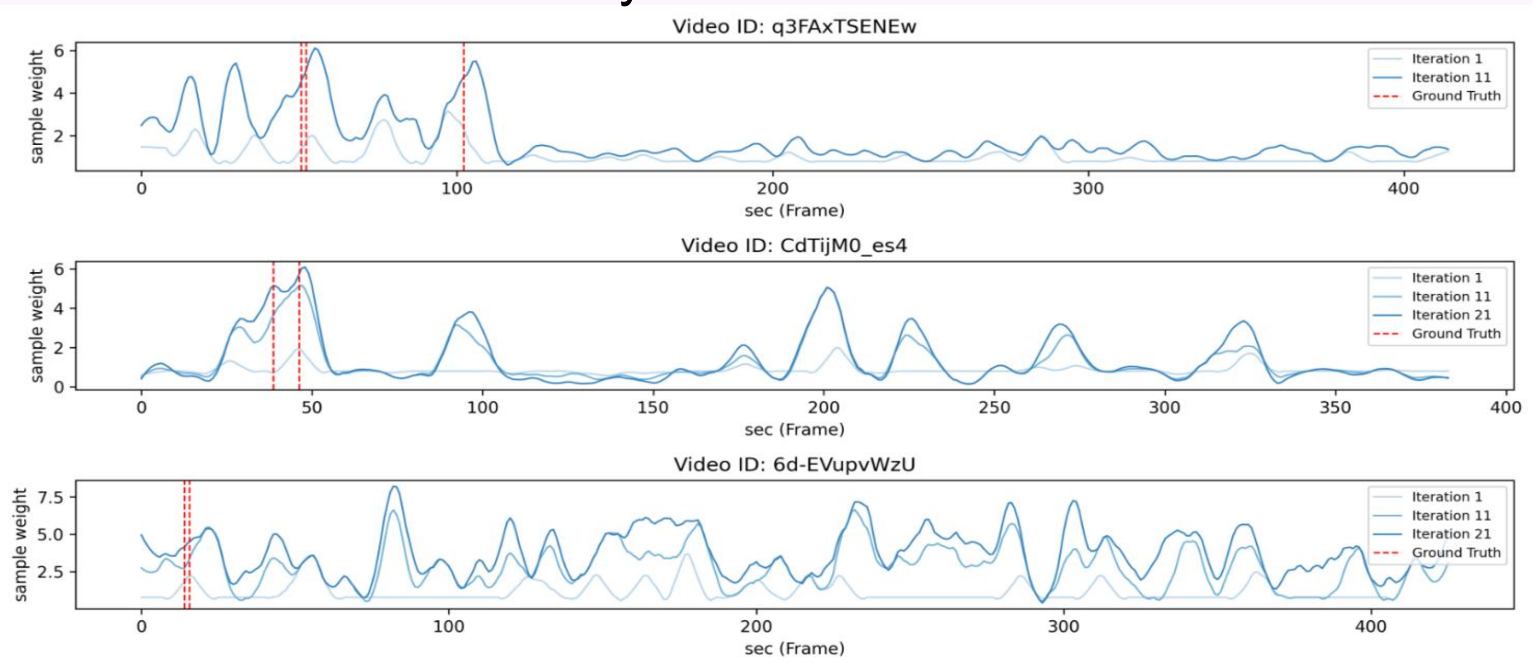
Reference: Qwen2-VL

## T*

## a light-weighted plug-in for temporal searching



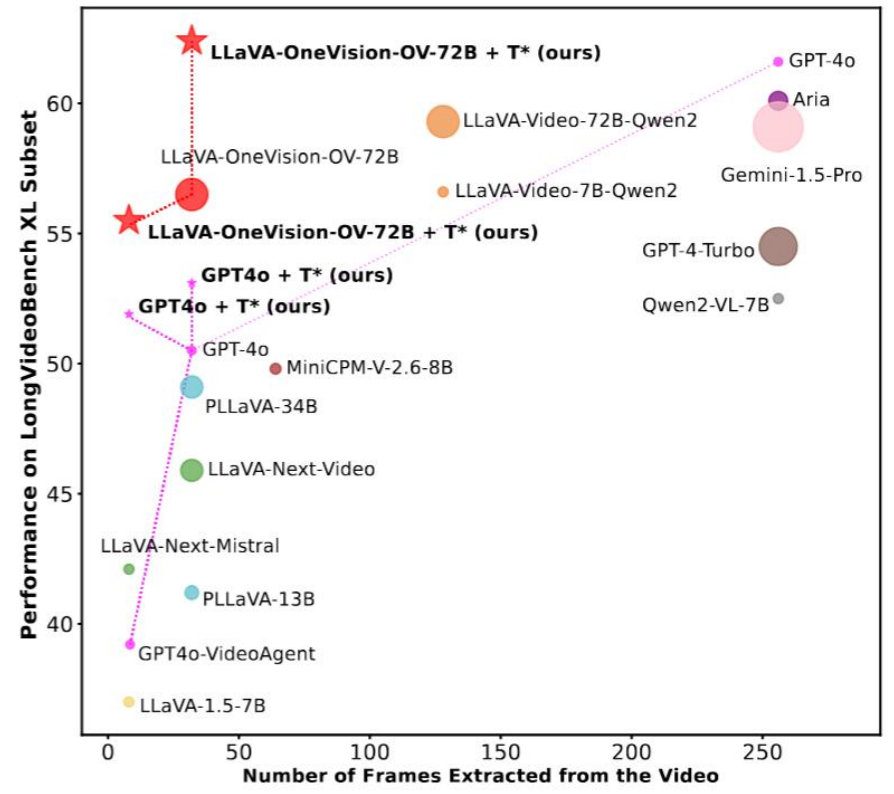

- Ground textual queries to visually descriptive objects
- **Spatial Searching:** Compose image grids from videos (sacrifice spatial accuracy), detect objects
- **Zooming in Temporally and Spatially:** Adapt temporal distribution based on detection scores (temporal upsample and spatial upsample), iteratively sample until all objects are found.
- Use these frames as keyframes for downstream tasks




Sampling weights *gradually align with* ground truth frames over iterations.

*T** allows the model to **zoom in** on distant relevant keyframes **simultaneously** (e.g., ~50s / ~100s for top plot).

## *T** can plug in any VLMs




T* can boost:

in ***extralong***-video(15m-1h):
- LLaVA-OV-72B 56.5→ **62.4%**
- GPT-4o 50.5→**53.1%**

in ***long***-video (2-10min):
- LLaVA-OV-72B 61.6→ **64.1%**
- GPT-4o 57.3→**59.4%**

in ***medium***-video (2-10min):
- LLaVA-OV-72B 77.4→ **79.3%**
- GPT-4o 73.5→**74.3%**

| Model | Frames | NExT-QA 0.7min | EgoSchema 3min |
|---|---|---|---|
| Baselines using Static Uniform Sampling | | | |
| InternVideo [69] | 90 | 49.1 | 32.1 |
| MVU [52] | 16 | 55.2 | 60.3 |
| LLoVi [86] | 90 | 67.7 | 57.6 |
| LangRepo [23] | 180 | 60.9 | 66.2 |
| LLaVA-OneVision-7B [28] | 32 | 79.4 | 65.4 |
| Baselines using Adaptive Frame Selecting | | | |
| SeViLA [83] | 32 | 63.6 | 25.7 |
| VideoAgent [67] | 8.4 | 71.3 | 60.2 |
| LVNet [49] | 12 | 72.9 | 66.0 |
| VideoTree [70] | 63.2 | 73.5 | 66.2 |
| VidF4 [37] | 8 | 74.1 | - |
| Ours: Plug in $T^*$ for Efficient Temporal Search | | | |
| LLaVA-OneVision-7B [28] | 8 | 76.4 | 63.6 |
| + *T** | 8 | **80.4** | **66.6** |

T* can also enhance ***short*** video understanding by **3-4%** on NExT- QA and EgoSchema

| Metric | Pearson Correlation | Pearson p-value | Spearman Correlation | Spearman p-value |
|---|---|---|---|---|
| Temporal $F_1$ | **0.901** | 0.037 | 0.700 | 0.188 |
| Temporal Precision | 0.828 | 0.084 | **0.975** | 0.005 |
| Visual $F_1$ | 0.829 | 0.083 | 0.600 | 0.285 |
| Temporal Recall | 0.655 | 0.231 | 0.700 | 0.188 |
| Visual Recall | 0.568 | 0.317 | 0.500 | 0.391 |
| Visual Precision | 0.327 | 0.591 | 0.100 | 0.873 |

| Method | Searching Efficiency | | | | Overall Task Efficiency | | |
|---|---|---|---|---|---|---|---|
| | Grounding | Matching | TFLOPs ↓ | Latency (sec) ↓ | TFLOPs ↓ | Latency (sec) ↓ | Acc ↑ |
| Baselines: Static Frame Sampling | | | | | | | |
| Uniform-8 [64] | N/A | N/A | N/A | 0.2 | 139.3 | 3.8 | 53.7 |
| Baselines: Adaptive Frame Selection | | | | | | | |
| VideoAgent [60] | GPT4×4 | CLIP-1B×840 | 536.5† | 30.2 | 690.7† | 34.9 | 49.2 |
| Retrieval-based | N/A | YOLO-110M×840 | 216.1 | 28.6 | 355.4 | 32.2 | 57.3 |
| Ours: $T^*$ for Efficient Keyframe Search | | | | | | | |
| Attention-based | LLaVA-72B×3 | N/A | 88.9 | 13.7 | 228.2 | 17.3 | 59.3 |
| Detector-based | LLaVA-7B×1 | YOLO-110M×49 | 33.3 | 7.5 | 172.6 | 11.1 | 59.8 |
| Training-based | LLaVA-7B×1 | YOLO-110M×38 | **30.3** | **6.8** | **169.6** | **10.4** | **60.3** |