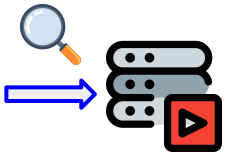
MAGNET: A Multi-agent Framework for Finding Audio-Visual Needles by Reasoning over Multi-Video Haystacks

Sanjoy Chowdhury[1], Mohamed Elmoghany[2], Yohan Abeysinghe[3], Junjie Fei[2], Sayan Nag[4], Salman Khan[3], Mohamed Elhoseiny[2], Dinesh Manocha[1]

[1]University of Maryland, College Park, [2]KAUST, [3]MBZUAI, [4]University of Toronto

Motivation

Considering the various songs and techniques presented—from basic chord strumming to lead riff execution, how can a learner effectively structure their practice sessions to gradually incorporate and master these skills?

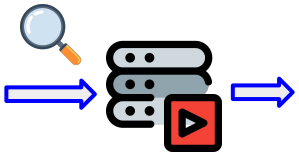


(1) Begin with mastering basic chords (e.g., E, A, D) using slow, deliberate practice and vocal counting. (Vid #1 [897-969s]) (2) Use looping exercises to reinforce muscle memory for specific passages, gradually increasing tempo. (Vid #2 [227-287s]) (3) Integrate new techniques (such as lead riff and advanced strumming) in small segments, ensuring each is perfected before combining them in full song play-alongs. (Vid #3 [570-641s]) (4) Regularly review challenging sections and incorporate practice drills and pauses as recommended in each lesson. (Vid #3 [570-641s], Vid #4 [702-766s, 811-955s])

Motivation

Step 1: Video temporal grounding

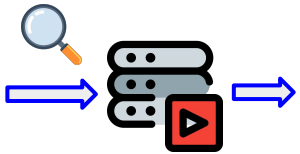
Considering the various songs and techniques presented—from basic chord strumming to lead riff execution, how can a learner effectively structure their practice sessions to gradually incorporate and master these skills?

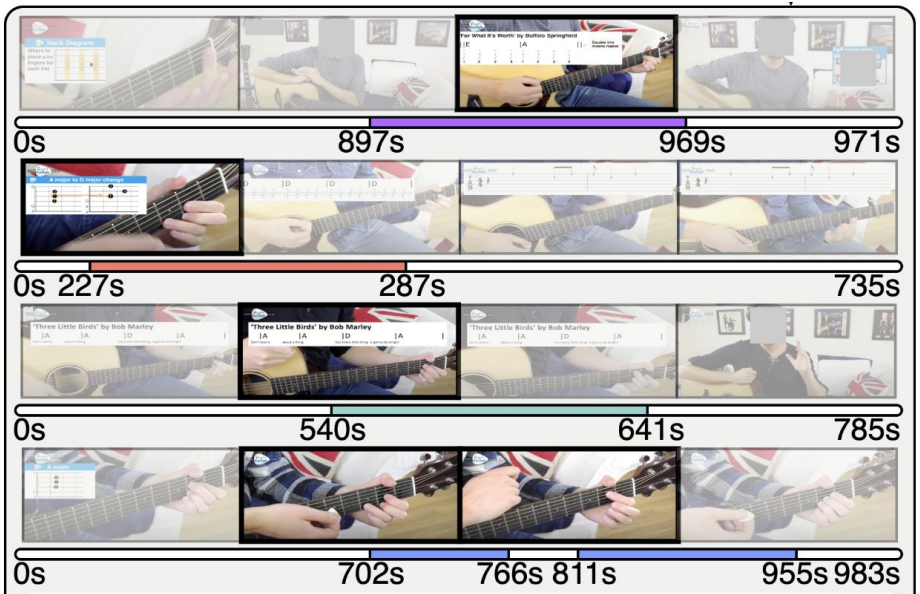


(1) Begin with mastering basic chords (e.g., E, A, D) using slow, deliberate practice and vocal counting. (Vid #1 [897-969s]) (2) Use looping exercises to reinforce muscle memory for specific passages, gradually increasing tempo. (Vid #2 [227-287s]) (3) Integrate new techniques (such as lead riff and advanced strumming) in small segments, ensuring each is perfected before combining them in full song play-alongs. (Vid #3 [570-641s]) (4) Regularly review challenging sections and incorporate practice drills and pauses as recommended in each lesson. (Vid #3 [570-641s], Vid #4 [702-766s, 811-955s])

Motivation

Considering the various songs and techniques presented—from basic chord strumming to lead riff execution, how can a learner effectively structure their practice sessions to gradually incorporate and master these skills?



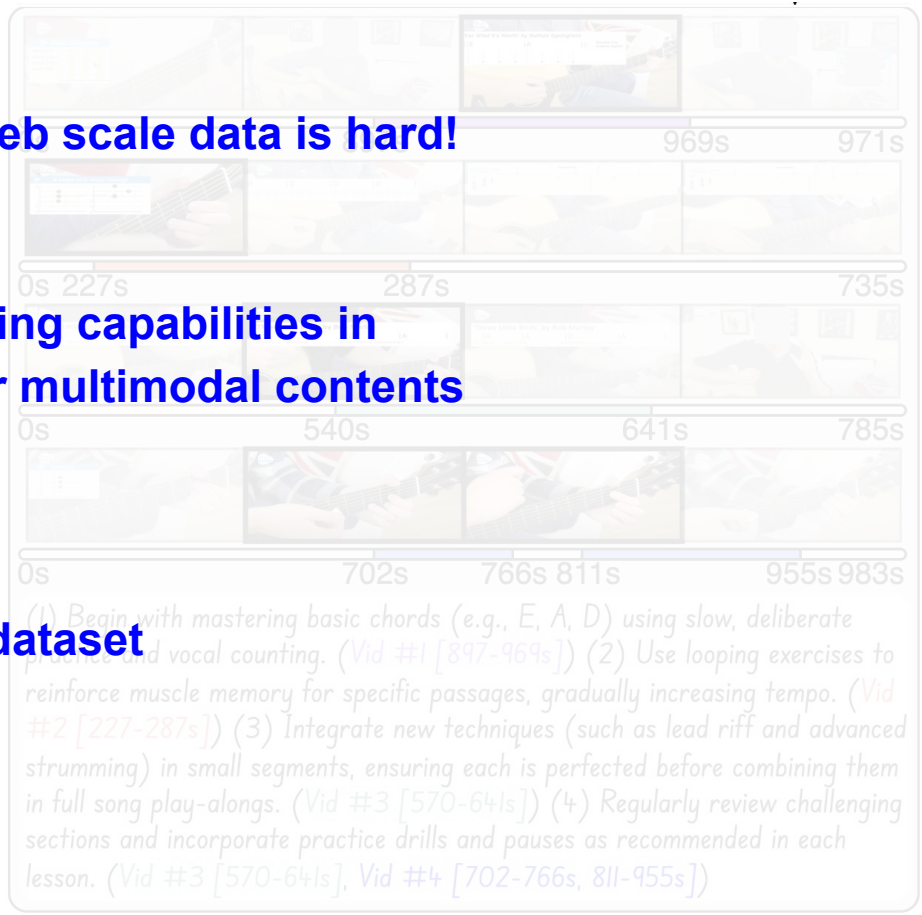


(1) Begin with mastering basic chords (e.g., E, A, D) using slow, deliberate practice and vocal counting. (Vid #1 [897-969s]) (2) Use looping exercises to reinforce muscle memory for specific passages, gradually increasing tempo. (Vid #2 [227-287s]) (3) Integrate new techniques (such as lead riff and advanced strumming) in small segments, ensuring each is perfected before combining them in full song play-alongs. (Vid #3 [570-641s]) (4) Regularly review challenging sections and incorporate practice drills and pauses as recommended in each lesson. (Vid #3 [570-641s], Vid #4 [702-766s, 811-955s])

Step 2: Response aggregation

Motivation

- **Browsing through web scale data is hard!**
- **Audio-Visual reasoning capabilities in existing MLLMs over multimodal contents is not satisfactory**
- **Lack of benchmark dataset**



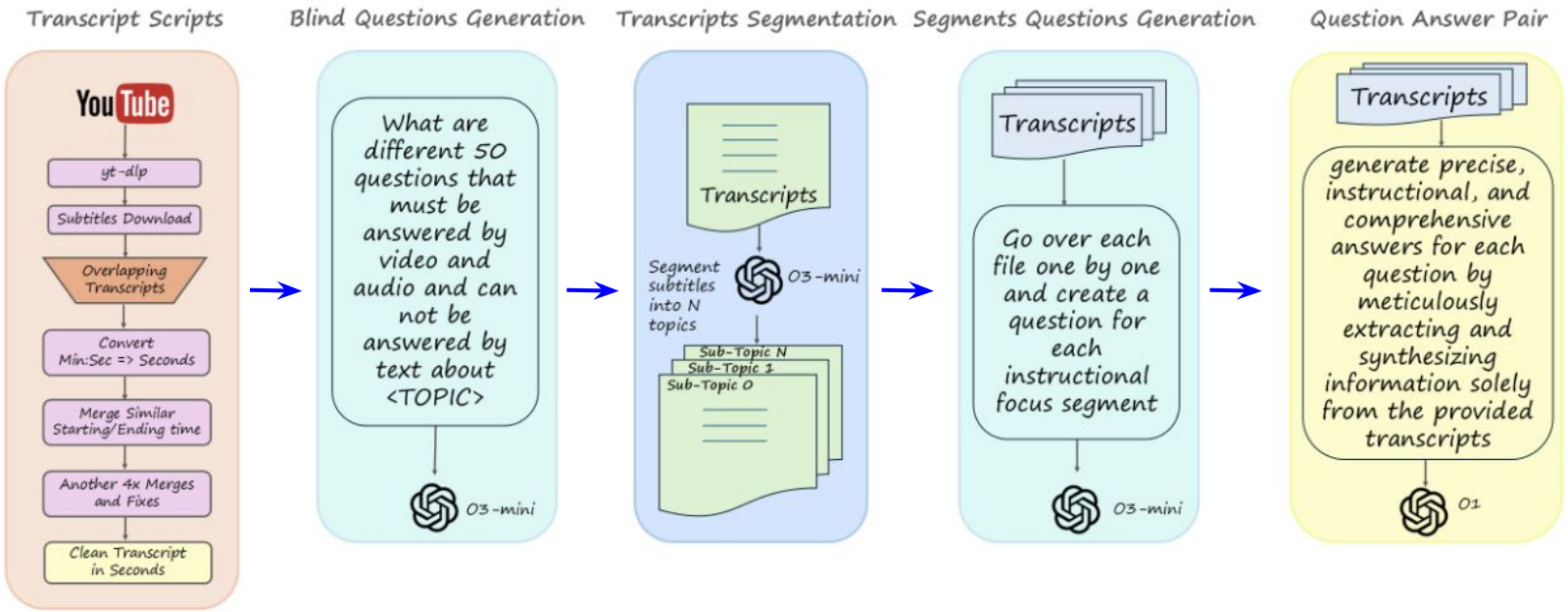
Our Contributions

- A *novel task*, *AVHaystacksQA*, and introduce AVHaystacks, new benchmark consisting of 3100 audio-visual QA pairs

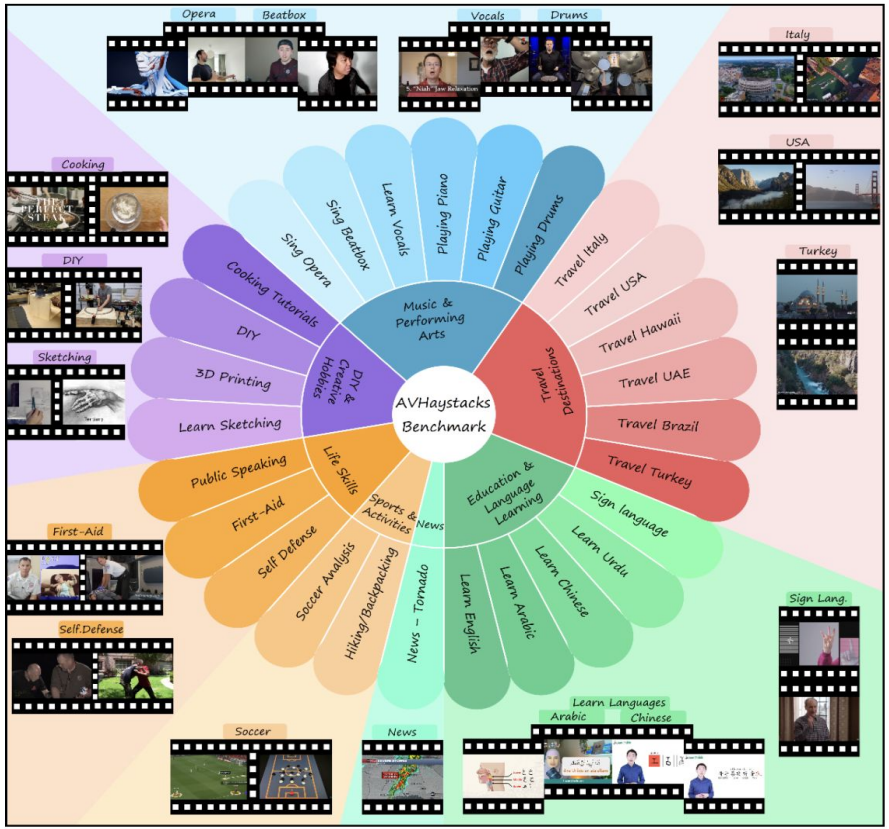
- Two *novel metrics*: STEM and MTGS

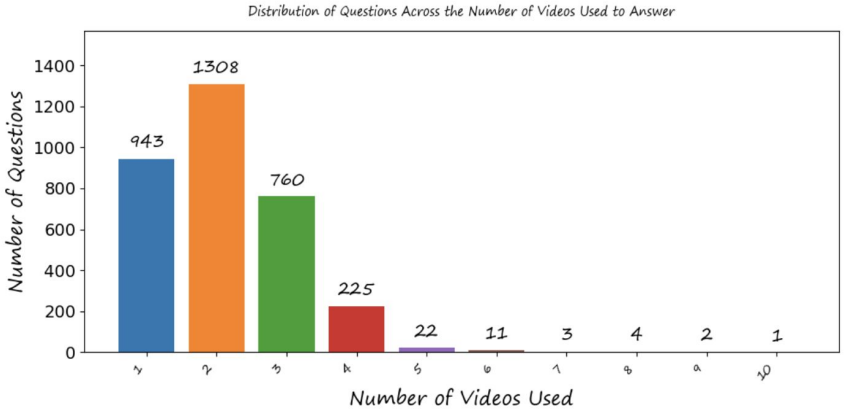
- Propose a *model-agnostic, multi-agent training strategy*, MAGNET

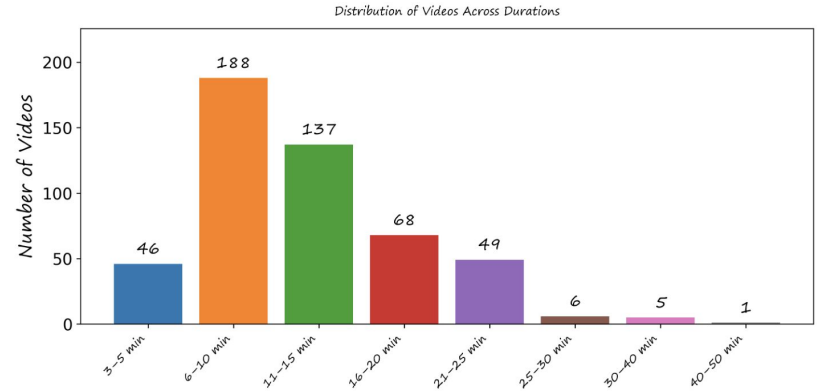
Benchmark Construction



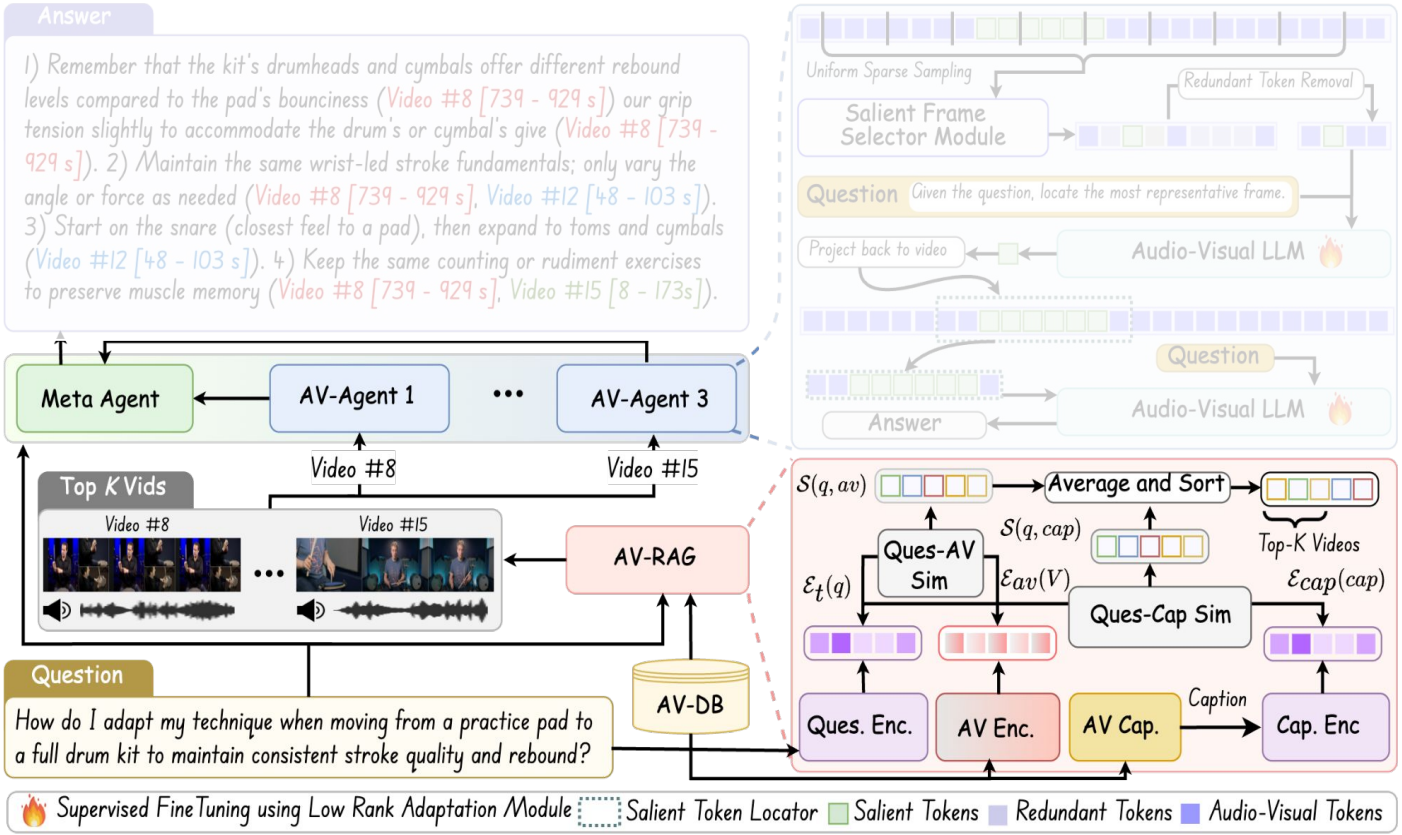
Benchmark Construction



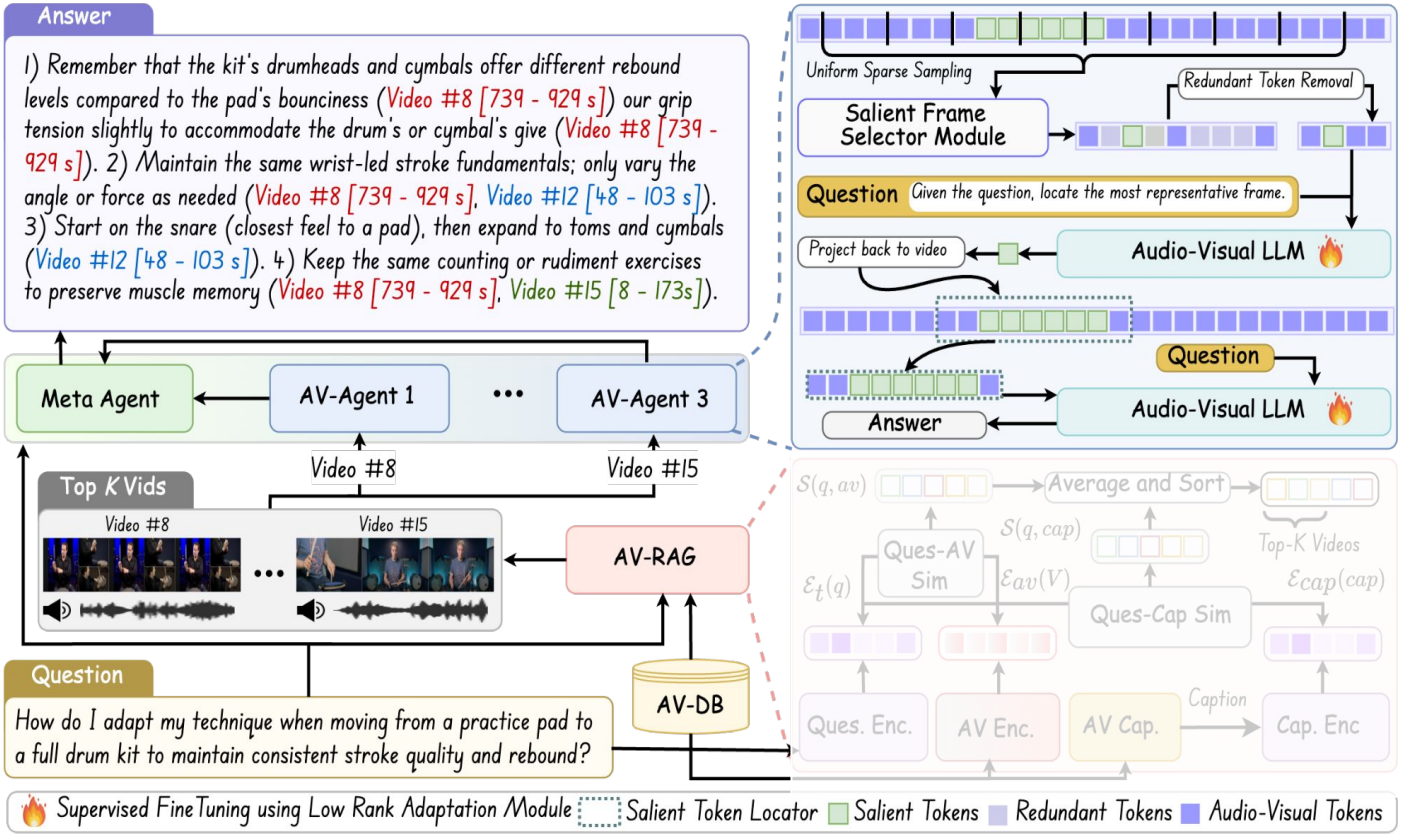




Our contributions



Our contributions



Frame Selection

Algorithm 1 SFS

Input: m total frames, target count k, matrix Q

Output: Selected frame indices

Initialize: $C[0 \dots m][0 \dots k] \leftarrow \infty, C[0][0] \leftarrow 0$
Initialize: $backtrack[0 \dots m][0 \dots k] \leftarrow -1$
for $j \in \{1, \dots, k\}$ **do**
    for $i \in \{j \dots m\}$ **do**
        for $p \in \{j-1 \dots i-1\}$ **do**
            if $C[p][j-1] + Q[p][i] < C[i][j]$ **then**
                $C[i][j] \leftarrow C[p][j-1] + Q[p][i]$
                $backtrack[i][j] \leftarrow p$
Initialize: $result \leftarrow [\,], j \leftarrow k, i \leftarrow m$
while $j > 0$ **do**
    $result$.append(i)
    $i \leftarrow backtrack[i][j], j \leftarrow j - 1$
return $result$.reverse()

Let I_t denote the t-th sampled frame, and $z_t \in \mathbb{R}^d$ its (Hadamard) fused audio-visual embedding from ImageBind. We compute the pairwise cosine similarity between all frame pairs:

$$\Gamma_{ab} = \frac{z_a^\top z_b}{\|z_a\|_2 \cdot \|z_b\|_2}, \quad \forall a, b \in \{1, \dots, m\}$$

To discourage temporally adjacent selections, we apply a temporal separation penalty to frame pairs, where γ is the separation penalty factor:

$$\Delta_{ab} = \gamma \left(\frac{1}{\sin\left(\frac{\pi}{2}|a-b|\right) + 1} - 1 \right)$$

The total affinity matrix is defined as $\mathbf{Q}_{ab} = \Gamma_{ab} + \Delta_{ab}$. We then select a sequence of k frame indices $\mathcal{T} = \{t_1, t_2, \dots, t_k\}$ such that $1 \leq t_1 < \dots < t_k \leq m$ and the total pairwise similarity is minimized (process detailed in Algorithm 1) using the following equation:

$$\mathcal{T} = \arg \min_{\substack{\mathcal{T} \subset \{1, \dots, m\} \\ |\mathcal{T}| = k}} \sum_{i=1}^{k-1} Q_{t_i t_{i+1}}$$

Robust Evaluation

Algorithm 2 STEM: **S**tep-wise **E**rror **M**etric

Input: Ground Truth Steps: $\{G_1, \ldots, G_n\}$, Predicted Steps: $\{P_1, \ldots, P_m\}$, Text Similarity Threshold: $\tau_s = 0.5$.

Output: Missing Step: S_M, Hallucinated Step: S_H, Wrong Step Order: S_O, Step wise Video ID False Positives and Negatives: S_{FP}, S_{FN}, Step-wise IoU on time intervals: S_{IoU}, Similarity Matrix: M_{sim}, Step Similarity Function: $\texttt{Sim}(\cdot)$, Hungarian Matching Algorithm: $\texttt{Hung}(\cdot)$, Matched Steps: $\hat{GT}, \hat{P}$

$M_{\text{sim}} \leftarrow \texttt{Sim}(G_i^{\text{text}}, P_j^{\text{text}})$ ▷ Compute similarity matrix

$\hat{G}, \hat{P} \leftarrow \texttt{Hung}(M_{\text{sim}}, \tau_s, G, P)$ ▷ Obtain matched pairs

for matched pairs $(\hat{G}_i, \hat{P}_j)$ **do**

  if $i \neq j$ **then**

    $S_O \leftarrow S_O + 1$ ▷ Wrong Step Order

  for groundings $(v_{\text{pred}}, t_{\text{start}}^{\text{pred}}, t_{\text{end}}^{\text{pred}})$ in P_j **do**

    if $v_{\text{pred}} \notin \{v_{\text{gt}} \in G_i\}$ **then**

      $S_{FP} \leftarrow S_{FP} + 1$ ▷ Video ID Mismatch

    else

      $S_{\text{IoU}} \leftarrow \text{IoU}\left([t_{\text{start}}^{\text{gt}}, t_{\text{end}}^{\text{gt}}], [t_{\text{start}}^{\text{pred}}, t_{\text{end}}^{\text{pred}}]\right)$

  for groundings $(v_{\text{gt}}, t_{\text{start}}^{\text{gt}}, t_{\text{end}}^{\text{gt}})$ in G_i **do**

    if $v_{\text{gt}} \notin \{v_{\text{pred}} \in P_j\}$ **then**

      $S_{FN} \leftarrow S_{FN} + 1$ ▷ Video ID Mismatch

for unmatched $(G - \hat{G})_i$ **do**

  $S_M \leftarrow S_M + 1$ ▷ Missing Step

for unmatched $(P - \hat{P})_j$ **do**

  $S_H \leftarrow S_H + 1$ ▷ Hallucinated Step

Quantitative Results

Method	AVHaystacks-50					AVHaystacks-Full				
	B@4 ↑	Cr ↑	Text Sim ↑	GPT Eval ↑	H Eval ↑	B@4 ↑	Cr ↑	Text Sim ↑	GPT Eval ↑	H Eval ↑
VideoRAG	43.16	119.78	5.31	6.32	3.42	41.59	115.97	5.15	6.13	3.32
Video-RAG	42.64	117.86	5.23	6.20	3.37	40.67	112.12	4.99	5.97	3.23
Qwen2.5 omni	10.84	28.59	1.90	2.11	1.07	-	-	-	-	-
Unified IO2	11.64	34.28	2.15	2.40	1.02	-	-	-	-	-
VideoSALMONN	11.90	32.32	2.07	2.39	0.91	-	-	-	-	-
MAGNET +VideoSALMONN-ZS	29.11	83.60	3.93	4.66	2.59	27.37	76.19	3.69	4.30	2.45
MAGNET +Unified IO2-ZS	28.78	81.79	3.85	4.52	2.54	27.95	76.1	3.69	4.35	2.45
MAGNET +Qwen 2.5 Omni -ZS	30.54	85.56	4.01	4.73	2.64	28.49	81.74	3.85	4.57	2.54
MAGNET +VideoSALMONN-FT	52.30	144.40	6.20	7.46	3.96	49.24	136.86	5.96	7.19	3.81
MAGNET +Unified IO2-FT	53.66	146.38	6.28	7.58	4.00	51.45	142.56	6.12	7.34	3.91
MAGNET +Qwen 2.5 Omni-FT	**55.82**	**153.98**	**6.53**	**7.84**	**4.15**	**53.69**	**146.30**	**6.28**	**7.56**	**4.01**
MAGNET +Gemini 1.5 Pro	57.67	157.72	6.69	8.03	4.25	55.80	153.95	6.53	7.80	4.15

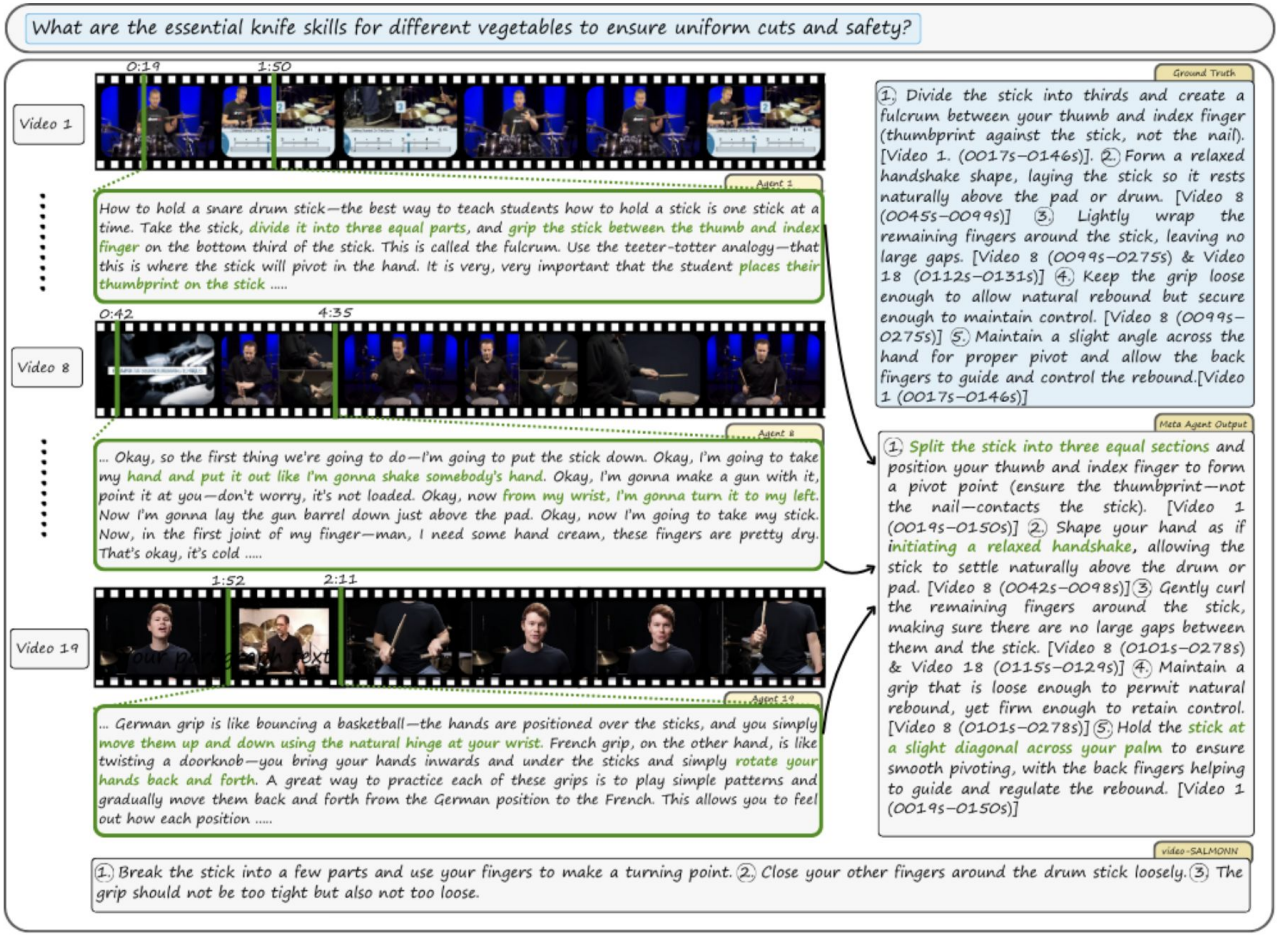
Response alignment scores

Quantitative Results

Method	AVHaystacks-50						AVHaystacks-Full					
	$MTGS_{avg}$ ↑	SM ↓	SH ↓	SO ↓	SFP ↓	SFN ↓	$MTGS_{avg}$ ↑	SM ↓	SH ↓	SO ↓	SFP ↓	SFN ↓
MAGNET +VideoSALMONN-ZS	0.48	0.35	0.34	0.35	0.31	0.25	0.45	0.41	0.33	0.43	0.36	0.33
MAGNET +Unified IO2-ZS	0.51	0.39	0.31	0.31	0.32	0.22	0.42	0.49	0.39	0.37	0.37	0.29
MAGNET +Qwen 2.5 Omni -ZS	0.54	0.37	0.28	0.32	0.28	0.21	0.49	0.43	0.34	0.39	0.33	0.27
MAGNET +VideoSALMONN-FT	0.81	0.12	0.16	0.19	0.18	0.11	0.75	0.13	0.18	0.23	0.19	0.14
MAGNET +Unified IO2-FT	0.79	0.14	0.16	0.17	0.18	0.14	0.72	0.15	0.18	0.20	0.21	0.18
MAGNET +Qwen 2.5 Omni-FT	**0.83**	**0.11**	**0.13**	**0.14**	**0.15**	**0.09**	**0.79**	**0.13**	**0.16**	**0.19**	**0.19**	**0.12**
MAGNET +Gemini 1.5 Pro	0.85	0.09	0.12	0.14	0.10	0.07	0.81	0.12	0.14	0.17	0.12	0.09

Grounding evaluation and Step-wise error results

Qualitative Results



Acknowledgement

Project Page

Thank you!