



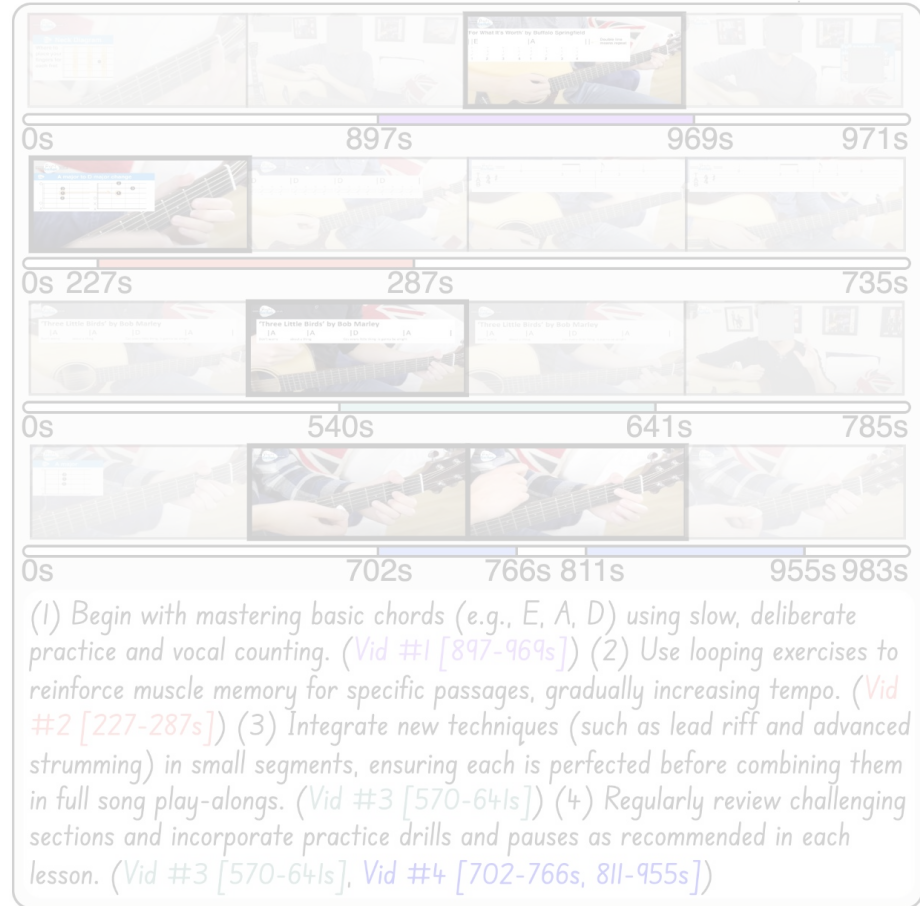
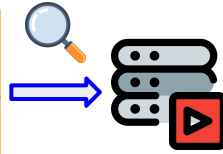
MAGNET: A Multi-agent Framework for Finding Audio-Visual Needles by Reasoning over Multi-Video Haystacks

Sanjoy Chowdhury¹, Mohamed Elmoghany², Yohan Abeyasinghe³, Junjie Fei², Soyan Nag⁴, Salman Khan³,
Mohamed Elhoseiny², Dinesh Manocha¹

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

Motivation

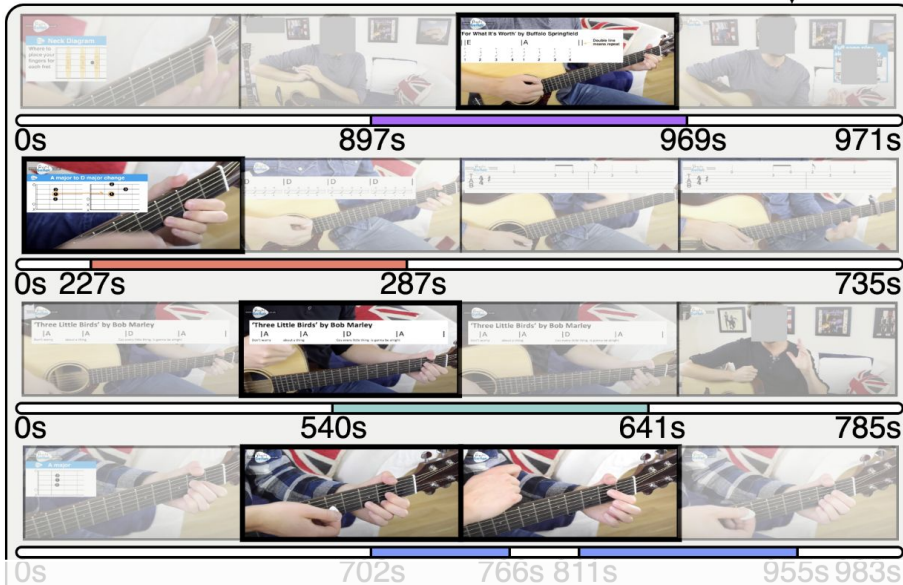
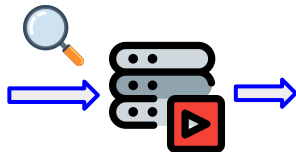
Considering the various songs and techniques presented—from basic chord strumming to lead riff execution, how can a learner effectively structure their practice sessions to gradually incorporate and master these skills?



Motivation

Step 1: Video temporal grounding

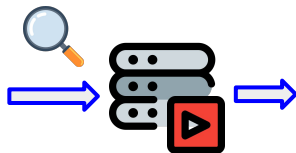
Considering the various songs and techniques presented—from basic chord strumming to lead riff execution, how can a learner effectively structure their practice sessions to gradually incorporate and master these skills?



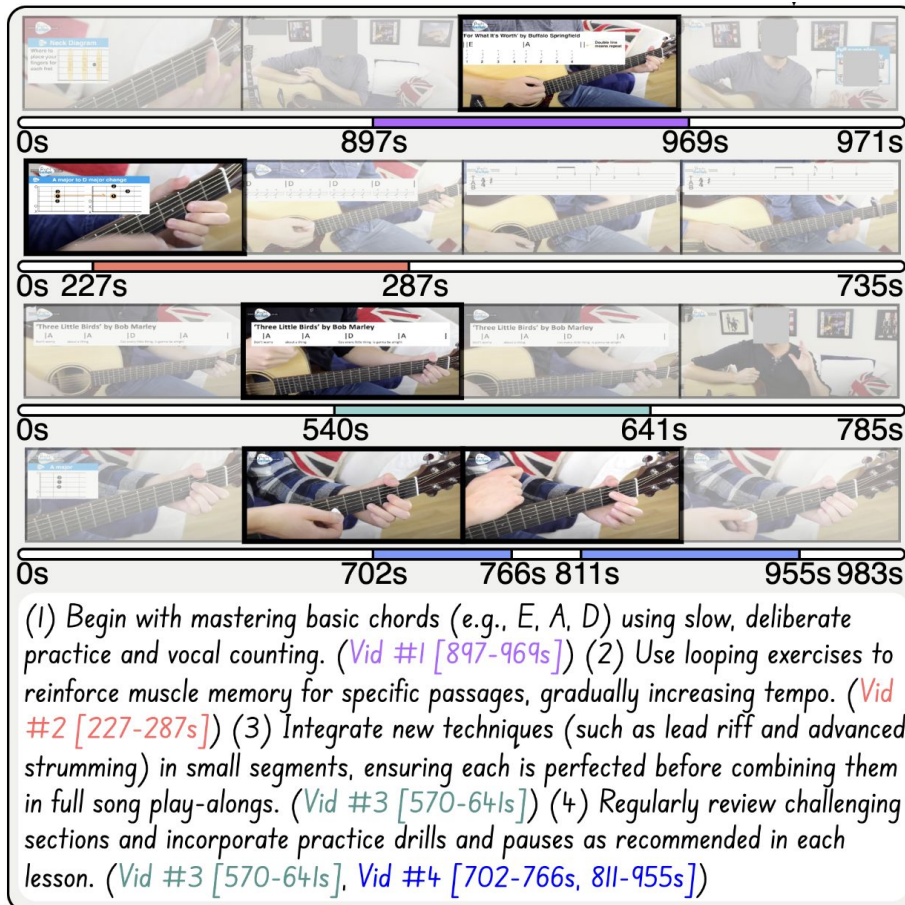
(1) Begin with mastering basic chords (e.g., E, A, D) using slow, deliberate practice and vocal counting. (Vid #1 [897-969s]) (2) Use looping exercises to reinforce muscle memory for specific passages, gradually increasing tempo. (Vid #2 [227-287s]) (3) Integrate new techniques (such as lead riff and 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?



Step 2: Response aggregation



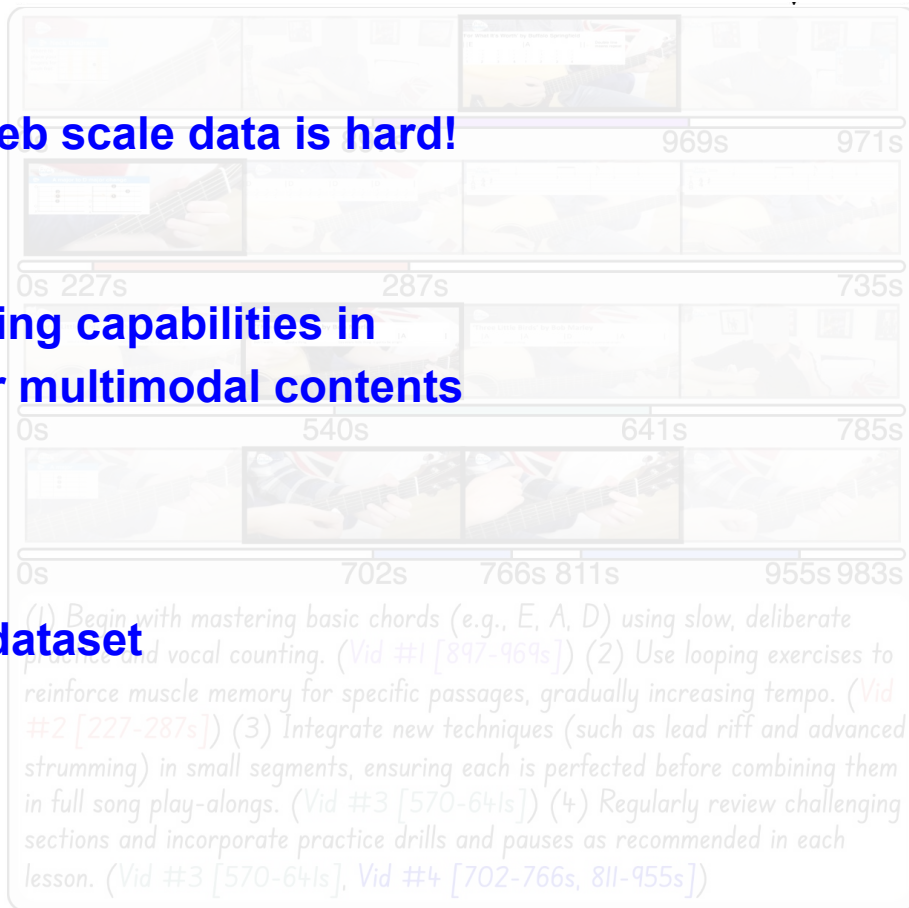
Motivation

- Browsing through web scale data is hard!

- Audio-Visual reasoning capabilities in existing MLLMs over multimodal contents is not satisfactory

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?

- Lack of benchmark dataset



Our Contributions

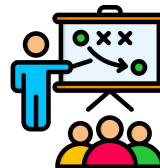
- ✓ A novel task, *AVHaystacksQA*, and introduce AVHaystacks, new benchmark consisting of 3100 audio-visual QA pairs



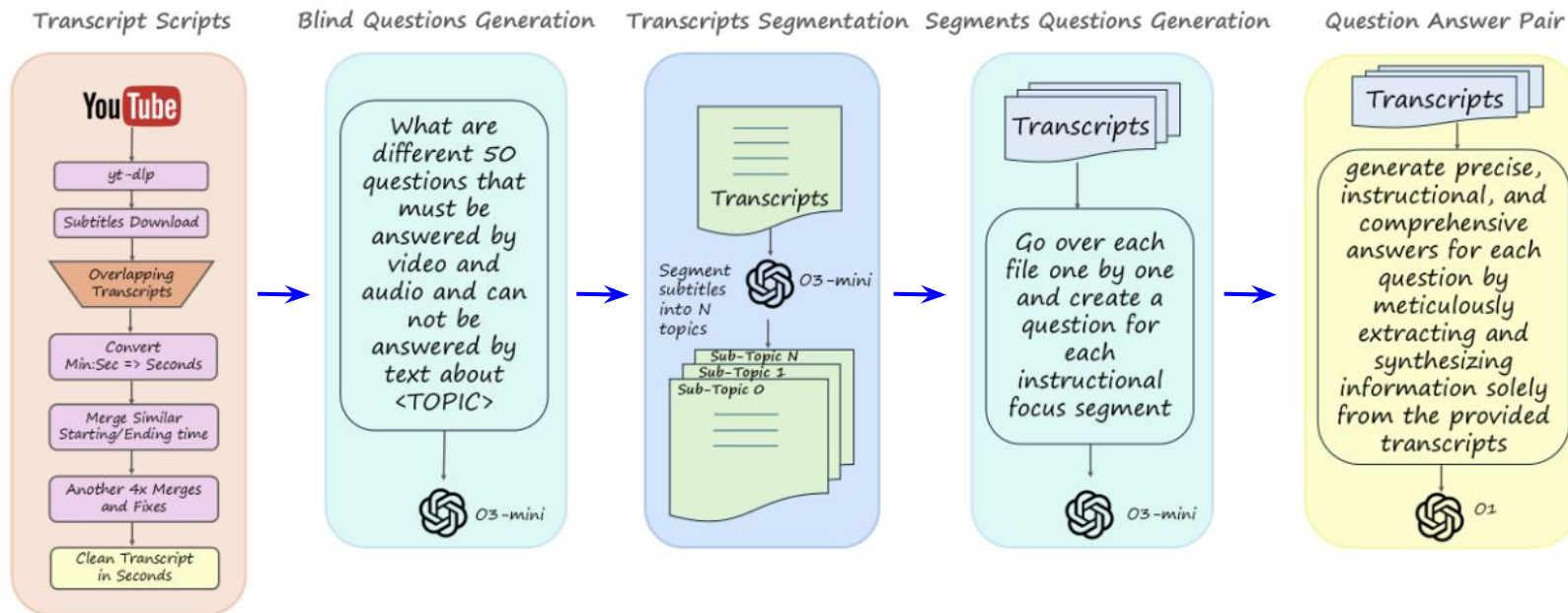
- ✓ Two *novel metrics*: STEM and MTGS



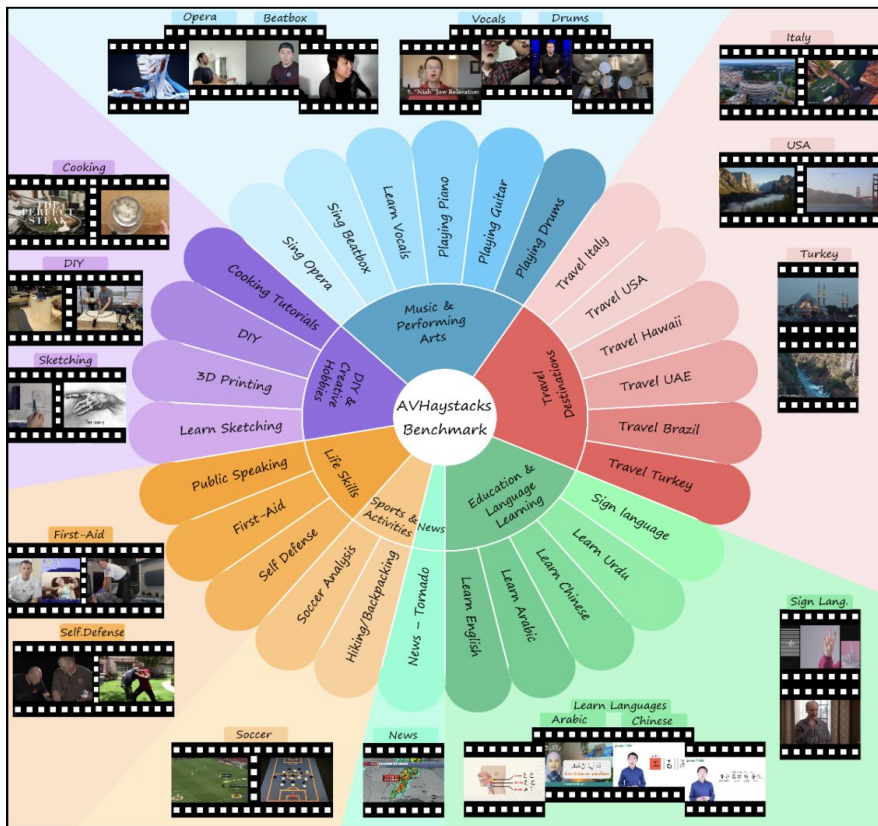
- ✓ Propose a *model-agnostic, multi-agent training strategy*, **MAGNET**



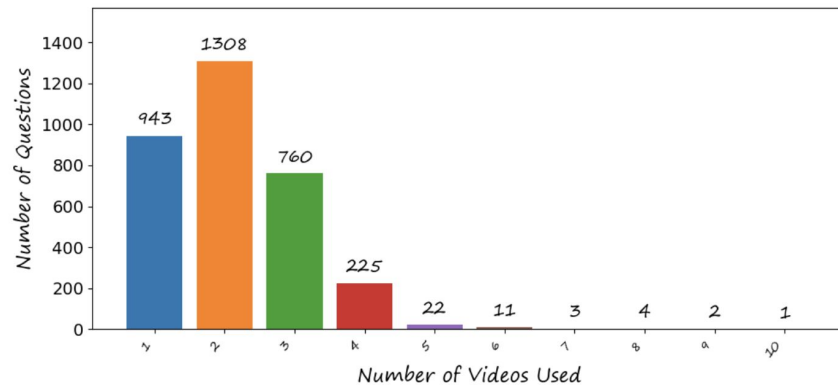
Benchmark Construction



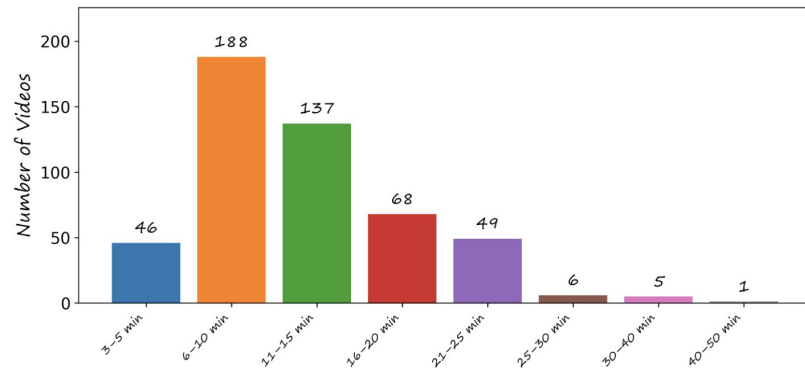
Benchmark Construction



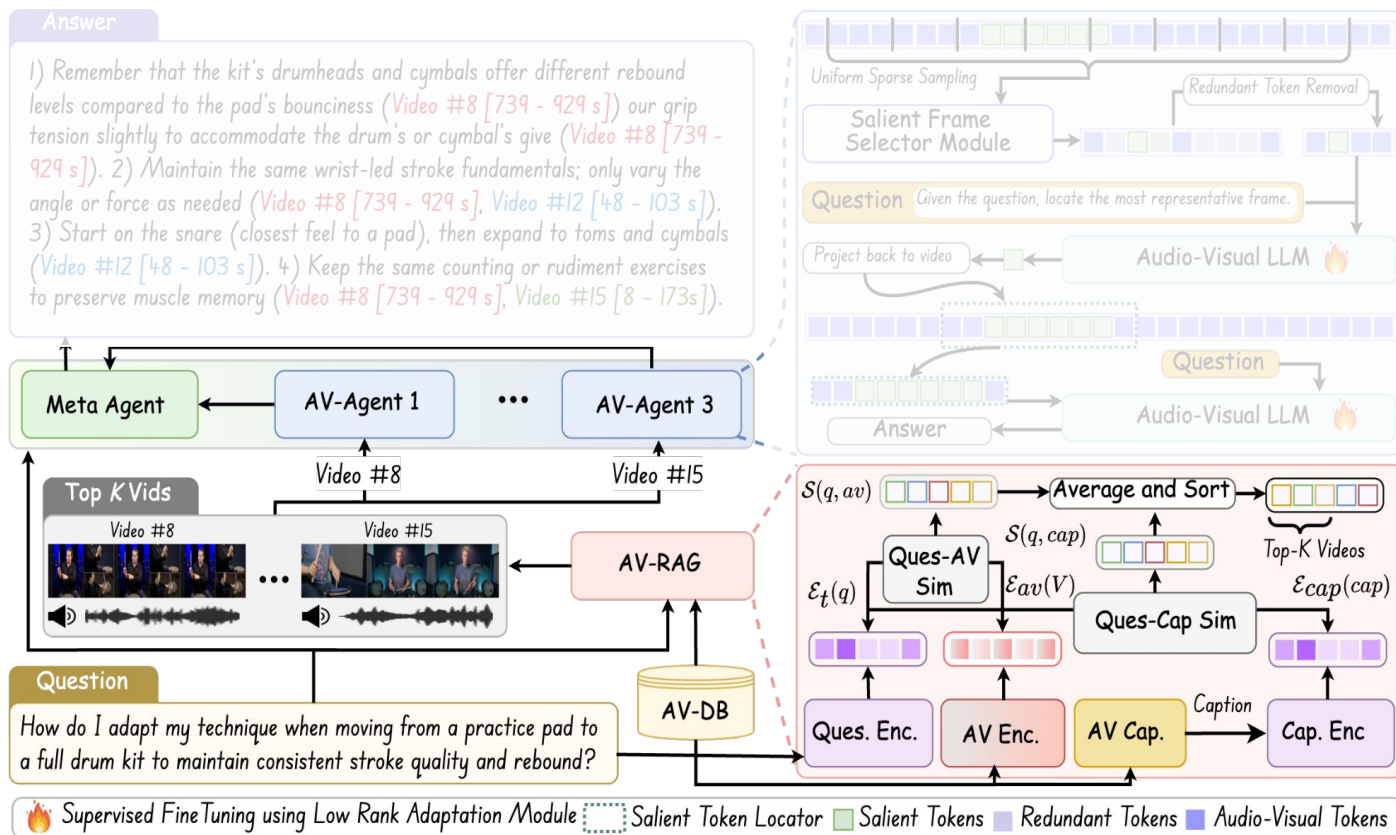
Distribution of Questions Across the Number of Videos Used to Answer



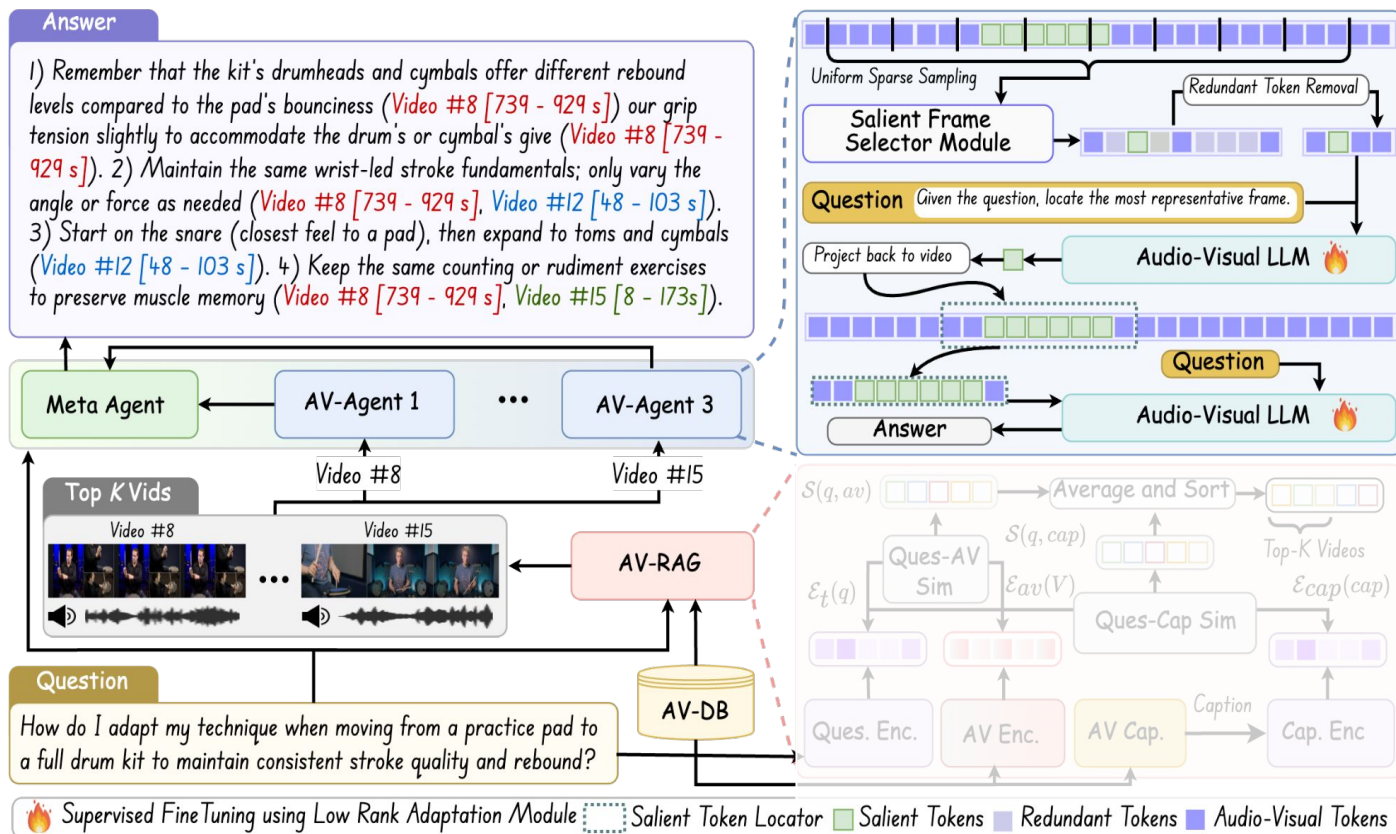
Distribution of Videos Across Durations



Our contributions



Our contributions



Frame Selection

Algorithm 1 SFS

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

Output: Selected frame indices

```

1: Initialize:  $C[0 \dots m][0 \dots k] \leftarrow \infty, C[0][0] \leftarrow 0$ 
2: Initialize:  $backtrack[0 \dots m][0 \dots k] \leftarrow -1$ 
3: for  $j \in \{1, \dots, k\}$  do
4:   for  $i \in \{j \dots m\}$  do
5:     for  $p \in \{j-1 \dots i-1\}$  do
6:       if  $C[p][j] - 1 + Q[p][i] < C[i][j]$  then
7:          $C[i][j] \leftarrow C[p][j] - 1 + Q[p][i]$ 
8:          $backtrack[i][j] \leftarrow p$ 
9: Initialize:  $result \leftarrow [], j \leftarrow k, i \leftarrow m$ 
10: while  $j > 0$  do
11:    $result.append(i)$ 
12:    $i \leftarrow backtrack[i][j], j \leftarrow j - 1$ 
13: 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:

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

Robust Evaluation

Algorithm 2 STEM: Step-wise Error Metric

Input: Ground Truth Steps: $\{G_1, \dots, G_n\}$, Predicted Steps: $\{P_1, \dots, P_m\}$, 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: $\text{Sim}(\cdot)$, Hungarian Matching Algorithm: $\text{Hung}(\cdot)$, Matched Steps: $\hat{G}T, \hat{P}$

```
1:  $M_{sim} \leftarrow \text{Sim}(G_i^{\text{text}}, P_j^{\text{text}})$   $\triangleright$  Compute similarity matrix
2:  $\hat{G}, \hat{P} \leftarrow \text{Hung}(M_{sim}, \tau_s, G, P)$   $\triangleright$  Obtain matched pairs
3: for matched pairs  $(\hat{G}_i, \hat{P}_j)$  do
4:   if  $i \neq j$  then
5:      $S_O \leftarrow S_O + 1$   $\triangleright$  Wrong Step Order
6:   for groundings  $(v_{\text{pred}}, t_{\text{start}}^{\text{pred}}, t_{\text{end}}^{\text{pred}})$  in  $P_j$  do
7:     if  $v_{\text{pred}} \notin \{v_{\text{gt}} \in G_i\}$  then
8:        $S_{FP} \leftarrow S_{FP} + 1$   $\triangleright$  Video ID Mismatch
9:     else
10:       $S_{IoU} \leftarrow \text{IoU}([t_{\text{start}}^{\text{gt}}, t_{\text{end}}^{\text{gt}}], [t_{\text{start}}^{\text{pred}}, t_{\text{end}}^{\text{pred}}])$ 
11:    for groundings  $(v_{\text{gt}}, t_{\text{start}}^{\text{gt}}, t_{\text{end}}^{\text{gt}})$  in  $G_i$  do
12:      if  $v_{\text{gt}} \notin \{v_{\text{pred}} \in P_j\}$  then
13:         $S_{FN} \leftarrow S_{FN} + 1$   $\triangleright$  Video ID Mismatch
14:  for unmatched  $(G - \hat{G})_i$  do
15:     $S_M \leftarrow S_M + 1$   $\triangleright$  Missing Step
16:  for unmatched  $(P - \hat{P})_j$  do
17:     $S_H \leftarrow S_H + 1$   $\triangleright$  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

What are the essential knife skills for different vegetables to ensure uniform cuts and safety?

Video 1

0:19

1:50




Agent 1

How to hold a snare drum stick—the best way to teach students how to hold a stick is one stick at a time. Take the stick, **divide it into three equal parts**, and **grip the stick between the thumb and index finger** on the bottom third of the stick. This is called the fulcrum. Use the teeter-totter analogy—that this is where the stick will pivot in the hand. It is very, very important that the student **places their thumbprint on the stick**

Video 8

0:42

4:35




Agent 8

... Okay, so the first thing we're going to do—I'm going to put the stick down. Okay, I'm going to take my hand and **put it out like I'm gonna shake somebody's hand**. Okay, I'm gonna make a gun with it, point it at you—don't worry, it's not loaded. Okay, now **from my wrist, I'm gonna turn it to my left**. Now I'm gonna lay the gun barrel down just above the pad. Okay, now I'm going to take my stick. Now, in the first joint of my finger—man, I need some hand cream, these fingers are pretty dry. That's okay, it's cold

Video 19

1:52

2:11



Agent 19

... German grip is like bouncing a basketball—the hands are positioned over the sticks, and you simply **move them up and down using the natural hinge at your wrist**. French grip, on the other hand, is like twisting a doorknob—you bring your hands inwards and under the sticks and simply **rotate your hands back and forth**. A great way to practice each of these grips is to play simple patterns and gradually move them back and forth from the German position to the French. This allows you to feel out how each position

Ground Truth

①. Divide the stick into thirds and create a fulcrum between your thumb and index finger (thumbprint against the stick, not the nail). [Video 1. (0017s–0146s)]. ②. Form a relaxed handshake shape, laying the stick so it rests naturally above the pad or drum. [Video 8 (0045s–0099s)] ③. Lightly wrap the remaining fingers around the stick, leaving no large gaps. [Video 8 (0099s–0275s) & Video 18 (0112s–0131s)] ④. Keep the grip loose enough to allow natural rebound but secure enough to maintain control. [Video 8 (0099s–0275s)] ⑤. Maintain a slight angle across the hand for proper pivot and allow the back fingers to guide and control the rebound. [Video 1 (0017s–0146s)]

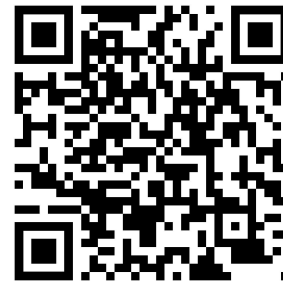
Meta Agent Output

①. **Split the stick into three equal sections** and position your thumb and index finger to form a pivot point (ensure the thumbprint—not the nail—contacts the stick). [Video 1 (0019s–0150s)] ②. Shape your hand as if **initiating a relaxed handshake**, allowing the stick to settle naturally above the drum or pad. [Video 8 (0042s–0098s)] ③. Gently curl the remaining fingers around the stick, making sure there are no large gaps between them and the stick. [Video 8 (0101s–0278s) & Video 18 (0115s–0129s)] ④. Maintain a grip that is loose enough to permit natural rebound, yet firm enough to retain control. [Video 8 (0101s–0278s)] ⑤. **Hold the stick at a slight diagonal across your palm** to ensure smooth pivoting, with the back fingers helping to guide and regulate the rebound. [Video 1 (0019s–0150s)]

video-SALMONN

①. Break the stick into a few parts and use your fingers to make a turning point. ②. Close your other fingers around the drum stick loosely. ③. The grip should not be too tight but also not too loose.

Acknowledgement



[Project Page](#)

Thank you!

