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
{sanjoyc, dmanocha}@umd.edu sayan.nag@mail.utoronto.ca mohamed.elhoseiny@kaust.edu.sa
{yohan.abeysinghe, salman.khan}@mbzuai.ac.ae m.osama.elmoghany@gmail.com
https://schowdhury671.github.io/magnet_project/

Abstract

Large multimodal models (LMMs) have shown remarkable progress in audio-visual understanding, yet they struggle with real-world scenarios that require complex reasoning across extensive video collections. Existing benchmarks for video question answering remain limited in scope, typically involving one clip per query, which falls short of representing the challenges of large-scale, audio-visual retrieval and reasoning encountered in practical applications. To bridge this gap, we introduce a **novel task** named **AVHaystacksQA**, where the goal is to identify salient segments across different videos in response to a query and link them together to generate the most informative answer. To this end, we present **AVHaystacks**, an audio-visual benchmark comprising 3100 annotated QA pairs designed to assess the capabilities of LMMs in multi-video retrieval and temporal grounding task. Additionally, we propose a model-agnostic, multi-agent framework **MAGNET** to address this challenge, achieving up to 89% and 65% relative improvements over baseline methods on BLEU@4 and GPT evaluation scores in QA task on our proposed AVHaystacks. To enable robust evaluation of multi-video retrieval and temporal grounding for optimal response generation, we introduce two new metrics, **STEM**, which captures alignment errors between a ground truth and a predicted step sequence and **MTGS**, to facilitate balanced and interpretable evaluation of segment-level grounding performance.

1 Introduction

Large Multimodal Models (LMMs) [1–6] have achieved remarkable progress in audio-visual understanding. However, they continue to face significant challenges [7] when it comes to retrieving and reasoning over large-scale multimedia collections particularly in tasks such as audio-visual retrieval-augmented generation (RAG) and multi-video temporal grounding. These limitations hinder their effectiveness in real-world applications, such as querying personal video archives, how-to repositories, or educational video libraries, where complex queries often require joint processing of both audio and visual modalities across multiple video segments.

Despite their growing capabilities, to the best of our knowledge, no existing task or benchmark systematically evaluates LMMs' ability to identify and integrate salient segments from multiple videos to construct informative, grounded responses. As a result, there remains a gap in properly assessing their performance on large-scale audio-visual retrieval and reasoning tasks. Existing benchmarks [8–11] are generally limited in scope typically associating each question with only a single short clip, as shown in Tab. 1. However, real-world information-seeking scenarios often

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

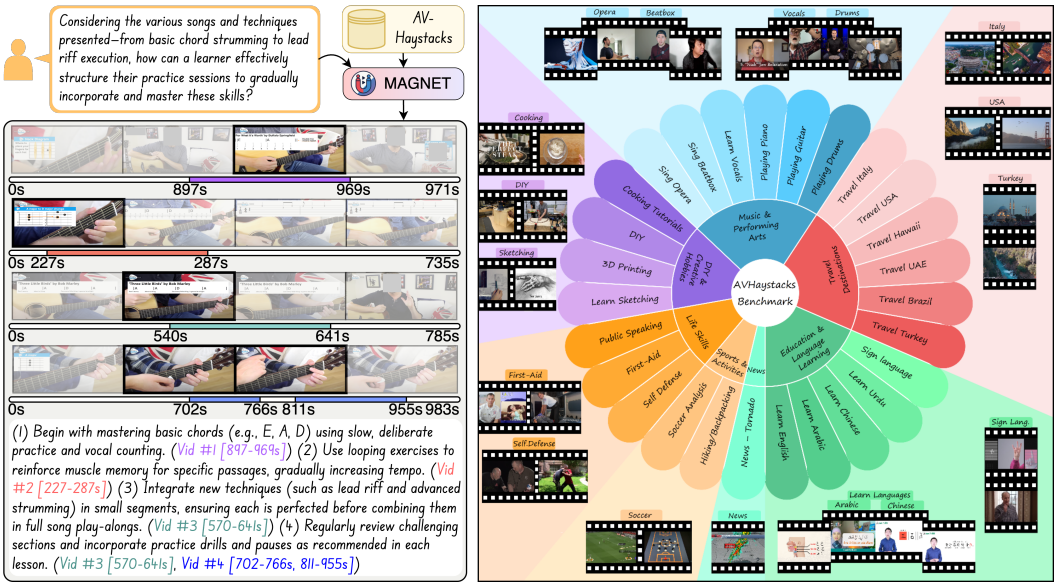


Figure 1: **A new task and a benchmark**. We introduce a novel task AVHaystacksQA, which requires multi-video linkage and reasoning to generate the most informative answer. To this end, we curate a new benchmark AVHaystacks comprising 3100 QA pairs obtained through careful inspection.

demand searching through hundreds or thousands of video segments, identifying the most relevant audio-visual snippets, and synthesizing coherent, evidence-backed answers from them.

To address this gap, we introduce a new benchmark **AVHaystacks** to facilitate novel task **AVHaystack-sQA**. This task is designed to comprehensively evaluate the LMMs ability to perform large-scale, multi-video audio-visual retrieval and reasoning. Each query in our benchmark is grounded in a massive collection of up to **500** video clips, requiring models to localise relevant segments both temporally and across sources and reason over their combined audio and visual signals. This setup reflects the complexity of real-world information needs more accurately than previous single-video setups.

A central challenge in building such a benchmark is designing specific, unambiguous questions that genuinely require both audio and visual understanding and cannot be accurately answered using a single modality or video clip alone. To address this, we implement a robust data filtering pipeline that combines the strengths of large language models (LLMs) and human annotators to remove generic, redundant, or overly broad queries. For instance, questions such as *"How do I improve my strumming while playing guitar?"* or *"How do you adjust for tightness in your lips to hinder a clear operatic tone?"* are carefully selected and validated to ensure they demand cross-modal reasoning and temporal grounding within specific video segments.

To enable LMMs to tackle this challenging task, we introduce **MAGNET** a novel retrieval-augmented, multi-agent framework designed to find audio-visual needles within multi-video haystacks. Our approach integrates multiple specialized audio and video encoders to capture rich multi-modal semantics, and incorporates a multi-agent framework that scores the relevance of retrieved video segments with respect to the input query. By tightly coupling retrieval with reasoning, MAGNET facilitates efficient and scalable exploration of large-scale video corpora. Extensive experiments demonstrate that our framework significantly enhances both retrieval accuracy and answer generation performance.

Main Contributions:

(1) We **propose a novel task**, AVHaystacksQA, and introduce AVHaystacks a new benchmark consisting of 3100 audio-visual QA pairs drawn from videos across diverse domains (Fig. 1). This benchmark pushes the boundaries of video retrieval and reasoning by requiring models to navigate and reason over large-scale video collections. To the best of our knowledge, no existing benchmark systematically evaluates multi-video keypoint detection and reasoning capabilities.

Dataset	Train	Test	MS	TA	MVL	AVR	AVD	RQA	AVQA	LC	Avg. Dur. (s)
Video Datasets											
ShareGPT4Video [12]	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	26
Cinepile [13]	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	160
NExT-QA [14]	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	48
Video-MME [9]	✗	✓	✓	✗	✗	✗	✗	✗	✗	✓	1020
LongVideoBench [10]	✗	✓	✓	✓	✗	✗	✗	✗	✗	✓	480
MovieChat [15]	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	420
Audio-Visual Datasets											
UnAV-100 [16]	✓	✓	✗	✓	✗	✗	✓	✗	✓	✗	42
VAST-27M [17]	✓	✓	✓	✗	✗	✗	✓	✗	✓	✗	20
AVQA [18]	✓	✓	✗	✓	✗	✗	✗	✗	✓	✗	60
AVInstruct [19]	✓	✗	✓	✓	✗	✗	✗	✗	✓	✗	115
Music-AVQA [20]	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	10
VGGSound [21]	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	10
AVBench+SAVEnVid [22]	✓	✓	✓	✓	✗	✗	✓	✗	✓	✗	182
AVHaystacks (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**738**

Table 1: **Comparison with prior video/audio-visual benchmarks.** MS: Model-Assisted; TA: Temporal Annotation; MVL: Multi-Video Linkage; AVR: Audio-Visual fine-grained Reasoning; AVD: Audio-Visual Description; RQA: Retrieval-based QA Answering; AVQA: Audio-Visual QA; LC: Long Context, where QA context spans over 5 mins.

(2) We **conduct extensive evaluations** of state-of-the-art audio-visual models on AVHaystacks, analyzing their performance across a range of multi-video retrieval and reasoning setups. Our findings reveal that current models perform suboptimally in retrieving relevant videos from large corpora and struggle to reason effectively across multiple clips to identify the key segments needed to answer complex queries.

(3) To enable robust evaluation of audio-visual retrieval and grounded temporal reasoning, we **introduce two novel metrics**: **STEM**, which quantifies alignment errors between the ground-truth and predicted step sequences in multi-video audio-visual answer generation; and **MTGS**, which provides a balanced and interpretable assessment of segment-level grounding performance.

(4) Finally, we propose a **model-agnostic, multi-agent training strategy**, MAGNET, designed to enhance model performance in identifying key segments across multi-video haystacks. Experimental results show that our framework achieves up to 89% and 65% relative improvements over baseline methods on BLEU@4 and GPT-based evaluation scores, respectively, on AVHaystacksQA.

2 AVHaystacks: Audio-Visual Benchmark for Multi-Video Temporal Grounding and Reasoning

The benchmark curation pipeline consists of five stages: (1) *Video curation:* Following careful manual inspection, we collect 500 videos from YouTube spanning 27 diverse categories, including how-to guides, cooking, travel, musical instrument tutorials, language instruction, and vocal lessons. Each video is selected to ensure its suitability for audio-visual QA tasks specifically, queries where answering correctly requires a strong understanding of both audio and visual modalities, with complementary cues essential for accurate reasoning. (2) *Blind question generation:* Using OpenAI O3-MINI with custom prompts (details in the supplementary), we generate 50 topic-agnostic questions per topic. This promotes comprehensive evaluation by introducing diversity in reasoning types, modality dependence, and task complexity. (3) *Transcript cleaning and segmentation:* Subtitle overlaps are resolved, and transcripts are segmented into coherent instructional subtopics to support fine-grained QA generation. (4) *Segment-aware QA prompting:* Segment-specific questions are automatically generated to require multimodal comprehension leveraging audio, visual, and textual cues. (5) *Answer grounding:* Answers are constructed in a step-wise manner, referencing at least two distinct video segments. Additional details are provided in the supplementary material.

Dataset Selection. We apply four filtering criteria for video curation: (1) synchronized audio, on-screen text, and visual changes; (2) clear procedural step sequences; (3) duration between 5–25 minutes; and (4) availability of English captions.

Grounded Audio Visual Question Answering. Each QA item includes: (i) a free-form question, (ii) a step-by-step answer, and (iii) a list of ⟨videoID, start, end⟩ references. Unlike prior single-clip datasets, 82% of our QA pairs require evidence from at least two distinct videos, making them well-suited for LMMs. Examples of extended answers are shown in supplementary.

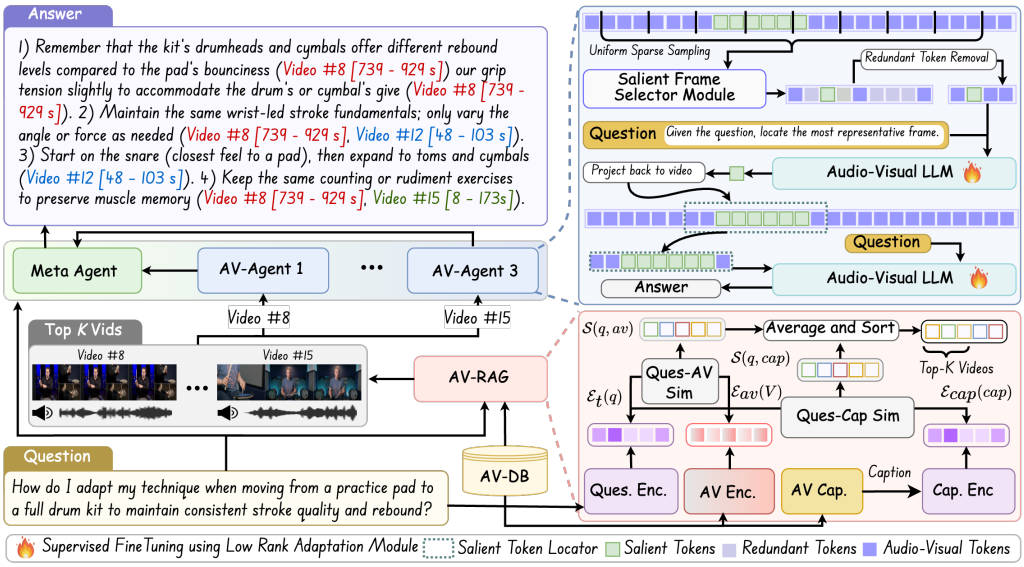


Figure 2: **Overview of MAGNET.** Given a user query, AV-RAG retrieves the top-K relevant videos (with audio), which are then processed by dynamically spawned Audio-Visual agents and a meta-agent aggregator to generate temporally grounded, step-wise responses. An adaptive, modality-agnostic frame selection module improves performance over baselines (see Tab. 2).

Multiple Audio-Visual Entity Linkage. To enable fine-grained, cross-video entity grounding without manual bounding boxes, we generated **3100** QA queries (train/test split 2k/1k) drawing evidence from a **500-video** pool. To facilitate experiments with baseline approaches we curate a small subset AVHaystacks-50. Instead of box-level labels, each entity is grounded via a sequence of multimodal spans: (i) textual segments from cleaned transcripts, (ii) audio intervals capturing distinctive sounds, and (iii) visual frame intervals showing the entity onscreen. These spans often span multiple videos in a defined order (e.g., steps 1–3 of a recipe across two clips), capturing both temporal and cross-video dependencies (details in supplementary).

3 Method

3.1 Task Definition: AVHaystacksQA

Given a question q and a collection of N videos $\mathcal{V} = \{V_1, \ldots, V_N\}$, our framework aims to retrieve the top-k most relevant videos to support AVLLMs understanding and answering the question q. MAGNET accomplishes this through a *two-step* retrieval process designed to effectively identify and reason over relevant videos for each question as demonstrated in Fig.2.

3.2 Audio Visual Preprocessing

AV RAG. We compute similarity scores between the averaged query representation q and (*i*) fused audio-visual features (using Hadamard fusion) $\mathcal{E}_{av}(\cdot)$ and (*ii*) encoded audio-visual captions $\mathcal{E}_{cap}(\cdot)$ for each video $V_j \in \mathcal{V}$ using cosine similarity, as defined in Eq. 1:

$$\mathcal{S}(q, \mathcal{V}) = \cos(\mathcal{E}_t(q), \mathcal{E}_f(V_j)) \quad \text{for} \quad V_j \in \mathcal{V}, \tag{1}$$

Here, $\mathcal{S}$ denotes the similarity score between query q and video set $\mathcal{V}$; cos is cosine similarity; $\mathcal{E}_t$ is the text encoder; and $\mathcal{E}_f \in \{\mathcal{E}_{av}, \mathcal{E}_{cap}\}$.

We use IMAGEBIND [23] to encode both the query and audio-visual features. Captions are generated using Gemini 1.5 Pro [24] and encoded with IMAGEBIND to obtain $\mathcal{E}_{cap}(\cdot)$; all embeddings are cached in our retrieval database to enable fast similarity computation without runtime re-encoding. To compute final relevance, we average $\mathcal{S}(av, q)$ and $\mathcal{S}(cap, q)$ to obtain Sim_{avg}, then rank videos in descending order and select the top-k most relevant videos.

Salient Frame Selector Module (SFS). Long videos often contain sparse but crucial moments relevant to a query. To efficiently localize these key events, we introduce a *salient 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()

Algorithm 2 STEM: **St**ep-wise **E**rror **M**etric

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: $\texttt{Sim}(\cdot)$, 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

module that focuses on visually and semantically important content. This ensures attention is directed to both the *right content* and the *right time*, facilitating efficient reasoning.

We begin by uniformly sampling m candidate frames without any prior. The objective is to select k representative frames that are both visually diverse and temporally dispersed.

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:

$$\Gamma_{ab} = \frac{z_a^\top z_b}{\|z_a\|_2 \cdot \|z_b\|_2}, \quad \forall a, b \in \{1, \dots, m\} \tag{2}$$

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

$$\Delta_{ab} = \gamma \left(\frac{1}{\sin\left(\frac{\pi}{2}|a-b|\right) + 1} - 1 \right) \tag{3}$$

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:

$$\mathcal{T} = \arg \min_{\substack{\mathcal{T} \subset \{1,\dots,m\} \\ |\mathcal{T}|=k}} \sum_{i=1}^{k-1} Q_{t_i t_{i+1}} \tag{4}$$

3.3 Grounded Question-Answering with Audio-Visual agents

Once the potentially relevant videos are shortlisted (i.e., those likely to be helpful in answering the given query), we deploy a dynamic agentic setup based on an AVLLM backbone. Due to its recent success in fine-grained audio-visual comprehension, we utilise Qwen 2.5 Omni [25] as the backbone for our AVLLM agents. For each shortlisted video, a dedicated instance of Qwen 2.5 Omni is spawned to process that video independently. Each agent analyses its assigned video and predicts the most relevant temporal segments along with a response that may contain an answer to the query. The outputs from all individual AVLLM agents, the identified time windows and corresponding partial responses are then aggregated by a meta-agent. In our setup, GPT-4o[26] acts as the meta-agent,

which ingests the agent responses and synthesises a coherent, contextually grounded final answer for the input query.

4 Experiments

4.1 Metrics

We utilize a set of 4 metrics to evaluate the outcomes of our approach which are as follows:

Response Alignment Scores. We evaluate the semantic alignment between the summarized outputs, comprising step-wise responses to a given query, and the corresponding ground truth using a combination of automated and human-centric metrics. For automated evaluation, we employ standard text-based similarity metrics such as BLEU@4 and CIDEr. Additionally, we utilize the GTE-L model [27] to extract sentence embeddings for both the predicted response and the ground truth, and compute their cosine similarity. Beyond these, we adopt the GPT-as-a-Judge framework to score the predicted responses against the ground truth on a 10-point scale, subsequently normalizing these scores for consistency. Finally, we include human evaluation scores (on a scale 1-5 with 1 being lowest) as a complementary metric (averaged across 20 evaluators) to assess the alignment and overall quality of the predicted answers with respect to the ground truth.

Retrieval Evaluation Scores. To evaluate how well our system retrieves relevant videos from the haystacks, we present recall values using *R@1*, *R@3*, and *R@5*. These metrics help us assess the video retrieval accuracy by examining the presence of relevant videos within the top few ranks.

Matched Temporal Grounding Score. To evaluate the alignment between predicted and ground truth temporal segments (along with video IDs), for each query, we propose the *Matched Temporal Grounding Score* (MTGS). This metric measures the average temporal overlap (IoU) between predicted and ground truth time intervals, but only for video instances where the video IDs match. If a video ID is not present in both prediction and ground truth, it is excluded from the computation. Furthermore, for each matched video ID, we compute the temporal IoU over the union of all corresponding intervals. This ensures that the metric reflects segment-wise grounding accuracy at the video level. Formally, let $\mathcal{V}_G$ and $\mathcal{V}_P$ denote the sets of video IDs present in the ground truth and predicted outputs, respectively. Let $\mathcal{V}_M = \mathcal{V}_G \cap \mathcal{V}_P$ be the set of matched video IDs. For each $v \in \mathcal{V}_M$, let $G_v = \{(t^{\text{gt}}_{\text{start}}, t^{\text{gt}}_{\text{end}})\}$ be the set of ground truth intervals and $P_v = \{(t^{\text{pred}}_{\text{start}}, t^{\text{pred}}_{\text{end}})\}$ be the set of predicted intervals. We define the temporal IoU for video v as: $\text{IoU}_v = \frac{\text{Duration}(\text{Intersection}(G_v, P_v))}{\text{Duration}(\text{Union}(G_v, P_v))}$, where the intersection and union are computed by merging overlapping intervals across G_v and P_v, and $\text{Duration}(\cdot)$ computes the total length of the resulting intervals. The final MTGS is then computed as the mean IoU across all matched video IDs: $\text{MTGS} = \frac{1}{|\mathcal{V}_M|} \sum_{v \in \mathcal{V}_M} \text{IoU}_v$. In cases where there is no matched video ID between prediction and ground truth (i.e., $|\mathcal{V}_M| = 0$), we define $\text{MTGS} = 0$. This metric provides a balanced and interpretable evaluation of segment-level grounding performance, with sensitivity to both partial and full overlaps, while ensuring fairness by averaging over matched video contexts. We report the average value of this score MTGS_{avg} in Tab. 3.

STep-wise Error Metric. The Step-wise Error Metric (STEM) quantifies alignment errors between a ground truth step sequence and a predicted step sequence in instructional or procedural data (detailed in Algorithm 2). Given Ground truth steps: $\{G_1, G_2, \ldots, G_n\}$, Predicted steps: $\{P_1, P_2, \ldots, P_m\}$, Text similarity threshold: $\tau_s \in [0, 1]$, typically set to 0.5, we begin by computing a similarity matrix $M_{\text{sim}} \in \mathbb{R}^{n \times m}$, where each entry is given by a cosine similarity function $\texttt{Sim}(G^{\text{text}}_i, P^{\text{text}}_j)$ computed between the step-wise text embeddings. We obtain a set of valid matched pairs of predicted and ground truth steps via Hungarian matching [28] which minimizes the total dissimilarity. If $i \neq j$, the prediction is out of order, contributing to the wrong step order count. Subsequently, to assess grounding mismatch, for each matched step, we compare video IDs and compute corresponding IoUs between prediction and ground truth. Finally, unmatched ground truth steps add to the missing step count, and unmatched predicted steps are considered as hallucinated steps.

4.2 Baselines

In our experiment, we have evaluated several open and closed-sourced Audio-Visual models on the retrieval and AVHaystacksQA performance. We extensively evaluate AVHaystacks on VideoRAG [8], Video-RAG [29], Qwen 2.5 Omni [25], Unified IO2 [4], Video-SALMONN [30]. We suitably adopt Video-RAG, VideoRAG for our task. To accommodate videos in Qwen-2.5-Omni, Unified-IO2, Video-SALMONN for AVHaystacks-50 we sparsely sample frames and downsize them to low resolution and also compress audio using [31].

Method	AVHaystacks-50					AVHaystacks-Full				
	B@4 ↑	Cr ↑	Text Sim ↑	GPT Eval ↑	H Eval ↑	B@4 ↑	Cr ↑	Text Sim ↑	GPT Eval ↑	H Eval ↑
VideoRAG [8]	43.16	119.78	5.31	6.32	3.42	41.59	115.97	5.15	6.13	3.32
Video-RAG [29]	42.64	117.86	5.23	6.20	3.37	40.67	112.12	4.99	5.97	3.23
Qwen2.5 omni [25]	10.84	28.59	1.90	2.11	1.07	-	-	-	-	-
Unified IO2 [4]	11.64	34.28	2.15	2.40	1.02	-	-	-	-	-
VideoSALMONN [30]	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

Table 2: **Response Alignment Scores.** Our proposed MAGNET offers significant gains over baseline approaches (first section) and our adapted baselines (second section) across multiple objective and subjective metrics on two dataset splits. B@4: BLEU@4, Cr: CIDEr, H Eval: Human Evaluation. Closed source model: as a reference for upperbound.

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

Table 3: **Grounding evaluation and Step-wise error results** on AVHaystack-50 and AVHaystack-Full datasets using MTGS and STEM (SM, SH, SO, SFP, SFN) metrics respectively.

Method	AVHaystacks-50		AVHaystacks-Full	
	R@3 ↑	R@5 ↑	R@3 ↑	R@5 ↑
ImageBind-RAG []	78.22	82.57	60.41	66.13
Text-RAG	82.84	84.87	66.54	71.60
Video-RAG	85.26	88.52	69.83	73.52
VideoRAG	85.57	89.79	70.43	74.96
Ours	**90.68**	**93.17**	**73.15**	**79.20**

Table 4: **Retrieval Evaluation Scores** on AVHaystack-50 and AVHaystack-Full datasets.

4.3 Main Results

Audio Visual QA. Experimental results in Tab. 2 demonstrate that among FT models MAGNET +Qwen 2.5 Omni-FT, achieves best performance across all automatic and human evaluation metrics. On both AVHaystacks splits, it achieves the highest scores outperforming both ZS and fine-tuned variants of MAGNET combined with other AVLLMs (e.g., VideoSALMONN, Unified IO2).

Grounding Evaluation and Step-wise Error Assessment. Tab. 3 indicates that our approach substantially improves the grounding capabilities and reduces step-wise error rates of the open-source AVLLMs, as reflected from the $MTGS_{Avg}$ and STEM values, respectively. To robustly validate our proposed STEM, we conduct a human evaluation and observe a strong correlation between human judgments and the metric. Refer to supplementary material for more details.

Notably, in both Tabs. 2 - 3, to provide an upper bound, we incorporate a strong closed-source model with powerful generative capabilities such as Gemini-1.5 -Pro [32] within multi-agent framework MAGNET, coupled with Salient Frame Selector module and report its performance. We observe that our best model MAGNET +Qwen 2.5 Omni-FT almost reaches the upper bound values for .

Audio Visual Retrieval. Our method sets a new benchmark on AV retrieval task across both the dataset splits (Tab. 4). Compared to existing approaches, it achieves the largest margin of improvement in R@3 and R@5, particularly on the more challenging AVHaystacks-Full, where gains over strong baselines like VideoRAG and Text-RAG are 2.7 points in R@3 and 7.6 points in R@5. These improvements indicate that our approach retrieves more relevant samples consistently and scales effectively to larger retrieval spaces. The results also highlight the benefit of leveraging richer modality integration and adaptive reasoning in our model over static retrieval pipelines.

The above results (Tabs. 2 - 4) demonstrate that employing our multi-modal RAG pipeline not only improves retrieval quality but also aligns better with human judgments.

4.4 Ablations

Importance of modalities. The results in Tab. 5 show that performance is generally best when both audio and visual modalities are used, highlighting the benefit of multi-modal information. Gemini-1.5-Pro consistently outperforms Qwen-2.5-Omni across all retrieval and response alignment scores metrics, indicating the benefits of MAGNET in formulating coherent and information-rich responses. Qualitative assessment indicates a strong correlation across tasks, underlining the utility of our RAG pipeline.

Sampling strategy. For both backbones, using SFS significantly improves performance across all metrics. For Qwen2.5-Omni-FT, switching from Uniform to SFS increases BLEU@4 score by 0.17

Method	Audio	Visual	BLEU@4 ↑	Text Sim ↑	GPT Eval ↑	Human Eval ↑
MAGNET +Qwen-2.5-Omni-FT	✗	✓	45.28	5.15	6.16	3.32
	✓	✗	38.96	4.82	5.78	3.13
	✓	✓	**53.64**	**6.28**	**7.52**	**4.01**
MAGNET +Gemini-1.5-Pro	✗	✓	48.48	5.55	6.64	3.57
	✓	✗	42.94	5.15	6.14	3.32
	✓	✓	**55.80**	**6.53**	**7.85**	**4.15**

Table 5: **Performance of MAGNET under different modality settings on AVHaystacks-Full.**

Method	Uniform	SFS	BLEU@4 ↑	Text Sim ↑	GPT Eval ↑	Human Eval ↑
MAGNET +Unified-IO2-FT	✓	✗	34.89	4.41	5.18	2.88
	✗	✓	**51.45**	**6.12**	**7.34**	**3.91**
MAGNET +Qwen-2.5-Omni-FT	✓	✗	36.58	4.58	5.42	2.98
	✗	✓	**53.61**	**6.28**	**7.53**	**4.01**
MAGNET +Gemini-1.5-Pro	✓	✗	41.94	4.74	5.64	3.08
	✗	✓	**55.87**	**6.53**	**7.81**	**4.15**

Table 6: **Effect of sampling strategy.** We systematically analyse our design choice replacing the sampling algorithm with 3 models on AVHaystacks-Full.

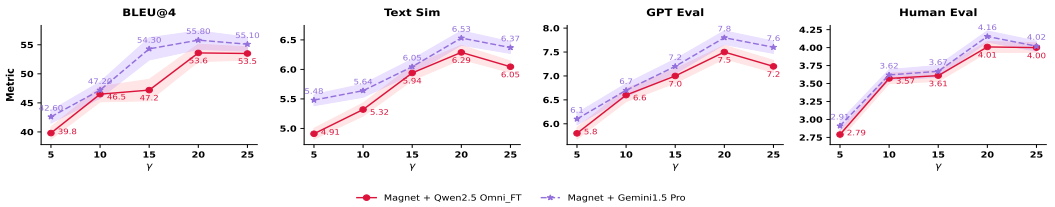


Figure 3: Effect of γ on eval metrics for MAGNET *+Qwen-2.5-Omni-FT* and MAGNET *+Gemini-1.5-Pro*.

and Human Eval score by 1.03. A similar trend is observed with Gemini 1.5 Pro where the best results are obtained with our proposed sampling strategy as seen in Tab. 6. These results underscore the advantage of semantically guided sampling over uniform strategies, as SFS more effectively captures informative segments, leading to better grounding, coherence, and human preference.

Penalty hyperparameter. Fig. 3 demonstrates a steady rise in performance across all metrics as γ increases, although a slight dip in performance is observed at $\gamma = 25$, notably in BLEU@4 and Human Eval for MAGNET + Qwen-2.5-Omni-FT, potentially indicate the onset of overfitting or increased parameter sensitivity in that region. The varying magnitudes of the dip across metrics indicate that the effect of γ is not uniform across different aspects of model performance.

4.5 Qualitative Results

Fig.4 showcases our system's ability to retrieve the most relevant videos and accurately localize temporal segments needed to answer a query using audio-visual cues. The meta-agent effectively highlights key instructional moments e.g., forming the fulcrum (`Video 1`), handshake shaping (`Video 8`), and diagonal pivoting (`Video 1`) in alignment with expert references. It also merges complementary segments from `Videos 8` and `18` to describe finger placement, demonstrating robustness to redundancy and variation. Overall, MAGNET excels at retrieving, grounding, and synthesising evidence across videos into coherent, temporally aligned responses compared to a recent baseline – which fails to retrieve suitable videos and subsequently fails to temporally ground the salient regions.

Additional qualitative and quantitative results are provided in the supplementary material.

5 Related Works

Video QA Benchmarks. Video Question Answering (VidQA) involves answering natural language queries using visual content alone or in combination with other modalities like audio [33–36]. Early benchmarks such as MovieQA [37] relied heavily on subtitles, with minimal visual grounding [38]. Datasets like ActivityNet-QA [39] and How2QA [40] target visual understanding in daily and instructional contexts. More recent efforts, including NeXT-QA [41], Perception Test [42], STAR [43], and AGQA [44], emphasize spatio-temporal and causal reasoning. EgoSchema [45] extends VidQA to long-form egocentric videos using LLM-generated questions. Other benchmarks address longer video reasoning [46, 36] and specialized domains like instructional [47, 48] and egocentric content [49–52]. Despite progress, most VidQA datasets are constrained by limited modalities or fixed time windows. Our work enables large-scale, cross-video retrieval and multimodal reasoning, bridging VidQA and broader audio-visual understanding.

MLLMs for Video Understanding. Open-source LLMs [53–55] have enabled Video MLLMs that connect visual encoders to LLMs via projection bridges [56, 34, 57, 58]. While effective on short clips, they struggle with long videos due to context limits and temporal complexity [11]. To address this, long-context LLMs [59–62] and token compression [63, 15, 64] scale input capacity and support agent-based decomposition and retrieval [65–67]. MovieChat [15], for example, uses hierarchical memory for frame-level summarization. However, these models lack the ability to reason over audio-visual content across multiple videos and generate coherent responses. Our multi-agent framework addresses this gap.

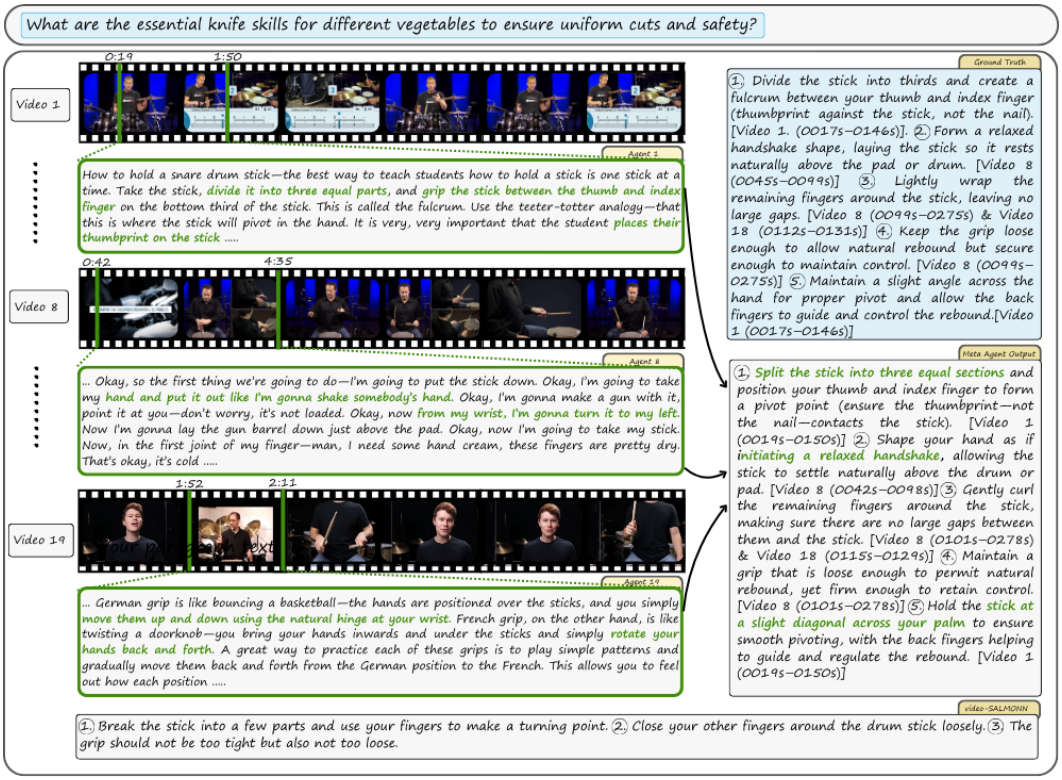


Figure 4: **Qualitative results.** Powered by efficient video retrieval pipeline and multi-agent configurations of MAGNET +Qwen-2.5-Omni-FT demonstrates strong reasoning abilities by first identifying the key videos followed by audio-visual temporal grounding to localise the salient regions across multiple videos when subjected to a how-to question.

Retrieval Augmented Generation (RAG). RAG improves generative models by integrating retrieval to inject external knowledge [68–73]. While well-studied in text domains [74–79], recent efforts have extended RAG to vision-language tasks [80–85]. MuRAG [81] uses non-parametric multimodal memory, and MIRAGE [82] employs CLIP-based retrieval. Other methods convert images to text via OCR, captioning, and detection before dense retrieval [86, 79]. Multimodal RAG has also been applied in domains like healthcare [87, 88], leveraging images as contextual input. While prior models focus on text or vision alone, MAGNET integrates off-the-shelf models with a custom retrieval-fusion pipeline for scalable audio-visual-language retrieval and generation.

6 Conclusions and Future Work

We introduced a novel benchmark and framework for evaluating LMMs in audio-visual retrieval and reasoning an area less explored than image or single-video based settings. Our task targets the challenging problem of retrieving relevant audio-visual segments from large video corpora, requiring joint temporal, auditory, and visual reasoning, akin to real-world multimedia search. To tackle this, we proposed MAGNET, a scalable retrieval-augmented generation system that identifies key moments across multiple videos and synthesizes grounded responses. It combines a sampling strategy, off-the-shelf models, and a multi-agent relevance scoring mechanism to extract and fuse salient content. Experiments show substantial gains over baselines, underscoring the promise of retrieval-augmented methods in audio-visual reasoning. We hope this work inspires richer benchmarks that push LMMs toward dynamic, temporally grounded multimodal understanding.

While MAGNET and AVHaystacks advance AV multi-video reasoning, several future directions remain. Replacing off-the-shelf components with end-to-end trainable modules could improve retrieval and frame selection. Enhancing agentic reasoning with collaborative mechanisms (e.g., planning or voting) may boost interpretability and performance. Lastly, integrating personalisation (e.g., user-driven retrieval) would support real-world applications like education and assistive tools.

References

[1] Henghui Du, Guangyao Li, Chang Zhou, Chunjie Zhang, Alan Zhao, and Di Hu. Crab: A unified audio-visual scene understanding model with explicit cooperation. *arXiv preprint arXiv:2503.13068*, 2025.

[2] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Meerkat: Audio-visual large language model for grounding in space and time. In *European Conference on Computer Vision*, pages 52–70. Springer, 2024.

[3] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024.

[4] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455, 2024.

[5] Shoubin Yu, Jaehong Yoon, and Mohit Bansal. Crema: Generalizable and efficient video-language reasoning via multimodal modular fusion. *arXiv preprint arXiv:2402.05889*, 2024.

[6] Yunlong Tang, Daiki Shimada, Jing Bi, and Chenliang Xu. Avicuna: Audio-visual llm with interleaver and context-boundary alignment for temporal referential dialogue. *arXiv e-prints*, pages arXiv–2403, 2024.

[7] Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E Gonzalez, Trevor Darrell, and David M Chan. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. *arXiv preprint arXiv:2407.13766*, 2024.

[8] Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. Videorag: Retrieval-augmented generation with extreme long-context videos. *arXiv preprint arXiv:2502.01549*, 2025.

[9] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

[10] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37: 28828–28857, 2024.

[11] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.

[12] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024.

[13] Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*, 2024.

[14] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.

[15] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024.

[16] Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22942–22951, 2023.

[17] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36:72842–72866, 2023.

[18] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3480–3491, 2022.

[19] Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. In *European Conference on Computer Vision*, pages 146–164. Springer, 2024.

[20] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118, 2022.

[21] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020.

[22] Jungang Li, Sicheng Tao, Yibo Yan, Xiaojie Gu, Haodong Xu, Xu Zheng, Yuanhuiyi Lyu, Linfeng Zhang, and Xuming Hu. Saven-vid: Synergistic audio-visual integration for enhanced understanding in long video context. *arXiv preprint arXiv:2411.16213*, 2024.

[23] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023.

[24] Google AI. Gemini: Google's multimodal ai model. *Google AI Research*, 2024. `https://fireflies.ai/blog/gemini-vs-gpt-4`.

[25] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[26] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[27] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

[28] Harold W Kuhn. Variants of the hungarian method for assignment problems. *Naval research logistics quarterly*, 3(4):253–258, 1956.

[29] Yongdong Luo, Xiawu Zheng, Xiao Yang, Guilin Li, Haojia Lin, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *arXiv preprint arXiv:2411.13093*, 2024.

[30] Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*, 2024.

[31] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36: 27980–27993, 2023.

[32] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[33] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

[34] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.

[35] Yaoyao Zhong, Junbin Xiao, Wei Ji, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges. *arXiv preprint arXiv:2203.01225*, 2022.

[36] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024.

[37] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016.

[38] Bhavan Jasani, Rohit Girdhar, and Deva Ramanan. Are we asking the right questions in movieqa? In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[39] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019.

[40] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*, 2018.

[41] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.

[42] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36: 42748–42761, 2023.

[43] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024.

[44] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297, 2021.

[45] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

[46] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37: 28828–28857, 2024.

[47] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1686–1697, 2021.

[48] Haiwan Wei, Yitian Yuan, Xiaohan Lan, Wei Ke, and Lin Ma. Instructionbench: An instructional video understanding benchmark. *arXiv preprint arXiv:2504.05040*, 2025.

[49] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, et al. Hd-epic: A highly-detailed egocentric video dataset. *arXiv preprint arXiv:2502.04144*, 2025.

[50] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022.

[51] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024.

[52] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018.

[53] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[54] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[55] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[56] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

[57] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[58] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[59] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024.

[60] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.

[61] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024.

[62] Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. Longllava: Scaling multi-modal llms to 1000 images efficiently via hybrid architecture. *arXiv preprint arXiv:2409.02889*, 2024.

[63] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023.

[64] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024.

[65] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. *arXiv preprint arXiv:2403.11481*, 2024.

[66] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. *arXiv preprint arXiv:2403.10517*, 2024.

[67] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. *arXiv preprint arXiv:2405.19209*, 2024.

[68] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[69] Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, 2023.

[70] Yang Bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Chun-Mei Feng. Sentence-level prompts benefit composed image retrieval. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. Spotlight Presentation.

[71] Chun-Mei Feng, Yang Bai, Tao Luo, Zhen Li, Salman Khan, Wangmeng Zuo, Xinxing Xu, Rick Siow Mong Goh, and Yong Liu. Vqa4cir: Boosting composed image retrieval with visual question answering. *arXiv preprint arXiv:2312.12273*, 2023. URL `https://arxiv.org/abs/2312.12273`.

[72] Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, et al. Searching for best practices in retrieval-augmented generation. *arXiv preprint arXiv:2407.01219*, 2024.

[73] Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue Liu, Tei-Wei Kuo, Nan Guan, et al. Retrieval-augmented generation for natural language processing: A survey. *arXiv preprint arXiv:2407.13193*, 2024.

[74] Yizheng Huang and Jimmy Huang. A survey on retrieval-augmented text generation for large language models. *arXiv preprint arXiv:2404.10981*, 2024.

[75] Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*, 2024.

[76] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023.

[77] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 3929–3938. PMLR, 2020.

[78] Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox, Helen Meng, and James Glass. SAIL: Search-augmented instruction learning. *arXiv preprint arXiv:2305.15225*, 2023.

[79] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics, 2020.

[80] Jun Chen, Dannong Xu, Junjie Fei, Chun-Mei Feng, and Mohamed Elhoseiny. Document haystacks: Vision-language reasoning over piles of 1000+ documents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24817–24826, 2025.

[81] Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5558–5570. Association for Computational Linguistics, 2022.

[82] Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E. Gonzalez, Trevor Darrell, and David M. Chan. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. *arXiv preprint arXiv:2407.13766*, 2024.

[83] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024.

[84] Yin Wu, Quanyu Long, Jing Li, Jianfei Yu, and Wenya Wang. Visual-rag: Benchmarking text-to-image retrieval augmented generation for visual knowledge intensive queries. *arXiv preprint arXiv:2502.16636*, 2025.

[85] Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. Vdocrag: Retrieval-augmented generation over visually-rich documents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24827–24837, 2025.

[86] Weizhe Lin and Bill Byrne. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*, 2022.

[87] Liwen Sun, James Zhao, Megan Han, and Chenyan Xiong. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. *arXiv preprint arXiv:2407.15268*, 2024.

[88] Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. Rule: Reliable multimodal rag for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1093, 2024.

[89] Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1593–1603, 2024.

[90] Nader Asadi, Mahdi Beitollahi, Yasser Khalil, Yinchuan Li, Guojun Zhang, and Xi Chen. Does combining parameter-efficient modules improve few-shot transfer accuracy? *Artificial Intelligence for Engineering*, 2024.

[91] Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with qlora. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, 2023.

[92] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[93] OpenAI. Gpt-4o: Enhanced multimodal language model. *OpenAI Research*, 2024. https://openai.com/index/hello-gpt-4o/.

[94] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=1tZbq88f27.

[95] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.

[96] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2024.

[97] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[98] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024.

[99] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[100] Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. Videorag: Retrieval-augmented generation with extreme long-context videos. *arXiv preprint arXiv:2502.01549*, 2025.

[101] Sanjoy Chowdhury, Sreyan Ghosh, Subhrajyoti Dasgupta, Anton Ratnarajah, Utkarsh Tyagi, and Dinesh Manocha. Adverb: Visually guided audio dereverberation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7884–7896, 2023.

[102] Sanjoy Chowdhury, Sayan Nag, KJ Joseph, Balaji Vasan Srinivasan, and Dinesh Manocha. Melfusion: Synthesizing music from image and language cues using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26826–26835, 2024.

[103] Xinhao Mei, Varun Nagaraja, Gael Le Lan, Zhaoheng Ni, Ernie Chang, Yangyang Shi, and Vikas Chandra. Foleygen: Visually-guided audio generation. In *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2024.

[104] Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. Codi-2: In-context interleaved and interactive any-to-any generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27425–27434, 2024.

[105] Sanjoy Chowdhury, Subhrajyoti Dasgupta, Sudip Das, and Ujjwal Bhattacharya. Listen to the pixels. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2568–2572. IEEE, 2021.

[106] Sanjoy Chowdhury, Aditya Patra, Subhrajyoti Dasgupta, and Ujjwal Bhattacharya. Audvisum: Self-supervised deep reinforcement learning for diverse audio-visual summary generation. In *BMVC*, page 315, 2021.

[107] Junyu Gao, Hao Yang, Maoguo Gong, and Xuelong Li. Audio–visual representation learning for anomaly events detection in crowds. *Neurocomputing*, 582:127489, 2024.

[108] Parthasaarathy Sudarsanam, Irene Martín-Morató, and Tuomas Virtanen. Representation learning for semantic alignment of language, audio, and visual modalities. *arXiv preprint arXiv:2505.14562*, 2025.

[109] Sanjoy Chowdhury, Hanan Gani, Nishit Anand, Sayan Nag, Ruohan Gao, Mohamed Elhoseiny, Salman Khan, and Dinesh Manocha. Aurelia: Test-time reasoning distillation in audio-visual llms. *arXiv preprint arXiv:2503.23219*, 2025.

[110] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Yaoting Wang, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Avtrustbench: Assessing and enhancing reliability and robustness in audio-visual llms. *arXiv preprint arXiv:2501.02135*, 2025.

[111] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[112] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*, 2024.

[113] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

[114] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024.

[115] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15516–15525, 2021.

[116] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*, 2023.

[117] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124, 2024.

[118] Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, et al. Av-odyssey bench: Can your multimodal llms really understand audio-visual information? *arXiv preprint arXiv:2412.02611*, 2024.

[119] Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, et al. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*, 2024.

[120] Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. *arXiv preprint arXiv:2411.19772*, 2024.

[121] Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. *arXiv preprint arXiv:2410.18325*, 2024.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have tried to be thorough while presenting the primary contributions of this work. We will provide more details in the supplementary

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We add limitations and future work in the conclusions section

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We don't provide any theoretical results. Our method and benchmark are explained in detail, and we back every claim with empirical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have tried to be thorough while explaining our approach, and we will publicly release the code and dataset upon acceptance

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will publicly release the code and dataset to ensure reproducibility

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have tried to add all such details

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Not applicable for us

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We will add these details in the supplementary

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We do

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: We don't deal with any sensitive information. The collected samples are under appropriate licence and free to be used

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work doesn't contain such risk

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We duly cite and refer to parts that has been used as a part of our study. We will further acknowledge all the baselines in the final version of the paper

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

Justification: We will release the data and code publicly and will add the documentation

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: We will add this in the supplementary

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: Our work doesn't involve any risk

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [No]

Justification: We don't involve the usage of LLM other than for experimental and data preparation purposes

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

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

Supplementary Material

The supplementary is organised as follows:

A Supplementary Video

In the supplementary video, we elaborate on our proposed task AVHaystacksQA with illustrative examples and demonstrate the intricacies involved in a multi-video linked QA setting. The video shows how the given question involves referring to multiple videos to obtain a comprehensive answer, followed by the meta agent summarising the responses to come up with the final answer. We also highlight the salient components of MAGNET and discuss the end-to-end flow. The use of headphones is recommended for a better audio-visual QA experience.

B Dataset Statistics

In this section, we provide additional details about AVHaystacks. Tab. B summarizes the topics from which the samples are collected, along with the number of questions per category and the corresponding video-to-question ratio. The benchmark comprises **103 hours** of video content from **500 video samples** across **27 diverse** categories, accompanied by carefully annotated QA pairs that temporally ground salient segments within the videos. **To the best of our knowledge, this is the first benchmark of its kind, as no prior work provides multi-video linked audio-visual QA pairs.**

Fig. 5 and 6 depict the distribution of total video hours across the various categories included in our benchmark. As shown, the samples are well distributed and sourced from a wide range of scenarios. A significant portion is drawn from *Travel Destinations*, *Education and Language Learning*, *Music and Performing Arts*, and *DIY and Creative Hobbies*—domains that demand strong audio-visual comprehension for effective temporally grounded AV QA.

Fig. 7 illustrates the distribution of sample durations (in minutes) within AVHaystacks. Most video samples fall within the 6–15 minute range, posing substantial challenges for temporal grounding tasks. Evaluation models must handle long context windows and effectively manage the increased complexity of processing extended multimodal sequences (vision and audio).

Additionally, Fig. 8 shows the distribution of the number of videos associated with each question. As observed, the majority of questions require referencing two or more videos to determine the answer, further emphasizing the complexity and richness of AVHaystacks.

Finally, it is to be noted that: (i) Our dataset annotation process requires manual inspection to ensure complete correctness. Although the sample curation process is automated, to ensure strict sanity we manually validate each sample, which is both tedious and time-consuming. (ii) The models evaluated in our study are already audio-visually informed, and we fine-tune them efficiently using LoRA adapters. Our salient frame selection module enhances context by focusing on key frames. With LoRA, we require fewer samples to fine-tune our audio-visual agents, as they are heavily pre-trained on audio-visual data. Importantly, we are only fine-tuning

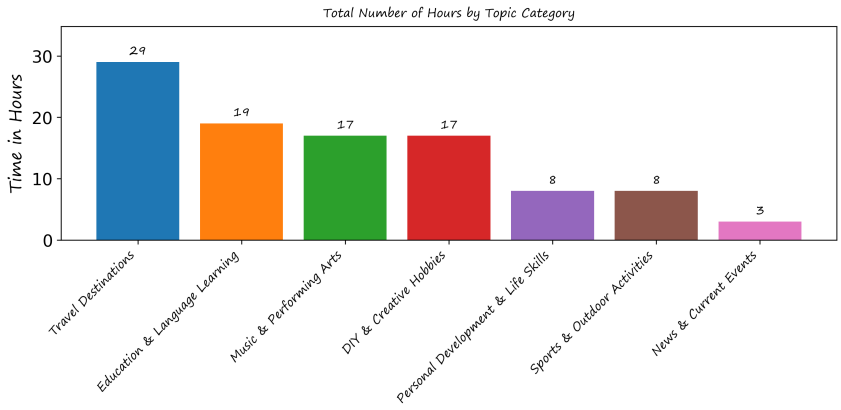


Figure 5: Number of hours of videos per topic category in AVHaystacks benchmark.

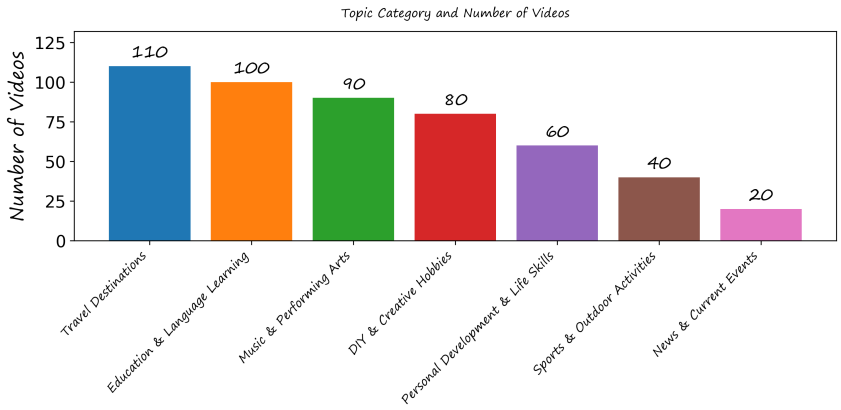


Figure 6: Number of videos per topic category in AVHaystacks benchmark.

the agents, not training the entire retrieval pipeline. Notably, the SFS module can be used with a closed-source model where finetuning was not done. As seen in Tables 2 and 3, these methods demonstrate strong performance even without any fine-tuning. Similar observations were made in recent literature [89–91] demonstrating that employing LoRA enables these models to perform well even with limited samples.

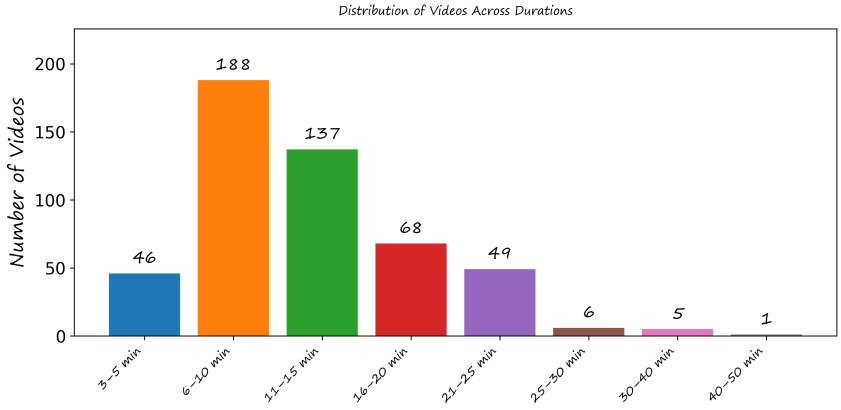


Figure 7: Distribution of videos based on their duration AVHaystacks benchmark.

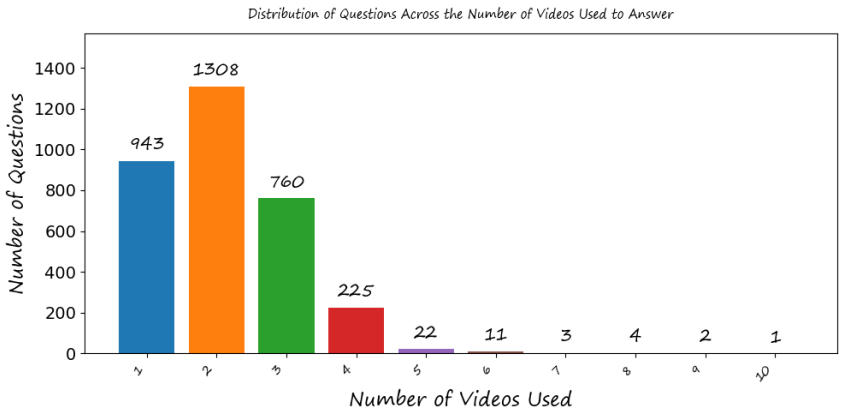


Figure 8: Number of videos referred to in each question AVHaystacks benchmark.

Topic	# Videos	# Questions	Ratio Videos:Questions
General Cooking Tutorials	20	117	1/6
Learn Vocals	20	51	2/5
Learn English	20	127	1/6
Learn Arabic	20	174	1/9
Learn Chinese	20	115	1/6
Learn Urdu	20	247	1/9
Travel Turkey	10	78	1/8
Travel Brazil	20	78	1/4
Travel UAE	20	105	1/5
Travel Hawaii	20	105	1/5
Travel USA	20	106	1/5
Travel Italy	20	106	1/5
Sing Beatbox	20	129	1/6
Learn Opera	20	123	1/6
DIY	20	187	1/9
Learn Sketching	20	89	1/4
3D Printing	20	189	1/9
Public Speaking	20	97	1/5
First-Aid	20	125	1/6
Self Defense	20	126	1/6
Soccer Analysis	20	139	1/7
News - Tornado	20	112	1/5
Sign Language	20	107	1/5
Hiking/Backpacking	20	151	1/7
Playing Music - Drums	20	80	1/4
Playing Music - String Instrument	5	34	1/7
Playing Music - Piano	5	50	1/10
Total	**500**	**3147**	**1/6**

Table 7: Distribution of videos and QA pairs across different topics.

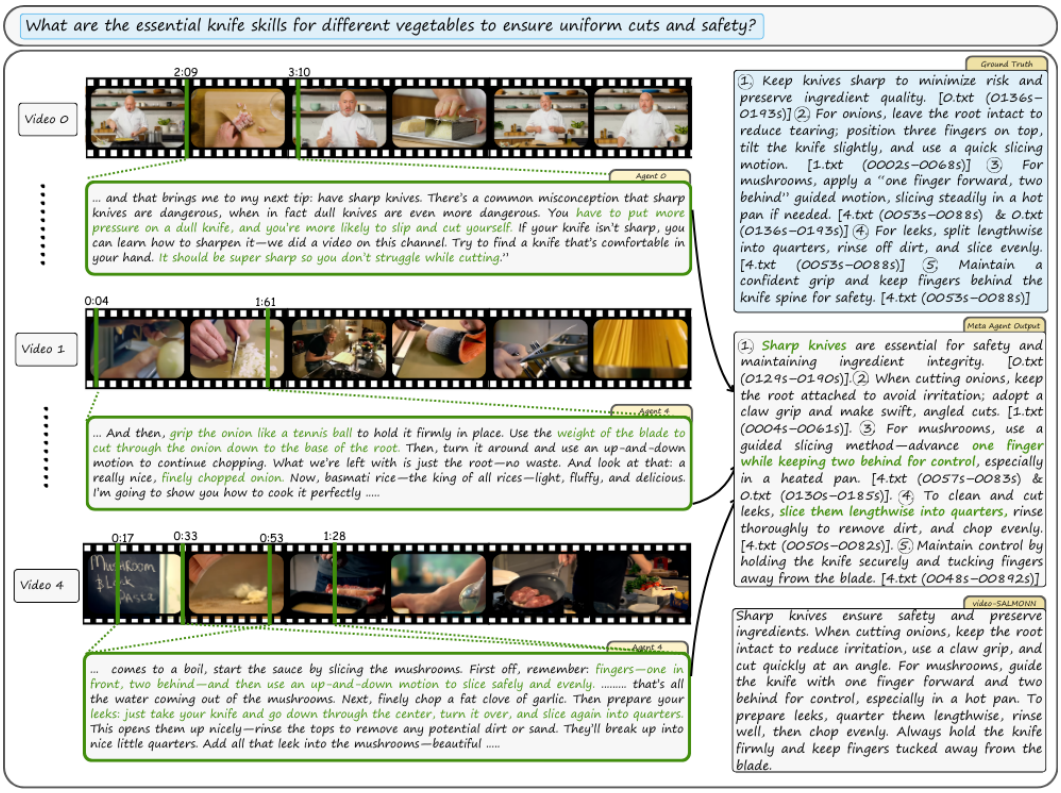


Figure 9: Performance comparison of MAGNET and Video-SALMONN on cooking video tutorial.

C Qualitative Results

The qualitative examples in Fig. 9 to Fig. 13 showcase a variety of scenarios, including cooking demonstrations, language-speaking tutorials, and 3D printing lessons. In these examples, we compare the responses generated by baseline models with those produced by MAGNET. As illustrated, AVLLMs enhanced with our retrieval and multi-agent audio-visual reasoning modules consistently outperform the baselines on AVHaystacksQA. While the baseline models often struggle to identify and retrieve the most informative video segments, our method accurately selects the most relevant videos for each question, exhibiting strong audio-visual comprehension and reasoning capabilities. Furthermore, the temporal windows identified by our approach are generally precise and focused, reflecting a nuanced cross-modal understanding and effective temporal grounding across videos.

D More Ablation Results

Top-k Video Selection. The top-k selection experiment on AVHaystacks-Full shows (Tab. 8) that both Qwen-2.5-Omni-FT and Gemini-1.5-Pro achieve peak performance at $k = 6$, where BLEU4, Text Similarity, GPT Eval, and STEM-Missing metrics all indicate optimal retrieval and generation quality. Gemini-1.5-Pro consistently outperforms Qwen across all k values, demonstrating stronger contextual understanding and output coherence. While increasing k generally improves performance by reducing missed content, values beyond $k = 6$ offer slightly diminishing returns and may introduce noise. Further, it leads to an increase in compute (since more AVLLM agents are required). These findings highlight the importance of selecting an appropriate k and leveraging more capable language models like Gemini for effective multimodal retrieval.

Meta Agents. Tab. 9 analyses the impact of different meta agents within the MAGNET framework on the AVHaystacks-Full benchmark. Across both base models: Qwen-2.5-Omni-FT and Gemini-1.5-Pro performance consistently improves as stronger meta agents are used, with Gemini achieving the best results in all metrics (BLEU4, Text Sim, GPT Eval, and STEM-Order). Notably, Gemini, as both the meta agent and core model, yields the highest overall performance, highlighting its superior capability in segment selection, coherence, and multimodal reasoning. These results underscore the critical role of the meta agent in guiding high-quality content synthesis.

Frame Sampling Function. Tab. 10 compares the effect of different frame sampling functions (SFS) on the AVHaystacks-Full benchmark using two base models: MAGNET $_{\text{+Qwen-2.5-Omni-FT}}$ and MAGNET $_{\text{+Gemini-1.5-Pro}}$.

Figure 10: Performance comparison of MAGNET and Unified IO2 on english pronunciation tutorial.

Across both models, we observe that the proposed SFS function (highlighted), which dynamically scales dissimilarity using an inverse sine function, consistently outperforms cosine and exponential alternatives. For instance, with Qwen-2.5-Omni-FT, the proposed SFS achieves the highest scores across all metrics. Similarly, when paired with Gemini-1.5-Pro, the same function yields the best performance, achieving high BLEU4 and a notable MTGS_{avg}. These results suggest that the sinusoidal-inverse-based SFS is more effective at capturing temporal importance for reasoning tasks, likely due to its sharper penalisation of semantically redundant or temporally proximal frames. This indicates that thoughtful frame selection plays a critical role in enhancing downstream multimodal generation quality.

Number of Video Frames Selection. Tab.11 analyzes the impact of varying the number of uniformly sampled frames m on AVHaystacks-Full using two base models. Across both models, performance consistently improves as the number of frames increases from 15 to 75. For instance, with Qwen-2.5-Omni-FT, BLEU@4 rises from 45.89 to 53.61, and MTGS_{avg} improves substantially from 0.43 to 0.83, indicating that denser frame sampling leads to more accurate and temporally grounded responses. Similarly, Gemini-1.5-Pro shows a consistent upward trend, reaching its peak performance at $m = 75$. These results suggest that richer frame coverage provides stronger contextual grounding for both models, reinforcing the need for higher temporal granularity in multi-video audio-visual reasoning tasks.

Text Similarity Threshold.

Tab. 12 presents the effect of varying the text similarity threshold τ_s used in frame filtering on AVHaystacks-Full. For both methods, performance peaks at $\tau_s = 0.5$, with **BLEU@4**, **Text Sim**, and **GPT Eval** achieving the

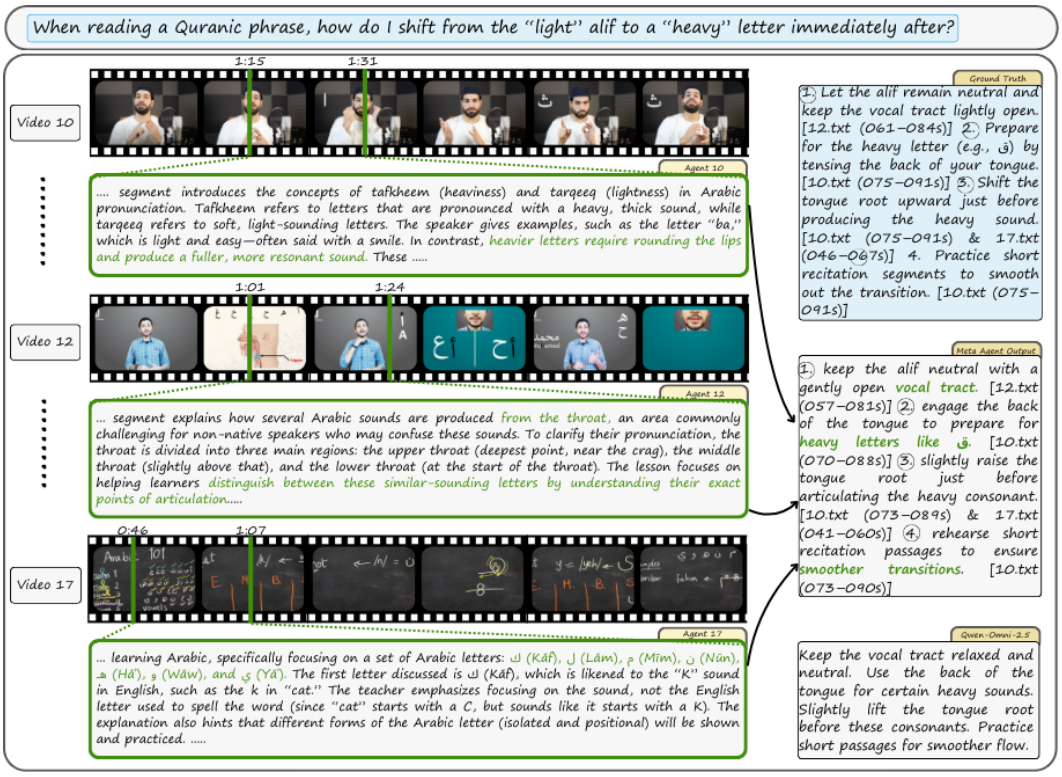


Figure 11: Performance comparison of MAGNET and Qwen-2.5-Omni on Urdu pronunciation tutorial.

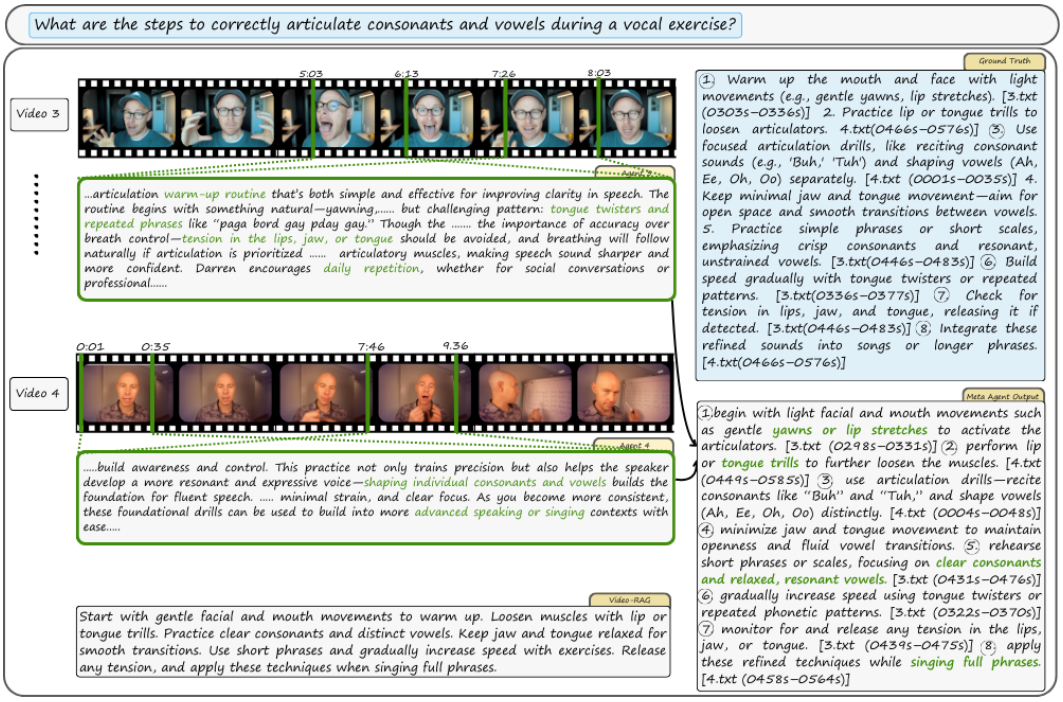


Figure 12: Performance comparison of MAGNET and Video-RAG on vocal exercise tutorial.

Figure 13: Performance comparison of MAGNET and VideoRAG on 3D printing tutorial.

highest scores, while the STEM-Missing and STEM-Order errors are minimized. Specifically, **Qwen-2.5-Omni-FT** sees improvements in BLEU@4 from **51.32** to **53.61** and a drop in STEM-Missing from **0.15** to **0.13** as τ_s increases from 0.3 to 0.5. Similarly, **Gemini-1.5-Pro** reaches optimal performance at $\tau_s = 0.5$, with the best overall textual coherence and minimal temporal grounding errors. However, setting $\tau_s = 0.7$ degrades performance across metrics, suggesting that overly strict filtering removes useful context. These findings highlight the importance of balancing informativeness and precision in frame selection by carefully tuning the similarity threshold.

E Human Evaluation on STEM:

Tab. 13 presents a comparative evaluation of various MAGNET model configurations using both automated STEM and human evaluation metrics averaged across 20 raters (Cohen's $\kappa = 0.82$). In particular, the fine-tuned MAGNET $_{+\text{ Qwen 2.5 Omni-FT}}$ model achieves strong overall performance, outperforming other models in most categories. Specifically, it ties for the lowest SM, achieves the lowest SH, and second-lowest SO scores under STEM, indicating improved semantic and syntactic alignment. It also performs competitively in human evaluations, with HM, HH, and HO scores close to or better than all other non-proprietary models. Although MAGNET $_{+\text{ Gemini 1.5 Pro}}$ demonstrates slightly superior performance across several metrics. These results underscore the efficacy of fine-tuning with Qwen 2.5 Omni, particularly in aligning model outputs with human judgments.

Method	K	BLEU@4 ↑	Text Sim ↑	GPT Eval ↑	STEM-Missing ↓
MAGNET +Qwen-2.5-Omni-FT	1	46.67	4.67	5.37	0.29
	3	50.44	5.93	6.88	0.20
	6	**53.61**	**6.28**	**7.53**	**0.13**
	10	53.38	6.18	7.36	0.14
MAGNET +Gemini-1.5-Pro	1	49.80	4.84	5.04	0.27
	3	52.09	5.22	6.31	0.19
	6	**55.87**	**6.53**	**7.81**	**0.12**
	10	55.63	6.41	7.64	0.12

Table 8: **Top-k selection** on AVHaystacks-Full.

F More Details on Benchmark Construction

In this section, we provide further details on AVHaystacks construction. We outline the steps involved in data preparation in Fig. 14. The complete benchmark creation pipeline details.

Method	Meta Agent	BLEU@4 ↑	Text Sim ↑	GPT Eval ↑	STEM-Order ↓
MAGNET +Qwen-2.5-Omni-FT	Reka	51.38	5.91	6.88	0.25
	Qwen-2.5-Omni-FT	51.98	5.98	6.96	0.24
	Claude	52.29	6.05	7.34	0.22
	GPT	52.93	6.11	7.37	0.20
	Gemini	**53.61**	**6.28**	**7.53**	**0.19**
MAGNET +Gemini-1.5-Pro	Reka	54.89	6.02	6.80	0.22
	Claude	55.02	6.13	7.09	0.20
	GPT	55.32	6.20	7.48	0.19
	Gemini	**55.87**	**6.53**	**7.81**	**0.17**

Table 9: **Effect of meta agents** on AVHaystacks-Full.

Method	SFS Function	BLEU@4 ↑	Text Sim ↑	GPT Eval ↑	MTGS_{avg} ↑
MAGNET +Qwen-2.5-Omni-FT	$\Delta_{ab} = \gamma\left(\cos\left(\frac{\pi}{2}\lvert a-b\rvert\right)-1\right);\ \gamma=10$	52.85	5.93	7.24	0.79
	$\Delta_{ab} = \gamma\left(e^{\lambda\lvert a-b\rvert}-1\right);\ \gamma=10, \lambda=5$	53.18	6.03	7.47	0.81
	$\Delta_{ab} = \gamma\left(\frac{1}{\sin\left(\frac{\pi}{2}\lvert a-b\rvert\right)+1}-1\right);\ \gamma=20$	**53.61**	**6.28**	**7.53**	**0.83**
MAGNET +Gemini-1.5-Pro	$\Delta_{ab} = \gamma\left(\cos\left(\frac{\pi}{2}\lvert a-b\rvert\right)-1\right);\ \gamma=10$	55.20	6.01	7.33	0.80
	$\Delta_{ab} = \gamma\left(e^{\lambda\lvert a-b\rvert}-1\right);\ \gamma=10, \lambda=5$	55.79	6.15	7.62	0.82
	$\Delta_{ab} = \gamma\left(\frac{1}{\sin\left(\frac{\pi}{2}\lvert a-b\rvert\right)+1}-1\right);\ \gamma=20$	**55.87**	**6.53**	**7.81**	**0.85**

Table 10: **Effect of different frame sampling functions** on AVHaystacks-Full.

Method	m	BLEU@4 ↑	Text Sim ↑	GPT Eval ↑	MTGS_{avg} ↑
MAGNET +Qwen-2.5-Omni-FT	15	45.89	4.46	5.10	0.43
	50	49.27	5.19	6.03	0.62
	75	**53.61**	**6.28**	**7.53**	**0.83**
MAGNET +Gemini-1.5-Pro	15	47.36	4.92	5.44	0.51
	50	51.56	5.81	6.62	0.69
	75	**55.87**	**6.53**	**7.81**	**0.85**

Table 11: **Effect of number of frames selection** on AVHaystacks-Full.

Method	τ_s	BLEU@4 ↑	Text Sim. ↑	GPT Eval ↑	STEM-Missing ↓	STEM-Order ↓
MAGNET +Qwen-2.5-Omni-FT	0.3	51.32	5.75	6.68	0.15	0.24
	0.5	**53.61**	**6.28**	**7.53**	**0.13**	**0.19**
	0.7	50.67	5.52	6.43	0.19	0.22
MAGNET +Gemini-1.5-Pro	0.3	54.13	6.31	7.56	0.13	0.22
	0.5	**55.87**	**6.53**	**7.81**	**0.12**	**0.17**
	0.7	53.95	6.30	7.51	0.17	0.20

Table 12: **Text similarity threshold** on AVHaystacks-Full.

F.1 Dataset Examples

Please find example videos with QAs from different categories as shared in the supplementary zip (refer to the *'AVHaystacks-dataset-samples'* folder). **The video files are compressed to fit within the supplementary material size limit**. Samples are collected from different areas (how-to, musical lessons, news etc) making our benchmark extremely diverse and considerably challenging for the models. The purpose of curating samples from such diverse sources is to robustly evaluate every model on their generalization capabilities.

Method	STEM			Human Eval.		
	SM ↓	SH ↓	SO ↓	HM ↓	HH ↓	HO ↓
MAGNET +VideoSALMONN-ZS	0.41	0.33	0.43	0.39	0.37	0.41
MAGNET +Unified IO2-ZS	0.49	0.39	0.37	0.46	0.35	0.36
MAGNET +Qwen 2.5 Omni -ZS	0.43	0.34	0.39	0.45	0.37	0.42
MAGNET +VideoSALMONN-FT	0.13	0.18	0.23	0.15	0.17	0.21
MAGNET +Unified IO2-FT	0.15	0.18	0.20	0.19	0.23	0.22
MAGNET +Qwen 2.5 Omni-FT	**0.13**	**0.16**	**0.19**	**0.17**	**0.18**	**0.16**
MAGNET +Gemini 1.5 Pro	0.12	0.14	0.17	0.14	0.18	0.13

Table 13: **STEMvs Human** on AVHaystack-Full dataset. SM: STEM- Missing, SH: STEM- Hallucination, SO: STEM- Order, HM: Human Eval - Missing, HH: Human Eval. - Hallucination, HO - Human Eval. - Order

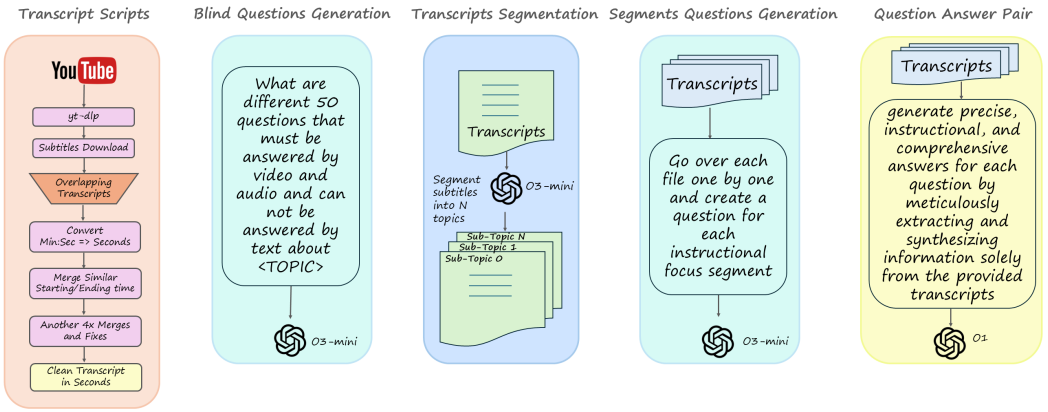


Figure 14: Steps involved in benchmark data collection.

F.2 Pipeline overview

Our multi-step data collection strategy is as follows in sequence: (1) download and curate 500 videos that satisfy the four-modality filter (Dataset Selection); (2) issue 50 blind prompts per topic (Listing 1) to GPT *without* revealing transcripts to encourage cross-video reasoning; (3) repair caption timing (Alg. 1) and convert HH:MM:SS→seconds for uniform indexing; (4) segment transcripts into sub-topics (Listing 4) and create one segment-aware question each (Listing 2); and (5) assemble QA pairs whose answers cite at least two videos (Listing 5).

F.3 Question-pipeline specifics

(i) *Blind phase*: prompt GPT with Listing 1; discard any question whose answer is general knowledge information or an answer that does not depend on audio and visuals; (ii) *Segment phase*: To expand the number of questions, we used segmented-phase questions. For each sub-topic, prompt GPT with Listing 2; ensure the generated question references audio, visual, and caption tokens.

Listing 1: Blind questions generation prompt

```
I am collecting data about how to <TOPIC>.
What are different 50 questions
that must be answer by a video, audio
and can not be answered by text only?
```

Listing 2: Segments-based questions generation

```
Go over each file one by one and create
a question for each segments
```

F.4 Caption repair and segmentation

(1) *Overlap fix*: downloaded subtitles had mis-matched intervals which needs to be fixed (example in Listing 3); (2) *Normalization*: map all times to seconds and drop duplicate caption lines; (3) *Segmentation*: apply the template in Listing 4 and retain segments whose duration lies in $[15, 120]$ s.

Listing 3: Transcripts timing mismatches

```
00:00:00 --> 00:00:**03**
hi everyone welcome to my youtube channel
00:00:**01** --> 00:00:05
about parrots. i am david. i'm here with
```

Listing 4: Transcripts segmentation output format

```
Segment [Segment Number]
Time: [Start Seconds] --> [End Seconds]
Title: [Concise Title]
Details:
   - Instructional Focus: Brief description
   - Key Steps and details:
      - [Step / Detail 1]
      - [Step / Detail 2]
      - [Step / Detail 3]
      - [Step / Detail 4]
      - [Step / Detail 5]
   - Audio Cues: Audio elements description
```

F.5 QA-pair generation

After transcripts are segmented, GPT is prompted to produce a triple consisting of (i) the question itself, (ii) a step-by-step answer, and (iii) a list of $\langle \text{videoID}, \text{start}, \text{end} \rangle$ references that support each step. Crucially, every answer must draw evidence from *multiple* videos, making cross-clip reasoning a core requirement of AVHaystacks (see Listing 5).

Listing 5: Question-Answer pair output format

```
Question 1?
Answer:
   1) step 1
   2) step 2
   3) step 3
   4) step 4
   5) step 5
References:
  1.txt 0017s > 0074s, 8.txt 0045s > 0270s,
  2.txt 0050s > 0100s, 3.txt 0110s > 0150s
```

G SFS Prompt

Below, we add the prompt used to select the key frames using the SFS algorithm. We provide step-by-step, clear instructions about the task, the reasoning process, and the expected output. Among various other prompts employed, this one produces the best results.

Figure 15: **Failure case of MAGNET.**

Salient Frame Selection Prompt

Task: Given a question and a set of key frames extracted from a video, identify the most relevant frames that best support answering the question.

Step 1: Reasoning Process Explain your selection by considering the following factors (in order of importance):
1) Presence of objects or actions explicitly mentioned in the question
2) Scenes that clearly align with the question's context
3) Visual elements directly related to the question details
4) Location or background context, even if the main object/action is not visible
5) Semantically related or typically co-occurring objects
6) Human motion or activity suggesting relevant events

Step 2: Output List the selected image indices using the format: [idx1, idx2, idx3, ...]

The objective is to select visual evidence useful for answering the question, not to answer the question itself.

H Failure case

Fig. 15 illustrates a failure case of MAGNET. Owing to the visual and auditory similarity between a violin and a cello, the retrieval module fails to accurately identify segments where both instruments are played simultaneously. As shown, the selected segment from the first video features only the cello, omitting the presence of the violin. The other two retrieved videos result from incorrect retrieval, leading to erroneous temporal grounding and highlighting a limitation of the system in distinguishing between acoustically and visually similar sources.

I Implementation Details

Training is done for 5 epochs on 4 A100 GPUs and the best checkpoint is selected for evaluation. Following the success of Low-Rank Adaptation (LoRA), we apply it with a rank of 8 and an alpha value of 32 for fine-tuning. AdamW is used as optimiser with a learning rate of 1e-4. We use a per-device batch size of 1 and a gradient accumulation step of 16. A cosine learning rate scheduler is employed with a warmup ratio of 0.05.

Dataset	MS	TA	MVL	AVR	AVD	RQA	AVQA	LC
Video Datasets								
Video-Bench [116]	✓	✗	✗	✗	✗	✗	✗	✗
EgoSchema [45]	✓	✗	✗	✗	✗	✗	✗	✓
MVBench [34]	✓	✗	✗	✗	✗	✗	✗	✓
MMBench-Video [117]	✗	✗	✗	✗	✗	✗	✗	✓
Audio-Visual Datasets								
AVOdesseyBench [118]	✓	✗	✗	✗	✗	✗	✓	✗
OmniBench [119]	✓	✗	✗	✗	✗	✗	✓	✗
LongVALE [120]	✓	✓	✗	✗	✓	✗	✓	✓
AVHBench [121]	✓	✗	✗	✗	✓	✗	✓	✗
AVHaystacks (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

Table 14: **Comparison with prior video/audio-visual benchmarks.** MS: Model-Assisted; TA: Temporal Annotation; MVL: Multi-Video Linkage; AVR: Audio-Visual fine-grained Reasoning; AVD: Audio-Visual Description; RQA: Retrieval-based QA Answering; AVQA: Audio-Visual QA; LC: Long Context, where QA context spans over 5 mins.

J More related works

Large Multimodal Models. Large Multimodal Models (LMMs) have advanced significantly in understanding and reasoning over single and multiple images [92, 93, 24, 94, 95], expanding vision-language capabilities across diverse tasks and domains [96–99]. Their strength lies in large-scale cross-modal alignment and powerful language modeling. However, LMMs still struggle with scaling to large image or video collections [82, 100] due to computational and representational challenges. Retrieval-based approaches address this by enabling efficient access, processing, and reasoning over extensive multimedia content, including video and audio.

(Audio)Video Benchmarks. We compare AVHaystacks with recent audio-visual and video QA benchmarks in Tab. 14. As shown, most existing benchmarks do not offer multi-video linked QA annotations, making AVHaystacks the first of its kind and inherently more challenging. Our data collection framework introduces a scalable, semi-automated, and richly annotated pipeline in real-world settings, providing significant advantages for future research. We hope this benchmark will inspire the community to explore and advance work in this promising direction.

Audio-Visual Learning. Powered by improved models and high-quality annotated data, audio-visual learning has advanced significantly in areas such as cross-modal generation [101–104], representation learning [105–108], multimodal large language models [109, 2, 110, 30, 57, 6, 111], and cross-modal integration [112–114]. Recent works have contributed to cross-modal generation by leveraging visual and/or language context to generate coherent, complex audio [101, 102]. The work on active audio-visual separation and embodied agents highlights the role of motion and egocentric perception in learning robust representations. These ideas extend naturally to audio-visual LLMs [115, 2], where perceptually grounded models interact with dynamic environments.

K Human Study Details

We conducted a human study involving 20 participants to evaluate the following: (i) the correctness and reliability of the samples collected in AVHaystacks, (ii) the quality of responses generated by MAGNET, and (iii) the correlation between the proposed metric STEMand human evaluation. Application details are presented in Fig. 16, and user consent information is provided in Fig. 17.

Each participant received detailed instructions outlining the goals of the study and their specific tasks. They were shown several samples obtained through our semi-automated data collection strategy and asked to rate the quality of each sample on a scale from 1 to 5. The aggregated ratings indicate high relevance and correctness, with an average score of 4.6/5 for the collected samples. Participants also evaluated the responses generated by MAGNET on the proposed AVHaystacksQA, as discussed in the main paper.

The user study protocol was approved by the Institutional Review Board (IRB). No personal information was collected, stored, or shared at any stage of the study.

USER STUDY APPLICATION

1. **Abstract:**

This study focuses on the evaluation of the quality of collected samples for a benchmark dataset, results generated by our novel framework for the project titled "MAGNET: A Multi-agent Framework for Finding Audio-Visual Needles by Reasoning over Multi-Video Haystacks" and assessing the quality of the proposed metric. We mainly use this technique for AI powered audio-visual processing applications. The evaluation is done by asking human participants to rate the quality of the samples and model responses.

2. **Subject Selection:**

   a. **Recruitment:** Hidden for anonymity

   b. **Eligibility Criteria:** Anyone with normal hearing and vision, is at least 18 years old, and is proficient in English is eligible. Participants with corrected-to-normal hearing and vision will also be eligible.

   We have two screening questions. Our screening question are 1) "Do you have normal hearing / corrected-to-normal hearing?" 2) "Are you proficient in English?". The participant who answers yes to both these questions will be allowed to continue with the survey.

   c. **Rationale:** Evaluations from people without normal hearing will bias our study results. Participants also need to be at least 18 years old and need to be proficient in English to understand the speech contents being played.

   d. **Enrollment Numbers:** 50 people max

   e. **Rationale for Enrollment Numbers**: This number gives reliable statistical results

3. **Procedures:**

   Using a headset is recommended possibly accompanied by videos and giving your preferences for each of them. The expected time to finish the procedure is 30 minutes. We are performing a survey/questionnaire and subject will only complete once. We will ask eligibility questions first and we will not collect survey data from ineligible participants. We will delete eligibility screening answers immediately for ineligible participants. You need to perform the following three tasks:

   First, the procedures involve you going over series of videos and assess the quality of the samples in the presented benchmark. Each sample has a question, a collection of videos and a response to it. Your task is to go over them carefully and rate the quality of the samples on a scale of 1-10 with 1 being the lowest. The samples are from diverse scenarios involving how-to, travelvlogs, new reading etc.

   Secondly, you need to go over model responses and rate them based on their accuracy, order, cohrence with the question asked.

   Finally, you need to evaluate the quality of our proposed metric and rate three aspects hallucination, order and missing steps.

4. **Risks:**

   There are no known risks.

5. **Benefits:**

   There are no direct benefits to participants. This study will provide researchers from audio-visual community a better understanding of how good the multi-video linked QA pipeline works and also evaluate them on various challenging cases.

6. **Confidentiality:**

   We do not collect information that can identify the participants. Any data collected will be stored on a password protected computer and will be securely wiped in 2 years from the day of creation. Only the investigators of this study have access to the data.

7. **Consent Process:**

   In our online study, we will first present screening questions before consenting to ensure that ineligible participants are not enrolled in your study. Then we present consent information to our participants, and they need to read and click a button that says "I agree" to indicate their consent and continue to our questions. We request a waiver of written consent for our online study based on following facts:
   (1) Our research only requires subjects to listen and view to videos, which involves no more than minimal risk to the subjects
   (2) We present participants with consent information electronically before the study. Subjects can choose to not continue if they do not give their consent. The waiver of written consent will not adversely affect the rights and welfare of the subjects.
   (3) The research could not practicably be carried out without the waiver or alteration because we need to collect responses from people in other regions in the ..... and will not be able to collect signatures from each subject.
   (4) We will provide our contact information during the study and encourage the subjects to contact us with any questions or concerns and they will be provided with additional pertinent information after participation.
   (5) Participant can save their signed consent form. For a fair comparison, we will not use any deception.

8. **Conflict of Interest:**

   No conflict of interest.

9. HIPAA Compliance:

Not Applicable.

10. Research Outside of the United States:

Not Applicable.

11. Research Involving Prisoners:

Not Applicable.

12. SUPPORTING DOCUMENTS

Your Initial Application must include a **completed Initial Application Part 1 (On-Line Document),** the information required in items 1-11 above, and all relevant supporting documents including: consent forms, letters sent to recruit participants, questionnaires completed by participants, and any other material that will be presented, viewed or read to human subject participants.

The consent forms in your approved IRBNet PACKAGE must be used. When creating or editing your consent form, please provide the most recent IRBNet package number at the bottom, right corner of the consent form. This ensures you are using the most "uptodate" version of the form.

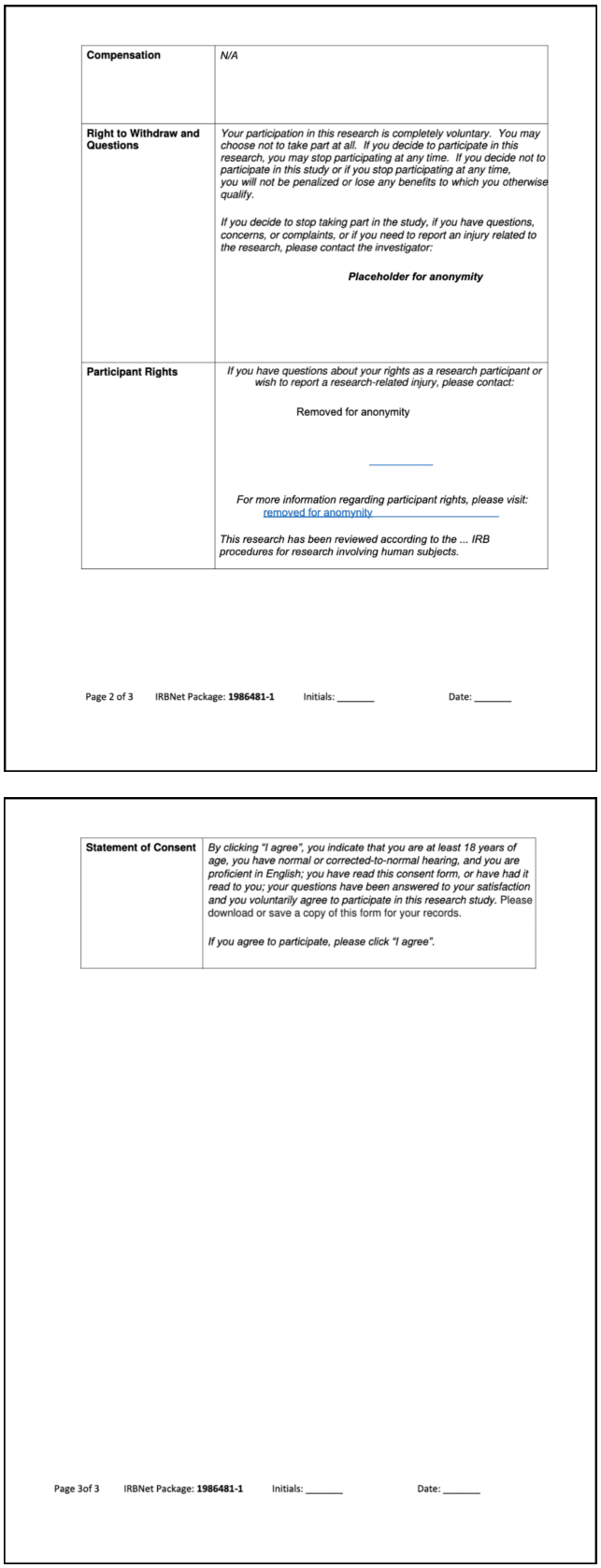
To find your IRBNet package number, go to the MY PROJECTS tab and click on the title of your project. In the PROJECT OVERVIEW page, your IRBNet package number will be listed at the top, next to your project title.

Figure 16: User study guidelines.

Initials: _____ Date: _____

CONSENT TO PARTICIPATE

Project Title	*A Multi-agent Framework for Finding Audio-Visual Needles by Reasoning over Multi-Video Haystacks*
Purpose of the Study	*This research is being conducted by ..... at ..... We are inviting you to participate in this research project because you have normal or corrected-to-normal hearing and visual abilities. The purpose of this research project is to evaluate the quality of our dataset, model response and correctness of our proposed metric.*
Procedures	*The procedures involve going over samples collected in the dataset and validate the correctness of the annotation. You will be presented with set of questions along with some videos and some responses relevant for them. You have to rate them based on their relevance, order and other aspects. You also need to go over the model response and evaluate its accuracy and provide a rating. Finally, you neede to rate the goodness of a metric by validating how good the score reported by the metric alignes with what you think is the correct response. The expected time to finish the procedure is 30 minutes. We will ask eligibility questions first and we will not collect survey data from ineligible participants. We will delete eligibility screening answers immediately for ineligible participants.*
Potential Risks and Discomforts	*There are no known risks from participating in this research study.*
Potential Benefits	*There are no direct benefits from participating in this research. We hope that, in the future, other people might benefit from this study through improved understanding of how to generate more realistic virtual sound for speech related applications.*
Confidentiality	*Any potential loss of confidentiality will be minimized by storing data in a password protected computer. No identifiable information will be collected and* the researchers will be the only individuals to have access *to the data.* *If we write a report or article about this research project, your identity will be protected. Your information may be shared with representatives of the ... or ... if you or someone else is in danger or if we are required to do so by law.*



Compensation	*N/A*
Right to Withdraw and Questions	*Your participation in this research is completely voluntary. You may choose not to take part at all. If you decide to participate in this research, you may stop participating at any time. If you decide not to participate in this study or if you stop participating at any time, you will not be penalized or lose any benefits to which you otherwise qualify.* *If you decide to stop taking part in the study, if you have questions, concerns, or complaints, or if you need to report an injury related to the research, please contact the investigator:* ***Placeholder for anonymity***
Participant Rights	*If you have questions about your rights as a research participant or wish to report a research-related injury, please contact:* Removed for anonymity *For more information regarding participant rights, please visit:* removed for anomynity *This research has been reviewed according to the ... IRB procedures for research involving human subjects.*

Page 2 of 3 IRBNet Package: **1986481-1** Initials: _____ Date: _____

Statement of Consent	*By clicking "I agree", you indicate that you are at least 18 years of age, you have normal or corrected-to-normal hearing, and you are proficient in English; you have read this consent form, or have had it read to you; your questions have been answered to your satisfaction and you voluntarily agree to participate in this research study.* Please download or save a copy of this form for your records. *If you agree to participate, please click "I agree".*

Page 3of 3 IRBNet Package: **1986481-1** Initials: _____ Date: _____

Figure 17: User consent application.