AFRL at TREC 2025: Zero-Shot Human-Free Video Search with Caption Expansion

Andrew Young[1], Emily Conway[2], Jeremy Gwinnup[2]
[1]Wright State University, [2]Air Force Research Laboratory
young.565@wright.edu, {emily.conway, jeremy.gwinnup.1}@us.af.mil

Abstract

We describe the Air Force Research Laboratory's submission to the TREC 2025 Ad-hoc Video Search (AVS) task. Our approach addresses the challenge of indexing and searching video content without human-generated metadata by employing modern multimodal large language models for caption generation. We upgrade from traditional short-caption baselines generating 7-word descriptions to state-of-the-art models producing up to 100-word descriptions with inter-frame context awareness. Our system integrates three components: caption expansion for richer semantic coverage, context-aware learning through weighted concept banks and unlikelihood training for better embeddings, and query decomposition with question-answering models for precise re-ranking. While our official submission encountered critical integration failures resulting in poor performance, we present our methodology, analyze the failures, and demonstrate the viability of our core approach through preliminary validation results.

1 Introduction

Modern video platforms like YouTube rely heavily on human-created titles, descriptions, content tags, and watch patterns to enable meaningful video search and discovery. However, these systems face significant challenges when such human-generated metadata is unavailable or incomplete. This limitation is particularly problematic for applications in security, intelligence, and educational video analysis where automated indexing is essential.

The Text REtrieval Conference (TREC) Video Retrieval Evaluation has provided a benchmark for evaluating video search systems for over two decades. The TREC 2025 Ad-hoc Video Search (AVS) task (Awad, 2026) challenges systems to retrieve relevant video segments from a large collection given only textual queries, without relying on pre-existing human annotations.

Our approach focuses on a the questions: Can machines understand video content well enough to generate complete and accurate text descriptions without any human input and can these descriptions be comprehensive enough to support arbitrary keyword searches?

We present a zero-shot video search system that leverages recent advances in multimodal large language models (LLMs) to generate rich, detailed video captions. To address the central question of whether machine-generated descriptions can support arbitrary keyword search, we make three technical contributions and one analytical contribution.

First, we systematically evaluate modern caption generation models for video retrieval, demonstrating a 10-fold expansion from traditional 7-word captions (BLIP baseline) to 72-102 word descriptions (InternVL3-8B) that capture richer visual detail, temporal context, and inter-frame relationships. Second, we integrate context-aware learning strategies—weighted concept banks that emphasize rare visual features and unlikelihood training that distinguishes mutually exclusive concepts—to improve embedding quality beyond generic pre-trained representations. Third, we develop a re-ranking framework that decomposes complex multi-attribute queries into simple yes/no questions, enabling more precise constraint satisfaction through lightweight question-answering models.

While our official submission encountered critical integration failures resulting in poor performance, our fourth contribution is a candid analysis of these failures alongside preliminary validation results that demonstrate the viability of the core approach when properly implemented.

2 Task Description

The TREC 2025 AVS task requires systems to search the V3C2 dataset (Awad et al., 2020), a large-scale collection of 9,760 videos (1.6 TB, approximately 1,300 hours) segmented into 1,425,454 shots with a mean video duration of 8 minutes. The dataset consists of creative commons videos from Vimeo, representing diverse visual content. Given 20 textual queries, systems must return up to 1,000 ranked shots per query.

Our submission was classified as **Class F** (fully automatic, no human intervention), **Training Type D** (any non-V3C2 training data allowed), **Task M** (main ad-hoc search task), and **Novelty C** (common/conventional approach). Evaluation uses Mean extended inferred Average Precision (xinfAP) (Yilmaz et al., 2008), calculated from sampled relevance judgments with 100% sampling for positions 1-300 and 25% sampling for positions 301-1000.

3 Related Work

Video retrieval has evolved from early concept-based approaches to modern embedding-based methods. Traditional systems relied on pre-defined concept detectors (Snoek et al., 2006), while recent work leverages vision-language models trained on large-scale image-text pairs (Radford et al., 2021).

BLIP (Li et al., 2022) introduced a bootstrapping approach for vision-language pre-training, generating short image captions. Recent multimodal LLMs like CogVLM (Wang et al., 2023) and InternVL (Chen et al., 2024) have pushed caption quality and length significantly higher through better architectural designs and training strategies.

For video-specific modeling, works like VideoLLaMA (Zhang et al., 2023) and Video-ChatGPT (Maaz et al., 2023) have explored temporal context modeling. Our work builds on these advances by systematically evaluating modern captioning models for the video retrieval task.

Context-aware learning in retrieval has been explored through various techniques. Weighted concept banks (Wu and Ngo, 2020) assign different importance to rare versus common concepts. Unlikelihood training (Welleck et al., 2020) helps models distinguish between semantically similar but mutually exclusive concepts.

Building on these advances in vision-language modeling and context-aware learning, we designed a four-stage video search pipeline that progressively refines retrieval results from caption generation through context-aware embedding to query-specific re-ranking.

4 System Overview

Our video search pipeline consists of four main stages:

4.1 Preprocessing

We extract keyframes from each video segment in both the training data (for concept learning) and the test collection (V3C2). Keyframes are sampled at regular intervals to capture visual diversity while maintaining computational feasibility.

4.2 Caption Generation

Each keyframe is processed through a multimodal LLM to generate detailed text descriptions. Unlike traditional single-frame captioning (e.g., BLIP generating 7-word descriptions), we employ models capable of understanding inter-frame relationships and generating longer, more comprehensive captions.

For the V3C2 dataset, we use ChatGPT-4o (OpenAI, 2024) for concept expansion on training captions to enrich the vocabulary and improve coverage of possible search terms.

4.3 Feature Embedding

Both video captions and text queries are embedded into a shared high-dimensional feature space using Vision Transformer (ViT) pre-trained embeddings. This enables semantic comparison between queries and video content.

We train context-aware embeddings using two complementary techniques. Our **weighted concept bank** assigns higher weights to rare concepts (e.g., “floral”) versus common ones (e.g., “dress”) in the MSRVTT training dataset (Xu et al., 2016), ensuring that distinctive visual features contribute more strongly to similarity scores. Complementing this, **unlikelihood learning** ensures mutually exclusive concepts (e.g., “man” vs “woman”) remain distinguishable in the feature space despite their semantic similarity in standard embeddings.

4.4 Retrieval and Re-Ranking

Initial retrieval uses cosine similarity between query and video embeddings to select the top 1,500 candidates.

The re-ranking stage decomposes complex queries into multiple simple questions that can be answered by lightweight question-answering models. For example, the query "Find a bald man with glasses" generates:

1. Is this a person?
2. Is this person a man?
3. Is this person bald?
4. Is this person wearing glasses?

ChatGPT-4o (OpenAI, 2024) is used for both question generation and question answering, with the final ranking determined by the number of affirmative answers.

Having outlined the overall pipeline architecture, we now detail the caption generation component, which forms the foundation of our approach by transforming visual content into searchable text.

5 Caption Generation Models

We evaluated four caption generation approaches:

5.1 BLIP (Baseline)

BLIP (Li et al., 2022) represents the 2022 state-of-the-art for image captioning. It generates 3 independent captions per keyframe, averaging 7 words each. However, it processes single images without inter-frame context, limiting its ability to capture temporal dynamics.

5.2 NVIDIA Describe Anything

NVIDIA's Describe Anything model (Lian et al., 2025) generates longer single captions averaging 57 words, achieving 67.3% accuracy on the DLC benchmark. While improving caption length, it still lacks explicit temporal modeling.

5.3 CogVLM2-Llama3-Caption

CogVLM2 (Wang et al., 2023) produces single captions averaging 105 words with 67.7% accuracy on MVBench, a video understanding benchmark. The model begins to capture some inter-frame relationships.

5.4 InternVL3-8B (Selected)

We selected InternVL3-8B (Chen et al., 2024) as our primary caption generation model for several reasons. It achieves state-of-the-art 75.4% accuracy on MVBench, a comprehensive video understanding benchmark, while providing flexible output options: either 3 captions at approximately 72 words each, or a single 102-word caption. Unlike earlier models, it explicitly models temporal context and inter-frame relationships, capturing motion and event progression within its 8K token context window. This context size supports approximately 31 keyframes, corresponding to 1-2 minutes of video content—a significant improvement over single-frame approaches. The longer, more detailed captions increase the likelihood that relevant attributes and objects appear in the generated text, enabling better keyword matching.

6 Context-Aware Learning

6.1 Weighted Concept Bank

In natural video collections, some concepts appear much more frequently than others. A simple bag-of-words approach treats all concepts equally, potentially drowning out rare but distinctive features.

We implement a weighted concept bank inspired by TF-IDF weighting, where concepts are weighted inversely proportional to their frequency in the MSRVTT (Xu et al., 2016) training corpus. For a query like "Find a floral dress", the rare term "floral" receives high weight while the common term "dress" receives lower weight, ensuring that the distinctive pattern feature dominates the retrieval score.

6.2 Unlikelihood Learning

Words that are semantically similar in the embedding space may be mutually exclusive in practice. For example, "man" and "woman" have similar embeddings (both refer to people, appear in similar contexts), but a video typically shows one or the other, not both.

Unlikelihood training (Welleck et al., 2020) explicitly pushes apart embeddings of mutually exclusive concepts. For a query "man with glasses",

this prevents videos of "woman with glasses" from ranking too highly despite the shared "with glasses" similarity.

6.3 Learning Rate Fusion

We combine multiple learning signals:

1. Interpretable concept similarities (weighted concept bank)
2. Uninterpretable feature embeddings (ViT pre-trained representations)
3. Task-specific relevance (from training queries and relevance judgments)

These are fused through a multi-task learning objective to learn richer feature representations that capture both semantic meaning and task-specific patterns.

Together, importance weighting, mutual exclusion enforcement, and multi-signal fusion create embeddings that go beyond generic pre-trained representations by incorporating both semantic knowledge and task-specific patterns from video retrieval training data.

While context-aware embeddings improve semantic matching, complex queries often require explicit constraint checking beyond similarity scoring, motivating our query decomposition approach.

7 Re-Ranking Strategy

Complex queries often combine multiple attributes or constraints. Simple embedding similarity may retrieve videos that match some but not all requirements.

Our re-ranking approach decomposes queries into atomic yes/no questions. For example, the query "Find a bald man with glasses" generates questions such as "Is this a person?", "Is this person a man?", "Is this person bald?", and "Is this person wearing glasses?" For the top 1,500 candidates from initial retrieval, we use ChatGPT-4o to answer each question for each video. Videos are re-ranked by the number of affirmative answers, with ties broken by the original similarity score.

This approach offers three key benefits: it reduces false positives from partial matches by explicitly checking all constraints; it enables use of simpler, faster QA models rather than complex joint reasoning models; and it provides interpretable reasoning for rankings through the question-answer chain.

However, the approach faces practical challenges. It requires automatic query parsing to identify adjectives and nouns for question generation, adds significant processing time proportional to the number of questions and candidates, and may overconstrain results when queries are ambiguous or use figurative language.

8 Experimental Setup

8.1 Training Data

Due to IT policy download restrictions at our institution, we were unable to obtain the complete V3C1 development dataset. This severely limited our ability to tune hyperparameters on representative data, validate our re-ranking strategies, and debug system integration issues. We instead relied on the MSRVTT dataset (Xu et al., 2016) for concept bank construction and general model training, which provides video captioning data but not retrieval relevance judgments similar to TREC.

8.2 Implementation Details

For caption generation, we configured InternVL3-8B to produce 3 captions per video at approximately 72 words each, maximizing semantic coverage while managing computational cost. These captions were embedded using ViT-L/14 pre-trained on CLIP, leveraging its strong vision-language alignment. Our retrieval pipeline first selected the top 1,500 candidates by cosine similarity, then applied re-ranking using ChatGPT-4o (OpenAI, 2024) for both question generation and answering. The final submission returned the top 1,000 results per query as required by the task specification.

8.3 System Configuration Issues

Our official submission (Run ID: F_M_C_D_AFRL_25_5) suffered from system misconfiguration issues that we identified only after the submission deadline. Specifically, re-ranking components were not fully integrated into the submission pipeline, audio information from videos was not incorporated into captions, and some preprocessing steps were not properly validated.

Metric	Score
Mean xinfAP	0.000
Inferred AP @ 0.0 Recall	0.037
Inferred Precision @ 5	0.020
Inferred Precision @ 10	0.010
True Shots Returned	184 / 30,939

Table 1: Official submission results for run F_M_C_D_AFRL_25_5. Mean xinfAP of 0.000 indicates critical system failure.

Despite these implementation challenges, we submitted our system to the official evaluation. The results, detailed below, reflect both the promise of our methodology and the consequences of incomplete system integration.

9 Results

Table 1 shows the performance of our official submission.

9.1 Performance Analysis

Our submission achieved a Mean xinfAP of 0.000, indicating critical system failure. While we successfully retrieved 184 true relevant shots out of 30,939 total, they were not ranked highly enough to contribute to the primary metric.

These results indicate catastrophic system failure rather than merely poor performance. While we successfully retrieved 184 relevant shots out of 30,939 total relevant shots in the ground truth, these were ranked so poorly that they contributed negligibly to the primary metric. This suggests that our core retrieval mechanism found some relevant content, but the ranking mechanism was fundamentally broken.

Figure 1 shows the complete TREC evaluation results, with the precision-recall curve demonstrating near-zero performance across all recall levels. The inferred precision at different depths (P@5: 0.020, P@10: 0.010) shows some ability to find relevant content, but poor ranking quality.

For context, typical TREC AVS systems achieve Recall@1 around 10%, with state-of-the-art reaching 29.1% (Dong et al., 2021). Our results fall far below this baseline.

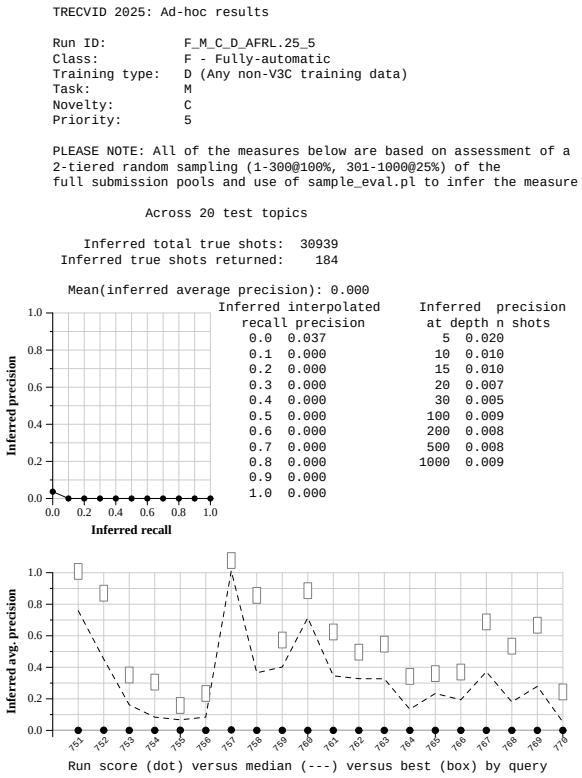

Figure 1: Official TREC 2025 evaluation results for run F_M_C_D_AFRL_25_5. The precision-recall curve (top left) shows near-zero performance across all recall levels. The score distribution (bottom) demonstrates consistently poor performance across all 20 test queries compared to median and best submissions.

9.2 Failure Analysis

We identified several critical issues:

Incomplete Training Data: The absence of V3C1 development data meant we could not validate our approach on similar content or tune system parameters effectively.

System Integration: The re-ranking component, while developed and tested in isolation, was not properly integrated into the final submission pipeline. Our submitted results likely reflect only the initial retrieval stage.

Audio Modality: Videos often contain crucial information in audio (speech, environmental sounds) that was completely ignored by our vision-only captioning approach.

Video Length Constraints: InternVL3-8B's 8K token limit restricts processing to approximately 1-2 minutes of video. Longer videos were truncated, potentially missing relevant content.

To understand whether our approach has merit despite the official failure, we examined system performance on a custom validation dataset where

integration was properly completed.

9.3 Preliminary Validation Results

On a custom Ukrainian news broadcast dataset (3,400 videos from February-April 2022), our system showed more promising results before the integration issues. For the query "military planes", we successfully retrieved relevant content, suggesting the core methodology has merit when properly implemented.

Similarly, for the query "a bald man with glasses", preliminary results included reasonable matches:

1. A man wearing glasses and a hat walking through a forest
2. A man with glasses wearing a bicycle helmet
3. A man with glasses and hair answering the phone

These results indicate that caption quality and retrieval logic were sound, but system integration failures prevented this capability from manifesting in the official submission.

These failures, while disappointing, provide valuable insights into system requirements for production video retrieval. We now discuss improvements that address each identified limitation.

10 Discussion and Future Work

Our failure analysis identified four categories of improvement: enhancing caption generation to handle longer videos and multiple modalities, advancing retrieval methods beyond simple similarity scoring, optimizing the re-ranking stage for computational efficiency, and strengthening system engineering practices to prevent integration failures.

10.1 Caption Generation Improvements

Longer Video Support: Extending caption generation beyond 1-2 minutes is critical. Potential approaches include:

- Hierarchical captioning (scene-level then video-level)
- Sliding window approaches with caption fusion
- More efficient video encoders supporting longer contexts

SEO-Style Indexing: Rather than generating grammatically correct sentences, we could generate keyword lists optimized for search coverage, similar to search engine optimization (SEO) practices.

Audio Integration: Incorporating automatic speech recognition (ASR) and audio event detection would provide crucial complementary information to visual captions.

Multi-Model Fusion: Combining outputs from multiple captioning models could improve coverage and reduce individual model biases.

10.2 Retrieval Method Enhancements

Advanced Similarity Metrics: Moving beyond cosine similarity to more sophisticated distance metrics in the feature space could improve ranking quality.

Multi-Word Tokenization: Grouping related words together (e.g., "wedding dress", "military aircraft") would enable phrase-aware matching beyond bag-of-words approaches.

Temporal-Aware Ranking: Shots from the same video that are temporally close often have similar content. De-duplicating or diversifying results could improve user experience.

Stochastic Embeddings: Rather than deterministic embeddings, modeling uncertainty in the embedding space could better capture ambiguity in visual content.

10.3 Re-Ranking Optimization

Automated Query Parsing: Implementing robust adjective-noun parsing would enable automatic question generation for arbitrary queries.

Model Selection: Investigating smaller, specialized QA models could reduce computational cost while maintaining effectiveness.

Model Distillation: Training a student model to mimic the re-ranking behavior could enable parallel processing and faster inference.

Question Generation Strategies: Optimizing the number and type of questions generated could balance precision gains against computational cost.

10.4 System Engineering

Comprehensive Testing: Establishing thorough testing protocols and validation datasets is essential to catch integration issues before submission.

Evaluation Metrics: Implementing standard metrics (R@1, R@5, R@10, Mean Average Precision) in our development pipeline would enable rapid iteration.

Data Pipeline Validation: Ensuring complete training data availability and proper data flow through all system components is critical.

11 Conclusion

We presented a zero-shot video search system for the TREC 2025 AVS task that leverages modern multimodal LLMs for caption generation. Our approach combines caption expansion (from 7-word to 72-word descriptions), context-aware learning through weighted concept banks and unlikelihood training, and query decomposition for re-ranking.

While our official submission encountered critical system integration issues resulting in poor performance, our methodology shows promise in preliminary validation. The core insight remains valid: longer, more detailed captions generated by state-of-the-art multimodal models can provide better coverage for arbitrary keyword searches.

Our experience highlights four critical lessons for video retrieval systems. First, complete and representative training data is essential, our lack of a complete V3C1 dataset severely limited validation and tuning. Second, rigorous system integration testing must be prioritized over component-level development, as our re-ranking component worked in isolation but failed in the submission pipeline. Third, multi-modal approaches that combine vision and audio are necessary for comprehensive video understanding, as visual captions alone miss crucial information. Finally, comprehensive evaluation throughout development, not just at submission time, is critical for catching integration failures early. Future work will focus on addressing the identified limitations, particularly proper system integration, longer video support, and audio incorporation, to realize the potential of this approach.

Limitations

Our work has several significant limitations:

Dataset Constraints: Due to institutional download restrictions, we could not obtain the complete V3C1 development dataset, severely limiting our ability to validate and tune our system on representative data.

Video Length Limitations: Current caption generation is limited to 1-2 minute segments due to the 8K token context window of InternVL3-8B. Longer videos require truncation or segmentation, potentially missing relevant content.

Vision-Only Processing: We do not incorporate audio information (speech, sounds, music), which often contains crucial information for video understanding and retrieval.

Computational Cost: Generating detailed captions for large video collections using LLMs is computationally expensive, limiting practical scalability.

System Integration: Our official submission suffered from integration failures, with the re-ranking component not properly incorporated into the final pipeline.

English-Only: Our system currently only processes English text queries and generates English captions, limiting applicability to multilingual video collections.

Query Complexity: The re-ranking approach requires manual query parsing or sophisticated NLP to decompose complex queries, which may not generalize to all query types.

Acknowledgments

We thank the TREC 2025 organizing committee and NIST for providing the evaluation framework and data. Andrew Young's graduate studies are supported by the AFRL Scholars Program.

References

George Awad. 2026. Trec 2025 ad-hoc video search (avs) track overview.

George Awad, Asad A Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Joy Zhang, Eliot Godard, Lukas Diduch, et al. 2020. V3c–a research video collec-

tion. In *International Conference on Multimedia Modeling*, pages 349–360. Springer.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. 2021. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4065–4080.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Long Lian, Yifan Ding, Yunhao Ge, Sifei Liu, Hanzi Mao, Boyi Li, Marco Pavone, Ming-Yu Liu, Trevor Darrell, Adam Yala, and Yin Cui. 2025. Describe anything: Detailed localized image and video captioning. *arXiv preprint arXiv:2504.16072*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

OpenAI. 2024. Chatgpt (gpt-4o). https://openai.com/chatgpt. Accessed: February 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Cees GM Snoek, Marcel Worring, Jan C Van Gemert, Jan-Mark Geusebroek, and Arnold WM Smeulders. 2006. The mediamill trecvid 2006 semantic video search engine. In *TRECVID workshop*, volume 70.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*.

Jiaxin Wu and Chong-Wah Ngo. 2020. Interpretable embedding for ad-hoc video search. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3357–3366.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. 2008. A simple and efficient sampling method for estimating ap and ndcg. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 603–610.

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