MAVIS: Multi-Agent Video Retrieval via Structured Video Understanding

Jie Zhang[1] , **Qilang Ye**[2], **Hao Zhou**[1,3], **Haochen Liang**[4], **Fei Luo**[1*]
[1]School of Computing and Information Technology, Great Bay University
[2]College of Computer Science, Nankai University
[3]Tsinghua Shenzhen International Graduate School, Tsinghua University
[4]Graduate School of Information Science and Technology, The University of Tokyo
jz@stu.cqut.edu.cn, luofei@gbu.edu.cn

Abstract

The dominant paradigm in video retrieval relies on embedding-based full-corpus scanning, which suffers from inherent computational inefficiency and the semantic asymmetry between information-dense videos and sparse textual queries. To bridge this gap, we introduce **MAVIS**, a novel multi-agent framework that rethinks retrieval as cooperative reasoning rather than brute-force search. MAVIS first bridges the granularity mismatch by parsing raw videos into a **Structured Semantic Library**, enabling explicit attribute-level indexing. During retrieval, a planner decomposes complex user intents into atomic sub-tasks, dispatching specialized agents to independently nominate candidates. Crucially, MAVIS employs a **Logic-aware Debate** mechanism with a strict veto protocol, where agents collaboratively prune logical mismatches to identify a compact set of "controversial" candidates for fine-grained verification. This agentic workflow effectively bypasses the inefficiency of full-library traversal. Extensive experiments on MSR-VTT, MSVD, and ActivityNet demonstrate that MAVIS achieves competitive performance without task-specific fine-tuning, offering a scalable and interpretable alternative to traditional dual-encoder approaches.

1 Introduction

The exponential growth of video content on the web has precipitated an urgent demand for intelligent retrieval systems capable of navigating massive multimedia archives (Wang et al., 2025a). While recent years have witnessed significant strides in video retrieval, the dominant paradigm remains *embedding-based full-matching*, where every query is compared against the holistic feature representations of all videos in the corpus (Wang et al., 2024; Tian et al., 2024; Tang et al., 2025).

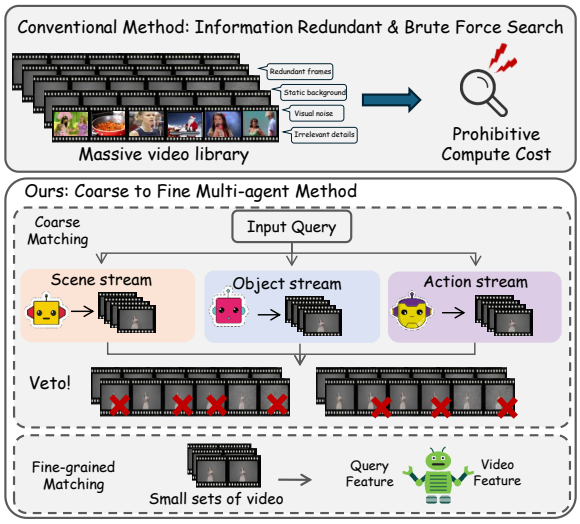

Figure 1: Overview of MAVIS. Specialized agents tackle distinct dimensions to prune the search space through independent proposal and collaborative debate. This process identifies a minimal set of high-quality candidates for final fine-grained matching, ensuring both high precision and computational efficiency.

This brute-force approach encounters a fundamental bottleneck: **semantic asymmetry and computational redundancy**. User queries are typically compact and intent-specific (e.g., "a red car turning left"), whereas videos are information-dense streams containing complex, multifaceted semantics. Relying solely on global embeddings for retrieval across the entire corpus is inherently prone to noise (Ma et al., 2022; Tian et al., 2024); the subtle visual cues required by the query are often statistically drowned out by the vast number of irrelevant distractors and the dominant background features of the video itself. Moreover, exhaustively scanning the entire library for every query incurs prohibitive computational costs. Consequently, there is a pressing need for a paradigm shift from "brute-force scanning" to "structured reasoning."

Despite these evident bottlenecks, the field has

*Corresponding Author

Findings of the Association for Computational Linguistics: ACL 2026, pages 21751–21764

yet to fundamentally challenge the exhaustive scanning paradigm. Mainstream solutions (Luo et al., 2022; Wu et al., 2023; Yang et al., 2024), predominantly driven by Vision-Language Models (VLMs) (Radford et al., 2021; Li et al., 2023), focus on optimizing joint embedding spaces to improve matching metrics. However, they inherently adhere to the $\mathcal{O}(N)$ brute-force search complexity. Moreover, achieving competitive performance with these methods typically relies on extensive task-specific fine-tuning, which limits their zero-shot generalization and applicability in open-world scenarios (Wang et al., 2023a; Wasim et al., 2023; Ren et al., 2024). Recently, emerging research has begun to explore Multimodal Large Language Models (MLLMs) (Ye et al., 2026b, 2025) for deeper reasoning. While these generative approaches offer a compelling training-free solution with superior semantic understanding, applying them to large-scale retrieval exacerbates the efficiency-accuracy dilemma: running heavy reasoners on a full corpus is computationally prohibitive (Li et al., 2024; Yu et al., 2023). Furthermore, methods that rely on a single agent to parse content often suffer from *Tunnel Vision*, where a solitary perspective overlooks subtle visual cues or hallucinates events due to cognitive overload (Du et al., 2023). This poses a fundamental challenge: *Can we synthesize the zero-shot reasoning capability of generative models with the scalability of retrieval systems, effectively bypassing the computational bottleneck of full-corpus traversal?*

To this end, we introduce **MAVIS**, a novel **M**ulti-**A**gent framework for **VI**deo **S**earch that redefines retrieval as a cooperative reasoning process. Instead of a mechanical brute-force scan, MAVIS operates like a team of human specialists tackling a complex investigation. To enable efficient lookup, we first employ a *Description Agent* to offload visual perception into a structured **Semantic Library**, effectively transforming raw pixels into a queryable textual index. Upon receiving a user query, a *Planner* first decomposes the complex intent into atomic semantic components (e.g., *Action*, *Object*, *Scene*). Guided by this breakdown, MAVIS dynamically assembles a team of specialized agents, activating a dedicated "expert" only for the semantic perspectives explicitly present in the query. In the retrieval phase, these agents work independently to nominate potential candidates based on their assigned focus. Crucially, to ensure precision, they convene for a **Logic-aware Debate**: employing a strict veto protocol, the agents act as peer reviewers for one another, collectively rejecting any candidate that logically contradicts a specific expert's view. This allows MAVIS to identify a compact set of *"controversial"* candidates for final fine-grained verification, effectively realizing the philosophy of "Inspect less, but understand more."

In summary, our main contributions are:

- We propose a **Structured Video Understanding** paradigm that explicitly parses videos into a semantic library, effectively bridges the granularity gap between dense video content and sparse textual queries.
- We design a **Collaborative Multi-Agent Framework** that dynamically assembles specialist agents based on query intent. We introduce a novel **Logic-aware Debate** where agents leverage a veto protocol to rigorously filter logical mismatches, significantly pruning the search space while preserving hard positives.
- Extensive experiments on MSR-VTT (Xu et al., 2016), MSVD (Guadarrama et al., 2013), and ActivityNet (Krishna et al., 2017) demonstrate that MAVIS turns generic pre-trained models into strong video retrievers, outperforming state-of-the-art fine-tuning methods.

2 Related Work

2.1 Text-to-video Retrieval

Video retrieval faces significant challenges due to the modality gap between visual content and natural language (Dong et al., 2022; Zhang et al., 2025, 2026). Traditional methods typically prioritize either retrieval accuracy through fine-grained cross-modal interaction (Jin et al., 2023; Wu et al., 2025), or inference efficiency by utilizing dual-encoder architectures (Liu et al., 2025; Fang et al., 2024). The emergence of large-scale Vision-Language Models (VLMs) such as CLIP (Radford et al., 2021) has revolutionized video retrieval by enabling effective cross-modal alignment (Tian et al., 2024; Liu et al., 2025; Yang et al., 2024). However, deriving optimal performance typically requires extensive fine-tuning, inevitably escalating the computational overhead. Recent research has also explored training-free approaches that leverage pre-

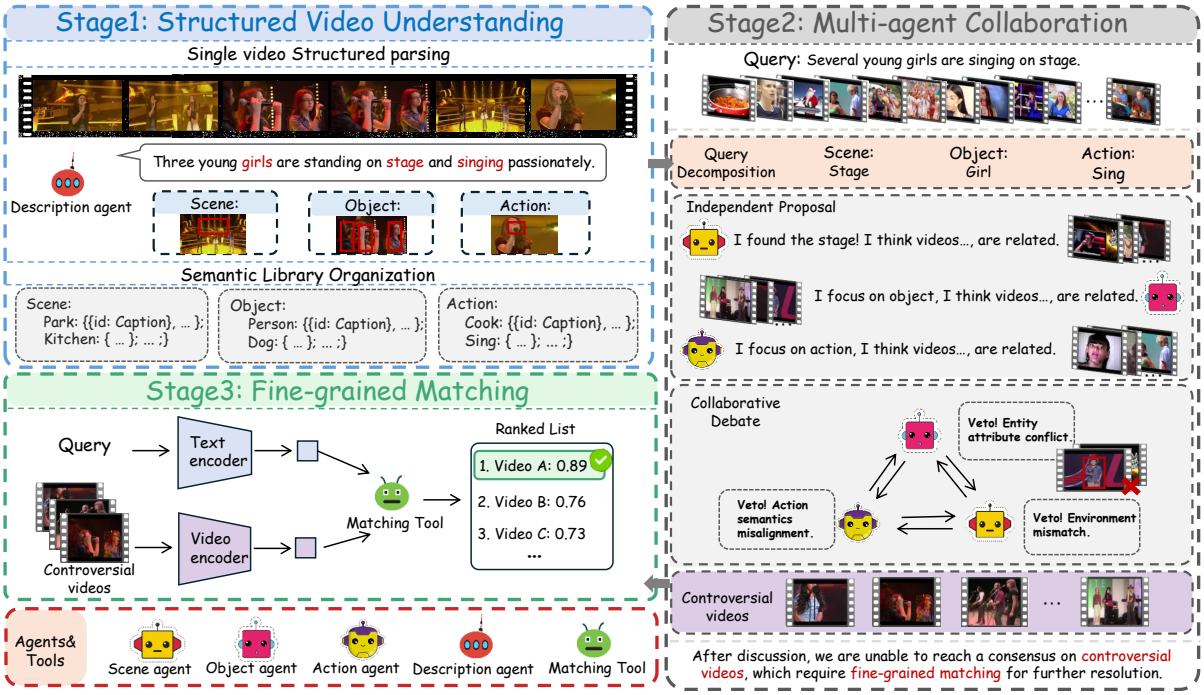

Figure 2: Overview of MAVIS. The pipeline consists of three progressive stages: (1) Structured Video Understanding: The Description Agent parses raw videos into structured semantic components (Scene, Object, Action) to construct a semantic library. (2) Multi-agent Collaboration: Given a user query, a planner decomposes the query. Role-specific agents independently propose candidates and engage in a collaborative debate to filter out mismatches. (3) Fine-grained Matching: For "controversial videos" where consensus is not reached, the Matching Tool utilizes fine-grained visual-text alignment to generate the final ranked list.

trained models without fine-tuning. Systems like Merlin (Han et al., 2024) rely on multi-round interactions and iterative reasoning, enabling effective video-query matching without the need for model retraining. However, it still requires traversing the entire video library, relying on a single model through multiple rounds of interaction. In contrast to these approaches, MAVIS avoids full-library traversal and leveraging multi-agent collaboration, leading to more efficient and accurate retrieval.

2.2 Agent and tool use

Recent advancements in large language models (LLMs) and multimodal LLMs (MLLMs) have significantly enhanced their reasoning and planning capabilities, driving the development of autonomous agents (Zhang et al., 2024; Luo et al., 2024; Ye et al., 2026a, 2024). Agent-based methods are effective at decomposing complex tasks (Yao et al.; Shen et al., 2023), and the multi-agent paradigm has achieved significant success in various tasks, such as long video understanding (Chen et al., 2025; Yang et al., 2025), where agent collaboration has substantially improved task comprehension and performance. Furthermore, by integrating external tools, these models have bridged the gap between general-purpose reasoning and real-world execution (Fan et al., 2024; Zhu et al., 2026b,a). However, existing frameworks typically rely on heavy iterative reasoning or analyze videos individually, making them inefficient for retrieving targets from massive libraries. To Bridge this gap, we introduce MAVIS, a framework that repurposes multi-agent collaboration for efficient retrieval. MAVIS enhances retrieval accuracy specifically by leveraging agent consensus, while avoiding the computational burden of full-library traversal and model fine-tuning.

3 Method

3.1 Overview

Given a natural language query q and a video corpus $\mathcal{V} = \{v_i\}_{i=1}^{N}$, the goal of MAVIS is to retrieve the top-k most relevant videos. Unlike traditional dual-encoder approaches, MAVIS adopts a *coarse-to-fine* agentic workflow. The framework consists of three phases: (1) **Structured Video Understanding**, which transforms videos into se-

mantic attributes; (2) **Collaborative Multi-Agent Pruning**, where specialized agents actively filter the corpus to identify a small set of "controversial" candidates; and (3) **Fine-grained Matching**, which performs fine-grained verification only on the survivors. The overall architecture is illustrated in Figure 2.

3.2 Structured Video Understanding

Conventional video-text retrieval paradigms typically align holistic video and text embeddings within a joint latent space. However, this global alignment is inherently susceptible to **semantic asymmetry**. As illustrated in Fig. 1, a raw video naturally encapsulates exhaustive details, including concurrent entities, background events, and atmospheric nuances. In contrast, user queries are typically concise and intent-focused . This **granularity mismatch** allows extraneous visual details to act as noise in the embedding space, diluting the matching score for the core intent. To address this, MAVIS introduces a structured video understanding process. Instead of ingesting noisy raw features, we leverage the visual reasoning capability of multimodal large language models (MLLMs) to actively *filter* video content into explicit semantic attributes and concise descriptions. This process comprises two phases: *Single Video Structured Parsing* and *Semantic Library Organization*.

Single Video Structured Parsing. For each video v_i, we employ a pre-trained MLLM with a specialized prompt designed to suppress irrelevant details and extract only salient events. Specifically, the model is instructed to directly output a concise caption c_i that mirrors the brevity of search queries, along with a structured semantic tuple:

$$\mathbf{s}_i = (s_i^{\text{SCN}}, s_i^{\text{OBJ}}, s_i^{\text{ACT}}) \tag{1}$$

where each component contains standardized keywords (e.g., $s_i^{\text{OBJ}} = \{\text{cartoon character}\}$). By bypassing the generation of verbose descriptions, MAVIS ensures that the stored representation c_i is semantically aligned with the granularity of potential text queries, minimizing the risk of length-induced mismatch.

Semantic Library Organization. We reorganize the parsed corpus into a structured semantic library $\mathcal{L}$. Given the open-vocabulary nature of MLLM outputs, raw tags often exhibit linguistic variations (e.g., *"chatting"*, *"communicating"*). To ensure retrieval robustness, we leverage the MLLM to normalize raw visual observations into a unified set of **Canonical Concepts** $\mathcal{K}$. During the parsing phase, the model is instructed to consolidate diverse descriptive terms into standardized keywords. This creates a coherent vocabulary $\mathcal{K}$ dynamically, ensuring that semantically identical content is indexed under the same key.

We structure the library into three domain-specific sub-libraries: $\mathcal{L}_{\text{SCN}}$ (Scene), $\mathcal{L}_{\text{OBJ}}$ (Object), and $\mathcal{L}_{\text{ACT}}$ (Action). Within each sub-library, videos are organized under specific semantic categories. Formally, the sub-library $\mathcal{L}_d$ functions as a map where each key k represents a specific concept, pointing to all relevant video instances:

$$\mathcal{L}_d[k] = \{(\text{id} = i, \text{cap} = c_i) \mid k \in s_i^d\}, \tag{2}$$

where c_i denotes the concise caption. This categorical organization empowers specialized agents to rapidly isolate relevant candidates by querying specific semantic keys, effectively converting the retrieval task from an exhaustive scan to a targeted lookup.

3.3 Collaborative Multi-Agent Pruning

To avoid the prohibitive cost of full-corpus verification, MAVIS employs a multi-agent pruning stage. Drawing inspiration from human collaboration, MAVIS mirrors the workflow of expert teams: rather than exhaustively scrutinizing every document, it prioritizes efficiency by rapidly filtering candidates via domain-specific heuristics and reserving debate for high potential cases. As illustrated in Fig. 3, we formalize this intuition into three step: *Query Decomposition*, *Independent Proposal*, and *Collaborative Debate*.

Query Decomposition. Upon receiving a query q, MAVIS first identifies the necessary semantic roles. A planner decomposes q into a set of sub-intents $\mathcal{Q} = \{q^r \mid r \in \mathcal{R}\}$, where $\mathcal{R} \subseteq \{\text{SCN}, \text{OBJ}, \text{ACT}\}$ represents the active semantic dimensions. For example, the query "*A dog running on the grass*" activates all three agents, whereas "*A dog running*" triggers the Object and Action agents (A_{OBJ} and A_{ACT}). This selective activation ensures computational resources are not wasted on irrelevant dimensions.

Independent Proposal. Each active agent A_r acts as a domain specialist tasked with retrieving candidates from its dedicated sub-library $\mathcal{L}_r$. Given the open-vocabulary discrepancy between the user's query q and the stored canonical keys, the agent

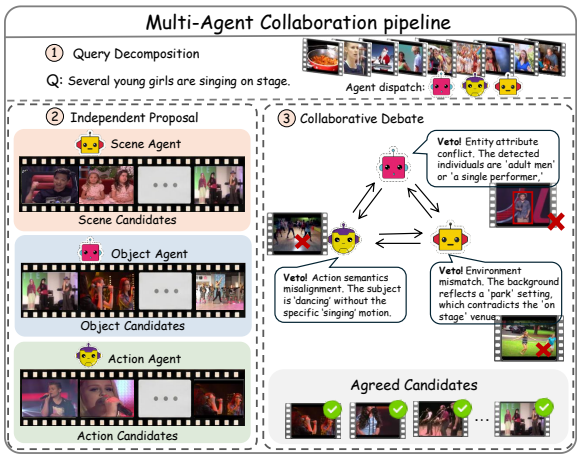

Figure 3: Multi-agent Collaboration pipeline: (1) Query Decomposition: Parsing the query into sub-tasks for agent dispatch. (2) Independent Proposal: Agents independently nominate initial candidates. (3) Collaborative Debate: A veto mechanism filters semantically inconsistent samples based on scene, object, and action constraints to reach a final consensus.

autonomously navigates the library index to identify the canonical category k that semantically aligns with the intent, resolving linguistic variations. Upon locating the correct category, the agent proceeds to verify the specific candidates. It acts as a strict evaluator, computing a confidence score $\phi_r(v_i)$ for each video by measuring the alignment between the query and the video's caption $C_{i,r}$:

$$\phi_r(v_i) = \cos\left(E^{\text{txt}}_{\text{VLM}}(q^r), E^{\text{txt}}_{\text{VLM}}(C_{i,r})\right), \quad (3)$$

where $E^{\text{txt}}_{\text{VLM}}(\cdot)$ denotes the text encoder of a VLM, and $\cos(\cdot)$ computes the cosine similarity. This metric quantifies the agent's judgment confidence. Based on this score, the agent generates a *Proposal Message* M_r:

$$M_r = \{(v_i, \phi_r(v_i)) \mid \phi_r(v_i) > \tau_{\text{soft}}\}, \quad (4)$$

where τ_{soft} is a lenient threshold designed to maximize recall before entering the debate stage.

Collaborative Debate. Instead of forcing agents to communicate point-to-point, the system serves as a central registry that asynchronously aggregates the independent proposal messages M_r from all agents into a unified preliminary candidate pool $\mathcal{V}_{\text{pool}}$. This union operation prioritizes recall, ensuring that a video proposed by *any* domain expert is considered a potential candidate:

$$\mathcal{V}_{\text{pool}} = \bigcup_{r \in \mathcal{R}} \{v_i \mid (v_i, \phi_r) \in M_r\}. \quad (5)$$

To support a rigorous debate, the blackboard ensures that every agent has a "vote" on every candidate in $\mathcal{V}_{\text{pool}}$. Since agents search independently, a video v_i proposed by the Object agent might initially lack a score from the Action agent. Therefore, for any $v_i \in \mathcal{V}_{\text{pool}}$, if an agent A_r has not yet computed a score (i.e., $v_i \notin M_r$), it performs an on-demand check. This synchronizes the blackboard, assigning a complete confidence vector $\Phi(v_i) = \{\phi_r(v_i) \mid r \in \mathcal{R}\}$ to every candidate.

With complete information available, the system applies a *Veto Protocol* to strictly filter out mismatched videos. The intuition is based on logical negation: a video remains valid only if *no* domain expert identifies a hard conflict. If the Object agent detects that a required entity is definitely missing ($\phi_{\text{OBJ}} < \tau_{\text{hard}}$), the video is discarded immediately, regardless of how well it matches the scene or action attributes. Formally, the surviving set is derived by subtracting the vetoed candidates:

$$\mathcal{V}_{\text{contro}} = \mathcal{V}_{\text{pool}} \setminus \{v_i \in \mathcal{V}_{\text{pool}} \mid \exists r, \phi_r(v_i) < \tau_{\text{hard}}\}, \quad (6)$$

where τ_{hard} is a strict lower bound representing logical negation. The remaining videos in $\mathcal{V}_{\text{contro}}$, termed *Controversial Candidates*, are semantically plausible but require fine-grained verification in the next stage.

3.4 Fine-grained Matching

The previous collaborative pruning stage effectively filters out the vast majority of irrelevant videos. For the remaining candidates $\mathcal{V}_{\text{contro}}$, we employ a powerful Vision-Language Model to perform fine-grained verification. The final relevance score is computed as:

$$S(q, v_i) = \text{Cos}(E^{\text{txt}}_{\text{VLM}}(q), E^{\text{vid}}_{\text{VLM}}(v_i)), \quad (7)$$

where $E^{\text{txt}}_{\text{VLM}}(\cdot)$ and $E^{\text{vid}}_{\text{VLM}}(\cdot)$ denote the text and video encoders of the VLM, respectively. We rank $v_i \in \mathcal{V}_{\text{contro}}$ by $S(q, v_i)$ to return the top-k results. This strategy optimizes the efficiency-accuracy trade-off: by reserving VLM computations exclusively for the most plausible candidates, MAVIS achieves deep visual understanding at a fraction of the computational cost required for full-corpus scanning.

4 Experiments

4.1 Datasets and Metrics

We conduct our evaluations on three widely used video retrieval datasets: MSR-VTT (Xu et al.,

Table 1: Performance comparison on MSR-VTT, MSVD, and ActivityNet. We categorize methods into supervised fine-tuning, large-scale foundation models, and agentic methods. **MAVIS** achieves superior performance without training.

Method	MSR-VTT			MSVD			ActivityNet		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Supervised / Fine-tuned Methods									
CLIP4Clip (Luo et al., 2022)	44.5	71.4	81.6	46.2	76.1	84.6	40.5	72.4	80.1
X-CLIP (Ma et al., 2022)	49.3	75.8	84.8	50.4	80.6	87.1	46.4	75.9	84.5
Cap4Video (Wu et al., 2023)	51.4	75.7	83.9	51.8	80.8	88.3	49.2	77.3	85.1
TeachCLIP (Tian et al., 2024)	46.8	74.3	82.6	47.4	77.3	84.5	42.2	72.7	84.0
T-MASS (Wang et al., 2024)	52.7	77.1	85.6	53.0	81.0	88.5	48.5	76.4	84.2
Large-scale Video Foundation Models									
VAST (Chen et al., 2023)	49.3	68.3	73.9	55.4	82.0	88.5	55.2	82.5	90.6
InternVideo2-6B (Wang et al., 2025b)	55.9	78.3	85.1	59.3	84.4	89.6	63.2	85.6	92.5
LanguageBind-H (Zhu et al., 2023)	44.8	70.0	78.7	53.9	80.4	87.8	41.0	68.4	80.8
VideoPrism-g (Zhao et al., 2024)	39.7	63.7	-	58.5	83.4	89.2	52.7	79.4	-
Marengo-2.6 (Twelve Labs, 2024)	49.35	73.47	-	61.0	85.3	91.5	55.36	82.55	-
Agentic Methods									
MERLIN (Round 0) (Han et al., 2024)	44.4	67.6	76.2	52.39	77.16	84.78	56.58	84.77	91.73
MERLIN (Round 3)	72.6	91.8	95.6	71.79	91.79	96.87	66.05	90.97	95.54
MERLIN (Round 5)	78.0	94.2	**96.8**	77.61	**94.48**	97.31	68.44	91.95	96.63
MAVIS (Ours)	**78.6**	**94.9**	96.3	**78.1**	94.25	**97.33**	**69.15**	**92.4**	**96.8**

2016), MSVD (Guadarrama et al., 2013), and ActivityNet (Krishna et al., 2017). To ensure a strictly fair comparison, we adhere to the specific evaluation settings and data partitions provided by MERLIN (Han et al., 2024). We evaluate performance using the standard recall at K ($R@K$) metric, which measures the percentage of queries for which the correct video appears in the top K retrieved results.

4.2 Implementation Details

For the initial structured parsing of the video corpus, we employ **Qwen3-omni-flash**[1] as the backbone for the Description Agent. This selection balances format accuracy with semantic alignment efficiency. The multi-agent retrieval team (Scene, Object, and Action) shares a unified reasoning backbone: **GPT-4o**[2]. We utilize role-specific system prompting to specialize this backbone into distinct domain experts.

For similarity scoring (ϕ_r) during the proposal phase, we utilize OpenAI's **text-embedding-3-large** model[3] to encode textual descriptions into high-dimensional vectors. For the fine-grained matching stage ($S(q, v_i)$), we employ **ViCLIP** (Wang et al., 2023b) to encode the query and video frames into a shared embedding space for precise visual verification.

Regarding the threshold settings, we empirically set the lenient threshold $\tau_{\text{soft}} = 0.5$ to maximize recall in the proposal phase, ensuring potential candidates are not prematurely discarded. For the final selection and debate phase, we implement a strict confidence threshold of $\tau_{\text{hard}} = 0.3$ to enforce logical consistency through our veto protocol. We implement our framework using PyTorch. All experiments are conducted on a single NVIDIA A100 GPU to ensure computational consistency.

4.3 Main Results

We evaluate MAVIS across three video-text retrieval benchmarks: MSR-VTT, MSVD, and ActivityNet, as summarized in Table 1. Compared to traditional supervised models, MAVIS demonstrates a substantial performance leap despite being entirely training-free. For instance, on MSR-VTT, MAVIS achieves 78.6% R@1, which is considerably higher than fine-tuned models such as T-MASS (52.7%) and Cap4Video (51.4%). While

[1] https://github.com/QwenLM/Qwen3

[2] https://openai.com/index/hello-gpt-4o/

[3] https://platform.openai.com/docs/guides/embeddings

models like InternVideo2-6B and VideoPrism-g excel in general video understanding, their broad-scale training objectives may not be fully optimized for the fine-grained cross-modal alignment required in retrieval tasks. These results suggest that for retrieval-specific challenges, structured alignment strategies are more critical than purely scaling up model size, as MAVIS better filters irrelevant visual noise to match textual queries. In comparison with agentic baselines, MAVIS shows superior efficiency and accuracy. While MERLIN starts with relatively poor performance in Round 0 (e.g., 56.58% R@1 on ActivityNet) and requires five iterative rounds to reach 68.44%, MAVIS achieves 69.15% R@1 in a single pass. This underscores the efficacy of our pipeline in handling temporal dependencies and filtering noise more efficiently than iterative refinement, while avoiding the heavy computational overhead of multiple reasoning rounds.

4.4 Ablation Study

To validate the contribution of each component in MAVIS, we conducted extensive ablation studies. We analyze the framework from three perspectives: (1) the effectiveness of the collaborative pruning mechanism, (2) the architectural design of agents and data representations, and (3) the trade-off between retrieval efficiency and accuracy.

Impact of Collaborative Debate. We first investigate the Logic-aware Debate mechanism (Table 2, Rows a-c). *Union Only* (a) achieves high recall potential but suffers from sub-optimal precision and a prohibitively large candidate pool, as it fails to filter irrelevant videos proposed by independent agents. Conversely, *Strict Intersection* (b) acts as an overly aggressive filter, significantly dropping R@1 to 58.4% by missing valid candidates that lack full consensus among all agents. While *Average Fusion* (c) serves as a strong baseline, it struggles with "hard negatives" where at least one specialized agent possesses a decisive veto signal. Our proposed **Union + Veto** strategy (MAVIS) achieves the optimal balance, outperforming Union Only by 3.4% in R@1 while effectively pruning the candidate pool size from 68 to 23, ensuring high efficiency for the subsequent fine-grained stage.

Necessity of Agent Specialization. Replacing our specialized agents with a *Single General Agent* (Row d) leads to a notable 6.5% drop in R@1. This performance decay suggests that a single agent suffers from increased cognitive load when parsing Composite queries, often leading to a "semantic blurring" effect where critical details are overshadowed by verbose context. In contrast, our multi agent design facilitates semantic isolation, enabling the Scene, Object, and Action agents to function as domain-specific experts. By isolating these semantic roles, each agent can selectively attend to the most discriminative cues.

Table 2: Comprehensive ablation study on MAVIS components in MSR-VTT. $|\mathcal{V}_{\text{contro}}|$ denotes the average number of videos retained per query for the fine-grained stage. The default MAVIS setting is highlighted in **gray**.

Method Setting	R@1	R@5	R@10	$\lvert\mathcal{V}_{\text{contro}}\rvert$
MAVIS (Full Model)	**78.6**	**94.9**	**96.3**	**16**
Q1: Impact of Collaborative Debate Strategy				
(a) Union Only (No Veto)	75.2	90.1	92.5	58
(b) Strict Intersection (∩)	58.4	72.3	78.2	**8**
(c) Average Fusion Score	76.8	92.5	94.1	31
Q2: Necessity of Agent Specialization				
(d) Single General Agent	72.1	88.4	90.8	57
Q3: Impact of Semantic Library				
(e) w/o Structured Attributes	70.5	84.2	87.5	69

Table 3: Efficiency-Accuracy Trade-off analysis on MSR-VTT. FLOPs are estimated for the inference of a single query against the full corpus. Speedup is calculated relative to the Fine Only (ViCLIP).

Paradigm	R@1	R@5	R@10	FLOPs (G)	Speedup
(i) Coarse Only (Agent debate)	68.0	82.5	88.4	**1**	**100,000**×
(ii) Fine Only (Brute-force ViCLIP)	50.8	74.8	82.9	100,000	1×
(iii) MAVIS (Coarse-to-Fine)	78.6	94.9	96.3	3200	31.2×

Impact of Semantic Library. Utilizing *Raw Captions* directly (Row e) degrades performance significantly. This validates the necessity of our *Structured Video Understanding* stage. Unlike free-form text which often contains redundant noise or loosely coupled descriptions, structured parsing provides explicit semantic guidance by routing categorical attributes to their corresponding agents. This modular specialization allows each agent to focus exclusively on verified attributes within its domain, effectively preventing "semantic bleeding" and minimizing hallucinations.

Efficiency-Accuracy Trade-off. We evaluate the "Inspect less, understand more" philosophy in Table 3. (i) **Coarse Only** retrieval (Agent debate) is extremely efficient, yielding a 100,000× speedup by matching structured captions in the text domain; however, it lacks final visual confirmation. (ii) **Fine**

Table 4: Ablation study of thresholds τ_{soft} and τ_{hard} on the MSR-VTT dataset. $|\mathcal{V}_{\text{contro}}|$ denotes the average number of candidates requiring fine-grained matching.

τ_{soft}	τ_{hard}	$\lvert\mathcal{V}_{\text{contro}}\rvert$	R@1 (↑)	R@10 (↑)
0.3	0.3	42.1	78.8	96.2
0.5	**0.3**	**16.4**	**78.6**	**96.3**
0.7	0.3	5.8	72.4	85.5
0.5	0.1	38.2	78.7	96.5
0.5	**0.3**	**16.4**	**78.6**	**96.3**
0.5	0.5	4.2	65.1	82.4

Only (Brute-force ViCLIP) represents the standard zero-shot visual baseline. Due to the lack of structured semantic reasoning, it achieves a lower precision while incurring a prohibitive computational cost of 100,000G FLOPs. (iii) **MAVIS** successfully bridges this gap and significantly outperforms the brute-force baseline. By leveraging the collaborative debate to prune the search space to just 16 candidates, MAVIS achieves a peak performance of **78.6% R@1**. This demonstrates that our collaborative pruning and reasoning mechanism is not only efficient but also essential for resolving complex semantic queries that monolithic VLMs struggle to handle.

Ablation on Retrieval Thresholds. We evaluate the sensitivity of MAVIS to the proposal threshold τ_{soft} and the veto threshold τ_{hard} in Table 4. The lenient threshold τ_{soft} is designed to maximize the coverage of the preliminary candidate pool $\mathcal{V}_{\text{pool}}$. While $\tau_{\text{soft}} = 0.3$ slightly boosts recall, it bloats the initial candidate pool to 42.1, increasing computational overhead; $\tau_{\text{soft}} = 0.5$ serves as the optimal filter to maintain a manageable pool size.

The strict threshold τ_{hard} represents the boundary for logical negation within our collaborative veto protocol. When τ_{hard} is set too high, the system becomes overly aggressive, leading to "over-vetoing" where valid positives are discarded due to minor semantic discrepancies. Conversely, a lower τ_{hard} weakens the logical filter, failing to eliminate irrelevant distractors and resulting in a larger $\mathcal{V}_{\text{contro}}$.

4.5 Case Study

To qualitatively evaluate the reasoning capabilities of MAVIS, we perform a comparative analysis against two representative paradigms: **InternVideo2**, a perception foundation model, and **Merlin**, an MLLM-based multi-round reasoning agent. As illustrated in Figure 4. InternVideo2 relies on global embeddings to match video features with

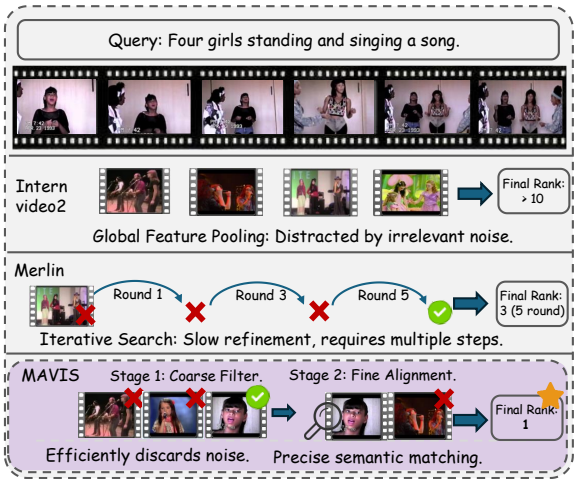

Figure 4: **Qualitative comparison of MAVIS, InternVideo2, and Merlin.** Case A: InternVideo2 suffers from retrieval biases as its global perception over-fits to background context rather than specific query intents. Case B: Merlin relies on sequential multi-round refinement. MAVIS successfully retrieves the ground truth via its logic-aware veto protocol.

textual queries. In cases of *semantic asymmetry*, InternVideo2 often yields false positives. The dominant scene features statistically overwhelm subtle object cues in the global latent space, leading to a loss of fine-grained precision. Merlin utilizes an MLLM to orchestrate external tools through multiple rounds of reasoning. However, it is prone to *tunnel vision* and reasoning hallucinations due to its single-agent architecture.

MAVIS resolves these challenges by replacing holistic matching with a collaborative veto protocol. By assigning specialized agents to verify Scene, Object, and Action dimensions independently, MAVIS ensures that any semantic mismatch triggers an immediate veto, effectively eliminating hallucinations and perception drowning.

5 Conclusion

We introduced MAVIS, a training-free framework for text-to-video retrieval that replaces exhaustive full-corpus matching with a structured, coarse-to-fine retrieval paradigm. By first transforming raw videos into a semantic library and then performing retrieval through query decomposition, specialized agent collaboration, and logic-aware debate, MAVIS effectively addresses the granularity mismatch between sparse textual queries and information-dense video content. This design allows the system to prune large amounts of irrele-

vant information before invoking expensive fine-grained matching, thereby improving both retrieval efficiency and semantic precision.

Experiments on MSR-VTT, MSVD, and ActivityNet show that MAVIS achieves strong retrieval performance without task-specific fine-tuning, while maintaining a favorable efficiency–accuracy trade-off. The ablation studies further verify that the structured semantic library, agent specialization, and veto-based collaborative pruning are all important to the overall effectiveness of the framework. Taken together, these results suggest that cooperative multi-agent reasoning is a promising alternative to conventional embedding-based retrieval, and may provide a scalable and interpretable direction for future multimodal retrieval systems.

6 Limitations

Despite its promising results, MAVIS has several limitations. First, as a training-free framework, its performance is bounded by the zero-shot perception and reasoning capabilities of the underlying foundation models. Errors introduced during the structured parsing stage, such as omitted entities, inaccurate actions, or hallucinated descriptions, may propagate through the retrieval pipeline, since the current framework does not include an explicit mechanism for correcting upstream mistakes.

Second, MAVIS relies on empirically chosen proposal and veto thresholds, and its retrieval quality is therefore sensitive to threshold calibration. Although the selected settings perform well on the evaluated benchmarks, they may require adjustment in domains with different semantic distributions or higher visual ambiguity. In addition, the current agent design mainly focuses on scene, object, and action, which may be insufficient for more complex retrieval scenarios involving temporal dynamics, audio cues, OCR, or fine-grained relational reasoning. Finally, because MAVIS builds on off-the-shelf MLLMs and VLMs, it may also inherit their biases and failure modes, which should be considered when deploying the framework in real-world applications.

Ethics Statement

In accordance with the ACL Ethics Policy, we provide the following statement regarding our work. We utilized an AI-based language model exclusively for rephrasing and polishing the textual content of this paper to improve its clarity and linguistic quality. We explicitly state that the core methodology, experimental designs, and empirical results presented in this work are entirely original and were not generated or altered by any AI assistant. MAVIS is a training-free framework that leverages off-the-shelf Multimodal Large Language Models (MLLMs), such as GPT-4o and Qwen3. These large-scale models may reflect societal biases (e.g., gender, age, or race) present in their massive pre-training corpora. While MAVIS introduces a logic-aware veto protocol to enhance retrieval precision, it may still inherit or propagate such biases during the semantic parsing or scoring phases. Users should be cautious when deploying our framework in sensitive real-world applications.

References

Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. 2025. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. *arXiv preprint arXiv:2503.10200*.

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

Jianfeng Dong, Xianke Chen, Minsong Zhang, Xun Yang, Shujie Chen, Xirong Li, and Xun Wang. 2022. Partially relevant video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 246–257.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. 2024. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pages 75–92. Springer.

Sheng Fang, Tiantian Dang, Shuhui Wang, and Qingming Huang. 2024. Linguistic hallucination for text-based video retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*.

Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2013. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot

recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2712–2719.

Donghoon Han, Eunhwan Park, Gisang Lee, Adam Lee, and Nojun Kwak. 2024. Merlin: Multimodal embedding refinement via llm-based iterative navigation for text-video retrieval-rerank pipeline. *arXiv preprint arXiv:2407.12508*.

Peng Jin, Hao Li, Zesen Cheng, Jinfa Huang, Zhennan Wang, Li Yuan, Chang Liu, and Jie Chen. 2023. Text-video retrieval with disentangled conceptualization and set-to-set alignment. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 938–946.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer.

Yang Liu, Shudong Huang, Deng Xiong, and Jiancheng Lv. 2025. Learning dynamic similarity by bidirectional hierarchical sliding semantic probe for efficient text video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5667–5675.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304.

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

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. 2022. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, pages 638–647.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180.

Haoran Tang, Meng Cao, Jinfa Huang, Ruyang Liu, Peng Jin, Ge Li, and Xiaodan Liang. 2025. Muse: Mamba is efficient multi-scale learner for text-video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7238–7246.

Kaibin Tian, Ruixiang Zhao, Zijie Xin, Bangxiang Lan, and Xirong Li. 2024. Holistic features are almost sufficient for text-to-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17138–17147.

Twelve Labs. 2024. Introducing Marengo 2.6: A state-of-the-art video foundation model for any-to-any search. https://www.twelvelabs.io/blog/introducing-marengo-2-6. Accessed: 2026-01-05.

Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. 2024. Text is mass: Modeling as stochastic embedding for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16551–16560.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023a. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560.

Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, and 1 others. 2025a. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8428–8437.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, and Yaohui Wang. 2023b. Internvid: A large-scale video-text dataset for multimodal understanding and generation.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, and Yansong Shi. 2025b. Internvideo2: Scaling foundation models formultimodal video understanding. In *European Conference on Computer Vision*.

Syed Talal Wasim, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. 2023. Vita-clip: Video and text adaptive clip via multimodal prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23034–23044.

Peng Wu, Wanshun Su, Xiangteng He, Peng Wang, and Yanning Zhang. 2025. Varcmp: Adapting cross-modal pre-training models for video anomaly retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8423–8431.

Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. 2023. Cap4video: What can auxiliary captions do for text-video retrieval? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10704–10713.

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*, pages 5288–5296.

Xiangpeng Yang, Linchao Zhu, Xiaohan Wang, and Yi Yang. 2024. Dgl: Dynamic global-local prompt tuning for text-video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6540–6548.

Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. 2025. Vca: Video curious agent for long video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20168–20179.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Qilang Ye, Zitong Yu, Rui Shao, Yawen Cui, Xiangui Kang, Xin Liu, Philip Torr, and Xiaochun Cao. 2025. Cat+: Investigating and enhancing audio-visual understanding in large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

Qilang Ye, Wei Zeng, Meng Liu, Jie Zhang, Yupeng Hu, Zitong Yu, and Yu Zhou. 2026a. When eyes and ears disagree: Can mllms discern audio-visual confusion? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 11955–11963.

Qilang Ye, Yu Zhou, Lian He, Jie Zhang, Xuanming Guo, Jiayu Zhang, Mingkui Tan, Weicheng Xie, Yue Sun, Tao Tan, and 1 others. 2026b. Sugar: Learning skeleton representation with visual-motion knowledge for action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 17930–17938.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2023. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36:76749–76771.

Jiayu Zhang, Pengjie Tang, Yunlan Tan, and Hanli Wang. 2025. Mgtr-miss: More ground truth retrieving based multimodal interaction and semantic supervision for video description. *Neural Networks*, page 107817.

Jiayu Zhang, Shuo Ye, Qilang Ye, Zihan Song, Jiajian Huang, and Zitong Yu. 2026. Retrieving to recover: Towards incomplete audio-visual question answering via semantic-consistent purification. *Preprint*, arXiv:2604.10695.

Kechi Zhang, Jia Li, Ge Li, Xianjie Shi, and Zhi Jin. 2024. Codeagent: Enhancing code generation with tool-integrated agent systems for real-world repo-level coding challenges. *arXiv preprint arXiv:2401.07339*.

Long Zhao, Nitesh B. Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J. Sun, Luke Friedman, Rui Qian, Tobias Weyand, and Yue Zhao. 2024. Videoprism: A foundational visual encoder for video understanding.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, and 1 others. 2023. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*.

Yijie Zhu, Jie He, Rui Shao, Kaishen Yuan, Tao Tan, Xiaochen Yuan, and Zitong Yu. 2026a. δ vla: Prior-guided vision-language-action models via world knowledge variation. *arXiv preprint arXiv:2603.08361*.

Yijie Zhu, Rui Shao, Ziyang Liu, Jie He, Jizhihui Liu, Jiuru Wang, and Zitong Yu. 2026b. H-gar: A hierarchical interaction framework via goal-driven observation-action refinement for robotic manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

A Appendix: Implementation Details

model setup. The performance of MAVIS heavily relies on the quality of the structured captions generated in the first stage. To select the optimal backbone for the Description Agent, we conducted a preliminary study comparing state-of-the-art Multimodal Large Language Models (MLLMs), including **GPT-4.1**, **Gemini-2.5-Pro**, and **Qwen3-omni-flash**. We constructed a dataset of 200 diverse videos sampled from the training set of MSR-VTT. For each model, we evaluated its instruction-following capability (i.e., adherence to the JSON

Table 5: Preliminary selection of the Description Agent.

Model	Format Accuracy	Alignment Score
GPT-4.1	96.4%	83.1
Gemini-2.5-Pro	97.8%	82.3
Qwen3-omni-flash	**98.5%**	**84.6**

output format) and the semantic accuracy of the extracted attributes. Specifically, we employ the *Caption-Query Alignment Score* as our evaluation metric, which measures the retrieval recall of the generated captions against ground-truth queries using OpenAI's **text-embedding-3-large** model. The motivation behind this metric is to ensure that the generated captions are highly semantic-aligned with potential user queries, thereby bridging the modality gap more effectively during the retrieval phase. As shown in Table 5, **Qwen3-omni-flash** achieves the highest scores in both format accuracy and semantic alignment. Consequently, we adopted Qwen3-omni-flash as the backbone for the Description Agent to construct the Semantic Library.

B visualization of Semantic Library

The MAVIS semantic library is organized as a structured inverted index, designed to replace conventional brute-force scanning with efficient targeted lookups. The library is partitioned into three domain-specific sub-libraries: Scene, Object, and Action, each utilizing normalized semantic concepts as primary keys.

In our implementation, each key points to a collection of video metadata entries. As shown in Figure 5, which visualizes a subset of the Scene library and Object library, each entry consists of a unique video ID and a concise caption generated by the Description Agent. This structure ensures that the textual granularity of the library is strictly aligned with potential user queries, facilitating high-precision matching while significantly reducing the number of candidates for the final fine-grained verification stage.

C Appendix: Prompt Settings

Since MAVIS operates as a training-free framework, the zero-shot reasoning and tool-invocation capabilities of the underlying MLLMs are primarily governed by precise prompt engineering. As summarized in Table 6, we design specialized instruction sets for each agent to ensure a robust workflow. Rather than relying on parameter updates, MAVIS elicits different functional behaviors through role-specific prompts, enabling the planner to decompose user queries into semantic sub-intents, the description agent to produce concise structured video descriptions, and the scene, object, and action agents to retrieve and verify candidates from their corresponding semantic sub-libraries.

These prompts are designed with two goals in mind. First, they encourage semantic consistency across different stages of the pipeline by enforcing standardized outputs and reducing irrelevant or overly verbose generations. Second, they improve the stability of multi-agent collaboration by clearly specifying each agent's scope, decision criteria, and expected response format, which is especially important for the subsequent veto-based debate process. We provide the main prompt templates in this appendix to improve reproducibility and to clarify how MAVIS translates general-purpose MLLMs into specialized retrieval agents without additional training.

Scene Library

Beach
- f-24IxG9ijw_25_40.avi: Along the seashore, a young boy is seen riding a motorcycle.
- g2IYQq7IkXc_23_32.avi: On a narrow strip of land in the ocean, a polar bear runs while appearing to drive away several walruses.
- jLgmCY1fEE8_16_26.avi: A man is engaged in a conversation with several women on the beach.
- lo4KcsBN--A_0_10.avi: In the ocean waters, a tortoise is observed chasing after fish.
- omIPdpxg--4_39_46.avi: A woman is captured in the moment of falling onto a sand castle at the beach.

Park
- f9_bP219ehQ_63_70.avi: At an event in the park, a man is undergoing an interview.
- fMFvOgb4k6E_35_43.avi: While hoeing the soil in the park, a woman pauses to remove an object from her foot.
- fkONJEgTNJY_25_35.avi: Positioned on the steps of a swimming pool, a dog plays with the foamy water sprayed from a garden hose.
- o_mWZWcm2r4_47_54.avi: Observed by other children, a girl pedals her bicycle around a parking area.
- q8t7iSGAKik_11_31.avi: A man demonstrates his skill by maneuvering a soccer ball with his feet in the park.

Indoor
- fBA_lxUiwSg_2_4.avi: An indoor cat is seen leaping onto the surface of a table.
- f_CvW22Eauc_16_23.avi: Dressed in a formal white suit, a man walks across the indoor floor while a woman remains seated.
- fr9H1WLcF1A_256_261.avi: A pair of young men are engaged in a game of table tennis indoors.
- fvBs0xpEZhQ_10_30.avi: Inside an office, a man juggles a small ball using his head and feet.
- gbW9f8xydks_0_10.avi: From a pan inside a cage, two animals eat what appear to be apple slices.

Street
- gqxpGOHUH9k_113_119.avi: A large group of men and women are seen dancing out on the street.
- jbzaMtPYtl8_48_58.avi: As spectators look on, two men roll massive tires sideways along a street.
- jlahRlo4jlU_30_36.avi: A man is observed walking along a road in the city.
- mZVPkPqwzR4_38_45.avi: Several deer are filmed crossing over a public road.
- mmSQTI6gMNQ_120_128.avi: From behind a bush, two men watch a pair of women standing on the street.

Object Library

dog
- eyhzdC936uk_15_27.avi: A young boy is having fun playing with a dog.
- fcvW1vr8hAs_96_102.avi: While holding a guitar, a man stops to pet two dogs.
- ficwZQYmRLE_5_20.avi: After a large white dog sniffs a small yellow duck, the duck pecks the dog's face.
- fkONJEgTNJY_25_35.avi: A dog stands upon the swimming pool steps, playing with foam from a garden hose.
- gjVBEJGHrXk_26_38.avi: A cat is seen actively chasing a dog around.
- glii-kazad8_21_29.avi: Walking toward the edge of the dock is a small dog.

person
- eyhzdC936uk_15_27.avi: A boy is interacting and playing with a dog.
- f-24IxG9ijw_25_40.avi: On the seashore, a boy is seen riding a motorcycle.
- f9_bP219ehQ_63_70.avi: During a park event, a man participates in an interview.
- fEsrO_poIUg_161_168.avi: A man places a large knife's tip into a vise before pulling the blade toward himself.
- fF89MasBFLw_321_326.avi: A young woman is shown applying stickers across her entire face.
- fGc6_D0JEIQ_31_46.avi: A trio of men are captured dancing together.

motorcycle
- f-24IxG9ijw_25_40.avi: A boy travels along the seashore on a motorcycle.
- izU1dDwnuMY_80_92.avi: Three men are seen riding motorcycles through the woods.
- k5OKBX2e7xA_19_32.avi: A man is performing various professional stunts and tricks on a motorcycle.
- klteYv1Uv9A_27_33.avi: While the motorcycle is in motion, a man is lying back on it.
- lAznAeFFldg_6_10.avi: A man is riding his bike over a steep hill in the forest.

cat
- f9Won2JpOEU_60_80.avi: A cat lies down comfortably while licking its own paw.
- fBA_lxUiwSg_2_4.avi: A cat leaps onto the top of a table.
- fJr2evLANsE_0_10.avi: A cat is seen playfully interacting with a turtle.
- gMqKUPeTAkg_17_30.avi: A cat is caught drinking water directly from the kitchen sink.
- gjVBEJGHrXk_26_38.avi: A cat is engaged in a pursuit, chasing a dog.

Figure 5: **Visualization of Semantic Library.** This figure illustrates samples from sub-libraries. Each entry maps a specific video ID to its corresponding concise semantic summary. This structured organization enables the Planner and specialized agents to perform targeted lookups rather than exhaustive scans.

Table 6: Agent roles and task specifications in MAVIS. Each agent has a specialized role: Description Agent parses videos into structured semantic components; Planner decomposes queries; Retrieval Agents independently nominate candidates; Debate Controller applies a veto protocol to filter logical mismatches.

Agent / Role	Instruction and Task Specification
Description Agent	**System:** Act as a video description agent specialized in generating concise, query-aligned captions. The output should be concise and intent-focused. Avoid including excessive details, background information, or atmospheric nuances that are not central to the main action.
(Video Parsing)	**Task:** Analyze video frames to produce two outputs: (1) A *Concise Caption* (5-20 words, matching the granularity of typical user queries, focusing on PRIMARY action, main objects, and key scene); (2) *Structured Semantic Components* (Scene, Object, Action) normalized into canonical forms. Scene identifies primary location. Object extracts main entities in canonical form (singular, e.g., "dog", "cat", "person", "motorcycle"). Action extracts primary activity in gerund form (e.g., "playing", "riding", "cleaning"). Use standardized keywords and normalize synonyms. **Output Format:** JSON with `{caption, scene, object, action}` keys, where `scene`, `object`, and `action` are arrays of standardized keywords.
Planner	**System:** Act as a query planner that decomposes user search queries into atomic semantic components. Identify which semantic dimensions (Scene, Object, Action) are present in the query and extract the corresponding sub-intents for each active dimension.
(Decomposition)	**Task:** (1) Identify which semantic dimensions are present: Scene, Object, and/or Action. (2) Extract the specific sub-intent for each active dimension. (3) If a dimension is not mentioned, exclude it from the output (selective activation). For example, "a dog running on the grass" activates Scene, Object, and Action; "a dog running" activates only Object and Action. **Output Format:** JSON containing `active_dimensions` (list of activated roles, e.g., `["Object", "Action"]`) and `sub_intents` (dictionary mapping each role to its sub-intent, e.g., `{"Object": "dog", "Action": "running"}`).
Retrieval Agent	**System:** Act as a domain specialist (Scene/Object/Action) retrieval agent. Each agent focuses on one semantic dimension to search the semantic library for videos matching the query's component in that dimension.
(Candidate Proposal)	**Task:** (1) *Category Matching:* Navigate the library index to identify the canonical category key that semantically matches the query sub-intent, resolving linguistic variations between query and stored canonical keys (e.g., "bike" $\rightarrow$ "motorcycle", "chatting" $\rightarrow$ "talking"). If no exact match, find the closest semantic category. (2) *Candidate Retrieval:* Retrieve ALL videos from the matched category. (3) *Similarity Computation:* For each candidate video, invoke a similarity tool to compute confidence scores $\phi_r(v_i) = \cos(E^{\text{txt}}_{\text{VLM}}(\text{query\_sub\_intent}), E^{\text{txt}}_{\text{VLM}}(\text{video\_caption}))$, where $E^{\text{txt}}_{\text{VLM}}(\cdot)$ is the VLM text encoder and $\cos(\cdot)$ computes cosine similarity. This quantifies the agent's confidence in the match. (4) *Proposal Generation:* Generate a proposal message M_r containing all videos where $\phi_r(v_i) > \tau_{\text{soft}}$ (threshold 0.5, designed to maximize recall). Act as a strict evaluator, computing confidence scores for each video. **Output Format:** JSON with `matched_category` (the canonical category key), `reasoning` (explanation of why this category matches the query intent), and `proposal` (list of `{id, cap, similarity}` tuples for all candidates that pass the soft threshold).
Debate Controller	**System:** Act as a logic-aware debate coordinator. Facilitate collaborative filtering among multiple retrieval agents using a strict *Veto Protocol* to filter out logically inconsistent candidates.
(Veto Mechanism)	**Task:** (1) *Complete Information:* Ensure every agent has a "vote" (score) on every candidate in the unified pool $\mathcal{V}_{\text{pool}}$. If an agent hasn't computed a score for a candidate, it performs an on-demand check. (2) *Veto Protocol:* Apply logical negation: a video is VETOED (removed) if ANY agent identifies a hard conflict, where $\phi_r(v_i) < \tau_{\text{hard}}$ (threshold 0.3). (3) *Controversial Candidates:* Identify videos that survive the veto. These are semantically plausible but require fine-grained verification. They have no hard conflicts from any agent, but may have mixed confidence scores. **Output Format:** JSON with `vetoed_candidates`, `controversial_candidates`, `veto_reasons` and `controversial_scores`.