From Alignment to Reason: Multi-Agent Debate for Tactical Badminton Video Retrieval

Yi-Xiang Zhang Yu-Shuen Wang
National Yang Ming Chiao Tung University
yixiang.ii13@nycu.edu.tw yushuen@cs.nycu.edu.tw

Abstract

Retrieving complex, tactical moments in specialized domains like badminton is a significant challenge that current Vision-Language Models (VLMs) are not equipped to handle. Trained on massive, generic datasets, VLMs learn pixel-to-text alignments but lack the domain-specific expertise to understand tactical intent or causality. We propose a novel framework that bypasses this VLM alignment, instead leveraging the rich domain knowledge of Large Language Models (LLMs) to interpret structured event data. Our method first employs domain-specific computer vision tools to decompose videos into structured, textual game logs. We demonstrate that these game logs are a remarkably potent asset for retrieval, as semantic search on these logs alone dramatically outperforms state-of-the-art VLM-based systems. To capture the causal reasoning that raw logs lack, we introduce a rigorous, multi-agent dialectic reasoning process where agents collaboratively debate the log, draft and revise a narrative, and verify its grounding to the original game log events. This "Generate-then-Retrieve" framework provides a step-function improvement in retrieval accuracy. Our system is not only more accurate but also fully interpretable, providing human-readable, grounded explanations for every result. Project page: yixiang1120.github.io/MADR-project-page

Searching for a winning smash created by successfully manipulating the opponent's court position.

Figure 1. Our system allows users to retrieve rally videos using high-level tactical queries. In this rally, the bottom-court player was repeatedly displaced: from the middle court to the left court and then to the front-right court. The top-court player then delivered a smash to the bottom player's rear-left court and scored. Because the paper can only present static frames, we encourage readers to watch the retrieved video for this rally, as well as additional retrieval examples, available in the supplemental material.

1. Introduction

Retrieving specific tactical moments from hours of broadcast match footage is a critical yet profoundly challenging task in computational sports analysis. For a badminton coach, queries are not about simple appearances, such as "a player in a red shirt", but instead involve complex causality and intent, as illustrated in Figure 1. The dominant paradigm for video retrieval, vision-language models (VLMs), attempts to solve this by learning a direct alignment between video pixels and text. However, this alignment is trained on massive, generic datasets (e.g., "a cat jumping", "a person cooking"), which lack the fine-grained, domain-specific annotations required for expert-level understanding. Consequently, a VLM may recognize a "hit", but it fundamentally fails to understand the tactical significance of that hit within the context of a badminton match.

We argue that this text-video alignment is not suitable for specialized domains. In contrast to VLMs, which rely heavily on image- or video-text pairs, the amount of high-quality alignment data available for training such models is substantially smaller than the purely textual corpora used to train modern large language models (LLMs). As a result, VLMs often lack the domain knowledge required

for tactical understanding. LLMs, by comparison, have already ingested vast quantities of specialized knowledge from the web, including sports rulebooks, tactical blogs, and match commentary. Accordingly, we adopt a decoupled perception-reasoning framework [43] that allows LLMs to apply their rich domain knowledge directly to the structured events extracted using domain-specific tools. This approach bypasses the feature alignment limitations in VLMs.

Our perception module is built upon established computer vision techniques that decompose the video into a structured, textual *game log*. This log contains a timestamped sequence of atomic events. Our critical finding is that this game log, by itself, is a remarkably potent retrieval asset. We demonstrate that semantic search directly on these structured logs dramatically outperforms state-of-the-art VLM-based retrieval systems [19, 30, 32, 47], proving that for specialized domains, a structured data representation is far superior to a generic visual embedding.

Although the game log effectively captures what happened, it still does not explain the reasoning or causality behind those events. To achieve this deeper level of understanding, we introduce a Multi-Agent Dialectic Reasoning (MADR) process. This process is a rigorous, multi-stage pipeline designed to transform the log into high-fidelity tactical narratives. First, an *Offense Analyst* and a *Defense Analyst* read the raw game log and engage in an adversarial debate, challenging hypotheses about player intentions and tactical effectiveness. A *Tactic Summarizer* then synthesizes this discussion into a first-draft narrative. This draft is passed to two *Reviewers*, who provide critical feedback, prompting the *Tactic Summarizer* to revise its output. Finally, a *Verification Agent* performs a grounding check, ensuring that every claim in the final narrative can be traced back to one or more events in the original game log.

This hierarchical approach forms a *Generate-then-Retrieve* framework. During an offline phase, we generate both the game logs (for factual event retrieval) and the fully verified narratives (for causal/tactical retrieval). Our experiments show that this multi-layered approach provides a step-function improvement in retrieval accuracy. Using the game logs alone surpasses VLMs, and using the narratives generated by our debate team further enhances semantic accuracy, particularly for complex, intent-based queries. This framework is not only accurate but also fully interpretable, providing a clear explanation for every retrieved result.

2. Related Works

Text-Video Understanding and Retrieval. Recent work in video-language learning has centered on aligning textual and visual modalities within a shared embedding space. Following the success of CLIP [35] in vision-language alignment, methods such as CLIP4Clip [25] extended this idea to the video domain by aggregating frame-level representations into unified video embeddings. This paradigm has since evolved into large-scale video foundation models [28, 49], which are trained on massive general-domain datasets to learn video and text embeddings jointly. These models demonstrate remarkable generalization across tasks such as retrieval [25, 44, 45], captioning [40], and question answering [11], all grounded in the same alignment principle. However, alignment-based models often struggle in domain-specific applications, such as tactical sports analysis, where causal relations are critical yet underrepresented in general-purpose data.

Perception-Reasoning Frameworks. Recent studies have explored decoupled perception-reasoning frameworks to better integrate structured perception with language-based reasoning. In these architectures, perception modules first extract task-relevant representations from raw multimodal inputs, while reasoning modules interpret these representations to support decision-making [1, 39] or causal understanding in complex domains, such as anomaly detection [48] and question answering [31]. Such frameworks mark a shift from a generic alignment-based approach toward goal-oriented reasoning, providing a more interpretable and adaptable foundation for domain-specific video understanding.

Multi-Agent Reasoning and Debate Frameworks. The concept of using multiple interacting agents to solve complex problems has recently gained significant traction within the LLM community. Frameworks have been proposed where LLM agents assume different roles, such as a planner, an executor, or a critic, to collaboratively accomplish tasks ranging from software development [33] to medical diagnosis [9], scientific discovery [16], and broader strategic reasoning applications [52], such as analyzing structured gameplay log data for tactical understanding [27]. This approach, often involving debate or discussion, has been shown to improve reasoning, reduce hallucinations, and yield solutions that integrate more perspectives and deeper Analyst than single-model approaches [5]. For example, agents may engage in consultative dialogues or adversarial debates to refine a plan or verify information. Representative frameworks such as SagaLLM [8], SocraSynth [6], and EVINCE [7] further demonstrate how structured validation, conditional statistics, and information-theoretic moderation can regulate multi-agent dialogue to achieve more robust, verifiable reasoning outcomes. While recent automated multi-agent workflow search frameworks, such as AFlow [51] and MaAS [50], have advanced this paradigm with powerful optimization capabilities, they rely on explicit reward functions to evaluate and evolve the generated workflows. This reliance limits their applicability in open-ended or subjective domains such as tactical sports analysis, where a definitive ground-truth reward for strategic narratives does not exist.

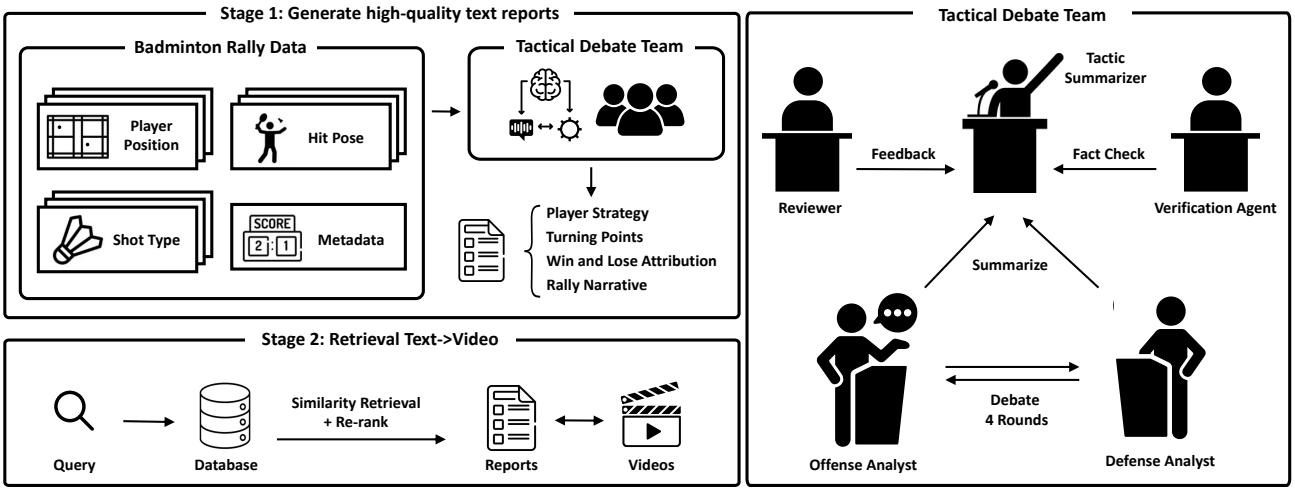

Figure 2. We propose a Generate-then-Retrieve framework for badminton rally videos, which enables users to search for videos using high-level tactical queries. In the first stage, game logs, such as shuttle trajectories and player positions, are extracted from the videos. A tactical debate team, consisting of multiple agents, analyzes player intentions and tactics based on these logs to transform the videos into tactical narratives. In the second stage, rally videos are retrieved by comparing user queries with the generated narratives.

Computational Sports Analysis. Prior research in computational sports analysis has provided essential tools for extracting low-level events from game footage. This includes foundational work in player/ball tracking [12, 21, 29], as well as fine-grained action recognition [17, 46] and stroke-type classification [10, 24, 41]. These methods are highly effective at identifying what action occurred at a specific moment. However, they are not designed to provide the tactical reasoning or causal explanations that connect these discrete actions into a meaningful rally narrative.

Our contribution lies in establishing an explicit and interpretable semantic layer for retrieval. We demonstrate, for the first time, how a multi-agent debate framework can transform raw badminton rally videos into high-level tactical narratives that encode intent and causal reasoning. These generated narratives become the central medium for retrieval, replacing opaque visual embeddings with searchable, human-readable semantics. This *Generate-then-Retrieve* paradigm redefines video retrieval as a reasoning-driven process, enabling complex, intent-oriented queries that embedding-based methods cannot easily support.

3. Methodology

To generate narratives that are both factually grounded and tactically insightful from raw badminton data, we propose a Multi-Agent Dialectic Reasoning (MADR) framework (Figure 2). The core design philosophy is to simulate the collaborative reasoning process of a professional coaching team. Unlike single-agent methods that rely on one LLM to produce a report in a single step, which often suffer from factual inaccuracies, MADR decomposes the complex analytical task into a sequence of well-defined subtasks. Each sub-task is assigned to an agent with a distinct persona and responsibility. Through structured debate, synthesis, peer review, and iterative refinement, these agents collectively construct the final analysis report.

3.1. Game Log Preprocessing

The tactical reasoning process in MADR does not rely directly on raw video frames but on a structured game log that summarizes each rally with precise spatiotemporal and event-level information. This design reflects how a professional coaching team discusses tactics: instead of reviewing every frame of a match, they reason over concise symbolic descriptions of what happened, where it occurred, and how the players responded. To construct this structured representation, we use the ShuttleSet dataset [42], which provides stroke-level annotations of badminton rallies. Each entry records the rally index, frame index, player positions, shot types, and rally outcomes. In practice, such game logs are not difficult to obtain. Modern event-based detection models can automatically extract similar logs from broadcast videos [12, 20, 24]. Since this data extraction step is not the focus of our work, we directly use the annotated game logs available in ShuttleSet as our input.

While the ShuttleSet dataset provides detailed rally annotations, its positional data is expressed in the coordinate system of the original video frames. Since different match videos are recorded from varying camera angles, analyzing trajectories directly in these video coordinates is ineffective. We therefore perform a geometric calibration by estimating a homography matrix for each match based on the detected court lines [26], and then transforming the 2D pixel coordi-

nates into a unified court view. This standardization ensures that player positions can be interpreted consistently across different recordings. Next, we dynamically assign player roles based on the serving side in the first stroke, defining the top-court and bottom-court players according to their Y-axis position relative to the net. This role normalization is crucial for tactical reasoning, as it allows subsequent agents to focus on strategic interactions.

While ShuttleSet provides detailed annotations of events and positions, it does not include information about player poses. We regard pose as an important complementary cue because it reveals how balanced and controlled a player's body is during a shot. This information enhances the understanding of both the current rally and the evolving game context, offering insight into momentum and tactical intent. To capture this aspect, we integrate a pose analysis stage using UniPose [2]. Guided by the coordinates recorded in the log, YOLOv8 [37] is used to crop key frames showing the active hitter. For rallies ending with a winning shot, we also capture the opponent's reaction by locating the frame closest to the shuttle's landing point. UniPose processes each cropped frame to generate concise textual pose descriptions, such as *"the player stands with knees slightly bent, right foot forward, and the racket arm raised."* These descriptions are added to the game log for the debate team to analyze. We provide an example of the structured game log in *Appendix A1* to help readers understand the type of information on which the tactical team bases its analysis.

3.2. Log-Based Reasoning and Verification

The MADR framework simulates human expert collaboration by decomposing the reasoning process into a series of interdependent cognitive steps. These steps progress from divergent debate (steps 1 and 2) to narrative generation (step 3), followed by peer review (step 4), refinement (step 5), and verification (step 6). All discussions within MADR are strictly grounded in the structured game log, which serves as the single reference for discussions. The following paragraphs describe each step in detail, while *Algorithm 1* and *Appendix A2* (both in the Appendix) provide additional implementation details and prompt configurations.

Step 1: Independent Analyses Based on the Game Log. The system first prompts two specialized agents, *Offense Analyst* and *Defense Analyst*, to interpret the game log independently. Each agent examines the rally from a distinct tactical perspective. *Offense Analyst* focuses on how the attacking player establishes initiative and converts opportunities, while *Defense Analyst* concentrates on how the opponent anticipates, defends, or counteracts. Both agents produce self-contained analytical commentaries that capture their reasoning about player decisions, stroke effectiveness, and situational intent. These initial interpretations serve as the foundation for later debate and discussion.

Step 2: Divergent Debate and Analytical Refinement. After completing their individual analyses, the agents exchange their commentaries and begin a structured debate. Each agent reviews the other's reasoning, identifies weaknesses or overlooked evidence, and formulates a response. This dialectical process encourages exploration of multiple tactical explanations rather than premature convergence on a single interpretation. It mirrors how human analysts challenge one another to refine collective understanding.

The dynamics of this debate are governed by a contentiousness parameter c, which modulates the tone and stance of interaction through prompt-based conditioning [5]. When c is high, both agents adopt a critical and confrontational tone, actively identifying contradictions or missing evidence in the opponent's reasoning. When c is low, the interaction becomes cooperative and integrative, as the agents work together to merge perspectives into a coherent tactical interpretation. This adaptive prompting strategy enables the debate to naturally evolve from adversarial challenge to constructive synthesis, resembling how an expert coaching team moves from disagreement to consensus.

Step 3: Rally Narrative Generation. After the debate concludes, the system assigns the synthesis task to an agent acting as the *Tactic Summarizer*. This agent does not introduce new arguments but serves as a neutral integrator that reconciles differing viewpoints. It receives the complete debate history and is instructed to evaluate the key arguments, resolve conflicting interpretations, and weave the most valuable insights from both sides into a single, fluent description of the rally. The tactic summarizer's role is to convert multi-agent dialogue into a coherent report $\mathbf{R}_{v1}$.

Step 4: Analytical Peer Review. The tactical interpretation of report $\mathbf{R}_{v1}$ may be limited in analytical depth. Two domain-specialized agents, acting as *Reviewers*, evaluate the quality of $\mathbf{R}_{v1}$ from offensive and defensive perspectives. They assess whether the narrative merely describes what happened or also explains why it happened and what tactical meaning it conveys. Each reviewer provides structured feedback in the form of enhancement instructions, denoted as $\mathbf{R}_{review}$, identifying sections that lack analytical depth and offering targeted revision suggestions.

Step 5: Integration and Revision. If either reviewer requests modification, the *Tactic Summarizer* agent is reengaged to perform the revision. It integrates the reviewers' suggestions $\mathbf{R}_{review}$ to generate a refined report $\mathbf{R}_{v2}$. In this stage, the *Tactic Summarizer* agent is expected to expand the analysis beyond descriptive accuracy and articulate deeper tactical reasoning, highlighting the intentions, strategies, and situational implications that underlie each rally.

Step 6: Rally Narrative Verification. While the *Tactic Summarizer* agent is encouraged to produce a contextually rich report $\mathbf{R}_{v2}$, it may introduce inaccuracies during generation or revision. To ensure alignment with the underlying

rally data, the system invokes a dedicated *Verification Agent* that maintains consistency between the generated narrative and the structured log. This agent performs a sentence-level comparison, verifying each statement against the corresponding events recorded in the log, with particular attention to player-stroke attributions and rally outcomes. Any discrepancies are immediately corrected, yielding a verified version $\mathbf{R}_{final}$. This design allows the framework to combine interpretive depth with data-grounded precision, supporting human-like analysis while preserving strict factual consistency.

3.3. Rally Video Retrieval

After rally videos are converted into textual narratives through the MADR framework, we enable tactical video search using a standard retrieval-augmented generation (RAG) architecture. The pipeline consists of two main components: *offline indexing* and *online retrieval*, as illustrated in *Algorithm 2* in *Appendix A3*.

Offline Indexing. Each verified report $\mathbf{R}_{final}$ is incorporated into the retrieval system through offline indexing. We encode each report into a high-dimensional vector using a text embedding model (e.g., Qwen3-Embedding-8B [53]), and store the vector with its metadata (e.g., the link to the corresponding rally video) in a database.

Online Retrieval. When a user submits a natural-language query $\mathbf{Q}$, the system performs retrieval in two steps to balance efficiency and precision. First, the query is encoded using the same embedding model to obtain a query vector, which is then used to perform a fast approximate nearest neighbor (ANN) search [3, 15] over the narrative embeddings. This coarse retrieval step quickly filters out irrelevant chunks and retrieves the top-K candidates most likely to match the query. In the second step, a more expressive re-ranking model (e.g., Qwen3-Reranker-8B [38]) compares $\mathbf{Q}$ with the full rally reports associated with these candidates and assigns refined relevance scores based on deeper semantic similarity.

4. Experiments and Results

4.1. Benchmark Dataset Generation

Narrative and Query Generation. To evaluate the performance of our retrieval system, we constructed a benchmark dataset that pairs natural-language queries with corresponding badminton rally videos. We began by employing a team of Gemini-2.5-Flash agents [13], configured as a tactical debate team following the process described in Section 3, to generate rally narratives for all matches in the ShuttleSet dataset. For each narrative, the Qwen3-VL-32B-Thinking [38] [1] model was used to generate multiple candidate queries that a coach or analyst could plausibly use to retrieve the corresponding video segment. To ensure query quality, each generated query was then evaluated by Qwen3-VL-32B-Thinking across four perspectives: clarity, specificity, challenge, and answerability, each scored on a five-point scale. The scores were combined into a composite quality metric, and the top 50% of the highest-scoring queries were retained for further filtering.

Although the selected queries were of high quality, many were semantically similar. To reduce redundancy, we applied the maximum marginal relevance (MMR) algorithm [4] to select the most representative subset. The MMR scoring considered both the quality score obtained from the previous step and the embedding similarity between queries, computed using Gemini-Embedding-001 [22]. By balancing these two factors, the algorithm greedily selected a compact set of diverse and informative queries for each rally. Finally, through manual inspection, we found that the resulting queries naturally fall into three categories: factual queries, relational queries, and strategic reasoning queries. Detailed descriptions of these categories, along with example queries, are provided in *Appendix A4*.

Query-Video Matching. The final stage focused on constructing answers that define the correct video rally for each query. Since evaluating every possible query-video pair would be computationally prohibitive, we first reduced the search space by building a candidate pool using a hybrid retrieval method that combines sparse retrieval (BM25 [23, 36]) and dense retrieval (Gemini Embedding [22] with FAISS [15]). The results from both methods were merged through reciprocal rank fusion [14, 34] to improve recall, ensuring that potential matches were not missed. This filtering step greatly accelerates the subsequent evaluation by retaining only a small set of promising candidates. We then applied a dual-verification process using the Gemini-2.5-Pro Thinking model [13] to ensure both tactical and factual correctness. The model assessed whether each candidate narrative reflected the tactical intent expressed in the query, and whether the described events were consistent with the underlying game log, including player-stroke attribution and rally outcomes. Only pairs that satisfied both tactical and factual criteria were retained as the final gold-standard query-video set for evaluation.

4.2. Benchmark Dataset Validation

We first assessed the quality of the dataset since the positive query-video pairs were generated automatically by the LLM. To verify the reliability of these ground-truth annotations, we randomly sampled 50 positive pairs from the dataset and created another 50 negative pairs by randomly mismatching queries and narratives, yielding a total of 100 pairs. The Gemini-2.5-Pro Thinking model, GPT-5.2 Thinking model, and three badminton experts (one team

[1] We used the text-encoder of the VLM in this process because its performance exceeds that of its LLM counterpart with the same model size.

Table 1. Performance comparison of text-to-video retrieval methods on the badminton dataset. We compare our approach with the state-of-the-art VLMs and a strong baseline using the same embedding model applied directly to raw game logs. The best results are highlighted in bold.

	Method	Hit@K (%) ↑			Recall@K (%) ↑			MAP (%) ↑	MdR ↓	MnR ↓
		H@1	H@5	H@10	R@1	R@5	R@10			
Baselines	*Multimodal Embeddings*									
	VLM2Vec (2B) [19]	3.04	10.43	15.65	0.01	0.15	0.41	3.40	104.0	319.79
	VLM2Vec (7B) [19]	4.35	10.00	17.39	0.03	0.16	0.80	3.50	105.0	350.84
	VLM2Vec-V2 (2B) [30]	0.87	11.30	16.96	0.00	0.23	0.64	3.21	115.5	341.00
	Ops-MM-v1 (2B)	2.17	10.00	13.48	0.03	0.17	0.28	3.42	125.0	313.76
	Ops-MM-v1 (7B)	2.17	8.26	10.43	0.10	0.21	0.27	3.50	125.0	328.21
	RzenEmbed-v2 (7B) [18]	5.65	12.17	18.26	0.11	0.41	0.62	4.10	95.5	308.20
	CLIP-based Video retrieval									
	CLIP-ViP (B/16) [47]	2.61	8.70	13.48	0.01	0.09	0.34	3.66	132.5	279.26
	CLIP4Clip (B/16) [25]	1.74	6.96	13.48	0.02	0.08	0.82	3.70	122.0	346.63
	XCLIP (B/16) [32]	2.17	8.26	14.35	0.05	0.09	0.75	3.64	99.0	326.07
	Raw Rally Data									
	Qwen3-Embedding (0.6B)	10.00	24.35	37.83	0.73	2.99	6.31	10.56	19.0	107.79
	Qwen3-Embedding (0.6B) + Rerank	19.57	36.52	43.91	1.96	6.57	9.38	12.96	18.5	106.48
	Qwen3-Embedding (4B)	10.87	27.83	37.83	0.84	3.76	7.08	10.10	18.0	89.01
	Qwen3-Embedding (4B) + Rerank	13.91	29.13	38.26	1.70	4.82	7.95	10.95	22.0	89.27
	Qwen3-Embedding (8B)	10.00	28.26	40.87	1.51	5.42	8.22	12.26	21.0	105.57
	Qwen3-Embedding (8B) + Rerank	21.30	40.00	46.09	3.94	9.33	10.96	15.79	13.0	103.52
Ours	*Ours (Tactic Debate Team: Qwen3-8B-without-thinking Agents)*									
	Qwen3-Embedding (0.6B)	18.70	34.35	44.78	2.71	5.53	7.86	12.48	13.0	94.58
	Qwen3-Embedding (0.6B) + Rerank	17.39	37.83	46.96	2.53	5.97	8.61	12.52	13.0	94.23
	Qwen3-Embedding (4B)	23.91	41.74	50.43	2.40	5.41	8.35	13.20	10.0	75.50
	Qwen3-Embedding (4B) + Rerank	27.39	45.65	51.30	3.36	7.12	9.75	14.73	9.0	75.13
	Qwen3-Embedding (8B)	17.83	39.13	50.00	1.40	4.39	8.60	11.15	10.5	83.19
	Qwen3-Embedding (8B) + Rerank	27.39	47.83	57.39	4.72	8.67	11.40	14.74	6.0	81.64
	Ours (Tactic Debate Team: Gemini-2.5-Flash-without-thinking Agents)									
	Qwen3-Embedding (0.6B)	39.57	63.48	72.17	6.92	12.24	16.92	22.45	2.0	25.05
	Qwen3-Embedding (0.6B) + Rerank	36.09	69.13	76.52	4.93	14.64	18.80	22.36	2.0	24.47
	Qwen3-Embedding (4B)	40.87	66.52	74.35	7.57	13.96	18.01	23.09	2.0	24.41
	Qwen3-Embedding (4B) + Rerank	**55.65**	**77.39**	**83.04**	**11.57**	**19.86**	**25.07**	**30.32**	**1.0**	**22.50**
	Qwen3-Embedding (8B)	33.91	61.30	69.13	4.76	11.17	15.73	19.18	3.0	41.00
	Qwen3-Embedding (8B) + Rerank	55.22	74.78	79.13	10.18	17.92	21.27	26.44	1.0	38.58

coach and two collegiate players) independently evaluated these pairs. The model produced two scores, analytical and factual, while the expert provided a single satisfaction score reflecting how well each retrieved video matched the intended query. All scores were normalized to the range [0, 1]. We then computed Pearson and Spearman correlations, as well as mean absolute error (MAE), between the experts' ratings and the models' scores. As shown in Figure 4 (left), the results indicate a strong consistency (correlation > 0.7) between human and both oracle models' judgments of retrieval quality, proving our automatic generation pipeline is reliable and unbiased.

4.3. Retrieval Performance

To evaluate the effectiveness of our system, we conducted experiments covering both the offline indexing and online retrieval stages, examining how different configurations perform on our benchmark dataset. For offline indexing, we compared three representations: (1) raw game logs converted from rally videos, (2) tactical narratives generated by the Qwen3 debate team, and (3) narratives produced by the Gemini-2.5-Flash debate team. For online retrieval, we evaluated multiple embedding model sizes and measured the impact of using a re-ranking module versus direct embedding-based search.

Metrics. We first describe the evaluation metrics and

MADR (Gemini-2.5-flash w/o Thinking)

The Bottom-court Player responded to the Top-court Player's initial short service with a net shot (Shot 2). This net shot set up the subsequent lob from the Top-court Player, which then became the perfect opportunity for the smash (Shot 4). Upon receiving a lob, the Bottom-court Player immediately transitioned to offense with a powerful smash (Shot 4). This aggressive shot forced a weak return, which the Bottom-court Player capitalized on with a rush (Shot 6) from the front court. The rush was specifically aimed at exploiting the Top-court Player's recovery time and position after the weak return net (Shot 5), thereby compressing the court and forcing another defensive, easily attackable shot. **The Bottom-court Player's strategy was characterized by an 'attack-and-follow' pattern**, where each aggressive shot was designed to set up the next, demonstrating deliberate timing and placement to generate the final winning opportunity. This culminated in a final decisive smash (Shot 8) after the Top-court Player's defensive return, demonstrating effective court compression and exploitation of spatial vulnerabilities, particularly **capitalizing on the Top-court Player's 'below the net' defensive return drive (Shot 7) to set up the winning opportunity.**

VLM (Gemini-2.5-Pro w/ Thinking)

The Bottom-court player controlled the initial phase, transitioning from a neutral rally to a strong attacking position with a well-executed straight smash. His follow-up cross-court net shot was aggressive **and well-placed. However, he was tactically outmaneuvered at the net by the Top-court Player's precise counter-shot, which forced him onto the defensive. The resulting lift was a necessary but ultimately fatal concession of the attack, and his subsequent defensive block was insufficient to handle the pressure, leading to an unforced error.**

Figure 3. We compared the narratives generated by our MADR system with those produced by a powerful cloud-based vision-language model. The tactic debate team consisted of Gemini-2.5-Flash non-thinking agents, while the VLM used Gemini-2.5-Pro thinking models. Despite its heavy computational capacity, the VLM failed to generate an accurate narrative, as the bottom-court player actually won the rally. For clarity, we highlight insightful sentences in green and incorrect ones in red, respectively.

baseline settings used in our study. Retrieval performance was evaluated using hit rate and recall at top-1, top-5, and top-10, as well as mean average precision (MAP), mean rank (MnR), and median rank (MdR). Hit rate and recall measure the proportion of queries for which at least one relevant video appears within the top-k results. MAP provides a comprehensive evaluation by averaging the precision at every rank where a correct video is retrieved, rewarding systems that not only retrieve all relevant results but also rank them higher. MnR computes the average rank of the first relevant result across successful queries (lower is better), while MdR reports the median of these ranks to mitigate the influence of outliers. The formulas of these metrics are provided in *Appendix A5*.

Baselines. We adopted several representative VLMs from the massive multimodal embedding benchmark leaderboard [30], selecting those with publicly available pre-trained weights. Specifically, we included VLM2Vec v2 [19, 30], Ops-MM-v1, CLIP-ViP [47], XCLIP [32], and RzenEmbed-v2 [18]. These models represent the prevailing alignment-based paradigm for text-video retrieval and provide a strong reference for evaluating our framework.

Results. As summarized in Table 1, standard VLM baselines show limited effectiveness for specialized tactical queries. Retrieved videos were often visually similar yet tactically irrelevant, revealing their difficulty in distinguishing surface-level appearances from the strategic intent implied by the query. This limitation is expected: alignment-based models lack an explicit mechanism for directly extracting tactical information from video pixels. To further probe this issue, we prompted the VLM version of the Gemini-2.5-Pro Thinking model to generate tactical narratives for individual rallies and compared its outputs with those produced by our MADR framework. As shown in Figure 3, even this strong thinking model exhibited unavoidable hallucinations and produced narratives that were noticeably generic and lacking in tactical specificity. This contrast highlights why alignment alone is insufficient for domain-level tactical understanding.

When retrieval was instead performed using structured textual game logs, accuracy improved substantially. Even a simple embedding similarity search between the query and log text yielded higher scores across all metrics, and a re-ranking stage provided an additional gain. This confirms that structured, domain-specific representations can significantly enhance retrieval precision. However, because these logs merely describe what happened, they remain insufficient for queries that require causal understanding or intent reasoning, such as *"Find rallies where a player was forced into a weak return."*

Further experiments examined retrieval based on the narratives produced by the MADR framework. Searching over these narratives yielded consistent performance improvements across all metrics (Table 1). Narratives generated by the Qwen3-8B team outperformed the game-log baseline, highlighting that explicit reasoning substantially benefits

Table 2. Ablation study of MADR components. All results are obtained using the Qwen3-4B embedding model with re-ranking.

Method Configuration	H@1	H@5	H@10
MADR (Full Pipeline)	**55.65**	**77.39**	**83.04**
w/o Review-Revision	50.87	70.00	75.65
w/o Debate	40.00	66.96	71.74
w/o Debate & Review-Revision	40.00	66.09	72.17

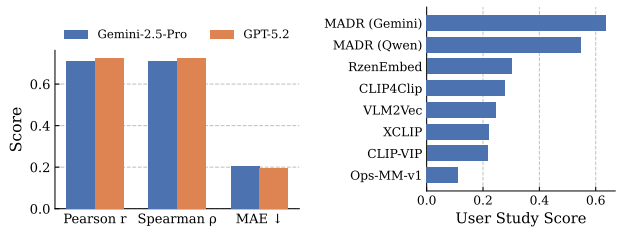

Figure 4. (Left) Correlation analysis of benchmark data quality, comparing scores produced by LLM-based evaluators with the judgments of human experts. (Right) Average satisfaction scores assigned by human experts to the retrieved results for each query.

tactical retrieval. Upgrading the debate agents to Gemini-2.5-Flash further improved results, suggesting that reasoning quality directly influences retrieval effectiveness.

Our experiment results also revealed a few system-level details. First, we noticed that the Qwen3-Embedding-4B (2560D) performed slightly better on the query task than the larger Qwen3-Embedding-8B (4096D). We suspect this may be due to the curse of dimensionality, where the much larger 4096D space made the similarity search less effective. Second, we analyzed the interplay between the re-ranker and text quality. We observed that when the narratives were of high quality (i.e., from the Gemini team), the additional benefit from a less-powerful re-ranker was minimal. Finally, while our Recall@K ($k = 1, 5, 10$) scores appear modest, this is an expected outcome because the total number of ground-truth-positive videos in the dataset far exceeds 10.

Ablation Study. To evaluate the contribution of each component within the MADR pipeline, we conducted an ablation study comparing the full system against several reduced variants. Specifically, we removed (1) the review-revision stage, (2) the debate stage, and (3) both debate and review-revision stages. As summarized in Table 2, removing any part of the reasoning pipeline led to noticeable drops in hit rate. The results indicate that both the adversarial debate and the iterative refinement stages play essential roles in producing high-quality tactical narratives.

User study. To assess retrieval quality from a user's perspective, we conducted a user study with a team coach and two collegiate players. We evaluated 25 queries in total, comprising evenly and randomly sampled queries from the categories in Section 4.1 and an additional 10 queries authored specifically by the coach. For each query, the top-1 retrieved video from every method was collected and presented on a single evaluation page. Two variants of the MADR-generated narratives were included: one produced by the Qwen3 debate team and the other by the Gemini-2.5-Flash team. Both variants used the same Qwen3-Embedding-4B model for online retrieval. The experts viewed each query and assigned a score between 0 and 1 to every retrieved video. All videos were shown in a randomly shuffled order to prevent positional bias. Figure 4 (right) shows the average scores. The state-of-the-art VLMs received low scores, consistent with their difficulty handling tactical intent. In contrast, our MADR received a clear user preference, achieving a leading score of 0.64.

4.4. Limitations

Although the proposed MADR framework substantially outperforms alignment-based VLM baselines, it still faces several limitations. First, the game logs used in our evaluation combine ShuttleSet annotations with a portion of automatically extracted pose descriptions. Despite the strong accuracy of the CV models, relying solely on automatically extracted logs can degrade retrieval performance. Second, even with high-quality narratives produced by the tactic debate team, during retrieval, the system continues to struggle with highly compositional tactical queries that require long-range reasoning or multi-step causal interpretation. For example, *"Find rallies where a player exploits an opponent's initial error to establish and maintain net dominance, systematically dismantling the opponent's defensive structure through aggressive shot selection and timing."* In such cases, the current embedding-based retrieval and lightweight re-ranking may fail to retrieve relevant matches.

5. Conclusions

This work demonstrates that retrieving high-level tactical semantics from sports videos requires reasoning rather than text-video alignment. While structured perception captures what happened, tactical queries depend on understanding why it happened, which generic VLMs cannot model. Our multi-agent debate framework fills this gap by converting game logs into grounded tactical narratives, enabling effective retrieval of complex, intent-driven scenarios and providing interpretable justifications linked to specific game events. Although the overall retrieval performance still has room to improve, particularly for highly compositional tactical queries, the results show that decoupling perception from reasoning is essential for expert-level tactical video search. Moreover, this paradigm can naturally extend to other domains that already possess domain-specific machine learning tools capable of producing structured representations for downstream reasoning.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments. We thank Edward Y. Chang and Chih-Chuan Wang for insightful discussions and for helping us with the user study. This work is partially supported by the National Science and Technology Council (NSTC), Taiwan, under Contract: 113-2221-E-A49 -149 -MY3, 114-2221-E-A49 -129 -MY3, and 115-2425-H-A49 -001 -, and by the Higher Education Sprout Project of the National Yang Ming Chiao Tung University and Ministry of Education (MOE), Taiwan.

References

[1] Mohamed Salim Aissi, Clémence Grislain, Mohamed Chetouani, Olivier Sigaud, Laure Soulier, and Nicolas Thome. Viper: Visual perception and explainable reasoning for sequential decision-making. *arXiv preprint arXiv:2503.15108*, 2025. 2

[2] Bruno Artacho and Andreas Savakis. Unipose: Unified human pose estimation in single images and videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7035–7044, 2020. 4

[3] Sunil Arya, David M Mount, Nathan S Netanyahu, Ruth Silverman, and Angela Y Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)*, 45(6):891–923, 1998. 5

[4] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, 1998. 5

[5] Edward Y Chang. Examining gpt-4's capabilities and enhancement with socrasynth. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 7–14. IEEE, 2023. 2, 4

[6] Edward Y Chang. Socrasynth: Multi-llm reasoning with conditional statistics. *arXiv preprint arXiv:2402.06634*, 2024. 2

[7] Edward Y. Chang. Evince: Optimizing multi-llm dialogues using conditional statistics and information theory, 2025. 2

[8] Edward Y Chang and Longling Geng. Sagallm: Context management, validation, and transaction guarantees for multi-agent llm planning. *arXiv preprint arXiv:2503.11951*, 2025. 2

[9] Jocelyn J Chang and Edward Y Chang. Socrahealth: enhancing medical diagnosis and correcting historical records. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1400–1405. IEEE, 2023. 2

[10] Kai-Shiang Chang, Wei-Yao Wang, and Wen-Chih Peng. Where will players move next? dynamic graphs and hierarchical fusion for movement forecasting in badminton. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6998–7005, 2023. 3

[11] 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. 2

[12] Yu-Jou Chen and Yu-Shuen Wang. Tracknetv3: Enhancing shuttlecock tracking with augmentations and trajectory rectification. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, pages 1–7, 2023. 3

[13] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 5

[14] Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759, 2009. 5

[15] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024. 5

[16] Alireza Ghafarollahi and Markus J Buehler. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37 (22):2413523, 2025. 2

[17] Kristina Host and Marina Ivašić-Kos. An overview of human action recognition in sports based on computer vision. *Heliyon*, 8(6), 2022. 3

[18] Weijian Jian, Yajun Zhang, Dawei Liang, Chunyu Xie, Yixiao He, Dawei Leng, and Yuhui Yin. Rzenembed: Towards comprehensive multimodal retrieval. *arXiv preprint arXiv:2510.27350*, 2025. 6, 7

[19] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024. 2, 6, 7

[20] Nyan-Ping Ju, Dung-Ru Yu, Tsi-Ui Ik, and Wen-Chih Peng. Trajectory-based badminton shots detection. In *2020 International Conference on Pervasive Artificial Intelligence (ICPAI)*, pages 64–71. IEEE, 2020. 3

[21] Stephanie A Kovalchik. Player tracking data in sports. *Annual Review of Statistics and Its Application*, 10(1):677–697, 2023. 3

[22] Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, et al. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*, 2025. 5

[23] Xianming Li, Julius Lipp, Aamir Shakir, Rui Huang, and Jing Li. Bmx: Entropy-weighted similarity and semantic-enhanced lexical search. *arXiv preprint arXiv:2408.06643*, 2024. 5

[24] Yun-Hsuan Lien, Chia-Tung Lian, and Yu-Shuen Wang. Shuttleflow: learning the distribution of subsequent badminton shots using normalizing flows. *Machine Learning*, 114(2):39, 2025. 3

[25] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 2, 6

[26] Hui Ma and Xuan Ding. Robust automatic camera calibration in badminton court recognition. In *2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, pages 893–898. IEEE, 2022. 3

[27] Weiyu Ma, Qirui Mi, Yongcheng Zeng, Xue Yan, Yuqiao Wu, Runji Lin, Haifeng Zhang, and Jun Wang. Large language models play starcraft ii: benchmarks and a chain of summarization approach. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 2

[28] Neelu Madan, Andreas Møgelmose, Rajat Modi, Yogesh S Rawat, and Thomas B Moeslund. Foundation models for video understanding: A survey. *arXiv preprint arXiv:2405.03770*, 2024. 2

[29] Mehrtash Manafifard, Hamid Ebadi, and H Abrishami Moghaddam. A survey on player tracking in soccer videos. *Computer Vision and Image Understanding*, 159:19–46, 2017. 3

[30] Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, et al. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025. 2, 6, 7

[31] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13235–13245, 2024. 2

[32] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European conference on computer vision*, pages 1–18. Springer, 2022. 2, 6, 7

[33] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186, 2024. 2

[34] Zackary Rackauckas. Rag-fusion: a new take on retrieval-augmented generation. *arXiv preprint arXiv:2402.03367*, 2024. 5

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[36] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. *Okapi at TREC-3*. British Library Research and Development Department, 1995. 5

[37] Mupparaju Sohan, Thotakura Sai Ram, and Ch Venkata Rami Reddy. A review on yolov8 and its advancements. In *International Conference on Data Intelligence and Cognitive Informatics*, pages 529–545. Springer, 2024. 4

[38] Qwen Team. Qwen3 technical report, 2025. 5

[39] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023. 2

[40] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. *Advances in neural information processing systems*, 35:5696–5710, 2022. 2

[41] Wei-Yao Wang, Hong-Han Shuai, Kai-Shiang Chang, and Wen-Chih Peng. Shuttlenet: Position-aware fusion of rally progress and player styles for stroke forecasting in badminton. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4219–4227, 2022. 3

[42] Wei-Yao Wang, Yung-Chang Huang, Tsi-Ui Ik, and Wen-Chih Peng. Shuttleset: A human-annotated stroke-level singles dataset for badminton tactical analysis. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5126–5136, 2023. 3

[43] Xin Wang, Haoyang Li, Zeyang Zhang, Haibo Chen, and Wenwu Zhu. Modular machine learning: An indispensable path towards new-generation large language models. *arXiv preprint arXiv:2504.20020*, 2025. 2

[44] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 2

[45] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2.5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025. 2

[46] Fei Wu, Qingzhong Wang, Jiang Bian, Ning Ding, Feixiang Lu, Jun Cheng, Dejing Dou, and Haoyi Xiong. A survey on video action recognition in sports: Datasets, methods and applications. *IEEE Transactions on Multimedia*, 25:7943–7966, 2022. 3

[47] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022. 2, 6, 7

[48] Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. Follow the rules: Reasoning for video anomaly detection with large language models. In *European Conference on Computer Vision*, pages 304–322. Springer, 2024. 2

[49] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 2

[50] Guibin Zhang, Luyang Niu, Junfeng Fang, Kun Wang, Lei Bai, and Xiang Wang. Multi-agent architecture search via

agentic supernet. In *Proceedings of the 42nd International Conference on Machine Learning*. JMLR.org, 2025. 2

[51] Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xiong-Hui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. AFlow: Automating agentic workflow generation. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[52] Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. Llm as a mastermind: A survey of strategic reasoning with large language models, 2024. 2

[53] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. 5