Nash-Pruned CredMAS: Dynamic Panel Pruning for VLM-MAS using Nash-based Selection and Doubly-Robust Credits

Yijia Fan[1], Mingyu Liu[1], Jing Yang[1], Jian Wang[2],
Keze Wang[1]*, Jusheng Zhang[1]*
[1]Sun Yat-sen University, [2]Snap Inc.
{kezewang, jushengzhang88889}@gmail.com

Abstract

Multi-round Vision-Language Model (VLM) Multi-Agent Systems (MAS) offer powerful reasoning capabilities but suffer from prohibitive costs due to static panel designs, where all N agents communicate at every T round. This approach is fundamentally inefficient, as it ignores the *context-dependent* and *diminishing* marginal utility of specific agents. To address this, we propose **Nash-CredMAS**, an economic framework that transforms agent selection into a **dynamic resource allocation game**. Unlike heuristic routing or one-time pruning, our method operates in two phases: (1) **Offline Causal Value Learning**, where we employ a **doubly-robust (AIPW) estimator** to train a context-aware value function from biased interaction logs, effectively learning the true marginal contribution of agents; and (2) **Online Dynamic Auctions**, where agents bid for communication slots based on their predicted utility. We formulate the inference-time selection as a **submodular maximization problem** under budget constraints, theoretically guaranteeing a $(1 - 1/e)$-approximation of the optimal coalition via a greedy strategy. Empirically, Nash-CredMAS achieves state-of-the-art results on challenging benchmarks, including MMMU and V*-Bench, while reducing token consumption by over 25% compared to static baselines. The system naturally converges to an economic equilibrium where agents actively remain silent when their marginal value does not justify the cost.

1 Introduction

The rapid evolution of Large Language Models (LLMs) (Radford et al., 2019; Gemini Team, 2025; DeepSeek-AI, 2025) and, more recently, Vision-Language Models (VLMs) (Bai et al., 2025; OpenAI, 2024; Liu et al., 2023) has catalyzed a new frontier in artificial intelligence: Multi-Agent Systems (MAS) (Guo et al., 2024; Chen et al., 2025; Tran et al., 2025). These systems, where multiple LLM and VLM agents collaborate, debate(Du et al., 2023; Khan et al., 2024), or compete, have shown remarkable promise in solving complex, multimodal problems far exceeding the capacity of a single model, from intricate visual reasoning to sophisticated game-theoretic scenarios(Zhang et al., 2025b; Wang et al., 2024).

However, this enhanced capability comes at a steep, often prohibitive, cost. Many prominent multi-round MAS frameworks, such as those based on structured debates or game-theoretic interactions, employ a *static panel* of experts. In this paradigm, all N agents are invoked to contribute at every T round. This design leads to a multiplicative cost explosion: the total operational cost scales linearly with both the number of agents and the number of interaction rounds, i.e., $O(N \times T)$. This static, all-speak approach is fundamentally inefficient, as it fails to account for the heterogeneous, context-dependent, and often diminishing marginal utility of each agent's contribution.

This inefficiency raises a critical question: **is it possible to remodel the static MAS panel as a *dynamic market-based system*?** Can we design a framework that intelligently allocates communication resources, allowing high-utility agents to bid for slots while silencing those who are redundant or irrelevant to the current context? Such a system could dramatically reduce communication overhead while simultaneously enhancing—or maintaining—task performance by filtering out noise.

To address this challenge, we argue that to make multi-round dialogue both cost-effective and performant, the MAS must learn to **dynamically allocate its budget** (i.e., perform round-by-round selection) based on economic consensus.

In this work, we propose **Nash-CredMAS**, a novel framework that operationalizes this "dynamic allocation" strategy. Our approach is bifurcated into two stages. First, in an **offline learning phase**,

* Corresponding authors.

we address the challenge of selection bias in interaction logs. We employ a doubly-robust estimator (AIPW) to provably and accurately estimate the marginal credit of each VLM agent's multi-modal message. These unbiased signals are then used to train a context-aware *Value Function*, which predicts agent-level economic metrics—specifically the potential Net Utility (NU)—serving as a reliable measure of an agent's "value density" given any conversation history.

During the **inference phase** on a new task, the system operates as a dynamic auction mechanism. At each round t, we frame the agent selection for round $t+1$ as a resource allocation game grounded in game-theoretic principles(Hosseini and Vazirani, 2023; Navon et al., 2022; Schuster, 2017), using the Value Function to evaluate agents' potential bids. A greedy algorithm provides a computationally efficient solution to this problem, treating agent selection as a submodular maximization task. This method is supported by theoretical guarantees, specifically the approximation guarantees for maximizing submodular functions under budget constraints. We observe that this dynamic mechanism naturally converges to a stable state (an economic equilibrium) where agents actively stop communicating when their marginal value outweighs the cost. Experience shows that our VLM-based multi-agent system achieves state-of-the-art results in benchmarks such as MMMU(Yue et al., 2024) and MMBench(Liu et al., 2024b).

2 Related work

VLM and Multi-Agent Systems. The progression from Large Language Models (LLMs) to Vision-Language Models (VLMs) has been a significant leap, endowing models with the ability to reason over and integrate multi-modal information(Liu et al., 2024a; An et al., 2025; Xue et al., 2025). This capability has unlocked new frontiers in complex reasoning tasks that bridge vision and text. A natural and powerful extension of this paradigm is the VLM-based Multi-Agent System (MAS). Early VLM-MAS frameworks have demonstrated remarkable potential in diverse domains, such as collaborative visual dialogu (Das et al., 2017; Hu et al., 2025), embodied navigation (Zhao et al., 2024; Goetting et al., 2024) , and solving complex scientific problems (e.g., ScienceQA(Lu et al., 2022)). These systems, often structured as debates or collaborative panels, aggregate the specialized visual and reasoning skills of multiple agents. However, this amplification of capability comes at the cost of massively increased computational overhead, as the processing of image or video tokens at each round by each agent is somehow expensive.

Efficiency in Multi-Agent Systems. As MAS frameworks scale, communication cost becomes a critical bottleneck (Zheng et al., 2025). Research into efficient MAS has explored several directions. A primary approach involves *static structural constraints*, such as pre-defining roles or fixed communication topologies (e.g., in AutoGen (Wu et al., 2023) or PHP (Zheng et al., 2024)), which limit interactions but lack adaptability. A second direction focuses on *dynamic routing or pruning*. Some methods employ heuristic-based pruning to remove redundant messages post-generation (Zhang et al., 2025a). More advanced techniques utilize learnable components, such as a centralized router or critic, to decide "who speaks next" (Liu et al., 2024d; Zhang et al., 2025c; Yue et al., 2025). These methods often require complex, costly-to-train routing policies or "pre-speak" mechanisms, where agents generate preliminary summaries to determine their relevance, adding another layer of inference cost. Our work, Nash-Pruned CredMAS, diverges significantly from these. We propose an *economic panel pruning* approach. Instead of costly per-message online routing, we perform a single, offline decision between rounds, dynamically selecting a high-utility panel. This selection is not based on heuristics but on robust, doubly-robust credit estimates, framing efficiency as an economic optimization problem rather than a routing problem.

Nash Equilibrium in Agent Selection. The concept of a Nash Equilibrium (NE), originating from non-cooperative game theory, describes a stable state in which no player (agent) can unilaterally improve their outcome by changing their strategy (Nash, 1951). In the context of MAS, this concept has been used to analyze agent interactions and find stable policies (Xie et al., 2025a; Fuente et al., 2024; Xie et al., 2025b). More recently, game-theoretic principles have been applied to agent coordination and selection. For instance, the Nash Bargaining Solution (NBS) provides a framework for finding a fair and Pareto-optimal outcome in a cooperative game. Our work draws inspiration from these principles, framing the agent selection problem as a search for a stable, high-utility panel.

We utilize a Nash-style bargaining objective (maximizing the log-sum of agent utilities) as the theoretical foundation for our selection process.

3 Theoretical Framework

We establish theoretical guarantees for our dynamic allocation framework. We first show that our offline value learning is unbiased despite selection bias in logs. We then prove that our online auction mechanism, modeled as submodular maximization, guarantees a near-optimal allocation at every round.

3.1 Consistency of Causal Value Learning

The foundation of our approach is the ability to learn the *true* marginal contribution of an agent from biased logs. Let $\mathcal{D} = \{(H_t, i, r_t)\}$ be the offline dataset. The observed reward r_t is biased because it depends on the specific policy that generated the data. To correct this, we rely on the Doubly Robust (DR) property of the AIPW estimator used to label our training data.

Let the outcome regression model be $\mu_a(X) = \mathbb{E}[S \mid X, A = a]$ and the propensity score model be $e(X) = P(A = 1 \mid X)$, where $A = 1$ is the "treatment" of including agent i's message and $X = H_t$ is the context.

Proposition 1 (Doubly Robust Consistency)**.** *The AIPW estimator $\hat{y}_{i,t}$ used to train our Value Function V_ϕ is a consistent and asymptotically unbiased estimator for the true marginal contribution $d_{i,t}^{true}$ if **either** (i) the outcome model $\mu_a(X)$ is correctly specified, **or** (ii) the propensity model $e(X)$ is correctly specified.*

Proof. (Standard proof from Robins et al. (1994)). This ensures that our Value Function V_ϕ converges to the true potential utility of an agent as the dataset size $N \to \infty$, even under policy shift. □

3.2 Optimality of the Online Auction

At inference time, our system solves a resource allocation problem at each round t. We now prove that our greedy auction strategy is not merely a heuristic but a theoretically grounded approximation algorithm.

Problem Formulation. Let $U(S \mid H_t) = \sum_{i \in S} V_\phi(H_t, i)$ be the collective utility of a selected panel S. In a multi-agent context, information is often redundant (e.g., two agents explaining the same visual feature). We capture this by modeling $U(\cdot)$ as a **non-negative monotone submodular function**. Submodularity formally captures the property of *diminishing returns*: for $A \subseteq B \subseteq \mathcal{A}$ and agent $x \notin B$, $U(A \cup \{x\}) - U(A) \geq U(B \cup \{x\}) - U(B)$.

Our auction solves the following constrained maximization problem:

$$\max_{S \subseteq \mathcal{A}, |S| \leq K} F(S) \quad \text{where} \quad F(S) = U(S) - \lambda \sum_{i \in S} c_i \tag{1}$$

Theorem 1 (Approximation Guarantee)**.** *For the problem of maximizing a monotone submodular function subject to a cardinality constraint K, the Greedy Algorithm (which iteratively selects the agent with the highest marginal gain) produces a solution set S_{greedy} that satisfies:*

$$F(S_{greedy}) \geq \left(1 - \frac{1}{e}\right) F(S_{opt}) \tag{2}$$

where S_{opt} is the optimal set and $1 - 1/e \approx 0.632$.

Proof. This is a classical result from Nemhauser et al. (1978). By framing our dynamic panel selection as a submodular maximization problem, we guarantee that our computationally efficient auction mechanism achieves at least 63% of the theoretical optimal utility at every single round. This provides a strong worst-case bound that simple heuristic pruning methods lack. □

3.3 Convergence to Economic Equilibrium

Finally, we connect our algorithmic stopping condition to game theory. In our dynamic system, we do not force agents to be silent; they choose to be silent based on the auction outcome.

Definition 2 (Economic Nash Equilibrium)**.** *A state S^* at round t is an Economic Nash Equilibrium if:*

1. *For any active agent $i \in S^*$, its bid covers its cost: $V_\phi(H_t, i) \geq \lambda c_i$.*
2. *For any silent agent $j \notin S^*$, its potential contribution does not justify the cost: $V_\phi(H_t, j) < \lambda c_j$.*

The stopping condition of Algorithm 1 (when $A_{\text{keep}} = \emptyset$) corresponds to reaching an Economic Nash Equilibrium where the marginal utility of all remaining candidates is below the cost threshold. At this point, further communication is economically irrational, and the system naturally converges.

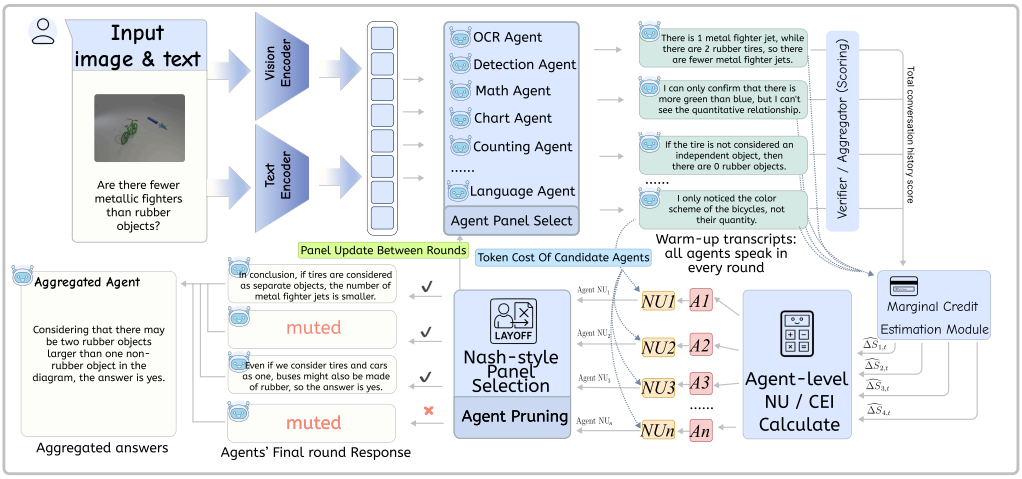

Figure 1: **Overview of Nash-CredMAS. (1) Offline:** We train a context-aware *Value Function* using unbiased AIPW credits derived from interaction logs. **(2) Online:** Agents bid for slots based on predicted utility. A greedy auction mechanism solves a **submodular maximization** problem to dynamically select the optimal coalition, achieving an economic equilibrium where marginal utility justifies the cost.

4 Method: Nash-Pruned CredMAS

We propose **NP-CredMAS**, a framework that transforms the static agent selection problem into a dynamic, round-by-round *resource allocation game*. Unlike prior works that rely on heuristic routing or one-time pruning, our approach is grounded in economic game theory and causal inference.

Formally, we model the inference process as a sequential decision-making problem where, at each round t, a central mechanism (the auctioneer) selects a subset of agents (the panel) $A_{\text{keep}}^{(t)} \subseteq \mathcal{A}$ to maximize the incremental task utility subject to a token budget.

The framework consists of two distinct phases:

1. **Offline Causal Value Learning:** We pre-train a lightweight *Value Function* V_ϕ using a Doubly-Robust (AIPW) estimator to learn the true marginal contribution of agents from offline interaction logs.

2. **Online Dynamic Pruning (Inference):** For a new task, we execute a *Context-Aware Auction* at every round. Agents "bid" based on their predicted utility from V_ϕ, and the system selects the optimal coalition via a greedy approximation of submodular maximization.

4.1 Phase 1: Learning the Causal Value Function (Offline)

A major challenge in training a router is the bias in existing interaction data (e.g., an agent might seem useless only because another agent spoke first). To address this, we employ the **Doubly-Robust AIPW estimator** not as an online metric, but as a *teacher* to train a context-aware value function.

Let $\mathcal{D} = \{(H_t, i, r_t)\}$ be an offline dataset of multi-agent interactions, where H_t is the conversation history, i is the agent, and r_t is the observed reward. We compute the *unbiased* marginal credit label $\hat{y}_{i,t}$ for each sample using the AIPW formulation:

$$\hat{y}_{i,t} = \hat{\mu}_a(H_t) + \frac{\mathbb{I}(A_t = a)}{e(H_t)} (r_t - \hat{\mu}_a(H_t)) \quad (3)$$

where $\hat{\mu}_a$ is the outcome model and $e(H_t)$ is the propensity score.

We then train a lightweight **Value Network** $V_\phi(H_t, i)$, parameterized by ϕ, to predict this causal credit. The training objective is to minimize the Mean Squared Error (MSE) between the predicted utility and the AIPW-estimated credit:

$$\mathcal{L}(\phi) = \mathbb{E}_{\mathcal{D}} \left[(V_\phi(H_t, i) - \hat{y}_{i,t})^2 \right] \quad (4)$$

This network V_ϕ serves as the "Critic" during inference, capable of predicting the *potential* value of an agent i given any context H_t, even if that agent was not active in the original trajectory.

4.2 Phase 2: Inference as a Submodular Auction Game

At inference time on a new task τ', we abandon the static "grace period" pruning. Instead, we perform **dynamic, per-round selection**. We model the selection of agents at round t as maximizing a utility

function $F(S)$ where $S \subseteq \mathcal{A}$ is the selected subset of agents.

The Auction Mechanism. At the start of round t, with history H_{t-1}:

1. **Bidding:** Each candidate agent $i \in \mathcal{A}$ queries the pre-trained Value Network to submit a bid $b_{i,t} = V_\phi(H_{t-1}, i)$. This bid represents the expected marginal utility of agent i speaking in the current context.

2. **Cost Constraint:** Each agent has an associated token cost c_i.

3. **Selection Strategy:** The system solves the following maximization problem:

$$\max_{S \subseteq \mathcal{A}, |S| \leq K} \sum_{i \in S} (b_{i,t} - \lambda \cdot c_i) \tag{5}$$

where λ is a hyperparameter balancing performance and cost.

Handling Redundancy via Submodularity. Crucially, simply summing bids implies additivity, which holds weak in MAS. To capture redundancy (e.g., two agents saying the same thing), we rely on the *context-aware* nature of V_ϕ. If the history H_{t-1} already contains information X, the Value Network predicts a low bid for agent i who would only repeat X. Thus, the objective function effectively behaves as a **submodular function** (diminishing returns).

Greedy Solution. Maximizing a submodular function under cardinality constraints is NP-hard. However, it is theoretically proven that a **Greedy Algorithm** achieves a $(1 - 1/e)$-approximation of the optimal solution. Therefore, our algorithm (Alg. 1) greedily selects agents with the highest *marginal utility-to-cost ratio* until the budget or context window is filled.

4.3 Connection to Nash Equilibrium

The stopping condition in Algorithm 1 (Line 16) corresponds to an **approximate Nash Equilibrium**. When no remaining agent in the pool has a bid $b_{i,t}$ that exceeds the cost threshold (or when the marginal gain of adding any agent is negative), the system reaches a stable state where no agent can unilaterally improve the collective utility by speaking. This aligns with the concept of a *cooperative game equilibrium* under resource constraints.

Algorithm 1 NP-CredMAS: Dynamic Submodular Auction

1: **Input:** Task q; Agent set $\mathcal{A}$; Value Net V_ϕ; Budget K.
2: **Initialize:** $H_0 \leftarrow \{q\}$
3: **for** $t = 1$ **to** T **do**
4: $A_{\text{pool}} \leftarrow \emptyset$; $A_{\text{keep}}^{(t)} \leftarrow \emptyset$
5: ▷ Phase 1: Bidding (Candidate Pruning)
6: **for** $i \in \mathcal{A}$ **do**
7: $b_{i,t} \leftarrow V_\phi(H_{t-1}, i)$
8: **if** $b_{i,t} >$ Threshold **then** $A_{\text{pool}} \leftarrow A_{\text{pool}} \cup \{i\}$
9: **end if**
10: **end for**
11: ▷ Phase 2: Greedy Selection (Submodular Max)
12: **while** $|A_{\text{keep}}^{(t)}| < K$ **and** $A_{\text{pool}} \neq \emptyset$ **do**
13: $i^* \leftarrow \text{argmax}_{i \in A_{\text{pool}}} \frac{U(A_{\text{keep}}^{(t)} \cup \{i\}) - U(A_{\text{keep}}^{(t)})}{c_i}$
14: $A_{\text{keep}}^{(t)} \leftarrow A_{\text{keep}}^{(t)} \cup \{i^*\}$; $A_{\text{pool}} \leftarrow A_{\text{pool}} \setminus \{i^*\}$
15: **end while**
16: ▷ Phase 3: Execution
17: **if** $A_{\text{keep}}^{(t)} = \emptyset$ **then break** ▷ Equilibrium (Silence)
18: **end if**
19: $(H_t, \mathbf{m}_t) \leftarrow \text{ExecuteRound}(H_{t-1}, A_{\text{keep}}^{(t)})$
20: **end for**
21: **return** Answer from H_T

5 Overall Performance and Cost Comparison

Experimental Setup. To validate the effectiveness of our **NP-CredMS** framework, we conduct a comprehensive evaluation against established single-agent and multi-agent baselines. We measure performance across a diverse suite of eight leading VLM benchmarks: **MMBench** (Liu et al., 2024b), **MMMU_val** (Yue et al., 2024), **OCR-Bench** (Liu et al., 2024c), **V*-Bench** (Wu and Xie, 2023), **ScienceQA** (Lu et al., 2022), **Real-WorldQA** (xAI Corp., 2024), **MathVista_mini** (Lu et al., 2024), and **CountBench** (Paiss et al., 2023). Our baselines include standard single-agent methods (CoT (Wei et al., 2023), ComplexCoT (Fu et al., 2023), SC (Wang et al., 2023)) and prominent Multi-Agent Systems (MAS), such as full-panel methods (PHP(Zheng et al., 2024), LLM-Debate (Du et al., 2023)) and SOTA pruning techniques (AgentPrune-R) (Zhang et al., 2025a) and DyLAN

(Liu et al., 2024d). All experiments are conducted using **Qwen-2.5VL (72B)** as the foundational large vision-language model for all agents, ensuring a fair comparison. We report accuracy (%) for performance, and total token consumption (summed across all agents and rounds) as the primary metric for cost. Please note that for all experiments, we set the maximum number of debate rounds to 5; however, experiments (see Sec 6) reveal that the system naturally converges to an economic equilibrium by round 3, where no further agents find it profitable to bid against the cost threshold. We used the llava-ov (Li et al., 2024) training set for training, and for all experiments, all MAS systems employed the same "5 agents per round" setup: an OCR agent, a visual understanding agent, a mathematical agent, a logical reasoning agent and abstract semantic agent. Please note that this experiment only counts the token cost during the inference process, excluding the training and warmup phases.

Results and Analysis. The main results, based on the **Qwen-2.5VL (72B)** backbone, are presented in Table 1 (Performance) and Table 2 (Cost-Performance). As shown in Table 1, **NP-CredMS** achieves new state-of-the-art (SOTA) performance across all eight challenging VLM benchmarks. This demonstrates its robust ability to enhance multi-modal reasoning. Critically, on complex, multi-step tasks that are known bottlenecks for VLMs, our method shows the largest gains. For instance, on **V*-Bench**, our method achieves a top score of 80.90%, surpassing the next-best pruning method (AgentPrune-R) by 0.85% and the strong Vanilla baseline by 3.9%. Similarly, on **MMMU_val**, our method attains 54.81%, a significant +3.51% gain over Vanilla, and a +0.81% gain over AgentPrune-R.

Table 2 illustrates the core design intent (cost reduction and efficiency gain) of our framework. On general-purpose benchmarks like **MMBench**, NP-CredMS not only achieves SOTA accuracy (85.4%) but also reduces total token consumption by 24% compared to AgentPrune-R and by over 50% compared to full-panel methods like LLM-Debate. This trend holds on the highly complex **V*-Bench**, where our method delivers the best performance (80.9%) while consuming 22.5% fewer tokens than AgentPrune-R. On **MMMU_val**, our method again achieves SOTA accuracy while reducing costs by 28% relative to AgentPrune-R. These results confirm that NP-CredMS, through its dynamic pruning based on causal value predictions, effectively executes dynamic resource allocation, simultaneously pushing performance to SOTA levels while autonomously curbing redundancy.

6 Dynamic Convergence and Qualitative Analysis

Experimental Setup. To validate the core hypothesis of our framework—that dynamic, per-round pruning at inference time is both fast and effective—we conducted a specialized analysis on the challenging **MMMU_val** benchmark. We initialized a full panel of $N = 5$ heterogeneous Qwen-2.5VL agents (e.g., with different system prompts or temperature settings). We tracked two key metrics across the task's duration ($T = 5$ rounds):

1. **Panel Size:** The number of active agents, $|A_{\text{keep}}^{(t)}|$.
2. **Panel Stability:** The Jaccard similarity $J(A_{\text{keep}}^{(t)}, A_{\text{keep}}^{(t-1)})$ between the active sets of consecutive rounds, measuring convergence.

This setup is designed to empirically test our claim of rapid convergence to an approximate Nash Equilibrium.

6.1 Dynamic Convergence Analysis

Analysis of Convergence. The results, averaged across all MMMU_val tasks, are presented in Figure 2. The findings strongly support our hypothesis of rapid convergence and aggressive, intelligent cost-saving. We observe two key trends:

1. **Economic Convergence (Cost Reduction):** The mean number of active agents (Avg. Panel Size) shows a dramatic contraction, stabilizing at **2.1 agents by** $t = 3$. This pruning has a profound effect on cost, as seen in the 'Cumulative Cost'. The cost growth decelerates sharply: the marginal cost (cost added per round) drops from 3.0 units in $t = 1$, to 1.3 ($4.3 - 3.0$) in $t = 2$, and just 0.6 ($4.9 - 4.3$) in $t = 3$. Critically, the marginal cost approaches zero in $t = 4$ (0.2) and $t = 5$ (0.1). This provides strong evidence for our claim: once the system converges to a stable panel and solution, **it actively ceases further communication not because of a hard stop**, but because the marginal utility of any additional message falls below the token cost, achieving a stable economic equilibrium.

Table 1: Overall performance comparison across eight VLM benchmarks. We report accuracy (%). The best score in each column is in **bold**. Values in parentheses show absolute gain over the Vanilla (single-agent Qwen-2.5VL) baseline.

Method	MMBench	MMMU_val	OCRBench	V*-Bench	ScienceQA	RealWorldQA	MathVista	CountBench
Single-Agent Methods (Qwen-2.5VL Backbone)								
Vanilla	83.40	51.30	84.20	77.00	88.80	68.50	68.60	86.40
CoT	83.91 (+0.51)	51.92 (+0.62)	84.81 (+0.61)	77.53 (+0.53)	89.11 (+0.31)	68.90 (+0.40)	69.11 (+0.51)	86.92 (+0.52)
ComplexCoT	84.20 (+0.80)	52.51 (+1.21)	85.13 (+0.93)	78.10 (+1.10)	89.40 (+0.60)	69.12 (+0.62)	69.53 (+0.93)	87.21 (+0.81)
SC	84.05 (+0.65)	52.34 (+1.04)	85.00 (+0.80)	77.92 (+0.92)	89.31 (+0.51)	69.00 (+0.50)	69.30 (+0.70)	87.15 (+0.75)
Multi-Agent Methods (Qwen-2.5VL Backbone)								
PHP	84.51 (+1.11)	53.05 (+1.75)	85.50 (+1.30)	79.03 (+2.03)	89.80 (+1.00)	69.51 (+1.01)	70.12 (+1.52)	87.80 (+1.40)
LLM-Debate	84.72 (+1.32)	53.30 (+2.00)	85.73 (+1.53)	79.31 (+2.31)	90.15 (+1.35)	69.70 (+1.20)	70.44 (+1.84)	88.01 (+1.61)
DyLAN	84.30 (+0.90)	52.81 (+1.51)	85.30 (+1.10)	78.88 (+1.88)	89.90 (+1.10)	69.40 (+0.90)	70.01 (+1.41)	87.73 (+1.33)
AgentPrune-R	85.01 (+1.61)	54.00 (+2.70)	86.10 (+1.90)	80.05 (+3.05)	90.50 (+1.70)	70.10 (+1.60)	71.00 (+2.40)	88.52 (+2.12)
NP-CredMS (Ours)	**85.42 (+2.02)**	**54.81 (+3.51)**	**86.53 (+2.33)**	**80.90 (+3.90)**	**91.20 (+2.40)**	**70.80 (+2.30)**	**71.71 (+3.11)**	**89.05 (+2.65)**

Table 2: Consolidated cost-performance comparison on MMBench, MMMU_val, and V*-Bench. Our **NP-CredMS**, highlighted in gray, achieves SOTA performance while maintaining superior token efficiency.

	MMBench		MMMU_val		V*-Bench	
Method	**Acc. (%)**	**Token Cons.**	**Acc. (%)**	**Token Cons.**	**Acc. (%)**	**Token Cons.**
NP-CredMS (Ours)	**85.4**	$\mathbf{7.50 \times 10^5}$	**54.8**	$\mathbf{9.0 \times 10^5}$	**80.9**	$\mathbf{12.0 \times 10^6}$
AgentPrune-R	85.0	9.80×10^5	54.0	12.5×10^5	80.1	15.5×10^6
LLM-Debate	84.7	15.0×10^5	53.3	20.0×10^5	79.3	24.0×10^6
PHP	84.5	16.0×10^5	53.1	22.0×10^5	79.0	25.0×10^6
DyLAN	84.3	11.0×10^5	52.8	14.0×10^5	78.9	18.0×10^6
Vanilla (Qwen-2.5VL)	83.4	2.00×10^5	51.3	3.0×10^5	77.0	4.0×10^6

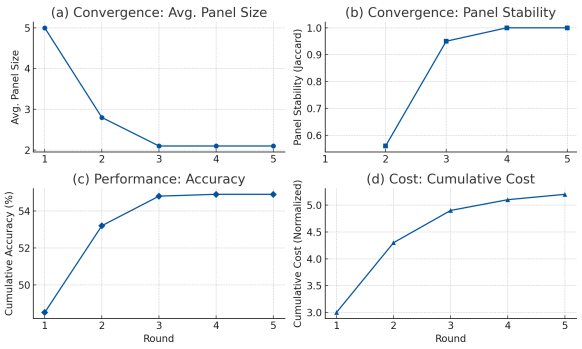

Figure 2: Dynamic panel convergence, performance, and cost on MMMU_val (Averaged, $N = 5$).

2. **Performance Convergence (Efficiency Gain):** Concurrently, the cumulative accuracy (the solution's quality at the end of each round) rises significantly from 48.5% at $t = 1$ to 53.2% at $t = 2$ as noisy and hallucinating agents are pruned. The accuracy converges to our SOTA performance of **54.8% by** $t = 3$, demonstrating that this cost-saving measure actively *improves* the final answer quality.

Finally, the Panel Stability (Jaccard) confirms this, reaching ≈ 1.0 by $t = 4$. This demonstrates that the system does not oscillate; it decisively identifies and converges upon a minimal, stable, and high-performance subset of contributors within just 3 rounds.

Table 3: Ablation study on MMBench and V*-Bench. Our full model (highlighted) demonstrates the best cost-performance balance, validating the necessity of all three core components.

	MMBench		V*-Bench	
Method Variant	**Acc. (%)**	**Token Cons.**	**Acc. (%)**	**Token Cons.**
NP-CredMS (Full Model)	**85.4**	$\mathbf{7.5 \times 10^5}$	**80.9**	$\mathbf{12.0 \times 10^6}$
w/o AIPW (Frozen Estimator)	83.1 (-2.3)	8.1×10^5	78.5 (-2.4)	14.5×10^6
Static Pruning (Non-Dynamic)	84.7 (-0.7)	7.9×10^5	80.1 (-0.8)	12.8×10^6
w/o Economic Metrics (Random-K)	82.5 (-2.9)	7.5×10^5	77.8 (-3.1)	12.0×10^6

7 Ablation Study on Core Components

Experimental Setup. To rigorously validate the necessity of our framework's key components, we conducted an extensive ablation study. We compare our full **NP-CredMS** model against three ablated variants on the MMBench (general) and V*-Bench (complex reasoning) benchmarks. The variants are:

1. **w/o AIPW (Frozen Estimator):** Replaces the doubly-robust AIPW credit estimator with a simpler, but biased, "frozen" path estimator (d^{frozen}), which does not use regeneration to correct for estimation errors.
2. **Static Pruning (Non-Dynamic):** Replaces our round-by-round auction with a one-time pruning decision at $t = T_0$. The panel remains fixed for all subsequent rounds, ignoring context shifts.
3. **w/o Economic Metrics (Random-K):** Replaces our entire economic selection logic. It randomly prunes $N - K$ agents, where K is set to the average panel size observed by our full model ($K \approx 2.1$).

All variants use the same Qwen-2.5VL backbone and $N = 5$ initial agents. We report both accuracy (%) and total token consumption to evaluate the cost-performance trade-off.

Results and Analysis. The results in Table 3 confirm that each component of NP-CredMS is critical for achieving the optimal balance of high performance and low cost.

The 'w/o AIPW' variant, which relies on a simpler, biased credit estimator, suffers a major performance drop (-2.3% on MMBench, -2.4% on V*-Bench) and also *increases* token consumption. This is because the biased "frozen" estimator fails to accurately distinguish valuable contributions from harmful ones, leading to suboptimal pruning decisions: it incorrectly prunes useful agents (hurting accuracy) and retains noisy agents (wasting tokens). **The 'Static Pruning' variant**, which fixes the panel after round T_0 without re-evaluating bids, shows a non-trivial performance degradation (-0.7% to -0.8%). This confirms the necessity of our **round-by-round dynamic auction**, as agent utility is highly context-dependent and fluctuates as the conversation evolves. **The 'w/o Economic Metrics (Random-K)'** variant shows the most significant performance collapse (-2.9% to -3.1%), even though its cost is identical to our full model by design. This clearly proves that *how* pruning is done matters more than just pruning itself. A naive heuristic like random selection is unable to differentiate between a critical expert and a harmful hallucinator.

8 Economic Controllability and Pareto Frontier

Experimental Setup. A core advantage of our framework is its controllability. Users should be able to "dial" their desired balance between performance and cost, rather than being forced into a single operating point. To demonstrate this, we analyze the impact of the key hyperparameter from our Nash-style selection: the panel size constraint K. We conduct a sweep on the **MMMU_val** benchmark, varying K from $K = 1$ (most aggressive pruning, selecting only the top agent) to $K = 5$ (no pruning, equivalent to a full panel, $N = 5$). We report the final cumulative accuracy (%) and the total token consumption (Token Cons.) for each configuration.

Results and Analysis. The results in Table 4 clearly illustrate the Pareto frontier of our controllable framework. As we increase K from 1 to 3, performance rises significantly from 52.0% to our SOTA-level result of **54.8%**, while cost increases moderately. This demonstrates the benefit of collaboration among a small, curated panel of high-value agents.

Crucially, increasing K beyond 3 yields sharply diminishing returns. From $K = 3$ to $K = 5$ (the full panel), the accuracy nearly saturates, increasing only marginally from 54.8% to 55.0%. However, this tiny 0.2% performance gain comes at a disproportionate cost: the token consumption more than *doubles*, increasing from 2.1x to 4.5x. The full panel is clearly Pareto-inefficient.

Table 4: Cost-performance trade-off on MMMU_val by varying the panel size constraint K (Full panel $N = 5$). The $K = 3$ setting (our default) represents the optimal "sweet spot" on the Pareto frontier, achieving near-peak performance before diminishing returns set in.

Panel Size (K)	**Accuracy (%)**	**Token Cost (Normalized)**
1	52.0	1.0x
2	54.1	1.5x
3 (Default)	**54.8**	**2.1x**
4	54.9	2.7x
5 (Full Panel)	55.0	4.5x

This analysis confirms two key findings: (1) NP-CredMS is a fully controllable system, allowing users to select their desired point on the cost-performance curve (e.g., $K = 2$ for high efficiency, $K = 3$ for the best performance/cost balance). (2) The optimal "sweet spot" ($K = 3$) achieves near-peak performance at a fraction of the cost of a full panel, validating that our economic pruning is essential for optimizing the cost-performance trade-off.

9 Conclusion

In this work, we addressed the critical challenge of prohibitive $O(N \times T)$ communication costs in multi-round VLM Multi-Agent Systems, a problem stemming from static "all-speak" panels that also suffer from performance degradation by low-value agents. We proposed **Nash-Pruned CredMAS (NP-CredMS)**, an economic framework that transforms these static systems into dynamic, self-pruning ones at inference time. Empirically, NP-CredMS achieves SOTA results on eight VLM benchmarks, including MMMU_val and V*-Bench, while significantly reducing token consumption.

10 Limitations

First, our causal value learning relies on the coverage of offline logs, where significant distribution shifts could degrade the accuracy of the doubly-robust estimator. Second, the greedy auction mechanism assumes strictly submodular agent utility, potentially failing to capture complex high-order dependencies where combined agents offer non-

diminishing returns. Finally, while effective on VLM benchmarks, our framework's generalizability to purely text-based agents or tool-use scenarios requires further empirical verification.

11 Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62276283, in part by the China Meteorological Administration's Science and Technology Project under Grant CMAJBGS202517, in part by Guangdong-Hong Kong-Macao Greater Bay Area Meteorological Technology Collaborative Research Project under Grant GHMA2024Z04, in part by Fundamental Research Funds for the Central Universities, Sun Yat-sen University under Grant 23hytd006 and 23hytd006-2, in part by Guangdong Provincial High-Level Young Talent Program under Grant RL2024-151-2-11, and in part by the Key Development Project of the Artificial Intelligence Institute, Sun Yat-sen University under Grant 2025RGZN009.

References

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, and 3 others. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *Preprint*, arXiv:2509.23661.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Shuaihang Chen, Yuanxing Liu, Wei Han, Weinan Zhang, and Ting Liu. 2025. A survey on llm-based multi-agent system: Recent advances and new frontiers in application. *Preprint*, arXiv:2412.17481.

Abhishek Das, Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. 2017. Learning cooperative visual dialog agents with deep reinforcement learning. *Preprint*, arXiv:1703.06585.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *Preprint*, arXiv:2305.14325.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. *Preprint*, arXiv:2210.00720.

Neil De La Fuente, Miquel Noguer i Alonso, and Guim Casadellà. 2024. Game theory and multi-agent reinforcement learning : From nash equilibria to evolutionary dynamics. *Preprint*, arXiv:2412.20523.

Gemini Team. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Dylan Goetting, Himanshu Gaurav Singh, and Antonio Loquercio. 2024. End-to-end navigation with vision language models: Transforming spatial reasoning into question-answering. *Preprint*, arXiv:2411.05755.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *Preprint*, arXiv:2402.01680.

Mojtaba Hosseini and Vijay V. Vazirani. 2023. Nash-bargaining-based models for matching markets: One-sided and two-sided; fisher and arrow-debreu. *Preprint*, arXiv:2105.10704.

Yupeng Hu, Changxing Ding, Chang Sun, Shaoli Huang, and Xiangmin Xu. 2025. Bilateral collaboration with large vision-language models for open vocabulary human-object interaction detection. *Preprint*, arXiv:2507.06510.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. *Preprint*, arXiv:2402.06782.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024b. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024c. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12).

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2024d. A dynamic llm-powered agent network for task-oriented agent collaboration. *Preprint*, arXiv:2310.02170.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *Preprint*, arXiv:2310.02255.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Preprint*, arXiv:2209.09513.

J.F. Nash. 1951. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. 2022. Multi-task learning as a bargaining game. *Preprint*, arXiv:2202.01017.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. 2023. Teaching clip to count to ten. *Preprint*, arXiv:2302.12066.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Stefan Schuster. 2017. A new solution concept for the ultimatum game leading to the golden ratio. *Scientific Reports*, 7(1):5642.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D. Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *Preprint*, arXiv:2501.06322.

Jin Wang, Shichao Dong, Yapeng Zhu, Kelu Yao, Weidong Zhao, Chao Li, and Ping Luo. 2024. Diagnosing the compositional knowledge of vision language models from a game-theoretic view. *Preprint*, arXiv:2405.17201.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Penghao Wu and Saining Xie. 2023. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *Preprint*, arXiv:2308.08155.

xAI Corp. 2024. Grok-1.5 vision preview: Connecting the digital and physical worlds with our first multimodal model. `https://x.ai/news/grok-1.5v`.

Qintong Xie, Edward Koh, Xavier Cadet, and Peter Chin. 2025a. Nash q-network for multi-agent cybersecurity simulation. *Preprint*, arXiv:2509.00678.

Yi Xie, Zhanke Zhou, Chentao Cao, Qiyu Niu, Tongliang Liu, and Bo Han. 2025b. From debate to equilibrium: Belief-driven multi-agent LLM reasoning via bayesian nash equilibrium. In *Forty-second International Conference on Machine Learning*.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, Shrikant Kendre, Jieyu Zhang, Shaoyen Tseng, Gustavo A Lujan-Moreno, Matthew L Olson, Musashi Hinck, David Cobbley, Vasudev Lal, Can Qin, and 14 others. 2025. xgen-mm (blip-3): A family of open large multimodal models. *Preprint*, arXiv:2408.08872.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Yanwei Yue, Guibin Zhang, Boyang Liu, Guancheng Wan, Kun Wang, Dawei Cheng, and Yiyan Qi. 2025. Masrouter: Learning to route llms for multi-agent systems. *Preprint*, arXiv:2502.11133.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. 2025a. Cut the crap: An economical communication pipeline for LLM-based multi-agent systems. In *The Thirteenth International Conference on Learning Representations*.

Jusheng Zhang, Yijia Fan, Wenjun Lin, Ruiqi Chen, Haoyi Jiang, Wenhao Chai, Jian Wang, and Keze Wang. 2025b. Gam-agent: Game-theoretic and

uncertainty-aware collaboration for complex visual reasoning. *Preprint*, arXiv:2505.23399.

Jusheng Zhang, Zimeng Huang, Yijia Fan, Ningyuan Liu, Mingyan Li, Zhuojie Yang, Jiawei Yao, Jian Wang, and Keze Wang. 2025c. Kabb: Knowledge-aware bayesian bandits for dynamic expert coordination in multi-agent systems. *Preprint*, arXiv:2502.07350.

Xinxin Zhao, Wenzhe Cai, Likun Tang, and Teng Wang. 2024. Imaginenav: Prompting vision-language models as embodied navigator through scene imagination. *Preprint*, arXiv:2410.09874.

Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. 2024. Progressive-hint prompting improves reasoning in large language models. *Preprint*, arXiv:2304.09797.

Yujia Zheng, Zhuokai Zhao, Zijian Li, Yaqi Xie, Mingze Gao, Lizhu Zhang, and Kun Zhang. 2025. Thought communication in multiagent collaboration. *Preprint*, arXiv:2510.20733.