SCALING VIDEO UNDERSTANDING: COMPACT LATENT COLLABORATION FOR BUDGET-CONSTRAINED DISTRIBUTED AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-modal large language models (MLLMs) advance vision–language understanding but face inherent limitations in long-video tasks due to bounded perception context budgets. Existing agentic methods mitigate this via rule-based preprocessing, yet often suffer from information loss, high cost, and reliance on textual intermediates. We propose MACF-Video, an end-to-end Multi-Agent Collaboration Framework that decouples per-agent perception budgets from global video complexity, enabling scalable video understanding while preserving visual fidelity. MACF-Video partitions videos into segments for locally budgeted agents and enables holistic reasoning via an agent-native latent communication protocol. Each agent encodes partial observations into compact, task-sufficient tokens in a shared embedding space, allowing efficient and information-preserving collaboration by a central coordinator. We introduce a curriculum training strategy that progressively enforces semantic alignment, evidence summarization, and cross-agent coordination. Extensive experiments on diverse video understanding benchmarks show that MACF-Video consistently outperforms state-of-the-art MLLMs and multi-agent systems under identical budget constraints, demonstrating the effectiveness of our latent collaboration for scalable video understanding.

1 INTRODUCTION

Multi-modal large language models (MLLMs) Alayrac et al. (2022); Li et al. (2023b); Hurst et al. (2024); Liu et al. (2023); Lin et al. (2026) have demonstrated remarkable capability in vision-language understanding Yin et al. (2024); Caffagni et al. (2024); Zhang et al. (2024a); Tang et al. (2025); Cao et al. (2025); Li et al. (2024); Yin et al. (2025). However, a single MLLM inherently remains constrained by limited input context budgets, hindering scalability to long-temporal or high-resolution perceptual inputs. This limitation is especially evident in video understanding, where raw visual streams are highly redundant, less information-dense than language, and weakly aligned with tasks, causing perceptual inputs to dominate the budget while contributing inefficiently to reasoning.

To address this limitation, agentic methods Zhou et al. (2025b); Yu et al. (2024); Luo et al. (2024); Long et al. (2025); Xu et al. (2026) augment MLLMs with auxiliary functions such as preprocessing, retrieval, and memory. A common preprocessing approach is perceptual sampling Han et al. (2024); Liang et al. (2024), which temporally and spatially downsamples videos to meet the input budget. Such approaches inevitably sacrifice temporal continuity and spatial fidelity, leading to substantial information loss, as illustrated in Fig. 1 (a). More advanced systems Fan et al. (2024); Wang et al. (2024) rely on retrieval to select a subset of "informative" frames, typically guided by caption-based matching, as illustrated in Fig. 1 (b). However, they incur high cost, depend heavily on intermediate text quality, and risking loss of visual details. Overall, these pipelines remain rule-driven and text-centric, limiting end-to-end alignment with native visual representations.

A natural solution is to distribute the perceptual workload across multiple agents, each under a local budget constraint, and to aggregate their partial knowledge through collaboration, as shown in Fig 1 (c). Related paradigms, such as coarse-to-fine Wang et al. (2024) or MapReduce Pang & Wang (2025) schemes, follow this intuition but still rely on expensive retrieval pipelines and textual communication across agents. Consequently, this distributed setup shift challenges to the inter-agent

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

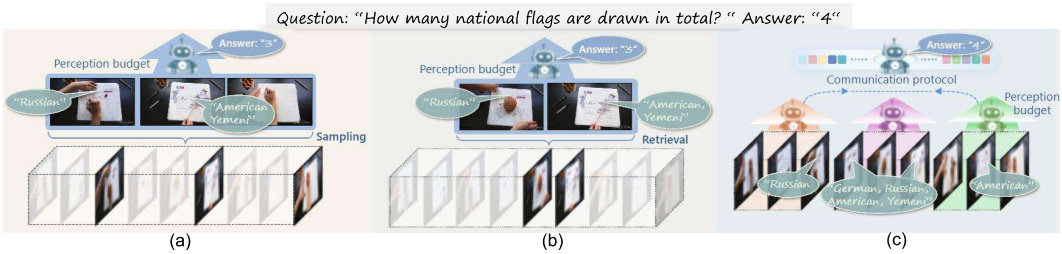


Figure 1: **Scaling video understanding with budgeted agents.** (a) Perceptual sampling meets the budget but discards temporal and spatial details (e.g., missing the "German flag"). (b) Retrieval selects key frames based on captions, incurring high cost and potential loss of visual fidelity. (c) MACF-Video partitions the video across local agents and communicates compact latent tokens in a shared space for coordination. (Text hints in green bubbles are for illustration only.)

communication protocol. It must address: (1) *Information fragmentation*, requiring aggregation across agents due to partial observations; (2) *Communication bandwidth constraints*, requiring exchanged messages to be compact; and (3) *Task-relevant information preservation*, ensuring minimal semantic loss during communication. Although text-based communication is commonly adopted Zuo et al. (2025); Kugo et al. (2025); Liu et al. (2025a), it struggles to convey fine-grained visual cues (e.g., motion, spatial relations, and textures) critical for video understanding. These limitations motivate the need for agent-native communication mechanisms that operate directly in the visual-semantic representation space Sukhbaatar et al. (2016); Foerster et al. (2016).

In this work, we propose an end-to-end Multi-Agent Collaboration Framework (MACF-Video) that scales video understanding by decoupling per-agent perception budgets from global video complexity. MACF-Video partitions the input video into multiple segments, each processed by a local agent within its perception budget. Our key innovation lies in a latent communication protocol that enables efficient and expressive cross-agent information exchange. Specifically, each local agent first maps its observations into compact, task-relevant communication tokens embedded in a shared continuous space. These agent-native tokens form a unified communication channel that preserves visual semantics often lost in textual descriptions. A central coordinator agent ingests tokens from all local agents under a communication bandwidth constraint and performs global understanding. As all messages reside in the same embedding space, cross-agent communication is both efficient and semantically aligned. To enable effective optimization, we introduce a three-stage curriculum training strategy that progressively establishes semantic alignment, query-aware evidence summarization, and cross-agent collaboration within the shared communication space. Extensive experiments on standard video understanding benchmarks demonstrate that MACF-Video consistently outperforms state-of-the-art MLLMs under identical budget constraints, validating the effectiveness of agent-native latent communication for scalable video understanding.

In summary, our contribution lies in three-fold:

- We formalize a budget-constrained multi-agent video understanding problem and propose MACF-Video, an end-to-end distributed framework that scales video understanding by decomposing global perception into locally budgeted agents with coordinated inference.
- We introduce an agent-native latent communication mechanism that transforms partial visual observations into compact, task-sufficient representations in a shared embedding space, enabling bandwidth-efficient and information-preserving aggregation across agents.
- We design a curriculum learning strategy that progressively enforces semantic alignment and coordination in the latent communication space, stabilizing optimization and enabling reliable global inference from fragmented, budget-limited evidence.

2 Related Work

Multimodal large language models for Understanding. Recent multimodal large language models (MLLMs) align pretrained LLMs with visual encoders to support instruction-following, grounding, and open-ended understanding over images Alayrac et al. (2022); Li et al. (2023b); Liu et al. (2023); Achiam et al. (2023). Building on this paradigm, video-oriented MLLMs extend the visual

front-end with temporal modeling and frame sampling to answer compositional queries that require motion cues and long-range temporal dependencies Li et al. (2025b); Zhang et al. (2023); Chen et al. (2023); Ren et al. (2024); Li et al. (2025a); He et al. (2024); Liu et al. (2025b). However, practical deployments face strict input-bandwidth or context limits, where aggressively downsampling frames or shortening clips can discard fine-grained evidence and break temporal reasoning in long videos Wu et al. (2024a); Wang et al. (2025). Our work targets this bottleneck by distributing perception across multiple segment-level VLM agents and aggregating their evidence by a coordinator.

Multi-agent communication. Most existing LLM-based agent frameworks communicate via free-form natural language Wu et al. (2024b); Li et al. (2023a); Qian et al. (2024); Hong et al. (2023), which has proven effective for a wide spectrum of language tasks such as planning, coding, multi-step reasoning, and tool orchestration. These agentic paradigms have also been extended to multi-modal tasks Fan et al. (2024); Wu et al. (2023); Shen et al. (2023); Yang et al. (2023); Pang & Wang (2025); Zhang et al. (2025b); Wang et al. (2026) by pairing an LLM coordinator or planner with specialized perception VLMs, enabling visual reasoning and content understanding in images and videos. However, when tasks require fine-grained visual evidence and temporal relations, discrete language communication becomes a bottleneck because not all visual evidence can be expressed precisely Su et al. (2025), and summarization often drop critical spatial, grounding and temporal cues. Beyond text messaging, emerging efforts in language domain Zou et al. (2025a); Fu et al. (2025b); Yu et al. (2026) explore latent or state-based coordination (e.g., transferring internal states such as KV-cache) to increase communication bandwidth and efficiency. In this paper, we study distributed multimodal evidence communication, where agents write visual evidence into a shared communication space and a coordinator integrates these distributed information for global answering. Unlike MapReduce Pang & Wang (2025), which discretizes evidence into lossy text, and LatentMAS Zou et al. (2025a) that forwards raw query-agnostic latent states in the text modality, MACF-Video couples budgeted distributed perception, query-conditioned multimodal latent communication, and curriculum optimization of the shared channel.

Context compression. Beyond parameter compression Frantar et al. (2022); Mozaffari et al. (2024), recent studies explore compressing long contexts into compact latent tokens that can be directly conditioned on by an LLM, alleviating context-length and inference-memory bottlenecks Ge et al. (2023); He et al. (2025). Such memory-slot based compression provides a model-native interface for retaining evidence under a fixed token budget, avoiding information loss caused by the natural language. Different from prior context compression which targets single model language long-context usage, we focus on *multi-agent* settings and compress each agent's *multimodal observations* into fixed-length communication tokens for communicating and downstream evidence aggregation. This design explicitly treats the latent tokens as a communication channel, enabling the coordinator to compose distributed visual evidence across temporal segments.

3 METHOD

3.1 PROBLEM FORMULATION

To formalize scalable video understanding in real-world scenarios, we consider the problem formulation in which each agent operates under explicit budget constraints on both perception and communication. We further discuss the necessity of this setup in App.I.

Let $V = \{I_i\}_{i=1}^{T}$ denote an input video consisting of T frames, where each frame I_i with original resolution $H \times W$. Given a textual query q, the objective is to predict an answer y, which may be a categorical label, a free-form textual response, or a structured output, depending on the task.

Perception budget. In practical deployments, each agent can process only a bounded amount of visual input per forward pass. We formalize this bound as a perception budget B^{per}, defined as the maximum number of pixels that can be ingested in a single inference. This pixel-based budget offers a simple and unified proxy for visual input capacity across different systems, models, or codec algorithms. A single agent is responsible for processing the video V under the perception budget:

$$\mathrm{pix}(X) \le B^{per}, \quad X = \mathcal{P}(V), \tag{1}$$

where $\mathrm{pix}(\cdot)$ counts the number of pixels in the processed input, $\mathcal{P}(\cdot)$ denotes a preprocessing function that may include temporal and spatial subsampling, including a trade-off between temporal coverage and spatial fidelity.

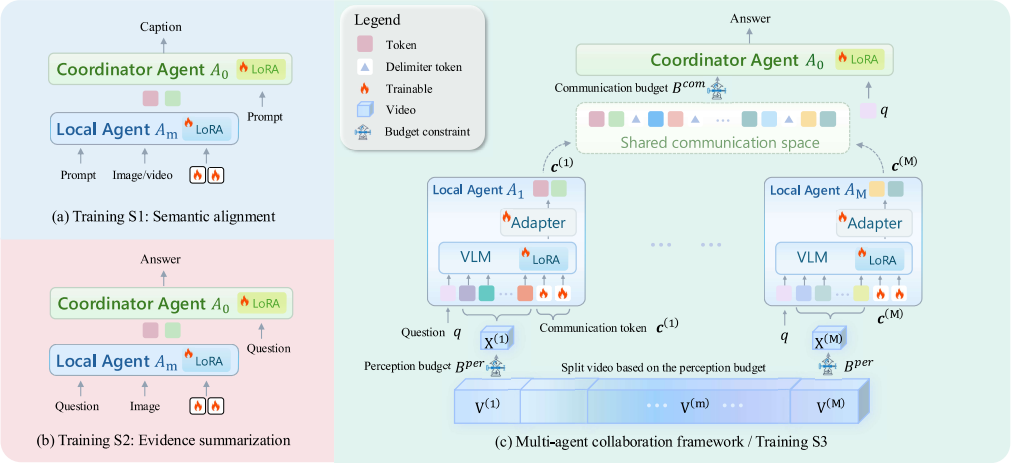


Figure 2: **Overview of MACF-Video.** MACF-Video distributes the video perception workload temporally across multiple local agents, each operating under a per-agent perception budget. Each agent compresses its observation into communication tokens extracted from the last-layer hidden states and projects them into a shared latent communication space via an adaptor module. A coordinator agent aggregates tokens from all agents to produce the final prediction. (a) and (b) illustrate the curriculum training, which includes semantic alignment and evidence summarization, followed by end-to-end coordination training shown in (c).

To overcome this perception bottleneck, we can distribute the perceptual workload across multiple agents. Given a set of M agents $\mathcal{A} = \{A_1, ..., A_M\}$, each A_m perceives a video segment $V^{(m)}$ via a local transformation $X^{(m)} = \mathcal{P}_m(V^{(m)})$, subject to a per-agent pixel budget, $\text{pix}(X^{(m)}) \leq B^{per}$. This decouples the global video length T from the per-agent input budget B^{per}, allowing the system to scale to substantially longer videos while preserving fine-grained visual information within each local segment.

Communication budget. Since no single agent observes the full video, agents must communicate local information for global coordination. This process is constrained by the finite context capacity of the coordinator A_0, which we model as a communication budget B^{com} representing the maximum information it can ingest per forward pass. Let $\mathbf{c}^{(m)}$ denote the communication message produced by agent A_m. The coordinator A_0 consumes a subset of the messages $\mathcal{C} \subseteq \{\mathbf{c}^{(1)}, ..., \mathbf{c}^{(M)}\}$ from all agents. The constraint is:

$$\sum_{\mathbf{c}^{(m)} \in \mathcal{C}} \left|\mathbf{c}^{(m)}\right| + |q| \;\leq\; B^{com}. \tag{2}$$

3.2 Multi-Agent Collaboration Framework (MACF-Video)

We introduce the Multi-Agent Collaboration Framework (MACF-Video) for scalable video understanding with compact latent collaboration. As illustrated in Fig. 2 (a), MACF-Video consists of three components: 1) distributed local agents with budgeted perception; 2) compact latent communication tokens for information exchange; 3) a coordinator agent fusing distributed evidence for global understanding. The entire framework is fully differentiable and trained end-to-end.

Local agents with budgeted perception Without loss of generality, we assume all local agents share the same architecture and operate under an identical perception budget. Given M available agents $\{A\}_{m=1}^{M}$, the input video is first partitioned into disjoint temporal segments $V = \{V^{(1)}, V^{(2)}, \ldots, V^{(M)}\}$, where each segment $V^{(m)}$ is assigned to agent A_m, for $m = 1, \ldots, M$. Focusing on multi-agent collaboration rather than sampling or compression strategies, we adopt a simple and consistent preprocessing scheme: each agent uniformly samples F frames from its segment and resizes them to a shared resolution (h, w). The resulting input satisfies the per-agent budget:

$$F \times h \times w \leq B^{per}. \tag{3}$$

Scaling the number of agents effectively redistributes the perception budget across time, allowing finer-grained spatiotemporal representation within each local segment under a fixed per-agent constraint.

Latent collaboration with communication tokens. Given the uniform video partitioning, each agent's local observation can contribute to the global task. We adopt a star-topology in which all local agents communicate directly with a central coordinator. Unlike prior multi-agent systems that rely on natural language for inter-agent communication Pang & Wang (2025), we avoid text-based messaging, which is often redundant and lossy for video understanding.

We propose to communicate using latent communication tokens. A local agent summarizes its visual evidence into compact continuous vectors encoding query-relevant evidence softly. These *communication tokens* are information-dense and directly consumable by the coordinator, enabling efficient and faithful collaboration in a shared representation space. Formally, each agent A_m encodes its local observation $X^{(m)}$ and query q into a fixed-capacity communication token sequence:

$$\mathbf{c}^{(m)} = A_m(X^{(m)}, q) \in \mathbb{R}^{K \times d}, \tag{4}$$

where K is the number of communication tokens per agent and d is the embedding dimension.

In the star-topology, the coordinator A_0 aggregates tokens from all agents to produce the final prediction:

$$\hat{y} = A_0(q, \mathbf{c}^{(1)}, ..., \mathbf{c}^{(M)}). \tag{5}$$

Communication is constrained by the coordinator's finite context capacity, modeled as a communication budget measured in tokens:

$$M \times K + \mathrm{token}(q) \;\leq\; B^{com}, \tag{6}$$

where $\mathrm{token}(\cdot)$ denotes number of tokens.

3.3 Optimization of MACF-Video

Since all local agents $A_{m=1}^{M}$ share the same input and output formats and model architectures, we tie their parameters and learn a single set of shared weights during training. This design reduces model complexity and enforces consistent behavior across agents processing different video segments. Notably, our framework can be extended to heterogeneous models, as each agent is equipped with an adaptor module that maps its communication tokens into a shared latent space consumable by the coordinator A_0.

Under the perception and communication constraints, the objective is to jointly learn the shared local-agent model A_m and the coordinator A_0 by minimizing the expected task loss:

$$\min_{\theta_A, \theta_0} \mathbb{E}_{(V,q,s)\sim\mathcal{D}}\big[\ell(\hat{y}, y)\big], \tag{7}$$

where θ_A denotes the parameters shared by all local agents and θ_0 the parameters of the cooridnator A_0, $\mathcal{D}$ the data distribution of triplets (V, q, y), $\ell(\cdot)$ measuring the discrepancy and often instantiated as autoregressive cross-entropy (CE).

Although MACF-Video is fully differentiable, direct end-to-end training is challenging due to a two-stage bottleneck that induces an effective min-cut in the information flow and the asymmetric roles of local agents and the coordinator. Local agents are responsible for mapping high-dimensional video inputs into concise communication tokens in a shared latent space, while the coordinator must aggregate these tokens and perform global coordination across agents. Training both components simultaneously from scratch often leads to unstable representations and ineffective collaboration. To address these challenges, we adopt a curriculum training strategy that progressively equips the model with the required capabilities, with three stages: : (1) semantic alignment to anchor communication tokens in a shared latent space; (2) evidence summarization to encode query-relevant information into compact tokens; and (3) cross-agent collaboration to enable effective global aggregation.

Semantic alignment. Before effective collaboration can emerge, communication tokens produced by local agents must be anchored in a unified semantic space that is consumable by the coordinator. Without such alignment, the communication tokens can drift across agents and lack semantic consistency. Accordingly, in the first training stage illustrated in Fig 2 (b), we initialize the *communication*

space with caption supervision, which provides dense, low-variance language targets and stabilizes the grounding of visual evidence. Given the preprocessed visual input X, a local agent A_m map it into communication tokens $\mathbf{c} = A_m(X)$, and the coordinator A_0 decodes the corresponding caption conditioned on these tokens. We optimize the standard autoregressive cross-entropy (CE) loss with respect to the ground-truth caption C:

$$\mathcal{L}_{\text{cap}} = \text{CE}\big(C,\ A_0(\mathbf{c})\big), \tag{8}$$

This first stage establishes a unified communication space, serving as the foundation for subsequent query-aware evidence summarization and cross-agent collaboration.

Evidence summarization. After semantic alignment, communication tokens lie in a shared latent space, but may not yet encode query-relevant information. Local agents must therefore learn to select and compress task-relevant evidence under a fixed communication budget. To this end, we perform supervised fine-tuning (SFT) on a high-quality Image-QA dataset. This stage encourages local agents to encode question-conditioned evidence into compact communication tokens, as illustrated in Fig 2 (c). Given an image X and question q, a local agent A_m produces communication tokens $\mathbf{c} = A_m(X, q)$ and the coordinator A_0 conditions on both the communication tokens and the query embedding to generate the target answer. Training is performed using a standard autoregressive cross-entropy loss on the target sequence:

$$\mathcal{L}_{\text{evi}} = \text{CE}\big(y, A_0([\mathbf{c}; \text{Emb}(q)]\big). \tag{9}$$

This stage trains local agents to produce compact, query-aware evidence summaries consumable by the coordinator.

Cross-agent collaboration. The ultimate goal of MACF-Video is to reason over distributed observations collected by multiple local agents. At this stage, the coordinator must not only interpret individual communication tokens, but also fuse evidence across agents and perform global coordination, as illustrated in Fig 2 (a). We train this capability using Video-QA datasets. Following the prediction formulation in Eq. 5, the final loss takes the form:

$$\mathcal{L}_{\text{col}} = \text{CE}\big(y, A_0([\mathbf{c}^{(1)}; ...; \mathbf{c}^{(M)}; \text{Emb}(q)])\big). \tag{10}$$

Overall, MACF-Video enables scalable video understanding under explict perception and communication budgets by distributing video perception across local agents and exchanging compact latent communication tokens with a coordinator. A three stage curriculum forms the shared communication space and equips the coordinator to fuse distributed evidence for global reasoning.

4 EXPERIMENT

To validate MACF-Video under budget-limited video understanding, we set the perception budget and conduct extensive experiments on video understanding task. In this section, we first details the setup of experiments. Then we report main results on four long video benchmarks (Sec. 4.2), followed by analyses on communication, efficiency, and ablations of our training curriculum (Sec. 4.3).

4.1 SETUP

Implementation. We use Qwen3-VL-8B Bai et al. (2025a), Qwen2.5-VL-7B Bai et al. (2025b) and LLaVA-OineVison1.5-8B An et al. (2025) as the local agents and the coordinator agent in the main results. During training on video data, each local agent receives $F = 16$ frames as input, and each frame is processed at a maximum resolution, $h \times w = 224 \times 224$. For the final training stage, we partition each video across $M = 4$ local agents for distributed understanding under the same per-agent input budge. At inference time, we increase the number of local agents to 6. All training is conducted on 4×NVIDIA A100 (80GB) GPUs. More details can be found in App. A.

Baselines. We mainly compare our method with twelve open-source and proprietary state-of-the-art MLLMs: GPT-4o Hurst et al. (2024), LLaVA-Next-Video-34BZhang et al. (2024b), ShareGPT4Video-8B Chen et al. (2024a), Kangaroo-8B Liu et al. (2024), VideoLLaMA3-7B Zhang et al. (2025a), Qwen2.5-VL-72B Bai et al. (2025b), Qwen2.5-VL-7B, LLaVA-OneVision1.5-8B An

Table 1: Quantitative comparisons of the state-of-the-art MLLMs on the diverse video understanding benchmark. The best and second best results are **bold** and underlined. All the numbers are presented in %. The perception budget is specified as the sampled frame number times processing resolution ($F \times h \times w$). * denotes that MapReduce and LatentMAS use Qwen3-VL-8B as the VLM engine.

Model	Video-MME	LongVideoBench	LVBench	MLVU-Test	Perc. budget
MLLMs without perception budget constraint					
GPT-4o	71.9	66.7	30.8	54.9	-
LLaVA-Next-Video-34B	52.0	50.5	32.2	-	-
ShareGPT4Video-8B	39.9	41.8	-	33.8	-
Kangaroo-8B	56.0	54.8	39.4	46.5	-
VideoLLaMA3-7B	66.2	59.8	45.3	47.7	-
MLLMs with perception budget constraint					
Qwen2.5-VL-72B	59.5	51.0	35.7	44.2	16*224*224
Qwen3-VL-30B	58.3	55.9	38.0	44.0	16*224*224
Qwen2.5-VL-7B	53.7	48.1	32.2	38.7	16*224*224
LLaVA-OneVision1.5-8B	56.1	54.4	36.4	41.8	16*224*224
Keye1.5-VL-8B	55.6	56.4	37.6	41.6	16*224*224
InternVL2.5-VL-8B	47.7	43.4	32.3	39.1	16*224*224
Qwen3-VL-8B	55.9	50.7	33.2	41.5	16*224*224
Multi-agent systems with perception budget constraint					
MapReduce*	46.7	38.2	30.5	33.3	16*224*224
LatentMAS*	55.7	48.5	33.2	42.3	16*224*224
Ours (Qwen2.5-VL-7B)	58.2	52.5	37.7	47.6	16*224*224
Ours (Qwen3-VL-8B)	**60.4**	56.8	40.2	49.2	16*224*224
Ours (LLaVA-OV1.5-8B)	59.9	**57.6**	**41.1**	**49.4**	16*224*224

et al. (2025), Keye1.5-VL-8B Yang et al. (2025a), InternVL2.5-VL-8B Chen et al. (2024b), Qwen-VL-30B and Qwen3-VL-8B Bai et al. (2025a). Moreover, to verify the effectiveness of our communication protocol, we implement two baseline and conduct experiments following the setups of MapReduce Pang & Wang (2025) and LatentMAS Zou et al. (2025a). For a fair comparison, we replace the closed-source model used in the original MapReduce implementation with Qwen3-VL-8B. We also adapt and modify the official LatentMAS code to support multimodal understanding. The implementation details of these two baselines are provided in the Appendix.

Benchmarks. We evaluate our method on four general-purpose long video understanding benchmarks: Video-MME Fu et al. (2025a), LongVideoBench Wu et al. (2024a), LVBench Wang et al. (2025) and MLVU-Test Zhou et al. (2025a). For MLVU-Test, we evaluate on its multiple-choice question set for stability and consistency. For all evaluations, we follow the decoding configuration used in the official demo code.

4.2 Main Results

As shown in Table 1, we compare against baselines on four standard benchmarks, demonstrating the effectiveness of MACF-Video for video understanding. We highlight two main findings below.

Superior performance. Under the input budget constraint, many existing methods exhibit noticeable performance degradation. In contrast, MACF consistently surpasses previous models across all the benchmarks. Compared with the Qwen3-VL-8B, MACF achieves consistent gains of **+4.5%**, **+6.1%**, **+7.0%**, and **+7.7%** on Video-MME, LongVideoBench, LVBench, and MLVU-Test, respectively. Notably, on MLVU-Test, MACF surpasses all compared open-source models, even when those baselines are evaluated without input budget constraints. Overall, these results validate the effectiveness of MACF's design for budget-limited long-form video understanding.

Generalization to other MLLMs backbone. To demonstrate the generalization of our proposed MACF, we conduct experiments with Qwen2.5-VL-7B and LLaVA-OneVision1.5-8B as further validations. As shown in Table 1, MACF consistently improves performance across all benchmarks when applied to different foundational backbones (e.g., Qwen3, LLaVA, and Qwen2.5). Specifically, our method yields an average absolute improvement of **+6.3%** for Qwen3-VL-8B and achieves an overall average gain of approximately **+5.6%**.

Effective communication representation. We further analyze the effectiveness of our communication protocol by comparing it with text-only communication and a LatentMAS KV-cache sharing baseline under same task setting. Across Video-MME, LongVideoBench, LVBench, and MLVU-

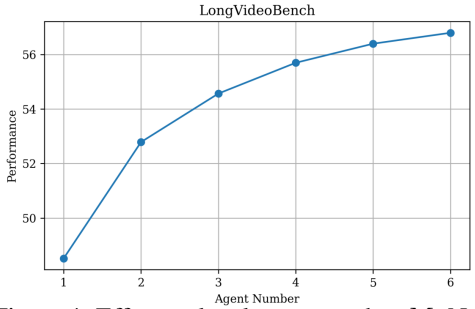


Figure 4: Effect on local agent number M. Note that we set $M = 4$ in training.

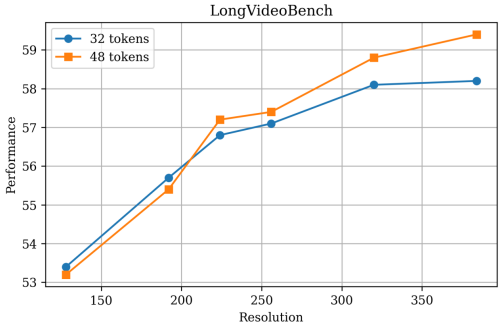


Figure 5: Effect of spatial capacity on LongVideoBench.

Table 2: Effect of communication bottleneck.

K	MLVU	LVBench	LongVideoBench
16	46.0	38.7	54.1
32	49.2	40.2	56.8
48	49.0	40.1	57.2

Table 3: Analysis of communication cost.

Method	Inference Time	Throughput
MapReduce	5.156s	387 tokens
LatentMAS	0.649s	784 KV-Cache
Ours	0.537s	192 tokens

Test, MACF outperforms text-based communication by **13.7%**, **18.6%**, **9.7%**, and **15.9%**, respectively. Compared with directly sharing KV-Caches as in LatentMAS, MACF still achieves consistent improvements of **9.3%**, **14.1%**, **10.5%**, and **10.3%** on the same benchmarks. These results indicate that explicitly designed communication tokens provide a more effective and structured medium for cross-agent information exchange than either textual summarization or raw latent sharing. It enable the coordinator to aggregate visual evidence more reliably under strict input budget constraints.

As shown in Fig. 3, we further conduct a qualitative experiment to compare text-based communication with our latent communication framework. In this example, the question asks for the rope color of the white dog, but the video clips in different agents contain multiple dogs. Under text-only communication, each agent must discretize its observation into a short caption-like statement. This compression not only drops visual details, but also easily introduces a reference binding error: an agent (agent6) may mistakenly associate the queried "white dog" with the yellow-white dog in its own segment and report the rope color as red. Once such error statement enters the shared text context, the coordinator may produces the incorrect answer. In contrast, MACF's communication tokens preserve fine-grained visual evidence in a structured latent form, enabling the coordinator to align information across agents more reliably and recover the correct answer.

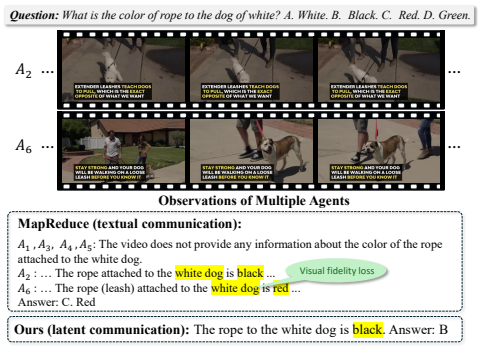


Figure 3: **Qualitative comparison of communication representations.** Compared with the text-based communication, our latent communication preserves fine-grained visual evidence and yields the correct result.

4.3 Discussion

Our framework is fundamentally governed by two information bottlenecks: a *perception bottleneck* at the input level and a *communication bottleneck* at the coordination level. From an information-theoretic perspective, these bottlenecks jointly determine how much task-relevant information can be preserved and transmitted under limited perceptual and communication budgets, and therefore control the system's ability to scale to long and complex videos.

Perception bottleneck. The perception bottleneck characterizes the total amount of visual information that can be processed by the system through all local agents. It is controlled by two complementary factors: temporal and spatial capacity. The total perception volume can be measured by the number of processed pixels, $(M \times F) \times (h \times w)$, where M is the number of agents, F the number

of frames per agent, and $h \times w$ the input frame resolution. Here, $(M \times F)$ represents temporal capacity, while $(h \times w)$ corresponds to spatial capacity.

(i) Effect of temporal capacity. Temporal capacity $(M \times F)$ is determined by the number of local agents M when the per-agent temporal input budget (i.e., the number of frames processed by each agent F) is fixed. Increasing the number of local agents expands the effective temporal coverage of the video, thereby reducing information loss caused by sparse temporal sampling. We study this effect on MLVU-Test by progressively scaling the number of local agents. As shown in Fig. 4, MACF-Video exhibits a consistent performance improvement as the number of agents increases.

(ii) Effect of spatial capacity. Spatial capacity is determined by the spatial resolution $(h \times w)$ of frames processed by each agent, which directly affects the fidelity of local visual representations. To isolate this factor, we fix both the number of local agents and the number of frames processed per agent, and increase the input bandwidth by allocating higher spatial resolution per frame. As shown in Fig. 5, MACF-Video exhibits a monotonic performance improvement as resolution increases, indicating reduced spatial distortion in local observations. However, as illustrated by the blue curve with $K = 32$ communication tokens, this improvement saturates beyond a certain point once the communication bottleneck, measured by $K \times M$ (see Sec. 4.3), becomes limiting. Notably, increasing the communication budget to $K = 48$ (orange curve) raises the performance ceiling, suggesting that higher-fidelity local representations require greater communication capacity for effective aggregation by the coordinator.

Communication bottleneck. The communication bottleneck characterizes the information transmission constraint between local agents $\{A\}_{m=1}^{M}$ and the coordinator A_0. The total communication capacity is bounded by $K \times M + \mathrm{token}(q)$, where M is the number of agents and K the number of communication tokens per local agent (we omit $\mathrm{token}(q)$ as it is data-dependent). For fixed M, this bottleneck is governed by K and can be viewed as a *rate constraint* on the information available to the global inference module. Under the main experimental setting wit fixed M, we evaluate communication budgets of $K \in \{16, 32, 48\}$. As shown in Table 2, $K = 16$ results in substantial performance degradation, indicating excessive compression and loss of task-relevant information. Increasing the budget to $K = 32$ yields a significant improvement, while further increasing to $K = 48$ provides only marginal gains. This saturation behavior suggests that once the communication rate exceeds the task-relevant information of the perceived observations, additional bandwidth offers diminishing returns, a hallmark of information bottleneck regimes.

Together, these two bottlenecks define a coupled rate-distortion trade-off between local perception and global aggregation. Insufficient perception bandwidth increases distortion at the input level, while insufficient communication capacity prevents the coordinator from reconstructing a task-sufficient global representation. By jointly controlling these two constraints, MACF-Video enables scalable video understanding that approaches an efficient operating point under strict perception and communication budgets.

Communication cost. As shown in Table 3, we report the communication throughout and response latency for different communication methods. Our approach achieves the lowest throughout and the shortest response latency. In contrast, LatentMAS Zou et al. (2025b) incurs the highest communication cost due to the necessity of transmitting high dimensional KV-Caches. Meanwhile, MapReduce Pang & Wang (2025) wit text-based communication suffers from the highest latency, primarily stemming from the autoregressive decoding process required for each agent interaction.

5 Conclusion

We present MACF-Video, an end-to-end multi-agent collaboration framework that enables scalable video understanding under constrained perception and communication budgets. By decomposing long videos across locally budgeted agents and introducing an agent-native latent communication protocol, MACF-Video decouples per-agent perception limits from global video complexity while preserving fine-grained visual semantics. A curriculum training strategy stabilizes optimization and promotes effective cross-agent coordination in the shared communication space. Extensive experiments demonstrate that MACF-Video consistently outperforms state-of-the-art MLLMs and multi-agent baselines under identical budget constraints. Together, these results highlight compact latent multi-agent collaboration as a promising paradigm for scalable and efficient video understanding.

AI USE STATEMENT

In this work, we evaluated multimodal large language models (MLLMs), including GPT-4o, as systems under evaluation on long video understanding benchmarks. These models generated predictions for the benchmark experiments described in the paper. We also used ChatGPT solely to improve the grammar, clarity, and readability of the manuscript. ChatGPT was not used to generate substantive scientific content, including research hypotheses, methodology, analyses, interpretations, claims, or conclusions. The MLLMs used in the benchmark experiments served only as systems under evaluation and were not used to design the research methodology or prepare the manuscript. Other tasks listed as requiring disclosure under the ICLR policy were either not performed with generative AI or were not applicable to this work. We take full responsibility for the experimental design, analysis, and final content of the paper.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Xiang An, Yin Xie, Kaicheng Yang, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: a survey. *arXiv preprint arXiv:2402.12451*, 2024.

Shidong Cao, Zhonghan Zhao, Shengyu Hao, et al. Efficient transfer from image-based large multimodal models to video tasks. *IEEE Transactions on Multimedia*, 2025.

Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023.

Lin Chen, Xilin Wei, Jinsong Li, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024a.

Zhe Chen, Weiyun Wang, Yue Cao, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Christopher Clark, Jieyu Zhang, Zixian Ma, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*, 2026.

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

Kaituo Feng, Kaixiong Gong, Bohao Li, et al. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025a.

Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. Cache-to-cache: Direct semantic communication between large language models. *arXiv preprint arXiv:2510.03215*, 2025b.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pp. 5267–5275, 2017.

Tao Ge, Jing Hu, Lei Wang, et al. In-context autoencoder for context compression in a large language model. *arXiv preprint arXiv:2307.06945*, 2023.

Wei Han, Hui Chen, Min-Yen Kan, and Soujanya Poria. Self-adaptive sampling for accurate video question answering on image text models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2522–2534, 2024.

Bo He, Hengduo Li, Young Kyun Jang, et al. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13504–13514, 2024.

Jie He, Richard He Bai, Sinead Williamson, et al. Clara: Bridging retrieval and generation with continuous latent reasoning. *arXiv preprint arXiv:2511.18659*, 2025.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*, 2023.

Aaron Hurst, Adam Lerer, Adam P Goucher, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Noriyuki Kugo, Xiang Li, Zixin Li, Ashish Gupta, Arpandeep Khatua, Nidhish Jain, Chaitanya Patel, Yuta Kyuragi, Yasunori Ishii, Masamoto Tanabiki, et al. Videomultiagents: A multi-agent framework for video question answering. *arXiv preprint arXiv:2504.20091*, 2025.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023a.

Guozhang Li, Xinpeng Ding, De Cheng, et al. Etc: Temporal boundary expand then clarify for weakly supervised video grounding with multimodal large language model. *IEEE Transactions on Multimedia*, 2025a.

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

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *Science China Information Sciences*, 68(10):200102, 2025b.

Yunxin Li, Baotian Hu, Xinyu Chen, Lin Ma, Yong Xu, and Min Zhang. Lmeye: An interactive perception network for large language models. *IEEE Transactions on Multimedia*, 26:10952–10964, 2024.

Jianxin Liang, Xiaojun Meng, Yueqian Wang, Chang Liu, Qun Liu, and Dongyan Zhao. End-to-end video question answering with frame scoring mechanisms and adaptive sampling. *arXiv preprint arXiv:2407.15047*, 2024.

Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *IEEE Transactions on Multimedia*, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024.

Runtao Liu, Ziyi Liu, Jiaqi Tang, Yue Ma, Renjie Pi, Jipeng Zhang, and Qifeng Chen. Longvideoagent: Multi-agent reasoning with long videos. *arXiv preprint arXiv:2512.20618*, 2025a.

Weijia Liu, Bo Miao, Jiuxin Cao, et al. Context-enhanced video moment retrieval with large language models. *IEEE Transactions on Multimedia*, 2025b.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. *arXiv preprint arXiv:2508.09736*, 2025.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, et al. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *arXiv preprint arXiv:2411.13093*, 2024.

Mohammad Mozaffari, Amir Yazdanbakhsh, and Maryam Mehri Dehnavi. Slim: One-shot quantization and sparsity with low-rank approximation for llm weight compression. *arXiv preprint arXiv:2410.09615*, 2024.

Ziqi Pang and Yu-Xiong Wang. Mr. video:" mapreduce" is the principle for long video understanding. *arXiv preprint arXiv:2504.16082*, 2025.

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)*, pp. 15174–15186, 2024.

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

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

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025.

Sainbayar Sukhbaatar, Rob Fergus, et al. Learning multiagent communication with backpropagation. *Advances in neural information processing systems*, 29, 2016.

Yunlong Tang, Jing Bi, Siting Xu, et al. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76. Springer, 2024.

Zeqing Wang, Wentao Wan, Qiqing Lao, et al. Towards top-down reasoning: An explainable multi-agent approach for visual question answering. *IEEE Transactions on Multimedia*, 2026.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

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

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024b.

Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. Long video understanding with learnable retrieval in video-language models. *IEEE Transactions on Multimedia*, 2026.

Biao Yang, Bin Wen, Boyang Ding, et al. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*, 2025a.

Jihan Yang, Shusheng Yang, Anjali W Gupta, et al. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025b.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

Yuehao Yin, Huiyan Qi, Bin Zhu, Jingjing Chen, Yu-Gang Jiang, and Chong-Wah Ngo. Foodlmm: A versatile food assistant using large multi-modal model. *IEEE Transactions on Multimedia*, 2025.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-voyager: Learning to query frames for video large language models. *arXiv preprint arXiv:2410.03226*, 2024.

Xinlei Yu, Zhangquan Chen, Yongbo He, et al. The latent space: Foundation, evolution, mechanism, ability, and outlook. *arXiv preprint arXiv:2604.02029*, 2026.

Boqiang Zhang, Kehan Li, Zesen Cheng, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024a.

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

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *arXiv preprint arXiv:2505.18079*, 2025b.

Yuanhan Zhang, Bo Li, haotian Liu, et al. Llava-next: A strong zero-shot video understanding model, April 2024b. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Yuanhan Zhang, Jinming Wu, Wei Li, et al. Video instruction tuning with synthetic data, 2024c. URL https://arxiv.org/abs/2410.02713.

Junjie Zhou, Yan Shu, Bo Zhao, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13691–13701, 2025a.

Yiyang Zhou, Yangfan He, Yaofeng Su, Siwei Han, Joel Jang, Gedas Bertasius, Mohit Bansal, and Huaxiu Yao. Reagent-v: A reward-driven multi-agent framework for video understanding. *arXiv preprint arXiv:2506.01300*, 2025b.

Jiaru Zou, Xiyuan Yang, Ruizhong Qiu, et al. Latent collaboration in multi-agent systems. *arXiv preprint arXiv:2511.20639*, 2025a.

Jiaru Zou, Xiyuan Yang, Ruizhong Qiu, et al. Latent collaboration in multi-agent systems. *arXiv preprint arXiv:2511.20639*, 2025b.

Jialong Zuo, Yongtai Deng, Lingdong Kong, Jingkang Yang, Rui Jin, Yiwei Zhang, Nong Sang, Liang Pan, Ziwei Liu, and Changxin Gao. Videolucy: Deep memory backtracking for long video understanding. *arXiv preprint arXiv:2510.12422*, 2025.

APPENDIX

A IMPLEMENTATION DETAILS.

MACF-Video Implementation Details. We introduce K special token <COM_K>. During training, we update the input embedding associated with this token. All parameters of the base model (Qwen3-VL-8B) are frozen, and we perform parameter-efficient fine-tuning with LoRA using r=16, α=32, and dropout 0.05. The adapter is implemented as a two-layer MLP. As mentioned in the Method section, the training proceeds in three stages. In Stage 1, we use caption data from the 0-30s subset of LLaVA-Video-178K Zhang et al. (2024c). In Stage 2, we use image data from Video-R1 Feng et al. (2025). In Stage 3, we train on video data from Video-R1 together with a subset of Molmo2 Clark et al. (2026). For captioning, we truncate the textual sequence to a maximum length of 256 tokens and compute the training loss only over the truncated tokens.

LatentMAS Implementation Details. We implement LatentMAS based on the official codebase. Following the MACF-Video setting, we uniformly split each video into $N = 6$ segments. For each segment, LatentMAS generates $K = 10$ latent tokens, and we cache the corresponding KV states. In our preliminary experiments, using only K latent tokens per segment was insufficient to capture the full video context. Therefore, we modify the inference pipeline by concatenating the KV caches from all six segments, then uniformly subsampling the KV entries $M = 724$. Finally, we concatenate the subsampled KV states with the latent tokens in temporal order and feed them to the decoder as the conditioning context. For fair comparison, we also choose the Qwen3-VL-8B as the understanding model.

Text-based Communication Implementation Details. We follow the same video sampling strategy and segment partitioning as in previous setting to ensure a fair comparison. Similar to MapReduce baselines, we use a vision language model (VLM) to generate a caption for each video segment. Concretely, we adopt Qwen3-VL-8B as the captioning backbone and also use the this model as the reducer for final prediction, matching the baseline setting. We use captions with a format consistent with prior work, and for each question we additionally prompt the VLM to produce a query-aware evidence summary from the segment captions, so that the reducer receives concise, question-relevant textual evidence rather than raw captions.

B MORE COMPARISON CASE.

As shown in Fig. 6, we provide an additional qualitative example comparing text-based communication with our latent communication framework. In this case, the evidence required to answer the question is distributed across two agents, so arriving at the correct answer depends on cross-segment reasoning and implicit state tracking to integrate information from multiple sources. Moreover, the key evidence is primarily at the concept level, which is difficult to represent faithfully and consistently using discrete textual summaries. For textual communication baseline, Agent2 correctly identifies the crucial concept in the image—quantitative easing—and captures it in its caption. However, Agent3, based on its own segment, infers that the topic is UK government debt. Because this information remains fragmented across agents, the final global prediction fails to associate Agent2's correct evidence with the question, and instead over-relies on Agent3's partial inference, ultimately producing an incorrect answer. In contrast, MACF-Video's communication tokens preserve fine-grained visual evidence in a structured latent form, enabling the coordinator to align information across agents more directly and recover the correct answer. This qualitative result highlights the effectiveness of our communication tokens for distributed video understanding.

C DATASET DETAILS.

Video-MME. Video-MME is a widely used benchmark for general video understanding. It contains videos collected from diverse domains, including movies, documentaries, sports, news, and daily activities, and evaluates models through multiple-choice question answering. The benchmark emphasizes comprehensive temporal understanding and long-context reasoning, covering tasks such as action recognition, event understanding, temporal ordering, and causal reasoning. Following the standard protocol, we report the average accuracy over the official evaluation split.



Question: *What does the combination of a bond certificate and the UK flag most likely symbolize in the video?*
A. The UK's dependence on foreign investments. *B. The UK's use of bonds for quantitative easing*
C. The UK's high national debt. *D. The UK's strong bond market.*

MapReduce (textual communication):

A_1, A_4, A_5, A_6: The video does not show a bond certificate combined with the UK flag.
A_2 : … The video does not show a bond certificate combined with the UK flag ... Quantitative easing…, a bank…, stacks of money……, and a printer printing money.
A_3 : … The video shows a bond certificate alongside the UK flag, followed by the text "£375bn since 2009." This strongly suggests a cumulative figure related to UK government debt or bond issuance over time....
Answer: C. The UK's high national debt.

Ours (latent communication): Answer: B. The UK's use of bonds for quantitative easing

Figure 6: Qualitative comparison of communication representations. Text-based communication (MapReduce) and our latent communication.

LongVideoBench. LongVideoBench is designed to evaluate understanding of long-duration videos with complex temporal dependencies. The benchmark includes videos with substantially longer durations than conventional short-video datasets and requires models to integrate information distributed across distant temporal segments. The evaluation mainly focuses on multi-hop reasoning, temporal event localization, and long-context comprehension under constrained visual budgets.

LVBench. LVBench is a benchmark targeting long-video reasoning and temporal understanding. LVBench contains challenging question-answer pairs that require models to aggregate evidence across multiple video segments rather than relying on local visual cues. The dataset particularly stresses temporal consistency, cross-scene reasoning, and the ability to maintain coherent representations over extended video contexts.

MVLU. MVLU evaluates multimodal video-language understanding abilities across a wide range of tasks, including event understanding, human-object interaction recognition, temporal reasoning, and fine-grained semantic comprehension. Compared with traditional short-video benchmarks, MVLU places stronger emphasis on compositional reasoning and the alignment between visual content and language queries.

VSI-Bench. VSI-Bench focuses on scene-level video understanding and holistic interpretation of long-form video content. The benchmark requires models to capture high-level semantic transitions, scene dynamics, and contextual relationships across temporally separated events. Such settings are particularly suitable for evaluating the effectiveness of communication mechanisms in distributed video understanding systems.

D EFFECT OF TRAINING STRATEGY.

To investigate the effect of our proposed training strategy, we conduct experiments on Video-MME and MLVU. As shown in the Table 4, the full training strategy achieves the best performance. Removing any stage leads to substantial performance drops, with average decreases of **13.9%**, **3.0%**, and **12.5%**, respectively. These ablation results provide three key findings. First, stage 1 demonstrates that aligning the communication tokens with the coordinator's

Table 4: Effect of training stages.

Stage1	Stage2	Stage3	Video-MME	MLVU
×	✓	✓	45.5	36.3
✓	×	✓	56.9	46.7
✓	✓	×	48.2	36.5
✓	✓	✓	60.4	49.2

Table 5: Analysis of communication budget.

Method	Video-MME	LongVideoBench	LVBench	MLVU-Test	Comm. budget (#tokens)
MapReduce	44.5	36.9	30.3	32.4	192
MapReduce	46.7	38.2	30.5	33.3	None
LatentMAS	51.1	42.7	29.7	38.9	192
LatentMAS	55.7	48.5	33.2	42.3	None
Ours	60.4	56.8	40.2	49.2	192

Table 6: Complex task evaluation.

Model	Obj. Count	Route Plan	Appr. Order	Charades
Qwen3	47.7	28.9	52.4	0.38 (mIoU)
Ours	**51.5**	**31.4**	**57.8**	**0.45** (mIoU)

representation space forms a well-initialized interface that is important for complex QA. Second, stage 2 shows that strengthening high-quality image QA understanding substantially benefits video understanding. Third, stage 3 highlights the importance of learning cross-agent collaboration for effective distributed communication.

E COMMUNICATION BUDGET.

To isolate the effect of the inter-agent communication mechanism, we fix the communication budget to 192 tokens and adopt the same backbone (Qwen3-VL-8B), video partitioning, and perception budget ($16 \times 224 \times 224$) across all compared protocols. As shown in Table 5, both protocols fall substantially behind our method under the same 192-token budget, with average gaps of 15.6 and 11.1 points across the four benchmarks for MapReduce and LatentMAS, respectively. The reasons differ between the two baselines. For MapReduce, communication is performed in natural language, so a 192-token budget forces aggressive textual summarization that discards fine-grained spatio-temporal evidence. For LatentMAS, although messages are exchanged in a continuous latent space, the latent representation is not explicitly optimized for compact transmission, so truncating it to 192 tokens removes a large fraction of task-relevant signal. In contrast, our protocol learns to concentrate task-relevant evidence into a compact set of latent tokens, yielding a markedly better accuracy-bandwidth trade-off.

F MORE COMPLEX TASK RESULTS.

To demonstrate the latent communication protocol has the capability to transfer to more complex task, we provide additional evaluation on VSI-Bench Yang et al. (2025b) (complex reasoning) and Charades-STA Gao et al. (2017) (temporal grounding) under the same perception budget. For VSI-Bench, we evaluate object counting, route planning, and appearance ordering without re-training. For Charades-STA, we prompt each local agent with its time segment and finetune using the temporal video grounding subset of TimeIT Ren et al. (2024). As shown in Table 6, MACF-Video consistently outperforms the single-agent baseline on both task types, confirming generalization to complex video understanding task.

G SCALING OF THE LOCAL AGENT NUMBER.

To evaluate scaling behavior, we conducted inference-time experiments with varying agent counts. Performance improves from $M = 2$ to $M = 7$ but saturates beyond $M = 7$. Note that MACF-Video is trained with $M = 4$ agents, so results for $M > 4$ represent inference-time extrapolation, where the coordinator becomes the bottleneck as it must aggregate more tokens than seen during training. Training with larger M would alleviate this, but requires more GPU memory (we currently train on a single device to avoid inter-device communication overhead). A promising future direction is

Table 7: Scaling behavior of MACF.

Agent Number	2	4	5	6	7	8	9	10
LongVideoBench	52.8	55.7	56.4	56.8	57.0	56.9	57.1	57.0

Table 8: Model agnostic. Performance of different local agents paired with a fixed coordinator agent (Qwen3-VL-8B).

Local Agent	Video-MME	LongVideoBench	LVBench	MLVU-Test
Qwen3-VL-8B	60.4	56.8	40.2	49.2
Qwen2.5-VL-7B	56.4	51.1	36.5	47.8
Qwen2.5-VL-7B & LLaVA-OV1.5-8B	56.6	52.0	37.0	48.1

autoregressive aggregation, where the coordinator ingests tokens from local agents autoregressively, decoupling training agent count from GPU memory constraints.

H MODEL AGNOSTIC.

To demonstrate the model-agnostic nature of MACF-Video, we conducted experiments with a fixed coordinator backbone (Qwen3-VL-8B) and different local agent backbones. Specifically, we evaluated two distinct configurations: one where all local agents utilize Qwen2.5-VL-7B, and another employing a heterogeneous mixture of Qwen2.5-VL-7B and LLaVA-OneVision1.5-8B. As shown in Table 8, the results confirm that MACF-Video generalizes beyond a single model family. This is enabled by the adapter module (2-layer MLP), which projects any backbone's hidden states into a shared latent space with fixed dimensionality $c^{(m)} \in \mathbb{R}^{K\times d}$, absorbing architectural differences across backbones.

I NECESSITY OF THE SETUP.

The constraint in Eq. 1 is not an artificial restriction but a faithful abstraction of how MLLMs are actually deployed. On the model side, every MLLM has a finite context that is easily exceeded by long videos with thousands of frames, and dense visual tokens further inflate activation and KV-cache memory beyond what a single forward pass can afford. On the deployment side, commercial APIs enforce strict per-request frame caps with non-trivial bandwidth overhead, while edge or privacy-sensitive scenarios must run locally with limited compute. Following Video-R1 Feng et al. (2025), which validates its reasoning design with only a few frames, we fix B^{per} to $16 \times 224 \times 224$ as a representative operating point to evaluate MACF-Video. This value serves as a stand-in for the effective per-forward visual context capacity rather than a hard model limit. MACF-Video is thus complementary to advances in context length: rather than chasing an ever-larger B^{per}, we make better use of whatever budget is available.