Proc. of International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA 2026)
5-7 February 2026, Boracay Island-Philippines

2026 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA) | 979-8-3315-7191-7/26/$31.00 ©2026 IEEE | DOI: 10.1109/ACDSA67686.2026.11468028

MAS-LLaVA: Motion-Aware Adaptive Sampling for Training-Free Video Large Language Models

Jialin Tang
Department of Computer Science
California State University Fullerton
Fullerton, USA
tjl_0516@csu.fullerton.edu

Yu Bai
Department of Electrical and Computer Engineering
California State University Fullerton
Fullerton, USA
ybai@fullerton.edu

***Abstract*—Recent advances in video large language models (VLLMs) have enabled strong zero-shot reasoning, yet most systems still rely on uniform frame sampling and fragile GPU execution pipelines. We propose MAS-LLaVA, a training-free enhancement to LLaVA that introduces uniform frame sampling, adaptive importance-based token sampling, and device-consistent inference. First, uniform frame sampling ensures balanced temporal coverage by selecting equidistant frames, maintaining compatibility with existing aggregation strategies. Second, an adaptive token selection module computes the feature magnitude and diversity of patch tokens, assigning probabilistic importance scores and sampling a compact subset of visual tokens under a fixed token budget. Third, a device-aware execution pipeline ensures that all intermediate tensors inherit the input frame's device and data type, allowing uniform and adaptive sampling strategies to run reliably across heterogeneous GPUs. Experiments on NExT-QA and IntentQA show that MAS-LLaVA improves accuracy and stability across diverse video inputs without any retraining. These findings demonstrate that smarter, training-free sampling and inference design can substantially improve both the practicality and robustness of VLLMs in real-world video–language understanding.**

Keywords*—*Uniform Frame Sampling, Adaptive Token Selection, Device-Aware Inference, Training-Free Video LLM, Visual Complexity, Video Question Answering, Efficiency, Robustness

I. Introduction

Recent advances in large multimodal models have significantly improved performance on visual–language understanding tasks such as captioning, retrieval, and question answering [1], [2]. In particular, *Video Large Language Models* (VLLMs) have emerged as powerful architectures that couple high-capacity vision encoders with pretrained large language models (LLMs) to interpret and reason over long-form video content [3]–[5]. By aligning tokenized spatiotemporal visual features with language representations, VLLMs extend image-based vision–language models to handle sequential, motion-rich, and context-dependent inputs.

Modern VLLMs such as Video-LLaVA [3], TS-LLaVA [4], and GPT-4V [6] integrate video frame encoders (e.g., CLIP, ViT, or EVA) with autoregressive LLMs through multimodal adapters that fuse visual and textual embeddings. This design enables complex tasks such as temporal event localization, human–object interaction analysis, and video-grounded dialogue. More recent efforts incorporate token compression, keyframe selection, and memory-aware attention to improve scalability and maintain context across long videos [7].

VLLMs have shown promise in diverse applications, including video question answering, autonomous driving perception, educational tutoring, surveillance summarization, and multimodal content generation. For example, models such as Video-ChatGPT [5] and the hardware-aware optimization strategy HAO [8] improve the efficiency and robustness of multimodal inference pipelines, while UniVideo-like unified modeling approaches [7] aim to support tasks such as captioning, action recognition, and summarization within a single framework. However, many VLLMs rely on extensive retraining or domain-specific fine-tuning, limiting their scalability and consistency across hardware platforms. Furthermore, the *sampling pipeline*, which governs how frames are selected before multimodal fusion, plays a crucial role in determining efficiency and temporal fidelity. Existing sampling methods are often static and assume homogeneous computational environments, which can lead to inference failures or inconsistent outputs on heterogeneous GPUs [9], [10]. In addition, recent advances in hardware-aware neural network optimization have demonstrated that efficient model design can significantly improve robustness under resource constraints. Representative works include binarized and stripe-wise optimization [11], tensor-train–based model compression for recurrent networks [12], focal-loss–enhanced compression strategies [13], and locality-sensing acceleration for real-time video inference on embedded platforms [14]. These efforts highlight the importance of efficiency-aware designs, echoing MAS-LLaVA's motivation to improve training-free VLLM inference without retraining. However, these issues are not merely implementation details—they fundamentally limit the practical deployment of VLLMs. Static or device-inconsistent sampling pipelines often lead to unstable inference, numerical divergence, or degraded visual grounding when deployed

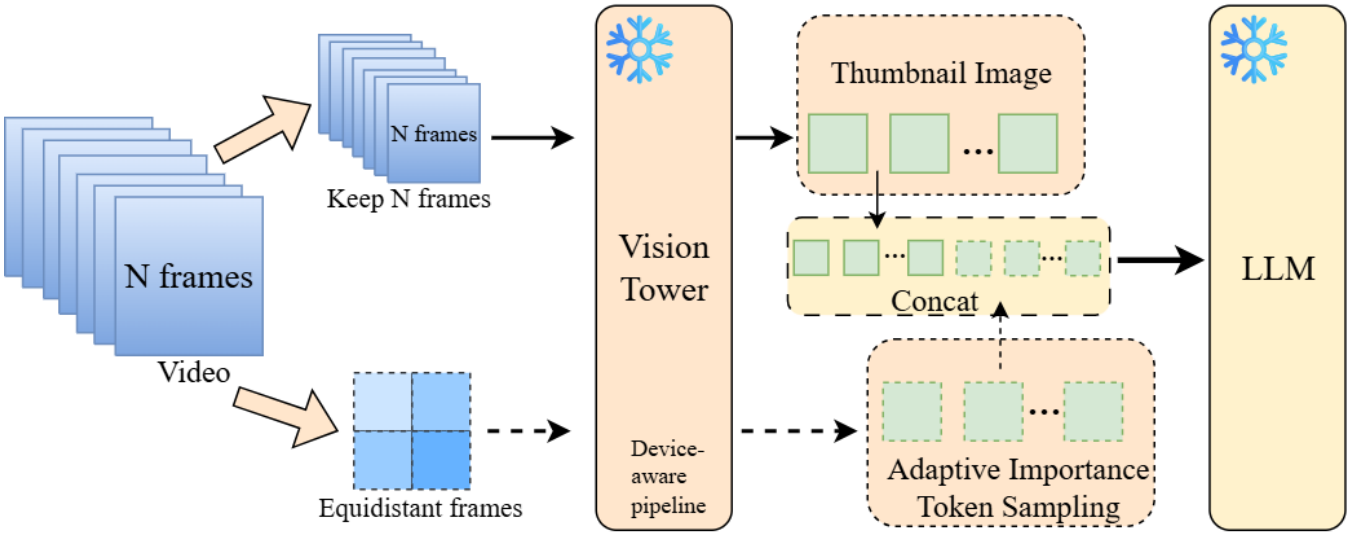

Fig. 1. Overall architecture of the proposed MAS-LLaVA framework. The uniform frame sampling ensures balanced temporal coverage by selecting equidistant frames, while the adaptive importance-based token sampling dynamically selects informative and diverse tokens based on feature magnitude and diversity. The device-aware pipeline guarantees cross-device consistency during vision feature extraction. Thumbnail images provide a stable visual reference, and all selected tokens are concatenated and fed into the LLM for video–language understanding.

across heterogeneous GPUs. Meanwhile, retraining or fine-tuning large-scale video–language models for every hardware environment is computationally prohibitive and data-inefficient. Therefore, there is a strong need for a training-free yet device-adaptive framework that can maintain robustness, efficiency, and semantic fidelity without retraining. To address these challenges, we propose **MAS-LLaVA** (Motion-Aware Adaptive Sampling), a *training-free* framework designed to improve the robustness and efficiency of VLLMs without modifying their core architecture, as illustrated in Fig. 1.

Our framework introduces three key components.

First, a **uniform frame sampling** module ensures balanced temporal coverage by selecting equidistant frames, providing a lightweight and compatible mechanism for long-form video inputs.

Second, an **adaptive importance-based token sampling (MAS)** module evaluates the feature magnitude and diversity of patch tokens extracted from these frames and probabilistically selects a compact, informative subset under a fixed token budget. This process effectively captures **importance- and diversity-aware** cues at the feature level, enhancing the representational richness of visual tokens.

Third, a **device-aware sampling pipeline** ensures stable execution across heterogeneous GPU environments by explicitly aligning tensor device and data type to avoid implicit casts.

Extensive experiments on multiple video question answering benchmarks demonstrate that MAS-LLaVA consistently improves reasoning quality, stability, and scalability while maintaining computational efficiency.

II. Related Work

Recent advances in multimodal large language models (LLMs) have substantially strengthened the integration of visual and linguistic understanding. Early image-centric frameworks such as BLIP-2 and InstructBLIP introduced efficient cross-modal alignment by employing frozen image encoders together with lightweight adapters (e.g., Q-Formers) to bridge visual representations and LLMs [2], [9]. Subsequent models, including Qwen-VL [15] and mPLUG-Owl [16], further expanded this paradigm through modularized multimodal fusion and multi-stage instruction tuning, leading to more robust performance across diverse vision–language tasks. Meanwhile, transformer-based architectures have also demonstrated efficiency gains beyond vision–language modeling. For instance, HyperEAST [17] proposed a lightweight transformer that replaces the conventional dot-product attention with a Linear Fusion Attention Mechanism, achieving compact spectral–spatial modeling while maintaining high classification accuracy. This underscores the versatility of efficient attention mechanisms across modalities.

In parallel, the LLaVA family of models adopts a lightweight projection-based multimodal alignment strategy. Video-LLaVA [3] extends the original LLaVA [18] design by connecting vision encoders to LLMs through an MLP-based adapter and performing alignment-before-projection for unified video representations. Later developments such as LLaVA-NeXT [19] further enhanced multimodal reasoning capabilities by leveraging stronger LLM backbones and improved instruction-following data. Similarly, MiniGPT-4 [20] demonstrated that frozen vision encoders and language models can be effectively aligned using simple linear projections, highlighting the scalability of lightweight multimodal adapters.

While image-based multimodal models rely on static paired image–text datasets, video LLMs extend these systems into the temporal domain, requiring explicit modeling of motion and long-range dynamics. Training-based approaches such

as Video-ChatGPT [21], Video-LLaMA [22], VideoLLaMA 2 [23], and Video-LLaVA [3] adapt image-centric LLMs to videos by fine-tuning on large-scale video–text corpora and incorporating temporal strategies such as frame pooling, cross-modal attention, or SlowFast-style dual-pathway encoders [24]. Although these models achieve strong performance on complex video understanding tasks, they require extensive training data and substantial computational resources, making them difficult to scale or deploy across heterogeneous hardware environments.

In contrast, training-free video LLMs reuse pre-trained image LLMs without additional fine-tuning, relying instead on efficient visual token compression or temporal restructuring. The IG-VLM framework [25] converts a sequence of frames into a grid-based image representation to remain compatible with image-centric LLMs, though the compression inevitably reduces spatial resolution and weakens temporal fidelity. FreeVA [26] explores lightweight token pooling and temporal aggregation to construct compact video representations under strict token budgets. More recently, SlowFast-LLaVA [27] incorporates a SlowFast-style dual-pathway design inspired by audiovisual recognition architectures [28], enabling training-free video understanding by jointly modeling long-term context and short-term motion cues without updating model parameters.

Building upon these insights, our proposed MAS-LLaVA introduces a **uniform frame sampling and adaptive token selection strategy** that unifies temporal coverage and feature-level adaptivity. Specifically, uniformly spaced frames are used to preserve global scene context, while adaptive importance-based token sampling selects the most informative and diverse tokens based on feature magnitude and diversity. This hybrid design enables MAS-LLaVA to achieve a favorable balance between token efficiency and representational richness. Empirically, MAS-LLaVA delivers consistent improvements over existing training-free video LLMs in both accuracy and stability, demonstrating superior generalization across diverse video–language benchmarks.

III. Overview of the Proposed MAS-LLaVA Framework

We propose **MAS-LLaVA** to enhance the robustness and efficiency of existing training-free video LLMs. While such models effectively compress video frames into visual tokens through a thumbnail-and-sampling mechanism, they still face two major limitations: (1) unstable execution across heterogeneous GPU devices due to inconsistent tensor alignment and memory handling, and (2) limited adaptability when representing videos with diverse spatial and semantic complexity.

To overcome these issues, MAS-LLaVA introduces three complementary components—a *Uniform Frame Sampling* module, an *Adaptive Importance-Based Token Sampling (MAS)* module, and a *Device-Aware Sampling Pipeline*—which together stabilize inference and strengthen semantic reasoning without modifying any pretrained parameters of the underlying video LLM. The first module ensures balanced temporal coverage by selecting equidistant frames, the second adaptively selects informative and diverse visual tokens to improve representation quality under a fixed context length, and the third guarantees device-consistent execution across heterogeneous hardware environments. Although the current implementation employs uniform frame sampling, the adaptive token selection effectively captures **importance- and diversity-aware** cues at the feature level, aligning with the overall design philosophy of MAS-LLaVA.

A. Device-Aware Sampling Pipeline

Training-free VLLMs frequently encounter cross-device execution issues such as mismatched tensor allocation or implicit `dtype` conversions, especially when running on different accelerators (e.g., RTX 3090, A100, L40). To address this, MAS-LLaVA establishes a unified device-aware sampling pipeline, in which every intermediate tensor explicitly inherits the device and data type from the input video frames. Formally, for any transformation operation $\mathcal{T}$ applied to a feature tensor $\mathbf{X}$ referenced to an input frame $\mathbf{R}$, we enforce:

$$\text{device}(\mathcal{T}(\mathbf{X};\mathbf{R})) = \text{device}(\mathbf{R}), \tag{1}$$

$$\text{dtype}(\mathcal{T}(\mathbf{X};\mathbf{R})) = \text{dtype}(\mathbf{R}). \tag{2}$$

This invariant guarantees numerical consistency, prevents cross-device casting errors, and ensures reproducible execution across all supported GPU architectures. Implementation-wise, this mechanism is realized by explicitly applying `.to(device, dtype)` to all intermediate tensors before feature projection. As a result, inference remains deterministic and memory-stable, even for long-form videos.

B. Adaptive Importance-Based Token Sampling (MAS)

After uniform frame sampling for constructing the grid thumbnail, **all frames** are encoded into patch tokens $\{\mathbf{f}_j\}_{j=1}^{N}$ using the frozen CLIP-ViT-L/14 encoder, where $N = T \times P$ (T is the total number of frames and P is the number of patches per frame). To retain only the most informative tokens under the fixed context length of LLaVA, we apply an adaptive importance-based *global* selection strategy:

$$I_j = \|\mathbf{f}_j\|_2 + \beta\|\mathbf{f}_j - \bar{\mathbf{f}}\|_2, \tag{3}$$

where $\bar{\mathbf{f}} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{f}_j$ is the mean token representation and β controls the diversity weighting. We always include the first and last tokens for temporal coverage and sample the remaining tokens only from the mid-sequence. Let $\mathcal{M} = \{2,\ldots,N-1\}$ denote the set of mid-sequence tokens. Selection probabilities are computed over $\mathcal{M}$ as

$$P_j = \frac{\exp(I_j)}{\sum_{m\in\mathcal{M}}\exp(I_m)}, \quad j \in \mathcal{M}. \tag{4}$$

To break ties between equally scored tokens, a small Gaussian noise is added before the softmax. We then sample $Q_s - 2$ tokens from $\mathcal{M}$ by multinomial sampling according to $\{P_j\}_{j\in\mathcal{M}}$, such that the total number of selected tokens equals the token budget Q_s. This design enforces a fixed computational cost while prioritizing tokens with high feature

Algorithm 1: MAS-LLaVA Sampling Strategy (Implementation-Accurate)

Input: Video $V = \{I_t\}_{t=1}^{T}$, frame budget K, sampled-token budget Q_s
Output: Selected visual tokens $\mathcal{T}$ for LLaVA inference

if *I_t are stored on heterogeneous devices* **then**
  `// Device/dtype consistency`
  Align device and dtype for all tensors via `.to(device, dtype);`

`// Grid branch: build a thumbnail from uniformly sampled frames`
Uniformly sample K frames: $\mathcal{F} = \{F_1, \ldots, F_K\}$;
Form a thumbnail grid $G = \mathrm{Grid}(\mathcal{F})$;
Extract grid features: $\mathbf{G} = \mathrm{Encode}(G)$;

`// Token branch: encode all frames, then sample globally over mid-sequence`
Encode all frames into patch tokens $\{\mathbf{f}_j\}_{j=1}^{N}$ with $N = T \times P$;
Compute importance $I_j = \|\mathbf{f}_j\|_2 + \beta\|\mathbf{f}_j - \bar{\mathbf{f}}\|_2$ for all tokens;
Add small Gaussian noise to I_j to break ties;
Always include the first and last tokens: $\mathcal{T}_s \leftarrow \{\mathbf{f}_1, \mathbf{f}_N\}$;
Let $\mathcal{M} = \{2, \ldots, N-1\}$ be mid-sequence indices;
Compute probabilities on $\mathcal{M}$: $P_j = \exp(I_j) / \sum_{m \in \mathcal{M}} \exp(I_m)$ for $j \in \mathcal{M}$;
Sample $Q_s - 2$ tokens from $\mathcal{M}$ via multinomial using $\{P_j\}_{j \in \mathcal{M}}$;
$\mathcal{T}_s \leftarrow \mathcal{T}_s \cup \mathrm{SampledTokens}(\{\mathbf{f}_j\}_{j \in \mathcal{M}})$;

`// Concatenate sampled tokens with grid features`
Aggregate sequence: $\mathcal{T} = \mathrm{Concat}(\mathcal{T}_s, \mathbf{G})$;
Feed $\mathcal{T}$ to LLaVA's multimodal adapter;
return $\mathcal{T}$;

magnitude and diversity, improving representation density without any retraining.

C. Integration and Training-Free Inference

All modules operate purely at inference time, reusing the original visual encoder, projector, and language backbone. By coupling uniform frame sampling, device-aware consistency, and adaptive importance-based token selection, MAS-LLaVA achieves stable, hardware-independent inference and more balanced feature representation, enabling robust reasoning in long-form and heterogeneous video scenarios.

IV. Experimental Settings and Results

A. Experimental Setup and Datasets

All experiments are conducted on a single NVIDIA A100 GPU (80 GB) using CUDA 13.0 and PyTorch 2.3.1. We build upon the official Video-LLaVA [3] implementation as our base framework to ensure consistency and fairness in evaluation. To contextualize the performance of MAS-LLaVA, we summarize the reported results of representative training-free or lightweight video–language models, including MovieChat+ [29], Vista-LLaMA [30], DeepStack-L [31], M^3 [32], IG-VLM [25], and SF-LLaVA [27]. Evaluations follow standard protocols on two widely used video question–answering benchmarks: NExT-QA [33] and IntentQA [34].

NExT-QA comprises approximately 5.4K videos with 52K multiple-choice questions requiring causal, temporal, and commonsense reasoning across multi-event sequences. IntentQA contains roughly 3.2K videos and 22K question–answer pairs emphasizing human intentions, motivations, and goal-oriented interactions. Unless otherwise specified, we sample 50 frames uniformly per video, extract features using **CLIP-ViT-L/14**, and maintain a training-free evaluation pipeline—no model parameters are updated at inference time. MAS-LLaVA builds upon the official LLaVA-1.6 Vicuna checkpoint by inserting the proposed *Adaptive Importance-Based Token Sampling (MAS)* module into the existing thumbnail-and-token compression interface. This integration allows the language backbone to accept MAS-selected tokens (together with thumbnail features) directly, preserving architectural compatibility and enabling robust inference without any fine-tuning.

All experiments are conducted under deterministic inference settings (temperature = 0) to ensure reproducibility, and results are reported from single runs. Note that the adaptive token sampling in MAS introduces a minimal stochastic perturbation ($\sigma = 0.01$) to resolve tie cases, which is an intended part of the method design.

B. Quantitative Results

Table I and Figs. 2 and 3 summarize the quantitative performance on the IntentQA and NExT-QA benchmarks, respectively. MAS-LLaVA attains 64.9% on NExT-QA and 60.8% on IntentQA, outperforming recent training-free models (e.g., SF-LLaVA, IG-VLM) by up to 1.0%. Notably, MAS-LLaVA maintains competitive performance on both reasoning- and intent-oriented datasets while improving temporal sensitivity on action-centric clips. Compared with these methods, MAS-LLaVA consistently yields higher accuracy and smoother inference across GPUs of varying compute capability. These results confirm that the proposed device-aware sampling and adaptive token selection mechanisms complement each other effectively, enhancing both robustness and semantic completeness without sacrificing computational efficiency.

C. Scalability and Efficiency

We further evaluated MAS-LLaVA across different hardware configurations, including consumer-grade and data-center GPUs (e.g., RTX 3090, A100, and L40). MAS-LLaVA introduces minimal computational overhead and exhibits stable behavior under varying memory bandwidth and tensor allocation conditions due to its unified device-aware execution pipeline, which ensures consistent tensor device and dtype inheritance across all sampling operations.

TABLE I.
COMPARISON OF VIDEO-LLMS (7B) WITH DIFFERENT VISUAL ENCODERS, SORTED BY PERFORMANCE. **BOLD** NUMBERS INDICATE THE BEST RESULTS.

Method	LLM Size	Video-Encoder	NExT-QA	IntentQA
MovieChat+	7B	CLIP-G	54.8	–
DeepStack-L	7B	CLIP-L	61.0	–
Vista-LLaMA	7B	CLIP-G	60.7	–
Video-LLaVA	7B	ViT-L	60.5	–
M^3	7B	CLIP-L	63.1	58.8
IG-VLM	7B	CLIP-L	63.1	60.3
SF-LLaVA	7B	CLIP-L	64.2	60.1
MAS-LLaVA (Ours)	**7B**	**CLIP-L**	**64.9**	**60.8**

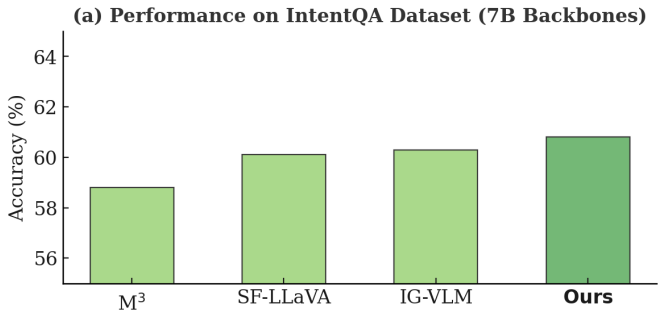

Fig. 2. Comparison of video–language models on the IntentQA benchmark using 7B backbones. The proposed **MAS-LLaVA** (Ours) achieves the best overall performance across all baselines.

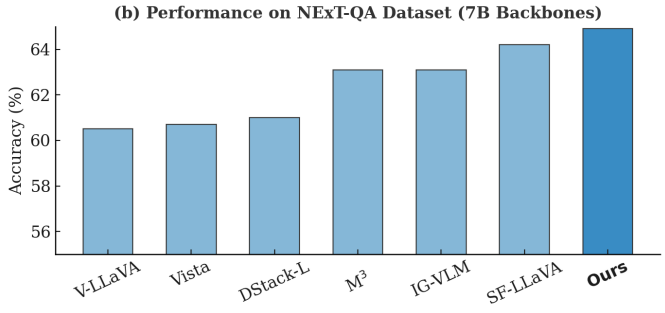

Fig. 3. Comparison of video–language models on the NExT-QA benchmark using 7B backbones. The proposed **MAS-LLaVA** (Ours) achieves the best overall performance across all baselines.

Importantly, MAS-LLaVA's adaptive token selection prioritizes patches with high feature magnitude and diversity, suppressing redundant or low-information tokens. This improves representation density under a fixed token budget while preserving architectural compatibility with the underlying LLaVA backbone. The MAS configurations provide a natural efficiency–accuracy trade-off in practice (e.g., lighter settings for speed, heavier settings for accuracy) without requiring any retraining. A comprehensive latency and scalability study across heterogeneous GPUs will be presented in future work to further validate deployment robustness.

V. CONCLUSION AND DISCUSSION

This work presented MAS-LLaVA, a training-free enhancement framework that improves the robustness and efficiency of video large language models (VLLMs) through device-aware consistency and adaptive importance-based token sampling. By stabilizing tensor allocation across heterogeneous GPUs and prioritizing importance- and diversity-salient tokens, MAS-LLaVA injects more informative visual representations into the thumbnail-and-token interface of existing VLLMs. Experiments on multiple video QA benchmarks demonstrate consistent gains on action-centric and dynamic videos while maintaining comparable performance on static or long-form scenarios—all without retraining. These results verify the practicality of inference-time adaptive sampling strategies for scalable, real-world video–language understanding.

While MAS-LLaVA achieves strong performance, several limitations remain. First, the sampling module currently relies primarily on low-level feature magnitude and diversity cues, which may be less effective for videos with subtle semantics or dialogue-driven content. Incorporating higher-level object, intent, or audio cues could yield a more holistic sampling policy. Second, MAS currently uses a fixed token and frame budget; dynamic scheduling based on video length, content complexity, or runtime constraints may further improve efficiency. Third, the gains are most pronounced on action-rich datasets such as NExT-QA and IntentQA; narrative-driven benchmarks may expose different strengths or weaknesses. MAS-LLaVA also does not yet leverage cross-modal guidance such as audio–visual alignment or language-coherent frame selection, which could enhance interpretability and temporal grounding. Finally, this study focuses on 7B-scale backbones; larger vision–language models may benefit even more from adaptive sampling.

Practical Impact. Beyond benchmark evaluations, MAS-LLaVA has practical implications for edge–cloud and embod-

ied AI applications where retraining or model modification is infeasible. Its inference-time adaptive sampling enables efficient deployment in scenarios such as real-time surveillance, autonomous driving, and robotic perception, where computational resources and hardware configurations vary widely. By maintaining device consistency and sampling adaptively according to visual importance, MAS-LLaVA can serve as a plug-and-play module to improve stability and efficiency across distributed or resource-constrained environments. This design makes it particularly suitable for deployment on heterogeneous systems, from embedded GPUs to cloud clusters, enabling scalable multimodal intelligence without additional training overhead.

Future work will explore integrating stronger vision backbones such as Swin Transformer and Vision Mamba, adaptive frame scheduling for edge–cloud deployment, and explainable importance selection through multi-scale temporal fusion. A broader evaluation on videos with low light, occlusion, or fast motion will further validate the robustness of MAS-LLaVA. Overall, MAS-LLaVA represents a lightweight yet effective step toward training-free, device-adaptive video–language understanding, paving the way for scalable and efficient multimodal systems.

REFERENCES

[1] J.-B. Alayrac *et al.*, "Flamingo: A visual language model for few-shot learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[2] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[3] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan, "Video-llava: Learning united visual representation by alignment before projection," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Miami, Florida, USA), pp. 5971–5984, Association for Computational Linguistics, 2024.

[4] T. Qu, M. Li, T. Tuytelaars, and M.-F. Moens, "Ts-llava: Constructing visual tokens through thumbnail-and-sampling for training-free video large language models," *arXiv preprint arXiv:2411.11066*, 2024.

[5] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "Videochat: Chat-centric video understanding," *Science China Information Sciences*, vol. 68, no. 10, p. 200102, 2025.

[6] OpenAI, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[7] C. Wei, Q. Liu, Z. Ye, Q. Wang, X. Wang, P. Wan, K. Gai, and W. Chen, "Univideo: Unified understanding, generation, and editing for videos," *arXiv preprint arXiv:2510.08377*, 2025.

[8] Z. Dong, Y. Gao, Q. Huang, J. Wawrzynek, H. K. H. So, and K. Keutzer, "Hao: Hardware-aware neural architecture optimization for efficient inference," *arXiv preprint arXiv:2104.12766*, 2021.

[9] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," *arXiv preprint arXiv:2305.06500*, 2023.

[10] Y. Rao, W. Zhao, B. Pan, J. Lu, and J. Zhou, "Dynamicvit: Efficient vision transformers with dynamic token sparsification," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[11] X. Ma, P. Sun, S. Luo, Q. Peng, R. F. DeMara, and Y. Bai, "Binarized l_1-regularization parameters enhanced stripe-wise optimization algorithm for efficient neural network optimization," *IEEE Journal of Emerging and Selected Topics in Industrial Electronics*, vol. 5, no. 2, pp. 790–799, 2024.

[12] M. Liu, M. Yin, K. Han, R. F. DeMara, B. Yuan, and Y. Bai, "Algorithm and hardware co-design co-optimization framework for lstm accelerator using quantized fully decomposed tensor train," *Internet of Things*, vol. 22, p. 100680, 2023.

[13] M. Liu, S. Luo, K. Han, R. F. DeMara, and Y. Bai, "Autonomous binarized focal loss enhanced model compression design using tensor train decomposition," *Micromachines*, vol. 13, no. 10, p. 1738, 2022.

[14] X. Ma, J. Tang, and Y. Bai, "Locality-sensing fast neural network (lfnn): An efficient neural network acceleration framework via locality sensing for real-time videos queries," in *2023 24th International Symposium on Quality Electronic Design (ISQED)*, pp. 1–8, 2023.

[15] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, 2023.

[16] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, *et al.*, "mplug-owl: Modularization empowers large language models with multimodality," *arXiv preprint arXiv:2304.14178*, 2023.

[17] J. Tang, N. Ma, C. Jia, R. Tian, and Y. Guo, "Hypereast: An enhanced attention-based spectral–spatial transformer with self-supervised pre-training for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 22241–22255, 2025.

[18] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems*, 2023.

[19] B. Li, K. Zhang, H. Zhang, D. Guo, R. Zhang, F. Li, Y. Zhang, Z. Liu, and C. Li, "Llava-next: Stronger llms supercharge multimodal capabilities in the wild." Blog post, LLaVA-VL project, 2024. Available at https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/.

[20] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[21] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 12585–12602, 2024.

[22] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553, 2023.

[23] Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, Y. Zhu, W. Zhang, Z. Luo, D. Zhao, and L. Bing, "Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms," *arXiv preprint arXiv:2406.07476*, 2024.

[24] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6202–6211, 2019.

[25] W. Kim, C. Choi, W. Lee, and W. Rhee, "An image grid can be worth a video: Zero-shot video question answering using a vlm," *arXiv preprint arXiv:2403.18406*, 2024.

[26] W. Wu, "Freeva: Offline mllm as training-free video assistant," *arXiv preprint arXiv:2405.07798*, 2024.

[27] M. Xu, M. Gao, Z. Gan, H.-Y. Chen, Z. Lai, H. Gang, K. Kang, and A. Dehghan, "Slowfast-llava: A strong training-free baseline for video large language models," *arXiv preprint arXiv:2407.15841*, 2024.

[28] F. Xiao, Y. J. Lee, K. Grauman, J. Malik, and C. Feichtenhofer, "Audiovisual slowfast networks for video recognition," *arXiv preprint arXiv:2001.08740*, 2020.

[29] E. Song, W. Chai, T. Ye, J. Hwang, X. Li, and G. Wang, "Moviechat+: Question-aware sparse memory for long video question answering," *arXiv preprint arXiv:2404.17176*, 2024.

[30] F. Ma, X. Jin, H. Wang, Y. Xian, J. Feng, and Y. Yang, "Vista-llama: Reducing hallucination in video language models via equal distance to visual tokens," *arXiv preprint arXiv:2312.08870*, 2023.

[31] L. Meng, J. Yang, R. Tian, X. Dai, Z. Wu, J. Gao, and Y.-G. Jiang, "Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms," *arXiv preprint arXiv:2406.04334*, 2024.

[32] M. Cai, J. Yang, J. Gao, and Y. J. Lee, "Matryoshka multimodal models," *arXiv preprint arXiv:2405.17430*, 2024.

[33] J. Xiao, X. Shang, A. Yao, and T.-S. Chua, "Next-qa: Next phase of question-answering to explaining temporal actions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9777–9786, June 2021.

[34] J. Li, P. Wei, W. Han, and L. Fan, "Intentqa: Context-aware video intent reasoning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11963–11974, 2023.