# RM-LLaVA: An Efficient Training-Free Video Understanding Framework with Redundancy-Minimized Frame Selection

**Chuqin Zhou**

Nanyang Technological University, Singapore
ZHOU0530@e.ntu.edu.sg

## Abstract

Vision-Language Models (VLMs) have made significant advances in image-language tasks, and many recent studies are now exploring their application to video understanding. However, video frames generate a large number of visual tokens, making efficient frame selection crucial. Some existing methods employ relatively simple frame sampling strategies, which often result in redundant frame selection and insufficient temporal coverage when handling complex and dynamic video content. To address this issue, we propose RM-LLaVA, a training-free video-language understanding framework based on Redundancy-Minimized Frame Selection (RMFS). RMFS consists of two stages: (i) structure-aware clustering to refine the candidate frame pool, and (ii) iterative semantic diversity maximization to select a compact and informative set of frames. The proposed method requires neither training nor fine-tuning, and can be seamlessly applied to off-the-shelf image-language models. We conduct extensive experiments on three standard VideoQA benchmarks, demonstrating that RM-LLaVA outperforms existing training-free approaches on two of them and surpasses many fine-tuned models. Ablation studies further verify the effectiveness and complementarity of the proposed components.

## Introduction

Recent advances in Vision-Language Models (VLMs) have opened new possibilities for video understanding without dedicated video-model training. By leveraging powerful pre-trained image-text models, one can attempt to solve video tasks by converting video data into a form digestible by image-based models. IG-VLM (Kim et al. 2024) is one such approach, which represents a video as a single image grid composed of a handful of frames. In IG-VLM, six frames uniformly sampled from the video are arranged in a grid and fed into a VLM to answer questions or describe the video. This simple training-free strategy achieved state-of-the-art zero-shot results on multiple VideoQA benchmarks, demonstrating the surprising power of treating video frames as "one big image."

Despite its success, uniform frame sampling in IG-VLM has clear limitations. Videos often contain redundant frames – adjacent frames may look nearly identical – and also brief but crucial events that a coarse uniform sampling might skip. With only six slots available, what frames we choose is critical. Uniform sampling does not account for frame content; it can over-sample long static scenes while missing brief but critical moments. Consequently, the six-frame grid might lack diversity, hurting the model's understanding of the video's full story. We hypothesize that a smarter frame selection strategy – one that avoids semantic redundancy and covers more distinct scenes – can significantly improve a grid-based VLM's performance without any model retraining.

In this paper, we introduce Redundancy-Minimized Frame Selection (RMFS), a frame sampling method designed to maximize the informational diversity of a fixed-size frame set. Our approach enhances frame selection by maximizing semantic diversity via a structured two-stage pipeline: structure-aware candidate refinement followed by iterative selection. This design ensures that frames are not only semantically distinct but also contextually complementary. In contrast to conventional uniform sampling—which often results in visually redundant or semantically irrelevant frames—RM-LLaVA actively selects frames that are both diverse and representative. Table 1 presents a comparison of representative vision-language models for video understanding, highlighting that our proposed RM-LLaVA uniquely achieves all three advantages: training-free, redundancy-aware, and capable of video understanding.

Our contributions are summarized as follows. First, we introduce a novel frame selection pipeline that reduces redundancy and enhances semantic coverage, allowing the model to better capture essential visual information. Second, we conduct extensive experiments on several standard VideoQA benchmarks, demonstrating the effectiveness of our proposed method. Third, we provide in-depth analyses and ablation studies to validate the contribution of each component within our framework. Lastly, the method is highly practical—it requires no model fine-tuning or external video metadata and adds minimal computational overhead, making it suitable for real-world deployment.

Table 1: Comparison of representative vision-language models for video understanding.

| Method | Training-free | Redundancy-aware | Video understanding |
|---|---|---|---|
| `Image-LLM` | ✓ | ✗ | ✗ |
| `Training-based Video LLM` | ✗ | ✓ | ✓ |
| `Training-free Video LLM` | ✓ | ✗ | ✓ |
| **`RM-LLaVA (proposed)`** | ✓ | ✓ | ✓ |

# Related Work

## Image Large Language Models

Image Large Language Models (Image LLMs) have rapidly evolved from early image-text alignment approaches (e.g., CLIP (Radford et al. 2021)) into powerful multimodal systems capable of supporting multilingual, multi-task, and interactive understanding. Models such as Flamingo (Alayrac et al. 2022), BLIP and BLIP-2(Li et al. 2022) (Li et al. 2023), CoCa (Yu et al. 2022), and PaLI (Chen et al. 2023) build efficient vision-language bridges through freezing strategies and large-scale pretraining. LLaVA(Liu et al. 2023) connects visual encoders with language models and fine-tunes on image-instruction data, enabling visual question answering and reasoning. GPT-4V (OpenAI et al. 2024), trained on large-scale multimodal data, demonstrates strong visual understanding across tasks such as image description, chart analysis, and meme interpretation. In the 3D domain, models like 3D-LLM (Hong et al. 2023), 4M(Mizrahi et al. 2023), Cube-LLM (Cho et al. 2024), and Grounded 3D-LLM (Chen et al. 2024b) further enhance spatial reasoning and multimodal referential grounding.

## Video Large Language Models

Video Large Language Models (Video LLMs) extend language models to temporal visual understanding by addressing challenges in sequence modeling and multimodal reasoning. Early works like MERLOT Reserve(Zellers et al. 2022) and InternVideo (Wang et al. 2022) demonstrated the value of large-scale video-text pretraining. Models such as Vid2Seq (Yang et al. 2023), VideoTree (Wang et al. 2025b), and VideoAgent (Wang et al. 2025a) explored hierarchical reasoning and agent-based control over long videos. Dialogue-based systems like VideoChat (Li et al. 2024a) and Video-ChatGPT (Maaz et al. 2024) integrate video encoders with LLMs for interactive understanding. Other advances include streaming processing (VideoLLM-Online (Chen et al. 2024a)) and thought-based reasoning (Video-of-Thought (Fei et al. 2024)). Notably, many models adapt image-based LLMs with minimal video-specific tuning, enabling strong zero-shot performance.

## Training-Free Video LLMs

Training-Free Video LLMs leverage pretrained Image LLMs without fine-tuning, enabling efficient video understanding through strategies like temporal pooling (FreeVA (Wu 2024)) and image grid representation (IG-VLM (Kim et al. 2024)). However, their reliance on a small number of frames and lack of explicit temporal modeling limits performance on complex videos. Recent advances such as SlowFast-LLaVA (Xu et al. 2024b) address this by introducing a dual-pathway design that separately captures spatial and temporal cues, enhancing temporal understanding while remaining training-free. Free Video-LLM (Han et al. 2024) achieves lightweight inference by leveraging prompt information and applying region-level cropping to reduce visual input.

# RM-LLaVA

## Preliminaries

Image Large Language Models (Image LLMs) extend pretrained large language models with vision capabilities by integrating an image encoder and a modality alignment module. These models are typically built by combining a powerful vision backbone (e.g., EVA(Fang et al. 2022), or SigLIP(Zhai et al. 2023)) with a frozen or lightly tuned LLM (e.g., LLaMA(Touvron et al. 2023), or GPT-like architectures). The core idea is to align visual embeddings with the LLM's language space, enabling the model to understand and generate language conditioned on image inputs.

Given an input image $I$, the visual encoder (denoted as $\mathrm{Visual}_{\mathrm{enc}}$) extracts visual tokens:

$$F_I = \mathrm{Visual}_{\mathrm{enc}}(I). \tag{1}$$

These tokens are then passed through a projector—typically a lightweight MLP or transformer-based adapter—to match the modality and dimensionality of the LLM's embedding space. The resulting aligned features are given by:

$$F_{\mathrm{proj}} = \mathrm{Projector}(F_I). \tag{2}$$

These features are then combined with a language prompt and fed into the LLM to perform vision-language tasks such as image captioning, visual question answering, or reasoning:

$$A = \mathrm{LLM}(\mathrm{Prompt}, F_{\mathrm{proj}}, Q), \tag{3}$$

where $Q$ is the user query or instruction, and $A$ is the model's output.

Recent works have proposed various architectures to improve image-text alignment and multimodal reasoning. For instance, BLIP-2(Li et al. 2023) introduces a Querying Transformer (Q-Former) to extract task-relevant features before projection. MiniGPT-4(Zhu et al. 2023) uses a linear layer to align CLIP features with Vicuna(Chiang et al. 2023), while LLaVA(Liu et al. 2023) leverages both pre-training and instruction tuning to achieve strong multi-turn VQA

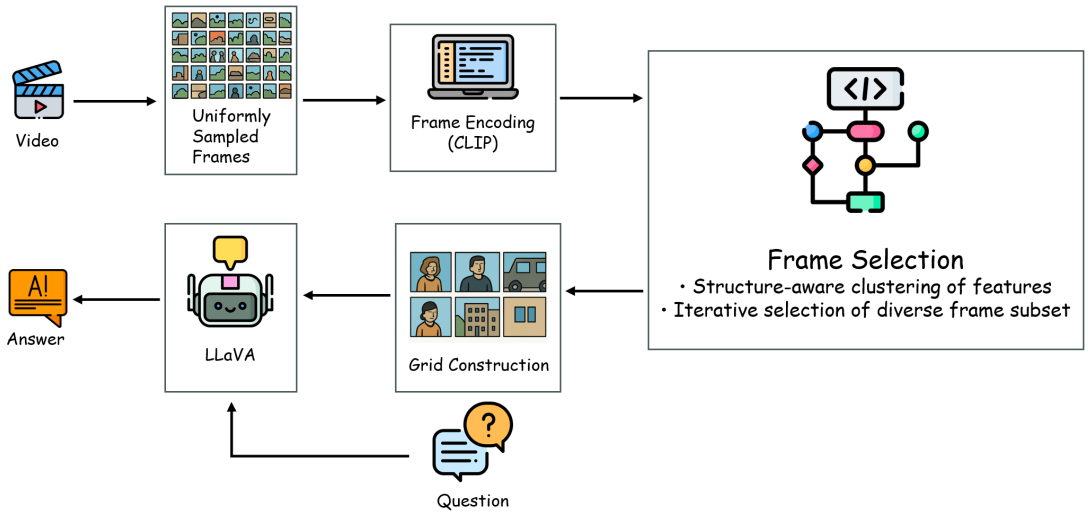



Figure 1: Overview of the training-free video understanding pipeline of RM-LLaVA, which incorporates a redundancy-minimized frame selection module.

capabilities. Despite architectural differences, most models follow a similar encoder–projector–LLM pipeline.

## RM-LLaVA Architecture

We propose RM-LLaVA, a training-free video language model framework that enhances semantic richness in frame selection through a two-stage process: clustering-based candidate refinement followed by iterative semantic diversity maximization. Unlike uniform frame sampling—which often introduces redundancy by repeatedly capturing similar frames or overlooks brief but critical events—RM-LLaVA explicitly prioritizes semantic diversity and informativeness. By filtering out visually redundant frames and selecting maximally distinct candidates, it enables more efficient and expressive use of the limited visual token budget, resulting in significantly enhanced temporal coverage and robustness in downstream tasks.

Conventional uniform sampling is oblivious to both visual content and task relevance. In relatively static scenes—such as a press conference where a speaker remains at the podium—uniformly sampled frames often capture nearly identical shots, resulting in redundant visual inputs and inefficient use of model capacity. In contrast, dynamic videos like cooking tutorials present the opposite challenge: uniform sampling tends to over-represent extended, repetitive segments such as chopping vegetables or stirring, while entirely missing brief but critical actions like pouring oil, adjusting heat, or plating the final dish—steps that may be essential for accurately answering a question or generating a meaningful caption.

To address these shortcomings, RM-LLaVA first performs clustering over the full frame pool to reduce coarse visual redundancy. It then incrementally selects a subset of semantically diverse and complementary frames. This two-stage process yields a compact yet informative frame set, better aligned with the underlying content structure and task requirements, thereby enhancing the model's ability to interpret and respond to diverse prompts. Figure 1 illustrates the architecture of RM-LLaVA.

**Frame Sampling and Feature Extraction** Given a video $V$, we uniformly sample $N$ candidate frames:

$$\mathcal{F} = \{f_1, f_2, \ldots, f_N\}, \tag{4}$$

Each frame $f_i \in \mathcal{F}$ is passed through a frozen image encoder $\phi(\cdot)$ to obtain L2-normalized visual features:

$$\mathbf{x}_i = \frac{\phi(f_i)}{\|\phi(f_i)\|_2}, \quad \mathbf{x}_i \in R^d. \tag{5}$$

Let $X = \{x_1, x_2, \ldots, x_N\}$ denote the set of normalized features.

**Redundancy-Minimized Frame Selection** To reduce redundancy and enhance information diversity within the visual input, we propose a redundancy-minimized frame selection strategy that operates through two stages: *structure-aware clustering* and *iterative diversity selection*. The goal is to identify a compact set of frames that are both semantically representative and mutually complementary, under a strict visual token budget.

We define the selection of the final frame subset $D^* \subseteq X$ of size $K$ as an optimization problem that jointly maximizes representativeness while minimizing redundancy:

$$D^* = \arg \max_{D \subseteq X | D| = K} \sum_{i \in D} \mathrm{Rep}(x_i) - \lambda \cdot \sum_{i,j \in D i \neq j} \mathrm{Sim}(x_i, x_j) \tag{6}$$

Here, $\mathrm{Rep}(x_i)$ denotes the semantic representativeness of frame $x_i$, which encourages coverage of distinct content regions in the video, and $\mathrm{Sim}(x_i, x_j)$ denotes the pairwise similarity between frames, discouraging redundancy. The hyperparameter $\lambda$ balances the trade-off between coverage and diversity.

While this optimization is intractable to solve exactly, we adopt a two-phase tractable instantiation. First, we perform structure-aware clustering over $X$ in the feature space to obtain a candidate pool $S = \{s_1, s_2, \ldots, s_C\}$ that reflects coarse semantic partitions of the video. Then, from $S$, we iteratively construct the final subset $D$ using a progressive diversity-aware selection strategy that approximates the above objective.

The process begins by selecting the globally most distinct candidate, i.e., the one with the lowest average similarity to all others:

$$D \leftarrow \left\{ \arg\min_{i \in S} \frac{1}{|S| - 1} \sum_{j \in S j \neq i} \mathrm{Sim}(s_i, s_j) \right\} \tag{7}$$

Then, we iteratively add new frames that exhibit maximal marginal gain in diversity:

$$s^* = \arg\min_{i \in S \setminus D} \frac{1}{|D|} \sum_{j \in D} \mathrm{Sim}(s_i, s_j), \quad D \leftarrow D \cup \{s^*\} \tag{8}$$

until $|D| = K$.

This unified selection strategy ensures that the final frame set is compact yet semantically rich, striking a balance between visual diversity and content relevance. It enables the downstream vision-language model to receive a more informative and non-redundant representation of the video, even under tight input constraints. A detailed investigation into the choice of the number of clusters $C$ and the final subset size $K$ is provided in our ablation studies, where we analyze their impact on downstream performance.

**Final fusion** The final selected frames $F' = \{f_1, f_2, \ldots, f_K\}$ are spatially arranged into a single image grid:

$$\mathcal{G} = \mathrm{GridConcat}(F') \tag{9}$$

This grid-format image $\mathcal{G}$ serves as the visual input to the vision-language model, paired with a prompt $p$ for inference:

$$y = \mathrm{VLM}(\mathcal{G}, p) \tag{10}$$

The use of an image grid preserves spatial coherence and enables the model to jointly attend to all selected frames within a single forward pass.

# Experiments

## Benchmarks and Backbone Model

In our experiments, we evaluate the effectiveness of the proposed RM-LLaVA framework on three widely used open-ended video question answering (VideoQA) benchmarks: MSVD-QA(Chen and Dolan 2011), MSRVTT-QA(Xu et al. 2016), and TGIF-QA(Jang et al. 2017). These datasets have been extensively used in prior work to assess the ability of models to understand video content and answer natural language questions. To ensure a fair and objective evaluation of model performance, we employ GPT-3.5-Turbo-0125 as an automatic evaluation agent, following standardized evaluation protocols commonly adopted in recent studies(Wu 2024)(Xu et al. 2024a). For each prediction, the agent provides both a binary correctness judgement and a 1–5 rating reflecting answer quality. We report two metrics: Acc (%), the percentage of answers judged correct, and Score, the average 1–5 quality rating assigned by GPT-3.5 over all samples. For the underlying vision-language backbone, our implementation builds upon LLaVA-v1.6 (7B)(Liu et al. 2023), using CLIP-L/14 as the visual feature extractor.

## Main Results

Table 2 presents a comprehensive comparison of RM-LLaVA against a series of state-of-the-art 7B-parameter video-language models on three standard VideoQA benchmarks. For fairness, we locally reproduced the results of several baseline methods using their released code, which may differ from the results reported in the original papers.

On MSVD-QA, RM-LLaVA achieves the highest accuracy of 78.7% and a score of 4.1, surpassing strong competitors such as IG-VLM (78.3%) and SF-LLaVA (78.2%). This result demonstrates RM-LLaVA's ability to capture fine-grained semantic cues in short, high-quality video clips.

On MSRVTT-QA, a dataset known for its greater temporal diversity and noisier web videos, RM-LLaVA attains 65.1% accuracy and 3.6 in score, outperforming all other models, including IG-VLM (63.4%) and SF-LLaVA (64.1%). The improvement indicates that RM-LLaVA's redundancy-minimized frame selection pipeline effectively enhances temporal coverage without sacrificing semantic quality.

Although TGIF-QA is a challenging benchmark focused on temporal reasoning and fine-grained actions, RM-LLaVA still achieves a strong 75.4% accuracy and 4.1 score. While it slightly trails SF-LLaVA (78.2%), it outperforms many other baselines, including IG-VLM (73.2%) and Video-LLaVA (70.0%). Notably, RM-LLaVA still shows a clear gain over IG-VLM, emphasizing the value of our redundancy-aware frame selection strategy even under dense temporal tasks.

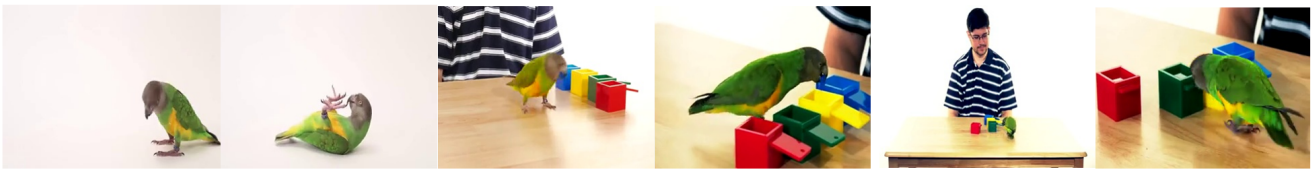
The superior performance of RM-LLaVA can be attributed to its redundancy-aware and training-free architecture. This ensures that selected frames are both semantically diverse and relevant, enabling the language model to focus on truly informative visual content while discarding redundant or uninformative frames. Consequently, RM-LLaVA makes more efficient use of the limited visual token budget, leading to better temporal coverage and stronger semantic alignment.

Furthermore, despite not requiring additional training, RM-LLaVA outperforms many training-based methods, demonstrating the power of task-aware input selection over brute-force fine-tuning. Here, our method's ability to capture key semantic moments without manual supervision or retraining proves particularly valuable. Figure 2 shows examples where RM-LLaVA answers a question based on the content of selected frames.

## Ablation studies

Table 2: Main results of the proposed method and comparison with other video LLMs. All models use 7B or comparable LLMs.

| Model | Training-Free | MSVD-QA | | MSRVTT-QA | | TGIF-QA | |
|---|---|---|---|---|---|---|---|
| | | Acc (%) | Score | Acc (%) | Score | Acc (%) | Score |
| Video-LLaMA (Zhang, Li, and Bing 2023) | ✗ | 51.6 | 2.5 | 29.6 | 1.8 | – | – |
| Video-ChatGPT (Maaz et al. 2024) | ✗ | 64.9 | 3.3 | 49.3 | 2.8 | 51.4 | 3.0 |
| Video-LLaVA (Lin et al. 2024) | ✗ | 70.7 | 3.9 | 59.2 | 3.5 | 70.0 | 4.0 |
| VideoChat (Li et al. 2024a) | ✗ | 56.3 | 2.8 | 45.0 | 2.5 | 34.4 | 2.3 |
| VideoChat2 (Li et al. 2024b) | ✗ | 70.0 | 3.9 | 54.1 | 3.3 | – | – |
| Vista-LLaMA (Ma et al. 2025) | ✗ | 65.3 | 3.6 | 60.5 | 3.3 | – | – |
| FreeVA (Wu 2024) | ✓ | 73.8 | 4.1 | 60.0 | 3.5 | – | – |
| IG-VLM (Kim et al. 2024) | ✓ | 78.3 | 4.1 | 63.4 | 3.5 | 73.2 | 4.0 |
| SF-LLaVA (Xu et al. 2024b) | ✓ | 78.2 | 4.1 | 64.1 | 3.4 | **78.2** | **4.2** |
| **RM-LLaVA (proposed)** | ✓ | **78.7** | **4.1** | **65.1** | **3.6** | 75.4 | 4.1 |



**Question 1:** What is walking on a table?
**RM-LLaVA:** In this video, a green parrot is walking on a table. The key frames show a person in the background observing the parrot's actions. The parrot appears to be exploring its surroundings and interacting with nearby objects.

**Question 2:** What does a green parrot then move as a seated man watches?
**RM-LLaVA:** In this video, a green parrot is seen interacting with a set of colorful blocks. Throughout the video, the parrot is the main character, and the man is observing the parrot's actions. The context of the video is a playful and engaging interaction between the parrot and the blocks, showcasing the parrot's curiosity and agility.

Figure 2: Examples of video-based question answering using RM-LLaVA.

Table 3: Ablation study on RM-LLaVA components on MSRVTT-QA. The Image-LLM used is LLaVA-v1.6-7B.

| Model Variant | MSRVTT-QA | |
|---|---|---|
| | Acc (%) | Score |
| **RM-LLaVA (full)** | **65.1** | **3.6** |
| w/o clustering | 64.5 | 3.5 |
| w/o iterative selection | 63.9 | 3.5 |

Table 4: Ablation study on selection strategy on MSRVTT-QA. The Image-LLM used is LLaVA-v1.6-7B.

| Selection Strategy | MSRVTT-QA | |
|---|---|---|
| | Acc (%) | Score |
| **Iterative** | **65.1** | **3.6** |
| Random | 60.4 | 3.3 |
| Edge Density | 62.8 | 3.4 |

**RM-LLaVA Components** To comprehensively evaluate the contribution of each module within RM-LLaVA, we conducted an ablation study on the MSRVTT-QA dataset. Table 3 presents the results of this analysis, focusing on the impact of the clustering-based candidate refinement and the iterative selection components.

When we remove the clustering stage and directly feed all uniformly sampled frames into the selection process, the accuracy drops to 64.5%. This decline highlights the importance of the clustering step in filtering out visually redundant or semantically similar frames. By grouping similar frames and selecting only a few representatives from each cluster, this stage significantly enhances the diversity and representativeness of the candidate pool. Without this step, the downstream model may receive inputs that are semantically repetitive, thereby reducing its ability to infer nuanced video se-

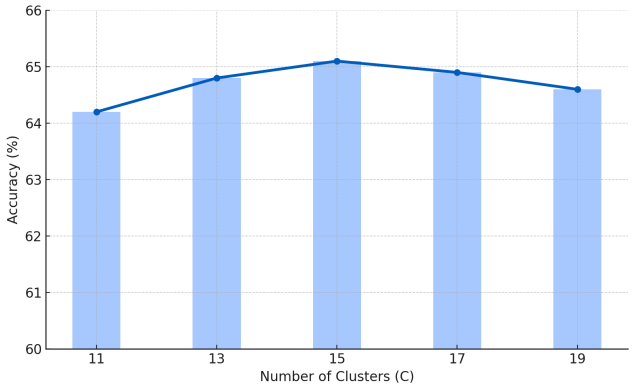



Figure 3: Effect of the number of clusters on MSRVTT-QA.

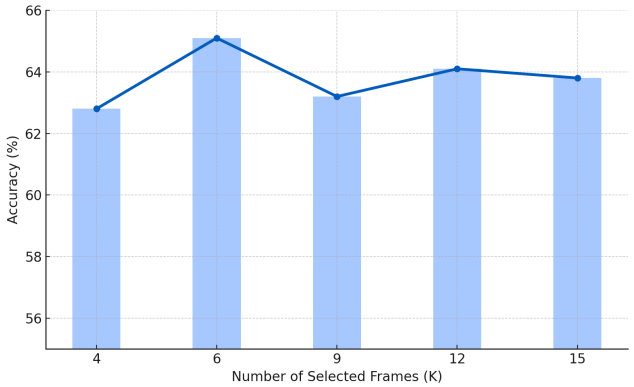



Figure 4: Effect of the number of selected frames on MSRVTT-QA.

mantics.

We further remove the iterative selection module, which is responsible for promoting inter-frame semantic diversity through progressive diversity-aware selection, resulting in an accuracy drop to 63.9%. This suggests that while clustering ensures semantic abstraction, the iterative selection step is essential for selecting frames that are not only representative but also complementary. It effectively avoids overrepresenting a single scene or event, ensuring temporal diversity and capturing a broader narrative context.

These findings confirm that both clustering and iterative selection are essential. Their synergy forms the core of RM-LLaVA's frame selection pipeline, enabling informative visual inputs under tight token limits.

**Selection Strategies** To further assess the effectiveness of our proposed iterative frame selection approach, we compare it against two widely used strategies: random selection and edge density selection, with results summarized in Table 4.

The random selection strategy uniformly samples frames from the video without any regard for content relevance or semantic distribution. As expected, this naive approach leads to the poorest performance, with an accuracy of 60.4% and a score of 3.3. The primary issue with random selection lies in its high likelihood of choosing visually redundant frames or missing semantically important moments, especially in videos with significant temporal or contextual variance.

The edge density selection strategy aims to improve upon random sampling by selecting frames with higher structural complexity. The edge density score is defined as the proportion of edge pixels within the frame. This method assumes that frames with more edge structures are more likely to contain informative visual content. Compared to random selection, it achieves better performance—62.8% accuracy and a score of 3.4—demonstrating its effectiveness in capturing visually rich frames. However, since it relies solely on low-level edge information and lacks high-level semantic understanding, it may still select redundant frames, especially when high-density edge patterns originate from visually similar contexts.

In contrast, our iterative frame selection method not only considers informativeness but also explicitly enforces semantic diversity among selected frames. As a result, it achieves the highest accuracy and score among all tested strategies.

This superior performance validates clearly our hypothesis: maximizing semantic complementarity, rather than solely informativeness or randomness, leads to more effective video understanding. The iterative frame selection approach enables the vision-language model to process a richer and more comprehensive representation of the video content, making it a key contributor to RM-LLaVA's success.

**Number of Clusters** To better understand the role of structure-aware clustering in our redundancy-minimized selection pipeline, we further investigate how the number of clusters $C$ used during candidate refinement affects downstream MSRVTT-QA performance. As shown in Figure 3, increasing $C$ from 11 to 15 consistently improves accuracy (64.2% $\rightarrow$ 65.1%). A moderate number of clusters provides an adequate partitioning of the feature space, allowing the refined candidate pool to capture diverse semantic regions of the video while effectively suppressing coarse visual redundancy.

When $C$ is increased beyond this point (e.g., 17 and 19), the accuracy slightly drops to 64.9% and 64.6%, respectively. We attribute this degradation to over-fragmentation: excessive clustering splits semantically coherent content into small, unstable partitions, reducing the representativeness of cluster centroids and weakening the subsequent diversity-aware selection stage.

**Number of Selected Frames** We further analyze how the number of selected frames $K$ affects accuracy on MSRVTT-QA. As shown in Figure 4, increasing $K$ from 4 to 6 yields a substantial improvement (62.8% $\rightarrow$ 65.1%), indicating that using too few frames under-represents the video and fails to capture essential events. The peak performance at $K = 6$ suggests that this setting achieves a good balance between temporal coverage and redundancy suppression.

As $K$ increases further to 9 and 12, the accuracy no longer improves and instead fluctuates within a lower range (63.2% and 64.1%), demonstrating diminishing returns. This indi-

cates that adding more frames beyond the essential subset introduces redundant or less informative visual content. When $K = 15$, accuracy slightly drops to 63.8% compared to $K = 12$, reinforcing that using too many frames introduces additional redundancy rather than helpful information.

## Conclusion

In this work, we presented RM-LLaVA, a training-free video understanding framework designed to overcome the limitations of uniform frame sampling in image-based vision-language models. We introduce a redundancy-minimized frame selection (RMFS) pipeline, consisting of two stages: structure-aware clustering and iterative semantic diversity selection. This approach effectively selects a compact yet semantically rich set of frames, enhancing temporal coverage and contextual diversity. Our experiments on three standard VideoQA benchmarks show that RM-LLaVA achieves state-of-the-art performance among all training-free methods on two of the benchmarks, while also outperforming most training-based models. Ablation studies further validate the complementary roles of clustering and iterative selection in improving model robustness and accuracy. Without requiring any model fine-tuning, RM-LLaVA provides a general, efficient, and scalable solution for video-language tasks. We believe this framework paves the way for developing more efficient, accessible, and interpretable video-language models in real-world applications.

## References


Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; et al. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 23716–23736. Curran Associates, Inc.

Chen, D.; and Dolan, W. 2011. Collecting Highly Parallel Data for Paraphrase Evaluation. In Lin, D.; Matsumoto, Y.; and Mihalcea, R., eds., *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 190–200. Portland, Oregon, USA: Association for Computational Linguistics.

Chen, J.; Lv, Z.; Wu, S.; Lin, K. Q.; Song, C.; Gao, D.; Liu, J.-W.; Gao, Z.; Mao, D.; and Shou, M. Z. 2024a. VideoLLM-online: Online Video Large Language Model for Streaming Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18407–18418.

Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A.; Padlewski, P.; et al. 2023. PaLI: A Jointly-Scaled Multilingual Language-Image Model. arXiv:2209.06794.

Chen, Y.; Yang, S.; Huang, H.; Wang, T.; Xu, R.; Lyu, R.; Lin, D.; and Pang, J. 2024b. Grounded 3D-LLM with Referent Tokens. arXiv:2405.10370.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Cho, J. H.; Ivanovic, B.; Cao, Y.; Schmerling, E.; Wang, Y.; Weng, X.; Li, B.; You, Y.; Krähenbühl, P.; Wang, Y.; and Pavone, M. 2024. Language-Image Models with 3D Understanding. arXiv:2405.03685.

Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2022. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. arXiv:2211.07636.

Fei, H.; Wu, S.; Ji, W.; Zhang, H.; Zhang, M.; Lee, M.-L.; and Hsu, W. 2024. Video-of-Thought: Step-by-Step Video Reasoning from Perception to Cognition. arXiv:2501.03230.

Han, K.; Guo, J.; Tang, Y.; He, W.; Wu, E.; and Wang, Y. 2024. Free Video-LLM: Prompt-guided Visual Perception for Efficient Training-free Video LLMs. arXiv:2410.10441.

Hong, Y.; Zhen, H.; Chen, P.; Zheng, S.; Du, Y.; Chen, Z.; and Gan, C. 2023. 3D-LLM: Injecting the 3D World into Large Language Models. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 20482–20494. Curran Associates, Inc.

Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; and Kim, G. 2017. TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering. arXiv:1704.04497.

Kim, W.; Choi, C.; Lee, W.; and Rhee, W. 2024. An Image Grid Can Be Worth a Video: Zero-Shot Video Question Answering Using a VLM. *IEEE Access*, 12: 193057–193075.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 19730–19742. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 12888–12900. PMLR.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2024a. VideoChat: Chat-Centric Video Understanding. arXiv:2305.06355.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; Wang, L.; and Qiao, Y. 2024b. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. arXiv:2311.17005.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2024. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. arXiv:2311.10122.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 34892–34916. Curran Associates, Inc.

Ma, F.; Jin, X.; Wang, H.; Xian, Y.; Feng, J.; and Yang, Y. 2025. Vista-LLaMA: Reducing Hallucination in Video Language Models via Equal Distance to Visual Tokens. arXiv:2312.08870.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. arXiv:2306.05424.

Mizrahi, D.; Bachmann, R.; Kar, O.; Yeo, T.; Gao, M.; Dehghan, A.; and Zamir, A. 2023. 4M: Massively Multimodal Masked Modeling. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 58363–58408. Curran Associates, Inc.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; et al. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Wang, X.; Zhang, Y.; Zohar, O.; and Yeung-Levy, S. 2025a. VideoAgent: Long-Form Video Understanding with Large Language Model as Agent. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Computer Vision – ECCV 2024*, 58–76. Cham: Springer Nature Switzerland. ISBN 978-3-031-72989-8.

Wang, Y.; Li, K.; Li, Y.; He, Y.; Huang, B.; Zhao, Z.; Zhang, H.; Xu, J.; Liu, Y.; Wang, Z.; Xing, S.; Chen, G.; Pan, J.; Yu, J.; Wang, Y.; Wang, L.; and Qiao, Y. 2022. InternVideo: General Video Foundation Models via Generative and Discriminative Learning. arXiv:2212.03191.

Wang, Z.; Yu, S.; Stengel-Eskin, E.; Yoon, J.; Cheng, F.; Bertasius, G.; and Bansal, M. 2025b. VideoTree: Adaptive Tree-based Video Representation for LLM Reasoning on Long Videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 3272–3283.

Wu, W. 2024. FreeVA: Offline MLLM as Training-Free Video Assistant. arXiv:2405.07798.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5288–5296.

Xu, L.; Zhao, Y.; Zhou, D.; Lin, Z.; Ng, S. K.; and Feng, J. 2024a. PLLaVA : Parameter-free LLaVA Extension from Images to Videos for Video Dense Captioning. arXiv:2404.16994.

Xu, M.; Gao, M.; Gan, Z.; Chen, H.-Y.; Lai, Z.; Gang, H.; Kang, K.; and Dehghan, A. 2024b. SlowFast-LLaVA: A Strong Training-Free Baseline for Video Large Language Models. arXiv:2407.15841.

Yang, A.; Nagrani, A.; Seo, P. H.; Miech, A.; Pont-Tuset, J.; Laptev, I.; Sivic, J.; and Schmid, C. 2023. Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10714–10726.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. CoCa: Contrastive Captioners are Image-Text Foundation Models. arXiv:2205.01917.

Zellers, R.; Lu, J.; Lu, X.; Yu, Y.; Zhao, Y.; Salehi, M.; Kusupati, A.; Hessel, J.; Farhadi, A.; and Choi, Y. 2022. MERLOT Reserve: Neural Script Knowledge Through Vision and Language and Sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16375–16387.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid Loss for Language Image Pre-Training. arXiv:2303.15343.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In Feng, Y.; and Lefever, E., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 543–553. Singapore: Association for Computational Linguistics.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv:2304.10592.