Towards an Efficient and Unified Strategy for Video Understanding Applications

Rahul Kumar
Dept of Electrical Engineering
Indian Institute of Technology, Hyderabad
Hyderabad, India
ee24mtech11020@iith.ac.in

Sumohana Channappayya
Dept of Electrical Engineering
Indian Institute of Technology, Hyderabad
Hyderabad, India
sumohana@ee.iith.ac.in

***Abstract*—Understanding long-range videos remains a key challenge in computer vision due to high temporal redundancy and computational burden. Despite strong performance of recent models, they are constrained in terms of scalability and generalization when applied to longer video sequences. In this work, we present Keyframe-based Spatio-Temporal Adaptive Representation (K-STAR), a redundancy-aware video summarization framework designed to generate compact and semantically rich representations that are effective in downstream tasks. The proposed method jointly models appearance and motion cues while filtering redundant frames. Importantly, it preserves critical temporal transitions while significantly reducing the number of processed frames. Additionally, each key frame is encoded using object, scene, and background-aware prompts, enabling richer semantic representation. Evaluated on the UCF-101 dataset, K-STAR achieves Top-1 accuracy of 93.06% and Top-5 accuracy of 98.73%, with 56× frame reduction and 11.5× faster inference, demonstrating competitive performance with substantially improved efficiency.**

***Index Terms*—Video Summarization, Action Recognition, Key Frame Extraction, Vision-Language Models**

I. Introduction

The rapid expansion of large-scale video collections and online video platforms has driven the demand for efficient video understanding systems. Unlike images, videos are long sequences of frames in which meaningful events unfold over time, requiring models to capture spatio-temporal representations for tasks such as action recognition and video retrieval. Yet, processing long videos remains computationally expensive due to the large number of frames involved. A key challenge is the high temporal redundancy in real-world videos—adjacent frames often differ only slightly and add limited new semantic information [1], [2]. As a result, processing is computationally and memory-intensive.

Previous approaches compensate for this with the help of a coarse temporal sampling strategy or through architecture-level trade-offs [3]. Other works focus on compressing video into a compact set of informative frames, which complement tasks such as video summarization and key-frame extraction [4], [5]. Classical approaches, such as histogram and feature clustering along with shot boundary detections, are also computationally light, but fail to capture high-level semantics due to their operation in frame-level pixel spaces [6]. The Deep Learning approach also aids in the process by improving semantic fidelity, but mainly relies on the labeled dataset, which is manually expensive. Further, these methods also suffer from scaling issues for long-term video representation [7]. In the present modern era of Vision–Language Models (VLMs) and Multimodal Large Language Models (MLLMs), these have demonstrated superior semantic reasoning techniques and better generalization abilities. [8]–[10]. However, using VLMs densely across long sequences of image frames makes it impractical. This is because the per-frame visual encoding and the language-conditioned reasoning scales linearly, which massively impacts inference time on long videos [2], [8]. This necessitates a hybrid strategy of a lightweight model, with a pre-processing redundancy-aware filtering approach that preserves the temporal coherence of the compact and informative keyframes. In this work, we propose K-STAR as a redundancy-aware video understanding framework that integrates motion-driven keyframe selection with structured semantic prompting to construct compact and semantically rich representations. The resulting keyframe set forms a compact representation of the original sequence and is processed sequentially using the VLM. The proposed approach first performs keyframe selection using a unified scoring mechanism, enabling the identification of informative and non-redundant frames. Building on the selected keyframes, structured semantic prompts are used to extract scene-specific semantics and action descriptors from each frame. A temporal summary from earlier frames is propagated to enable progressive reasoning across the sequence, enabling an efficient yet semantically rich video understanding. The proposed model is evaluated on the UCF-101 [11] dataset that reports Top-1 and Top-5 accuracy. The results show that K-STAR achieves competitive performance while reducing the number of processed frames by up to 56× and improving the inference efficiency by 11.5×, highlighting a favourable trade-off between accuracy and computation cost. The following are the main contributions of this work.

1) *K-STAR framework.* A redundancy-aware video understanding framework that integrates keyframe selection with semantic prompting for efficient long video processing.
2) *Motion-aware keyframe selection.* A two-stage frame selection strategy combining temporal gradient-based

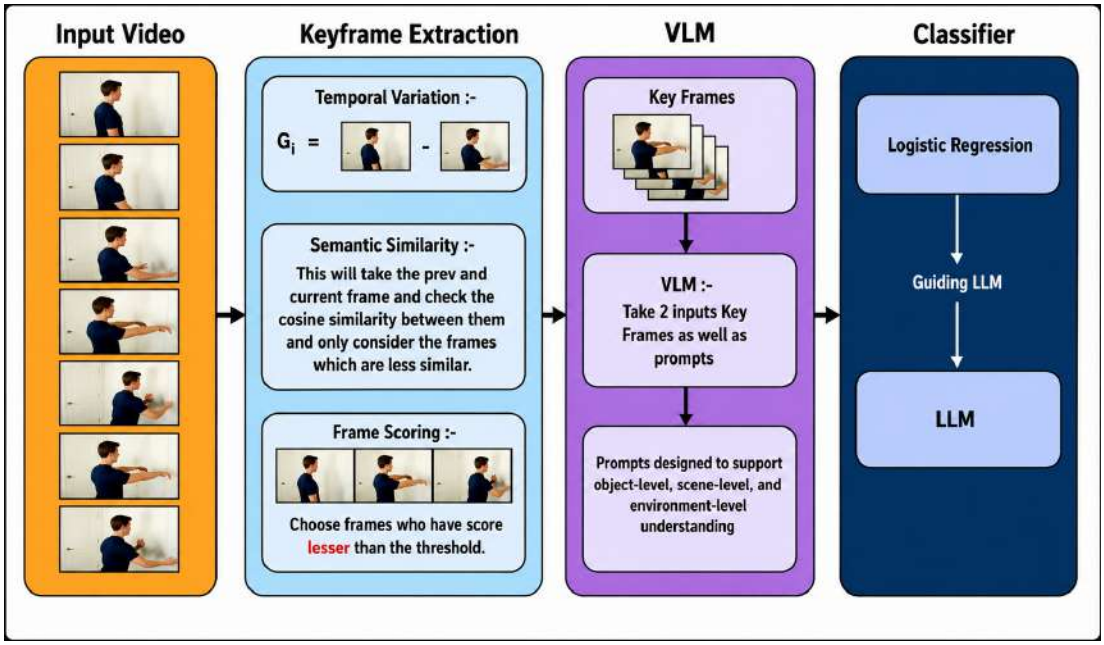

Fig. 1. Overview of the proposed K-STAR framework. The pipeline performs motion-aware keyframe selection to remove redundancy , followed by structured semantic prompting and sequential VLM-based reasoning to generate compact and semantically rich video representations.

motion estimation with feature-space cosine similarity to identify informative and non-redundant frames.

3) *Structured semantic prompting.* A prompting mechanism that captures object-level, scene-level, and background context for enhanced semantic representation.
4) *Efficient sequential reasoning.* A sequential VLM-based inference pipeline that propagates temporal summaries across frames for progressive video understanding.
5) *Efficiency-performance trade-off.* Demonstration of competitive accuracy with significantly reduced frame processing and inference cost.

The overall pipeline and the strategies used in the work is shown in Figure 1.

II. Related Work

A. Video Summarization and Keyframe Selection

The concept of video summarization takes a long-range video as input, produces a compact representation, and compresses it to a subset of frames preserving essential semantic and temporal content. The naive methods rely on empirical approaches such as shot boundary detection, where histogram differences are used to identify probable keyframe [12], [13]. In deep learning settings, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs) were commonly used to capture temporal dependencies [14], [15]. These models learn frame-level importance scores by modeling the sequential relationships between frames. Further works incorporated attention mechanisms to improve the temporal modeling which helped to selectively focus on salient temporal segments under reduced supervision [16].

Further, the Graph Neural Networks (GNNs) approaches were also explored for modeling the inter-frame relationships. One such work, SUM-GAN [17] and SumGraph [18] used this approach to capture the long-range dependencies across frames. The integration of transformers here further aided in enabling the overall reasoning across the video sequences [19].

More recent approaches incorporate multi-modal information to guide summarization. The Query-based video summarization [20], [21] introduces textual queries to steer the selection of relevant frames. In addition, contrastive-based VLMs such as CLIP [22] have been used to measure semantic similarity between frames and text descriptions for both zero-shot summarization and keyframe extraction.

Our work differs from these approaches by combining motion-aware keyframe extraction with embedding-space redundancy suppression and sequential multimodal reasoning, enabling efficient summarization while preserving semantic continuity.

B. Efficient Video Understanding

Further analysis in the field of video understanding comes at the cost of computational complexity when processing longer sequences. The 3D convolution network-based approaches compensate for this issue through temporal sampling as well as by hierarchical feature aggregation. Some works adopts dual-stream architecture, which helps in separating both the spatial and motion streams [3], [23].

Subsequent works focus on architectural strategies for video modeling. TimeSformer [24] splits the attention into separate spatial and temporal stages, which substantially reduces computational. On the other hand, Video Swin Transformer [25] incorporates the concept of local window and performs attention across these frames. Further memory-based architectures such

as MemViT [26] improve scalability by maintaining temporal memory representations.

Techniques such as Token compressions were also proposed to reduce computational cost. Token Merging (ToMe) [27] progressively merges similar tokens at the time of inference, which enables efficient processing. On a different note, MViTv2 [28] incorporated a pooled attention mechanism to reduce token counts in deeper layers.

Despite these approaches improving computational efficiency, they typically operate on dense frame sequences. In contrast, our method focuses on reducing the computational overhead by explicitly removing redundant frames and significantly reducing the number of processed tokens by downstream models.

C. Video-based Vision-Language Models

The understanding of tasks can be leveraged in a multimodal setup with the recently introduced VLMs. VideoBERT [29] and UniViLM [30] learn the joint representations of video and text through masked token prediction objectives.

Inspired by the contrastive setup of CLIP [22], CLIP4Clip and X-CLIP [31] extend this to the video setup by aggregating at the frame-level embedding and computing the similarities between them. Similarly, InternVideo [32] introduces a unified masking approach to leverage the multimodal representations.

The Video-LLMs, with the inspiration of the Large Language Models(LLMs), integrate the video-based encoders with the LLM-based decoders. One of the common models, Flamingo [33], helps in arranging the visual features with language tokens with the help of a gated cross-attention layer. Further works like VideoChat [9] and Video-LLaMA [34] extend the above approach to video understanding tasks. VideoChatGPT [10] and Video-LLaVA, on the other hand, aim at better conversation capabilities.

Despite sigificant progress in this field, the scaling issue with long videos remains a constant challenge due to the presence of large visual tokens. To address this, methods such as LLaMA-VID [34], MovieChat [35], TimeChat [36] address it using uniform token compression mechanisms. But, in a real-world scenario, these tokens might not be uniformly distributed, which can further lead to extracting redundant tokens or the loss of the informative tokens.

We introduce a redundancy-aware frame selection strategy here that can handle the above issue prior to multimodal reasoning. Thereby, combining this with motion-based keyframe extraction along with a semantic similarity filtering approach, our framework provides not only a compact but also an informative frame sequence for VLMs.

III. Proposed Method

We propose the framework, **K-STAR**, which consists of three main components, namely *motion-aware keyframe selection*, *structured semantic prompting*, and *sequential vision-language reasoning*, respectively.

A. Overview

Given an input video $V = \{f_1, f_2, \ldots, f_T\}$ with T frames, the objective is to construct a compact representation that preserves both semantic and temporal information while removing redundant frames.

The overall formulation is expressed as:

$$\mathbb{H}(V) = \mathbb{R}\big(\mathbb{P}(\mathbb{M}(V))\big) \tag{1}$$

where $\mathbb{M}$ denotes motion-based frame selection, $\mathbb{P}$ represents structured semantic prompting, and $\mathbb{R}$ corresponds to sequential multimodal reasoning. $\mathbb{H}(V)$ denotes the final semantic-temporal representation of the video.

B. Motion-Aware Keyframe Selection

The keyframe selection module aims to identify informative and non-redundant frames while preserving temporal dynamics. This is achieved through a unified scoring mechanism that combines temporal variation and semantic similarity.

a) Temporal Variation: The temporal variation between consecutive frames is computed as :

$$G_i = \|\nabla f_{i+1} - \nabla f_i\| \tag{2}$$

where ∇f_i denotes the spatial gradient of frame f_i.

b) Semantic Similarity: To capture redundancy in feature space, cosine similarity is computed between frame embeddings:

$$S_i = \frac{\phi(f_i) \cdot \phi(f_{i+1})}{\|\phi(f_i)\| \, \|\phi(f_{i+1})\|} \tag{3}$$

where $\phi(\cdot)$ denotes the visual embedding function.

c) Frame Scoring: The final importance score is defined as:

$$\text{Score}_i = \lambda G_i + (1 - \lambda)(1 - S_i) \tag{4}$$

where $\lambda \in [0, 1]$ balances motion sensitivity and semantic redundancy. The frames are selected as keyframes if $\text{Score}_i < \delta$, where δ is a predefined threshold.

C. Structured Semantic Prompting

Given the selected keyframes $K = \{k_1, k_2, \ldots, k_M\}$, where $\mathbb{M} \ll T$, each frame is transformed into a structured semantic representation using a prompting function:

$$p_t = \mathbb{P}(k_t) \tag{5}$$

The prompting mechanism encodes each frame with respect to object-level, scene-level, and background context, yielding richer, more interpretable representations than raw visual features.

D. Sequential Vision-Language Reasoning

To preserve temporal context across the selected keyframes, we employ a sequential reasoning strategy using a Vision-Language Model (VLM). For each keyframe k_t, the generated semantic description from the previous step is incorporated into the prompt of the current frame.

The reasoning state is updated as

$$s_t = \mathcal{R}(k_t, p_t, s_{t-1}) \tag{6}$$

where p_t denotes the structured semantic prompt associated with keyframe k_t and s_{t-1} represents the accumulated textual summary generated from preceding keyframes.

Unlike recurrent architectures, the VLM itself is not stateful. Temporal continuity is maintained by propagating the generated summary as textual context to subsequent reasoning steps. The final representation s_M is used for video-level understanding.

E. Classifier-Guided Semantic Prior

To constrain the reasoning space and reduce ambiguity, we introduce a classifier-guided prior. Given the input video V, a lightweight classifier predicts a set of top-K candidate classes:

$$\mathcal{C}_{topK} = \mathbb{C}(V) \tag{7}$$

This candidate set is incorporated into the prompting stage:

$$p_t = \mathbb{P}(k_t, \mathcal{C}_{topK}) \tag{8}$$

This guidance enables the reasoning module to focus on semantically plausible actions, improving both accuracy and robustness.

IV. Experiments

A. Experimental Setup

a) Datasets: The proposed method is evaluated on the standard action recognition benchmark **UCF-101**, which consists of 101 action classes with various motion patterns and scene variations.

b) Evaluation Metrics: Performance is measured using the accuracy of the Top-1 and Top-5 classifications. In addition, efficiency is evaluated using the frame reduction ratio and inference speedup. The frame reduction ratio is defined as:

$$\text{Reduction Ratio} = \frac{T}{M} \tag{9}$$

where T and M denote the total number of frames and the number of selected frames, respectively.

c) Implementation Details: Visual embeddings are extracted using a pre-trained vision encoder, and cosine similarity is computed in the embedding space. Hyperparameters were empirically selected on a validation subset. We use $\lambda = 0.7$ to equally balance motion and semantic redundancy, and $\delta = 0.93$ as the frame selection threshold. The sequential reasoning module is implemented using Qwen2-VL-2B-Instruct. All experiments were conducted on a single NVIDIA TITAN Xp GPU with 12GB memory.UniFormer is pretrained on Kinetics-400 and operates on dense full-video inputs. K-STAR employs pretrained visual encoders and a pretrained Vision-Language Model. Only the lightweight classifier used for generating the classifier-guided semantic prior is trained on UCF-101 and processes significantly fewer frames, highlighting its efficiency.

B. Comparison with State-of-the-Art

We compare K-STAR with representative baselines, including: (i) uniform frame sampling, (ii) dense processing, and (iii) keyframe-based summarization methods.

TABLE I
COMPARISON OF VARIOUS MODELS ON UCF-101.

Method	Top-1 (%)	Top-5 (%)	Frame Reduction	Speedup
Pali Gemma-3B	62.9	96.3	–	–
Meat Chameleon 7B	78.1	97.3	–	–
Uniform Sampling	89.01	92.7	14×	6.6×
Keyframe Baseline	90.5	93.8	24×	8.9×
UniFormer-S†	98.3	–	1×	–
K-STAR (Ours)	**93.06**	**98.73**	**56**×	**11.5**×

C. Efficiency Analysis

We evaluate the efficiency of the proposed method in terms of frame reduction and inference speed.

TABLE II
EFFICIENCY COMPARISON OF K-STAR.

Method	Frames Processed	Reduction (×)	Speedup (×)
Dense Processing	167	1×	1×
Uniform Sampling	12	14×	6.6×
K-STAR (Ours)	**3**	**56**×	**11.5**×

The results demonstrate that K-STAR reduces the number of processed frames by up to **56**× while achieving **11.5**× faster inference compared to dense processing. This highlights the effectiveness of the proposed keyframe selection strategy in minimizing redundancy while preserving critical temporal information.

D. Semantic Prompting

We qualitatively evaluate the proposed semantic prompting by conditioning the VLM on a fixed set of decomposed queries $\mathcal{P} = \{Q_i\}_{i=1}^{5}$ for each keyframe k_t. The model produces structured outputs $\mathcal{A} = \{A_i\}_{i=1}^{5}$, where each response captures a distinct semantic component, including action, motion, object interaction, discriminative cues, and scene context.

This explicit decomposition encourages multi-faceted reasoning and yields consistent and interpretable descriptions across keyframes. As shown in Fig. 2, the responses align with complementary visual cues, demonstrating the effectiveness of structured prompting in capturing fine-grained semantics.

E. Limitations

Despite its effectiveness, several limitations remain. The motion-based scoring mechanism can be sensitive to noise and abrupt visual changes, which may affect keyframe selection. The framework also depends on the capability of the underlying Vision-Language Model and may miss subtle temporal cues important for fine-grained understanding.

Additionally, the current sequential design limits parallelization and may increase latency. Finally, the method relies

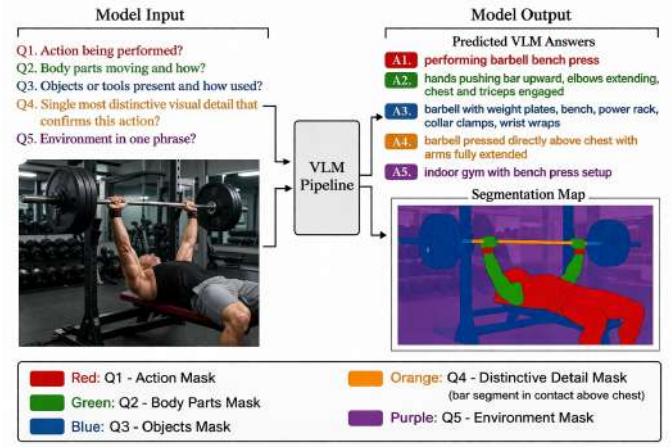

Fig. 2. Semantic prompting pipeline. Given an input image, a structured set of queries is passed to the VLM, producing decomposed semantic responses corresponding to action, motion, object interaction, discriminative cues, and scene context.

solely on visual and textual modalities, without leveraging complementary information such as audio, which could further enhance performance.

V. Conclusion & Future Work

We introduced **K-STAR**, a redundancy-aware video understanding framework that improves efficiency by filtering redundant frames before semantic reasoning. By combining motion-aware keyframe selection with embedding similarity, along with structured semantic prompting and sequential vision-language reasoning, the approach generates compact yet semantically rich video representations. Experimental results demonstrate competitive performance with significantly reduced frame processing and improved inference efficiency.

Future work will focus on addressing the identified limitations: this includes improving robustness through adaptive thresholding, incorporating multimodal signals such as audio, and designing more parallelizable architectures to enhance scalability and real-time performance.

References

[1] Y. X. Limin Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," 2016. [Online]. Available: https://arxiv.org/abs/1608.00859

[2] Y. S. Zhan Tong *et al.*, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," 2022. [Online]. Available: https://arxiv.org/abs/2203.12602

[3] C. Feichtenhofer *et al.*, "Slowfast networks for video recognition," 2019. [Online]. Available: https://arxiv.org/abs/1812.03982

[4] T. Liu and J. R. Kender, "Computational approaches to temporal sampling of video sequences," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 3, no. 2, p. 7–es, May 2007. [Online]. Available: https://doi.org/10.1145/1230812.1230813

[5] H. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Systems*, vol. 1, pp. 10–28, 1993. [Online]. Available: https://api.semanticscholar.org/CorpusID:28791109

[6] M. K. Zakaryan, "Video shot detection using histogram comparison," *MPCS*, vol. 44, pp. 77–84, 2021.

[7] C. F. Chao-Yuan Wu *et al.*, "Long-Term Feature Banks for Detailed Video Understanding," in *CVPR*, 2019.

[8] J. W. K. Alec Radford *et al.*, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[9] Y. H. KunChang Li *et al.*, "Videochat: Chat-centric video understanding," 2024. [Online]. Available: https://arxiv.org/abs/2305.06355

[10] B. Lin *et al.*, "Video-llava: Learning united visual representation by alignment before projection," *arXiv preprint arXiv:2311.10122*, 2023.

[11] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," 2012. [Online]. Available: https://arxiv.org/abs/1212.0402

[12] Mundur *et al.*, "Keyframe-based video summarization using delaunay clustering," vol. 6, no. 2. Springer, 2006, pp. 219–232.

[13] de Avila *et al.*, "Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 56–68, 2011.

[14] Zhang *et al.*, "Video summarization with long short-term memory," in *ECCV*, 2016, pp. 766–782.

[15] Zhao *et al.*, "Hierarchical recurrent neural network for skeleton based action recognition," in *CVPR*, 2017, pp. 1110–1119.

[16] J. Fajtl *et al.*, "Summarizing videos with attention," 2019. [Online]. Available: https://arxiv.org/abs/1812.01969

[17] Mahasseni *et al.*, "Unsupervised video summarization with adversarial lstm networks," in *CVPR*, 2017, pp. 202–211.

[18] J. Park *et al.*, "Sumgraph: Video summarization via recursive graph model," in *ECCV*, 2020, pp. 647–663.

[19] Z. Ji *et al.*, "Video summarization with attention-based encoder-decoder networks," *IEEE TCSVT*, vol. 30, no. 6, pp. 1709–1717, 2020.

[20] M. Narasimhan *et al.*, "Clip-it! language-guided video summarization," in *NeurIPS*, vol. 34, 2021, pp. 13 988–14 000.

[21] W. Moon *et al.*, "Query-dependent video representation for moment retrieval and highlight detection," in *CVPR*, 2023, pp. 23 023–23 033.

[22] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[23] K. Simonyan *et al.*, "Two-stream convolutional networks for action recognition in videos," in *NeurIPS*, 2014.

[24] G. Bertasius *et al.*, "Is space-time attention all you need for video understanding?" in *ICML*, vol. 139, 2021, pp. 813–824.

[25] Z. Liu *et al.*, "Video swin transformer," in *CVPR*, 2022, pp. 3202–3211.

[26] C.-Y. Wu *et al.*, "Memvit: Memory-augmented multiscale vision transformer," in *CVPR*, 2022, pp. 13 587–13 597.

[27] D. Bolya *et al.*, "Token merging: Your vit but faster," in *ICLR*, 2023.

[28] Y. Li *et al.*, "Mvitv2: Improved multiscale vision transformers," in *CVPR*, 2022, pp. 4804–4814.

[29] A. M. Chen Sun *et al.*, "Videobert: A joint model for video and language representation learning," 2019. [Online]. Available: https://arxiv.org/abs/1904.01766

[30] L. J. Huaishao Luo *et al.*, "Univl: A unified video and language pre-training model for multimodal understanding and generation," 2020. [Online]. Available: https://arxiv.org/abs/2002.06353

[31] G. X. Yiwei Ma *et al.*, "X-clip: End-to-end multi-grained contrastive learning for video-text retrieval," 2022. [Online]. Available: https://arxiv.org/abs/2207.07285

[32] K. L. Yi Wang *et al.*, "Internvideo: General video foundation models via generative and discriminative learning," 2022. [Online]. Available: https://arxiv.org/abs/2212.03191

[33] J. D. Jean-Baptiste Alayrac *et al.*, "Flamingo: a visual language model for few-shot learning," 2022. [Online]. Available: https://arxiv.org/abs/2204.14198

[34] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," 2023. [Online]. Available: https://arxiv.org/abs/2306.02858

[35] E. Song *et al.*, "Moviechat: From dense token to sparse memory for long video understanding," *arXiv preprint arXiv:2307.16449*, 2023.

[36] L. Y. Shuhuai Ren *et al.*, "Timechat: A time-sensitive multimodal large language model for long video understanding," *ArXiv*, vol. abs/2312.02051, 2023.