

Flexible Frame Selection for Efficient Video Reasoning

Shyamal Buch Arsha Nagrani Anurag Arnab Cordelia Schmid
 Google DeepMind

Abstract

Video-language models have shown promise for addressing a range of multimodal tasks for video reasoning, such as video question-answering. However, the inherent computational challenges of processing long video data and increasing model sizes have led to standard approaches that are limited by the number of frames they can process. In this work, we propose the Flexible Frame Selector (FFS), a learnable policy model with a new flexible selection operation, that helps alleviate input context restrictions by enabling video-language models to focus on the most informative frames for the downstream multimodal task, without adding undue processing cost. Our method differentiates from prior work due to its learnability, efficiency, and flexibility. We verify the efficacy of our method on standard video-question answering and reasoning benchmarks, and observe our model can maintain or improve base video-language model accuracy while significantly reducing the number of downstream processed frames.

1. Introduction

Reasoning about events to answer questions about videos is a long-standing challenge for multimodal AI systems. In the image domain, vision-language models (VLMs) have shown great promise for multimodal reasoning tasks [32, 42, 44, 58, 60]. However, transferring this success from large image-language models [9, 31, 32, 55] to the video-language setting [9, 39, 46, 72, 85] has proven challenging, largely due to the computational complexities arising simultaneously from (1) increasing model sizes and (2) the volume of video data that must be processed for a single pass.

Thus, current state-of-the-art approaches for video-language understanding are often constrained to using relatively few frames from the original video data, usually obtained through uniform downsampling [2, 8, 9, 12]. This approach, while straightforward, often ignores the nuances of the fundamental nature of videos. In particular, videos include highly correlated input frames and depict many

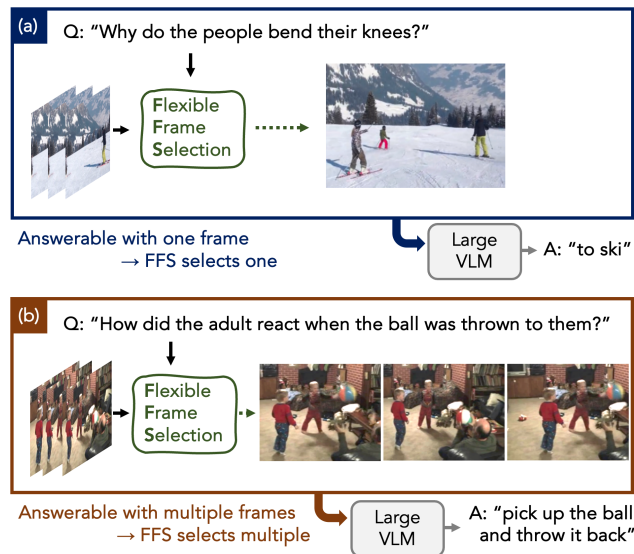


Figure 1. **Adaptive frame selection for VideoQA.** Video question-answering (VideoQA) is a key task for long video understanding, encompassing causal, temporal, descriptive event reasoning. Large vision-language models (VLMs) are often computationally constrained by expensive video input context. However, not all frames are useful or necessary to perform the task. We propose FFS, a lightweight policy for *flexible frame selection*. FFS is *adaptive*, learning to use multimodal context to select only what is necessary for the downstream, large VLM, whether it is (a) a single frame for simple questions, or (b) multiple frames for complex queries. Unlike other adaptive selection approaches [45, 70], our FFS *improves* overall efficiency, requiring no expensive LLM/VLM API calls of its own.

events, from which only a few may be semantically relevant to the task. Thus, sampling uniformly can potentially result in frame inputs that are either redundant, containing information in multiple frames where one would suffice [5, 30], or irrelevant, with respect to the input question [38, 85]. Processing these inputs can thus lower overall efficiency. In addition, for video reasoning tasks, key events of interest can occur for relatively short timespans within the larger video, and this uniform sampling method can miss providing the VLM this critical information and thereby lower accuracy. The interplay of these characteristics of videos pro-

vides an opportunity: by focusing on frames that *matter* to the reasoning task, we can potentially improve both accuracy and efficiency simultaneously.

This opportunity has motivated the development of multimodal frame selection models, which build on early work in unimodal systems [28, 76] designed for simpler tasks (e.g. action recognition [6, 29, 59]). These models aim to select a fixed number of input frames [5, 85] to pass to downstream multimodal models [32, 55] that can perform the final reasoning task. However, there are two limitations of these approaches: (1) state-of-the-art approaches for frame selection can be as computationally expensive as the downstream models [70, 85], so while there are accuracy improvements, they come at substantial efficiency *cost*, and (2) they often pick a fixed number of frames [5, 85], regardless of the specific nature of the video or question. As illustrated in Fig. 2, this can lead to cases of specific (video, question) inputs where the sampler’s fixed value is set too high, picking redundant or semantically irrelevant frames and reducing efficiency, or set too low, missing key information and reducing accuracy.

To address these limitations, we propose the Flexible Frame Selector (FFS), an efficient, adaptive selection model based on a new *flexible, controllable selection* operation. The core insight for our approach is motivated by the fact that adaptive selection for video reasoning is inherently *contingent on multimodal context*: videos may require few or more frames depending on the question complexity, and some questions are more readily answerable with certain videos. In Fig. 1, we illustrate some examples: for simple reasoning questions, it can be sufficient to process one or two frames, while for multi-event reasoning questions, the model may need multiple frames sampled across the video. Accordingly, our adaptive FFS model is able to dynamically determine the number of frames needed and select only what is necessary. Central to our new efficient design is our introduction of a novel “[CTRL]” register token to the selector, which *accumulates* the necessary multimodal context and enables us to leverage efficient “fixed” selection design components in an adaptive manner (Fig. 2). This new token and its integration to the model enables the user to explicitly control the minimum/maximum bounds for the number of selected frames, but gives the model the flexibility (in a differentiable, learnable manner) to determine *both* the number of frames it should pick and the discrete selection itself, conditioned on the input video/question context. Critically, our overall design improves overall efficiency, making no expensive VLM/LLM calls of its own like other adaptive video reasoning and selection frameworks [45, 70]. We demonstrate the effectiveness of our approach on improving efficiency-accuracy on standard multimodal video question answering and event reasoning benchmarks, including IntentQA [33], NExt-QA [77], and EgoSchema [41].

2. Related Work

Multimodal Video Reasoning. Video Question-Answering (videoQA) is an important task for assessing multimodal video understanding systems’ ability to reason about videos [41, 77, 81, 82, 90]. End-to-end VLMs for this task typically consists of strong vision and language encoders/decoders [2, 7–10, 32, 34, 39, 88]. For state-of-the-art VLMs, memory limits often lead to significant down-sampling: e.g. temporally sampling a subset frames with large strides [8, 10, 34, 65] or spatially subsampling each frame to a single token [68, 83, 91]. Importantly, these methods do not intelligently select frames according to what may be useful to the task itself. As such, recent video reasoning model frameworks [45, 70, 87] have considered an adaptive approach, leveraging expensive VLMs and LLMs as APIs [1] to iteratively select and reason about context to improve performance. These approaches, however, come at substantial efficiency cost, leaving an opportunity to improve efficiency while maintaining accuracy gains.

Unimodal Selection and Compression. The ideas surrounding selection of key and relevant information has a rich history in video analysis and understanding. Video compression relies on the idea that there are large amounts of visual redundancy between frames, such that only a significantly smaller set of keyframes are necessary to process fully [74, 75]; in our work, we extend this idea further, with the observation that *even fewer* frames may be necessary to fully process in the context of a specific question about a video. In earlier work, accelerating action recognition in short videos through adaptive selection was extensively explored [16, 19, 28, 76]. While these settings are unimodal (vision-only) and limited to a closed set of actions and activities, these techniques have formed the basis for inspiring later work in multimodal video tasks [5, 38].

Multimodal Selection for VLMs. Recent work has focused on leveraging language in conjunction with video information to guide selection more effectively [5, 18, 85]. However, these approaches all share some key limitations. First, prior work is largely constrained to rigid selection designs, where a fixed number of selected frames or segments must be determined ahead of time, and applied uniformly to all methods; e.g., in [5] this number is set to 1, and in [85] it is set to 4. In contrast, our method adaptively determines how many selected frames are actually needed *after* observing the video and question input. While some prior work leverage differentiable operations [5, 18], other state-of-the-art techniques have a strictly non-differentiable design, and require binary groundtruth supervision for each frame, denoting whether it should be selected or not [85]. For example, [85] requires pre-training with groundtruth frame-level annotations to denote explicitly whether a given frame should be selected or not, and fine-tuning with VLM-generated pseudolabels [32] which

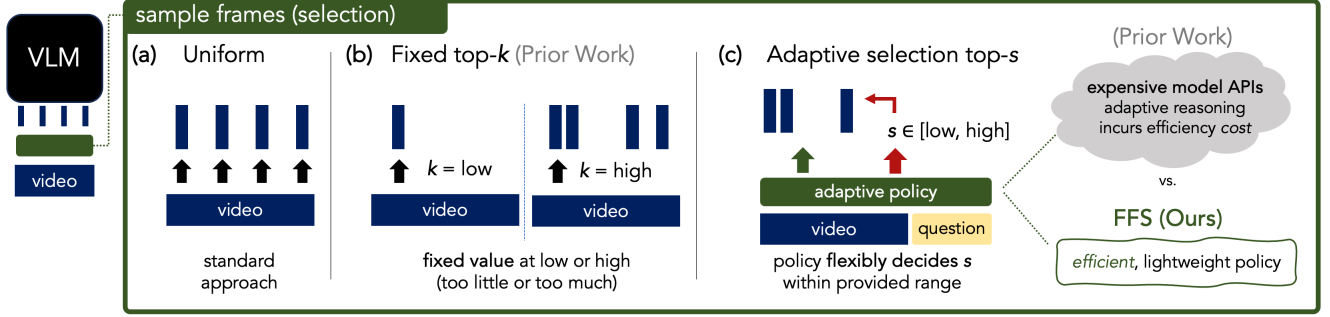


Figure 2. **Selection approaches (overview).** (a) The standard approach for selecting frames for large VLMs is to draw a sparse sample uniformly across the video, which can lead to missing key relevant information or inclusion of semantically irrelevant content to the input question [30, 38]. (b) Prior work [5, 85] improved on this with multimodal selection, usually with a top- k policy that is fixed to low or high values; this leads to cases where the selection is either too broad or too narrow for the specific (video, question) input. (c) Our work improves on this further by creating an efficient, end-to-end learnable policy which can explicitly decide the value of k (within a provided [low, high] range) *conditioned* on the input (video, question), achieving a better efficiency and accuracy trade-off. Importantly, unlike recent work on adaptive video reasoning [45, 70], our policy does not make API calls to expensive VLM or LLM models (e.g. GPT-4).

can be noisy and give limited improvement in some settings. Finally, recent work like [45, 70, 85] leverage expensive VLM foundation models [1] to perform selection, which are themselves as heavy as the downstream VLM. While approaches like [70] show promising adaptive selection behavior, this substantial model cost obviates any potential efficiency gain. In contrast, we propose a more effective and lightweight design that can learn end-to-end from the downstream VLM task. We discuss potential synergies of our work with additional concurrent related work (e.g., [51, 64, 69, 86]) in our supplement.

3. Technical Approach

Overview. We here describe our technical approach for our Flexible Frame Selector (FFS), aimed at improving efficiency and accuracy for downstream video-language reasoning models. In Sec. 3.1 we present an overview of our approach. Our model has three key design aspects: First, the model is *lightweight*, adding marginal computational processing ahead of the downstream model to achieve overall efficiency and accuracy gains. Second, the model is end-to-end differentiable and can be supervised by task-specific VLMs without the need for groundtruth frame labels [85]. We describe how we achieve both of these first two design goals in Sec. 3.2, leveraging “fixed” top- k machinery [24, 27, 40]. Third, our model predicts an explicit, variable-sized selection set of frames conditioning on the input; we describe in Sec. 3.3 our new [CTRL] register token and its novel integration to create flexible behavior from traditionally inflexible architecture components, without compromising the first two design aspects.

3.1. Task and Goal Overview

We first briefly describe our task setting and overall goals.

Task and Inputs. Our work focuses on the video question-answering task to measure video-language reasoning capabilities, where the input to the model is a video frame sequence $V = \{v_1, \dots, v_n\}$ with n frames and a natural language question q_{quest} . The goal is to output a final answer a satisfying the question. In the discriminative formulation of this task, a set of m answer candidates $q_{\text{opts}} = \{q_{\text{opt}_1}, \dots, q_{\text{opt}_m}\}$ from which the correct answer $a \in q_{\text{opts}}$ is determined. Thus, the overall language input is constructed as $L = \{q_{\text{quest}}\} \cup q_{\text{opts}}$.

Goal. The goal of our method is to address the underlying motivation that the number of frames necessary to perform a given task is dependent on the specific context of the input question and video: for some videos where the question is simple, or the video is highly redundant, sampling one or two frames may suffice, while for multi-event reasoning or counting questions, multiple frames may be required to achieve full accuracy (Sec. 1). Thus, our goal is to create a selection model (FFS) that selects s frames from the original sequence, given multimodal inputs (V, L) :

$$\text{FFS}(V, L) : \{v_1, \dots, v_n\} \mapsto \{v_{i_1}, \dots, v_{i_s}\} \quad (1)$$

where $\{i_1, \dots, i_s\}$ are the selected indices, $s \in [k_{\text{low}}, k_{\text{high}}]$ is within pre-specified low/high bounds for the selection and $s \ll n$ represents a significant reduction in overall visual input size (number of frames). Our design is intended to be *flexible*; rather than a fixed value k that must be applied to *all* (video, question) pairs (regardless of need or relevance), our model is able to consider what is relevant first before a downstream (heavy) VLM is applied to an adaptive selection to achieve a better accuracy/efficiency balance. In the rest of this section, we discuss the key design elements that enable this behavior, and satisfy our overall design goals for our selector.

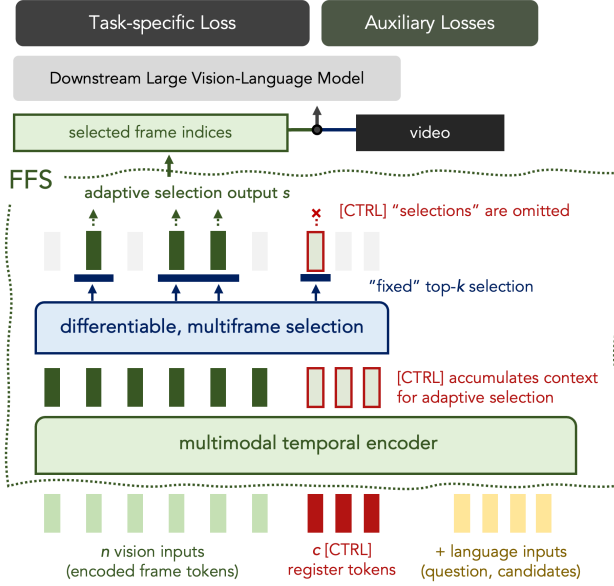


Figure 3. **Model overview.** We introduce a new, lightweight model (FFS) for input frame selection. Different from prior work, our model *flexibly* determines the number of selected frames $s \in [k_{low}, k_{high}]$ rather than a fixed value for k , where bounds are set by the user, and the exact value is adaptively determined by the model on specific (video, language) input without expensive VLM or LLM reasoning calls. The model makes the joint decision of how many and which frames to select by accumulating selection-relevant context in [CTRL] register tokens with a multimodal temporal encoder. Our overall design is end-to-end trainable, supervised by a downstream heavy vision-language model and auxiliary losses that help to guide selection behavior. See Sec. 3.2 and 3.3.

3.2. Flexible Frame Selector (FFS): A “Fixed” Base

We now discuss the initial, base design of the overall selector. Our overall Flexible Frame Selector (FFS) model design is constructed initially on an inflexible design (*i.e.*, selects a *fixed* number of frames, leveraging “fixed” top- k machinery [24, 27, 40]) that satisfies the initial design goals around having a lightweight and end-to-end learnable model. Our *full design* then incorporates some new key modifications, described further in Sec. 3.3, which preserve these attributes but add flexibility.

Overview. As illustrated in Fig. 3, our model has a two stage design: first, we have the *multimodal encoder*, which takes input tokens from the vision V and language L modalities, and outputs a list of selector representations that capture the semantic *informativeness* of each frame relative to the video and the question. The second stage then takes these representations and performs a discrete (differentiable) selection of the embeddings with the highest relevance to the downstream query.

We first describe the vanilla “fixed” version of our model that satisfies the conditions of being end-to-end differen-

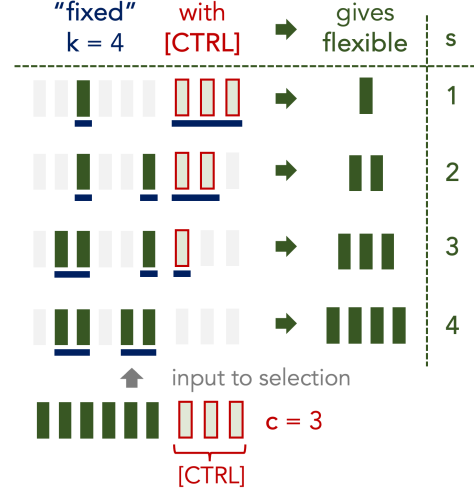


Figure 4. **Flexible Frame Selection with [CTRL].** We illustrate mechanically how introducing the [CTRL] register tokens enables flexibility with “fixed” differentiable selection. After the selector encoder step, the [CTRL] register tokens contain contextual information relevant for selection from the input video and language (question). While the top- k selected shown here is “fixed” at 4 (blue underlines), the resulting number of output selected tokens s can range from $[1, 4]$ depending on how many of the $c = 3$ [CTRL] tokens are also selected. Thus, the model can adaptively decide how many frames to select by adjusting what representation is accumulated in [CTRL]. See Sec. 3.3 for details.

tiable and lightweight, but is not yet flexible. In Sec. 3.3, we describe how introducing the [CTRL] register token and auxiliary losses unlock adaptive behavior.

FFS Encoder (Fixed Size Component). Our encoder model takes as input visual features $V_F = \{f_1, \dots, f_n\}$ corresponding to the n frames, which can be pre-computed or obtained from a lightweight image visual encoder model and textual features $L_T = \{t_{quest}, t_{opt_1}, \dots\}$ (mapping to the $m+1$ language inputs in L) from a similarly lightweight text encoder (*e.g.*, BERT [14] from CLIP [55]). We augment these representations to preserve temporal order (with respect to the original video) by adding learnable position encodings, and to both vision and language embeddings we add a learnable modality embedding to indicate the input type. Thus, the encoder is defined as:

$$\text{FFS}_{enc}(\{V_F, L_T\}) : (\{v_1, \dots, v_n\}, \{t_{quest}, t_{opt_1}, \dots\}) \mapsto \{f'_1, \dots, f'_n\} \quad (2)$$

where the output is a sequence of length n and each encoding f'_i is the selector encoding corresponding to original vision input v_i . This encoder is parameterized by a lightweight multimodal transformer, adding marginal parameter overhead relative to the downstream task model.

FFS Selector (Fixed Size Component). Our selector component takes the encoded output sequence $\{f'_1, \dots, f'_n\}$

from Eq. 2 as its input. These representations are mapped to a final logits vector $l = \text{MLP}(\{f'_1, \dots, f'_n\})$ of size n using an MLP network applied to each of the n embeddings. In a traditional fixed size design (*i.e.* number selected $s = k$), we would perform top- K selection on these embeddings to output a final set of the indices $\{i_1 \dots, i_s\}$ corresponding to the selected frames, as desired. Specifically,

$$\text{FFS}_{\text{sel}}(\{f'_1, \dots, f'_n\}) : \arg \max_k(l) \mapsto \{i_1, \dots, i_{s=k}\} \quad (3)$$

We describe our design modifications to enable flexibility through the [CTRL] token in Sec. 3.3 below.

FFS Learnability and Training. During training, our non-differentiable, discrete ($\arg\max$) top- k operation is made end-to-end differentiable by generalizing the Gumbel-Max straight-through estimator (STE) trick for Top-1 selection [24] to k values, per [27, 40]. Briefly, we add noise from the Gumbel(0,1) distribution [22] to the logits l to obtain a perturbed logits l' before “sampling” the top- k values on the forward pass. We apply the STE formulation to ensure what is then forwarded is the discrete selection, but on the backward pass the gradients flow through the original logits, as a first-order approximation [37]. The selected indices are represented as a binary mask of length n , which is then applied to the original frame set to obtain the final selected frames for the downstream VLM to perform the task. Through this formulation, we are able to directly integrate our model with a downstream VLM, and use supervision from the final task itself to learn what frames are worth selecting given example inputs, *without* needing explicit frame-level annotations (*e.g.*, localization labels, which can be expensive to obtain). In the context of discriminative videoQA, the supervising loss is cross entropy over the candidate option set ($|q_{\text{opts}}| = m$) and this loss is backpropagated end-to-end to the learned flexible selection policy, as illustrated in Fig. 3.

3.3. FFS with [CTRL]: From “Fixed” to Flexible

As initially described in Sec. 3.2, the base components for our FFS model are *inflexible*, requiring a fixed value for the number of selections $s = k$. The core challenge at the heart of our model remains: how do we create flexible behavior ($s \in [k_{\text{low}}, k_{\text{high}}]$) in such a setting, while preserving key properties of end-to-end differentiability and a lightweight design? Here, we describe how introducing a new token “[CTRL]” with its specific integration to our model allows us to address this challenge. We illustrate the mechanics of the resulting *flexible* selection in Fig. 3 (right).

Intuition. Intuitively, a flexible selector should, within the context of minimum, k_{low} , and maximum, k_{high} , size windows, determine that if the number of “informative” frames is *lower* than k_{high} , then it should only select a number $s \leq k_{\text{high}}$ of these frames for what is actually helpful. Thus, the problem reduces to building an adaptive filter conditioned on the vision-language inputs, such that anything

above this filter should be considered for final selection, and anything below it should be disregarded.

[CTRL] Modifications. We create this “threshold” by introducing our learnable [CTRL] token. At the input encoder level, we modify Eq. 2 to include as input a sequence of c [CTRL] tokens, as shown in Fig. 3, concatenated to the original input sequence. The encoder then outputs $n + c$ representations, instead of only n . Thus, Eq. 2 updates to:

$$\begin{aligned} \text{FFS}_{\text{enc}}^{\text{flex}}(\{V_F, [\text{CTRL}], L_T\}) : \\ (V_F, \{c_1, \dots, c_c\}, L_T) \mapsto \{f'_1, \dots, f'_n, c'_1, \dots, c'_c\} \end{aligned} \quad (4)$$

During multimodal encoding, the model can accumulate a representation in the [CTRL] register tokens that intuitively captures a more general video/question-level value, and not to any particular frame input. This representation is then implicitly compared during the (fixed) top- k selection operation, and only indices corresponding to *non*-[CTRL] representations (*i.e.*, actual frames) are finally output from FFS, with any selections to [CTRL] being discarded. Thus, we make the following modification to Eq. 3:

$$\begin{aligned} \text{FFS}_{\text{sel}}^{\text{flex}}(\{f'_1, \dots, f'_n, c'_1, \dots, c'_c\}) : \\ \arg \max_k(l) \mapsto \{i_1, \dots, i_{s \in [k_{\text{low}}, k_{\text{high}}]}\} \end{aligned} \quad (5)$$

where the final logits $l = \text{MLP}(\{f'_1, \dots, f'_n, c'_1, \dots, c'_c\})$ is expanded to size $n + c$, and selections of [CTRL] are dropped by keeping the first n indices of the selection mask.

By modulating k (in the selector) and c (number of [CTRL] tokens), we are able to explicitly encode the $[k_{\text{low}}, k_{\text{high}}]$ bounds, where $k_{\text{high}} = k$, and $k_{\text{low}} = k - c$. Intuitively, this k_{low} value is reached when the model chooses to select $k - c$ tokens from the n frames, and its “leftover budget” is accounted for by selecting *all* c [CTRL] tokens. In Fig. 3 and Fig. 4, we illustrate the case where $k = 4$ and $c = 3$; here, we are able to flexibly select any number between 1 and 4 frames (*e.g.*, in Fig. 3, we show a selection of 1 [CTRL] token and 3 frames). In this way, the [CTRL] token stands for “control” in two aspects. First, the [CTRL] token enables the user to *control* the bounds of k_{low} and k_{high} , such that $s \in [k_{\text{low}}, k_{\text{high}}]$, and its inclusion enables the model to control (within these bounds) how many tokens are finally selected. This is useful for cases when the downstream model performs better with a particular minimum number of inputs. Second, the post-encoder representations c'_* contain a “baseline control” reference that dynamically thresholds which frame tokens matter to the downstream task and are filtered through downstream.

[CTRL] Initialization and Training. We initialize the [CTRL] representations with mean embeddings of the input visual tokens $\{v_1, \dots, v_n\}$, combined with a learnable embedding $e_{[\text{CTRL}]_c}$. In addition to the task-specific loss described in Sec. 3.2 backpropagated from the downstream VLM, we supervise the selection behavior directly with a

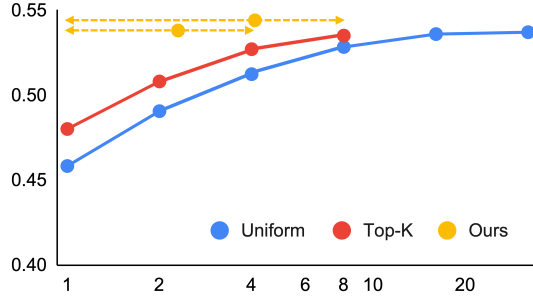


Figure 5. **Selection analysis (Ablation).** We perform an ablation analysis of our method against standard uniform selection (blue) and “fixed” top- k selection ($k \in \{1, 2, 4, 8\}$). The x -axis (log-scale) indicates the number of *selected* frames from an initial pool of 32 frames, and y -axis indicates accuracy on a validation subset of [33]. We show two flexible policies in yellow ($s \in [1, 4]$ and $[1, 8]$), where arrows indicate range and dots indicate average.

set of fixed- k policies of varying k (e.g., $k = \{1, 2, 4, 8\}$). We find that the initialization and incorporation of these auxiliary forms of supervision help to guide training and improve generalization at test time, and provide additional details, equations, and ablations in the supplement.

4. Experiments and Discussion

Here, we describe the videoQA datasets and evaluation metrics used (Sec. 4.1), our key baselines (Sec. 4.2) and implementation details (Sec. 4.2), and our discussion of results and analysis (Sec. 4.3). We focus our main analysis on video reasoning (QA) datasets highlighted in recent work [70, 87], with additional analysis in the supplement.

4.1. Datasets

We consider standard challenging video-language datasets for videoQA, developed specifically to benchmark temporal and causal video reasoning across multiple timescales: IntentQA [33], NExT-QA [77], and EgoSchema [41].

IntentQA [33] is focused on understanding the ability of videoQA systems to effectively answer questions specifically in cases where *intent* matters to the dynamics of events and event relationships, particularly with a focus on temporal and causal relationships. There are three main subsets: Temporal-Past/Next (TP/TN), Causal-How (CH), and Causal-Why (CW). Each video contains one question and 5 candidate answers, and results are reported using accuracy on the test set of 2134 video-question pairs, following [33]. **NExT-QA** [77] is focused on characterizing the ability of videoQA systems to effectively answer questions across three types: causal (C), temporal (T) and descriptive (D). We focus on the same multiple choice setting reported in prior work for video selection [5, 85], where each video clip contains one question and 5 candidate answers. NExT-QA

Table 1. Comparison with prior methods on IntentQA [33] test set (% accuracy). We observe improvements with FFS both overall and across the different causal (CW, CH) and temporal (TP/TN) subsets. EVA-V-L* extends the EVA-V-L model with the same LLM [48] priors used in CaVIR [33], for comparison consistency. @X denotes number of input frames, and $A \in [B, C]$ denotes average A selected with range $[B, C]$. Models that use 100B+ parameter, closed-source models are marked in grey.

Model	CW	CH	TP/TN	Overall
HGA [25]	44.9	51.0	39.6	44.6
HME [15]	46.1	54.3	40.8	46.2
HQGA [78]	48.2	54.3	41.7	47.6
VGT [79]	51.4	56.0	47.6	51.3
CaVIR* [33]	58.4	65.5	50.5	57.6
LLoVi (GPT-4) [87]	68.4	67.4	51.1	64.0
EVA-V-L*@32	67.0	66.3	53.7	63.6
EVA-V-L*@4	65.9	64.3	51.4	62.1
+ (Ours; @2.4 $\in [1, 4]$)	68.5	67.6	51.9	64.2

Table 2. Efficiency analysis. Instead of processing all uniformly sampled frames densely, our lightweight FFS processes the initial set of frames with marginal up front cost, resulting in overall efficiency gains (in frames/FLOPs) with comparable or improved accuracy (% NExT-QA, Tab. 3). Expanded analysis in supplement.

Model	Params	Avg. Frames	FLOPs	Acc %
EVA-V-L (baseline uniform)	8.2G	32	246.8T	61.6
	8.2G	4	31.1T	60.2
FFS (selection only)	0.2G	32	0.15T	-
EVA-V-L + FFS (combined)	8.4G	2.3 from 32	18.2T	62.2

and IntentQA complement each other in terms of question types and event reasoning, where NExT-QA aims to provide broader coverage on types of event reasoning, while IntentQA focuses exclusively on intent understanding with temporal and causal questions. The results are similarly reported using accuracy on the validation set of 4996 video-question pairs, following [5, 62, 72, 85].

EgoSchema [41] contains *long* egocentric videos (180s) based on the Ego4D [21] benchmark with multiple-choice questions, designed specifically to assess long video understanding (e.g., 100s human-verified “temporal certificates”). EgoSchema is focused entirely on evaluation: the hidden test set consists of 5000 videos via an evaluation server, of which 500 were publicly released for validation. Following prior work [4, 45, 50, 70, 87], we report results (accuracy) on the main (full, 5k) test set for overall results, and ablations on a subset of the validation.

4.2. Implementation Details and Baselines

Downstream Models. We consider multiple downstream models in this work for supervision and inference, with a primary focus on integrations with open-source models.

Table 3. Comparison with prior methods on NExT-QA [77] (% accuracy). “Hard” refers to the subset from [5], focused on (video, question) pairs that cannot be answered by a single frame. Prior work that uses significantly larger, closed-source models are grey. @X denotes number of input frames, and FFS selection $A \in [B, C]$ indicates an average of A with range $[B, C]$ frames. Our flexible video selector offers higher or comparable accuracy on fewer input frames for both EVA-V-L and OpenLLOVi.

Model	Desc.	Temp.	Caus.	Hard	All
HGA [25]	59.3	50.7	46.3	44.1	49.7
Temp[ATP] [5]	65.0	49.3	48.6	37.6	51.5
VGT [79]	64.1	55.1	52.3	-	55.0
MIST-CLIP [18]	66.9	56.6	54.6	-	57.1
Paxion (InternVid) [72, 73]	68.5	56.0	53.0	38.5	57.0
CoVGT [80]	69.3	57.4	58.8	-	60.0
SeViLa (T5-XL) [85]	75.6	61.5	61.3	-	63.6
ViperGPT (GPT-3.5) [62]	-	-	-	53.6	60.0
LLOVi (GPT-4) [87]	75.5	61.0	69.5	-	67.7
VideoAgent (GPT-4) [70]	81.1	64.5	72.7	58.4	71.3
EVA-V-L@32	72.9	53.4	61.2	44.3	61.6
EVA-V-L@4	71.9	51.3	60.0	42.6	60.2
+ FFS (Ours; @2.3 $\in [1, 4]$)	73.1	53.8	62.1	44.5	62.2
OpenLLOVi (Gemma; @32)	78.5	58.8	67.8	59.5	66.5
OpenLLOVi (Gemma; @8)	75.8	57.9	66.7	57.5	65.3
+ FFS (Ours; @8.1 $\in [1, 16]$)	78.2	58.8	68.1	59.9	66.7

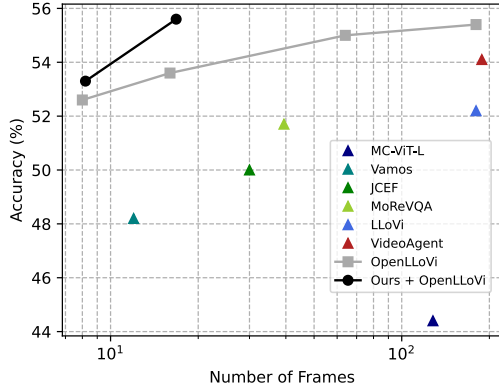


Figure 6. **Efficiency-accuracy analysis (EgoSchema).** We visualize the efficiency-accuracy trade-off against key baselines on the EgoSchema long video benchmark [41]. The OpenLLOVi curve (gray) is uniformly sampled, and for FFS (Ours + OpenLLOVi), dots are at the average number of selected frames. See Table 4.

Our primary model for task supervision is the contrastive VLM from VFC [46]. We call this model EVA-V-L, since we have updated the backbone to the same state-of-the-art EVA-CLIP-8B-plus [17, 61] used in recent adaptive, long video reasoning work [70]. We also consider a long video reasoning model based on LLOVi [87], which takes input frames and passes it to a base captioner [23, 36, 89] and then downstream LLM. In this work, we re-implement a version of this model based purely on open-source models, calling it “OpenLLOVi”: we keep consistent base VLMs as

Table 4. Comparison with prior work on long-video EgoSchema benchmark [41] (% accuracy, full set). We find our model improves on accuracy and efficiency. For multi-stage systems (e.g., [70]), “frames” includes inputs processed by large model (4B+ param) components; detailed analysis in supplement.

Method	Frames	Test Acc.
SeViLa [85]	36	22.7
LongViViT [50]	256	33.3
MC-ViT-L [50]	128	44.4
Vamos (GPT-4) [67]	12	48.2
JCEF (PaLM 2) [45]	30	50.0
MoReVQA (PaLM 2) [45]	39.4	51.7
LLOVi (GPT-3.5) [87]	180	52.2
VideoAgent (GPT-4) [70]	188.4	54.1
OpenLLOVi (Gemma; @180)	180	55.4
OpenLLOVi (Gemma; @16)	16	53.6
+ FFS (Ours; @16.8 $\in [4, 32]$)	16.8	55.6

in [70, 87] and using Gemma 2 Large [63] as a substitute LLM. These models are representative of recent successful model designs for videoQA; we describe additional details, variations, and integrations in the supplement.

Flexible Frame Selector (FFS). Our FFS model is significantly lightweight relative to the downstream models (e.g., in Table 2, 40x smaller). For our default design, our input visual encodings v_* are obtained with a CLIP-ViT-B/32 [54] backbone, and the corresponding CLIP BERT-Base [14] model for encoding language inputs t_* . Our multimodal selector encoder is a similarly lightweight transformer, and we explore different values of k_{low}, k_{high} (i.e., k and c) in our ablation studies, selecting from a default initial pool of 32 frames for NExT-QA and IntentQA [33, 77] and 64 frames for EgoSchema [41]. We provide additional training and architecture details in the supplement.

Baseline selection approaches. Our baseline comparisons are standard uniform selection (the default technique for downstream VLMs) and different fixed- k selection policies, where we ablate our [CTRL] token and maintain the rest of our selection model design. We also compare with SOTA adaptive selection methods like VideoAgent [70] which use expensive, closed-source models to perform iterative selection and reasoning, showing competitive performance with consistent or weaker models, and efficiency improvements. Experiments are primarily implemented in PyTorch [52].

4.3. Results and Analysis

Selection Ablations and Analysis. In Fig. 5, we plot the performance of our method and key ablation baselines against the number of selected frames for the downstream EVA-V-L model. The baseline curve represents a uniform selection of frames, with increasing accuracy as more frames are considered and the chance that informative frames are included in this set grows. However, by fo-

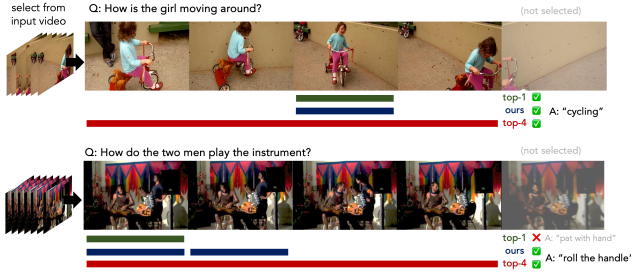


Figure 7. **Qualitative results.** Best viewed in PDF. Examples from NExT-QA [77], showing ours vs. fixed $k = \{1, 4\}$ selection; see supplement for longer visualizations (e.g., EgoSchema [41]). **(top)** we observe that one selected frame is sufficient to correctly answer the relatively simple question, and our model is able to avoid selecting any redundant or low-signal frames. **(bottom)** multiple frames needed to observe the interaction dynamics between the men and their instruments. Top-1 is insufficient, and leads to incorrect prediction “pat with hand”. Our model selects enough frames to obtain the correct answer “roll the handle”.

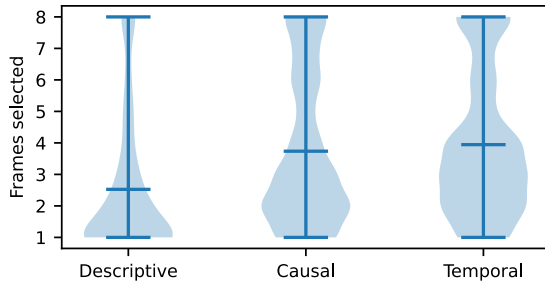


Figure 8. **Adaptive Selection Distribution.** We observe that the adaptive selection distribution for an FFS $s \in [1, 8]$ policy corresponds with different question types in NExT-QA [77]: descriptive questions are often answerable with fewer frames, while more complex causal and temporal questions require additional frames.

cusing on a specific selection, we can see improvements. The red curve is a *fixed* top- k policy (k unique outputs) and an ablation of our method that removes the [CTRL] tokens and flexibility, consistent with selection operation designs in prior work [5, 18, 85]. At the low end, the top-1 selection ablation mirrors the design of [5], and indicates what can be answerable with a single frame. In contrast, we observe our flexible approach ranges adaptively according to the specific needs of the video and question, within a provided range, offering a better overall efficiency trade-off. We visualize the adaptive distribution of our model in Fig. 8, and observe FFS maps well to different question types in [77].

Efficiency Analysis. In Table 2, we show an efficiency analysis of our lightweight policy in the context of a downstream integration. FFS processes the initial set of 32 frames with almost negligible marginal cost, relative to the downstream model, but results in substantial overall effi-

ciency gains: relative to the original 32 uniformly sampled baseline, we improve 13.5x on downstream frames and FLOPs with comparable accuracy, and relative to the 4 frame baseline, we have a 40% reduction (1.7x efficiency gain) with simultaneous accuracy improvement. In Figure 6, we show the relative efficiency-accuracy gains compared to SOTA methods that leverage closed-source models for reasoning, including models like VideoAgent that specifically do this for adaptive frame selection. These results suggest our method is well-suited for long video settings such as those in EgoSchema. We provide additional efficiency analysis details (latency, etc.) in the supplement.

Prior work comparisons. In Tables 1, 3, 4 we make comparisons of our method against our baseline models and prior work. Across all datasets, we observe that our model improves over our downstream model alone, in terms of efficiency and accuracy. On NExT-QA and IntentQA, we observe that our model is able to perform well even with selecting overall relatively few frames on average. On EgoSchema, a longer video benchmark, we observe our model performs well with fewer frames relative to prior work. This result suggests our technique is promising for long video settings with more input frames. We provide additional comparisons with prior work in the supplement.

Qualitative Results. Finally, in Fig. 7, we illustrate some qualitative results of our model and its selections. We specifically contrast our full model with flexible selection against the ablation model selecting a fixed top-1 and a fixed set of 4. In the top example, the question is quickly answered with a single well-chosen frame, while the top- k model includes irrelevant or redundant frames in its selection. Meanwhile, in the bottom example, the single frame selector is unable to gather sufficient information to correctly answer the question, since multiple frames are needed to understand the dynamics of the interaction between the players and their instruments. Meanwhile, our flexible model adapts to what is necessary of this more complex input, and selects an appropriately higher number of frames so that the downstream VLM can correctly answer the question. We include additional visualizations of longer video settings (e.g., EgoSchema [41]) in the supplement.

5. Conclusion

In this work, we proposed Flexible Frame Selector (FFS), a learnable policy model with a new *flexible* selection operation that enables video-language models to focus on the most informative frames for the downstream multimodal reasoning task in an efficient, adaptive, and learnable manner. We verified the efficacy of our method on standard video-question answering and reasoning benchmarks, highlighting FFS adaptive selection behavior and efficacy against key baselines. We discuss broader impacts, acknowledgments, and limitations in the supplement.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 3
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1, 2
- [3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A Video Vision Transformer. In *ICCV*, 2021. 1
- [4] Ivana Balažević*, Yuge Shi*, Pinelopi Papalampidi*, Rahma Chaabouni, Skanda Koppula, and Olivier J Hénaff. Memory consolidation enables long-context video understanding. *ICML*, 2024. 6, 2
- [5] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the “video” in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2917–2927, 2022. 1, 2, 3, 6, 7, 8, 5
- [6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2
- [7] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 2
- [8] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023. 1, 2
- [9] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2023. 1
- [10] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 2
- [11] Joonmyung Choi, Sanghyeok Lee, Jaewon Chu, Minhyuk Choi, and Hyunwoo J Kim. vid-tldr: Training free token merging for light-weight video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18771–18781, 2024. 5
- [12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 1
- [13] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *ICLR*, 2024. 5
- [14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 4, 7, 1
- [15] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007, 2019. 6
- [16] Hehe Fan, Zhongwen Xu, Linchao Zhu, Chenggang Yan, Jianjun Ge, and Yi Yang. Watching a small portion could be as good as watching all: Towards efficient video classification. In *IJCAI*, 2018. 2
- [17] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023. 7
- [18] Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14773–14783, 2023. 2, 7, 8
- [19] Shreyank N Gowda, Marcus Rohrbach, and Laura Sevilla-Lara. Smart frame selection for action recognition. In *AAAI*, 2021. 2
- [20] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 5
- [21] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, et al. Ego4d: Around the world in 3,000 hours of egocentric video, 2022. 6, 2, 5
- [22] Emil Julius Gumbel. Statistical theory of extreme values and some practical applications. *Nat. Bur. Standards Appl. Math. Ser. 33*, 1954. 5
- [23] Wenyi Hong, Weihang Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290, 2024. 7, 1
- [24] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 3, 4, 5
- [25] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11109–11116, 2020. 6, 7

- [26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2
- [27] Wouter Kool, Herke Van Hoof, and Max Welling. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In *International Conference on Machine Learning*, pages 3499–3508. PMLR, 2019. 3, 4, 5
- [28] Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6232–6242, 2019. 2
- [29] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 2
- [30] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022. 1, 3, 5
- [31] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1
- [32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2
- [33] Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. Inten-tqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11963–11974, 2023. 2, 6, 7, 1, 3, 4, 5
- [34] Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *EMNLP*, 2024. 2
- [35] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022. 3
- [36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 7, 1
- [37] Liyuan Liu, Chengyu Dong, Xiaodong Liu, Bin Yu, and Jianfeng Gao. Bridging discrete and backpropagation: Straight-through and beyond. In *NeurIPS*, 2023. 5
- [38] Haoyu Lu, Mingyu Ding, Nanyi Fei, Yuqi Huo, and Zhiwu Lu. Lgdn: Language-guided denoising network for video-language modeling. *Advances in Neural Information Processing Systems*, 35:25198–25211, 2022. 1, 2, 3
- [39] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 1, 2
- [40] Chris J Maddison, Daniel Tarlow, and Tom Minka. *a** sampling. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014. 3, 4, 5
- [41] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *arXiv preprint arXiv:2308.09126*, 2023. 2, 6, 7, 8, 1, 3, 5
- [42] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, pages 3195–3204, 2019. 1
- [43] Yue Meng, Chung-Ching Lin, Rameswar Panda, Prasanna Sattigeri, Leonid Karlinsky, Aude Oliva, Kate Saenko, and Rogerio Feris. Ar-net: Adaptive frame resolution for efficient action recognition. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 86–104. Springer, 2020. 5
- [44] Thomas Mensink, Jasper Uijlings, Lluís Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. *ICCV*, 2023. 1
- [45] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. MoReVQA: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3, 6, 7
- [46] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15579–15591, 2023. 1, 7
- [47] Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. Spoken moments: Learning joint audio-visual representations from video descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14871–14881, 2021. 1
- [48] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. 6
- [49] Rameswar Panda, Chun-Fu Richard Chen, Quanfu Fan, Ximeng Sun, Kate Saenko, Aude Oliva, and Rogerio Feris. Adamml: Adaptive multi-modal learning for efficient video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7576–7585, 2021. 5
- [50] Pinelopi Papalampidi, Skanda Koppula, Shreya Pathak, Justin Chiu, Joe Heyward, Viorica Patraucean, Jiajun Shen, Antoine Miech, Andrew Zisserman, and Aida Nematzadeh. A simple recipe for contrastively pre-training video-first encoders beyond 16 frames. *CVPR*, 2024. 6, 7, 2
- [51] Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryoo, Donghyun Kim, and Michael S Ryoo. Too many frames, not all useful: Efficient strategies for long-form video qa. *arXiv preprint arXiv:2406.09396*, 2024. 3
- [52] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming

- Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 7
- [53] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023. 3
- [54] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 7, 1
- [55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *ICML*, 2021. 1, 2, 4
- [56] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020. 4
- [57] Ishaan Singh Rawal, Alexander Matyasko, Shantanu Jaiswal, Basura Fernando, and Cheston Tan. Dissecting multimodality in videoqa transformer models by impairing modality fusion. In *Forty-first International Conference on Machine Learning*, 2024. 5
- [58] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022. 1
- [59] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2
- [60] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Annual Meeting of the Association for Computational Linguistics*, 2017. 1
- [61] Quan Sun, Jinsheng Wang, Qiyang Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*, 2024. 7, 1
- [62] D  dac Sur  s, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023. 6, 7, 2
- [63] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L  onard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram  , et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. 7, 1
- [64] Haibo Wang, Chenghang Lai, Yixuan Sun, and Weifeng Ge. Weakly supervised gaussian contrastive grounding with large multimodal models for video question answering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5289–5298, 2024. 3
- [65] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 2
- [66] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. All in one: Exploring unified video-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608, 2023. 5
- [67] Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwunjoon Lee, and Chen Sun. Vamos: Versatile action models for video understanding. *ECCV*, 2024. 7, 1
- [68] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6847–6857, 2021. 2
- [69] Xijun Wang, Junbang Liang, Chun-Kai Wang, Kenan Deng, Yu Lou, Ming C Lin, and Shan Yang. Vila: Efficient video-language alignment for video question answering. In *European Conference on Computer Vision*, pages 186–204. Springer, 2024. 3
- [70] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. *ECCV*, 2024. 1, 2, 3, 6, 7, 5
- [71] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12186–12195, 2022. 5
- [72] Yi Wang, Kunchang Li, Yizhuo Li, Yanan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 1, 6, 7, 2
- [73] Zhenhailong Wang, Ansel Blume, Sha Li, Genglin Liu, Jaemin Cho, Zineng Tang, Mohit Bansal, and Heng Ji. Paxion: Patching action knowledge in video-language foundation models. *Advances in Neural Information Processing Systems*, 36, 2023. 7
- [74] Olivia Wiles, Joao Carreira, Iain Barr, Andrew Zisserman, and Mateusz Malinowski. Compressed vision for efficient video understanding. In *ACCV*, 2022. 2
- [75] Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R Manmatha, Alexander J Smola, and Philipp Kr  henb  hl. Compressed video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6026–6035, 2018. 2
- [76] Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S Davis. Adaframe: Adaptive frame selection for fast video recognition. In *CVPR*, 2019. 2, 5
- [77] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng 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)*, pages 9777–9786, 2021. [2](#), [6](#), [7](#), [8](#), [1](#), [4](#), [5](#)
- [78] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2804–2812, 2022. [6](#)
- [79] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *European Conference on Computer Vision*, pages 39–58. Springer, 2022. [6](#), [7](#)
- [80] Junbin Xiao, Pan Zhou, Angela Yao, Yicong Li, Richang Hong, Shuicheng Yan, and Tat-Seng Chua. Contrastive video question answering via video graph transformer. *arXiv preprint arXiv:2302.13668*, 2023. [7](#)
- [81] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017. [2](#)
- [82] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1686–1697, 2021. [2](#)
- [83] Antoine Yang, Arsha Nagrai, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*, 2023. [2](#)
- [84] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. Hitea: Hierarchical temporal-aware video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15405–15416, 2023. [5](#)
- [85] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *NeurIPS*, 2023. [1](#), [2](#), [3](#), [6](#), [7](#), [8](#), [4](#)
- [86] Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-voyager: Learning to query frames for video large language models. *ICLR*, 2025. [3](#)
- [87] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. *EMNLP*, 2024. [2](#), [6](#), [7](#), [1](#), [3](#), [4](#)
- [88] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *EMNLP 2023 Demo*, 2023. [2](#)
- [89] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *CVPR*, 2023. [7](#), [1](#), [2](#), [3](#), [4](#)
- [90] Yaoyao Zhong, Junbin Xiao, Wei Ji, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges. *arXiv preprint arXiv:2203.01225*, 2022. [2](#)
- [91] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, 2018. [2](#)