Vamos: Versatile Action Models for Video Understanding

Shijie Wang[1]   Qi Zhao[1]   Minh Quan Do[1]   Nakul Agarwal[2]

Kwonjoon Lee[2]   Chen Sun[1]

[1]Brown University   [2]Honda Research Institute USA

Abstract. What makes good representations for video understanding, such as anticipating future activities, or answering video-conditioned questions? While earlier approaches focus on end-to-end learning directly from video pixels, we propose to revisit text-based representations, such as general-purpose video captions, which are interpretable and can be directly consumed by large language models (LLMs). Intuitively, different video understanding tasks may require representations that are complementary and at different granularity. To this end, we propose versatile action models (Vamos), a learning framework powered by a large language model as the "reasoner", and can flexibly leverage visual embedding and free-form text descriptions as its input. To interpret the important text evidence for question answering, we generalize the concept bottleneck model to work with tokens and nonlinear models, which uses hard attention to select a small subset of tokens from the free-form text as inputs to the LLM reasoner. We evaluate Vamos on five complementary benchmarks, Ego4D, NeXT-QA, IntentQA, Spacewalk-18, and EgoSchema, on its capability to model temporal dynamics, encode visual history, and perform reasoning. Surprisingly, we observe that text-based representations consistently achieve competitive performance on all benchmarks, and that visual embeddings provide marginal or no performance improvement, demonstrating the effectiveness of text-based video representation in the LLM era. We also demonstrate that our token bottleneck model is able to select relevant evidence from free-form text, support test-time intervention, and achieves nearly 5 times inference speedup while keeping a competitive question answering performance. Code and models are publicly released at https://brown-palm.github.io/Vamos/.

1 Introduction

Building a generative model for everyday human activities has long been a dream for researchers working on video understanding. Central to this problem are capturing the interactions between humans and the environment [11,70], modeling the temporal dynamics of activities [14,68], and encoding the hierarchical structures among atomic actions [19,60], activities [6,7], and events [31,33]. Once constructed, the generative model of actions can be applied to a wide range of tasks, including activity and event recognition [57], future behavior prediction [49], goal and intent inference [58], and temporal reasoning [75].
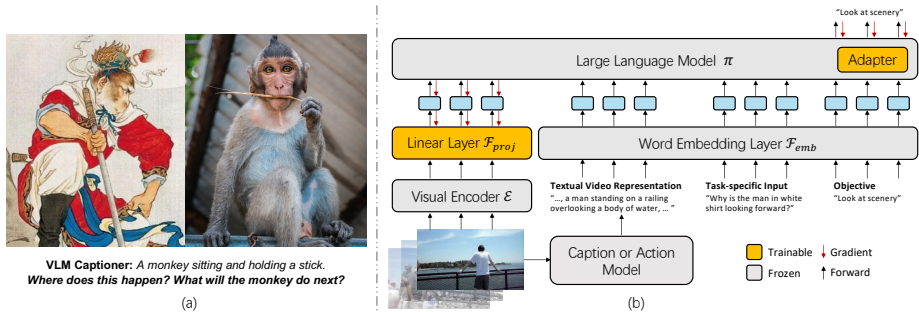
Fig. 1: (a) Visual observations with vastly diverse appearances may be described with the same captions. Our work explores the use of general-purpose text descriptions of video data for action anticipation and question answering. We propose Vamos, a versatile reasoning framework that allows us to study the impact of latent visual representations and free-form text descriptions for downstream applications. (b) Overview of the Vamos framework. It flexibly unifies distributed visual features and text-based representations such as video captions, and can be applied to diverse video understanding tasks.

Despite its desirable properties, generative modeling of actions from video observations remains challenging, hindered by two open research questions: First, what makes good video representations? Earlier attempts often relied on manually defining the actions and the objects being interacted with [1, 19, 60]. They require task-specific prior knowledge, and cannot generalize to the "open vocabulary" scenarios in the wild. Alternative approaches aim to model the temporal dynamics of human pose [32,44,52] or latent representations encoded by deep neural networks [28,68], which are either too fine-grained, or not directly intepretable. Second, what makes a good model of human actions? While earlier approaches attempted to apply rule-based generative action grammars [21, 26, 55, 57], they may not be able to capture the diverse, even peculiar ways of how events would unfold over time. More recent approaches adopt a data-driven framework and directly learn autoregressive models [38] on visual tokens [64,78], where the visual domain is often specialized (*e.g.* cooking, or robotics).

We aim to address both challenges by exploring an unconventional idea: Can **task-agnostic** natural language descriptions, such as those generated by off-the-shelf image caption models [40,48] on sampled video frames, serve as useful video representations for action modeling from videos? And if so, can we then leverage a pre-trained large language model (LLM) [67] as the generative model of actions, represented as free-form text? If both answers are yes, visual reasoning can then enjoy recent advances in LLM research, where they have been shown to be capable of learning context-free grammar [3] with long-range dependencies, predicting time-series [18,53], and performing reasoning [4,12,13,74], all of which are indispensable for action modeling. Since we assume the text descriptions are

general purpose, they can be extracted once and reused for different downstream tasks, similar to pre-computed visual embeddings [59].

To rigorously validate this idea, we propose versatile action models (**Vamos**), a framework that flexibly unifies three representations, namely distributed visual embeddings and free-form text descriptions, and can be applied to various applications by leveraging large language models, such as Llama-2 [67]. The visual embeddings are linearly projected into the same language space following standard practice [40, 54]. As illustrated in Figure 1(a), the same caption can sometimes be used to describe visually diverse inputs, it is thus essential for Vamos to be able to leverage one or multiple representations simultaneously, to understand the impact of individual representation type. Vamos can directly leverage an LLM's next token prediction capability for action anticipation [17]. We also ask Vamos to perform video question answering [75], by appending the question to the video representation as inputs to the LLM reasoner.

One inherent benefit of text-based video representation is its interpretability. Inspired by interpretable object classifiers such as concept bottleneck models (CBM) [36], we aim to understand which words serve as important evidence for question answering. However, CBM requires a pre-defined list of discrete visual concepts, and requires a linear classifier to achieve interpretability. We generalize this framework and learns hard attention to select a small subset of text inputs to feed to the LLM reasoner, where the text inputs are tokenized text as opposed to pre-defined concepts. We call our generalized formulation token bottleneck models (TBM). TBM naturally supports the incorporation of multimodal information, and allows users to perform causal intervention.

We perform extensive evaluations on five benchmarks, including the Ego4D dataset [17] for long-term action anticipation, NeXT-QA [75] and IntentQA [39] for video question answering, Spacewalk-18 [37] for long-form procedural video understanding, and EgoSchema [50] for zero-shot long-form video question answering. We observe that for the direct application of Vamos in the action anticipation task, the text-based representation outperforms its counterpart based on visual embeddings. We further observe that free-form video descriptors serve as an effective long-video representation that generalizes well in zero-shot setting, outperforming the strongest video-language model [71] by 66%. We then confirm that our observations are general, that text-based representation consistently provides competitive performance across all tasks, and that adding visual embeddings surprisingly results in marginal or no performance gains. Finally, we demonstrate that the token bottleneck model is able to select semantically relevant evidence for question answering, and achieves 5x speedup at inference time while maintaining the question answering accuracy.

2   Related Work

Vision-Language Foundation Models. Models such as CLIP [59] and ALIGN [29] bridge the vision and language modalities by learning a text encoder and an image encoder jointly with a contrastive loss on image and caption pairs. Another

line of vision-language models [10,43,45,63] combines the masked language modeling objective with image-text contrastive learning, and focus on downstream tasks such as visual question answering, visual commonsense reasoning, and text-guided object detection. For videos, VIOLET [16] trains an end-to-end transformer for video and language modalities, by representing videos as visual tokens and performing joint masked token modeling. To perform visual-language joint training, speech transcripts are often used as the language modality for videos [16, 64, 85, 86]. The objectives can be combined [82] and the encoders for different modalities can be shared [69]. Compared to existing VLMs, Vamos imposes an "information bottleneck" when text-based representation is used: It converts visual inputs into discrete action labels and free-form text descriptions.

Visually-augmented LLMs. Apart from joint visual-language pre-training, existing large language models (LLMs) can also be augmented to incorporate visual inputs. For example, VisualGPT [9] and Flamingo [2] directly fuse visual information into the layers of a language model decoder using a cross-attention mechanism instead of using images as additional prefixes to the language model. Other approaches, such as instructional tuning [48], prompting large language models for knowledge retrieval [62], or linearly projecting the visual embeddings into the input space of LLMs [46, 54], have also been explored. Vamos largely follows this approach to incorporate visual embeddings, with the goal to understand if and how they are complementary to text-based video representations.

Additionally, tool-using large language models have been recently proposed to invoke and incorporate the use of task-specific modules [23, 61], where visual perceptions consist a substantial subset of the tools. Notably, VisualProgram [20] and ViperGPT [65] propose to apply LLMs to generate symbolic programs based on pre-selected computer vision modules for visual question answering. VidIL [72] leverages expert knowledge to design object and action concepts for few-shot captioning and video question answering. Closest to our work is Socratic Models [87], where the authors propose to use natural language as the common interface to connect foundation models with different input and output modalities. Finally, several concurrent works [42,51,88] share similar motivations and methods to Vamos in utilizing text-based representations and modules for video understanding, without the incorporation of visual embeddings.

3 Method

We now describe how the text-based representation is constructed and incorporated into versatile action models.

3.1 Text-based Video Representation

A video often contains complex and dynamic information including context and interactions. While prior works [36, 73, 84] have demonstrated the effectiveness of condensing images into text-based representations such as visual concepts, it remains unclear if videos can also be condensed into text-based representations.

To answer this research question, we consider text descriptions that are task-agnostic, and can potentially be applied in diverse video understanding tasks.

Concretely, we rely on general-purpose captioning models to generate free-form text descriptions to characterize objects, scenes, and actions, which succinctly summarize the essential elements depicted in the video. We employ off-the-shelf image captioning models such as BLIP-2 [41] that generate image-level captions from the sampled video frames. These captions are subsequently concatenated to form a comprehensive video-level caption.

For certain tasks, prior knowledge might be helpful to guide the model learning. For example, when the goal is to model the long-term temporal dynamics of verbs and nouns for the long-term action anticipation task, it would be beneficial to trim the inputs to only contain discrete action labels. In practice, this can be achieved through the application of action recognition models such as Transformer encoders that operate in the pre-defined action space.

3.2   Versatile Action Models

Large language models (LLMs) have demonstrated strong capability for temporal reasoning [90] and even some potential for causal reasoning [34], both of which are crucial for video understanding. We introduce Vamos, a simple yet effective framework to utilize LLMs to unify video dynamic modeling tasks, including comprehending historical content (video question answering, VQA) and future prediction (long-term action anticipation, LTA). As shown in Figure 1 (b), given a video V and a pretrained LLM π, the input sequence $\mathbf{x}_t = [\mathbf{x}_{\text{tvr}}, \mathbf{x}_{\text{task}}]$ consists of the textual video representations $\mathbf{x}_{\text{tvr}}$ of V and other task specific language inputs $\mathbf{x}_{\text{task}}$ (e.g., instructions, questions, targets). The frozen word embedding layer $\mathcal{F}_{\text{emb}}$ first generate the corresponding text tokens $\mathbf{z}_t = \mathcal{F}_{\text{emb}}(\mathbf{x}_t) \in \mathbb{R}^{L_t \times D}$, where L_t is the sequence length of $\mathbf{x}_t$, D is the feature dimension.

Vamos incorporates the *residual* information not entirely captured by $\mathbf{x}_{\text{tvr}}$ via representations encoded directly from the visual modality, such as CLIP visual embedding. We adopt a learnable linear projection layer $\mathcal{F}_{\text{proj}}$ to align visual features with the language space. Specifically, the frozen vision backbone $\mathcal{E}$ takes in N_v frames $[v_1, ... v_{N_v}]$ sampled from V to generate the visual features. These visual features are then fed into the projection layer $\mathcal{F}_{\text{proj}}$ to produce visual tokens $\mathbf{z}_v = \mathcal{F}_{\text{proj}}(\mathcal{E}(v_1, ... v_{N_v})) \in \mathbb{R}^{N_v \times D}$. To combine information from the visual and textual representations, we adopt the early fusion strategy and concatenate $\mathbf{z}_v$ and $\mathbf{z}_t$ as the inputs to the LLM π. When labeled training data is available for task-specific fine-tuning, we update the weights of the LLM π either with LoRA [22] or LLaMA-Adapter [89].

Vamos can accommodate diverse video understanding tasks by formulating each task as sequence completion given an appropriate task description $\mathbf{x}_{\text{task}}$, the LLM π can then be optimized with the standard language modeling objectives. Specifically, for the VQA task, $\mathbf{x}_{\text{task}}$ is composed of instructions, questions, and answers, with the answer being the training objective. During inference, the answer that maximizes sequence modeling likelihoods is selected for multiple-choice QA, or directly generated for open-ended QA. For the LTA task, $\mathbf{x}_{\text{task}}$
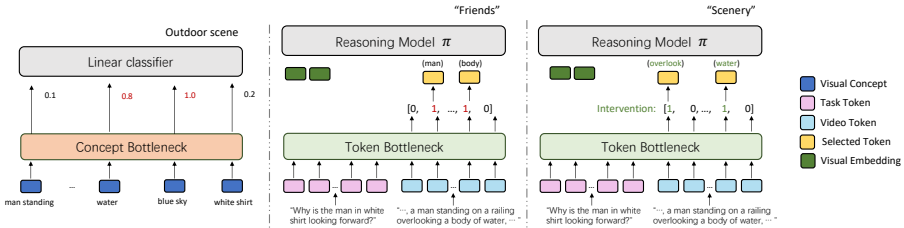
Fig. 2: An illustration of the token bottleneck model (TBM). We are inspired by the concept bottleneck models (CBM) [36], which achieve interpretable object classification by inspecting the weights of the learned linear classifier (left). Unlike CBM, Vamos does not require pre-defining a list of concepts. It directly works with tokenized text inputs. To provide input tokens to the reasoning model (an LLM), we leverage hard attention to generate binary rather than continuous weights (middle). The token bottleneck can be interpreted directly. It can also be intervened with human inputs (right), or augmented with residual visual information.

is composed of instructions and future actions, where the training objective is the future action sequence. During inference, the fine-tuned LLM is tasked to generate sequences of future actions based on the history actions.

Discussion: By design, Vamos naturally incorporates temporal information as the captions are timestamped. Since the text-based representation is general-purpose, the framework is efficient since once extracted, the text-based representation can be reused for different questions, just like the CLIP embeddings commonly used by VLMs. Vamos is a two-stage framework with decoupled "perception" and "reasoning" modules, in addition to the benefits on interpretability (of the intermediate representation) and efficiency (reuse of the intermediate features). Another conceptual benefit is its generalizability: While the visual distributions of different datasets may differ, the reasoning module may be shared. We show this is indeed the case for EgoSchema [50], when the end-to-end vision-language models performance significantly worse than Vamos.

3.3   Token Bottleneck Models

We aim to understand how the text-based representations are utilized by the LLM reasoners. As an LLM operates like a black box, we aim to enhance the interpretability of the overall framework by first understanding how it selects evidence to solve the downstream tasks. As illustrated in Figure 2, we are inspired by the success of interpretable object classifiers, such as the concept bottleneck model (CBM) [36]. CBM relies on pre-defining a list of concepts and building (often supervised) concept detectors for each of them. As our model's inputs are free-form text, we propose to directly work with word tokens as opposed to pre-defined concepts. In addition, CBM relies on linear classifiers to achieve model interpretability. Each weight in a learned classifier indicates the impor-

tance of the corresponding concept for making the prediction. We hypothesize that linear classifiers are not sufficiently expressive when solving tasks that require a stronger "reasoner". To strike a balance between model interpretability and expressiveness, we generalize the CBM framework to learn binary attention on the input tokens as opposed to continuous weights used by the CBM linear classifier. As illustrated in Figure 2, The binary weights indicate which tokens are to be selected and fed to the more expressive LLM for solving the target tasks. We name this generalized framework as token bottleneck models (TBMs).

To implement TBM, we design a lightweight token selector as an add-on module for Vamos (Figure 2 middle). It takes the tokenized embeddings for the text-based video representations $\mathbf{z}_{\text{tvr}}$ and the task-specific tokens $\mathbf{z}_{\text{task}}$ as its inputs. It learns to pick a single token among the candidate tokens $\mathbf{z}_{\text{tvr}}$ by optimizing the objective for a given downstream task. To select a sequence of tokens, we assume the important information is even distributed across the input sequence, and uniformly divide the input sequence into k segments $\{\mathbf{z}_{\text{tvr}}^{(1)}, ..., \mathbf{z}_{\text{tvr}}^{(k)}\}$, each of which contains n tokens. Each segment $\mathbf{z}_{\text{tvr}}^{(i)} = \{z_1^{(i)}, ..., z_n^{(i)}\}$ is fed into the token selector, from which one token $z^{(i)}$ is selected for the task $\mathbf{z}_{\text{task}}$.

Within the token selector, $\{z_1^{(i)}, ..., z_n^{(i)}\}$ are first projected to a lower dimension, and then provided as inputs to a shallow transformer encoder to obtain encodings $\{s_1^{(i)}, ..., s_n^{(i)}\}$. A linear layer then takes these encodings and generates the logits $\mathbf{g}^{(i)} \in \mathbb{R}^n$ for final selection. During training, we apply Gumbel-Softmax [27] on the logits $\mathbf{g}^{(i)}$ to pick the token $z^{(i)}$ for each segment $\mathbf{z}_{\text{tvr}}^{(i)}$ while ensuring the module is differentiable. In this way, k tokens are selected as the condensed representation of the original tokenized input sequence $\mathbf{z}_{\text{tvr}}$.

The token selector in TBM allows us to inspect the important evidence selected for the downstream tasks, and to intervene the wrongly recognized or selected tokens with the correct ones with human in the loop (Figure 2 right). Practically, the token selector also can also speed up the inference time due to its own light-weight implementation, and that only a much smaller subset of the tokens (e.g. 6%) are processed by the computationally heavy LLM.

4   Experiments

In this section, we conduct experiments on two tasks and four datasets with both quantitative and qualitative analysis.

4.1   Task and Datasets

Long-term action anticipation. The LTA task asks a model to predict a sequence of actions in form of a verb-noun pairs in a long future window based on video observations of the past actions. In LTA, a long video V is first split into a number of annotated video segments. Given video observation before segment i, our task is to predict the future actions in sequences of verb-noun pairs of the next Z segments allowing K candidate sequences. The correctness of the predicted sequence is measured with edit distance. We evaluation on:

Ego4D [17] is comprised of 3,670 hours of egocentric videos in hundreds of scenarios of daily life activity. The Ego4D LTA v2 benchmark we focus on includes a total duration of around 243 hours of videos annotated into 3472 clips with 117 verbs and 521 nouns. We follow the official dataset splits and adopt the official parameters of the evaluation metric, with $Z = 20$ and $K = 5$.

Video question answering. Given a set of videos V, and a corresponding set of language-based questions Q_v and their candidate answers A_q. The goal of video question answering (VQA) task is to predict the correct answer A for each video-question pair. The performance is measured by accuracy. For VQA, we evaluate on three datasets:

EgoSchema [50] is annotated on Ego4D videos for long-form video QA. Each video is around 3 minutes and the *temporal certificate* for humans to solve each task is around 100 seconds. It has 5,031 videos and each video is annotated with a multiple-choice question. All examples are for zero-shot evaluation.

Spacewalk-18 [37] is a long-form procedural video understanding benchmark collected on 18 spacewalk videos. The total duration is over 96 hours. We evaluate on the step recognition task which has a temporal certificate of 140 seconds. We follow the zero-shot setup with 1-minute context window, and report step recognition accuracy on the test set.

NeXT-QA [75] is a popular multiple choice video question answering benchmark that tests video understanding in terms of describing and reasoning the temporal actions. It contains 5,440 video clips and 47,692 questions, grouped into causal (48%), temporal (29%) and descriptive (23%).

IntentQA [39] is a multiple choice VQA dataset built on top of NeXT-QA but focuses on intent reasoning. The authors select the videos related to causal and temporal questions from NeXT-QA and constructed their own questions and answers to focus on testifying models's performance on reasoning questions.

4.2   Implementation

Generating Action Labels. To generate action labels for videos, we use a recognition model pretrained on Ego4D LTA. It is a 3-block 6-head transformer-encoder that takes 4 CLIP features and outputs two logits for verb and noun respectively. It predicts actions in the action space pre-defined by Ego4D LTA. For each video, our action recognition model samples 4 frames for each 8s segment splits uniformly and output a verb and noun pair for the segment.

Generating Video Captions. We generate zero-shot, task-agnostic video captions using BLIP-2 for Ego4D LTA, IntentQA, and EgoSchema. We use LLaVA-1.5 to generate captions for NeXT-QA and Spacewalk-18 due to its better performance (see Table 5). For Ego4D LTA, we sample the center frame for each video segment to generate its caption. The captions for the 8 observed segments are then concatenated as the video representations. For VQA benchmarks, we first uniformly sample a fixed number of frames for each video, and then generate and concatenate the frame-level captions. We sample 6, 6, 12, and 12 frames for NeXT-QA, IntentQA, Spacewalk-18, and EgoSchema, respectively. For Spacewalk-18, we additional use its provided speech narrations.

Vamos for Temporal Modeling. For full-shot VQA and LTA, We use LLaMA-7B and LLaMA2-7B [67] respectively as the temporal model for video understanding. During training, we use LLaMA Adapter [89] or low-rank adaption [22] (LoRA) to perform parameter-efficient fine-tuning on the training set. For vision input, we use the frozen CLIP [59] ViT-L/14 to extract image features. For the zero-shot long-form VQA on EgoSchema, we use several popular LLMs including OpenAI GPT-3.5-turbo, GPT-4, GPT-4o and LLaMA2-chat-13B.

Table 1: Vamos with various video representations. Results are reported on Ego4D LTA test set, the metric is edit distance. Results are reported on NeXT-QA validation set, and IntentQA test set, the evaluation metric is accuracy. On all datasets, text-based representations achieve competitive performance.

Input	Ego4D-LTA ↓			NeXT-QA ↑				IntentQA ↑			
	Verb	Noun	Action	Cau.	Tem.	Des.	All	CW	CH	TP&TN	All
vision	0.653	0.673	0.884	69.6	67.2	74.7	69.6	68.9	71.6	58.0	66.7
text	0.661	0.651	0.878	**75.5**	**71.3**	81.1	**75.0**	**74.0**	**78.6**	**67.5**	**73.2**
vis+text	**0.643**	**0.650**	**0.868**	74.5	71.0	**81.7**	74.5	73.5	76.6	64.3	71.7

4.3   LLM as Long-term Video Temporal Reasoner

We first apply Vamos on the long-term action anticipation task, which requires direct modeling of video temporal dynamics by predicting future actions tokens based on video observation. We fine-tune Vamos on Ego4D LTA dataset with continuous visual embeddings or text-based representation. We observe that action-based representation slightly outperforms the free-form captions due to the nature of the task. As shown in the first three columns of Table 1, we observe that the text-based representation outperforms the vision-based input, and combining the two further improves the performance.

Table 2: zero-shot VQA on Egoschema. *: 500 question subset.

Model	Input Type	Acc.
InternVideo [71]	frame	32.1%
GPT-4	text	**48.26%**
GPT-4*	gt-narration	81.80%

Table 3: Ablation on the number of frames on EgoSchema with GPT-3.5 turbo.

# Frames	Full Set Acc.
1	37.83%
4	38.36%
12	41.24%

We now consider the more challenging task, long-form video question answering, by conducting zero-shot experiment on the recently collected EgoSchema benchmark. We employ OpenAI GPT-4 as the video "reasoner". We observe that free-form captions extracted by BLIP-2 consistently outperform the action-based representation, presumably because richer information is retained in the captions. We report the free-form caption-based performance from now on, unless otherwise mentioned. From Table 2, we observe that Vamos with text-

based video representation largely outperforms the state-of-the-art InternVideo model [71], which is jointly trained on vision and language inputs. We attribute the performance gain due to the decoupling of perception and reasoning, the latter of which we hypothesize is easier to generalize even at zero-shot, thanks to the LLM pre-training. To better understand the performance upper-bound of the text-based representation, we use 500 ground-truth video narrations provided by the authors of EgoSchema to perform evaluation on this subset. Remarkably, the LLM achieves an impressive accuracy of 81.8% with the oracle captions. Although not directly comparable with the full-set performance, the result confirms the potential of the Vamos framework for reasoning over broad time spans in the long-form video question answering task, and that better empirical performance may be achieved by improving the captioning models.

4.4   What Makes Good Video Representation?

In addition to text features, Vamos can also integrate vision-language features. We then investigate whether different modalities encode complementary information on Ego4D, IntentQA, and NeXT-QA. We perform parameter efficient fine-tuning to update the weights of the LLM, whose gradients are used to jointly train the linear projection layer to incorporate the visual embeddings. We observe that naively fine-tuning with text and visual inputs lead to model overfitting, and hence perform modality dropout (i.e. randomly discard the entire sequence of visual embeddings) when fine-tuning the vis+text models.

While visual and text inputs are complementary in the Ego4D LTA task, we observe in Table 1 that the caption-based representation significantly outperform the vision-only baseline in the NeXT-QA and IntentQA benchmarks for video question answering, and that adding visual features only marginally affects the performance. This suggests that pre-trained visual embeddings such as CLIP may not encode residual information useful for the video QA task (e.g. fine-grained visual and motion features). Extracting visual representations that would complement the information encoded by the task-agnostic captions is thus an important future work.

4.5   Token Bottleneck Models

Although the text-based video representation is directly interpretable, it remains unclear which tokens are selected as evidence for making predictions. The token bottleneck models (TBMs) enable us to reveal the selected evidence and also improve the model's inference speed.

VQA Performance. We first tokenize and embed the input sequence $\mathbf{x}_t$ to obtain $\mathbf{z}_t$, which is uniformly partitioned into $k = 20$ or 40 segments. TBM is then applied on each segment to select one token. In this way, we condense a long sequence (644 tokens on average for NeXT-QA) into k tokens. TBM is jointly optimized with Vamos with LLaMA-3 during training.

For NeXT-QA dataset, we follow previous setting to use captions from 6 frames generated by LLaVA-1.5. For comparison, we also show the vision-only
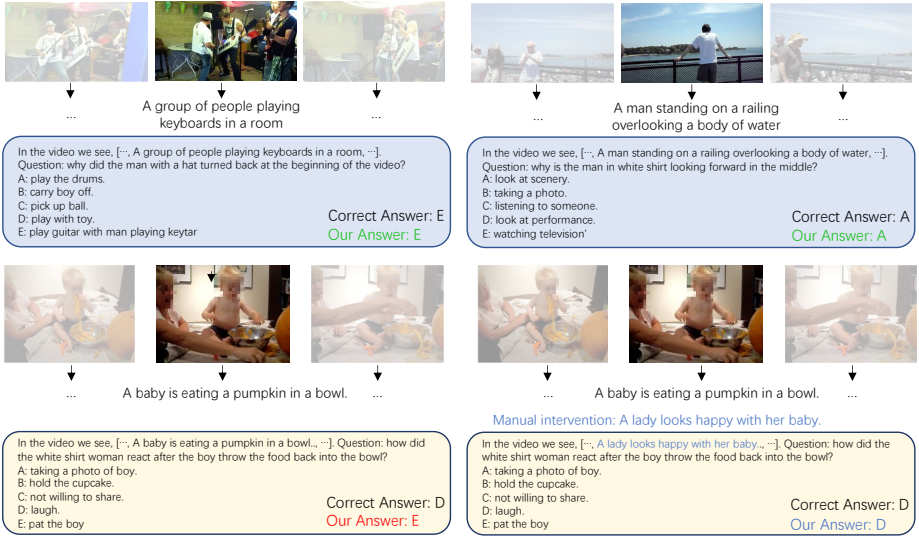
Fig. 3: Visualizations of example Vamos predictions and manual intervention.

performance taking in 12 frames and the performance with unselected caption-based input. Table 4 shows that the TBM leads to an expected performance drop due to discarding over 90% of the input tokens, and increasing k improves the TBM performance. When $k = 40$, Vamos achieves a competitive 69.6% accuracy, while only adding 0.7M parameters to the Vamos framework and achieving **5x inference speedup** from 1.41s to 0.29s per sample on a single A6000 GPU.

Table 4: Condensing captions with token bottleneck models on NeXT-QA.

Input	Selected tokens (k)	Cau.	Tem.	Des.	All
vision	all	71.9	67.4	75.6	71.0
text	all	77.2	75.3	81.7	77.3
text	20 / 644	68.1	64.6	70.8	67.4
text	40 / 644	70.1	67.4	72.2	69.6

Visualization of Vamos predictions. Figure 3 provides two positive and one negative examples selected from IntentQA. The predictions are made by Vamos without TBM. From the two positive examples we can see that the generated captions manage to describe the scene and activities happening in the video ("overlooking water" and "playing keyboards"), thus providing strong clues for LLMs to answer the question about description ("look at scenery") and reasoning ("play guitar with man playing keytar"). However, a sub-optimal caption will also

Fig. 4: Illustration of the token bottleneck model and the impact of intervention on model's predictions. The selected tokens are highly related to the task (examples *a* and *b*) and can be manually intervened to correct the model's prediction (example *c*).

cause the Vamos's failure of reasoning. In the third example, we observe that the captions successfully describe the baby's action of eating but fail to describe the presence and potential actions of a woman in the corner of the scene. This omission leads to an incorrect prediction regarding the woman's reaction.

Visualization of the Selected Tokens. TBM not only helps accelerate Vamos' inference speed, but also allows us to interpret which pieces of evidence are used by the model to make predictions. In Figure 4, we show two positive examples and one negative example of Vamos with TBM from NeXT-QA. The two positive examples demonstrate Vamos' ability to select tokens that serve as evidence for question answering. The negative example is challenging: even highly related tokens such as "snow" and "glide" are selected by Vamos, it still fails to reason the correct activities in the video.

Test Time Intervention. Text-based video representation is not only directly interpretable as used by Vamos and TBM, it also allows the users to conduct post-hoc, test-time intervention, which is pivotal for diagnosing and fixing failed predictions without retraining. In Figure 3 and Figure 4, we show examples of test-time intervention performed on the negative examples for Vamos and Vamos

with TBM, respectively. By providing more accurate and related captions or tokens, Vamos is able to correct its wrong predictions.

Table 5: Ablation on captioning model and frame numbers on NeXT-QA.

Caption	# Frame	Cau.	Tem.	Des.	All
LLaVA-13B	1	69.9	67.2	74.8	69.8
LLaVA-13B	3	73.0	70.4	79.4	73.1
LLaVA-13B	6	**75.5**	71.3	81.1	75.0
LLaVA-7B	6	75.2	**71.9**	**81.6**	**75.1**
BLIP-2	6	72.7	68.9	78.8	72.4

Table 6: Comparison of different LLaMA models on NeXT-QA.

Model	Cau.	Tem.	Des.	All
LLaMA1-7B	75.5	71.3	81.1	75.0
LLaMA2-7B	74.8	72.3	81.6	75.0
LLaMA3-8B	**77.2**	**75.3**	**81.7**	**77.3**

4.6   Design Choices and Ablation Study

Caption Models. We study the impact of caption models on the video QA performance on NeXT-QA. We compare two captioning models: BLIP-2 and LLaVA-1.5. We observe that captions generated by BLIP-2 are generally more concise (less than 20 tokens), while captions generated by LLaVA-1.5 are more detailed (around 100 tokens on average). Results shown in Table 5 shows that captions from LLaVA-1.5 achieve better performance. We also investigate the influence of caption model size by comparing LLaVA-1.5 7B and 13B versions. Interestingly, scaling LLaVA from 7B to 13B does not lead to improvement.
Number of Frames. We study the impact of sampled frame numbers for captioning on NeXT-QA. As shown in Table 5, we found that using more captioned frames leads to better performance. A similar trend can be observed on EgoSchema in Table 3. However, we observe diminishing return when 12 frames are used, and hence not worth the speed and accuracy trade-off.
Impact of LLMs on VideoQA. We study the impact of different LLaMA versions on NeXT-QA. As shown in Table 6, Vamos directly benefits from a more advanced LLM, which is a desirable property for practitioners.
Impact of LLMs on Long-form VideoQA. We study the impact of LLMs on the EgoSchema benchmark. As shown in Table 10, GPT-4o achieves significant improvements comparing with GPT-4 and LLaMA2-Chat-7B baselines, which again demonstrates that Vamos can directly benefit from advances in LLMs.

4.7   Comparison with State-of-the-art

We compare our proposed Vamos with other state-of-the-art models in Tables 7, 8, 9, and 10. We train Vamos with LLaMA2-7B on Ego4D LTA and LLaMA3-8B on NeXT-QA and IntentQA. On EgoSchema zero-shot VQA, our approach based on GPT-4o outperforms the best vision-language model by 66.8% on the full set. On NeXT-QA, Vamos achieves the best performance and significantly outperforms LLaMA-VQA [35] and SeViLA [83], even though the former uses a LLM with 33B parameters and the latter is trained on additional dataset with temporal localization supervision. On IntentQA, Vamos also outperforms all baselines, with a 28.7% accuracy improvement compared to the best performing prior method. On Ego4D LTA, Vamos outperforms previous works, without

relying on domain-specific video encoders [47]. Finally, on Spacewalk-18, we use GPT-4o as the LLM and achieves 18.6% accuracy, which significantly outperforms the prior best zero-shot performance of 13.6%.

Table 7: Comparison on NeXT-QA benchmark. * with additional supervision.

Model	Cau.	Tem.	Des.	All
ATP [5]	53.1%	50.2%	66.8%	54.3%
HiTeA [80]	62.4%	58.3%	75.6%	63.1%
Intern Video [71]	62.5%	58.5%	75.8%	63.2%
BLIP-2 [41]	70.1%	65.2%	80.1%	70.1%
SeViLA* [83]	74.2%	69.4%	81.3%	73.8%
LLaMA-VQA-7B [35]	72.7%	69.2%	75.8%	72.0%
LLaMA-VQA-33B [35]	76.2%	72.6%	78.8%	75.5%
Vamos (ours)	**77.2%**	**75.3%**	**81.7%**	**77.3%**

Table 8: Comparison with SOTA on IntentQA.

Model	CW	CH	TP&TN	ALL
HGA [30]	44.88%	50.97%	39.62%	44.61%
HQGA [76]	48.24%	54.32%	41.71%	47.66%
VGT [77]	51.44%	55.99%	47.62%	51.27%
BlindGPT [56]	52.16%	61.28%	43.43%	51.55%
CaVIR [39]	58.4%	65.46%	50.48%	57.64%
Vamos (ours)	**75.14%**	**77.44%**	**69.58%**	**74.16%**

Table 9: Comparison with SOTA on Ego4D LTA v2 test set.

Model	verb	noun	action
Slowfast [15]	0.717	0.736	0.925
VideoLLM [8]	0.721	0.725	0.921
PaMsEgoAI [25]	0.684	0.679	0.893
Palm [24]	0.696	0.651	0.886
AntGPT [90]	0.650	**0.650**	0.877
Vamos (ours)	**0.643**	**0.650**	**0.868**

Table 10: Egoschema VQA zero-shot performance on full set and subset.

Model	Input Type	Full	Subset
FrozenBiLM [79]	frame	26.9%	-
mPLUG-Owl [81]	frame	31.1%	-
InternVideo [71]	frame	32.1%	-
Vamos (LLaMA2-13B)	caption	36.73%	38.20%
Vamos (GPT-3.5)	caption	41.24%	47.60%
Vamos (GPT-4)	caption	48.26%	51.20%
Vamos (GPT-4o)	caption	**53.55%**	**57.20%**

5   Conclusion

We study different forms of video representations and propose versatile action models (Vamos) as a unified framework to utilize visual- and text-based representations for video understanding. We conduct extensive experiments on long-term action anticipation and video question answer benchmarks. Surprisingly, we observe that direct applications of free-form, general-purpose text-based video representations, such as captions, serve as strong video representation for all benchmarks we consider. Vamos utilizes large language models to perform zero-shot reasoning, and incorporate human feedback via test-time intervention. We further propose the token bottleneck models, which allow the users to interpret the evidence selected by Vamos, and speed up inference by nearly 5x. Vamos achieves state-of-the-art results on Ego4D LTA, IntentQA, NeXT-QA, Spacewalk-18, and outperforms the best vision-language model by over 66% on EgoSchema.

Limitations: Although our results show the promise of free-form text-based representations, we believe visual information is still essential for complex video understanding and reasoning. We expect future work to investigate alternative visual encoders [66] to extract fine-grained visual information beyond what has been captured by the captions, and to propose pre-training paradigms that better align visual inputs with the input space of LLMs. We also believe better benchmarks that require fine-grained, structured visual understanding are needed to rigorously evaluate the impact of representations for video understanding.

Acknowledgements

References

1. Aein, M.J., Aksoy, E.E., Wörgötter, F.: Library of actions: Implementing a generic robot execution framework by using manipulation action semantics. The International Journal of Robotics Research (2019)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. In: NeurIPS (2022)
3. Allen-Zhu, Z., Li, Y.: Physics of language models: Part 1, context-free grammar. arXiv preprint arXiv:2305.13673 (2023)
4. Boix-Adserà, E., Saremi, O., Abbe, E., Bengio, S., Littwin, E., Susskind, J.: When can transformers reason with abstract symbols? arXiv preprint arXiv:2310.09753 (2023)
5. Buch, S., Eyzaguirre, C., Gaidon, A., Wu, J., Fei-Fei, L., Niebles, J.C.: Revisiting the "video" in video-language understanding. In: CVPR (2022)
6. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: CVPR (2015)
7. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
8. Chen, G., Zheng, Y.D., Wang, J., Xu, J., Huang, Y., Pan, J., Wang, Y., Wang, Y., Qiao, Y., Lu, T., et al.: Videollm: Modeling video sequence with large language models. arXiv preprint arXiv:2305.13292 (2023)
9. Chen, J., Guo, H., Yi, K., Li, B., Elhoseiny, M.: Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In: CVPR (2022)
10. Chen, L.H., Zhu, Y., Shen, Y.C., Gao, H., Liu, X., Shen, X., He, Z., Henao, R., Miao, R., Guo, Y., et al.: Uniter: Universal image-text representation learning. In: ECCV (2020)
11. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: ECCV (2018)
12. Dasgupta, I., Lampinen, A.K., Chan, S.C., Creswell, A., Kumaran, D., McClelland, J.L., Hill, F.: Language models show human-like content effects on reasoning. arXiv preprint arXiv:2207.07051 (2022)
13. Ding, D., Hill, F., Santoro, A., Reynolds, M., Botvinick, M.: Attention over learned object embeddings enables complex visual reasoning. In: NeurIPS (2021)
14. Epstein, D., Wu, J., Schmid, C., Sun, C.: Learning temporal dynamics from cycles in narrated video. In: ICCV (2021)
15. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV (2019)

16. Fu, T.J., Li, L., Gan, Z., Lin, K., Wang, W.Y., Wang, L., Liu, Z.: Violet: End-to-end video-language transformers with masked visual-token modeling. arXiv preprint arXiv:2111.12681 (2021)

17. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR (2022)

18. Gruver, N., Finzi, M., Qiu, S., Wilson, A.G.: Large language models are zero-shot time series forecasters. arXiv preprint arXiv:2310.07820 (2023)

19. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: CVPR (2018)

20. Gupta, T., Kembhavi, A.: Visual programming: Compositional visual reasoning without training. In: CVPR (2023)

21. Hongeng, S., Nevatia, R., Bremond, F.: Video-based event recognition: activity representation and probabilistic recognition methods. Computer Vision and Image Understanding (2004)

22. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)

23. Hu, Z., Iscen, A., Sun, C., Chang, K.W., Sun, Y., Ross, D.A., Schmid, C., Fathi, A.: Avis: Autonomous visual information seeking with large language models. arXiv preprint arXiv:2306.08129 (2023)

24. Huang, D., Hilliges, O., Van Gool, L., Wang, X.: Palm: Predicting actions through language models@ ego4d long-term action anticipation challenge 2023. arXiv preprint arXiv:2306.16545 (2023)

25. Ishibashi, T., Ono, K., Kugo, N., Sato, Y.: Technical report for ego4d long term action anticipation challenge 2023. arXiv preprint arXiv:2307.01467 (2023)

26. Ivanov, Y.A., Bobick, A.F.: Recognition of visual activities and interactions by stochastic parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence (2000)

27. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144 (2016)

28. Jayaraman, D., Ebert, F., Efros, A.A., Levine, S.: Time-agnostic prediction: Predicting predictable video frames. arXiv preprint arXiv:1808.07784 (2018)

29. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021)

30. Jiang, P., Han, Y.: Reasoning with heterogeneous graph alignment for video question answering. In: AAAI (2020)

31. Jiang, Y.G., Bhattacharya, S., Chang, S.F., Shah, M.: High-level event recognition in unconstrained videos. International journal of multimedia information retrieval (2013)

32. Kalakonda, S.S., Maheshwari, S., Sarvadevabhatla, R.K.: Action-gpt: Leveraging large-scale language models for improved and generalized zero shot action generation. arXiv preprint arXiv:2211.15603 (2022)

33. Ke, Y., Sukthankar, R., Hebert, M.: Event detection in crowded videos. In: ICCV (2007)

34. Kıcıman, E., Ness, R., Sharma, A., Tan, C.: Causal reasoning and large language models: Opening a new frontier for causality. arXiv preprint arXiv:2305.00050 (2023)

35. Ko, D., Lee, J.S., Kang, W., Roh, B., Kim, H.J.: Large language models are temporal and causal reasoners for video question answering. In: EMNLP (2023)
36. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: ICML (2020)
37. Krishnan, R.M., Tang, Z., Yu, Z., Sun, C.: Spacewalk-18: A benchmark for multimodal and long-form procedural video understanding in novel domains. arXiv preprint arXiv:2311.18773 (2023)
38. Lester, J., Choudhury, T., Kern, N., Borriello, G., Hannaford, B.: A hybrid discriminative/generative approach for modeling human activities. In: IJCAI (2005)
39. Li, J., Wei, P., Han, W., Fan, L.: Intentqa: Context-aware video intent reasoning. In: ICCV (2023)
40. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
41. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: bootstrapping language-image pretraining with frozen image encoders and large language models. In: ICML (2023)
42. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. arXiv preprint arXiv:2311.17005 (2023)
43. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
44. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: ICCV (2021)
45. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: ECCV (2020)
46. Lin, J., Yin, H., Ping, W., Lu, Y., Molchanov, P., Tao, A., Mao, H., Kautz, J., Shoeybi, M., Han, S.: Vila: On pre-training for visual language models. arXiv preprint arXiv:2312.07533 (2023)
47. Lin, K.Q., Wang, J., Soldan, M., Wray, M., Yan, R., XU, E.Z., Gao, D., Tu, R.C., Zhao, W., Kong, W., et al.: Egocentric video-language pretraining. In: NeurIPS (2022)
48. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
49. Liu, M., Tang, S., Li, Y., Rehg, J.M.: Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In: ECCV (2020)
50. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. arXiv preprint arXiv:2308.09126 (2023)
51. Min, J., Buch, S., Nagrani, A., Cho, M., Schmid, C.: Morevqa: Exploring modular reasoning models for video question answering. arXiv preprint arXiv:2404.06511 (2024)
52. Minderer, M., Sun, C., Villegas, R., Cole, F., Murphy, K.P., Lee, H.: Unsupervised learning of object structure and dynamics from videos. In: NeurIPS (2019)
53. Mirchandani, S., Xia, F., Florence, P., Ichter, B., Driess, D., Arenas, M.G., Rao, K., Sadigh, D., Zeng, A.: Large language models as general pattern machines. arXiv preprint arXiv:2307.04721 (2023)
54. Moon, S., Madotto, A., Lin, Z., Nagarajan, T., Smith, M., Jain, S., Yeh, C.F., Murugesan, P., Heidari, P., Liu, Y., et al.: Anymal: An efficient and scalable anymodality augmented language model. arXiv preprint arXiv:2309.16058 (2023)

55. Nevatia, R., Hobbs, J., Bolles, B.: An ontology for video event representation. In: CVPR Workshop (2004)
56. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback, 2022. URL https://arxiv. org/abs/2203.02155 **13** (2022)
57. Pastra, K., Aloimonos, Y.: The minimalist grammar of action. Philosophical Transactions of the Royal Society B: Biological Sciences (2012)
58. Pei, M., Jia, Y., Zhu, S.C.: Parsing video events with goal inference and intent prediction. In: ICCV (2011)
59. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
60. Sadanand, S., Corso, J.J.: Action bank: A high-level representation of activity in video. In: CVPR (2012)
61. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., Scialom, T.: Toolformer: Language models can teach themselves to use tools. arXiv preprint arXiv:2302.04761 (2023)
62. Shao, Z., Yu, Z., Wang, M., Yu, J.: Prompting large language models with answer heuristics for knowledge-based visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14974–14983 (2023)
63. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D.: Flava: A foundational language and vision alignment model. In: CVPR (2022)
64. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: ICCV (2019)
65. Surís, D., Menon, S., Vondrick, C.: Vipergpt: Visual inference via python execution for reasoning. In: ICCV (2023)
66. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. arXiv preprint arXiv:2401.06209 (2024)
67. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
68. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: NeurIPS (2016)
69. Wang, A.J., Ge, Y., Yan, R., Yuying, G., Lin, X., Cai, G., Wu, J., Shan, Y., Qie, X., Shou, M.Z.: All in one: Exploring unified video-language pre-training. In: CVPR (2023)
70. Wang, X., Farhadi, A., Gupta, A.: Actions˜ transformations. In: CVPR (2016)
71. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al.: Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191 (2022)
72. Wang, Z., Li, M., Xu, R., Zhou, L., Lei, J., Lin, X., Wang, S., Yang, Z., Zhu, C., Hoiem, D., et al.: Language models with image descriptors are strong few-shot video-language learners. In: NeurIPS (2022)
73. Wei, C., Liu, C., Qiao, S., Zhang, Z., Yuille, A., Yu, J.: De-diffusion makes text a strong cross-modal interface. arXiv preprint arXiv:2311.00618 (2023)
74. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: NeurIPS (2022)

75. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: CVPR (2021)
76. Xiao, J., Yao, A., Liu, Z., Li, Y., Ji, W., Chua, T.S.: Video as conditional graph hierarchy for multi-granular question answering. In: AAAI (2022)
77. Xiao, J., Zhou, P., Chua, T.S., Yan, S.: Video graph transformer for video question answering. In: ECCV (2022)
78. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021)
79. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Zero-shot video question answering via frozen bidirectional language models. In: NeurIPS (2022)
80. Ye, Q., Xu, G., Yan, M., Xu, H., Qian, Q., Zhang, J., Huang, F.: Hitea: Hierarchical temporal-aware video-language pre-training. In: ICCV (2023)
81. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)
82. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)
83. Yu, S., Cho, J., Yadav, P., Bansal, M.: Self-chained image-language model for video localization and question answering. arXiv preprint arXiv:2305.06988 (2023)
84. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. arXiv preprint arXiv:2205.15480 (2022)
85. Zellers, R., Lu, J., Lu, X., Yu, Y., Zhao, Y., Salehi, M., Kusupati, A., Hessel, J., Farhadi, A., Choi, Y.: Merlot reserve: Neural script knowledge through vision and language and sound. In: CVPR (2022)
86. Zellers, R., Lu, X., Hessel, J., Yu, Y., Park, J.S., Cao, J., Farhadi, A., Choi, Y.: Merlot: Multimodal neural script knowledge models. In: NeurIPS (2021)
87. Zeng, A., Attarian, M., Ichter, B., Choromanski, K., Wong, A., Welker, S., Tombari, F., Purohit, A., Ryoo, M., Sindhwani, V., et al.: Socratic models: Composing zero-shot multimodal reasoning with language. arXiv preprint arXiv:2204.00598 (2022)
88. Zhang, C., Lu, T., Islam, M.M., Wang, Z., Yu, S., Bansal, M., Bertasius, G.: A simple llm framework for long-range video question-answering. arXiv preprint arXiv:2312.17235 (2023)
89. Zhang, R., Han, J., Liu, C., Gao, P., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Qiao, Y.: Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:2303.16199 (2023)
90. Zhao, Q., Wang, S., Zhang, C., Fu, C., Do, M.Q., Agarwal, N., Lee, K., Sun, C.: Antgpt: Can large language models help long-term action anticipation from videos? In: ICLR (2024)

A Additional Experiments and Results

Textual representation format. On the Ego4D LTA task, we observe that explicitly regularizing the text-based inputs to contain only verbs and nouns lead to slightly improved performance (0.878 versus 0.890 action edit distance). We hypothesize that this is due to the nature of the task, which focuses solely on predicting future verbs and nouns. In Table A1, we again compare the action-based and caption-based text representations, but on the EgoSchema zero-shot VQA task. In contrast to the LTA performance, the caption-based representation performs much better, intuitively because solving the video question answering would require more details about the video, which can be provided by the general-purpose captions.

Frame number on EgoSchema. In Table 3 we compare different frame numbers and the zero-shot performance on EgoSchema. We uniformly sample 1, 4, and 12 frames from the videos and concatenate the captions of each frame to form the text video representations. We use GPT-3.5 as the reasoner. The results show that using more frames leads to better performance as more information and temporal evidence are provided.

Temporal ordering information. We investigate the influence of temporal information for long-form video understanding by shuffling the frame order when concatenating captions. In Table A2, we compare the ordered and shuffled 12-frame captions. Surprisingly, the performance drop (2%) by shuffling the video captions is not as significant as we expect. We suspect that LLMs may have strong capability of "auto-correcting" the order of the input sentences, even if they are shuffled.

Table A1: Comparison of action and caption for **zero-shot** VQA on Egoschema.

Model	Input Type	Full Set Acc.
GPT-4	action	38.12%
GPT-4	caption	**48.26%**

Table A2: Ablation on the influence of temporal information. 12 frames are sampled.

Shuffle	Full Set Acc.
✔	39.22%
	41.24%

B Addition Implementation Details

Vamos. We train Vamos on 4 A6000 GPUs for 2, 5, 10 epochs on Ego4D, IntentQA, and NeXT-QA respectively. For NeXT-QA, we set the maximum sequence length of pure vision input as 128, and 1200 for captions and captions + vision inputs, for IntentQA, the sequence lengths are set as 128 and 512 respectively.

Token bottleneck model. For the token bottleneck model (TBM), we use a 2-layer transformer encoder with 2 attention heads, and a hidden size of 128. No additional positional encoding is added by the token bottleneck model. In order to condense the input sequence for interpretability and select the most relevant tokens, the TBM is *task dependent*, namely taking the questions and

the candidate answers (when available) as inputs along with the video captions. The selected tokens, as opposed to their corresponding encoded embeddings by the token bottleneck, are fed into the LLM, as illustrated in Figure 2 (middle) in the main paper.

C   Additional Visualization

Figure A3 illustrates the input format for long-term action anticipation and video question answering tasks. In the LTA task, the task-specific inputs are the discrete action labels and the target output is a future action sequence. For supervised VQA, the task-specific input is composed of instructions, video representations (vision and text), question and choices. The target output is the chosen answer to the question. For zero-shot VQA on EgoSchema, the video representation only consists of the text descriptions. Figure A3 shows the prompt designs for all tasks.

Figure A4 shows four example video captions generated by LLaVA and BLIP-2 on the NeXT-QA dataset, respectively. We observe that in general the LLaVA-generated captions are longer and more detailed. Recall that Table 5 shows that the more detailed LLaVA captions also lead to higher VQA performance.

In Figure A1 and Figure A2, we provide additional prediction examples from Vamos and Vamos with the token bottleneck model respectively.
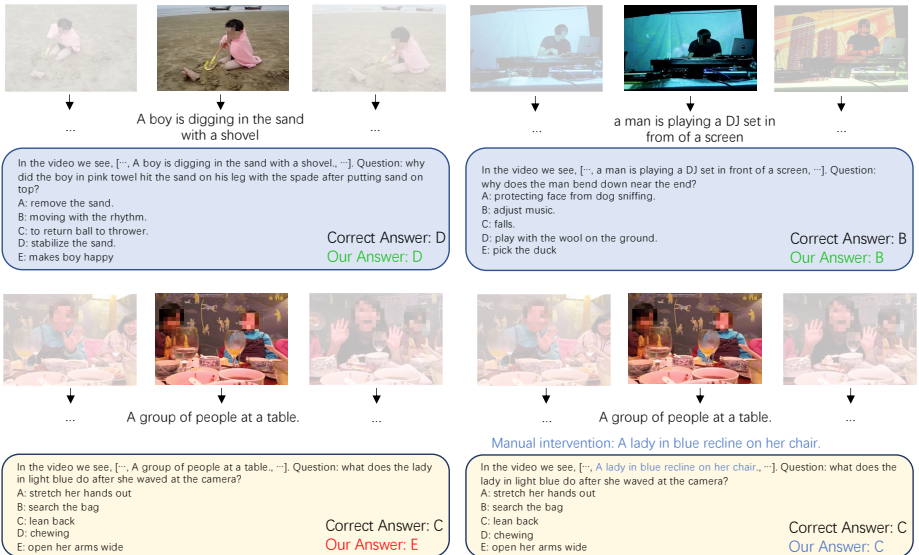

Fig. A1: More examples of Vamos video question answering and manual intervention.

(a) Positive Example

Frames

Question: "Why did the children throw stones into the sea? Choices: (A) playing on beach. (B) swimming around the pond. (C) excited and playing. (D) got on the board. (E) want to feel the water"
Tokens: ['**beach**', '**ocean**', 'cold', '**beach**', 'boy', '**beach**', 'girl', '**beach**', 'hand', 'closer', 'girl', 'beach', 'girl', 'back', '**beach**', '**beach**', 'girl', 'girl', '**beach**', '**beach**', 'red', '**crash**', '**ocean**', '**beach**', 'each', 'objects', 'closer', 'back', 'items', '**beach**', '**ocean**', 'cold', 'something', 'bag' '**beach**']
Prediction: (A) playing on beach.

(b) Positive Example

Frames

Question: "What does the guitarist do after shaking his right arm a few times at the start? Choices: (A) hat. (B) play the guitar. (C) white. (D) moves away from the microphone. (E) put guitar on table top"
Tokens: ['**guitar**', '**singing**', 'micro', 'stage', 'dark', '**singing**', 'micro', 'in', 'stage', 'in', '**playing**', '**guitar**', 'front', 'crowd', 'stage', 'in', 'I', 'micro', 'image', 'image', 'stage', '**singing**', 'micro', 'drum', '**musical**', 'drum', '**singing**', 'micro', 'stage']
Prediction: (B) play the guitar.

(c) Negative Example

Frames

Question: "Why does the person in yellow crouch over at the middle of the video? Choices: (A) touch the boy in stripes. (B) pick up something. (C) looking down at elephant. (D) talk to child. (E) to help"
Tokens: ['**baby**', 'car', 'play', 'room', 'icy', 'books', 'room', 'bow', 'play', '**baby**', 'keyboard', '**baby**', 'frame', 'floor', 'car', 'car', 'wheel', 'background', 'play', '**child**', 'play', 'car', 'play', 'background', 'ball', 'play', 'car', 'ride', 'baby', 'super', 'bow', 'other', 'car', 'car', 'book', '**baby**', 'closer', 'adding']
Prediction: (E) to help
Intervened Tokens: ['**baby**', 'car', 'play', 'room', 'icy', 'books', 'room', 'bow', 'play', '**baby**', 'keyboard', '**baby**', 'frame', 'floor', 'car', 'car', 'wheel', 'background', '**talk**', '**child**', 'play', 'car', '**talk**', 'background', 'ball', 'play', 'car', 'ride', '**child**', 'super', 'bow', 'other', 'car', 'car', 'book', '**child**', 'closer', 'adding']
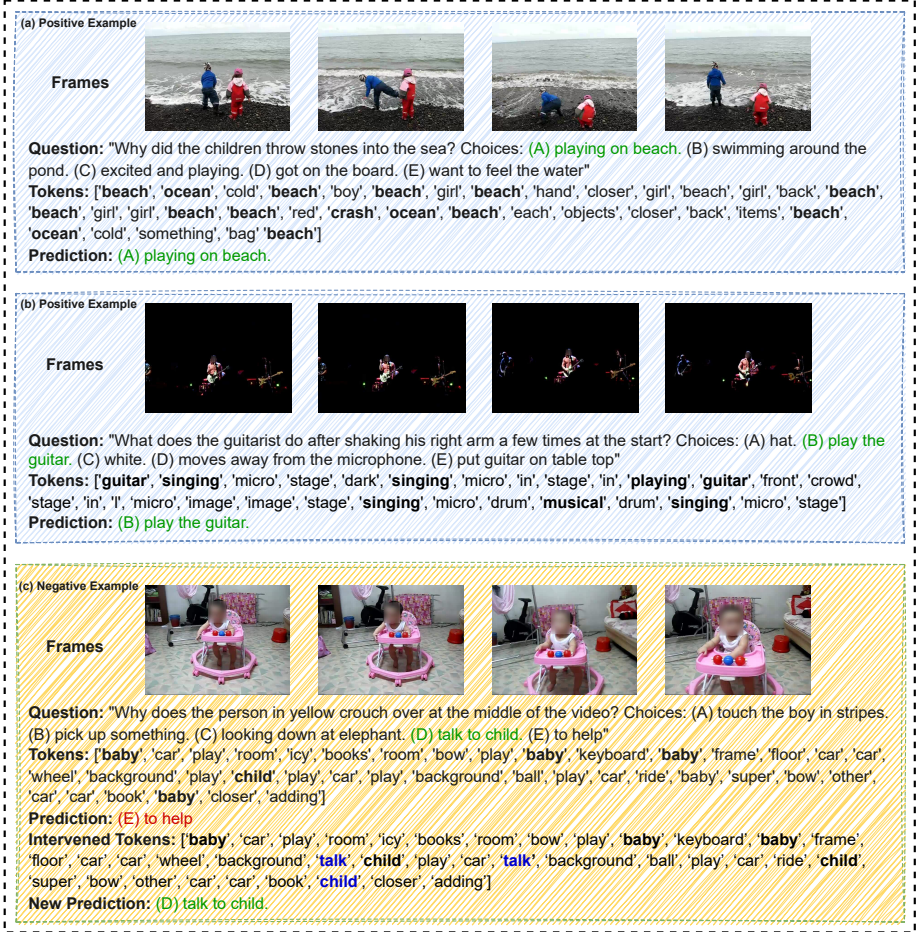**New Prediction:** (D) talk to child.

Fig. A2: Additional illustration of predictions with TBM and manual intervention. The token bottleneck selects tokens highly related to the question and intervention corrects wrong predictions without training.

Task 1: Long term action anticipation

Prompt:

| Touch Dough | Adjust Tray | Take Tray | Put Tray |

Completion:

lift sheet, hold dough, touch dough, touch dough, clean table, take bowl, touch bowl, put bowl, open cloth, fold cloth, put cloth, open door, open cabinet, touch dough, close door, close cabinet, move bowl, take bowl, put bowl, take cloth

Task 2: Video QA

Supervised

Question Text:

"Answer the question based on the observed scenes from a video:<v><v><v><v><v><v><v><v><v><v><v><v>
Video: a young child sitting on a chair, holding a green cup in their hands. The child appears to be playing with the cup, possibly enjoying a drink or a snack …<more text>…
Question: what did the baby do after throwing the green cup away while on the floor near the end?
Choices: (A) clap proudly. (B) the lady sitting down. (C) lay on floor. (D) just picked it up. (E) crawl
Answer:"

BLIP-2: "a baby sitting on a blue chair"

Target Result:

"(C) lay on floor"

Zero Shot

General Prompt Structure:

1. **Instruction**

"Suppose we have <narrations/actions> of a video describing <scenes/person/actions >. We would like to answer the choose the right option <number/letter> to answer the question correctly based on the <… >.
After {{<Correct Answer Number/Letter>:}}, you will predict the correct option <Number/Letter> from the 5 options in the form of a SINGLE <Number/Letter> with no text after. You always have to output just a single <Number/Letter> for the right option. <…> \n"

2. **Narrations + Question (Example from EgoSchema)**

"Narrations/Scene/Actions: <narrations/scene/actions>; Question: Although the video is predominantly focused on one recurring action, there is an interruption in c's activity. briefly describe this interruption and its significance within the video.; Answer Options: <options>; <Correct Answer Number/Letter>: "

Fig. A3: Input of Vamos for long-term action anticipation and video question answering. `<v>` in prompts denotes vision tokens.

Example 1: Both Model Captions Correct

BLIP-2 Caption:
"a person washing a dish in a sink"

LlaVA Caption:
"The image depicts a person washing a dirty dish in a kitchen sink. The person is using a sponge to clean the dish, which is placed in a bowl. The sink is located in the middle of the scene, with the person standing on the left side of the image.\n\nIn the background, there are two bottles, one positioned above the sink and the other one below it. Additionally, there is a cup placed near the top right corner of the image. The scene captures the everyday activity of cleaning dishes in a kitchen setting."

Example 2: Both Model Captions Correct

BLIP-2 Caption:
"a man is holding a piece of cloth in his hands"

LlaVA Caption:
"The image features a person holding a blue cloth in their hand, possibly preparing to fold or unfold it. The cloth is placed in the center of the scene, with the person's hand positioned above it. There are several other items in the image, including a handbag located towards the left side, a backpack in the middle-left area, and a book placed on the right side. Additionally, there is a cup situated near the top-right corner of the image. The scene appears to be a casual, everyday setting where someone is handling their belongings."

Example 3: LLaVA more accurate

BLIP-2 Caption:
"a woman is pouring rice into a pot"

LLaVA Caption:
"The image features a person in a green shirt, standing in a room and holding a large container filled with grains. They are in the process of pouring the grains into a smaller container, possibly a bowl. The person is using a spoon to scoop the grains and transfer them into the smaller container.\n\nThere are two bowls in the scene, one being the large container and the other being the smaller container. The person is positioned towards the right side of the image, with the large container occupying the majority of the space. The grains are scattered around the room, with some visible on the floor and others in the air, indicating an active and ongoing process of transferring the grains."

Example 4: LLaVA more accurate

BLIP-2 Caption:
"a person typing on a laptop in a living room"

LLaVA Caption:
"The image features a person sitting at a dining table with a laptop computer in front of them. The laptop is open and displaying a webpage, likely for work or leisure. The person is using a mouse to navigate the content on the screen. \n\nThe dining table is surrounded by chairs, with one on the left side and another on the right. There are also two cups placed on the table, one closer to the left side and the other near the center. Additionally, a potted plant is located near the left edge of the table, adding a touch of greenery to the scene."

Fig. A4: Example of video captions generated by BLIP-2 and LLaVA on the NeXT-QA dataset, respectively.