ViLA: Efficient Video-Language Alignment for Video Question Answering

Xijun Wang[1], Junbang Liang[2], Chun-Kai Wang[2], Kenan Deng[2], Yu Lou[2], Ming Lin[1,2], and Shan Yang[2]

[1] University of Maryland, College Park, USA
[2] Amazon, USA
{xijun, lin}@umd.edu, {junbangl, kenanden, ckwang, ylou, ssyang}@amazon.com

Abstract. In this work, we propose an efficient **Vi**deo-**L**anguage **A**lignment (ViLA) network. Our ViLA model addresses both efficient frame sampling and effective cross-modal alignment in a unified way. In our ViLA network, we design a new learnable text-guided Frame-Prompter together with a cross-modal distillation (QFormer-Distiller) module. Pre-trained large image-language models have shown promising results on problems such as visual question answering (VQA). However, how to efficiently and effectively sample video frames when adapting pre-trained large image-language model to video-language alignment is still the major challenge. Compared with prior work, our ViLA model demonstrates the capability of selecting key frames with critical contents, thus improving the video-language alignment accuracy while reducing the inference latency (**+3.3%** on NExT-QA Temporal with **3.0**× speed up). Overall, our ViLA network outperforms the state-of-the-art methods on the video question-answering benchmarks: **+4.6%** on STAR Interaction, **+2.2%** on STAR average with **3.0**× speed up, ours 2-frames out-perform SeViLA 4-frames on the VLEP dataset with **4.2**× speed-up. Code will be available at https://github.com/xijun-cs/ViLA.

1 Introduction

"If a picture is worth thousands of words, what is a video worth?" [36] Video watching has become a new social norm. Statistics show YouTube has approximately 122 million daily active users, based all over the world. Visitors spend on average 19 minutes per day on YouTube. An average of close to 1 million hours of video are streamed by YouTube users each and every minute. As video data continue to grow through internet viewing, video information retrieval becomes more demanding. Video data has tremendous capacity to store a vast variety of useful information. Compared to image question answering (Q&A) problem, video QA is more challenging due to one extra temporal dimension. How to efficiently sample relevant frames from a video with the computing resource constraint remains a long-standing problem in video QA research.

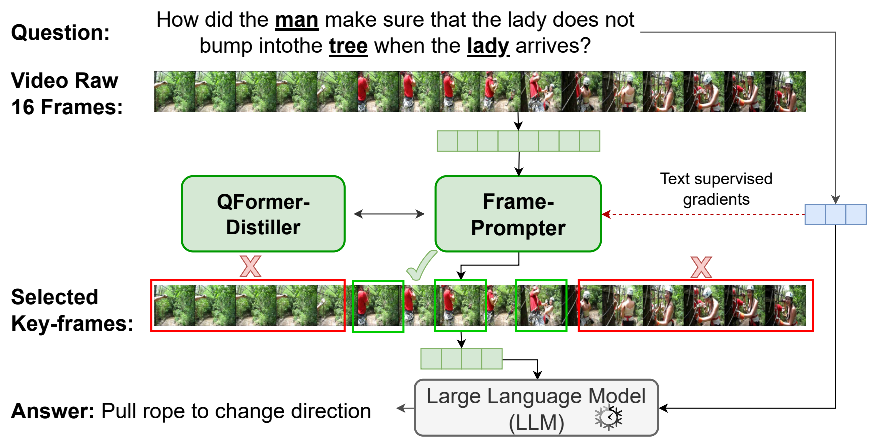

Fig. 1: Our efficient **Vi**sion-**L**anguage **A**lignment (**ViLA**) model via Frame-Promper and distilling contains two new modules: a *text-guided* Frame-Prompter and a cross-modal QFormer-Distiller. It learns to extract the most question-related frames while keeping the inference latency low.

Recent advances in pre-trained large-scale language models [6,46] have greatly boosted the performance of the vision-language models, especially on cross-modality tasks. Many state-of-the-art image-language models [1,8,27,49] leverage pre-trained LLMs. These models [10,60] achieve excellent performance on visual-language tasks such as image captioning [10], visual question answering [60] and more. Inherently, many video-language models [33, 65] are built from these pre-trained image-language models. These image-based video-language models treat a video as a series of multi-channel images sampled randomly or uniformly [27]. While this strategy works well for short videos, for long videos or videos with non-uniform information distribution, random or uniform frame sampling may miss critical information. When it comes to video-language alignment, frame sampling efficiency and effectiveness go hand-in-hand. One needs to not only reduce the number of sampled frames but also select frames that are most related to the input question. Previous work such as SeViLA [65] trains a separate keyframe localizer, which is not friendly for the real-time inference and introduces more parameters to the whole model. Besides video representation, cross-modality alignment while leveraging LLMs is another challenge. The critical problem lies in how to efficiently transfer video information to the LLM's input domain.

Main Contributions: To address these challenges, we propose a new network, ViLA. Compared to the state-of-the-art video-language models [10,27,65], ViLA consists of a new Frame-Prompter together with a QFormer-Distiller. Our Frame-Prompter learns to *select the most important frames influenced by the corresponding question text and and supervised by the VQA loss.* Meanwhile, the Frame-Prompter is meticulously designed to keep lightweight so as to be efficient. To effectively and efficiently transfer video information to LLM input domain, we add a new distillation on top of the QFormer, named QFormer-Distiller. The QFormer is the cross-modal query-visual Transformer proposed in previous BLIP models [10, 27, 28] for cross-modal fusion. We train our Frame-Prompter

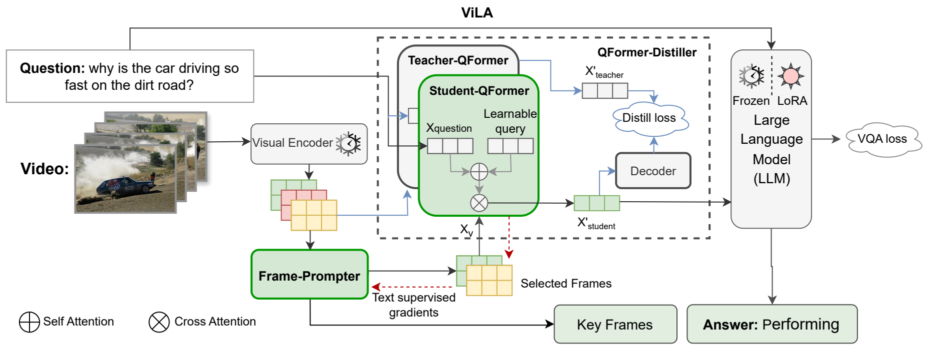

Fig. 2: Model Overview. Our ViLA model includes 4 sub-modules: the visual encoder, text-supervised Frame-Prompter (FP), QFormer-Distiller (QFD), and a LLM. We encode the video frames through a frozen visual encoder. Then we train the Teacher-QFormer using all the frame features. After that, we train the Student-QFormer and Frame-Prompter end-to-end. Unlike the Teacher-QFormer, our Student-QFormer is trained with masked frames features from a text-supervised Frame-Prompter. Finally, the input question text and QFormer transformed visual features go through a frozen large language model to generate the answer. Our network supports both leveraging LLM through proper visual prompting without affecting the original LLM (Frozen) ability on language tasks and finetuning LLMs(LoRA) simultaneously to get optimal performance on specific tasks.

and QFormer-Distiller end-to-end. The cross-modal temporal distiller *teaches* a smaller (i.e. fewer frames) QFormer. To the best of our knowledge, this work is the first to propose a Frame-Prompter and a distiller on top of the cross-modal alignment for video-language learning with pre-trained LLMs.

We validate our ViLA model on the Video Question Answering benchmark datasets. This includes the NExT-QA [56], STAR [54], How2QA [31], TVQA [25] and VLEP [26]. Our work outperforms previous strong SOTA methods across all the benchmarks, while reducing inference latency. Comparing with SOTA video-language model SeViLA [65], we reduce significant amount (50%) of training parameters and inference latency ($1.35-4.2\times$ speed up), while improving the accuracy. We also performed ablation study on our text-guided Frame-Prompter and QFormer-Distiller. As shown in Table 4, both the text-guided Frame-Prompter and the QFormer-Distiller play critical roles in making our method effective.

To sum up, the key novelty is a new text-guided Frame-Prompter and question-relevant QFormer-Distiller (trained from end-to-end). The Frame-Prompter enhances the *efficiency*, while the later bolsters the *effectiveness*. Together they optimize the selection of frames for video-language alignment learning.

2 Related Work

2.1 Visual-Language Alignment

Visual-Language Pre-training Vision-Language cross-modal pre-training has greatly improved over the past couple of years. Various network architectures and pre-training objectives have been proposed for different downstream tasks, including the dual-encoder architecture with image-text contrastive learning [42], the fusion-encoder architecture with image-text matching [29], and unified transformer architecture with masked language modeling [50]. These methods, along with others, focus on the ability to find image-text affinity [61], correlation [4], and/or completion [64], and need to pre-train the model end-to-end. To address the incompatibility with pre-trained unimodal models such as LLMs [6], recent works [27] proposed to train a QFormer to bridge the domain gap between two frozen pre-trained models. Inspired by its flexibility, more downstream tasks and applications have been proposed, including instruction-based image generation [55] and image question-answering [60].

While most of the previous work focus on image-text alignment, very few have discussed the extension to videos until most recently when temporal modeling starts to be included for better reasoning capabilities. HiTeA [62] jointly trains pairs in long and short view to capture temporal relations between moments and event. Smaug [32] introduces sparse image patch masks to reduce pre-training costs and improve cross-modal alignment. EgoVLPv2 [40] proposes cross-attention between the backbone encoders to improve both the pre-training efficiency and the downstream task performance.

Our method can leverages the pre-trained model during the video-language alignment, which naturally provides reasoning capability with less training cost.

Image-to-Video Transfer Learning Due to the high cost of large video dataset collection, many works leverage successful pre-trained image models and transfer the knowledge to video task. Previous works such as [3, 5, 18, 30] utilize a pre-trained ViT [12] and aggregate the temporal image feature sequence using transformer block to adapt for video understanding task. For Video-Language task, many works turn to large-scale Vision-Language models as the starting point, such as CLIP [42], BLIP [28]. Many works choose to adapt a pre-trained CLIP model for text-to-video retrieval task, by either augmenting the frame-level spatial features with temporal information [11, 57], or manipulate the cross-modal similarity calculation to get better video-language alignment [14,34,35,53]. Other works [39,41] focus on parameter efficient fine-tuning for video task by inserting trainable temporal modules into the pre-trained transformer architecture while keeping the rest of model frozen. Recent work SeViLA [65] proposes a language-aware frame localizer to sample relevant key frames from videos. In this paper, we propose a trainable text-guided Frame-Prompter and a QFormer-Distiller module, which help focus more on the important temporal and spatial information from the 2D frames. These techniques help to efficiently bridge the gap between image-language and video-language learning.

Video Question Answering One major downstream task of Video-Language pre-training is Video Question Answering (VQA). Early works [2, 22, 63] often rely on human annotated datasets, while recent works [58, 66, 67] make better use of large-scale data from public. Along with the quick advances in VQA methods, A lot of benchmark datasets have also been introduced to standardize the model performance comparison, including NExT-QA [56], STAR [54], How2QA [31], TVQA [25], and VLEP [26]. We benchmark our network mainly on the Video Question and Answering task.

2.2 Knowledge Distillation

One of our key component is cross-modal distillation. Knowledge Distillation [19, 51, 69] is original proposed for small and efficient models to mimic the softened class distributions, features of large teachers. For multi-modalities, researchers explore how to utilize the prior knowledge between different modalities [17, 44, 45]. On video domain, knowledge distillation has been used for efficient video inference [23,37], video captioning [38], video question answering [43]. Out of the supervised learning methods mentioned above, the idea of knowledge distillation has also been leveraged in many self-supervised methods for self-supervised video representation learning [48].

2.3 Frame Selection for Video QA

Early approaches in Video QA relied heavily on dense sampling methods, where frames are extracted at a fixed interval throughout the video. While straightforward, such methods can lead to excessive computational costs and memory requirements without significantly improving performance. Zhang et al. [68] proposed a more selective strategy, using attention mechanisms to identify key frames that are more likely to contain information relevant to the question. Following this, adaptive frame sampling methods [7, 13] have gained popularity. These methods aim to dynamically select frames based on the content's relevance to the question, thus optimizing the trade-off between computational efficiency and answer accuracy [13].

3 Method

Our ViLA model tackles the following challenges in large-scale Video-Language learning: how to sample question related frames and how to efficiently transfer video information for pre-trained frozen LLMs.

3.1 Model Architecture

As illustrated in Fig. 2, ViLA has four components: a pre-trained frozen large scale visual encoder E_v, a Frame-Prompter F_p, a QFormer-Distiller Q_d (Querying Alignment Transformer [10,27,28] with distillation) to fuse and extract text-conditioned visual information, and a pre-trained frozen LLM. We train our Frame-Prompter, QFormer-Distiller and two other frozen modules end-to-end. Our training objective includes a distillation loss L_{distill} and a task loss L_{vqa}. For the multi-choice Video Question Answer task, the task loss is the cross-entropy.

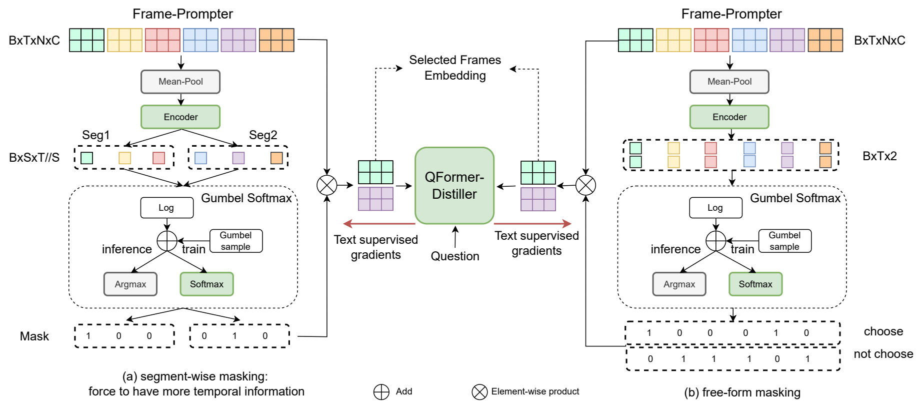

Fig. 3: Text-guided Frame-Prompter. Here we show the details of our learnable text-guided Frame-Prompter. We design a learnable Frame-Prompter to sample the most text query-related frames, with two design choises (a and b). We choose design (a) for diversified temporal sampling. We first encode the mean-pooled segment features. We then apply the Gumbel Softmax to compute the segment mask to guarantee the differentiability. The selected frames embedding then goes through the QFormer-Distiller. Here B means batch size, T means frame number, $N \times C$ means the frame feature sequences. The Frame-Prompter is learned with the text-supervised gradient. When VQA loss is applied, the input question text-related gradient further flows to the Frame-Prompter. The question text-related gradient guides the Frame-Prompter to select the most critical frames.

3.2 Text-guided Frame-Prompter Learning

For the four-dimensional video data, it's impractical and non-efficient to input all frames into a visual encoder model. Here we design a text-guided Frame-Prompter for efficient and effective frame sampling. It is designed to learn attending to fewer but more question-related frames. We start from the uniformly sampled video frames $\{f_1, \ldots, f_T\}$, T is frame number. As shown in Fig. 3, these raw frames first go through the pre-trained visual encoder E_v,

$$X = \{x_i | x_i = E_v(f_i), i \in [1, T]\}, \tag{1}$$

where x_i is the visual feature extracted from raw frames and T is the number of frames. We perform mean-pooling on the channels per-patch. The Frame-Prompter input shape is (B, T, N, C), B batch size, T temporal, (N, C) as the frame feature dimension. Mean-pool averages over the C channel. After mean-pool, the feature dimension becomes (B, T, N). Then we divide the feature into S segments $(B, S, T//S \cdot N)$, with $(T//S \cdot N)$ as the segment feature dimension. Then the segmented features goes through FC layer to project to $(B, S, T//S)$ dimension, ready for the gumbel-softmax computation. For free-form frame sampling Fig. 3(b), after mean-pool, we use FC layer to transfer (B, T, N) to $(B, T, 2)$. Then we encode the mean-pooled visual features onto an embedding using a fully-connected (FC) layer with a layer normalization (LN),

$$\hat{x_i} = W_2 * \mathrm{ReLU}(\mathrm{LN}(W_1 * \mathrm{Mean}(x_i))), \tag{2}$$

where W_1, W_2 are the learnable weights and LN is the layer normalization.

Our Frame-Prompter is a differentiable frame selecter. This allows the text-supervised gradients from the QFormer to guide the learning our frame selecter via backpropagation. The differentiability is achieved through the Gumbel-Softmax [21]. We have two choices (shown in Fig. 3 a and b) for frame selection before the Frame-Prompter: a) strategic sampling (like: segmented uniform sampling) and b) free-form sampling. We choose option (a) to force diverse temporal sampling. We first convert a segment of concatenated frames feature $s_k = [\hat{x_i}, \ldots, \hat{x}_{T//S}]$ into a categorical distribution π through the softmax operation,

$$\pi = \left\{ p_k \mid \frac{\exp(s_i)}{\sum_{j=1}^{S} \exp(s_j)}, k \in [1, T/S], i, j \in [1, S] \right\}, \tag{3}$$

where S is the frame number of one segment. Then we compute our segment mask M_S using the segment probability p_k and the g_k is sampled from the Gumbel$(0, 1)$ distribution,

$$M_k = \mathcal{H}\left(\arg\max\left[g_k + \log p_k\right]\right) k \in [1, T/S], \tag{4}$$

where $\mathcal{H}$ is the one-hot encoding operation. During training, we replace the argmax with the softmax for differentiability,

$$\frac{\exp((\log p_i + g_i)/\tau)}{\sum_{j=1}^{S} \exp((\log p_j + g_j)/\tau)}, i, j \in [1, S], \tag{5}$$

where τ is a tunable temperature. Full mask is $M = \{M_i, \cdots, M_{T//S}\}$. We apply mask M to input frames and then conduct CrossAttention to obtain text guidance by

$$X_{LLM} = \text{CrossAttn}(X \cdot M, X_t), \tag{6}$$

where X_t is the text information. This CrossAttention can help the Frame-Prompter learn to choose rich text frames. Finally, we use the task loss below to supervise the learning and choose the frame selection to answer our question:

$$L_{\text{vqa}} = \text{MSE}(LLM(X_{LLM}, X_t), QA_{\text{answer}}). \tag{7}$$

3.3 Cross-Modal Distillation

In this section, we present in detail our cross-modal distillation module, designed to selectively transfer video information ready for LLM consumption. Video-language alignment performance, unlike image-language alignment, is closely related the selected frames. Meanwhile, to leverage powerful pre-trained LLMs, we need to transform the selected visual information to the LLM's input domain. We adopt the QFormer proposed in the BLIP models [10, 27, 28] for the cross-modal transformation learning.

The original QFormer is designed for image-text fusion, and we add a distillation module to make it efficient for video-text fusion. Like the traditional distillation, our QFormer-Distiller includes a teacher-QFormer and a student-QFormer. We train the teacher-QFormer first (without the student-QFormer) on a wider receptive field. The student-teacher learning mechanism further encourages both the QFormer and the Frame-Prompter to learn to attend to the most relevant visual information given the input question text. We study how QFormer-Distiller affects Frame-Prompter (Sec. 4.3). And we find our FP+QFD 4-frames(Acc.**74.4%**) model's key-frame selection overlaps **92.3%** with the key-frame selected by the FP 8-frames(Acc.**74.1%**) model.

Specifically, during student learning process, both the teacher-QFormer and the student-QFormer will take in the video X_v and question text X_q. The concatenated tokenized video and question text will go through the QFormer cross-attention layers:

$$X' = \text{CrossAttn}(E_t(X_q), \text{SelfAttn}(E_v(X_v), \mathbf{q})), \tag{8}$$

where E_t is a text tokenizer, E_v is the visual tokenizer, $\mathbf{q}$ is the learnable query token. We utilize a decoder D to transform the student-QFormer output, ensuring consistency and recoverability to the teacher's feature. This supervision bolsters the effectiveness and performance of the model with fewer frames. Then the distillation objective is defined by:

$$L_{\text{distill}} = \text{MSE}(D(X'_{\text{student}}), X'_{\text{teacher}}). \tag{9}$$

We carefully design the decoder to be a simple Fully Connected (FC) layer with a layer normalization. We show through an ablation study (Sec. 4.3) that this combination works the best. Meanwhile, our FP+QFD 4-frames boosts accuracy by **1.8%** compared with FP 4-frames model and the key-frame selected overlaps 56.8% with that of the FP 4-frames model. This helps verify our QFD's ability to both boost performance and enhance the key-frame selection.

4 Experiments

4.1 Implementation Settings

For training, we conduct experiments with 8 $\times$ 40GB A100 GPUs. All the models in our experiments are trained using AdamW with cosine learning scheduler. For all the experiments, we use ViT-G(1B) [15] as the visual encoder and Flan-T5 XL (3B) [9] as large language model. For all the datasets, we use accuracy of choosing the right answer as the metric, and test on the validation dataset. For the temperature τ in the Frame-Prompter, it changes from 1 to 0.01 by $(0.01)^{(\text{current-step}/\text{total-training-step})}$ and is set as 0.01 during testing. And all inference (Infer.) time is evaluated on a single A100 GPU with batch size 4. Please refer supplementary for more training details.

Benchmark: To demonstrate the effectiveness of our proposed method on the video QA task, we compare our algorithm with the state-of-the-art (SOTA) methods on 5 datasets including 1 on video event prediction: 1) **NExT-QA** [56], a multi-choice VideoQA benchmark, with 3 types of questions: Causal (Why, How), Temporal (Previous/Next, Present), and Description (Binary, Location, Count and Other). 2) **STAR** [54], a multi-choice VideoQA benchmark for Situated Reasoning. It has four kinds of questions: Interaction, Sequence, Prediction, and Feasibility. 3) **How2QA** [31], a multi-choice VideoQA benchmark with 44k QA pairs for 22k 60-second clips selected from 9035 videos. 4) **TVQA** [25] is a large-scale video QA dataset with 152K questions along with 21k video clips from 460 hours of video. 5) **VLEP** [26] is a video event prediction benchmark, with 28,726 future event prediction cases. Following SeViLA [65], we formulate this task as a multi-choice Question Answering.

Method	# Frames	Temporal	Causal	Description	Average	Intro. Param.	Infer. Time (ms/video)
Just Ask [58] (ICCV2021)	20	51.4	49.6	63.1	52.3	-	-
All-in-One [47] (CVPR2023)	32	48.6	48.0	63.2	50.6	-	-
MIST [16] (CVPR2023)	32	56.6	54.6	66.9	57.1	-	-
HiTeA [62] (ICCV2023)	16	58.3	62.4	75.6	63.1	-	-
InternVideo [52] (Dec 2022)	8	58.5	62.5	75.8	63.2	-	-
ViLA (Ours)	**1**	66.5	69.3	78.0	70.5	188M	64
BLIP-2 [27] (ICML2023)	4	67.2	70.3	79.8	71.5	-	-
ViLA (Ours)	**2**	70.2	71.6	79.4	72.8	188M	72
SeViLA [65] (NeurIPS2023)	4	<u>67.7</u>	<u>72.1</u>	<u>82.2</u>	<u>73.4</u>	376M	301
SeViLA [65] (NeurIPS2023)	8	<u>67.0</u>	<u>73.8</u>	**81.8**	<u>73.8</u>	376M	306
ViLA (Ours)	4 (8 to 4)	**71.0**	72.9	**82.7**	74.3	188M	**99 (3.04× ↑)**
ViLA (Ours)	4 (32 to 4)	70.1	**73.8**	82.1	**74.4**	188M	208 (1.45× ↑)
ViLA (Ours)	8	**71.4**	**73.6**	<u>81.4</u>	**74.8**	188M	227 (1.35× ↑)
ViLA+LoRA (Ours)	**4**	71.4	74.5	80.3	75.1	188M	208
ViLA (Ours)	32	**71.8**	**75.3**	82.1	**75.6**	188M	-

Table 1: Comparison Results on NExT-QA dataset. Here we measure the accuracy of choosing the right answer. Especially on Temporal and Causal type of questions, our ViLA (using only 4 frames) improves **3.3%** and **1.7%** respectively, compared with SeViLA. We use **bold-face** font to indicate the best results and <u>underline</u> on the second best using the same number of frames (brown box for 4 frames and blue box for 8 frames). ViLA using 2-frames only out-performs BLIP-2 using 4-frames by **1.3%**. ViLA also achieves upto **3.04**× speedup. It needs to be noted that our ViLA achieves 75.1% average accuracy with only 4 frames when we finetune LLM with LoRA [20].

Baselines: We compare the performance of our ViLA model with several recent prominent models in the field. These include SeViLA [65], BLIP-2 [27], and InternVideo [52], within the context of fine-tuning scenarios. For fair comparison with SeViLA and BLIP-2, we employ the Vision Transformer-Global (ViT-G) as the visual encoder and Flan-T5-XL models as the Large Language Model.

4.2 Results

Here we show both quantitative and qualitative (Sec. 4.2) comparison results between our ViLA and SOTA methods on Video QA and Video Event Prediction task (Sec. 4.2). Together we present an in-depth analysis on the results.

Then we conduct ablation study (Sec. 4.3) on our proposed Frame-Prompter and QFormer-Distiller module and the design choices within each module.

Comparison Results on Video QA and Video Event Prediction: Overall, we demonstrate that the **cross-modal key-frame selection matters**. Our ViLA model out-performs strong SOTA methods across the 4 Video QA benchmark datasets and 1 Video Event Prediction while keeping the inference latency low. Especially, we highlight that our models' performance stands out on temporal (including Causal, Interaction, Prediction) type of questions, NExT-QA Temporal (+3.3%, 3.04× speed up), NExT-QA Causal(+1.7%, 1.45× speed up), STAR Interaction (+**4.6%**, 3.04× speed up), STAR Prediction (+2.8%, 3.04× speed up), How2QA (+0.3%, 3.04× speed up), VLEP (+0.7% with 1.45× speed up, +0.3% with 4.18× speed up) and TVQA (+1.8%, 3.04× speed up).

On the NExT-QA [56] dataset, compared with the SOTA SeViLA [65] on 4-frame and 8-frame setting, our ViLA improves by 1.0% on average accuracy while achieving a **3.04**× speed up (see Table 1).

Method (Frames Number)	Interaction	Sequence	Prediction	Feasibility	Avg.	Infer. Time (ms/video)
Flamingo-9B 4-shot [1] (30) (NeurIPS2022)	-	-	-	-	42.8	-
All-in-One [47] (32) (CVPR2023)	47.5	50.8	47.7	44.0	47.5	-
MIST [16] (32) (CVPR2023)	55.5	54.2	54.2	44.4	51.1	-
InternVideo [52] (8) (Dec 2022)	62.7	65.6	54.9	51.9	58.7	-
BLIP-2 [27] (4) (ICML2023)	<u>65.4</u>	69.0	59.7	54.2	62.0	-
ViLA (2) (Ours)	65.0	65.4	62.2	58.8	62.9	72
SeViLA [65] (4) (NeurIPS2023)	63.7	**70.4**	<u>63.1</u>	**62.4**	<u>64.9</u>	301
ViLA (4) (Ours)	**70.0**	**70.4**	**65.9**	<u>62.2</u>	**67.1**	99 (3.04× ↑)

Table 2: Comparison Results on STAR Video QA benchmark. For Interaction type of question, our ViLA improved **4.6%**. On average, our ViLA out-performs the SOTA method by 2.2% when using 4 frames with **3.04**× speed up. Note that our ViLA using 2-frames out-performs BLIP-2 using 4-frames.

We also test our ViLA model on the STAR [54] dataset. This dataset is challenging due to its diversified type of questions. As shown in Table 2, our ViLA model outperforms the several strong SOTA models on average by **2.2%** with **3.04**× speed up. Especially on the STAR Interaction and Prediction type of questions, we outperform SeViLA [65] by **4.6%** and **2.8%**. This result further demonstrate the importance of key-frame selection. And our model's advantages in extracting temporal and causal related key-frames.

We further test our model on the longer and larger-scale Video QA benchmark datasets: TVQA [25] and How2QA [31], as shown in Table 3. On the TVQA [25] dataset, our ViLA outperforms the SOTA method by **1.8%** at the 4-frame setting. On How2QA, our ViLA improvement is 0.3% with 3.04× speed up compared with SeViLA [65]. This is partially due to the limited 32-frame teacher-QFormer training. On the other hand, compared with Blip-2 [27], ViLA outperforms by **1.7%**. This difference again shows the key-frames selected by our ViLA model aligns better with the input question compared with uniform

Method	Frames Numbers	How2QA	VLEP	TVQA
FrozenBiLM [59] (NeurIPS2022)	10	81.5	-	57.5
InternVideo [52] (Dec 2022)	8	79.0	63.9	57.2
BLIP-2 [27] (ICML2023)	4	82.2	67.0	54.5
SeViLA [65] (NeurIPS2023)	4	83.6	68.9	61.6
ViLA (Ours)	**2**	82.8	69.2	60.6
ViLA (Ours)	4	**83.9**	**69.6**	**63.4**

Table 3: Comparison Results on How2QA, VLEP, and TVQA Video QA Benchmarks. ViLA improves performance over SeViLA by **1.8%** with **3.04**× speed up on TVQA dataset, 0.7% with 1.45× speed up on VLEP dataset, and 0.3% with 3.04× speed up on How2QA dataset at 4 frames setting. Ours 2-frames out-perform SeViLA 4-frames on VLEP by 0.3% with **4.2**× speed up.

sampling. To test our algorithm's capabilities, particularly for event prediction, we conduct an additional series of evaluations on VLEP [26] dataset. At the 4-frame setting, ViLA improves over SeViLA [65] by 0.7% with 1.45× speed up. it's noteworthy that ours 2-frames out-perform SeViLA 4-frames on the VLEP dataset by 0.3% with **4.2**× speed up (Table 3).

Qualitative Results: We qualitatively compare the key-frames selected by our ViLA with the ones from SeViLA [65] on different type of questions. As shown in Figure 4, our network select more question-related key-frames across different question types (including Causal, Temporal and Description or Factual). In the first (col 1 row 1) example in Figure 4, our ViLA locates the frames that visibly show the "car and the dirt", but the frames selected by SeViLA focuses on the "road". In the second (col 2 row 1) example, we locate the frames with the "the man not on the bull at the end", but the frames selected by SeViLA focuses on the "man on the bull at the end". And in the fourth (col 2 row 2) example, we locate 3 frames with the "four people" according to the question, but none of the frames selected by SeViLA shows "four people". For Temporal type of questions, our ViLA is also capable of selecting frames that are closer to the action specified in the question. In the seventh (col 1 row 4) and eighth (col 2 row 4) example in Figure 4, we locate key-frames around the "the men on stage" and the "black dog" (vs. SeViLA has 2 frames focusing on the "brown dog").

In addition, we qualitatively check the key-frames selected though our QFormer-Distiller. We show in Figure 5 training our QFormer-Distiller helps select the most question related frames. In the second example, we select frames that shows the "deep into the whole". And in the third example, out of the 16 frames, we pick up the frames that have both the "man in grey and the stick falling".

4.3 Ablation Study

Here we demonstrate the effectiveness of each proposed module: the text-guided Frame-Prompter and the question-relevant QFormer-Distiller. We also validate the decoder design choice within the QFormer-Distiller.

Fig. 4: Key-frame Selection Comparison Results (select 4 frames from 32 frames). We compare frames selected by our ViLA compared with that from the SOTA SeViLA [65] method. Across different type of questions, especially the Causal, Temporal type questions, keyframes selected by our network is more relevant and better related to the question.

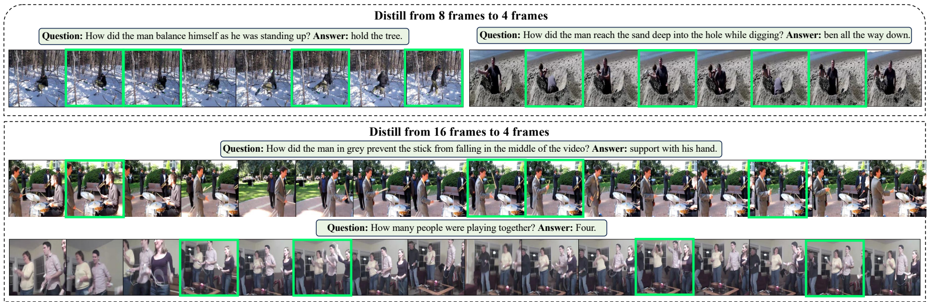

Fig. 5: QFormer-Distiller Results Visualization. Here we visualize the keyframes selected after cross-modal distillation. After distillation, we can select the most question-relevant frames even from 16 frames.

Components	STAR	VLEP	TVQA	NExT-QA
base (BLIP-2)	62.0	67.0	54.5	71.5
base+QFormer-Distiller	64.9	68.6	62.2	73.5
base+Frame-Prompter	65.3	68.8	62.7	73.6
base+QFormer-Distiller+Frame-Prompter	66.5	69.6	63.4	74.4

Table 4: Frame-Prompter and QFormer-Distiller Ablation Results. Across all four VideoQA datasets, we observe that both Text-aware Frame-Prompter and cross-modal QFormer-Distiller contribute significantly to our final performance. We highlight that on STAR, adding our QFormer-Distiller improves the accuracy by **2.9%**. Our Frame-Prompter further boost the accuracy by **1.6%**.

Frame-Prompter and QFormer-Distiller Ablation: Here we ablate our new QFormer-Distiller (QFD) and Frame-Prompter (FP). To summarize, our QFormer-Distiller and our Frame-Prompter each contributes 50% to the overall performance improvement across the 4 Video QA benchmarks. Specifically, our QFormer-Distiller boost performance by **2.9%** and our Frame-Prompter by **1.6%** on the STAR [54] dataset, as shown in Table 4. This shows both modules are critical to achieving the desirable performance.

We further explore how QFormer-Distiller (QFD) affect Frame-Prompter (FP). We find our FP+QFD 4-frames(Acc.**74.4%**) model's key-frame selection overlaps **92.3%** with the key-frame selected by the FP 8-frames(Acc.**74.1%**) model. Meanwhile, our FP+QFD 4-frames boosts accuracy by **1.8%** compared with FP 4-frames model and the key-frame selected overlaps 56.8% with that of the FP 4-frames model. This helps verify our QFD's ability to both boost performance and enhance the key-frame selection.

QFormer-Distiller Decoder Ablation: We study the design choice of the decoder of the QFormer-Distiller. One of the key components is the decoder before computing the distillation loss. Design choices for this component includes the number of Fully Connected layer (FC) and where to put a layer normalization (LN). From Table 5, we show that the simple FC with a LN (after FC) works the best. The FC layer provides the necessary computational structure, while the LN

Frame Prompter Decoder	Temporal	Causal	Description	Average
FC	68.5	70.9	79.3	72.4
FC+LN	70.1	73.8	82.1	74.4
FC+LN+GELU+FC	69.7	73.1	81.6	74.1

Table 5: QFormer-Distiller Decoder Ablation on NExT-QA. We find that a simple Fully Connected layer (FC) with a Layer Normalization (LN) works best across Temporal, Causal, Description. It is efficient and effective. GELU is activation function.

layer stabilizes the learning process. This configuration balances the effectiveness and operational efficiency, making it a well-suited choice for our method.

4.4 More Discussion

More recent works [24, 65] have shown powerful networks such as LLMs learns most of the data distribution during the pre-training stage. This is one of the major reason why prompt-learning is very effective. Our work is inspired from the prompt-learning. We hope to leverage LLM through proper visual prompting without affecting the generalization ability of the LLM, and this won't affect LLMs' original ability on language tasks.

However, for the VQA task itself, optimal performance in this task often necessitates training the Language Model (LLM). Therefore, we conduct an ablation study on NExT-QA that fine-tunes the LLM by using LoRA during the training. Our ViLA achieves 75.1% average accuracy with only 4 frames. This demonstrates that our ViLA has a strong adaptation ability.

Furthermore, to evaluate the generalization of our ViLA, we replace Flan-T5 with llama on NExT-QA, baseline with llama (68.6%) vs. ViLA with llama (72.7%), which shows that ViLA can adapt to different LLMs.

5 Conclusion, Limitation and Future Work

How to properly ingest visual information to LLMs to utilize its capability effectively in cross-modal tasks? In this work, we present a new ViLA network with a new text-guided Frame-Prompter to smartly sample important frames, together with a cross-modal temporal distillation (QFormer-Distiller) for efficient and effective video-language alignment. From our experiments, our ViLA out-performs SOTA methods on four video question answering benchmarks and one video event prediction benchmark, especially on the temporal (including Temporal, Causal, Interaction and etc.) type of questions. For the limitation, we don't have a change to evaluate on LLMs larger than 13b due to the resource constraints. We plan to continue research on the design of our Frame-Prompter, especially on long video-language alignment.

References

1. 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. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022) 2, 10
2. Amrani, E., Ben-Ari, R., Rotman, D., Bronstein, A.: Noise estimation using density estimation for self-supervised multimodal learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 6644–6652 (2021) 4
3. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846 (2021) 4
4. Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O.K., Aggarwal, K., Som, S., Piao, S., Wei, F.: Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. Advances in Neural Information Processing Systems **35**, 32897–32912 (2022) 4
5. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML. vol. 2, p. 4 (2021) 4
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 2, 4
7. Chen, K., Li, Y., Zhang, Z., Ren, X., Li, H.: Sas video-qa: Self-adaptive sampling for efficient video question-answering. arXiv preprint arXiv:2307.04192 (2023) 5
8. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al.: Pali: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794 (2022) 2
9. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022) 8
10. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500 (2023) 2, 5, 7
11. Deng, C., Chen, Q., Qin, P., Chen, D., Wu, Q.: Prompt switch: Efficient clip adaptation for text-video retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15648–15658 (October 2023) 4
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 4
13. Fan, C., Zhuo, T., Zhang, P., Li, X.: Heterogeneous memory enhanced multimodal attention model for video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1999–2008 (2019) 5
14. Fang, H., Xiong, P., Xu, L., Chen, Y.: Clip2video: Mastering video-text retrieval via image clip. arXiv preprint arXiv:2106.11097 (2021) 4
15. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19358–19369 (2023) 8

16. Gao, D., Zhou, L., Ji, L., Zhu, L., Yang, Y., Shou, M.Z.: 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. pp. 14773–14783 (2023) 9, 10
17. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2827–2836 (2016) 5
18. Heigold, G., Minderer, M., Gritsenko, A., Bewley, A., Keysers, D., Lučić, M., Yu, F., Kipf, T.: Video owl-vit: Temporally-consistent open-world localization in video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13802–13811 (October 2023) 4
19. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. Advances in Neural Information Processing Systems (NIPS) (2015) 5
20. 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) 9
21. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144 (2016) 7
22. Jiang, J., Chen, Z., Lin, H., Zhao, X., Gao, Y.: Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11101–11108 (2020) 4
23. Khani, M., Hamadanian, P., Nasr-Esfahany, A., Alizadeh, M.: Real-time video inference on edge devices via adaptive model streaming. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4572–4582 (2021) 5
24. Ko, D., Lee, J.S., Kang, W., Roh, B., Kim, H.J.: Large language models are temporal and causal reasoners for video question answering. arXiv preprint arXiv:2310.15747 (2023) 14
25. Lei, J., Yu, L., Bansal, M., Berg, T.L.: Tvqa: Localized, compositional video question answering. arXiv preprint arXiv:1809.01696 (2018) 3, 5, 9, 10
26. Lei, J., Yu, L., Berg, T.L., Bansal, M.: What is more likely to happen next? video-and-language future event prediction. arXiv preprint arXiv:2010.07999 (2020) 3, 5, 9, 11
27. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023) 2, 4, 5, 7, 9, 10, 11
28. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022) 2, 4, 5, 7
29. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021) 4
30. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Wang, L., Qiao, Y.: Uniformerv2: Unlocking the potential of image vits for video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1632–1643 (October 2023) 4
31. Li, L., Chen, Y.C., Cheng, Y., Gan, Z., Yu, L., Liu, J.: Hero: Hierarchical encoder for video+ language omni-representation pre-training. arXiv preprint arXiv:2005.00200 (2020) 3, 5, 9, 10

32. Lin, Y., Wei, C., Wang, H., Yuille, A., Xie, C.: Smaug: Sparse masked autoencoder for efficient video-language pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2459–2469 (2023) 4
33. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023) 2
34. Liu, Y., Xiong, P., Xu, L., Cao, S., Jin, Q.: Ts2-net: Token shift and selection transformer for text-video retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022) 4
35. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing **508**, 293–304 (2022) 4
36. MacFarland, S.: If a picture is worth a thousand words, what is a video worth? (2014), https://www.huffpost.com/entry/if-a-picture-video-production_b_4996655 1
37. Mullapudi, R.T., Chen, S., Zhang, K., Ramanan, D., Fatahalian, K.: Online model distillation for efficient video inference. In: Proceedings of the IEEE/CVF International conference on computer vision. pp. 3573–3582 (2019) 5
38. Pan, B., Cai, H., Huang, D.A., Lee, K.H., Gaidon, A., Adeli, E., Niebles, J.C.: Spatio-temporal graph for video captioning with knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10870–10879 (2020) 5
39. Pan, J., Lin, Z., Zhu, X., Shao, J., Li, H.: St-adapter: Parameter-efficient image-to-video transfer learning. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 26462–26477. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/a92e9165b22d4456fc6d87236e04c266-Paper-Conference.pdf 4
40. Pramanick, S., Song, Y., Nag, S., Lin, K.Q., Shah, H., Shou, M.Z., Chellappa, R., Zhang, P.: Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5285–5297 (2023) 4
41. Qing, Z., Zhang, S., Huang, Z., Zhang, Y., Gao, C., Zhao, D., Sang, N.: Disentangling spatial and temporal learning for efficient image-to-video transfer learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13934–13944 (October 2023) 4
42. 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. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 4
43. Shao, Z., Wan, J., Zong, L.: A video question answering model based on knowledge distillation. Information **14**(6), 328 (2023) 5
44. Shen, Y., Wang, X., Gao, P., Lin, M.: Auxiliary modality learning with generalized curriculum distillation (2023) 5
45. Shen, Y., Yang, L., Wang, X., Lin, M.C.: Small-shot multi-modal distillation for vision-based autonomous steering. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 7763–7770. IEEE (2023) 5
46. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: open and efficient foundation language models, 2023. URL https://arxiv. org/abs/2302.13971 2

47. Wang, J., Ge, Y., Yan, R., Ge, Y., Lin, K.Q., Tsutsui, S., Lin, X., Cai, G., Wu, J., Shan, Y., et al.: All in one: Exploring unified video-language pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6598–6608 (2023) 9, 10
48. Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., Yuan, L., Jiang, Y.G.: Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6312–6322 (2023) 5
49. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al.: Cogvlm: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079 (2023) 2
50. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pretraining for all vision and vision-language tasks. arXiv preprint arXiv:2208.10442 (2022) 4
51. Wang, X., Liu, D., Kan, M., Han, C., Wu, Z., Shan, S.: Triplet knowledge distillation. arXiv preprint arXiv:2305.15975 (2023) 5
52. 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) 9, 10, 11
53. Wang, Z., Sung, Y.L., Cheng, F., Bertasius, G., Bansal, M.: Unified coarse-to-fine alignment for video-text retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2816–2827 (October 2023) 4
54. Wu, B., Yu, S., Chen, Z., Tenenbaum, J.B., Gan, C.: Star: A benchmark for situated reasoning in real-world videos. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021) 3, 5, 9, 10, 13
55. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671 (2023) 4
56. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9777–9786 (2021) 3, 5, 8, 10
57. Xue, H., Sun, Y., Liu, B., Fu, J., Song, R., Li, H., Luo, J.: Clip-vip: Adapting pre-trained image-text model to video-language representation alignment (2023) 4
58. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Just ask: Learning to answer questions from millions of narrated videos. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1686–1697 (2021) 4, 9
59. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Zero-shot video question answering via frozen bidirectional language models. Advances in Neural Information Processing Systems **35**, 124–141 (2022) 11
60. Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., Wang, L.: Mm-react: Prompting chatgpt for multimodal reasoning and action. arXiv preprint arXiv:2303.11381 (2023) 2, 4
61. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: Filip: Fine-grained interactive language-image pre-training. arXiv preprint arXiv:2111.07783 (2021) 4
62. Ye, Q., Xu, G., Yan, M., Xu, H., Qian, Q., Zhang, J., Huang, F.: Hitea: Hierarchical temporal-aware video-language pre-training. arXiv preprint arXiv:2212.14546 (2022) 4, 9

63. Ye, Y., Zhao, Z., Li, Y., Chen, L., Xiao, J., Zhuang, Y.: Video question answering via attribute-augmented attention network learning. In: Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 829–832 (2017) 4
64. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arxiv 2022. arXiv preprint arXiv:2205.01917 4
65. 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) 2, 3, 4, 9, 10, 11, 12, 14
66. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16375–16387 (2022) 4
67. Zellers, R., Lu, X., Hessel, J., Yu, Y., Park, J.S., Cao, J., Farhadi, A., Choi, Y.: Merlot: Multimodal neural script knowledge models. Advances in Neural Information Processing Systems **34**, 23634–23651 (2021) 4
68. Zhang, P., Zhuo, T., Li, X.: Learning to select key frames for video question answering. ACM Multimedia Conference (2018) 5
69. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4320–4328 (2018) 5