VLAP: Efficient Video-Language Alignment via Frame Prompting and Temporal Distilling

Anonymous CVPR submission

Paper ID 0000

Abstract

Pre-trained large vision-language models especially large language models have shown promising results for language related tasks like question and answering. Most state-of-the-art video-language models are built from image-language models. But videos, unlike images have one more temporal dimension. With computing resource constrains, how to efficiently and effectively sample image frames from a video is the main challenge for video related tasks. With new advances in LLMs, new challenges emerge for cross-modal tasks like how to properly ingest visual information from videos to LLMs and what information to feed to LLMs. In this work, we propose an Efficient Video-Language Alignment (VLAP) network that tackles efficient frame sampling and cross-modal alignment in one. In our VLAP network, we design a learnable frame prompter module to sample the most important frames and introduce new a cross-modal temporal distillation model to reduce inference computation cost while keep the temporal information. Meanwhile, we introduce a Text-Visua-Text molding strategy to best align across the visual and language modality and leveraging the pre-trained LLMs. We show through ablation study that this molding strategy creates best alignment cross modalities. Overall, our VLAP network outperforms state-of-the-art methods on the video question answering benchmarks and video captioning benchmark.

1. Introduction

"If a picture worth thousands of words, what is a video worth?" Video watching is growing into a new social norm. Statistic shows 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, video information retrieval becomes more and more demanding. Video data has tremendous capacity to store vast variety of useful information. To enable video information retrieval, automatic video understanding is needed. Compared to images, video understanding is difficult in the one extra dimension. How to efficiently sample important frames or learn temporal information from a video with the computing resource constrain remains the long standing problem in video understanding research. Cross-modal alignment is another challenge especially with the advancement in LLM. How to best leverage LLM for video-language alignment is another emerging challenge.

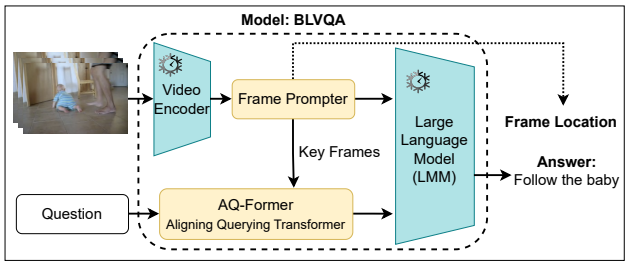

Figure 1. **[TODO: Overview of Boostrapping Lanuage-Video Molding.][YS: remove outline, we show here examples of the visualization of the frame prompter like the chosen frames, and the distillation.]**

Recent advances in large-scale pre-trained language models [5, 22] have greatly boost the performance in the vision-language models. Especially for image-language pre-training [15], many state-of-the-art image-language models leverage pre-trained LLMs to achieve the best on visual-language related tasks like image captioning, visual question answering and so on. Inherently, many video-language model build from those pre-trained image-language models. For those image-based video-language models, they treat video as a multi-channel images. That strategy works well for short videos with uniform or random frame sampling. But for long videos or videos with non-uniform information distribution, treating video as a multi-channel images is very limited. The challenge for sampling frames, information localization and efficient training remains even with

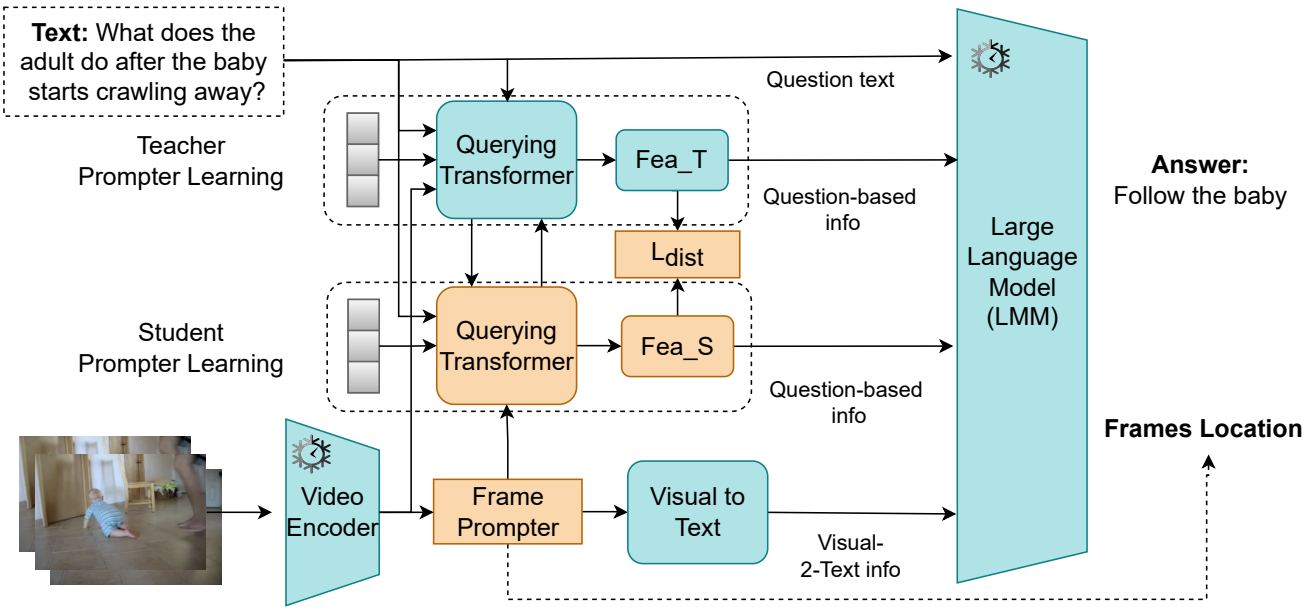

Figure 2. **[TODO: update the fig][YS: add frame prompter, distillation, cross-modal molding as module to the image, update the font size, clarify the distillation steps]**

bootstrapping from large pre-train image-language models.

Cross visual-language alignment has also gained huge improvement with Large-language models. However how to leverage pre-trained powerful LLMs for visual-language is another challenge in video-language learning. The critical problem lies in how to transfer video information to the LLM input domain. Previous work like BLIP-2 [15] transfer visual information using pre-trained visual-to-text module. More recently, in InstructBLIP [6], authors propose a QFormer to fuse the two modality before input to the LLM. For InstructBLIP (image-language model), the cross-modality alignment happens partially with the QFormer. How to

To address these challenges, we propose a new network VLAP. Our VLAP model tackles the problem of efficient video frame sampling and how to best align the two modalities leveraging LLM. Compared against state-of-the-art video-language models [6, 37], we innovate by proposing a new frame prompter, a cross-modal temporal distiller together with a Text-Visual-Text 3-way molding strategy. This frame prompter learns to pick frames that are most informative. The cross-modal temporal distiller teaches a smaller (uses less frames) QFormer for both efficient temporal learning and reduction in model train/inference cost. Our 3-way Text-Visual-Text cross-modal molding strategy best leverage the pre-trained LLM by feeding both text and visual information. Our contribution includes:

- a new instruction-aware video frame prompter to smartly sample important frames together with a cross-modal temporal distillation for efficient and effective temporal learning;
- a 3-way fusion strategy to best align vision and language leveraging pre-trained LLM;
- perform experiments on our strategy that out-performs state-of-the-art method on video question answering and captioning benchmarks.

2. Related Work

Vision-Language Pre-training Vision-Language cross-modal pre-trainig has a large improvement over the past couple of years. Different network architectures and pre-training objectives have been proposed for different downstream tasks, including the dual-encoder architecture with image-text contrastive learning [21], the fusion-encoder architecture with image-text matching [16], and unified transformer architecture with masked language modeling [24]. These methods along with others, focus on the ability to find image-text affinity [33], correlation [3], and/or completion [36], and need to pre-train the model end-to-end. To address the incompatibility with pre-trained unimodal models such as LLMs [5], recent works [15] proposed to train a Q-Former 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 [27] and image question-answering [32].

While most of the previous work focus on image-text alignment, very few have discussed the extension to videos.

Our method is the first to propose video-language pretraining, which naturally provides more capability to reasoning thanks to the temporal information provided by the video.

Image-to-Video Transfer Learning Due to the inherent large computational cost, many recent works have been leveraging image-to-video models transfer learning. Previous works such as [2, 4] utilized a pretrained ViT [7] and aggregate the temporal image feature sequence using transformer block for video understanding task. Given the success of CLIP [21] in image-language domain, many works such as [8, 19, 29] make use of a pre-trained CLIP model and manipulate the cross-modal similarity calculation for video-language alignment. [18, 20] focus on parameter efficient fine-tuning on videos by inserting a temporal module into the transformer architecture to integrate the temporal frame level information while freezing the rest of model. Recent work [37] proposed a language-aware frame localizer to sample relevant keyframes from videos. It adopt a pre-trained image-lanaguage model BLIP2 [15] as the frozen backbone and only the adapter Q-former are trained. In this paper, we propose a instruction-aware frame prompter and a distillation module. These helps to bridge the gap between image-language and video-language learning

3. Method

Our VLAP model aims to tackle the challenges in large scale Video-Language learning, especially on how to sample important frames in a video, efficient training and inference, how to align language and video so as to prepare video information for pre-trained LLMs.

3.1. Model Architecture

VLAP comprises of a frozen visual encoder E_v, an efficient instruction-aware frame prompter F_p learned from a teacher-student pattern, a Querying Alignment Transformer (QA-Former, noted as Q) that extracts question-based visual information and transfers them into LLMs friendly format, a frozen large language model (noted as LLM).

3.2. Instruction-aware Frame Prompter

For video data format information, it's impractical to input all frames into visual models for the efficiency consideration and there is lots of redundant information too. And the widely used uniform/random sampling methods don't have any instruction, which may lose important frames for VQA task and sample frames that are unimportant/irrelevant to language queries when using fewer frames.

To solve this problem, we propose instruction-aware Frame Prompter, which can sample language queries related frames in a teacher-student manner. As shown in Fig. 3, we represent a raw video as $\{f_1, \ldots, f_t\}$ from a uniform

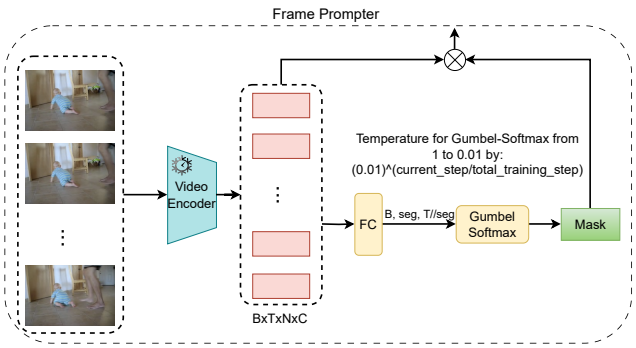

Figure 3. Caption.

sampling at frame rate as 32, which contains enough visual information. First, these raw frames go through the visual encoder first,

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

where x_i is the visual feature extracted from raw frames with shape $B \times T \times N \times C$. B is the batch size, T is the temporal frames number, N is the patch size, C is the feature dimension. Then we perform convolution to transform the dimension for frame selection,

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

where W_1 and W_2 are the convolutional layer weights, $\times$ stands for convolution operation. After Mean operation, the feature shape is $B \times T \times N$. We reshape the feature into $B \times S \times T/SN$, S is the segment number that divides the video into several segments, we pick one frame in one segment. After convolution W_1, the feature shape is $B \times S \times T/S$. Then the feature will go through Layer Normal layer and ReLU. Convolution W_2 generates the logits $\hat{x_i}$ for frames in a segment which denotes which frame to select.

To make this selection process leanable, we need to guarantee all the operations differentiable. Therefore, we apply Gumbel-Softmax [11] to do the frame selection. Specifically, we first generate a categorical distribution by using Softmax in a segment,

$$\pi = \left\{ p_i \mid p_i = \frac{\exp(x_i)}{\sum_{j=1}^{T/S} \exp(x_j)}, i, j \in [1, T/S] \right\}, \tag{3}$$

then we draw samples z from the categorical distribution with class probabilities π,

$$P = \mathrm{one\_hot}\left(\arg\max_i [g_i + \log \pi_i] \right) \tag{4}$$

where g_i are samples drawn from Gumbel(0, 1). $g_i = -log(-logGi)$ and G_j is sampled from uniform distribution at range (0,1). To remove the non-differtialble operation

$\arg\max$, we use the softmax function as a differentiable approximation for backpropagation:

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

where τ is the temperature hyperparameter.

And we design a student-teacher pattern for the training to make the student prompter chooses the most informative frames by optimization objective:

$$L_{F_p} = \mathrm{MSE}(Q(E_v(P \cdot X)), Q(E_v(X))) \quad (6)$$

where Q is QA-Former, MSE is mean squared error loss.

3.3. Cross-Modal Alignment

In this section, we explore the cross-modal molding for LLMs, specially for how to transform visual information to the right format which is friendly for LLMs inputs. We propose a TVT (Text-Visual-Text) pattern comprising of Question (Text), Aligned Question-related Visual Feature (visual: feature from QA-Former), Visual-2-Text (Text: key words).

3.3.1 QA-Former: Querying Alignment Transformer

Language and video alignment is very important for video question answering task. We need to conduct visual information transformation so that LLMs can make full use of visual information for generating answers. BLIP2 [15] proposed Q-former which is first pretrained with the frozen image encoder for vision-language representation learning and then adapted the output of Q-Former as soft visual prompts for text generation with a frozen LLM. InstructBLIP [6] adds task-related instruction text tokens as additional input to encourage the extraction of task-relevant image features.

We go a further step to proposed a QA-Former (Q), which applies question text as additional input with a student-teacher leaning paradigm to extract question-relevant video features. Our student-teacher leaning paradigm can encourage our QA-Former to learn more temporal information within few frames.

For teacher video input X and question input X_t, we first combine question input X_t with query X_q:

$$\mathrm{Query} = \mathrm{Cat}(\mathrm{E_t}(x_t), X_q)), \quad (7)$$

then query feature Query and video feature $E_v(X)$ will go through self-attention Self_A and cross-attention Cross_A:

$$X_q = \mathrm{FF}(\mathrm{Cross\_A}(\mathrm{Self\_A}(Query, E_v(X)))) \quad (8)$$

FF() is feed forward layers. For student video input $P \cdot X$,

$$X_q^s = \mathrm{FF}(\mathrm{Cross\_A}(\mathrm{Self\_A}(Query, E_v(P \cdot X)))). \quad (9)$$

We utilize a decoder to transform the student feature output, ensuring dimension consistency and recoverability to the teacher's feature. Then the optimization objective:

$$L = \mathrm{MSE}(D(X_q^s), X_q). \quad (10)$$

Teacher has a wider receptive field than student in terms of temporal modeling, by learning form the teacher's feature and collaborating with frame prompter, student can better model the temporal information with few frames.

3.3.2 Cross-modal Temporal Distillation

To solve the problem of efficient training and inference, we propose a new cross-modal distiller in our VLAP network.

3.3.3 Text-Visual-Text Fusion

For LLMs, the best input should be in text format. Most existing works feed LLMs by representing videos using continuous feature vectors or discrete text tokens. In our frame work, we combine them together to grantee most valid information has been input to LLMs. We input discrete text tokens coupled with a pretrained contrastive text model to represent the video information in a text format. **[TODO: add a fig for different molding comparison]**

4. Experiment Setup

4.1. Implementation Details

4.2. Benchmark

Video Question and Answering We compare our algorithm with the state-of-the-art (SOTA) methods on Five VideoQA datasets in terms of different aspects. Causal & Temporal in NExT-QA and How2QA, Interaction in STAR, Large scale in TVQA, Prediction in VLEP. Our results demonstrate that our proposed method can effectively address these challenges in video QA task.

NExT-QA [28] is a benchmark for causal and temporal reasoning in terms of multi-choice VideoQA. It has different kinds of questions: Causal (Why, How), Temporal (Previous/Next, Present), and Description (Binary, Location, Count and Other). It contains a total of 5.4K videos with an average length of 44s and approximately 52K questions.

STAR [26] is a multi-choice VideoQA benchmark for Situated Reasoning which contains 22K video clips with an average length of 12s along with 60K questions. STAR contains four different kinds of questions: Interaction, Sequence, Prediction, and Feasibility.

How2QA [17] is a a multi-choice VideoQA benchmark contains 44k QA pairs for 22k 60-second clips selected from 9035 videos. It provides the start and end points for the relevant moment for each question.

TVQA [13] is a large-scale video QA dataset based on 6 popular TV shows (Friends, The Big Bang Theory, How I Met Your Mother, House M.D., Grey's Anatomy, Castle). It contains 152K questions along with 21k video clips from 460 hours of video. It also provides the start and end points for the relevant moment for each question.

Video Event Prediction VLEP [14] is a video event prediction benchmark that requires the model to predict two future events based on the video premise. It contains 28,726 future event prediction cases from 10,234 diverse TV Shows and YouTube Lifestyle Vlog video clips. Following SeViLA [37], we formulate this task as a multi-choice QA.

Video Captioning Flickr30K [35] is obtained by extending Hodosh et al. [10]'s corpus. Flickr30k dataset contains 31,000 images and 158 915 captions, in which each image has 5 reference sentences provided by human annotators. These images cover daily activities, events, and scenes.

Video Moment Retrieval We test our frame prompter's ability on localizing important frames on the QVHighlight [12] dataset.

4.3. Metrics

For video question answering datasets, NExT-QA, STAR, TVQA, VLEP, and How2QA, we use accuracy of choosing the right answer and test on the validation dataset. For key frame detection dataset, QVHighlights, we report the accuracy on the hidden test set. For video key frame detection dataset, QVHighlights, we use mAP over IoU thresholds [0.5: 0.05: 0.95] as in [37], and R@1 with a positive prediction defined by high IoU ($\geq$0.5 or $\leq$0.7) with a ground truth moment. For video captioning, **[TODO: TBD]**

4.4. Baselines

We evaluate our VLAP against SeViLA [37], BLIP-2 [15], and InternVideo [25] in fine-tuning scenarios. For SeViLA and BLIP-2, we use the ViT-G and Flan-T5-XL as the visual encoder and LLM as in VLAP. Following [37], to adapt BLIP-2 to video input, we concatenate the visual feature from Q-former and input to Flan-T5-XL.

5. Results

5.1. Comparison Results on Video QA and Event Prediction Task

We first evaluate our algorithms on NEXT-QA dataset. As shown in Table 1, VLAP improves performance over the state-of-the-art by 1.0% at 4 frames setting and push the accuracy to reach 75.5 % on this dataset at 32 frames setting. For different types of questions, we ...

Results on STAR Then we evaluate our algorithms on STAR dataset. As shown in Table 3, VLAP improves performance over the state-of-the-art by 1.6% at 4 frames setting and achieved new SOTA at 67.9 % on this dataset. For different types of questions, we ...

Results on VLEP To explore the event prediction ability, we further evaluate our algorithms on VLEP dataset. As shown in Table **??**, VLAP improves performance over the state-of-the-art by 0.7% at 4 frames setting.

Results on TVQA we also further evaluate our algorithms on large scale TVQA dataset. As shown in Table **??**, VLAP improves performance over the state-of-the-art by 1.8% at 4 frames setting.

5.2. Comparison Results on Video Captioning Task

5.3. Ablation Study

5.3.1 Components Effectiveness Ablation

We evaluate the effectiveness of each component in our method, as shown in Table 7.

5.3.2 Instruction-aware Frame Prompter Ablation Results

We evaluate the effectiveness of frame prompter in our method, as shown in Table 5.

5.4. Cross-modal Distillation Ablation Results

We evaluate the effectiveness of distillation in our method, as shown in Table 4.

5.5. Different Frame number Ablation Results

We evaluate the effectiveness of VLAP in term of different frames, as shown in Table 6.

6. Conclusion and Future Work

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 6

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 3

Method (Frames Number)	Temporal	Causal	Description	Average
Just Ask [30] (20) (ICCV2021)	51.4	49.6	63.1	52.3
All-in-One [23] (32) (CVPR2023)	48.6	48.0	63.2	50.6
MIST [9] (32) (CVPR2023)	56.6	54.6	66.9	57.1
HiTeA [34] (16) (Dec 2022)	58.3	62.4	75.6	63.1
InternVideo [25] (8) (Dec 2022)	58.5	62.5	75.8	63.2
BLIP-2 [15] (4) (ICML2023)	67.2	70.3	79.8	71.5
SeViLA [37] (4) (May 2023)	67.7	72.1	82.2	73.4
SeViLA [37] (8) (May 2023)	67.0	73.8	81.8	73.8
VLAP (4) (Ours)	70.1	73.8	82.1	**74.4**
VLAP (8) (Ours)	71.4	73.6	81.4	**74.8**
VLAP (16) (Ours)	69.5	74.0	81.7	**75.0**
VLAP (32) (Ours)	72.3	74.9	82.1	**75.5**

Table 1. **VLAP Results on Next-QA.**

Method (Frames Number)	Interaction	Sequence	Prediction	Feasibility	Average
Flamingo-9B 4-shot [1] (30) (NeurIPs2022)	-	-	-	-	42.8
All-in-One [23] (32) (CVPR2023)	47.5	50.8	47.7	44.0	47.5
MIST [9] (32) (CVPR2023)	55.5	54.2	54.2	44.4	51.1
InternVideo [25] (8) (Dec 2022)	62.7	65.6	54.9	51.9	58.7
BLIP-2 [15] (4) (ICML2023)	65.4	69.0	59.7	54.2	62.0
SeViLA [37] (4) (May 2023)	63.7	70.4	63.1	62.4	64.9
VLAP (4) (Ours)	69.3	70.0	63.9	64.3	**66.5**
VLAP (8) (Ours)	70.6	74.1	66.3	60.6	**67.9**

Table 2. **BLVQA Results on STAR.**

Method	F #	VLEP	TVQA
FrozenBiLM [31] (NeurIPs2022)	10	-	57.5
InternVideo [25] (Dec 2022)	8	63.9	57.2
BLIP-2 [15] (ICML2023)	4	67.0	54.5
SeViLA [37] (May 2023)	4	68.9	61.6
VLAP (Ours)	4	**69.6**	**63.4**

Table 3. **Our VLAP Results on VLEP and TVQA. F # means frames number.**

Teacher's Frame number	T	C	D	Average
4 frames	71.2	73.0	80.6	73.8
8 frames	71.0	72.9	82.5	74.3
16 frames	70.7	73.4	80.1	73.6
32 frames	70.1	73.8	82.1	74.4

Table 4. **VLAP Results on Next-QA.** Fewer frame teacher can maintain better temporal alignment. Lager frame teacher can offer better causal and detail description instruction. Temporal (T), Causal (C), Description (D)

Frame Prompter	T	C	D	Average
LN	68.5	70.9	79.3	72.4
LN+BN	70.1	73.8	82.1	74.4
LN+BN+LN	69.7	73.1	81.6	74.1

Table 5. **VLAP Results on Next-QA for frame prompter decoder.** Temporal (T), Causal (C), Description (D),

[3] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022. 2

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 3

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2

Methods	T	C	D	Average
VLAP (4)	70.1	73.8	82.1	74.4
VLAP (8)	70.8	74.0	82.4	74.7
VLAP (16)	-	-	-	-
VLAP (32)	-	-	-	-

Table 6. **VLAP Results on Next-QA for frame prompter decoder.** Temporal (T), Causal (C), Description (D),

Components	STAR	VLEP	TVQA	NextQA
base	62.0	67.0	54.5	71.5
base+CMA	64.9	68.6	62.2	73.5
base+CMA+IFP	66.5	69.6	63.4	74.4

Table 7. **VLAP components ablation, Cross-Modal Alignment (CMA), Instruction-aware Frame Prompter (IFP).**

[6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 2, 4

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[8] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 3

[9] Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14773–14783, 2023. 6

[10] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 5

[11] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 3

[12] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021. 5

[13] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 5

[14] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. *arXiv preprint arXiv:2010.07999*, 2020. 5

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2, 3, 4, 5, 6

[16] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[17] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 4

[18] Yuqi Liu, Luhui Xu, Pengfei Xiong, and Qin Jin. Token mixing: parameter-efficient transfer learning from image-language to video-language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1781–1789, 2023. 3

[19] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 3

[20] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 26462–26477. Curran Associates, Inc., 2022. 3

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[22] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: open and efficient foundation language models, 2023. *URL https://arxiv. org/abs/2302.13971*. 1

[23] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. All in one: Exploring unified video-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608, 2023. 6

[24] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2

[25] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 5, 6

[26] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 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. 4

[27] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 2

[28] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 4

[29] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment, 2023. 3

[30] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1686–1697, 2021. 6

[31] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35:124–141, 2022. 6

[32] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 2

[33] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 2

[34] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. Hitea: Hierarchical temporal-aware video-language pre-training. *arXiv preprint arXiv:2212.14546*, 2022. 6

[35] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 5

[36] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. arxiv 2022. *arXiv preprint arXiv:2205.01917*. 2

[37] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *arXiv preprint arXiv:2305.06988*, 2023. 2, 3, 5, 6