Analyzing Zero-Shot Abilities of Vision-Language Models on Video Understanding Tasks

Avinash Madasu
Intel Labs
avinash.madasu@intel.com

Anahita Bhiwandiwalla
Intel Labs
anahita.bhiwandiwalla@intel.com

Vasudev Lal
Intel Labs
vasudev.lal@intel.com

Abstract

Foundational multimodal models pre-trained on large scale image-text pairs or video-text pairs or both have shown strong generalization abilities on downstream tasks. However unlike image-text models, pretraining video-text models is always not feasible due to the difficulty in collecting large-scale clean and aligned data, and exponential computational costs involved in the pretraining phase. Therefore, the pertinent question to ask is: Can image-text models be adapted to video tasks and is there any benefit to using these models over pretraining directly on videos? In this work, we focus on this question by proposing a detailed study on the generalization abilities of image-text models when evaluated on video understanding tasks in a zero-shot setting. We investigate 9 foundational image-text models on a diverse set of video tasks that include video action recognition (video AR), video retrieval (video RT), video question answering (video QA), video multiple choice (video MC) and video captioning (video CP). Our experiments show that image-text models exhibit impressive performance on video AR, video RT and video MC. Furthermore, they perform moderately on video captioning and poorly on video QA. These findings shed a light on the benefits of adapting foundational image-text models to an array of video tasks while avoiding the costly pretraining step. The code is available at https://github.com/IntelLabs/multimodal_cognitive_ai/tree/main/Video-Zeroshot.

1 Introduction

Foundational vision-language (VL) models [39, 48, 52, 24, 26, 33] have achieved state-of-the-results on multi-modal tasks like image-retrieval [50, 34] image captioning [1, 34] and visual question answering [4, 11] etc. These models are pretrained on large amounts of image-text pairs using contrastive learning and then evaluated on these tasks in a zero-shot or few-shot setting. However, pre-training such large scale models is extremely hard for videos because: (i) Most of the current large scale video-text datasets [36] are very noisy and mis-aligned (i.e. video frames and captions don't match temporally). Therefore, its difficult to collect clean and aligned video-text data without human-in-the-loop unlike image-text data. (ii) Pre-training models on such video-text datasets require heavy computational resources. Hence given the remarkable transfer capabilities of contrastive image-text models, the pertinent question to explore is:

How beneficial are foundational image-text models when applied on video tasks in a zero-shot setting?
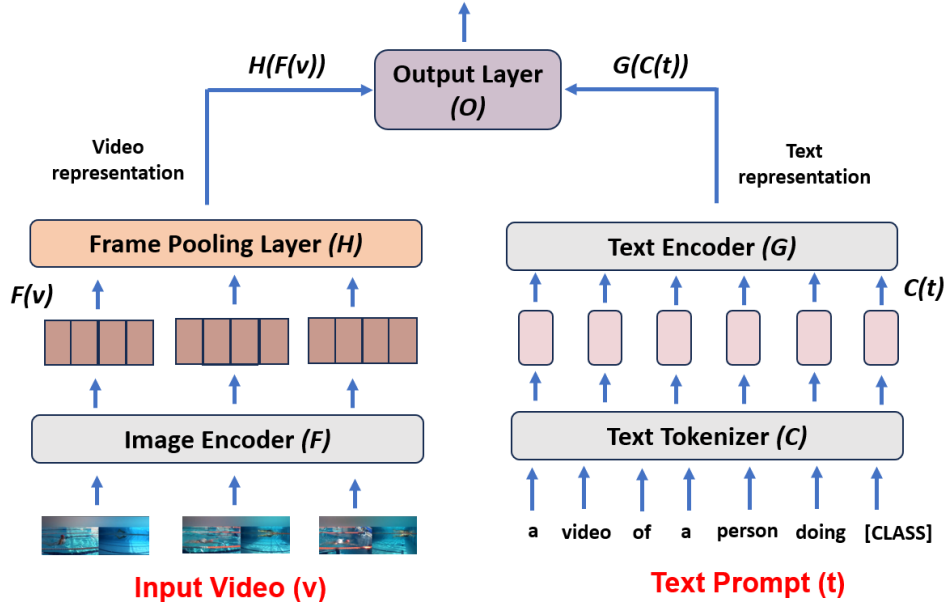
Figure 1: Overview of the architecture. It consists of four components (i) Image encoder *(F)* to get frame representations for the input video, (ii) Text encoder *(G)* to get text representations for the input text prompt, (iii) Frame pooling layer *(H)* to obtain pooled frame representations, (iv) Final output layer *(O)* to acquire the output.

To investigate this, we conduct a systematic study by evaluating a diverse set of image-text models on a wide range of video tasks. We propose a simple architecture (Fig 1) to effectively adapt the image-text models for video-tasks without any changes to the architecture. For this study, we utilize 9 foundational image-text models and test these extensively on 5 video understanding tasks namely video action recognition (video AR), video retrieval (video RT), video question answering (video QA), video multiple choice (video MC) and video captioning (video CP) in a zero-shot setting. In addition, we also benchmark the zero-shot performance of image-text to the existing state-of-the-art video-text models on each of the tasks. Our zero-shot evaluation results demonstrate that:

1. Foundational image-text models perform competitively to SotA video-text models on video action recognition, video retrieval and video multiple choice tasks.

2. Image-text models struggle on complex video reasoning tasks such as video question answering while demonstrating reasonable performance on video captioning.

3. The performance is heavily dependent on the size of the data used in pre-training image-text models and a frame count ranging from 12 to 20 is sufficient for the optimal score.

2   Related Work

Vision-Language Models (VLM): There has been a body of work [30, 28, 32, 27, 31, 13]that explore the phenomenon of pre-training on large image-text datasets [40, 8] and fine-tuning on downstream tasks. The goal is to learn generalized multi-modal representations applied to image understanding tasks [12, 7, 15, 20, 4, 1, 50]. Unlike the previous works, we explore the generalization abilities of image-text models on a broad range of video understanding tasks in a zero-shot setting.

Video understanding: Several benchmarks have been introduced such as video action recognition [42, 21, 41], video retrieval [46, 3, 9]video question answering [45, 2], video captioning [23, 54] etc. Transformer based models [5, 16, 29] pre-trained on large scale video datasets [5, 36] have been introduced and tested on these benchmarks. Our aim is to avoid the costly pre-training step and effectively adapt the foundational image-text models to videos.

Task	Dataset	Split	Test size	Metric	Frames	Classes
Action Recognition	Kinetics 700-2020 [41]	Val	30850	Accuracy	16	700
	UCF101 [42]	Test	13320	Accuracy	16	101
	HMDB51 [21]	Test	3783	Accuracy	16	51
Video Retrieval	MSRVTT [46]	Test	1000	R@1	16	1000
	MSVD [9]	Test	670	R@1	16	670
	DiDeMo [3]	Test	1002	R@1	32	1002
Video QA	MSRVTT [45]	Test	67770	Accuracy	16	1500
	MSVD [45]	Test	11983	Accuracy	16	1000
Video MC	TGIF Transition [19]	Test	6232	Accuracy	16	5
	TGIF Action [19]	Test	6232	Accuracy	16	5
	TGIF Frame [19]	Test	6232	Accuracy	16	1540
	MSRVTT [51]	Test	2981	Accuracy	16	4
Video Captioning	MSRVTT [46]	Test	2990	B-4, METEOR	16	Generation
	MSVD [9]	Test	690	B-4, METEOR	16	Generation

Table 1: **Summary of the tasks and the datasets used in zero-shot evaluation.** The table also illustrates the dataset split, size of the test data, evaluation metric, number of video frames and number of classes for each of the datasets. R@1 denotes Recall-1, B-4 is the average of BLEU-1, BLEU-2, BLEU-3 and BLEU-4.

3 Methodology

The primary question we aim to investigate in this work is: **How beneficial are image-text representations to downstream video tasks in a zero-shot setting?** To explore this, we propose a simple architecture that takes video and prompt text as inputs and returns the most similar class label or generates a text caption. As shown in the Figure 1, it consists of four components: image encoder *(F)*, text encoder *(G)*, frame pooling layer *(H)* and final output layer *(O)*. We leverage the image and text encoders of the respective models listed in Section 4.2. Additionally we add a frame pooling layer and also an output layer.

Image Encoder *(F)*. Given a video v, we sample N frames denoted as $v = [v_1, v_2, .., v_N]$. We then employ a pre-trained image encoder *(F)* to get the representations for each of the sampled video frames $F(v) = [F(v_1), F(v_2), ..., F(v_N)]$.

Text Encoder *(G)*. Given a text prompt t, we use a pre-trained text encoder *(G)* to obtain the text representations from the tokenized input text.

Frame Pooling Layer *(H)*. To obtain the final video representation, we leverage a frame pooling layer which computes the mean of all the frame representations.

$$H(F(v)) = \frac{\sum_{i=1}^{N} F(v_i)}{N} \tag{1}$$

Output Layer *(O)*. The output layer takes the video and text representations and outputs a class label with the maximum cosine similarity score for classification tasks or generates a text for captioning tasks.

4 Experimental Settings

4.1 Tasks

We analyze five different video understanding tasks: video action recognition (video AR), video retrieval (video RT), video question answering (video QA), video multiple choice (video MC) and video captioning (video CP). Table 1 summarizes the tasks, datasets for each task category, dataset split, metrics used for evaluation, number of frames and classes for each dataset. Below, we list the evaluation datasets for each of the tasks.

Task	Dataset	Prompt

Action Recognition	Kinetics 700-2020
	UCF101
	HMDB51
		`a video of a person doing {CLASS NAME}`
		`a video of a person doing {CLASS NAME}`
		`a demonstration of a person doing {CLASS NAME}`

Video Retrieval	MSRVTT MSVD DiDeMo	`{QUERY}` `{QUERY}` `{QUERY}`
Video QA	MSRVTT - QA MSVD - QA	`Question: {QUESTION} Answer: {ANSWER CANDIDATE}` `Question: {QUESTION} Answer: {ANSWER CANDIDATE}`
Video MC	TGIF Transition TGIF Action TGIF Frame MSRVTT - MC	`{QUESTION} {ANSWER CANDIDATE}` `{QUESTION} {ANSWER CANDIDATE}` `{QUESTION} {ANSWER CANDIDATE}` `{QUESTION} {ANSWER CANDIDATE}`
Video Captioning	MSRVTT MSVD	`A video of` `A video of`

Table 2: **Summary of the tasks and prompts used in zero-shot evaluation.** The curly brackets in a prompt is replaced with all the class names for action recognition and answer candidates for Video QA and Video MC tasks.

Video Action Recognition (Video AR): Kinetics 700-2020 [41], UCF-101 [42] and HMDB-51 [21].
Video Retrieval (Video RT): MSRVTT [46], MSVD [9] and DiDeMo [3].
Video Question Answering (Video QA): MSRVTT-QA [45] and MSVD-QA [45].
Video Multiple Choice (Video MC): TGIF-Transition, TGIF-Action, TGIF-Frame [19] and MSRVTT-MC [51].
Video Captioning (Video CP): MSRVTT [46] and MSVD [9].

4.2 Models

For experiments we use the following nine foundational image-text models: ALIP [49], CLIP [39], OpenCLIP [18], SLIP [37], LaCLIP [14], BLIP-2 [25], InstructBLIP [10], OFA [44] and Pix2Struct [22].

4.3 Evaluation Metrics

For the tasks of video action recognition, video question answering and video multiple choice we use accuracy as the evaluation metric. We report the results using the standard R@1 (Recall@1) metric in case of video retrieval. We employ multiple metrics: B-4 [38], METEOR [6] and CIDEr [43] for the video captioning task. B-4 is the average of BLEU-1, BLEU-2, BLEU-3 and BLEU-4.

4.4 Implementation

Table 2 summarizes the text prompts used for each of the video tasks.

Video Action Recognition: For the action recognition task, we use the prompts "`a video of a person doing {CLASS NAME}`" and "`a demonstration of a person doing {CLASS NAME}`". The field `{CLASS NAME}` is substituted with all the class names present in the dataset and the cosine similarity is measured for the input video and all the classes. The predicted class is the one with maximum cosine similarity among all the candidate classes.

Video Retrieval: In case of video retrieval, the input query is used as the prompt and its similarity is calculated with all the videos in the database.

Video QA and Video MC: Regarding the tasks of video QA and video MC, the fields `{Question}` and `{ANSWER CANDIDATE}` in the prompt "`Question: {QUESTION} Answer: {ANSWER CANDIDATE}`" are replaced with input question and all the possible answers respectively.

Model	Type	#PT	Video AR			Video Retrieval			Video QA	
			K700	UCF-101	HMDB-51	MSRVTT	MSVD	DiDeMo	MSRVTT	MSVD
ALIP [49]	IT	15M	15.7	30.3	21.2	15.3	33.9	14.3	0.1	**7.7**
CLIP [39]	IT	400M	**40.5**	**63.6**	**46.2**	29.9	44.2	22.6	1.5	6.5
OpenCLIP [18]	IT	2B	37.4	56.4	38.1	31.7	**48.2**	22.6	**2.9**	**7.7**
SLIP [37]	IT	15M	13	15.2	19.8	14.1	30.3	17.2	1.5	4.1
LaCLIP [14]	IT	400M	39.1	59.5	38.8	**32**	43.6	**23.3**	0.1	5.5
VIOLET [17]	VT	186M	-	-	-	34.5	48.3	32.6	-	-
LAVENDER [29]	VT	30M	-	-	-	-	-	-	2.7	9.2
MAXI [35]	VT	GPT3	-	78.2	52.3	-	-	-	-	-
Δ	-	-	-	-14.6	-5.9	-2.8	-0.1	-9.3	+0.2	-1.5

Table 3: **Results of zero-shot video action recognition, video retrieval and video question answering. Type** denotes whether the model is pre-trained on image-text (IT) or video-text (VT) data. **#PT** implies the size of data used in pre-training. K700 is the kinetics 700-2020 [41] dataset. Δ is the difference between the highest performance of image-text models and state-of-the-art video-text model.

Model	Type	TGIF-Transition	TGIF-Action	TGIF-Frame	MSRVTT-MC
ALIP [49]	Image-text	39.1	38.5	0.1	62.4
CLIP [39]	Image-text	**51.3**	**50.5**	**5.8**	74.3
OpenCLIP [18]	Image-text	41.2	45.1	4.2	73.3
SLIP [37]	Image-text	37.2	35.9	2.0	61.9
LaCLIP [14]	Image-text	41.6	46.1	2.6	**74.4**
LAVENDER [29]	Video-text	53.8	55.1	19.6	87.2
Δ	-	-2.5	-4.6	-13.8	-12.8

Table 4: **Results of zero-shot evaluation on video MC.** Δ is the difference between the highest performance of image-text models and state-of-the-art video-text model.

The answer candidate for which the similarity between the input video and the text prompt is maximum is chosen as the predicted answer.

Video Captioning: As for video captioning, the prompt "A video of" along with the video are used as inputs for generating captions.

5 Results and Discussion

5.1 Zero-shot image-text models are competitive to SOTA video-text models on video AR, RT and MC tasks

In Table 3, we present the results of image-text models on video action recognition and video retrieval. Table 4 illustrates the results on video multiple choice. From the tables, we observe that image-text models perform competitively to state-of-the-art video-text models on most of the tasks. On the task of video action recognition, image-text models lag behind SotA video-text model by 14.6% for UCF-101 and 5.9% for HMDB-51 dataset. Note that, MAXI [35] is pretrained on significantly larger video datasets compared to image-text models and is explicitly modelled to learn actions in them.

For the task of video retrieval, image-text models achieves comparable results to SotA video-text model. In fact, image-text models trail only by 2.8% and 0.1% for MSRVTT and MSVD respectively. In-case of DiDeMo, the difference in performance (9.2%) is quite significant. Observe that, MSRVTT and MSVD are shorter length video datasets whereas DiDeMo is a paragraph to video retrieval dataset.

In the case of video MC, we see that image-text models under perform by just 2.5% and 4.6% in case of TGIF-transition and TGIF-action datasets respectively. TGIF is a GIF dataset and are typically easier compared to videos and in addition the number of choices to select from in these datasets are limited to just 5. Hence, the image-text models were able to achieve similar performance to SotA video-text models on these relatively easier datasets. In contrast even though TGIF-frame is a

Model	Type	#PT	MSRVTT			MSVD		
			B-4	METEOR	CIDER	B-4	METEOR	CIDER
BLIP2 [25]	Image-text	129M	**35**	**18**	**19**	**35.6**	**22.2**	**27.4**
InstructBLIP [10]	Image-text	129.5M	26.2	12.9	11.8	27.1	16.2	17.8
OFA [44]	Image-text	36.27M	18.7	10.5	4.8	17.3	11.8	-
Pix2Struct [22]	Image-text	80M	23.3	14.4	-	19.3	15.3	-
Vid2Seq [47]	Video-text	1B + FT	-	30.8	64.6	-	45.3	146.2
Δ	-	-	-	-12.8	-45.6	-	-23.2	-118.8

Table 5: **Results of zero-shot evaluation on video captioning.** Note that Vid2Seq model is pretrained on YT-Temporal-1B [53] and then fine-tuned on MSRVTT and MSVD datasets respectively. Δ is the difference between the highest performance of image-text models and state-of-the-art video-text model.

(a) Video Action Recognition
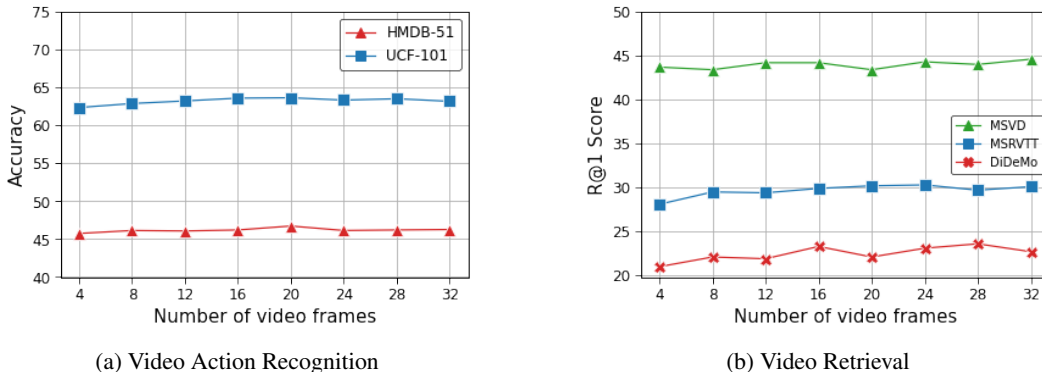

(b) Video Retrieval

Figure 2: Figure shows the effect of number of video frames on the zero-shot performance of CLIP for video action recognition and video retrieval. Figures for the additional tasks and image-text models are included in the appendix.

GIF dataset, the frequency of choices is 1000 which are significantly higher. Therefore, we spot a notable difference (13.8%) in performance between image-text and video-text models. We would like to add that the video-text models also perform considerably worse on TGIF-frame compared to TGIF-transition and TGIF-action. On MSRVTT-MC despite the limited multiple choices (i.e. 4), SotA video-text models evidently outperforms image-text models by 12.8%.

5.2 Zero shot image-text models perform reasonably on video captioning.

In Table 5, we provide a comparison between zero-shot image-text models and fine-tuned video-text models on the task of video captioning. As seen in the table, it is clear that zero-shot image-text models (BLIP-2) perform reasonably to video-text models despite pre-training on just 13% data and not fine-tuning on respective datasets.

5.3 Image-text and video-text models perform worse on video QA in a zero-shot setting

Table 3 illustrates the zero-shot results of image-text and video-text models on the video QA task. From the table, it is clear that video-text models marginally outperform image-text models in zero-shot video QA. It is worth noting that both image-text and video-text perform poorly on QA datasets. Video QA is a complex task which requires modelling of object and attribute relationships in the videos. This could be the reason zero-shot models show subpar performance.

5.4 Size of pre-training data matters for image-text models

We analyze the effect of the size of the pre-training data on the zero-shot performance of image-text models on the various video tasks. We compare image-text models like ALIP and SLIP pretrained

(a) Video Action Recognition
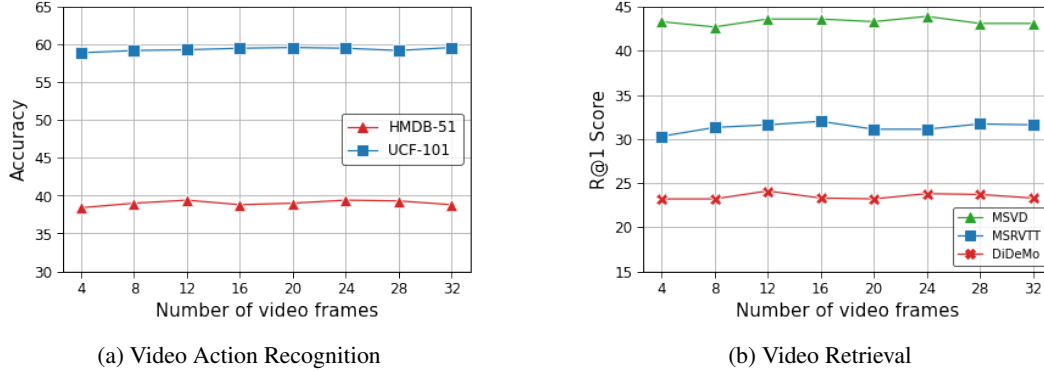
(b) Video Retrieval

Figure 3: Figure shows the effect of number of video frames on the zero-shot performance of LaCLIP for video action recognition and video retrieval.

on 15M image-text pairs, CLIP and LaCLIP pretrained on 400M and OpenCLIP pretrained on 2B image-text pairs. On video AR and video MC as summarized in the Tables 3 and 4 respectively, we observe models trained with 400M data consistently significantly outperform the models trained with 15M data across all three evaluation datasets. These models also slightly outperform OpenCLIP trained on 400% more training data. On the task of video retrieval, we note that larger pre-training data does have more impact on the zero-shot performance gains on the MSVD dataset. In case of MSRVTT and DiDeMo, the models pretrained on 400M and 2B are comparable in performance. In all the evaluations, we observe that models trained on 15M image-text pairs performs 50% worse than that of the models pretrained on 400M. On video QA, we observe that OpenCLIP performs remarkably well on MSRVTT compared to its counterparts which are trained on comparatively less data. However, on MSVD we see comparable performance between OpenCLIP, ALIP and SLIP trained on much lesser image-text pairs. When analyzing the performance on video captioning tasks in 5, we see BLIP2 trained with 129M image-text pairs outperforms OFA and Pix2Struct trained with lesser data (36.27M and 80M respectively). Overall we do observe a correlation between the amount of data used during pre-training and the zero-shot performance. However, the few outliers that we note do invite further analysis.

5.5 Impact of frame count on zero-shot performance of image-text models

In this section, we study the impact of video frame count on the zero-shot performance of image-text models. Figure 2 demonstrates the results of this study for CLIP model on video action recognition and video retrieval tasks. Our results indicate that the zero-shot performance increases with higher frame frequency and reaches an optimal point without further increase. We observe this effect for all the datasets of video action recognition and video retrieval. These results illustrate that frame count ranging from 12 to 20 offer the optimal zero-shot performance for most of the video tasks.

6 Conclusion

In this work, we presented a comprehensive analysis on the generalization capabilities of foundational image-text models on a broad range of video understanding tasks in a zero-shot setting. We evaluate nine foundational image-text models on five video tasks namely video action recognition, video retrieval, video question answering, video multiple choice and video captioning and also benchmark them against the state-of-the-art video-text models on these tasks. Our experiments demonstrate that image-text models accomplish competitive results to video-text models on video action recognition, video retrieval and video multiple choice. Moreover, the results also show that image-text models show reasonable performance on video captioning and perform worse on video question answering. We think that the future works can benefit from these findings in designing models for video understanding tasks.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019.

[2] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh. Audio-visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021.

[6] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[7] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer, 2014.

[8] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[9] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011.

[10] 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.

[11] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022.

[14] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *arXiv preprint arXiv:2305.20088*, 2023.

[15] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

[16] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. 2021.

[17] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. An empirical study of end-to-end video-language transformers with masked visual modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22898–22909, 2023.

[18] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below.

[19] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017.

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[21] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[22] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023.

[23] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer, 2020.

[24] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963, 2022.

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. 2023.

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven CH Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. 2022.

[27] 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.

[28] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, 2020.

[29] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23119–23129, 2023.

[30] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: Asimple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[31] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.

[32] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.

[33] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2021.

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[35] Wei Lin, Leonid Karlinsky, Nina Shvetsova, Horst Possegger, Mateusz Kozinski, Rameswar Panda, Rogerio Feris, Hilde Kuehne, and Horst Bischof. Match, expand and improve: Unsupervised finetuning for zero-shot action recognition with language knowledge. *arXiv preprint arXiv:2303.08914*, 2023.

[36] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019.

[37] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer, 2022.

[38] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[39] 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.

[40] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[41] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the kinetics-700-2020 human action dataset. *arXiv preprint arXiv:2010.10864*, 2020.

[42] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[43] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[44] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.

[45] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.

[46] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[47] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10714–10726, 2023.

[48] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19163–19173, June 2022.

[49] Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Alip: Adaptive language-image pre-training with synthetic caption. *arXiv preprint arXiv:2308.08428*, 2023.

[50] 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.

[51] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018.

[52] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[53] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022.

[54] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598, 2018.
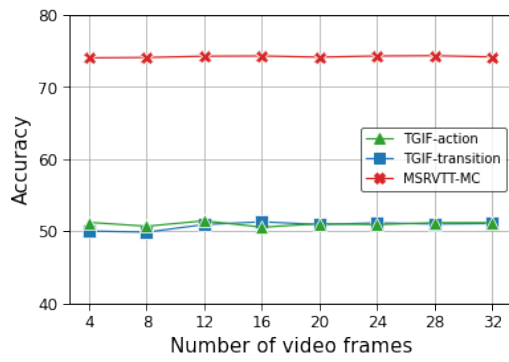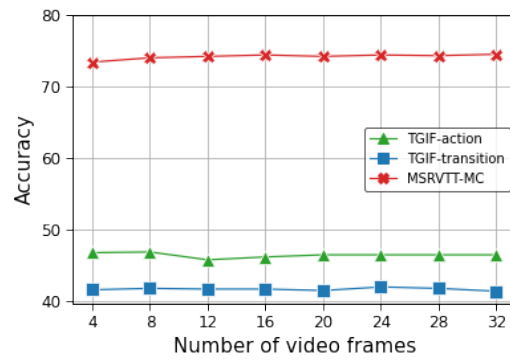
(a) CLIP   (b) LaCLIP

Figure 4: Figure shows the effect of number of video frames on the zero-shot video multiple choice performance of CLIP and LaCLIP.