Supplementary Materials

We first discuss the limitations, ethcial concerns and broader impact of this work (Section A). We detail the datasets (Section B), models (Section C), and training setups (Section D) in the supplementary materials to improve this work's reproducibility. Besides, Section E includes more experimental studies to strengthen the main text.

A Limitation, ethical concern, and broader impact

Limitation. VideoGLUE covers various unimodal video tasks and could be strengthened by adding multimodal tasks like video question answering. We chose three representative FM adaptation methods and used them to provide as uniform experiment protocols for different FMs as possible. However, some of our observations could be flipped with the evolution of adaptation methods, which are an active research area. We proposed a scalar score, VideoGLUE Score (VGS), to capture the efficacy and efficiency of an FM on video understanding. However, VGS might be dominated by one or a few datasets — when it becomes a serious issue, we should probably improve the score and/or retire the other datasets from future versions of VideoGLUE. Indeed, VGS is not a perfect score that covers all aspects of FMs in a comprehensive manner. For example, it does not account for an FM's model size, model architecture, etc. We hope future research will lead to new metrics to complement VGS and a more comprehensive evaluation of FMs for visual tasks.

Ethical concern. We evaluate FMs on three video tasks, eight datasets in total. We select the tasks and datasets based on their popularity and representativeness. Although carefully designed, our benchmark inevitably inherited some ethical concerns from those datasets. For instance, many of the datasets are curated by crawling videos from the Internet, which do not proportionately represent the experiences of the global population and can potentially lead to biased evaluations of FMs. Moreover, the video datasets involve human daily activities, leading to privacy concerns about the human actors in the videos. How to evaluate FMs for video understanding in a fair and privacy-preserving manner could be an important direction for future research.

Broader impact. Our research reveals the need and tremendous opportunities to research video-first FMs by improving pretraining video data and methodologies. Our studies on different adaptation methods on versatile tasks confirms that both tasks and adaptation methods matter when it comes to the evaluation of FMs, shedding light on the already vibrant area of FM adaptations. Finally, we hope our research could inspire research on foundation models development and video understanding in general, along with their applications in the real world.

B Video understanding datasets

B.1 Appearance-focused action recognition

Video classification is a task of classifying videos into pre-defined labels, with the major focus on human actions.

Kinetics400 (Kay et al., 2017) (K400) is a large-scale, high-quality video dataset widely used as a standard video classification benchmark. It contains more than 250k video clips with annotations of 400 human daily actions. The actions are human focused and cover a broad range of classes including human-human interactions and human-object interactions. Although the video clips span 10 seconds on average, many studies (Sevilla-Lara et al., 2021; Wang et al., 2018) have pointed out the task could be easily solved on the Kinetics datasets by inferring from the static objects appeared or background environment — motion information is less important than the visual appearance. Hence, we categorize Kinetics400 as an appearance-focused action classification dataset.

Moments-in-Time (Monfort et al., 2019) (MiT) is a large-scale video event classification dataset, with one million human annotated short video clips (around 3 seconds each). The temporal span corresponds to the averaged duration of human working memory and is a temporal envelope holding meaningful actions between people, objects, and phenomena. Videos in MiT are annotated with 339 most used verbs in the English vocabulary.

B.2 Motion-focused action recognition

Videos contain much more commonsense knowledge than still images do, such as an object's motion patterns and the causal consequences of an action, just to name a few. However, appearance-based benchmarks do not evaluate a model's understanding of such commonsense knowledge, complex scenes, and situations. In observance of this, some video datasets have been proposed and studied in recent years with the focus on motions and common-sensing reasoning that are prosperous in video data.

Something-something v2 (Goyal et al., 2017) (SSv2) is a collection of around 200k videos of human performing pre-defined, basic actions with everyday objects. There are 174 unique labels in total depicting atomic hand manipulations, like putting something into something, turning something upside down or covering something with something. This dataset benchmarks a model's fine-grained understanding capability of object motions and scene changes by making the label space atomic-action-focused and background-invariant.

Diving48 (Li et al., 2018) (D48) is introduced to evaluate a model's dynamic reasoning capability. The video clips in this dataset are obtained by segmenting online videos of major diving competitions. In total, there are around 18k videos annotated with 48 classes. Because of its standardization, the diving scenario is purposefully chosen to avoid the scene, object, and person biases.

B.3 Multi-label daily action classification

Most of current action classification datasets involve video clips with a clean snapshot of a single action. In contrast, humans perform daily complex activities step-by-step, simultaneously, or in an interleaving manner. Towards more comprehensive human daily activity reasoning, Charades (Sigurdsson et al., 2016) is introduced. Different from web-collected datasets whose contents are more structured, Charades is collected by crowd-sourcing from hundreds of actors recording their videos in their own homes, acting out casual everyday activities. Charades brings in more diversity into the video classification task due to its close-to-daily-life setting. Its videos are 30 seconds long on average and have multi-label annotations testing models' understanding of complex daily activities with multiple steps. Charades provides 110k videos with 157 action classes for training and evaluation.

B.4 Temporal action localization

Natural long videos contain scene changes and semantic shifts, while most of the existing video benchmarks formulate problems to focus on trimmed video clips. Such a gap introduces evaluation bias as clip-level benchmarks could not reflect a model's temporal feature discriminativeness, which is of key importance to solve long-form video understanding tasks. To comprehend the study on foundation models' video capabilities, we include the temporal action localization (TAL) task in our evaluation. The task of TAL is to predict not only the action labels but also each action instance's temporal boundary in untrimmed videos. We adopt ActivityNet v1.3 (Fabian Caba Heilbron & Niebles, 2015) as the dataset for the TAL task, which contains $10,002$ untrimmed videos in training and $4,985$ in validation. The video length in this dataset is between 5-10 minutes. In total, there are 200 types of activities annotated.

B.5 Spatiotemporal action localization

Spatiotemporal Action Localization (STAL) is a person-centric task that asks a system to localize actors and predict their atomic actions (Barker & Wright, 1955; Gu et al., 2018) in a transitory duration.

In AVA (Gu et al., 2018), 15 minutes long movie clips are densely annotated at 1Hz. In the key frames, every person is localized using a bounding box and labels corresponding to actions being performed by the actor. The label vocabulary consists of 80 different atomic visual actions. There are 430 different movies in total.

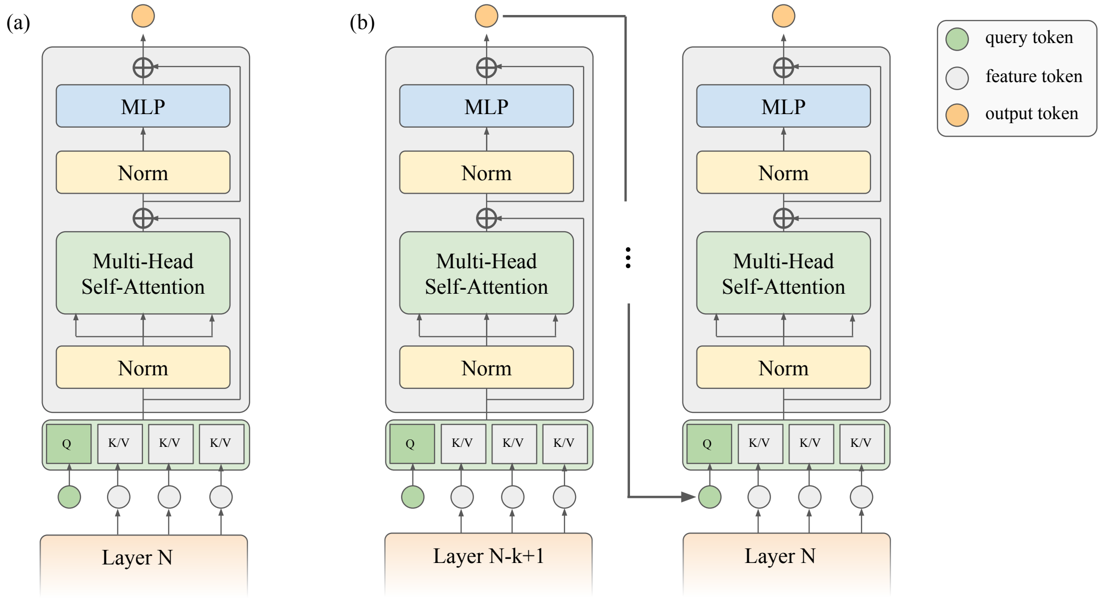

Figure 1: (a) Single-layer pooler head and (b) multi-layer attention pooling head for video classification and spatiotemporal action localization.

Table 1: Early vs. late fusion on image-native FMs. In this experiment, the frozen feature with a single-layer pooler head is used.

	K400		SSv2	
Method	Early	Late	Early	Late
CoCa	72.7	61.4	41.5	33.3
CLIP	70.5	75.2	38.1	41.0
FLAVA	67.9	71.3	40.4	40.6

AVA-Kinetics (Li et al., 2020) follows the same labeling protocol as AVA, while its data source comes from the Kinetics700 (Kay et al., 2017) video pool. The dataset contains over 230k clips annotated with the 80 AVA action classes for each of the humans in key frames.

C Model details

C.1 Task head architectures

In Figure 1, we plot the task heads used in our video classification and spatiotemporal action localization experiments, namely, the simple pooler head and multi-layer attention pooling head. For temporal localization, please refer to (Xu et al., 2020) for the task head's detailed architecture.

Figure 2 illustrates the encoder adapter layer's architecture. In the the adapter layer, only the down-sample layer, up-sample layer, and the scaling factor are tunable.

C.2 Image-to-video adaptation

Adapting image backbones to video tasks requires us to fuse the image embeddings at some point in the network and also introduce additional temporal information.

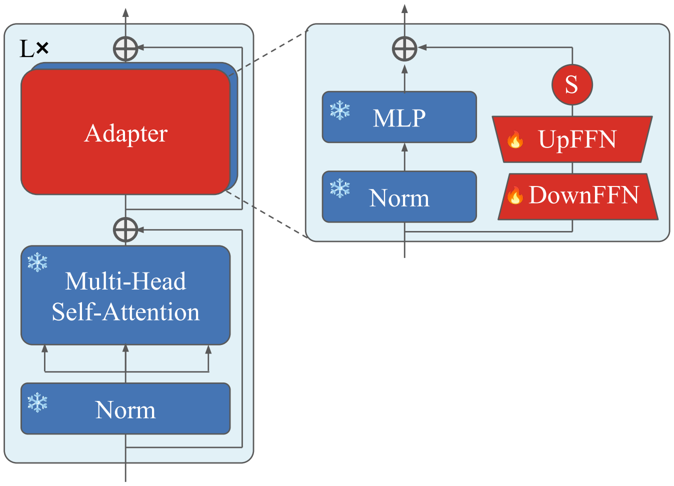

Figure 2: The adapter used in vision transformer. In the adapter layer, only the down-sample layer, up-sample layer, and the scaling factor are tunable. Between the down-sample layer and up-sample layer, an activation function is applied, which in our case is ReLU.

Table 2: Ablation study on the temporal positional embedding for image-to-video adaption. We choose FLAVA (Singh et al., 2022) with the frozen feature setting in this experiment.

Temporal Positional	VC (A)		VC (M)		VC (ML)
Embedding	K400	MiT	D48	SSv2	Charades
✗	71.3	29.7	41.6	30.3	10.7
✓	71.3	29.7	45.9	40.6	12.6

We consider two choices, early-fusion and late-fusion, and ablate them in the frozen feature setting in Table 1. In both early-fusion and late-fusion, we first apply the projection layer on each frame independently to embed pixel patches into embedding tokens. We then average-pool the embedding tokens from nearby frames to reduce the sequence length to $n \times h \times w$. In the early-fusion setting, we pass all tokens *together* to the image backbone to extract video features. In late-fusion, we pass each set of $h \times w$ tokens *independently* to the image backbone. Empirically, we find that the FLAVA (Singh et al., 2022) and CLIP (Radford et al., 2021) models do better with late-fusion while CoCa (Yu et al., 2022) does better with early-fusion.

Furthermore, we ablate the importance of temporal information using the frozen-features from FLAVA (Singh et al., 2022). In Table 2, we find that adding temporal positional embedding to the input is essential for D48 (Li et al., 2018), SSv2 (Goyal et al., 2017), and Charades (Sigurdsson et al., 2016) while not necessary for K400 (Kay et al., 2017) and MiT (Monfort et al., 2019). This supports our grouping that K400 and MiT are appearance-focused datasets.

Based on these findings, we use late-fusion for FLAVA (Singh et al., 2022) and CLIP (Radford et al., 2021) and early-fusion for CoCa (Yu et al., 2022). We add learnable temporal positional embeddings for all the image-native FMs.

D Task-specific hyperparameters

In the following, we provide experiment settings and hyperparamters we used in this study. In Table 3, we list the hyperparameters we applied in the video classification task. In Table 4, we present the hyperparameters we used on spatiotemporal action localization. In Table 5, we present the hyperparameters we used on temporal action localization task.

Table 3: Experimental configurations for video classification tasks. We let learning rate and weight decay to be tunable per model to allow some flexibility for task adaptations.

Config	**Kinetics400**	**Sth-sth v2**	**MiT**	**Diving48**	**Charades**
batch size	256	256	256	256	256
training epochs	150	50	50	100	50
ViT sequence length	8 × 14 × 14	8 × 14 × 14	8 × 14 × 14	8 × 14 × 14	8 × 14 × 14
optimization					
optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
optimizer momentum	0.9	0.9	0.9	0.9	0.9
learning rate schedule	cosine decay	cosine decay	cosine decay	cosine decay	cosine decay
warmup ratio	5%	5%	5%	5%	5%
data augmentations					
random horizontal flip	true	false	true	true	false
aspect ratio	(0.5, 2.0)	(0.5, 2.0)	(0.5, 2.0)	(0.5, 2.0)	(0.5, 2.0)
area ratio	(0.3, 1.0)	(0.3, 1.0)	(0.3, 1.0)	(0.3, 1.0)	(0.3, 1.0)
RandAug	(9, 0.5)	(9, 0.5)	-	-	-
MixUp	0.8	0.8	-	-	-
CutMix	1.0	1.0	-	-	-
evaluation					
multi-clips	4	1	4	4	4
multi-views	3	3	3	3	3
segment-based sample	false	true	false	false	false

We performed a greedy search on the learning rate and weight decay in all our experiments while keeping most other hyperparameters (e.g., data augmentation magnitude, dropout rate, drop path rate, etc.) consistent across different models and datasets. Specifically, we start with learning rate 1e-4 and weight decay 1e-5 and uniformly sample learning rates and weight decay factors with a rate of 5 and 10, respectively, centered around the starting points. After the first round, we pick the best-identified learning rate and weight decay factor as the new starting point and conduct another round of sampling with a rate of 2. We repeat another two to three rounds of hyperparameter search (with a rate of 2) until the model's performance converges. This process is a trade-off between computation costs and thoroughly examining an FM's performance under each experiment setup. The search ranges for the learning rate and weight decay are [4e-5, 2.5e-3] and [1e-6, 1e-4], respectively. We found that the learning rate is the most crucial factor when adapting an FM to downstream video understanding tasks.

E More studies

E.1 Large model adaptations

For the completeness of this report and reader's reference, in Table 6 we report experimental results under our settings with large FMs under the frozen backbone with one pooler head setup.

VideoMAE-v2-B/DL (Wang et al., 2023) denotes the ViT-B model distilled from ViT-g on the Kinetics710 datasets[1]. VideoMAE-v2-g (Wang et al., 2023) is the model that pretrained on UnlabeledHybrid dataset, while VideoMAE-v2-g/FT (Wang et al., 2023) conducts further finetuning using supervised training on Kinetics710. InternVideo-v2-g (Wang et al., 2024) and VideoPrism-g (Zhao et al., 2024) are two video

[1] https://github.com/OpenGVLab/VideoMAEv2/blob/master/docs/MODEL_ZOO.md

Table 4: Experimental configurations for spatiotemporal action localization.

Config	**AVA v2.2**	**AVA-Kinetics**
batch size	256	256
training epochs	50	50
ViT sequence length	8 × 16 × 16	8 × 16 × 16
optimization		
optimizer	AdamW	AdamW
optimizer momentum	0.9	0.9
layer decay	0.75	0.75
learning rate schedule	cosine decay	cosine decay
warmup ratio	5%	5%
data augmentations		
random horizontal flip	true	true
random scale	(0.5, 2.0)	(0.5, 2.0)
random color augmentation	true	true

Table 5: Experimental configurations for temporal action localization.

Config	**ActivityNet v1.3**
batch size	32
training epochs	10
feature extraction	
fps	15
per-clip length	16
clip stride	16
optimization	
optimizer	AdamW
optimizer momentum	0.9
learning rate schedule	cosine decay

Table 6: Evaluating large-scale FMs when using frozen feature with a one-layer pooler head. We report the Top-1 accuracy on K400, MiT, D48, SSv2 and MAP on Charades.

	VC (A)		VC (M)		VC (ML)
Model	K400	MiT	D48	SSv2	Charades
InternVideo-L	78.6	33.7	69.6	67.4	20.9
VideoMAE-v2-B/DL	86.7	38.9	61.4	57.7	33.2
VideoMAE-v2-g	59.7	20.7	42.5	44.2	12.7
VideoMAE-v2-g/FT	82.1	35.0	60.5	56.1	22.4
InternVideo-v2-g	85.0	43.0	53.1	61.6	40.9
VideoPrism-g	86.6	44.7	66.1	67.4	61.0

foundation models with multi-stage pre-training on curated in-house web video data. For InternVideo-v2-g, we use their stage-2 checkpoint[2]. For videoPrism-g, we use their final checkpoint.

[2] https://github.com/OpenGVLab/InternVideo/blob/main/InternVideo2/multi_modality/MODEL_ZOO.md

Table 7: Benchmark FMs adaptation on video understanding tasks under sample-efficient transfer learning. This table shows Top-1 classification accuracy and the relative accuracy (shown in the bracket). Results are achieved by using frozen features with pooler head.

	K400			SSv2		
Method	1%	10%	100%	1%	10%	100%
CoCa	27.1(37.8%)	48.9(67.0%)	73.1	5.6(13.4%)	20.9(50.4%)	41.5
CLIP	36.9(46.2%)	66.8(83.6%)	79.0	8.7(19.3%)	25.1(55.5%)	45.3
FLAVA	14.4(20.2%)	35.8(50.3%)	71.3	7.2(17.7%)	14.3(35.3%)	40.6
VideoMAE	15.5(23.9%)	32.0(49.2%)	65.0	13.7(25.4%)	30.3(56.2%)	53.9
InternVideo	20.4(29.5%)	50.2(72.4%)	69.3	19.5(33.6%)	41.1(70.7%)	58.2
VATT	34.1(45.4%)	63.7(84.8%)	75.1	12.9(22.4%)	37.6(65.0%)	57.8

E.2 Sample-efficient transfer learning

A strong FM should be able to adapt to downstream tasks with a few training samples. In this section, we test the adaption ability of FMs in a sample-efficient transfer learning setting. Particularly, we freeze backbones and train a pooler head to adapt the FMs on K400 and SSv2. For either dataset, we sample 1% and 10% data from the training set uniformly for training and evaluate on the full evaluation dataset.

We show our experimental results in Table 7. To better understand the data efficiency, we also show the relative Top-1 accuracy for each model (shown in the bracket), which is defined as the ratio between accuracy with fewer training examples and the accuracy achieved using all the training data. A higher relative Top-1 accuracy means the performance of the model is closer to its "full" capacity under the sample-efficient setting. We notice that the best performed model on each dataset in fully fine-tuned model also performs best in the few-shot setting. Especially, CLIP (Radford et al., 2021) achieves 46.2% and 83.6% relative Top-1 accuracy on K400 using only 1% and 10% of the training data, respectively. On SSv2, InternVideo (Wang et al., 2022) achieves 33.6% and 70.6% relative Top-1 accuracy with only 1% and 10% of the training data.

References

Roger G Barker and Herbert F Wright. Midwest and its children: The psychological ecology of an american town. *Marriage and family living*, 1955.

Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–970, 2015.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pp. 5842–5850, 2017.

Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6047–6056, 2018.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Ang Li, Meghana Thotakuri, David A Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*, 2020.

Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 513–528, 2018.

Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–8, 2019. ISSN 0162-8828. doi: 10.1109/TPAMI.2019.2901464.

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*, pp. 8748–8763. PMLR, 2021.

Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. Only time can tell: Discovering temporal data for temporal modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 535–544, 2021.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 510–526. Springer, 2016.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15638–15650, 2022.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11): 2740–2755, 2018.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14549–14560, 2023.

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.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024.

Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10156–10165, 2020.

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

Long Zhao, Nitesh B Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, et al. Videoprism: A foundational visual encoder for video understanding. *arXiv preprint arXiv:2402.13217*, 2024.