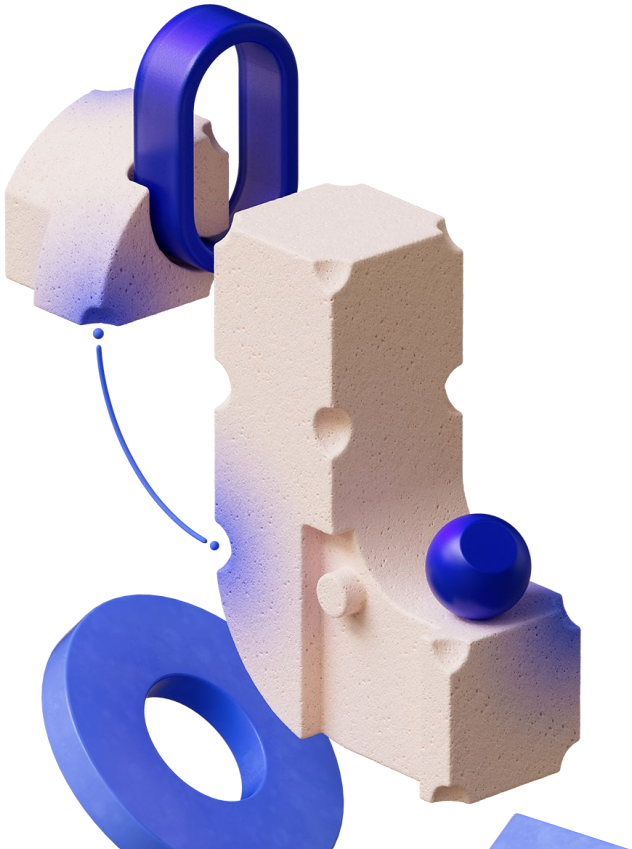
Google DeepMind



VideoGLUE: Video General Understanding Evaluation of Foundation Models

Liangzhe Yuan*†, Nitesh B. Gundavarapu*, Long Zhao*, Hao Zhou*, Yin Cui‡, Lu Jiang‡, Xuan Yang, Menglin Jia‡, Tobias Weyand, Luke Friedman, Mikhail Sirotenko, Huisheng Wang, Florian Schroff, Hartwig Adam, Ming-Hsuan Yang, Ting Liu, Boqing Gong†

* Equal contributions. † Corresponding authors. ‡ Work done at Google.

ICLR International Conference On Learning Representations

TRANSACTIONS tmlr on ML RESEARCH

Paper Code

Motivation

We evaluate the video understanding capabilities of existing foundation models (FMs) using a carefully designed experiment protocol consisting of three hallmark tasks (action recognition, temporal localization, and spatiotemporal localization), eight datasets well received by the community, and four adaptation methods tailoring an FM for downstream tasks.

Furthermore, we jointly profile FMs' efficacy and efficiency when adapting to general video understanding tasks using cost measurements during both training and inference.

Foundation Model	Pretraining Modality	Pretraining Data	Pretraining Objective
CoCa (Yu et al., 2022)	Image + Text	JFT3B + ALIGN	Contrastive + Captioning
CLIP (Radford et al., 2021)	Image + Text	WebImageText	Contrastive
FLAVA (Singh et al., 2022)	Image + Text	PMD	Contrastive + MIM + MLM
DINOv2 (Oquab et al., 2023)	Image	LVD-142M	MIM + DINO
VideoMAE (Tong et al., 2022)	Video	K400	MVM
InternVideo (Wang et al., 2022b)	Video	UnlabeledHybrid	MVM + Contrastive
VATT (Akbari et al., 2021)	Video + Audio + Text	HT100M	Contrastive

Task	Dataset	# of videos (train/validation)	Avg. length	Source	Notes
VC	Kinetics-400	235,693 / 19,165	10 secs	Web	Holistic, appearance
	Moments in Time	791,246 / 33,898	3 secs	Web	Holistic, appearance
	Something-Something v2	168,913 / 24,777	2 ∼ 6 secs	Crowdsource	Holistic, motion
	Diving48	15,027 / 1,970	5 secs	Web	Holistic, motion
	Charades	7,811 / 1,814	30 secs	Crowdsource	Multi-label, long-clip
TAL	ActivityNet v1.3	10,002 / 4,926	5 ∼ 10 mins	Web	Temporal
STAL	AVA v2.2	210,634 / 57,371	15 mins	Movie	Spatiotemporal, instance
	AVA-Kinetics	354,201 / 91,919	10 secs	Web	Spatiotemporal, instance

Paper Code

Main Observations

- **Task-specialized models** significantly **outperform** the seven **foundation models (FMs)** studied in this work, in sharp contrast to what FMs have achieved in natural language and image understanding.
- **Video-native FMs**, whose pre-training data mainly contains the video modality, are generally better than image-native FMs in classifying **motion-rich** videos, **localizing actions** in time, and understanding a video of **more than one action**.
- **Video-native FMs** can perform well on video tasks under **light adaptations** to downstream tasks (e.g., freezing FM backbones), while **image-native FMs** win in **full end-to-end fine-tuning**.

ICLR International Conference On Learning Representations
TRANSACTIONS tmlr on ML RESEARCH

Paper Code

Adaptation Methods

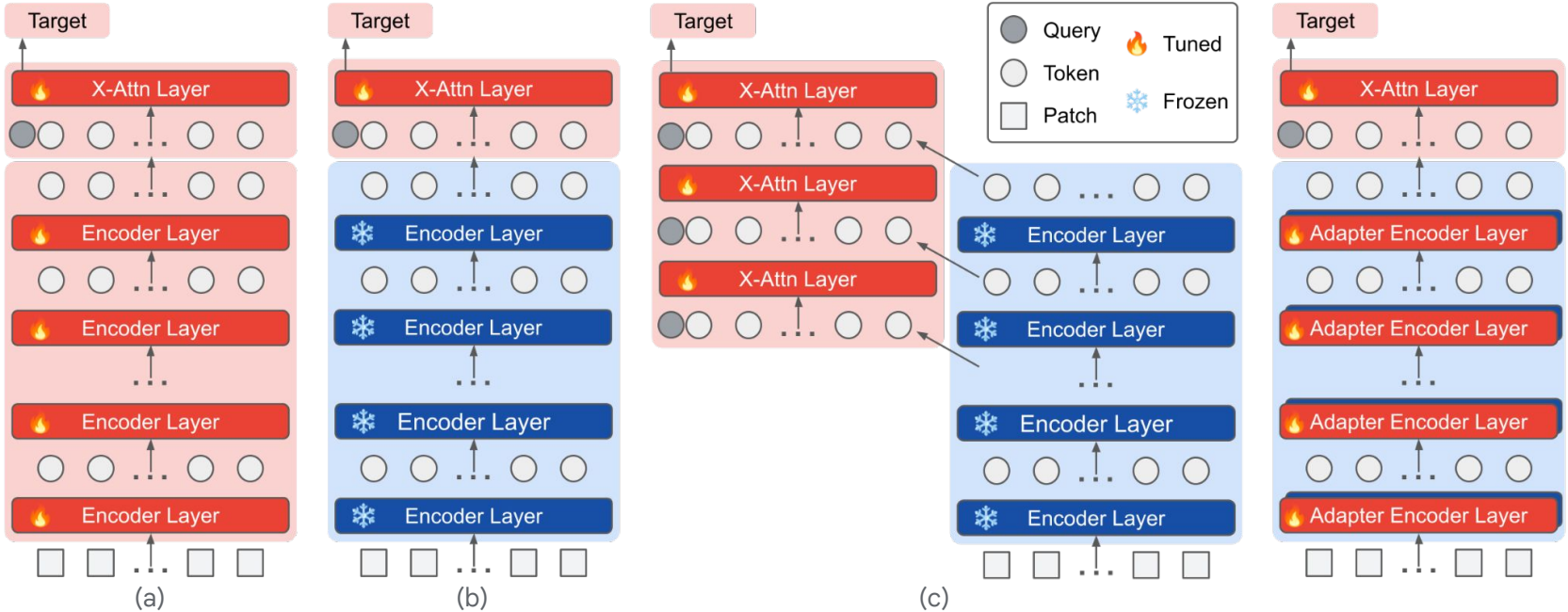
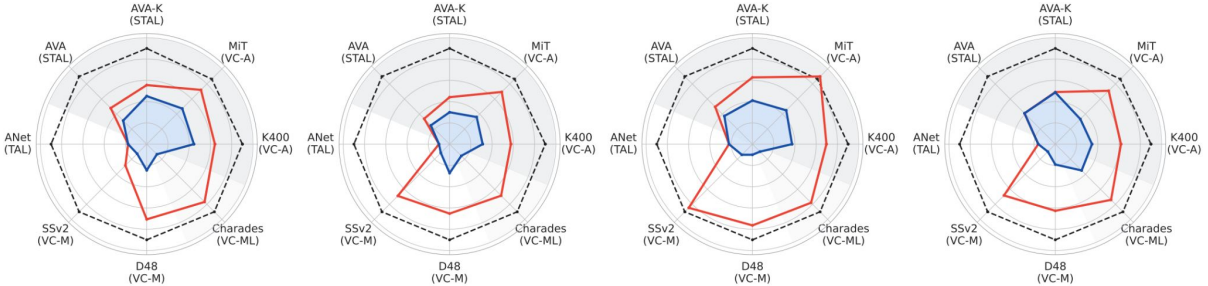


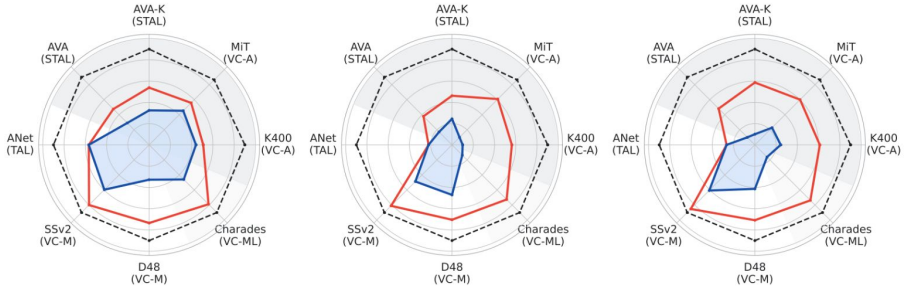
Fig. We study four adaptation methods to apply an FM to video understanding downstream tasks: (a) end-to-end fine-tuning, (b) frozen backbone, (c) frozen backbone with multi-layer attention pooler (MLAP), and (d) a low-rank adapter (LoRA).

Paper Code

Benchmark Results



(a) CLIP (b) FLAVA (c) CoCa (d) DINOv2



(e) VATT (f) VideoMAE (g) InternVideo

- Performance of FMs with end-to-end fine-tuning (red) and frozen backbone (blue), in comparison with state-of-the-art task-specialized models (black).
- We use gray shades to represent tasks that are more focused on appearance understanding more than motion.
- FMs generally fall behind task-specialized models; FMs that are trained with video data are generally better than image-native FMs on motion-focused tasks under the frozen backbone setting.
- Image-native FMs can generally catch up when fine-tuned end-to-end on the target dataset.

ICLR International Conference On Learning Representations
TRANSACTIONS tmlr on ML RESEARCH

Paper Code

VideoGLUE Score

We use **trainable parameters** and **inference FLOPs** to approximately represent the training and inference costs of an FM.

To rank FMs with the multi-dimensional assessments, we propose the following term to aggregate FMs' video understanding performance into a **single score**, termed VideoGLUE Score (VGS), where $\mathscr{S}_i$ is an FM's average performance score over our video tasks under the i-th adaptation method, and $\mathscr{C}_i^k$ is the corresponding cost value under the k-th developmental scenario.

$$VGS^k = \sum_{i=1}^{N} w_i^k \mathcal{S}_i, \text{ where } w_i^k = \frac{\mathcal{A}_i^k}{\sum_{j=1}^{N} \mathcal{A}_j^k} \text{ and } \mathcal{A}_i^k = \frac{1}{\log_{10} C_i^k},$$

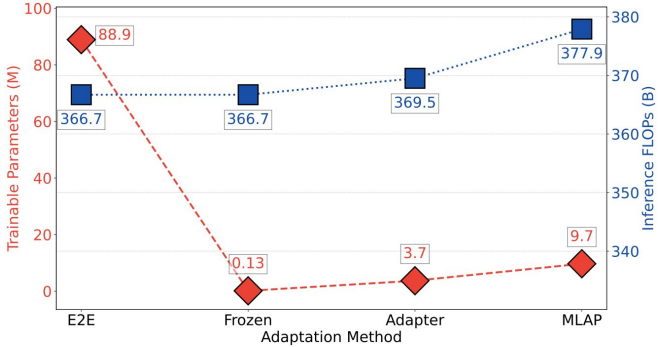
ICLR International Conference On Learning Representations
TRANSACTIONS tmlr on ML RESEARCH

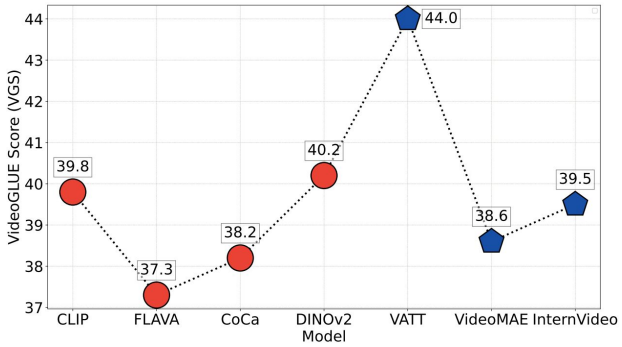
Paper Code

VideoGLUE Score

We use **trainable parameters** and **inference FLOPs** to approximately represent the training and inference costs of an FM.

To rank FMs with the multi-dimensional assessments, we propose the following term to aggregate FMs' video understanding performance into a **single score**, termed VideoGLUE Score (VGS), where $\mathscr{S}_i$ is an FM's average performance score over our video tasks under the i-th adaptation method, and $\mathscr{C}_i^k$ is the corresponding cost value under the k-th developmental scenario.





ICLR International Conference On Learning Representations
TRANSACTIONS tmlr on ML RESEARCH

Thank you.

Google DeepMind