International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02081-z

VLG: General Video Recognition with Web Textual Knowledge

Jintao Lin[1] · Zhaoyang Liu[2] · Wenhai Wang[3] · Wayne Wu[2] · Limin Wang[1]

Received: 14 March 2023 / Accepted: 13 April 2024

Abstract
Video recognition (action recognition) in an open world is quite challenging, as we need to handle different settings such as closed-set, long-tail, few-shot, and open-set. The majority of existing works often address each individual setting separately using various frameworks. However, these separate investigations would ignore the possibility of knowledge sharing across different settings, and stymie progress in video recognition as well as its application in the real world. By leveraging semantic knowledge from noisy text descriptions crawled from the Internet, we focus on the *general video recognition (GVR)* task of solving recognition problems of different settings within a unified framework. The core contribution of this paper is twofold. First, we build a comprehensive video recognition benchmark to facilitate the research of GVR, called *Kinetics-Text*. This dataset covers the mentioned four common settings, and provides multi-source text descriptions for all action classes for utilizing external textual knowledge from the Internet. Second, inspired by the flexibility of language representation, we analyse the correspondence between the video and text descriptions of its category and present a unified visual-linguistic framework *(VLG)* to solve the problem of GVR with an effective two-stage training paradigm. Our VLG is first pre-trained on video and language datasets to learn a shared feature space, and then devises a flexible bi-modal attention head to collaborate high-level semantic concepts under different settings. Extensive results show that our VLG obtains the state-of-the-art performance under four settings, and the superior performance demonstrates the effectiveness and generalization ability of our proposed framework. We hope our work makes a step towards the general video recognition and could serve as a baseline for future research. Code and datasets have been released in https://github.com/MCG-NJU/VLG.

Keywords General video recognition · Multi-modal analysis · Few-shot classification · Open-set classification

1 Introduction

Similar to image classification, the existing video recognition (action recognition) tasks are roughly grouped into four settings: closed-set (Kay et al., 2017; Carreira et al., 2019; Monfort et al., 2019, 2021), long-tail (Zhang et al., 2021c), few-shot (Zhu & Yang, 2018, 2020c; Zhu et al., 2021) and open-set (Acsintoae et al., 2021; Wang et al., 2021c), to mimic the realistic scenarios in practice. With multiple video benchmarks (Kay et al., 2017; Soomro et al., 2012; Goyal et al., 2017; Caba Heilbron et al., 2015; Carreira et al., 2019), a number of works (Wang et al., 2016; Liu et al., 2021b; Arnab et al., 2021; Zhang et al., 2021c; Zhu & Yang, 2018, 2020c; Bao et al., 2021) have been developed to study video recognition in these diverse scenarios.

Though various video benchmarks and frameworks have been established in the last few years, most existing works still follow a closed-set learning setting, where all the categories are pre-defined. Such method is unrealistic for many

Communicated by Karteek Alahari.

✉ Limin Wang
lmwang@nju.edu.cn

Jintao Lin
jintaolin@smail.nju.edu.cn

Zhaoyang Liu
zyliumy@gmail.com

Wenhai Wang
wangwenhai@pjlab.org.cn

Wayne Wu
wuwenyan0503@gmail.com

[1] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

[2] SenseTime Research, Shanghai, China

[3] Shanghai AI Laboratory, Shanghai, China

Published online: 25 May 2024

real-world applications, such as automatic tagging of web videos, where information regarding new video categories is not available during training. It is thus very challenging for closed-set methods to train a classifier for recognizing unseen or unfamiliar categories. In addition, most works (Feichtenhofer et al., 2016; Carreira & Zisserman, 2017; Tran et al., 2018; Kumar Dwivedi et al., 2019; Shu et al., 2018) focus on addressing individual settings separately with different frameworks. These separate investigations would ignore the potential sharing of knowledge among different settings, and severely impede the advance in video recognition as well as its application in the real world. Accordingly, we aim to present a single video benchmark covering all these settings, and propose a simple framework to handle these different sub-problems by leveraging semantic knowledge from noisy text description crawled from the Internet.

Since some works (Radford et al., 2021; Jia et al., 2021; Li et al., 2022; Yuan et al., 2021) have shown the efficacy of using natural language to supervise the visual representation learning, we intend to draw some extra knowledge (i.e., web text information) into our benchmark to facilitate the development of GVR. The extra web knowledge is expected to provide new cues for GVR. However, obtaining the paired text data for each video is prohibitively expensive. As shown in the Fig. 1, we observe that there are some connections between the video and text descriptions of its corresponding category. Specifically, the text descriptions for a specific video category exhibit some high-level semantic concepts to represent the static characteristics (e.g., scene) in space and dynamics (e.g., the steps to shooting) in time. In this sense, we hope that the text descriptions of video categories could provide useful clues to learn a more general representation for GVR under different settings. As a result, we build a new benchmark *Kinetics-Text* by extending the original Kinetics (Carreira & Zisserman, 2017) dataset to provide abundant text descriptions per-category in our benchmark to facilitate the research of GVR by crawling from the Internet. Moreover, in order to dig deeply into the general video recognition problem, we also hope this new video benchmark can cover a wide range of settings including closed-set, long-tail, few-shot and open-set (as shown in the Fig. 1). Accordingly, we curate different sub-datasets from the Kinetics-Text with four sub-settings: Kinetics-Close, Kinetics-LT, Kinetics-Fewshot and Kinetics-Open, to mimic the video distribution of different scenarios in real-world applications. Thees four kinds of sub-set on Kinetics-Text aim to provide a solid benchmark to verify the performance of video recognition models under different distributions.

Instead of dealing with each setting of video recognition with different methods, we develop a unified framework to address general video recognition. The unified framework would greatly reduce the work of hand-crafted design specific to each setting, and potentially increase its generalization ability due to the comprehensive consideration of all settings. We find some recent visual-linguistic representation works, e.g. CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), can learn transferable visual models from natural language supervision, and show a promising performance on image recognition under different settings. However, there is still a lack of work to bridge the gap between video and text for

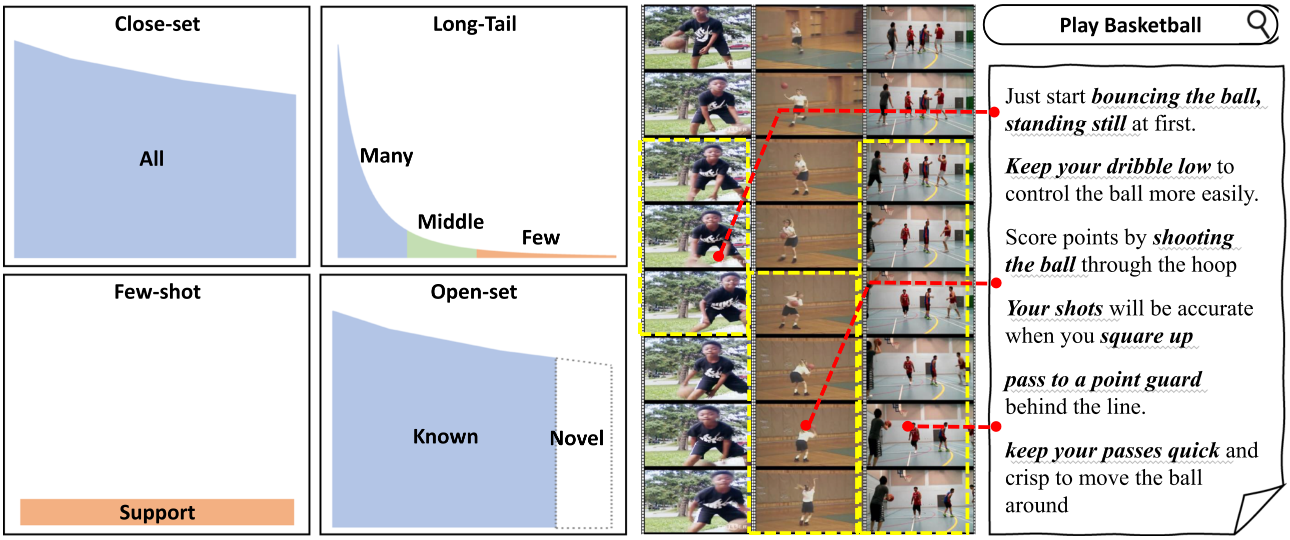

Close-set
All
Long-Tail
Many
Middle
Few
Few-shot
Support
Open-set
Known
Novel
Play Basketball
Just start ***bouncing the ball, standing still*** at first.
Keep your dribble low to control the ball more easily.
Score points by ***shooting the ball*** through the hoop
Your shots will be accurate when you ***square up***
pass to a point guard behind the line.
keep your passes quick and crisp to move the ball around

Fig. 1 Video label distribution of different scenarios and different modalities. As shown in the left, videos in GVR tasks have arbitrary distributions similar to natural data, such as closed-set, long-tail, few-shot and open-set. Most works only focus on coping with one aspect of them, while our method can use a unified framework to address the GVR task by combining the advantages of video and text modalities. The right part of the figure provides intuitive explanations for the correspondence between the videos and text modalities

general recognition under different scenarios. we consider using text as the supervision signals to learn a new video representation for general recognition scenarios, including long-tail, zero-shot, few-shot, and fully-supervised. Accordingly, we develop a video-language framework for general video recognition, termed as *VLG*. VLG could benefit from the visual-linguistic models pretrained on the large-scale image-text pairs (e.g., CLIP (Radford et al., 2021)), and connect video and text through customized temporal modeling. Our VLG leverages the rich semantic information of web text descriptions to guide the spatio-temporal feature learning. Specifically, our method primarily contains four components: (1) The frame encoder to learn the visual representation for each frame. (2) The temporal module to model temporal features across frames for video domain adaption; (3) The language encoder to learn the textual representation for each sentence of category description. (4) The bi-modal attention head to perform general video recognition under different settings. As text descriptions are directly collected from the Internet, they may include some noisy information. Thus we design a two-stage procedure to train our VLG: *Stage I* is to perform video-language pretraining, adapting the encoders from the image domain to the video domain to learn a visual-linguistic representation. *Stage II* filters out noisy texts and trains the bi-modal attention module to produce our final prediction. As demonstrated in experiments, our proposed VLG can effectively handle GVR under different settings of closed-set, long-tail, few-shot, and open-set.

In summary, we make the following contributions:

1. We formulate the task of general video recognition (GVR) and establish a comprehensive benchmark to fairly test the performance of video recognition models under different data distributions. The benchmark for general video recognition comprises closed-set, long-tail, few-shot and open-set, which shows different data distribution in practice.
2. To facilitate the research of GVR, we extend the original Kinetics into Kinetics-Text by elaborately collecting abundant text descriptions for each category. These extra textual knowledge exhibits more rich and high-level semantic concepts to represent the characteristics both in time and space, and contributes to the development of GVR.
3. We develop a unified video-language framework for general video recognition (VLG), which leverages the extensive web textual knowledge to effectively handle GVR under our customized two-stage learning strategy.
4. Extensive experiments demonstrate the effectiveness of our VLG on the Kinetics-Text for general video recognition under four settings.

We hope the findings together with the open-source code can inspire and facilitate future research on general video recognition, which enables us to examine the generalization ability of video recognition models in real-world applications.

2 Related Work

2.1 Video Representation

Video recognition has made rapid progress from the early hand-craft descriptors (Dollár et al., 2005; Klaser et al., 2008; Wang et al., 2013) to current deep networks. Deep neural networks can capture more general spatio-temporal representation from early two-stream networks, 3D-CNNs, and light-weight temporal modules to current transformer-based networks. Two-stream networks (Feichtenhofer et al., 2016; Simonyan & Zisserman, 2014; Wang et al., 2016) used two inputs of RGB and optical flow to separately model appearance and motion information in videos with a late fusion. Cao et al. (2020a) also proposed a composite two-stream framework based on a pre-training multi-channels self-attention model. 3D-CNNs (Carreira & Zisserman, 2017; Diba et al., 2018; Feichtenhofer et al., 2019; Stroud et al., 2020a) proposed 3D convolution and pooling to model space and time jointly. Light-weight temporal modules (Tran et al., 2018; Xie et al., 2018; Zhou et al., 2018; Jiang et al., 2019; Kumawat et al., 2021; Li et al., 2020b; Liu et al., 2021b) were designed as simple but powerful plugins to achieve the trade-off between efficacy and efficiency. Recently, several works (Arnab et al., 2021; Bertasius et al., 2021; Neimark et al., 2021; Fan et al., 2021) try to employ and adapt strong vision transformers to encode the spatial and temporal features jointly. The aforementioned methods mostly focus on addressing the video recognition problem only using visual modality in a supervised way, while ignoring the potentiality of natural language.

2.2 Visual-Textual Learning

Visual-Textual Pretraining has made great progress on several *down-stream vision tasks*. Mori et al. (1999) utilized paired text documents to connect images and words. Frome et al. (2013) and Weston et al. (2011) explored the image-text representation with class name annotations. Zhang et al. (2021b) obtained stronger visual embeddings from large object detectors, using visual-linguistic pretraining. Li et al. (2020a) learned powerful representation from a large-scale language corpus, with a visual-textual transformer. vadjust

Aligned with the success of image-language learning, lots of efforts have been made toward video-language learning (Miech et al., 2020a; Li & Wang, 2020a; Stroud et al., 2020b; Wang et al., 2021b; Ju et al., 2022). Li and Wang (2020b) learned powerful video representation from large-scale video-text pairs, with a contrastive learning method of CPD. Miech et al. (2020b) proposed a new learning loss to address misalignments inherent in narrated videos. Akbari et al. (2021) proposed a framework for learning multimodal representations for unlabeled data. Some works also focus on a specific type of downstream tasks, e.g. video-text VQA (Wang et al., 2020; Kant et al., 2020; Singh et al., 2019), video-text retrieval (Dong et al., 2021; Liu et al., 2021a; Yang et al., 2020). Specifically, Wang et al. (2021b) and Ju et al. (2022) adopted prompt engineering to reformulate their tasks into the same format as the pretraining objectives.

Video and language pretraining has also been an extensively studied topic. Wang et al. (2023) introduced an effective token rolling operation to encode temporal representations from video clips in a non-parametric manner. Xu et al. (2021) introduced new pretraining masking schemes to better mix across modalities and maintain separability for each modality. Zhu and Yang (2020a) and Zhu et al. (2020) directly modeled both global and local visual cues for fine-grained visual and linguistic relation learning in a self-supervised way. Ruan and Jin (2022) also provided a comprehensive overview of transformer-based pre-training methods for Video-Language learning.

Recently, CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) adopted simple noisy contrastive learning to obtain visual-linguistic representation from large-scale image-text web data. Inspired by this progress, there are also a lot of *CLIP-based video understanding* works. Ni et al. (2022) incorporated cross-frame attention mechanism and multi-frame integration transformer to enhance the spatio-temporal capability. Lin et al. (2022b) employed a lightweight Transformer decoder and learned a query token to dynamically collect frame-level spatial features from the CLIP image encoder, and adopted a local temporal module in each decoder layer to discover temporal clues from adjacent frames and their attention maps. Pan et al. (2022) used commonly adopted primitive operators for enabling a less studied parameter-efficient image-to-video transfer learning. Wu et al. (2023) revised the role of the linear classifier and replaced the classifier with different kinds of projection matrix from the pre-trained CLIP model. Luo et al. (2022) transferred the knowledge of the CLIP model to video-language retrieval in an end-to-end manner. Bain et al. (2022) found that simply using the baseline of weighted-mean of frame embeddings via query-scoring can achieve better performance for long video retrieval. Kahatapitiya et al. (2023) guided the learned latent space with freely-available auxiliary semantic information in the form of visually-grounded texts (e.g., object or scene information). Qian et al. (2022) designed a cross-modal fusion mechanism to aggregate complimentary multimodal information from video, audio, and optical flow for multimodal open-vocabulary video classification.

However, these methods cannot exploit the values of the noisy text descriptions data from the Internet, leading to an unsatisfactory performance on real-world applications. To mitigate these issues, Tian et al. (2022) proposed to adopt class-wise text descriptions for long-tailed image recognition, while our method seeks to learn video-language representations and further extends the framework to varied video recognition settings. Furthermore, our framework only utilizes a simple temporal module, which can be replaced with a more advanced spatiotemporal modeling module to further enhance the model's recognition capabilities.

2.3 General Video Recognition

While GVR has not been defined in the existing literature, we briefly summarize these sub-tasks of GVR: long-tailed classification, few-shot learning and open-set classification.

Long-tailed classification has been extensively studied based on re-sampling data, re-weighting loss, and transfer learning. Re-sampling data (Buda et al., 2018; Chawla et al., 2002; Chu et al., 2020; Drumnond & Holte, 2003; Han et al., 2005; Shen et al., 2016) aims to generate a class-balanced distribution, re-weighting loss (Cao et al., 2019; Cui et al., 2019; Huang et al., 2016; Khan et al., 2017; Wang et al., 2017; Tan et al., 2020) focuses on developing specific loss functions, and transferring strategy (Kang et al., 2019; Liu et al., 2019; Yin et al., 2019; Zhou et al., 2020; Zhu & Yang, 2020b) manages to transfer knowledge learned from head classes to tail classes. Specifically, Zhang et al. (2021c) proposed to dynamically sample frames for long-tailed video recognition.

As for *Few-shot classification*, it can be roughly divided into generative methods, initialization based methods, and metric based methods. Generative methods (Zhang et al., 2018; Kumar Dwivedi et al., 2019) generate additional task-specific training data to finetune a networks, initialization based methods (Finn et al., 2017, 2018) provide great initialization to learn novel classes quickly with gradients updates, and metric based methods (Bishay et al., 2019; Cao et al., 2020b; Zhu & Yang, 2018; Vinyals et al., 2016; Snell et al., 2017) compare the query set and support set with fixed feature representations. Specifically, Zhu and Yang Zhu & Yang (2018) proposed compound memory network to obtain an optimal video representation in a larger space, and Kumar Dwivedi et al. (2019) conditioned a conditional generative adversarial network with class prototype vectors to synthesize additional examples for novel categories.

Open-set recognition derives from face recognition (Li & Wechsler, 2005) and is firstly formalized in (Scheirer et al., 2012). In the early stage, lots of works (Scheirer et al., 2012, 2014; Jain et al., 2014) design variants of support vector machine (SVM) to reject unknown classes. With the recent development of deep learning, methods based on deep neural networks (Bendale & Boult, 2016; Ge et al., 2017; Ditria et al., 2020; Neal et al., 2018; Oza & Patel, 2019; Sun et al., 2020) are widely used in open-set recognition. As for open set video recognition, Shu et al. (2018) proposed ODN to gradually append new classes to the classification head, Krishnan et al. (2018), Subedar et al. (2019) and Krishnan et al. (2020) adopted Bayesian deep learning to acknowledge unknown classes, and Bao et al. (2021) incorporated evidential learning for large-scale and uncertainty-aware video recognition.

Compared with some current visual-linguistic approaches (Wang et al., 2021b; Ju et al., 2022; Ni et al., 2022; Lin et al., 2022b; Pan et al., 2022) for tasks related to video recognition, our method can not only provide a comprehensive video-language representation to bridge the gap between videos and texts in different cases, but also effectively utilize noisy web text annotations in practical applications.

3 The Kinetics-Text Benchmark

To leverage textual content to enhance video representation, we extend the original Kinetics400 dataset by crawling text descriptions for each category from the Internet to form a more comprehensive video benchmark called Kinetics-Text. In addition, to simulate the real-world video recognition from different scenarios, we further curate the *Kinetics-Text* into different kinds of sub-datasets, consisting of Kinetics-Close, Kinetics-LT, Kinetics-Fewshot, and Kinetics-Open. The dataset has been released in https://github.com/MCG-NJU/VLG.

Text descriptions The text descriptions are mainly crawled from Wikipedia (Wikipedia, 2022) and wikiHow (wikiHow, 2022). Following Tian et al. (2022), we first use the label name as the keyword to search for the best matching entry. Then, we filter out some unrelated parts of the entries, such as "references", "external links", and "bibliography", etc., to obtain the external text descriptions for each class. In addition, we also append 96 prompt sentences for each class as basic descriptions, which are generated by filling the pre-set templates, like ‘`a video of a {label}`’, with label names. In Fig. 8, we display a part of text descriptions collected for our benchmarks. We see that it is inevitable to include some noisy text descriptions, since these texts are all crawled from the Internet without fine-grained cleaning. In addition, we also report the detailed statistics of the collected text descriptions in Table 1. It can be seen that the text quantity of different classes varies significantly.

Kinetics-Close We directly adopt the original Kinetics400 (Kay et al., 2017) for closed-set setting, which contains daily activities and has around 250k trimmed videos covering 400 categories. These clips last around 10 s, and only the RGB frames are used to capture the visual cues, without transcripts and audio. Because of the expirations of some YouTube links, some original videos are missing over time. Our copy includes 240436 training videos and 19796 validation videos.

Kinetics-LT For the long-tailed case, we construct the Kinetics-LT dataset, which is a long-tailed version of Kinetics400 by sampling a subset following the Pareto distribution (Reed, 2001) similar to ImageNet-LT (Liu et al., 2019), with 930 $\sim$ 5 videos per class from the 400 classes of Kinetics400 dataset. Videos are randomly selected based on the distribution values of each class, and the 400 classes are randomly split into 109 many-shot classes, 209 medium-shot classes, and 82 few-shot classes. These splits are non-overlapping. We randomly select 20 training videos per class from the original training set as the validation set. The original validation set of Kinetics400 is used as the testing set in this paper. The dataset specifications are shown in Fig. 2.

Kinetics-Fewshot For the few-shot case, we conduct two kinds of few-shot settings, i.e., *5-shot-5-way* and *5-shot-C-way*. For the *5-shot-5-way setting*, we adopt the few-shot version of Kinetics (Zhu & Yang, 2018, 2020c), which has been frequently used to evaluate few-shot video recognition in previous works (Zhu & Yang, 2018, 2020c; Bishay et al., 2019; Zhang et al., 2020; Cao et al., 2020b; Perrett et al., 2021). In this setup, 100 videos from 100 classes are selected, with 64, 12 and 24 classes used for train/val/test. We conduct 200 trials with random samplings, to ensure the statistical significance. For the *5-shot-C-way setting*, we follow (Ju et al., 2022) to sample 5 videos from all categories to construct the training dataset, and measure the performance on the standard validation set, i.e. all videos from all categories in the validation set of Kinetics400. For statistical significance, we also conduct 10 random sampling rounds to choose training videos.

Kinetics-Open For open-set video classification, previous benchmarks will adopt UCF-101 testing set as known samples, and the testing splits of HMDB-51 and MiT-v2 datasets as two sources. Note there are a few overlapping classes between UCF-101 and the other two datasets, which will lead to information leakage during training. Therefore, we split the Kinetics400 into two parts, with 250 categories for training and the remaining 150 categories for evaluation. The 250-150 split setting was inspired by Ju et al. (2022), which divided Kinetics700 into a 400 train split and a 300 open-set val split. Following a similar rationale, we applied a proportionate split to Kinetics400, resulting in the 250-150 split setting. Videos in the training set and validation set are from different categories.

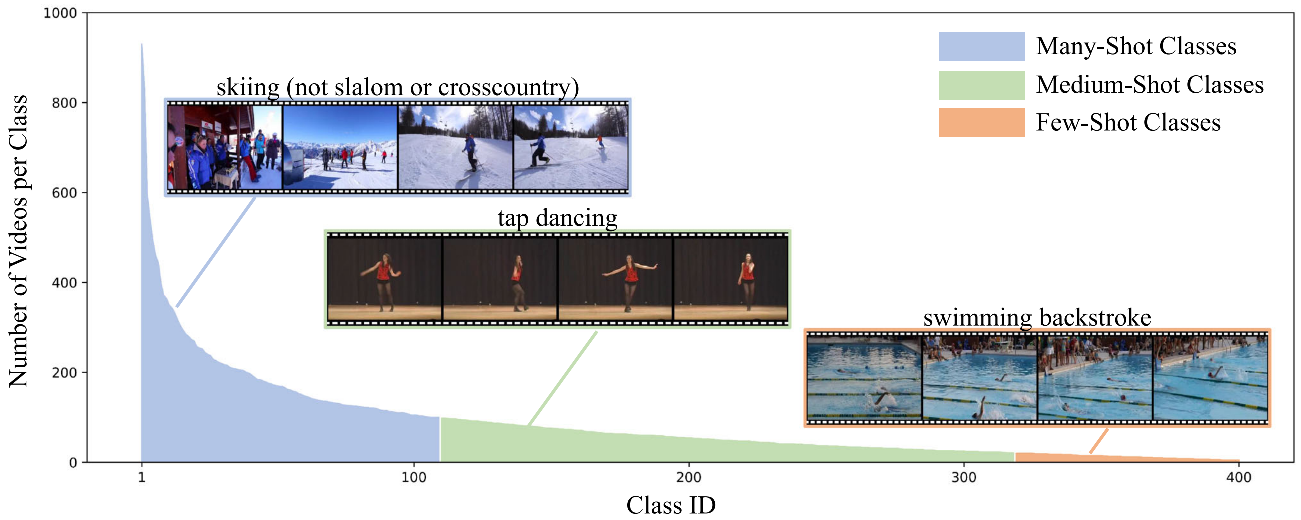


Fig. 2 The dataset statistics of Kinetics-LT. In this dataset, videos are randomly selected based on values of Pareto distribution for each class. The 400 classes are randomly split into 109 many-shot classes, 209 medium-shot classes, and 82 few-shot classes

Table 1 Detailed statistics of the text descriptions

Datasets	N_{min}	N_{max}	N_{mean}	N_{Med}	M_{min}	M_{max}	M_{mean}	M_{Med}	L_{Avg}
Kinetics400	7	634	143	99	252	20340	4011	2605	28

where N_{min}, N_{max}, N_{mean}, and N_{Med} denote for minimum, maximum, mean, and median number of sentences of classes respectively. M_{min}, M_{max}, M_{mean}, and M_{Med} denote for minimum, maximum, mean, and median number of words of classes respectively. L_{Avg} denotes the average number of tokens per sentence

4 Method

We first introduce the architecture of our proposed framework in Sect. 4.1, and then discuss its training strategy in Sect. 4.2. Finally, we present how to adapt our framework for different tasks in Sect. 4.3.

4.1 Overview

To effectively connect the video and language such that language concepts can relate to visual representations for general video recognition, we adopt a transformer-based network architecture (Radford et al., 2021), consisting of a video encoder $\Phi_{\text{video}}(\cdot)$ and a language encoder $\Phi_{\text{text}}(\cdot)$, to provide the visual representation and linguistic representation respectively. Specifically, the video encoder $\Phi_{\text{video}}(\cdot)$ is constructed with a frame encoder $\Phi_{\text{img}}(\cdot)$ followed by a temporal module $\Phi_{\text{temp}}(\cdot)$, which aggregates spatial features obtained from $\Phi_{\text{img}}(\cdot)$ over the temporal dimension.

As shown in the top of the Fig. 3, we first randomly sample a batch of videos $\mathcal{V} = \{V_i\}_{i=1}^{N}$, and the corresponding text sentences $\mathcal{T} = \{T_i\}_{i=i}^{N}$, where V_i and T_i are of the same class, N denotes the batch size, and each video contains F frames $V = \{I_i\}_{i=1}^{F}$. For texts $\mathcal{T}$, they are fed to the language encoder $\Phi_{\text{text}}(\cdot)$ to yield text embeddings E^T, while for videos $\mathcal{V}$, they are fed to the video encoder $\Phi_{\text{video}}(\cdot)$ to yield video embeddings E^V, by extracting frame features with $\Phi_{\text{img}}(\cdot)$ and then aggregating features along the temporal dimension with $\Phi_{\text{temp}}(\cdot)$:

$$E_i^T = \Phi_{\text{text}}(T_i), \tag{1}$$

$$E_i^V = \Phi_{\text{video}}(V_i) = \Phi_{\text{temp}}(\{\Phi_{\text{img}}(I_1), \ldots, \Phi_{\text{img}}(I_F)\}). \tag{2}$$

After that, we use a bi-modal attention head to aggregate the visual and linguistic features and then obtain the final prediction, as shown in the bottom of Fig. 3.

As raw text descriptions crawled from the Internet are noisy, it is necessary to obtain the salient sentences (namely, clean text descriptions) described in Sect. 4.2. The salient sentences reduce the impacts of noises for final prediction, which has been demonstrated in experiments in the Sect. 5.7. The bi-modal attention head dynamically fuses the video embeddings and text embeddings of salient sentences based on the attention weights. specifically, given video embedding $E^V \in \mathbb{R}^D$ and salient text embeddings of a certain class $E^T \in \mathbb{R}^{M \times D}$, we first calculate the query $\widetilde{Q} \in \mathbb{R}^D$, key $\widetilde{K} \in \mathbb{R}^{M \times D}$ and value $\widetilde{V} \in \mathbb{R}^{M \times D}$ of the attention operation.

$$\widetilde{Q} = \text{Linear}(\text{LayerNorm}(E^V)), \tag{3}$$

$$\widetilde{K} = \text{Linear}(\text{LayerNorm}(E^T)), \tag{4}$$

$$\widetilde{V} = E^T, \tag{5}$$

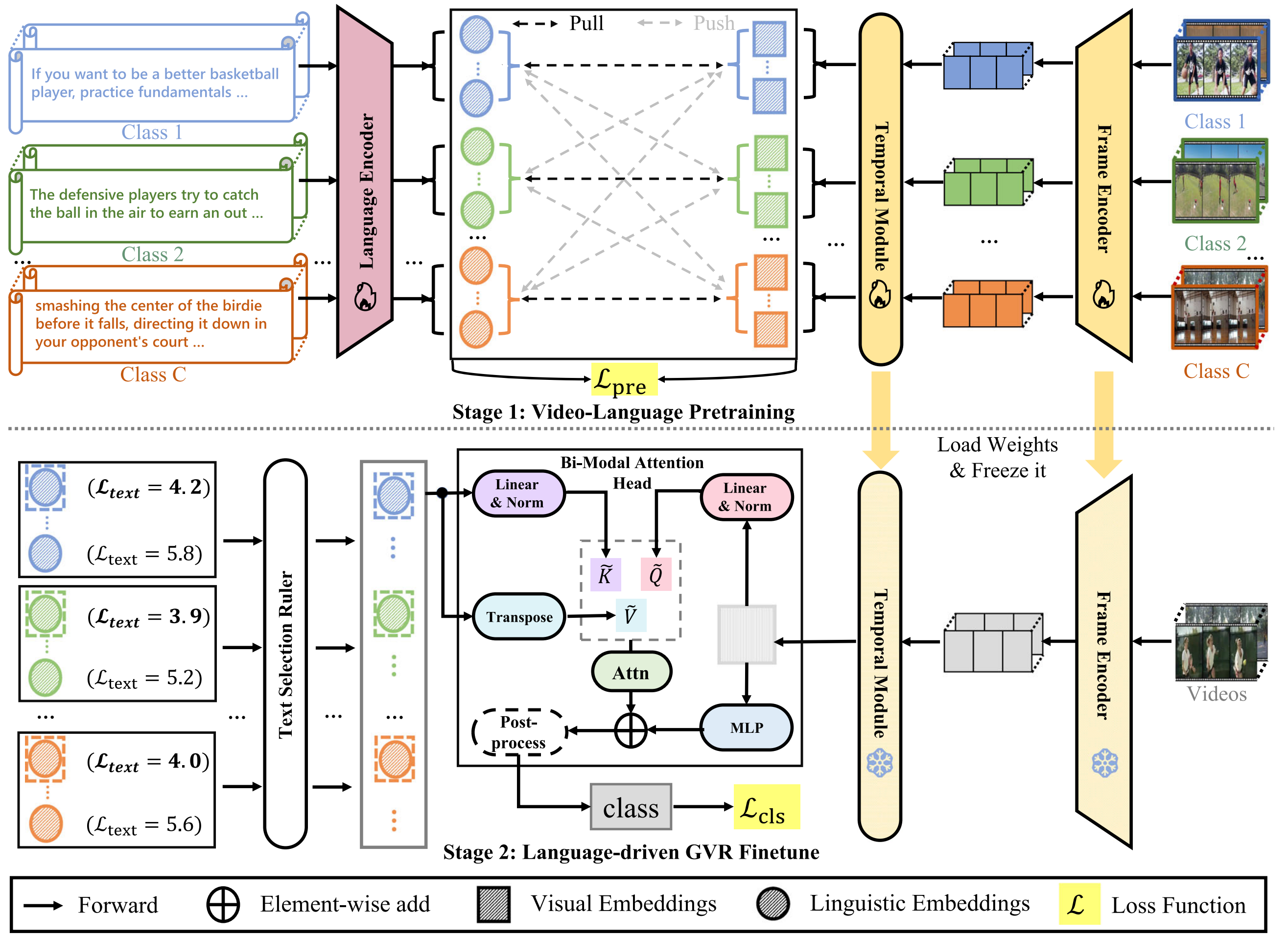

If you want to be a better basketball player, practice fundamentals ...
Class 1
The defensive players try to catch the ball in the air to earn an out ...
Class 2
smashing the center of the birdie before it falls, directing it down in your opponent's court ...
Class C
Language Encoder
Pull
Push
Temporal Module
Frame Encoder
Class 1
Class 2
Class C
$\mathcal{L}_{\text{pre}}$
Stage 1: Video-Language Pretraining
Load Weights & Freeze it
$(\mathcal{L}_{text} = 4.2)$
$(\mathcal{L}_{\text{text}} = 5.8)$
$(\mathcal{L}_{text} = 3.9)$
$(\mathcal{L}_{\text{text}} = 5.2)$
$(\mathcal{L}_{text} = 4.0)$
$(\mathcal{L}_{\text{text}} = 5.6)$
Text Selection Ruler
Bi-Modal Attention Head
Linear & Norm
Linear & Norm
$\tilde{K}$
$\tilde{Q}$
Transpose
$\tilde{V}$
Attn
MLP
Post-process
class
$\mathcal{L}_{\text{cls}}$
Temporal Module
Frame Encoder
Videos
Stage 2: Language-driven GVR Finetune
Forward
Element-wise add
Visual Embeddings
Linguistic Embeddings
$\mathcal{L}$ Loss Function

Fig. 3 The pipeline of VLG. The framework has two training stages. In the first stage, video-language pretraining (VLP) takes both the videos and text descriptions of each category as inputs, learning to link the two modalities through contrastive learning. In the second stage, embeddings of salient sentences, determined by the text selection ruler, are fed into the bi-modal attention head to make final predictions

where C is the class number, and M is the maximum number of sentences for each class, corresponding to the number of sampled salient sentences. Next, we adopt an attention operation to gather these M salient sentence embeddings for $\widetilde{G} \in \mathbb{R}^{D}$:

$$\widetilde{G} = \text{Softmax}(\frac{\widetilde{Q}\widetilde{K}^{\mathsf{T}}}{\sqrt{D}})\widetilde{V}. \tag{6}$$

Then, we perform broadcasting to gather the salient sentences embeddings over all the classes for $G \in \mathbb{R}^{C \times D}$, where C is the class number. The final classification probabilities are obtained based on the video embeddings E^{V} and enhanced text embeddings G:

$$P^{V} = \text{Softmax}(\text{MLP}(E^{V})), \tag{7}$$

$$P^{T} = \text{Softmax}(\text{sim}(E^{V}, G)/\tau), \tag{8}$$

$$P = P^{V} + P^{T}, \tag{9}$$

where P is the classification probability of the video, consisting of two terms, respectively for classification probability based on video representation P^{V}, and classification probability based on language representation P^{T}. sim$(\cdot, \cdot)$ denotes cosine similarity and τ is a learned parameter.

4.2 Training

We train our framework in two stages, namely Video-Language Pretraining (*VLP*) and Language-driven GVR Finetune, and design specific loss functions, i.e. $\mathcal{L}_{\text{pre}}$ and $\mathcal{L}_{\text{cls}}$, respectively for pretraining and classification.

Stage I: Video-Language Pretraining We jointly optimize the language encoder and video encoder together with the temporal module. The video features would be pulled together to their related category descriptions with higher similarity, and pulled away from irrelated sentences. Specifically, we use two contrastive learning NCE losses respectively for video embeddings E^V and text embeddings E^T:

$$\mathcal{L}_{\text{text}} = -\frac{1}{|\mathcal{V}_i^+|} \sum_{V_j \in \mathcal{V}_i^+} \log \frac{\exp(\text{sim}(E_j^V, E_i^T)/\tau)}{\sum_{V_k \in \mathcal{V}} \exp(\text{sim}(E_k^V, E_i^T)/\tau)}, \tag{10}$$

$$\mathcal{L}_{\text{video}} = -\frac{1}{|\mathcal{T}_i^+|} \sum_{T_j \in \mathcal{T}_i^+} \log \frac{\exp(\text{sim}(E_j^T, E_i^V)/\tau)}{\sum_{T_k \in \mathcal{T}} \exp(\text{sim}(E_k^T, E_i^V)/\tau)}, \tag{11}$$

where $\mathcal{L}_{\text{video}}$ and $\mathcal{L}_{\text{text}}$ represent the video and language losses respectively. $\mathcal{V}_i^+$ indicates a subset of $\mathcal{V}$, where all videos are of the same category with the text T_i. Similarly, all texts in $\mathcal{T}_i^+$ share the same class with the video V_i.

To effectively promote our framework for learning to connect the cross-modal information with limited text corpus, we adopt CLIP (Radford et al., 2021) pretrained model as the teacher model to distill knowledge for better visual-linguistic representation. To aggregate the frame features along the temporal dimension, the teacher model replaces the temporal module with the average pooling and outputs the same dimensions of embeddings as the student model. Their visual-linguistic similarities are used as soft targets for training weights associated with the student networks by the following objective:

$$S_{\mathcal{V}} = \frac{\exp(\text{sim}(E_i^V, E_i^T)/\tau)}{\sum_{V_j \in \mathcal{V}} \exp(\text{sim}(E_j^V, E_i^T)/\tau)}, \tag{12}$$

$$S_{\mathcal{T}} = \frac{\exp(\text{sim}(E_i^T, E_i^V)/\tau)}{\sum_{T_j \in \mathcal{T}} \exp(\text{sim}(E_j^T, E_i^V)/\tau)}, \tag{13}$$

$$\mathcal{L}_{\text{dist}} = -S'_{\mathcal{V}} \cdot \log S_{\mathcal{V}} - S'_{\mathcal{T}} \cdot \log S_{\mathcal{T}}, \tag{14}$$

where S and S' are cosine similarity scores respectively produced by our model and the frozen CLIP model. *With this pretraining stage, our framework can not only learn great video-language representation, but also reduce the risk of overfitting limited text corpus data.* Therefore, we optimize the video encoder and language encoder via pretraining loss $\mathcal{L}_{\text{pre}}$, defined as a weighted sum of $\mathcal{L}_{\text{video}}$, $\mathcal{L}_{\text{text}}$ and $\mathcal{L}_{\text{dist}}$:

$$\mathcal{L}_{\text{pre}} = \alpha \cdot (\mathcal{L}_{\text{video}} + \mathcal{L}_{\text{text}}) + (1 - \alpha) \cdot \mathcal{L}_{\text{dist}}. \tag{15}$$

Here, α is used to balance $\mathcal{L}_{\text{video}}$, $\mathcal{L}_{\text{text}}$ and $\mathcal{L}_{\text{dist}}$, which is set to 0.5 in our experiments.

Stage II: Language-driven GVR Finetune In order to take advantage of the valid semantic information and video-language feature, the second stage aims to select the salient sentences by filtering out the noisy texts, and then finetune the bi-modal attention head with the ground truth label.

To filter out noisy texts, we design a training-free text selection ruler (*TSR*) after obtaining the text embeddings, to sample the most discriminative sentences for each category. Specifically, we randomly choose λ videos of each class to construct a video batch V'. Then, we calculate $\mathcal{L}_{\text{text}}$ between each sentence and video batch V'. Finally, we select M sentences with the smallest $\mathcal{L}_{\text{text}}$ for the following classification. Note that the TSR only needs to be performed once at stage II.

To finetune the bi-modal attention head, we adopt two Cross Entropy losses $\mathcal{L}_{\text{CE}}$ for P^V and P^T (see Eqs. 7 and 8) respectively:

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{CE}}(P^V, \mathbf{y}) + \mathcal{L}_{\text{CE}}(P^T, \mathbf{y}), \tag{16}$$

where $\mathbf{y}$ is the ground truth label.

4.3 VLG for General Video Recognition

In most cases, given a query video and pre-selected text embeddings of salient sentences, we first feed the query video into the video encoder to obtain video embeddings. Then, the final result is predicted with the video embeddings and text embeddings of salient sentences through a video-language attention head. We follow this procedure in both the close-set setting and long-tailed setting. For the few-shot setting, we use base videos to pretrain the encoders during the first stage. Then, we use support videos to select salient sentences when combining the linguistic features, or directly use the video embeddings from VLP for linear probe testing. For the open-set setting, we follow the common procedure to train the framework, and insert a post-process step, which can be instantiated as the off-the-shelf open-set procedures (e.g., OpenMax (Bendale & Boult, 2016), Softmax with threshold, etc.), to recognize the novel videos during inference.

5 Experiments

We first introduce the evaluation metrics for different settings in Sect. 5.1 and implementation details in Sect. 5.2, before presenting state-of-the-art results over all these four benchmarks: Kinetics-Close, Kinetics-LT, Kinetics-Fewshot, and Kinetics-Open, respectively in Sects. 5.3, 5.4, 5.5 and 5.6. Next, we provide ablation studies in Sect. 5.7. Then, we present some representative visualization in Sect. 5.8.

5.1 Evaluation Metrics

We evaluate the performance of our framework under all of these four benchmarks. Besides the top-1 classification accuracy over all classes, for the *long-tailed* setting, we also report the accuracy of three disjoint subsets: *many-shot classes* (more than 100 training videos in each class), *medium-shot classes* (20∼100 training videos in each class), *few-shot classes* (less than 20 training videos in each class). For the *open-set* setting, we use *F-measure* score as a balance between precision and recall.

5.2 Implementation Details

Data Pre-processing If not specified, we use the segment-based input frame sampling strategy (Wang et al., 2016) with 8 frames. During training, we follow (Wang et al., 2021b) to process all frames to 224×224 input resolution. During inference, we resize all frames to 256×256 and center-crop them to 224×224.

Network Architectures If not specified, the video encoder adopts the pre-trained CLIP (Radford et al., 2021) visual encoder (ViT-B/16 (Sharir et al., 2021)) as our frame encoder. For the temporal module, we use a smaller version of the transformer with 6-layers and 8-head self-attention as default. To indicate the temporal order, we also add learnable temporal positional encoding onto the frame features as input. The language encoder also follows that of CLIP (Radford et al., 2021), which is a 12-layer transformer, and the maximum length of text tokens is set to 77 (including `[SOS]` and `[EOS]` tokens). We initialize the frame encoder and language encoder with pretrained weights of CLIP (Radford et al., 2021) during the first stage.

Training Hyper-parameters In our implementation, we always train the models using an AdamW (Loshchilov & Hutter, 2017) optimizer with the cosine schedule (Loshchilov & Hutter, 2016), a weight decay of 5×10^{-2}, and a momentum of 0.9 for 50 epochs. During the first stage, the size of the mini-batch is set to 16, and α is set to 0.5. The initial learning rate is set to 1×10^{-5} for frame encoder and language encoder, and set to 1×10^{-3} for the temporal module. During the second stage, the size of the mini-batch is set to 128. Both encoders are kept frozen, and the only trainable part is the bi-modal attention head. The learning rate of which is set to 1×10^{-3}. The number of selected sentences per class M is set to 64, and λ is set to 50. We conduct all experiments on 8 V100 GPUs.

Details for different settings For *closed-set*, *long-tail*, *5-shot-C-way*, and *5-shot-5-way without linear probe* (see Sect. 5.5), we adopt the proposed two-stage training pipeline. During inference, given a query video and pre-selected text embeddings of salient sentences, we feed the query video into the video encoder to obtain video embeddings. Then, the result is predicted with the video embeddings and text embeddings of salient sentences through a video-language attention head.

For *5-shot-5-way with linear probe*, we adopt the first pre-train stage for the base-split data training. During inference, we directly use the visual encode with the temporal module to perform linear probe testing.

For *open-set*, we adopt the first pre-train stage for the known split video data training. During inference, we follow the common procedure to train the framework, and insert a post-process step, which can be instantiated as the off-the-shelf open-set procedures (e.g., OpenMax (Bendale & Boult, 2016), Softmax with threshold, etc.), to recognize the novel videos during inference.

5.3 Experiments on Kinetics-Close

In Table 2, we compare our proposed methods with prior methods on Kinetics-Close, i.e. Kinetics400. There are mainly traditional CNN-based methods (e.g., X3D, SlowFast, TSM, etc.), Transformer-based methods (e.g., MViT, ViViT, Swin-L, etc.) and CLIP-based methods (e.g., ActionCLIP). These methods are pretrained with different strategies, including random initialization, ImageNet-1K/21K pretrainig, web-scale image pretraining. Compared to the CNN-based methods pretrained on ImageNet-1K, our VLG-ViT-B/16 surpasses TDN-R101 by 2.4% top-1 accuracy using fewer frames and views. It can be seen that Transformer-based methods and CLIP-based methods achieve better performance than traditional methods. Particularly, our models achieve higher accuracy than other competitors. For example, using only 8 frames and 1×1 view, our VLG-ViT-B/16 model achieves a higher top-1 accuracy compared to ViViT-L/16×2 (320), while using 20× fewer GFLOPs. In addition, our method achieves 82.9% top-1 accuracy with ViT-B/16 frame encoder, which exceeds ActionCLIP, a CLIP-based method, with 2.5× fewer video views. For a fair comparison, we further test our 16 frame ViT-B/16 on the val list of ActionCLIP, and our VLG achieves a higher accuracy performance of 83.5%. Moreover, when using ViT-L/14 as our visual backbone, our VLG can further achieve a higher accuracy of 86.4%, with lower resolution (224px) and fewer computational costs than MViTv2-L (312). The underlying reason is that the proposed temporal module can effectively model temporal relation among video frames and the joint language-image representation is successfully transferred to video domain with the help of our Video-Language pretraining design.

To further demonstrate the superiority of the proposed VLG, we propose CLIP-Raw and CLIP-Close as our baselines on Kinetics-Close to make fair comparisons. CLIP-Raw directly adopts the original CLIP weights and model with only prompt sentences to validate the accuracy performance, while CLIP-Close removing language encoder consists of

Table 2 Results on Kinetics-Close

Method	Pretrain	Frame	Views	Top-1	Top-5	GFLOPs (per-view)	Param (M)
X3D-XL (Feichtenhofer, 2020)	None	–	10 × 3	79.1	93.9	48.4	11.0
SlowFast, R101+NL (Feichtenhofer et al., 2019)	–	16	10 × 3	79.8	93.9	234.0	59.9
MViT-B, 64×3 (Fan et al., 2021)		64	3 × 3	**81.2**	95.1	455.0	36.6
TSM, ResNeXt101 (Lin et al., 2019)	IN-1K	8	10 × 3	76.3	–	–	–
TANet, R152 (Liu et al., 2021b)		16	4 × 3	79.3	94.1	242.0	–
TDN, R101 (Wang et al., 2021a)		24	10 × 3	**79.4**	94.4	198.0	88.0
ViViT-L/16x2 (Arnab et al., 2021)	IN-21K	32	4 × 3	80.6	94.7	–	–
TimeSformer-L (Bertasius et al., 2021)		8	1 × 3	80.7	94.7	2380.0	121.4
ViViT-L/16x2 (320) (Arnab et al., 2021)		32	4 × 3	81.3	94.7	3992.0	310.8
Swin-L (384) (Liu et al., 2022)		32	10 × 5	84.9	96.7	2107.0	200.0
MViTv2-L (312) (Li et al., 2021b)		40	5 × 3	**86.1**	97.0	2828.0	217.6
ViViT-H/16x2 (Arnab et al., 2021)	JFT	32	4 × 3	84.8	95.8	–	–
TokenLearner 16at18 (L/10) (Ryoo et al., 2021)		-	4 × 3	85.4	96.3	4076.0	450.0
MTV-H (Yan et al., 2022)		32	4 × 3	85.8	96.6	3706.0	–
CoVeR (Zhang et al., 2021a)		16	1 × 3	**86.3**	–	–	–
CLIP-Raw, R50 (Radford et al., 2021)	CLIP*	8	1 × 1	46.2	60.8	52.1	102.0
CLIP-Raw, ViT-B/16 (Radford et al., 2021)		8	1 × 1	55.0	67.5	144.0	150.0
CLIP-Close, R50 (Radford et al., 2021)		8	1 × 1	68.1	87.7	49.7	115.0
CLIP-Close, ViT-B/16 (Radford et al., 2021)		8	1 × 1	78.9	93.5	141.0	106.0
ActionCLIP, ViT-B/16 (Wang et al., 2021b)		16	10 × 3	82.6	96.2	563.1	141.7
VideoPrompt, ViT-B/16 (Ju et al., 2022)		16	5 × 1	76.6	93.3	–	–
Text4Vis, ViT-L/14 (Wu et al., 2023)		32	4 × 3	87.1	97.4	–	–
X-CLIP, ViT-L/14 (Ni et al., 2022)		8	4 × 3	87.1	97.6	658.0	–
ST-Adapter, ViT-L/14 (Pan et al., 2022)		32	1 × 3	87.2	97.6	2749.3	–
VicTR, ViT-L/14 (Kahatapitiya et al., 2023)		8	4 × 3	87.0	–	656.0	–
EVL, ViT-L/14 (Lin et al., 2022b)		16	1 × 3	87.0	–	1348.0	–
VLG, R50	CLIP*	8	1 × 1	72.3	90.8	76.7	148.0
VLG, ViT-B/16		8	1 × 1	81.8	95.3	148.0	121.0
VLG, ViT-B/16		16	1 × 1	82.4	95.8	282.3	121.0
VLG, ViT-B/16		16	4 × 3	82.9	96.1	282.3	121.0
VLG, ViT-L/14		8	1 × 1	85.5	96.3	650.3	371.0
VLG, ViT-L/14		8	4 × 3	**86.4**	**97.0**	650.3	371.0

Bold indicates the best result

By introducing the class-wise text descriptions, our model achieves superior performance to the existing approaches. “IN” denotes ImageNet and "K400" denotes Kinetics400. "-" indicates the numbers are not available for us. "CLIP*" denotes that the model is initialized with the weights pretrained on 400 M image-text pairs provided in CLIP (Radford et al., 2021). The total GFLOPs are calculated by the number of views and GFLOPs (per-view)

the frame encoder loading CLIP pretrained weights, temporal module and a linear classifier layer and is finetuned on Kinetics-Close for 100 epochs. One can observe that our method also gets absolute accuracy gain against the baselines with ResNet-50 (72.3% vs. 68.1% vs. 46.2%) and ViT-B/16 (81.8% vs. 78.9% vs. 55.0%) backbones. The results are desirable since our framework can take the advantage of the semantic information in the text descriptions in the first step of video-language pretraining, and is able to filter irrelated and noisy text descriptions in the second step of Language-driven general video recognition finetuning.

Currently, a lot of works are based on CLIP weights for video classification, like X-CLIP (Ni et al., 2022), EVL (Lin et al., 2022b), ST-Adapter (Pan et al., 2022), etc. Some of these works achieve higher recognition accuracy than our VLG. However, they usually design more complex network modules and require more frames within a clip for temporal modeling, while our VLG only adopts a simple temporal transformer as the temporal module. These improvements are

Table 3 Results on Kinetics-LT

Method	Pretrain	Backbone	Accuracy (%) Overall	Many	Medium	Few
TSN (Wang et al., 2016)	ImageNet	ResNet-50	47.2	59.3	49.4	23.6
TSM (Lin et al., 2019)			46.0	66.3	46.1	17.3
TANet (Liu et al., 2021b)			45.8	66.8	45.4	17.4
SlowOnly (Feichtenhofer et al., 2019)			44.8	67.7	44.1	14.4
NCM (Kang et al., 2019)	CLIP*	ResNet-50	41.8	53.0	42.3	24.6
cRT (Kang et al., 2019)			43.7	58.9	43.8	22.3
τ-normalized (Kang et al., 2019)			43.9	63.8	43.1	18.5
LWS (Kang et al., 2019)			45.1	58.6	44.8	27.1
SSD-LT (Li et al., 2021a)			48.3	59.6	49.1	30.0
PaCo (Cui et al., 2021)			50.1	60.1	50.3	35.8
CLIP-Raw (Radford et al., 2021)	CLIP*	ResNet-50	46.2	48.3	44.8	46.7
CLIP-LT (Radford et al., 2021)			53.4	70.3	53.3	31.1
VLG			60.8	71.7	60.4	47.2
CLIP-Raw (Radford et al., 2021)	CLIP*	ViT-B/16	55.0	57.1	53.7	55.5
CLIP-LT (Radford et al., 2021)			63.8	79.7	63.8	42.8
VLG			**70.7**	**81.9**	**69.7**	**58.3**

Bold indicates the best result

Traditional Long-tailed methods use the same visual backbone. CLIP* denotes that the model is initialized by the CLIP (Radford et al., 2021) weights. We report the overall accuracy and the accuracy of three disjoint subsets

out scope of our paper, whose focus is on general video recognition under different settings. In comparison, our framework VLG aims at utilizing a single pipeline to address video recognition problems under different settings, leveraging web-scale text information, and filtering out irrelevant text in a training-free manner. Furthermore, our framework only utilizes a simple temporal module, which can be replaced with a more advanced spatiotemporal modeling module to further enhance the model's recognition capabilities.

5.4 Experiments on Kinetics-LT

In Table 3, we can see that our VLG models are superior to conventional vision-based methods (including re-sampling data, re-weighting loss, and transfer learning) with the same video encoders (ResNet-50). Since there are few long-tailed methods specific to video domain, we re-implement and report the performance of some representative image-based long-tailed methods on Kinetics400-LT, such as τ-normalized, cRT, NCM, LWS (Kang et al., 2019), PaCo (Cui et al., 2021), and SSD-LT (Li et al., 2021a), which are all initialized with CLIP pretrained weights to ensure the fairness of experimental comparisons. It is worthwhile mentioned that we also add an additional temporal pooling for each image-based long-tailed method without introducing any new parameters to aggregate features along the temporal dimension for them. In addition, we also build CLIP-LT and CLIP-Raw as our simple baseline based on CLIP to corroborate our method. CLIP-LT is built the same as CLIP-Close.

It can be seen that our proposed method is superior to prior visual-based methods with the same backbone. For example, using the same ResNet-50 backbone, the overall accuracy of VLG reaches 60.8%, which outperforms SSD-LT by 12.5 points (60.8% vs. 48.3%), and 10.7% better than PaCo (60.8% vs. 50.1%). In addition to the overall accuracy, our VLG also outperforms the prior visual-based long-tailed methods on different shots settings. For example, our VLG reaches the top-1 accuracy of 71.7%, 60.4%, 47.2% respectively in the Many-shot, Medium-shot, Few-shot setting for the Kinetics-LT dataset, which significantly surpasses the state-of-the-art PaCo method by 11.6%, 10.3%, 11.4% points respectively in different kinds of settings. Moreover, when compared to CLIP baseline models, the performance of our method is also promising, which is 7.4% better than the CLIP-LT, and 14.6% better than the CLIP-Raw in the overall setting (60.8% vs. 53.4% vs. 46.2%). Moreover, our method is 7.1% better than the CLIP-LT, and 15.6% better than the CLIP-Raw in the medium-shot setting (60.4% vs. 53.3% vs. 44.8%). When it comes to the few-shot setting, our method also surpasses the CLIP-LT and CLIP-Raw. When using ViT-B/16 as the backbone, the overall accuracy of VLG can further boost up to 70.7%, with the top-1 accuracy of 81.9%, 69.7%, 58.3% respectively in the setting of many-shot, medium shot and few-shot. The competitive performance of our method can be attributed to the usage of textual knowledge from the text descriptions, which serves as a clue to enhance the semantic representation for the tail categories.

Table 4 Results on Kinetics-Fewshot

Method	Backbone	K-shot	N-way	Top-1
CMN (Zhu & Yang, 2018)	ResNet-50	5	5	78.9
TARN (Bishay et al., 2019)		5	5	78.5
ARN (Zhang et al., 2020)		5	5	82.4
VLG-L		5	5	84.6
VLG		5	5	**94.0**
E-Prompt (Ju et al., 2022)	ViT-B/16	5	5	96.4
VLG		5	5	**96.9**
E-Prompt (Ju et al., 2022)	ViT-B/16	5	$\mathcal{C}_{ALL}$	58.5
VLG		5	$\mathcal{C}_{ALL}$	**62.8**

Bold indicates the best result
Here, $\mathcal{C}_{ALL}$ denotes the model is tested on all categories of the corresponding dataset, rather than only 5-way classification. In Kinetics-fewshot, $\mathcal{C}_{ALL} = 400$. “VLG-L” denotes our method with linear probe testing

5.5 Experiments on Kinetics-Fewshot

Following (Ju et al., 2022), we conduct two kinds of few-shot settings, i.e., *5-shot-5-way* and *5-shot-C-way*.

5-shot-5-way For a fair comparison, this setting adopts the publicly accessible few-shot splits. During training, we simply use the base split for our first video-language pretraining stage, without any meta-learning paradigms. During the evaluation, we report average results over 200 trials with random sampling on the test split. Table 4 presents the average top-1 accuracy, and our method clearly achieves significant performance. Following CLIP (Radford et al., 2021), we directly adopt the linear probe to test the visual representation output from the video encoder, which obtains 84.6% top-1 accuracy and is higher than the traditional few-shot learning methods (2.2% higher than ARN and 6.1% higher than TARN). When combining the linguistic features, the performance can further boost up to 94.0% top-1 accuracy. We also use the same network settings with textual information following (Ju et al., 2022), and achieve better performance (96.9% vs. 96.4%), which can be attributed to the reason that the text descriptions with different expression crawled from the Internet can serve as more complex textual prompts than the self-designed continuous prompts.

5-shot-C-way We further investigate a more challenging experiment setting, which samples 5 videos from the training set for each class as the base split, and then directly evaluates the model on the standard Kinetics400 testing split. For statistical stability, we report the average results over 10 trials to ensure the reliability of results. It can be seen that our model still obtains a superior performance (62.8% vs. 58.5%), which is also higher than (Ju et al., 2022), indicating the robustness of our method even with fewer samples on multiple categories.

5.6 Experiments on Kinetics-Open

Openset video recognition aims to not only accurately classify known categories which have appeared in training, but also recognize unknown categories which are not seen in training. Without any other modifications to our framework, we only adopt softmax with *thresholds* and *OpenMax* (Bendale & Boult, 2016) as a post-process on the prediction logits to obtain the classification results, as described in Sect. 4.3. In addition, we also re-implement the OLTR with the same CLIP initialization and visual backbone (ResNet-50) as a comparison. As shown in Table 5, we outperform OLTR (Liu et al., 2019) among all different threshold numbers, indicating the significance of our video-language representation. Specifically, when using the same threshold post-process after obtaining the logits, our method surpasses OLTR significantly with 0.120, 0,135, 0.141, 0.108 improvement on the F-measure score respectively when the threshold is set to 0.1, 0.2, 0.3, 0.5. When switching to the OpenMax post-process, our method can still surpasses OLTR by a large margin. Moreover, when using the ViT-B/16 as the backbone, the F-measure scores of VLG can further boost up to 0.694, 0.698, 0.703, 0.721, 0.697 respectively for different thresholds. The underlying reason is that natural language is used to reference learned visual concepts(or describe new ones), thus enabling zero-shot transferring of the models to unknown categories.

5.7 Ablation Study

In this subsection, we conduct extensive ablation studies on the Kinetics-Close dataset to provide a deep analysis of our proposed method. We compare the original VLG with different variants to investigate the effect of VLG’s components and the superiority of the collected text descriptions from the Internet. In all of these experiments, we use ViT-B/16 as the default backbone and adopt the segment-based input frame

Table 5 Results on Kinetics-Open

Method	Post-process	F-measure thr=0.1	thr=0.2	thr=0.3	thr=0.5	thr=0.7
OLTR, R50 (Liu et al., 2019)	Threshold	0.490	0.504	0.513	0.502	0.458
VLG, R50	Threshold	0.610	0.639	**0.654**	**0.610**	**0.469**
VLG, R50	OpenMax	**0.616**	**0.641**	0.651	0.614	0.465
VLG, ViT-B/16	Threshold	0.657	0.672	0.694	**0.721**	0.697
VLG, ViT-B/16	OpenMax	0.694	0.698	**0.703**	0.699	0.633

Bold indicates the best result

OLTR and VLG are both initialized by CLIP weight. With the same backbone, VLG outperforms OLTR among all thresholds

Table 6 Ablation studies on Kinetics-Close

#	Pretrain	CLIP Weights	Fine-tuning Head	Ruler	Top-1
1	✓	✓	Bi-M	TSR	**81.8**
2	–	✓	Bi-M	TSR	76.0
3	✓	–	Bi-M	TSR	32.6
4	✓	✓	FC	–	79.5
5	✓	✓	KNN	–	79.9
6	✓	✓	Bi-M	RAND	80.0
7	✓	✓	Bi-M	BASIC	78.9

Bold indicate the best result

"Head" denotes the classification head used in stage II, "Bi-M" denotes the bi-modal attention head, "TSR" denotes the proposed text selection ruler, "RAND" denotes random selection strategy, and "BASIC" denotes only using basic prompted sentences

sampling strategy with 8 frames. All the other settings remain the same as Sect. 5.2 unless specifically mentioned.

Video-Language Pre-training To examine the effectiveness of our video-language pretraining (VLP) in the first stage of the framework, we remove it by directly performing the finetuning process on the pretrained weights of CLIP (Radford et al., 2021). As reported in the #1 and #2 of Table 6, the model with VLP outperforms the one without VLP by 5.8 points on the top-1 accuracy (81.8% vs. 76.0%). Such a gap might be attributed to the difficulties in learning temporal information and semantic inconsistency between videos and text representation, which can be alleviated through the designed video-language pretrainig in our framework and semantic information in the web texts. Additionally, we also train our method with randomly initialized weights to further analyze the influence of CLIP pre-trained weights in the video-language pretraining stage. Comparing the #1 and #3 of Table 6, we can see that initializing with CLIP pre-trained weights can significantly benefit our approach. The model with CLIP pretrained weights outperforms the one with randomly initialized weights by 49.2 points on the top-1 accuracy (81.8% vs. 32.6%). This phenomenon is mainly caused by the limited web-collected text corpus for pre-training. Specifically, there are only 400 class descriptions (about 95K sentences collected from the Internet for per-category descriptions) for Kinetics400, and it is easy to overfit a video to a specific set of sentences without a pretrained linguistic encoder. In comparison, using the CLIP weights for pretraining is more robust, since the checkpoints are trained on 400 million filtered web image-text pairs, which is cleaner and almost 4K times more than our collected text descriptions.

Module Design To further demonstrate the effectiveness of the components in our framework, we also carry out some ablation studies respectively in some vital modules (e.g., bi-modal attention head, text selection ruler, and temporal module) in our VLG.

We investigate the effectiveness of bi-modal attention head by comparing it with other recognition heads, including FC (only video-based, with additional linear projection), and KNN (video-language based, no additional learnable parameters). Specifically, FC means we directly append a linear layer to the visual backbone for finetuning. KNN means we apply K-Nearest Neighbors Algorithm, a non-parametric supervised learning classifier, to the similarity scores calculated from video embedding and textual embedding after the pre-training stage. As reported in #1, #4 and #5 of Table 6, the proposed head performs better than FC and KNN by 2.3 and 1.9 points respectively (81.8% vs. 79.5% vs. 79.9%). It is also notable that KNN is a video-language based bi-modal head, which is the same as the bi-modal attention head, and KNN also works better than FC. These phenomenons indicate the superiority of bi-modal attention head and the power of video-language representation.

Furthermore, we also study the significance of the text selection ruler the sampled salient sentences by replacing them with those sampled by "Random" and "Basic" strategies. For "Random", we randomly select M sentences per category from text descriptions. For "Basic", we only use the basic prompt sentences for each category as the salient sentences. As shown in Table 6, the model with TSR (See the #1 of Table 6) outperforms the model with other strategies on the Top-1 accuracy (81.8% vs. 80.0% vs. 78.9%), which indicates the effectiveness of our TSR to filter out some noisy sentences. It can be also seen that the "Random" strategy also outperforms the "Basic" strategy, which can be attributed to

Table 7 Ablation studies on the number of layers in Temporal Module

(a) Ablation studies of Temporal Module with different numbers of layers on Kinetics-Close

Number of layers	0	1	2	4	6	8
Top-1 Acc	79.8	80.9	81.2	81.4	**81.8**	80.6

(b) Ablation studies of Temporal Module with different numbers of layers on Kinetics-Fewshot

Number of layers	0	1	2	4	6	8
Top-1 Acc	**96.9**	96.5	95.6	95.8	96.1	95.2

Bold indicates the best result

We evaluate the Top-1 accuracy on Kinetics-Close and Kinetics-Fewshot by using different numbers of layers in the temporal module

Table 8 Effectiveness of Distillation Loss in the first Video-Language Pretraining stage

Distill kind	Dataset	Backbone	$1-\alpha$	Top-1	Many	Medium	Few
-	Kinetics-LT	ResNet-50	0	57.5	70.8	57.2	40.8
Logits	Kinetics-LT	ResNet-50	0.1	58.4	71.9	58.2	40.7
			0.5	**60.8**	71.7	**60.4**	**47.2**
			0.9	54.7	64.4	54.4	42.5
Feature	Kinetics-LT	ResNet-50	0.1	58.1	71.4	58.1	40.7
			0.5	59.7	72.2	59.6	43.0
			0.9	58.2	71.5	57.6	41.7

Bold indicates the best result

We evaluate the performance on Kinetics-LT to investigate the effectiveness of distillation loss

Table 9 Ablation studies of loss terms in the second finetune stage

(a) Ablation studies of loss terms on Kinetics-Close

Method	Backbone	Operation	Top-1	Top-5
VLG	ViT-B/16	Only P^V	79.5	94.7
		Only P^T	80.7	95.1
		P^V and P^T	**81.8**	**95.3**

(b) Ablation studies of loss terms on Kinetics-LT

Method	Backbone	Operation	Overall	Many	Medium	Few
VLG	ResNet-50	Only P^V	56.9	**73.2**	57.1	34.7
		Only P^T	59.8	67.2	60.2	**48.8**
		P^V and P^T	**60.8**	71.7	**60.4**	47.2

Bold indicates the best result

We investigate the effectiveness of the two terms in Eq. 4 of the main body by adopting three operations: "only P^V", "only P^T", and "both P^V and P^T"

the more complex expressions and rich semantic information from the web text descriptions.

In addition, we also demonstrate the effectiveness of the temporal module by using different numbers of layers in the temporal module. As shown in Table 7, the model achieves the highest top-1 recognition accuracy on Kinetics-Close with 6 transformer layers in the temporal module. An interesting phenomenon is that increasing the number of layers in the temporal module leads to a significant rise in accuracy performance at the beginning, but the accuracy falls when the temporal module has more than 6 layers. It may be attributed to the overfitting caused by using the transformer with too many layers. In the few-shot setting, the model achieves the highest recognition accuracy without the additional temporal module, since there are few videos to feed the data-hungry Transformer layers in the few-shot case.

Loss Design We verify the effectiveness of our loss design respectively in the two stages in our framework.

In the first pre-training stage, we adopt the distillation loss to reduce the risk of overfitting caused by limited text corpus. To better investigate the effectiveness of distillation loss, we conduct the ablation study on Kinetics-LT, which is more appropriate than the Kinetics-Close in this case, since it has fewer videos to avoid the influence of excessive visual information. We use ResNet-50 as the backbone for Kinetics-LT. As shown in Table 8, our method with distillation loss

Table 10 Ablation studies of text descriptions

(a) Ablation studies of text descriptions on Kinetics-Close

Method	Backbone	Operation	Top-1	Top-5
CLIP-Close	ViT-B/16	NO TEXT	78.9	93.5
VLG		BASIC	78.9	94.8
VLG		FULL	**81.8**	**95.3**

(b) Ablation studies of text descriptions on Kinetics-LT

Method	Backbone	Operation	Overall	Many	Medium	Few
CLIP-LT	ResNet-50	NO TEXT	53.4	70.3	53.3	31.3
VLG		BASIC	57.8	71.4	56.9	35.8
VLG		FULL	**60.8**	**71.7**	**60.4**	**47.2**

Bold indicates the best result

"NO TEXT" denotes using no text descriptions for training, same as the CLIP-Close and CLIP-LT. "BASIC" denotes only using basic prompted sentences for training, and "FULL" denotes using basic prompted sentences and crawled text descriptions for training

Table 11 Ablation studies of different splitting strategies

Method	Backbone	Operation	Overall	Many	Medium	Few
VLG	ResNet-50	RAND	60.8	71.7	60.4	47.2
		GOOGLE	61.2	71.4	62.7	44.2

"RAND" denotes using the original splits in our Kinetics-LT, which are randomly chosen. "GOOGLE" denotes using the splits sorted by the number of entries in Google search

achieves higher performance in medium-shot, few-shot and overall cases, compared to the one without distribution loss. It indicates that the distillation loss helps the model learn better video-language representation with limited data. To further study the influence of distillation in the pre-training stage, we try to use the pre-trained CLIP model as the teacher model to distill the video and language encoder of our model at the feature level, in addition to the logits distillation. As shown in Table 8, both feature distillation and logits distillation with α of 0.5 can improve the performance in many-shot, medium-shot, few-shot and overall cases. And our method achieves the highest performance on Kinetics-LT when using logits distillation with the loss weight α of 0.5.

In the second finetune stage, we calculated two CrossEntropy loss items as shown in Eq. 9. The first term P^V is based on the video-only embedding E^V, and the second term P^T is based on the enhanced text embedding G. The first term adopts the MLP to obtain the classification probability, and the second term calculates the cosine similarity between the video-only embedding E^V and the enhanced text embedding G. To study the effectiveness of these two terms, we add experiments by adopting the first term, the second term, or both in the closed-set and long-tailed set. It can be seen in Table 9 that in the closed-set, the model with both of the two terms performs better than the others, indicating the power of video-language representation when given abundant training data. In the long-tailed case, the model with only P^V performs well in the "Many" case but performs poorly in the "Few" case, while the model with only P^T performs well in the "Few" case but performs poorly in the "Many" case. By contrast, the model with both of the two terms serves as the trade-off without sacrificing too much performance for all cases, and further improves the overall accuracy for the long-tailed datasets. Therefore, we hold that both the two terms are necessary.

Dataset To further verify the rationality of the design of our Kinetics-Text benchmark, we also carry out ablation studies for the collected text descriptions and the splits for the labels in the Kinetics-LT.

We study the significance of our collected text descriptions by replacing them with "Using no sentences" operation and "Only using basic prompted sentences" operation, both in Kinetics-Close and Kinetics-LT. As shown in Table 10, using the extra class-wise text description crawled from Wiki and Wikihow can significantly improve the performance in both Kinetics-Close and Kinetics-LT. Specifically, in the few-shot classes of Kinetics-LT, one can observe that VLG with both basic prompted sentences and crawled text description gets absolute accuracy gain against VLG with only basic prompts and VLG without text descriptions (47.2% vs. 35.8% vs. 31.3%). It indicates the validity of text descriptions from the Internet, and effectiveness of leveraging abundant semantic knowledge to make up for the lack of video data.

The Kinetics-LT is curated by sampling a subset following the Pareto distribution and its categories are also randomly split into different parts (many-case, medium-case, few-case), which is similar to ImageNet-LT (Liu et al., 2019). To demonstrate the rationality of the label splitting in our

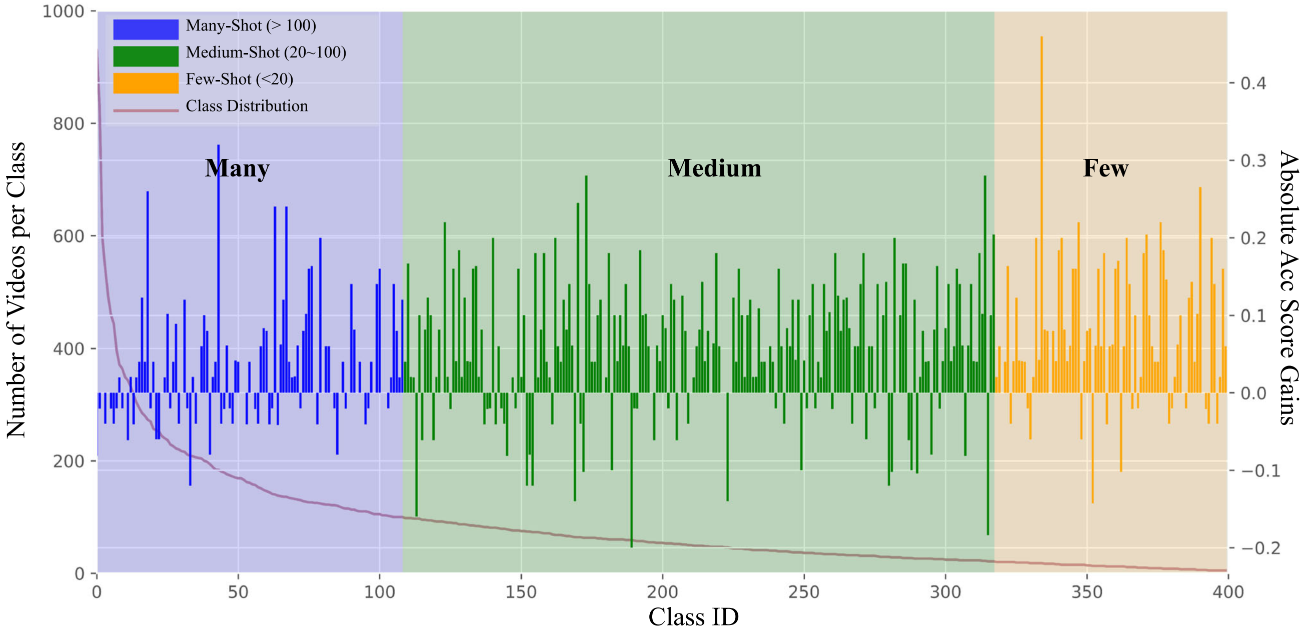


Fig. 4 Absolute accuracy score of our method over the baseline using ViT-B/16 as the backbone on Kinetics-LT. Our method enjoys more performance gains in classes with fewer video samples

Kinetics-LT, we adopt another strategy to re-split the classes into different parts according to their number of entries in Google search. As shown in the Table 11, there are no apparent changes in the recognition results, indicating the rationality of our label splitting in the long-tailed case. The reason for this phenomenon can be that the meaning of "Long tail" to a model consists in the distribution of the fed training data rather than the natural characteristics, and our video-language framework is also capable of using the class-wise text descriptions as a bridge to enhance the representation for the categories with fewer samples.

5.8 Visualization

Class-level Performance Improvement In Fig. 4, we visualize the class-level performance improvement on Kinetics-LT, which is measured by the absolute accuracy gains of our method against the baseline, both of which use ViT-B/16 as the visual backbone. We observe that our VLG outperforms the baseline on all cases of the Kinetics-LT dataset. Specifically, compared with the many-shot classes and medium-shot cases, there are more gains in the few-shot classes, indicating the introduced text descriptions can help mitigate the long-tail problem. It can be attributed to the rich semantic knowledge brought from the web text descriptions for the video representation in few-shot cases.

Visualization of Performance We use a radar chart to summarize the results across all regimes in Fig. 5. The shape and area of the radar chat can serve as the total result to quantify the effectiveness and generalization ability of VLG. We compare our method with current state-of-the-art methods in the radar chat, indicating the superiority of VLG over all settings.

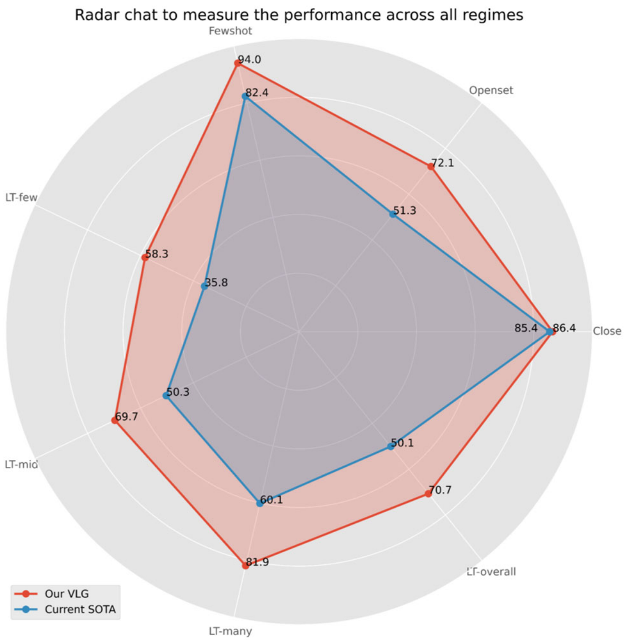


Fig. 5 Radar chat to measure the performance across all regimes. It can be seen that VLG outperforms current state-of-the-art methods for all settings

Case Analysis For more case analysis about the caption and categories, we present example videos and captions, along with their corresponding scores, for Kinetics400 cat-

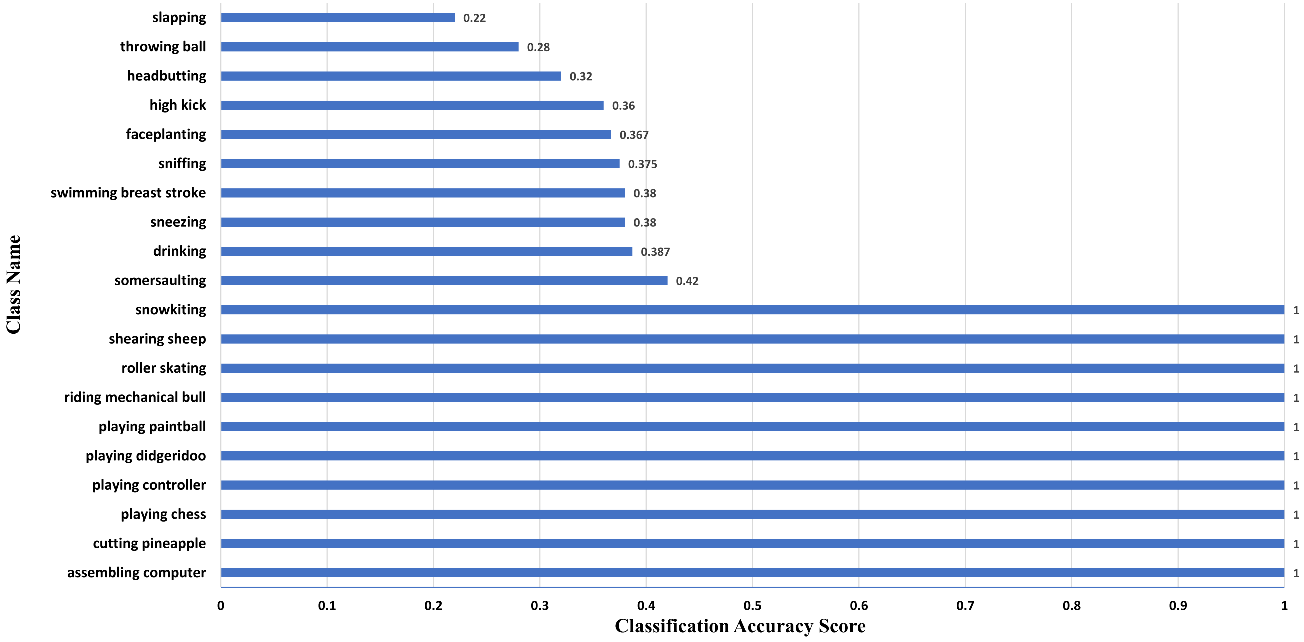


Fig. 6 The Top-10 classes in Kinetics400 that gain the highest and lowest classification scores. Categories with poor classification performance often have less related textual descriptions and typically involve subtle actions (e.g., *headbutting*, *slapping*, *faceplanting*, etc.) that occur in a brief moment, making it difficult to capture salient frames during the frame sampling process. Corresponding samples are provided in Fig. 10

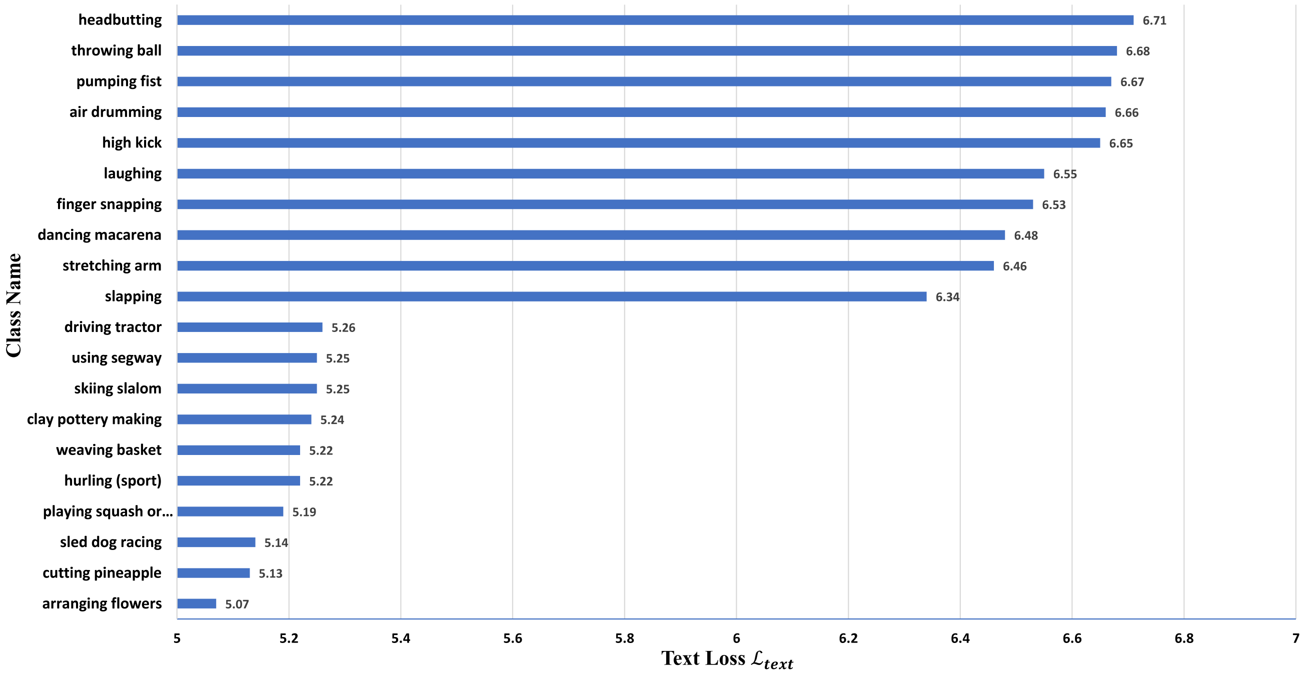
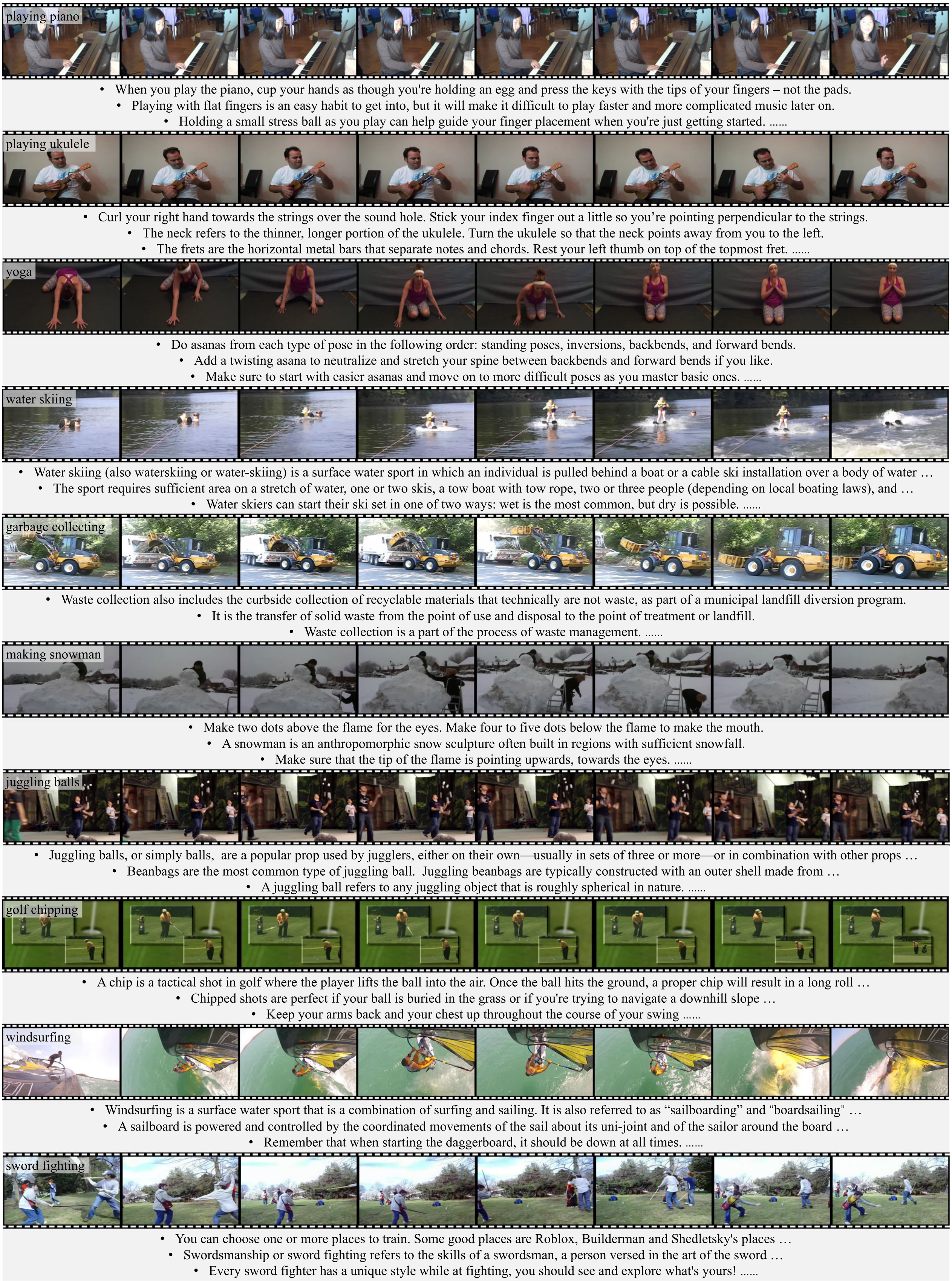


Fig. 7 The Top-10 classes in Kinetics400 that have the highest and lowest median of text loss $\mathcal{L}_{\text{text}}$. By comparing Figs. 6 and 7, it can be observed that categories with the highest median of text loss (e.g., *headbutting*, *throwing ball*, *high kick*, etc.) tend to have poorer classification performance



- When you play the piano, cup your hands as though you're holding an egg and press the keys with the tips of your fingers – not the pads.
- Playing with flat fingers is an easy habit to get into, but it will make it difficult to play faster and more complicated music later on.
- Holding a small stress ball as you play can help guide your finger placement when you're just getting started. ......

- Curl your right hand towards the strings over the sound hole. Stick your index finger out a little so you're pointing perpendicular to the strings.
- The neck refers to the thinner, longer portion of the ukulele. Turn the ukulele so that the neck points away from you to the left.
- The frets are the horizontal metal bars that separate notes and chords. Rest your left thumb on top of the topmost fret. ......

- Do asanas from each type of pose in the following order: standing poses, inversions, backbends, and forward bends.
- Add a twisting asana to neutralize and stretch your spine between backbends and forward bends if you like.
- Make sure to start with easier asanas and move on to more difficult poses as you master basic ones. ......

- Water skiing (also waterskiing or water-skiing) is a surface water sport in which an individual is pulled behind a boat or a cable ski installation over a body of water …
- The sport requires sufficient area on a stretch of water, one or two skis, a tow boat with tow rope, two or three people (depending on local boating laws), and …
- Water skiers can start their ski set in one of two ways: wet is the most common, but dry is possible. ......

- Waste collection also includes the curbside collection of recyclable materials that technically are not waste, as part of a municipal landfill diversion program.
- It is the transfer of solid waste from the point of use and disposal to the point of treatment or landfill.
- Waste collection is a part of the process of waste management. ......

- Make two dots above the flame for the eyes. Make four to five dots below the flame to make the mouth.
- A snowman is an anthropomorphic snow sculpture often built in regions with sufficient snowfall.
- Make sure that the tip of the flame is pointing upwards, towards the eyes. ......

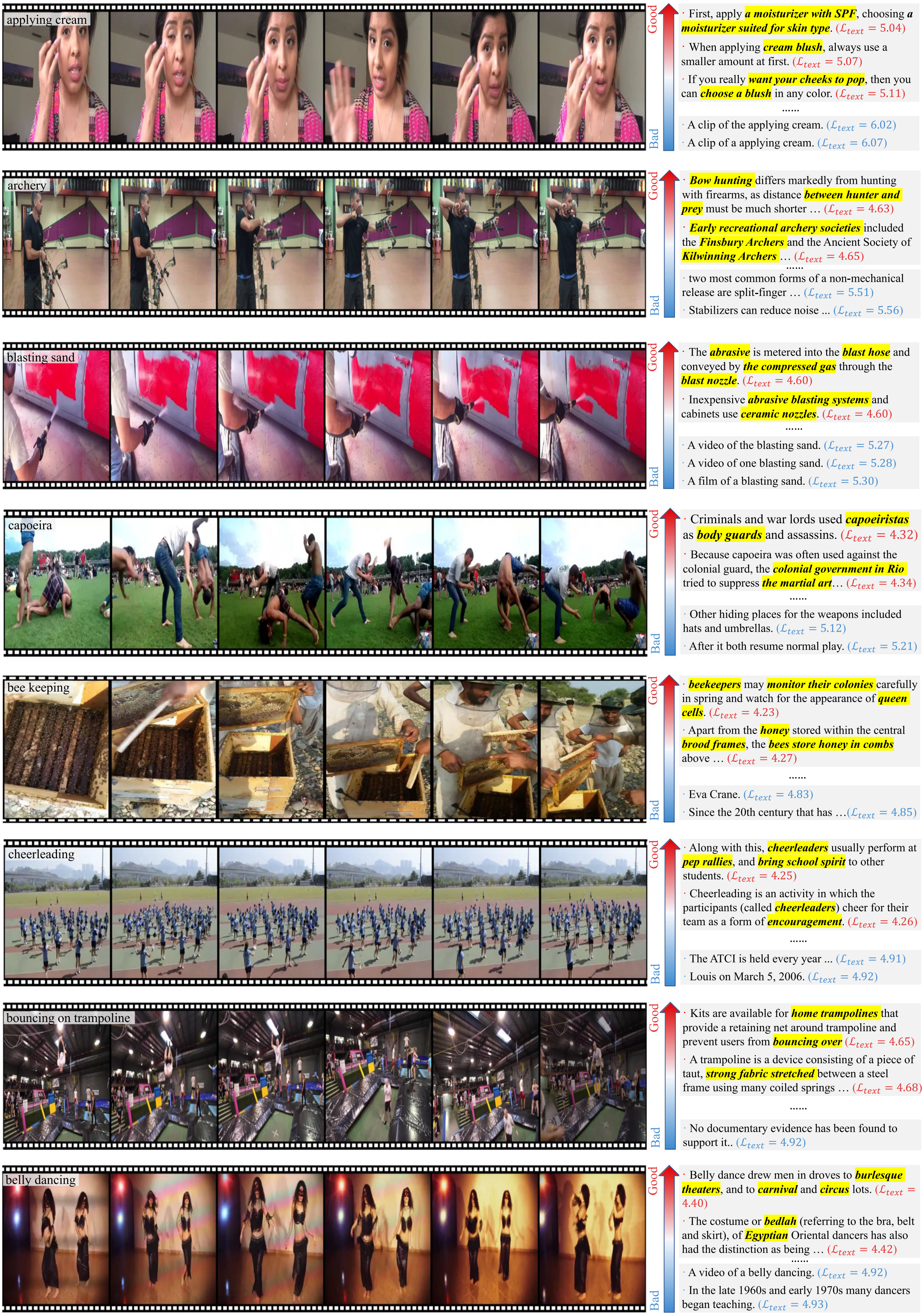
- Juggling balls, or simply balls, are a popular prop used by jugglers, either on their own—usually in sets of three or more—or in combination with other props …
- Beanbags are the most common type of juggling ball. Juggling beanbags are typically constructed with an outer shell made from …
- A juggling ball refers to any juggling object that is roughly spherical in nature. ......

- A chip is a tactical shot in golf where the player lifts the ball into the air. Once the ball hits the ground, a proper chip will result in a long roll …
- Chipped shots are perfect if your ball is buried in the grass or if you're trying to navigate a downhill slope …
- Keep your arms back and your chest up throughout the course of your swing ......

- Windsurfing is a surface water sport that is a combination of surfing and sailing. It is also referred to as “sailboarding” and “boardsailing” …
- A sailboard is powered and controlled by the coordinated movements of the sail about its uni-joint and of the sailor around the board …
- Remember that when starting the daggerboard, it should be down at all times. ......

- You can choose one or more places to train. Some good places are Roblox, Builderman and Shedletsky's places …
- Swordsmanship or sword fighting refers to the skills of a swordsman, a person versed in the art of the sword …
- Every sword fighter has a unique style while at fighting, you should see and explore what's yours! ......

Fig. 8 Examples of text descriptions crawled from Wikipedia and wikiHow for Kinetics400. Both useful and redundant information can be found in this text corpus



· First, apply ***a moisturizer with SPF***, choosing ***a moisturizer suited for skin type***. ($\mathcal{L}_{text}$ = 5.04)

· When applying ***cream blush***, always use a smaller amount at first. ($\mathcal{L}_{text}$ = 5.07)

· If you really ***want your cheeks to pop***, then you can ***choose a blush*** in any color. ($\mathcal{L}_{text}$ = 5.11)

......

· A clip of the applying cream. ($\mathcal{L}_{text}$ = 6.02)

· A clip of a applying cream. ($\mathcal{L}_{text}$ = 6.07)

· ***Bow hunting*** differs markedly from hunting with firearms, as distance ***between hunter and prey*** must be much shorter … ($\mathcal{L}_{text}$ = 4.63)

· ***Early recreational archery societies*** included the ***Finsbury Archers*** and the Ancient Society of ***Kilwinning Archers*** … ($\mathcal{L}_{text}$ = 4.65)

......

· two most common forms of a non-mechanical release are split-finger … ($\mathcal{L}_{text}$ = 5.51)

· Stabilizers can reduce noise ... ($\mathcal{L}_{text}$ = 5.56)

· The ***abrasive*** is metered into the ***blast hose*** and conveyed by ***the compressed gas*** through the ***blast nozzle***. ($\mathcal{L}_{text}$ = 4.60)

· Inexpensive ***abrasive blasting systems*** and cabinets use ***ceramic nozzles***. ($\mathcal{L}_{text}$ = 4.60)

......

· A video of the blasting sand. ($\mathcal{L}_{text}$ = 5.27)

· A video of one blasting sand. ($\mathcal{L}_{text}$ = 5.28)

· A film of a blasting sand. ($\mathcal{L}_{text}$ = 5.30)

· Criminals and war lords used ***capoeiristas*** as ***body guards*** and assassins. ($\mathcal{L}_{text}$ = 4.32)

· Because capoeira was often used against the colonial guard, the ***colonial government in Rio*** tried to suppress ***the martial art***… ($\mathcal{L}_{text}$ = 4.34)

......

· Other hiding places for the weapons included hats and umbrellas. ($\mathcal{L}_{text}$ = 5.12)

· After it both resume normal play. ($\mathcal{L}_{text}$ = 5.21)

· ***beekeepers*** may ***monitor their colonies*** carefully in spring and watch for the appearance of ***queen cells***. ($\mathcal{L}_{text}$ = 4.23)

· Apart from the ***honey*** stored within the central ***brood frames***, the ***bees store honey in combs*** above … ($\mathcal{L}_{text}$ = 4.27)

......

· Eva Crane. ($\mathcal{L}_{text}$ = 4.83)

· Since the 20th century that has …($\mathcal{L}_{text}$ = 4.85)

· Along with this, ***cheerleaders*** usually perform at ***pep rallies***, and ***bring school spirit*** to other students. ($\mathcal{L}_{text}$ = 4.25)

· Cheerleading is an activity in which the participants (called ***cheerleaders***) cheer for their team as a form of ***encouragement***. ($\mathcal{L}_{text}$ = 4.26)

......

· The ATCI is held every year ... ($\mathcal{L}_{text}$ = 4.91)

· Louis on March 5, 2006. ($\mathcal{L}_{text}$ = 4.92)

· Kits are available for ***home trampolines*** that provide a retaining net around trampoline and prevent users from ***bouncing over*** ($\mathcal{L}_{text}$ = 4.65)

· A trampoline is a device consisting of a piece of taut, ***strong fabric stretched*** between a steel frame using many coiled springs … ($\mathcal{L}_{text}$ = 4.68)

......

· No documentary evidence has been found to support it.. ($\mathcal{L}_{text}$ = 4.92)

· Belly dance drew men in droves to ***burlesque theaters***, and to ***carnival*** and ***circus*** lots. ($\mathcal{L}_{text}$ = 4.40)

· The costume or ***bedlah*** (referring to the bra, belt and skirt), of ***Egyptian*** Oriental dancers has also had the distinction as being … ($\mathcal{L}_{text}$ = 4.42)

......

· A video of a belly dancing. ($\mathcal{L}_{text}$ = 4.92)

· In the late 1960s and early 1970s many dancers began teaching. ($\mathcal{L}_{text}$ = 4.93)

Fig. 9 More visualization of text descriptions with corresponding $\mathcal{L}_{\text{text}}$. The values of $\mathcal{L}_{\text{text}}$ reflect the saliency of these sentences, indicating the effectiveness of our proposed TSR

egories with the best and worst performance in Fig. 6 and Fig. 10. It can be inferred that the quality of the text descriptions corresponding to the categories, as well as the property of actions, will determine the final classification accuracy. Categories with poor classification performance often have less related textual descriptions to the visual frames. For example, the text description in Wikipedia of label *'faceplanting'* refers to a takedown move in professional wrestling, rather than the act of landing face first, as a result of an accident or error. In addition, categories with poor classification performance tend to involve subtle actions (e.g., *headbutting*, *slapping*, *throwing ball*, etc.) that occur in a brief moment, making it difficult to capture salient frames during the frame sampling process over the whole video.

Additionally, we also calculate the median of text loss $\mathcal{L}_{\text{text}}$ for each category and illustrate the distribution of categories with the highest and lowest median of text loss $\mathcal{L}_{\text{text}}$ in Fig. 7. It can be observed that categories with the highest median of text loss (e.g., *headbutting*, *throwing ball*, *high kick*, etc.) tend to have poorer classification performance, which corresponds to the analysis on Fig. 6. Accordingly, we can use the median of text loss $\mathcal{L}_{\text{text}}$ for each category to quantify the level of noise.

Visualization of Text Corpus In this section, we provide some visualization of the collected text corpus in Fig. 8. It can be seen that these texts contain not only some noisy information within them, but also some static characteristics, dynamic evolution, and logical definition of the corresponding categories. In addition, to intuitively demonstrate the effectiveness of our text selection ruler (TSR), we provide some examples of sentences reserved or dropped by our TSR of different categories in Fig. 9. We observe that our method can sample useful texts or filter out the useless ones. It can also be seen that VLG can learn specific concepts or steps for each class, such as "Bow hunting" for "archery" and "queen cells" for "bee keeping". The salient sentences commonly contain these words of specific concepts in the category.

5.9 Additional Experiments

To validate the generalization of VLG, we conducted experiments on the commonly used UCF-101 and HMDB-51 datasets for few-shot tasks. We use the same data splits from STRM (Thatipelli et al., 2022).

It can be seen from the Tables 12 and 13 that despite using a relatively simple spatiotemporal information modeling module, our proposed method still achieves good performance on UCF-101 and HMDB-51 (VLG-L). It is also noted that on the HMDB51 dataset, the few-shot results are not very good when linear probe testing is not used. This could be due to the fact that most labels in HMDB51 are vague generalizations of actions (e.g., kick, push, throw), with relatively

Table 12 Results on UCF101-Fewshot

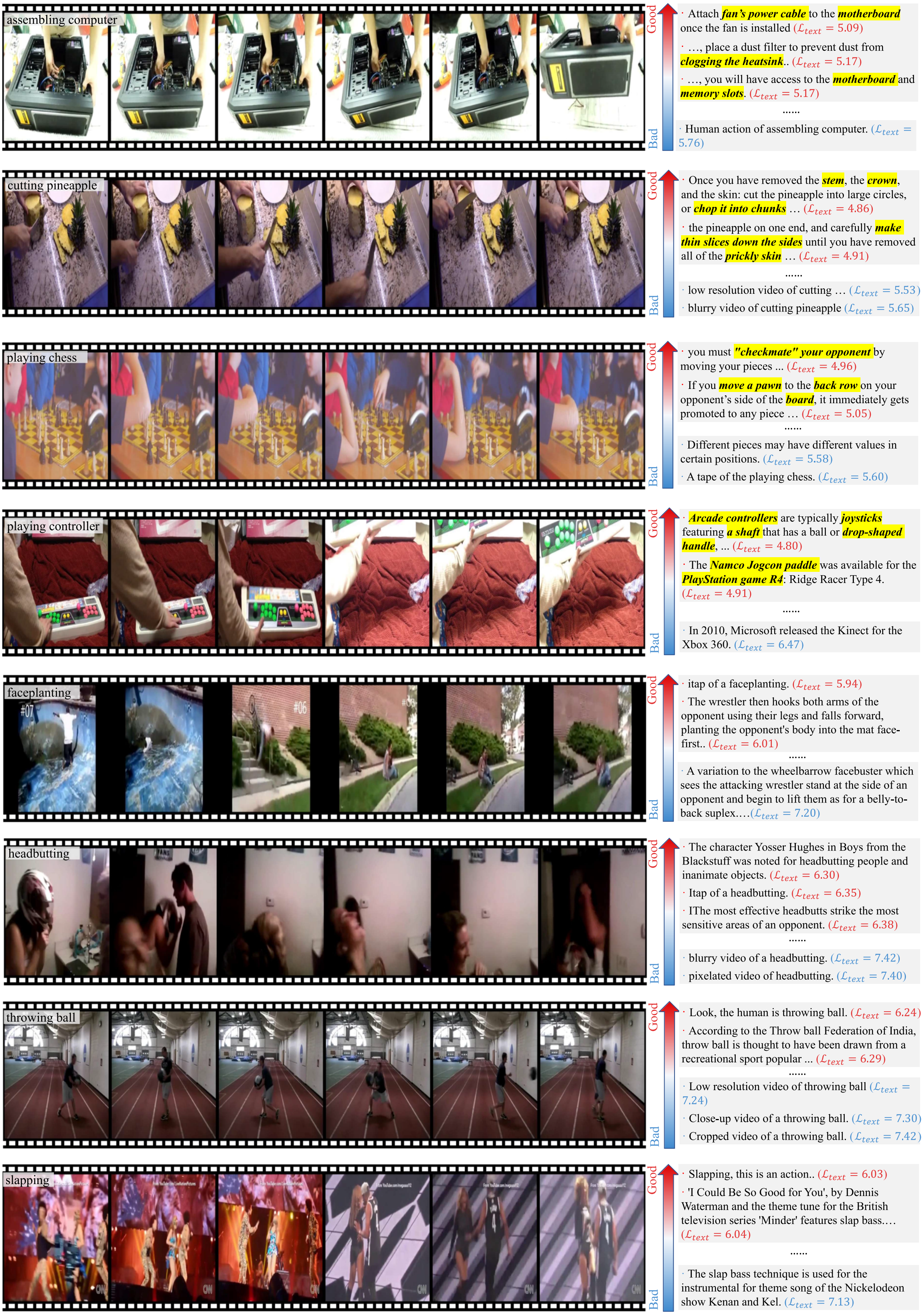
Method	Backbone	K-shot	N-way	Top-1
GenApp (Mishra et al., 2018)	ResNet-50	5	5	78.6
ProtoGAN (Kumar Dwivedi et al., 2019)		5	5	80.2
ARN (Zhang et al., 2020)		5	5	83.1
HF-AR (Kumar & Narang, 2021)		5	5	86.4
HyRSM (Wang et al., 2022)		5	5	94.7
STRM (Thatipelli et al., 2022)		5	5	96.8
VLG-L		5	5	90.4
VLG		5	5	93.8

Here, we use 'VLG-L' to denote our method with linear probe testing, and use 'VLG' to denote our proposed two-stage training framework

Table 13 Results on HMDB51-Fewshot

Method	Backbone	K-shot	N-way	Top-1
GenApp (Mishra et al., 2018)	ResNet-50	5	5	52.5
ProtoGAN (Kumar Dwivedi et al., 2019)		5	5	54.0
ARN (Zhang et al., 2020)		5	5	60.6
HF-AR (Kumar & Narang, 2021)		5	5	62.2
HyRSM (Wang et al., 2022)		5	5	76.0
STRM (Thatipelli et al., 2022)		5	5	77.3
VLG-L		5	5	76.0
VLG		5	5	34.2

Here, we use 'VLG-L' to denote our method with linear probe testing, and use 'VLG' to denote our proposed two-stage training framework



· Attach ***fan's power cable*** to the ***motherboard*** once the fan is installed ($\mathcal{L}_{text} = 5.09$)

· …, place a dust filter to prevent dust from ***clogging the heatsink***.. ($\mathcal{L}_{text} = 5.17$)

· …, you will have access to the ***motherboard*** and ***memory slots***. ($\mathcal{L}_{text} = 5.17$)

……

· Human action of assembling computer. ($\mathcal{L}_{text} = 5.76$)

· Once you have removed the ***stem***, the ***crown***, and the skin: cut the pineapple into large circles, or ***chop it into chunks*** … ($\mathcal{L}_{text} = 4.86$)

· the pineapple on one end, and carefully ***make thin slices down the sides*** until you have removed all of the ***prickly skin*** … ($\mathcal{L}_{text} = 4.91$)

……

· low resolution video of cutting … ($\mathcal{L}_{text} = 5.53$)

· blurry video of cutting pineapple ($\mathcal{L}_{text} = 5.65$)

· you must ***"checkmate" your opponent*** by moving your pieces ... ($\mathcal{L}_{text} = 4.96$)

· If you ***move a pawn*** to the ***back row*** on your opponent's side of the ***board***, it immediately gets promoted to any piece … ($\mathcal{L}_{text} = 5.05$)

……

· Different pieces may have different values in certain positions. ($\mathcal{L}_{text} = 5.58$)

· A tape of the playing chess. ($\mathcal{L}_{text} = 5.60$)

· ***Arcade controllers*** are typically ***joysticks*** featuring ***a shaft*** that has a ball or ***drop-shaped handle***, ... ($\mathcal{L}_{text} = 4.80$)

· The ***Namco Jogcon paddle*** was available for the ***PlayStation game R4***: Ridge Racer Type 4. ($\mathcal{L}_{text} = 4.91$)

……

· In 2010, Microsoft released the Kinect for the Xbox 360. ($\mathcal{L}_{text} = 6.47$)

· itap of a faceplanting. ($\mathcal{L}_{text} = 5.94$)

· The wrestler then hooks both arms of the opponent using their legs and falls forward, planting the opponent's body into the mat face-first.. ($\mathcal{L}_{text} = 6.01$)

……

· A variation to the wheelbarrow facebuster which sees the attacking wrestler stand at the side of an opponent and begin to lift them as for a belly-to-back suplex.…($\mathcal{L}_{text} = 7.20$)

· The character Yosser Hughes in Boys from the Blackstuff was noted for headbutting people and inanimate objects. ($\mathcal{L}_{text} = 6.30$)

· Itap of a headbutting. ($\mathcal{L}_{text} = 6.35$)

· IThe most effective headbutts strike the most sensitive areas of an opponent. ($\mathcal{L}_{text} = 6.38$)

……

· blurry video of a headbutting. ($\mathcal{L}_{text} = 7.42$)

· pixelated video of headbutting. ($\mathcal{L}_{text} = 7.40$)

· Look, the human is throwing ball. ($\mathcal{L}_{text} = 6.24$)

· According to the Throw ball Federation of India, throw ball is thought to have been drawn from a recreational sport popular ... ($\mathcal{L}_{text} = 6.29$)

……

· Low resolution video of throwing ball ($\mathcal{L}_{text} = 7.24$)

· Close-up video of a throwing ball. ($\mathcal{L}_{text} = 7.30$)

· Cropped video of a throwing ball. ($\mathcal{L}_{text} = 7.42$)

· Slapping, this is an action.. ($\mathcal{L}_{text} = 6.03$)

· 'I Could Be So Good for You', by Dennis Waterman and the theme tune for the British television series 'Minder' features slap bass.… ($\mathcal{L}_{text} = 6.04$)

……

· The slap bass technique is used for the instrumental for theme song of the Nickelodeon show Kenan and Kel. ($\mathcal{L}_{text} = 7.13$)

Fig. 10 More visualization of categories with best and worst accuracy performance. The top four lines correspond to the categories with the best classification performance, while the bottom four lines correspond to the categories with the worst classification performance

little visual commonality across HMDB51 videos, making it difficult to correspond to specific textual descriptions. The overly ambiguous labels in HMDB51 make it challenging to find well-matched textual descriptions on platforms like Wiki and WikiHow. Furthermore, there are hierarchical relationships between categories (e.g., "sword" v.s. "sword exercise"), causing a single description to correspond to two labels.

6 Conclusions and Future Work

Video recognition in an open and wild world is extremely difficult. Despite the fact that multiple video benchmarks and works have been developed to study video recognition in various scenarios, these separate investigations would ignore the possibility of knowledge sharing across settings, leading to inefficiency and impeding progress in video recognition as well as its application in the real world.

In this paper, we have studied the general video recognition (GVR) under four different settings: close-set, long-tail, few-shot and open-set. The GVR task enables us to examine the generalization ability of a video recognition model in real-world applications. To facilitate the research of GVR, we build comprehensive video benchmarks of Kinetics-GVR containing text descriptions for all action classes. Then, we propose a unified visual-linguistic framework (VLG) to accomplish the task of GVR. In particular, we present an effective two-stage training strategy to effectively adapt the image-text representation to video domain for GVR. Extensive results demonstrate that our VLG obtains the state-of-the-art performance under all settings on the Kinetics-GVR benchmark.

Although our VLG achieves superior performance on multiple general video recognition settings, it still needs a two-stage training paradigm and cannot be end-to-end trained. To tackle this, we can apply reinforcement learning with reward functions (Lin et al., 2022a; Meng et al., 2020) or gumbel-softmax tricks (Jang et al., 2016) to further improve the non-differentiable text selection parts. In addition, it might be difficult to crawl suitable descriptions of labels from Wiki or WikiHow for subtle actions, like "Put the glass on top of the table". Probably, it needs to participle phrases and crawl definitions from some dictionary websites as supplementary to improve the text descriptions. In summary, we hope the empirical findings and insights presented in this work could pave the way for future research on general video recognition which is still a nascent research topic.

Acknowledgements This work is supported by the National Key R & D Program of China (No. 2022ZD0160900), the National Natural Science Foundation of China (No. 62076119, No. 61921006), the Fundamental Research Funds for the Central Universities (No. 020214380091), and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

Data Availability The datasets generated during and/or analysed during the current study are available in the Kinetics repository: https://www.deepmind.com/open-source/kinetics.

References

Acsintoae, A., Florescu, A., Georgescu, M. I., Mare, T., Sumedrea, P., Ionescu, R. T., Khan, F. S., Shah, M. (2021). Ubnormal: New benchmark for supervised open-set video anomaly detection. arXiv preprint arXiv:2111.08644

Akbari, H., Yuan, L., Qian, R., Chuang, W. H., Chang, S. F., Cui, Y., & Gong, B. (2021). Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems, 34*, 24206–24221.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6836–6846)

Bain, M., Nagrani, A., Varol, G., & Zisserman, A. (2022). A clip-hitchhiker's guide to long video retrieval. arXiv preprint arXiv:2205.08508

Bao, W., Yu, Q., & Kong, Y. (2021). Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 13349–13358)

Bendale, A., & Boult, T. E. (2016). Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1563–1572)

Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? In *ICML*, vol 2 (pp. 4)

Bishay, M., Zoumpourlis, G., & Patras, I. (2019). Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. arXiv preprint arXiv:1907.09021

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks, 106*, 249–259.

Caba Heilbron, F., Escorcia, V., Ghanem, B., & Carlos Niebles, J. (2015). 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)

Cao, D., Xu, L., & Chen, H. (2020a). Action recognition in untrimmed videos with composite self-attention two-stream framework. In *Pattern recognition: 5th Asian conference, ACPR 2019, Auckland, New Zealand*, November 26–29, 2019, Revised Selected Papers, Part II 5 (pp. 27–40). Springer

Cao, K., Wei, C., Gaidon, A., Arechiga, N., & Ma, T. (2019). Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in neural information processing systems* 32

Cao, K., Ji, J., Cao, Z., Chang, C. Y., & Niebles, J. C. (2020b). Few-shot video classification via temporal alignment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10618–10627)

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6299–6308)

Carreira, J., Noland, E., Hillier, C., & Zisserman, A. (2019). A short note on the kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Chu, P., Bian, X., Liu, S., & Ling, H. (2020). Feature space augmentation for long-tailed data. In *European conference on computer vision* (pp. 694–710). Springer

Cui, J., Zhong, Z., Liu, S., Yu, B., & Jia, J. (2021). Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 715–724)

Cui, Y., Jia, M., Lin, T. Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9268–9277)

Diba, A., Fayyaz, M., Sharma, V., Arzani, M. M., Yousefzadeh, R., Gall, J., & Van Gool, L. (2018). Spatio-temporal channel correlation networks for action classification. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 284–299)

Ditria, L., Meyer, BJ., & Drummond, T. (2020). Opengan: Open set generative adversarial networks. In *Proceedings of the Asian conference on computer vision*

Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *2005 IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance* (pp. 65–72). IEEE

Dong, J., Li, X., Xu, C., Yang, X., Yang, G., Wang, X., & Wang, M. (2021). Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2021.3059295

Drumnond, C., & Holte, R. (2003). Class imbalance and cost sensitivity: Why undersampling beats oversampling. In *ICML-KDD 2003 workshop: Learning from imbalanced datasets*, vol 3

Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., & Feichtenhofer, C. (2021). Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6824–6835)

Feichtenhofer, C. (2020). X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 203–213)

Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1933–1941)

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6202–6211)

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning, PMLR* (pp. 1126–1135)

Finn, C., Xu, K., & Levine, S. (2018). Probabilistic model-agnostic meta-learning. In *Advances in neural information processing systems* 31

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., & Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems* 26

Ge, Z., Demyanov, S., Chen, Z., & Garnavi, R. (2017). Generative openmax for multi-class open set classification. arXiv preprint arXiv:1707.07418

Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., & et al. (2017). 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)

Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing* (pp. 878–887). Springer

Huang, C., Li, Y., Loy, C. C., & Tang, X. (2016). Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5375–5384)

Jain, LP., Scheirer, W. J., & Boult, T. E. (2014). Multi-class open set recognition using probability of inclusion. In *European conference on computer vision* (pp. 393–409). Springer

Jang, E., Gu, S., & Poole, B. (2016). Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144

Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., Le, Q., Sung, Y. H., Li, Z., & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning, PMLR* (pp. 4904–4916)

Jiang, B., Wang, M., Gan, W., Wu, W., & Yan, J. (2019). Stm: Spatiotemporal and motion encoding for action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2000–2009)

Ju, C., Han, T., Zheng, K., Zhang, Y., & Xie, W. (2022). Prompting visual-language models for efficient video understanding. In *European conference on computer vision* (pp. 105–124). Springer

Kahatapitiya, K., Arnab, A., Nagrani, A., & Ryoo, M. S. (2023). Victr: Video-conditioned text representations for activity recognition. arXiv preprint arXiv:2304.02560

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., & Kalantidis, Y. (2019). Decoupling representation and classifier for long-tailed recognition. arXiv preprint arXiv:1910.09217

Kant, Y., Batra, D., Anderson, P., Schwing, A., Parikh, D., Lu, J., & Agrawal, H. (2020). Spatially aware multimodal transformers for textvqa. In *European conference on computer vision* (pp. 715–732). Springer

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., & et al. (2017). The kinetics human action video dataset. arXiv preprint arXiv:1705.06950

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., & Togneri, R. (2017). Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems, 29*(8), 3573–3587.

Klaser, A., Marszałek, M., & Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British machine vision conference* (pp. 275–1). British Machine Vision Association

Krishnan, R., Subedar, M., & Tickoo, O. (2018). Bar: Bayesian activity recognition using variational inference. arXiv preprint arXiv:1811.03305

Krishnan, R., Subedar, M., & Tickoo, O. (2020). Specifying weight priors in Bayesian deep neural networks with empirical Bayes. In*Proceedings of the AAAI conference on artificial intelligence, 34*, 4477–4484.

Kumar, N., & Narang, S. (2021). Few shot activity recognition using variational inference. arXiv preprint arXiv:2108.08990

Kumar Dwivedi, S., Gupta, V., Mitra, R., Ahmed, S., & Jain, A. (2019). Protogan: Towards few shot learning for action recognition. In *Proceedings of the IEEE/CVF International conference on computer vision workshops*

Kumawat, S., Verma, M., Nakashima, Y., & Raman, S. (2021). Depthwise spatio-temporal STFT convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2021.3076522

Li, F., & Wechsler, H. (2005). Open set face recognition using transduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(11), 1686–1697.

Li, T., & Wang, L. (2020a). Learning spatiotemporal features via video and text pair discrimination. arXiv preprint arXiv:2001.05691

Li, T., & Wang, L. (2020b). Learning spatiotemporal features via video and text pair discrimination. CoRR arXiv: 2001.05691

Li, T., Wang, L., & Wu, G. (2021a). Self supervision to distillation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 630–639)

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., & et al. (2020a). Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European conference on computer vision* (pp. 121–137). Springer

Li, Y., Ji, B., Shi, X., Zhang, J., Kang, B., & Wang, L. (2020b). Tea: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 909–918)

Li, Y., Wu, CY., Fan, H., Mangalam, K., Xiong, B., Malik, J., & Feichtenhofer, C. (2021b). Improved multiscale vision transformers for classification and detection. arXiv preprint arXiv:2112.01526

Li, Z., Fan, Z., Tou, H., & Wei, Z. (2022). Mvp: Multi-stage vision-language pre-training via multi-level semantic alignment. arXiv preprint arXiv:2201.12596

Lin, J., Gan, C., & Han, S. (2019). Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7083–7093)

Lin, J., Duan, H., Chen, K., Lin, D., & Wang, L. (2022a) Ocsampler: Compressing videos to one clip with single-step sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13894–13903)

Lin, Z., Geng, S., Zhang, R., Gao, P., de Melo, G., Wang, X., Dai, J., Qiao. Y., & Li, H. (2022b) Frozen clip models are efficient video learners. In *European conference on computer vision* (pp. 388–404). Springer

Liu, Y., Chen, Q., & Albanie, S. (2021a) Adaptive cross-modal prototypes for cross-domain visual-language retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14954–14964)

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., & Yu, S. X. (2019) Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2537–2546)

Liu, Z., Wang, L., Wu, W., Qian, C., & Lu, T. (2021b) TAM: Temporal adaptive module for video recognition. In *ICCV* (pp. 13688–13698). IEEE

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2022) Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3202–3211)

Loshchilov, I., & Hutter, F. (2016) Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983

Loshchilov, I., & Hutter, F. (2017) Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101

Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., & Li, T. (2022). Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing, 508*, 293–304.

Meng, Y., Lin, CC., Panda, R., Sattigeri, P., Karlinsky, L., Oliva, A., Saenko, K., & Feris, R. (2020). Ar-net: Adaptive frame resolution for efficient action recognition. In *European conference on computer vision* (pp. 86–104). Springer

Miech, A., Alayrac, J. B., Smaira, L., Laptev, I., Sivic, J., & Zisserman, A. (2020a). End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*

Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., & Zisserman, A. (2020b). End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9879–9889)

Mishra, A., Verma, V. K., Reddy, M. S. K., Arulkumar, S., Rai, P., & Mittal, A. (2018) A generative approach to zero-shot and few-shot action recognition. In *2018 IEEE winter conference on applications of computer vision (WACV)* (pp. 372–380). IEEE

Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C., et al. (2019). Moments in time dataset: One million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(2), 502–508.

Monfor, M., Pan, B., Ramakrishnan, K., Andonian, A., McNamara, B. A., Lascelles, A., Fan, Q., Gutfreund, D., Feris, R., & Oliva, A. (2021). Multi-moments in time: Learning and interpreting models for multi-action video understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2021.3126682

Mori, Y., Takahashi, H., & Oka, R. (1999). Image-to-word transformation based on dividing and vector quantizing images with words. In *First international workshop on multimedia intelligent storage and retrieval management* (pp. 1–9). Citeseer

Neal, L., Olson, M., Fern, X., Wong, W. K., & Li, F. (2018). Open set learning with counterfactual images. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 613–628)

Neimark, D., Bar, O., Zohar, M., & Asselmann, D. (2021). Video transformer network. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3163–3172)

Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., & Ling, H. (2022) Expanding language-image pretrained models for general video recognition. In *European conference on computer vision* (pp. 1–18). Springer

Oza, P., & Patel, V. M. (2019). C2ae: Class conditioned auto-encoder for open-set recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2307–2316)

Pan, J., Lin, Z., Zhu, X., Shao, J., & Li, H. (2022). St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems, 35*, 26462–26477.

Perrett, T., Masullo, A., Burghardt, T., Mirmehdi, M., & Damen, D. (2021). Temporal-relational crosstransformers for few-shot action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 475–484)

Qian, R., Li, Y., Xu, Z., Yang, M. H., Belongie, S., & Cui, Y. (2022). Multimodal open-vocabulary video classification via pre-trained vision and language models. arXiv preprint arXiv:2207.07646

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., & Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PMLR

Reed, W. J. (2001). The pareto, zipf and other power laws. *Economics letters, 74*(1), 15–19.

Ruan, L., & Jin, Q. (2022). Survey: Transformer based video-language pre-training. *AI Open, 3*, 1–13.

Ryoo, M. S., Piergiovanni A, Arnab, A., Dehghani, M., & Angelova, A. (2021). Tokenlearner: What can 8 learned tokens do for images and videos? arXiv preprint arXiv:2106.11297

Scheirer, W. J., de Rezende, R. A., Sapkota, A., & Boult, T. E. (2012). Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(7), 1757–1772.

Scheirer, W. J., Jain, L. P., & Boult, T. E. (2014). Probability models for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*(11), 2317–2324.

Sharir, G., Noy, A., Zelnik-& Manor, L. (2021). An image is worth 16x16 words, what is a video worth? arXiv preprint arXiv:2103.13915

Shen, L., Lin, Z., & Huang, Q.(2016). Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision* (pp. 467–482). Springer

Shu, Y., Shi, Y., Wang, Y., Zou, Y., Yuan, Q., & Tian, Y. (2018). Odn: Opening the deep network for open-set action recognition. In *2018 IEEE international conference on multimedia and expo (ICME)* (pp. 1–6). IEEE

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems* 27

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., & Rohrbach, M. (2019). Towards VQA models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8317–8326)

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. In *Advances in neural information processing systems* 30

Soomro, K., Zamir, A. R., & Shah, M (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402

Stroud, J., Ross, D., Sun, C., Deng, J., & Sukthankar, R. (2020a). D3d: Distilled 3d networks for video action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 625–634)

Stroud, J. C, Ross, D. A., Sun, C., Deng, J., Sukthankar, R., & Schmid, C. (2020b). Learning video representations from textual web supervision. CoRR abs/2007.14937, https://arxiv.org/abs/2007.14937, 2007.14937

Subedar, M., Krishnan, R., Meyer, P. L., Tickoo, O., & Huang, J. (2019). Uncertainty-aware audiovisual activity recognition using deep Bayesian variational inference. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6301–6310)

Sun, X., Yang, Z., Zhang, C., Ling, K. V., & Peng, G. (2020). Conditional gaussian distribution learning for open set recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13480–13489)

Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., & Yan, J. (2020). Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11662–11671)

Thatipelli, A., Narayan, S., Khan, S., Anwer, R. M., Khan, F. S., & Ghanem, B. (2022). Spatio-temporal relation modeling for few-shot action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19958–19967)

Tian, C., Wang, W., Zhu, X., Dai, J., & Qiao, Y. (2022). VL-LTR: Learning class-wise visual-linguistic representation for long-tailed visual recognition. In *European conference on computer vision* (pp. 73–91). Springer

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6450–6459)

Vinyals, O., Blundell, C., Lillicrap, T., & Wierstra, D., et al. (2016). Matching networks for one shot learning. In *Advances in neural information processing systems* 29

Wang, H., Kläser, A., Schmid, C., & Liu, C. L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision, 103*(1), 60–79.

Wang, J., Ge, Y., Yan, R., Ge, Y., Lin, K. Q., Tsutsui, S., Lin, X., Cai, G., Wu, J., & Shan, Y., et al. (2023). All in one: Exploring unified video-language pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6598–6608)

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Gool, L. V. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision* (pp. 20–36). Springer

Wang, L., Tong, Z., Ji, B., & Wu, G. (2021a). TDN: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 1895–1904)

Wang, M., Xing, J., & Liu, Y. (2021b). Actionclip: A new paradigm for video action recognition. arXiv preprint arXiv:2109.08472

Wang, W., Feiszli, M., Wang, H., & Tran, D. (2021c). Unidentified video objects: A benchmark for dense, open-world segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10776–10785)

Wang, X., Liu, Y., Shen, C., Ng, C. C., Luo, C., Jin, L.,& Chan, C. S., Hengel, A., & Wang, L. (2020). On the general value of evidence, and bilingual scene-text visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10126–10135)

Wang, X., Zhang, S., Qing, Z., Tang, M., Zuo, Z., Gao, C., Jin, R., & Sang, N. (2022). Hybrid relation guided set matching for few-shot action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19948–19957)

Wang, Y. X., Ramanan, D., & Hebert, M. (2017). Learning to model the tail. In *Advances in Neural Information Processing Systems* 30

Weston, J., Bengio, S., & Usunier, N. (2011). Wsabie: Scaling up to large vocabulary image annotation. In *Twenty-second international joint conference on artificial intelligence*

wikiHow. (2022). wikhow, the most trusted how-to site on the internet. https://www.wikihow.com/Main-Page, Retrieved from May 19, 2022

Wikipedia. (2022). Wikipedia, the free encyclopedia. https://www.wikipedia.org/, Retrieved from May 19, 2022

Wu, W., Sun, Z., & Ouyang, W. (2023). Revisiting classifier: Transferring vision-language models for video recognition. In *Proceedings of the AAAI conference on artificial intelligence, 37*, 2847–2855.

Xie, S., Sun, C., Huang, J., Tu, Z., & Murphy, K. (2018). Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 305–321)

Xu, H., Ghosh, G., Huang, P. Y., Arora, P., Aminzadeh, M., Feichtenhofer, C., Metze, F., & Zettlemoyer, L. (2021). VLM: Task-agnostic video-language model pre-training for video understanding. arXiv preprint arXiv:2105.09996

Yan, S., Xiong, X., Arnab, A., Lu, Z., Zhang, M., Sun, C., & Schmid, C. (2022). Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3333–3343)

Yang, X., Dong, J., Cao, Y., Wang, X., Wang, M., Chua, T. S. (2020). Tree-augmented cross-modal encoding for complex-query video retrieval. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 1339–1348)

Yin, X., Yu, X., Sohn, K., Liu, X., & Chandraker, M. (2019). Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5704–5713)

Yuan, L., Chen, D., Chen, Y. L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., & Li, C. et al. (2021). Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432

Zhang, B., Yu, J., Fifty, C., Han, W., Dai, A. M., Pang, R., & Sha, F. (2021a). Co-training transformer with videos and images improves action recognition. arXiv preprint arXiv:2112.07175

Zhang, H., Zhang, L., Qi, X., Li, H., Torr, P. H., & Koniusz, P. (2020). Few-shot action recognition with permutation-invariant attention. In *European conference on computer vision* (pp. 525–542). Springer

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., & Gao, J. (2021b). Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5579–5588)

Zhang, R., Che, T., Ghahramani, Z., Bengio, Y., & Song, Y. (2018). Metagan: An adversarial approach to few-shot learning. In *Advances in neural information processing systems* 31

Zhang, X., Wu, Z., Weng, Z., Fu, H., Chen, J., Jiang, Y. G., & Davis, L. S. (2021c). Videolt: Large-scale long-tailed video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7960–7969)

Zhou, B., Cui, Q., Wei, X. S., & Chen Z. M. (2020). Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9719–9728)

Zhou, Y., Sun, X., Zha, Z. J., & Zeng, W. (2018). Mict: Mixed 3d/2d convolutional tube for human action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 449–458)

Zhu, L., & Yang, Y. (2018). Compound memory networks for few-shot video classification. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 751–766)

Zhu, L., & Yang, Y. (2020a). Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8746–8755)

Zhu, L., & Yang, Y. (2020b). Inflated episodic memory with region self-attention for long-tailed visual recognition. *In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 4344–4353

Zhu, L., & Yang, Y. (2020). Label independent memory for semi-supervised few-shot video classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(1), 273–285.

Zhu, Y., Li, X., Liu, C., Zolfaghari ,M., Xiong, Y., Wu, C., Zhang, Z., Tighe, J., Manmatha, R., & Li, M. (2020). A comprehensive study of deep video action recognition. arXiv preprint arXiv:2012.06567

Zhu, Z., Wang, L., Guo, S., & Wu, G., (2021). A closer look at few-shot video classification: A new baseline and benchmark. arXiv preprint arXiv:2110.12358

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.