Verbs in Action: Improving verb understanding in video-language models

Liliane Momeni[1] Mathilde Caron[2] Arsha Nagrani[2] Andrew Zisserman[1] Cordelia Schmid[2]
[1] Visual Geometry Group, University of Oxford, UK
[2] Google Research

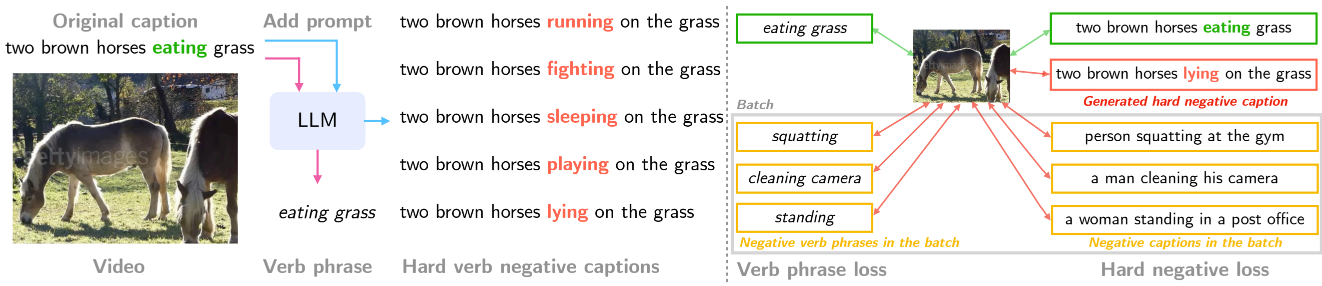

Figure 1. **Verb-Focused Contrastive (VFC) learning**: (Left): Given a video and its corresponding caption, we leverage a Large Language Model (LLM) to output (1) hard negative captions, where only the verb has been changed while keeping the remaining context, and (2) verb phrases which succinctly describe the action in the video. (Right): To encourage better verb reasoning, we subsequently enforce (1) a *calibrated* hard negative loss, using our generated hard negative captions and other captions in the batch, and (2) a fine-grained, verb phrase loss. We show that VFC improves verb understanding of video-language models compared to the standard contrastive loss.

Abstract

Understanding verbs is crucial to modelling how people and objects interact with each other and the environment through space and time. Recently, state-of-the-art video-language models based on CLIP have been shown to have limited verb understanding and to rely extensively on nouns, restricting their performance in real-world video applications that require action and temporal understanding. In this work, we improve verb understanding for CLIP-based video-language models by proposing a new Verb-Focused Contrastive (VFC) framework. This consists of two main components: (1) leveraging pretrained large language models (LLMs) to create hard negatives for cross-modal contrastive learning, together with a calibration strategy to balance the occurrence of concepts in positive and negative pairs; and (2) enforcing a fine-grained, verb phrase alignment loss. Our method achieves state-of-the-art results for zero-shot performance on three downstream tasks that focus on verb understanding, including video-text matching, video question-answering and video classification; while maintaining performance on noun-focused settings. To the best of our knowledge, this is the first work which proposes a method to alleviate the verb understanding problem, and does not simply highlight it. Our code is publicly available at [16] : scenic/projects/verbs_in_action.

1. Introduction

Large-scale visual-language models (VLMs) such as CLIP [58] have shown strong performance on multiple video-language tasks such as text-to-video retrieval [44], video question-answering, and open-set action recognition [42]. These models perform surprisingly well on these tasks in a zero-shot setting, despite being trained only on image-language pairs (with no access to temporal data), even outperforming strong video-specific models [5, 87].

A recently highlighted and well-documented problem with such models, however, is their strong *noun* or *object* bias, as evidenced by their lower performance in distinguishing between *verbs* in natural language descriptions [31, 53, 93]. This was first studied in images alone by the SVO-Probes benchmark [31], which shows that *image*-language models struggle to distinguish between different verbs, and often rely on the nouns instead. This problem persists with *video*-language models that inherit these VLMs, even after they are fine-tuned on video-text datasets [62, 85]. For example, Park *et al.* [53] similarly propose evaluation sets with hard verb negatives, and show that CLIP-based models, even when fine-tuned on video datasets, have difficulties discriminating verbs in a multi-choice setting where the context remains unchanged. Yuksekgonul *et al.* [93] further highlight limitations of vision-language models at understanding attribute, relationship, and order information. This deficiency in verb understanding limits the model's applicability for real-world tasks. Verbs encapsulate how people and objects interact with each other, and the environment, via actions in space and time.

We believe that there are two probable causes for this deficiency, even after fine-tuning on video-text data: (i) existing visual-text datasets have a strong bias towards single-frame concepts such as *objects* and *backgrounds* as well as *static* actions [9, 37, 67]. Models are hence less incentivized to understand dynamics and temporal actions [67], biasing them towards noun understanding; and (ii) the limitations of the cross-modal contrastive pretraining objective used by most current vision-language models [93]. In contrastive learning, the model is trained to distinguish correct video-caption pairs from incorrect ones. Since it is unlikely that existing datasets contain many examples with captions of *similar* context but *different* verbs, the task can be solved by taking little verb information into account. This relates to shortcut learning in deep neural networks [27].

In an attempt to mitigate this problem, we propose a novel training framework for tackling the task of verb understanding in vision-language models. Our framework, called **V**erb-**F**ocused **C**ontrastive pretraining (VFC), consists of two novel technical modifications to the contrastive learning framework. We first introduce a method to automatically generate negative sentences for training where only the verb has changed, keeping the context the same. This is done using LLMs [23, 59], in an automatic and scalable manner. Note that we *generate* hard negative captions, unlike works that simply mine hard negatives from an existing paired dataset [57], or change the order of words [93]. For example, given the caption '*two brown horses eating grass*', we generate the negative caption '*two brown horses running on the grass*' (see Fig. 1). While this improves performance on some downstream tasks, we find that introducing concepts simply in *negative* examples can also lead to an imbalance in the contrastive objective, favouring certain concepts in the feature space. To solve this, we propose a simple but effective *calibration strategy* to balance the occurrence of verbs in both positive and negative captions.

Secondly, inspired by recent works on *grounding* concepts in vision-language learning [10, 35], we also introduce a verb phrase loss that explicitly isolates the verb from a caption for more focused training. For example, we extract the verb phrase '*eating grass*' from the caption '*two brown horses eating grass*' (see Fig. 1). We find that this helps particularly for zero-shot performance on downstream tasks that do not use long sentences in their evaluation [28]. Verb phrases are also extracted from sentences using LLMs.

We then train a CLIP-based model [44] on a video-language dataset with this novel training framework. We show that a *single model* trained in this way transfers well to diverse downstream tasks that focus particularly on verb understanding, including three video benchmarks (multiple choice video-text matching on MSR-VTT [85], video question answering on Next-QA [82], action recognition on Kinetics [11]) and one image benchmark (SVO-probes [31]), achieving state-of-the-art performance compared to previous works in *zero-shot* settings (and often with fine-tuning as well); while maintaining performance on noun-focused settings. On Kinetics, we also introduce a verb split of the data which specifically highlights classes that are challenging to distinguish without fine-grained verb understanding ('*brushing hair*' vs '*curling hair*') and show that our model particularly improves performance on this split.

2. Related works

LLMs for video-text tasks. LLMs have been used for various vision applications, for example to initialise vision-text models [12, 45, 66]. Recent works further use frozen LLMs via prompting for tackling vision-language tasks [3, 24, 70, 76, 89, 91, 95]. LLMs have also been used in creative ways to obtain better supervision for training for various tasks [41, 64, 88, 94, 97]. For example, [88] use LLMs to generate question-answer pairs from transcribed video narrations, while [94] use LLMs to rephrase questions into sentences. [41] use LLMs to match noisy speech transcriptions to step descriptions of procedural activities. [51] train BERT [17] to predict action labels from transcribed speech segments and use this to scale up training data for action classification. [97] use pretrained LLMs conditioned on video to create automatic narrations. Recent works [64, 97] also show the benefits of using LLMs to paraphrase captions for data augmentation for video-language pretraining. [39] use LLMs to generate negative captions by manipulating event structures. Our work differs to [39] in that we focus specifically on verb negatives, and videos instead of images. Most closely related to our work, [53] construct a test set for verb understanding by leveraging T5 [59] and highlight the poor performance of current video-language models. Our work is substantially different: (i) we automatically construct hard negative captions for *training* (not testing), (ii) we compare the use of different LLMs, (iii) we show that training with such negative captions can improve verb understanding on various verb-focused benchmarks.

Hard negatives for contrastive pretraining. Hard negatives have been used to improve performance in metric representation learning and contrastive learning [30, 34, 80]. Recent works mine hard negatives from an existing paired dataset [57, 84, 90]. In comparison, in our work, we *generate* hard negative captions and propose a careful calibration mechanism for training effectively with such unpaired data. We also verify here the benefit of the HardNeg-NCE loss [57] when training with generated hard negative captions. [93] construct hard negative captions by shuffling words from the original caption to improve order and compositionality understanding. Our work differs by (i) focusing specifically on *verb* reasoning, as opposed to object-attribute relationships, (ii) using LLMs to construct hard verb text negatives as opposed to perturbing the word order, (iii) focusing on *video*-language models.

Learning from parts-of-speech in video. Recent works use parts-of-speech (PoS) tags for video understanding [25, 28, 63, 79, 86]. [79] learn multi-label verb-only representations, while other works focus on learning adverb representations [21, 22]. [2] use verb-noun pairs for unsupervised learning with instructional videos, while [25] leverage such pairs to generate data augmentations in the feature space. Other works exploit PoS for fine-grained or hierarchical alignment between video and text [14, 96]. [78] learn a separate multi-modal embedding space for each PoS tag and then combine these embeddings for fine-grained action retrieval. [14] construct a hierarchical semantic graph and use graph reasoning for local-global alignments. Most closely related to our work, [90] use a PoS based token contrastive loss. Our work differs in that: (i) we apply a verb phrase contrastive loss, as opposed to separate verb and noun losses; (ii) we extract verb phrases using a LLM and show this performs better than PoS tagging with NLTK [8] (Tab. 5); (iii) we evaluate our methods on verb-focused downstream tasks. Similarly to [28], we find that training with verb phrase supervision helps for zero-shot performance on tasks with shorter sentences.

Temporal understanding in videos. A long term goal in computer vision is temporal understanding in videos [11, 18, 29, 65, 68, 81, 98]. However, current training and test datasets have a strong visual bias towards *objects* and *backgrounds* as well as *static* actions [32, 67], with some works [9, 37] demonstrating strong results with a *single* frame. Despite these challenges, many recent works in video-only self-supervised learning propose pretext tasks for improving temporal modelling [1, 6, 7, 15, 19, 36, 40, 47, 54, 56, 60, 72, 73, 77, 92]. Unlike these works that use only video, [10, 69] focus on fine-grained temporal video-text alignment via localization of text sub-tokens. [4] also leverage before/after relations in captions to create artifical training samples for video-text. Differently to these works (which create augmented video negatives or positives), we approach the problem of improving *verb understanding* in video-language models from the language side, by leverging the strong generalization capabilities of LLMs.

3. Method

Our goal is to adapt large-scale vision-language pretrained models (such as CLIP) to understand *verbs*. We aim to do this without requiring such models to be retrained from scratch, but by simply fine-tuning them on a video-language dataset. However, given the pitfalls with using the standard video-text contrastive setup [58] on existing video-language datasets, we propose a new framework which we call **V**erb-**F**ocused **C**ontrastive pretraining (VFC). It consists of two components, both using the power of LLMs: (i) a novel calibrated hard negative training method where we train with synthetic verb-focused hard negative captions, and (ii) an additional verb phrase loss where videos are contrasted against isolated verb phrases as opposed to the entire caption. Note that a 'verb phrase' can be a single verb or verb-noun pair depending on the caption (see Fig. 1).

3.1. Preliminaries

Large Language Models (LLMs) are generative text models with impressive capacities, in particular for few-shot or prompt-based learning [23]. In our work, we design prompts to instruct a LLM to (i) create verb-focused hard negative captions and (ii) isolate verb phrases from the captions of a dataset. LLMs allow scalability and generalisation, and as we show in the ablations (see Tab. 2 and Tab. 5), are preferable to manual or rule based methods (eg. NLTK [8]). In particular, we use PaLM [23], a state-of-art autoregressive model, throughout this paper. However, our framework is agnostic to this choice and other LLMs can be used instead (see Tab. 2).

Video-language contrastive pretraining works by learning to distinguish between aligned and non-aligned video-text pairs. Given a dataset of N pairs $\{(V_i, T_i)\}_{i\in N}$ with video V_i and caption text T_i, we extract normalised feature representations v_i and t_i by using a video encoder f and text encoder g: we have $v_i = f(V_i)$ and $t_i = g(T_i)$. We use the InfoNCE loss [71] to make aligned ('positive') pairs close in feature space and all other pairwise combinations in the batch further apart [58]. We optimize for video-to-text L^{v2t} and text-to-video L^{t2v} alignments:

$$L_i^{t2v} = -t_i^\top v_i/\sigma + \log \sum_{j=1}^{B} \exp(t_i^\top v_j/\sigma) \qquad (1)$$

where B is the batch size and σ a temperature parameter controlling the sharpness of the distribution. L^{v2t} is obtained by inverting v and t in Eq. 1.

Architecture: adapting image-text models to videos. We leverage CLIP [58] for video-language tasks following the CLIP4CLIP 'seqTrans' protocol [44]. Both single-modal encoders (video f and text g) are initialized with CLIP weights, with four additional temporal frame aggregation transformer blocks stacked on top of the image encoder (see Sec. C.2 of the appendix for more details). Our approach is agnostic to model architecture and so any state-of-the-art video-language architecture could be potentially used.

3.2. Verb-Focused Contrastive Pretraining (VFC)

We describe both our calibrated hard negative training (Sec. 3.2.1) and the proposed verb phrase loss (Sec. 3.2.2).

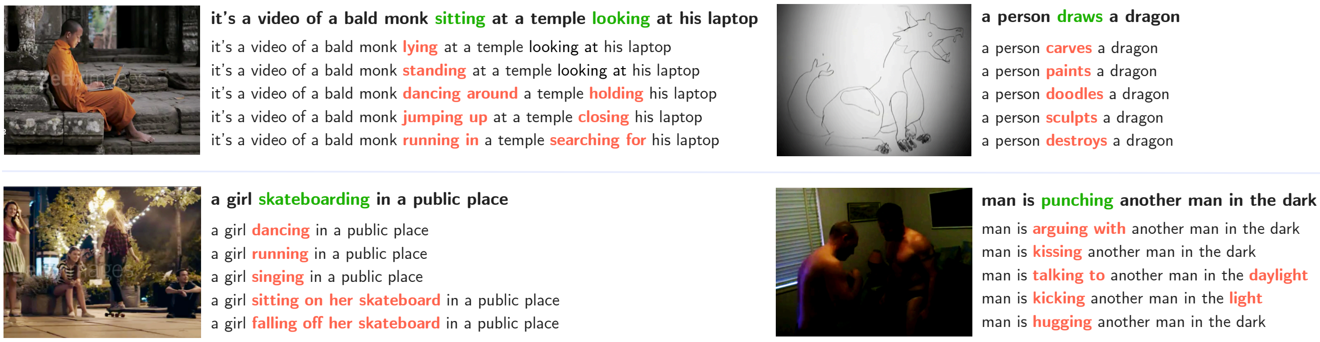

Figure 2. **Qualitative examples of hard negatives generated by PaLM.** We show a single frame per video and the corresponding caption in bold, with the verb highlighted in green. We see that PaLM can effectively generate hard negatives where the verb has changed (changes in red). When there are several verbs in the caption (see top left), PaLM may replace one or all verbs. As a failure case (bottom right), we show an example where PaLM can change more than just the verb, which could make it an easier negative (replacing 'punching' by 'talking' but also 'dark' by 'daylight').

3.2.1 Calibrated Hard Negative training

In regular contrastive learning, given a video-caption pair, other captions in the batch are simply pushed further in the feature space. Since it is unlikely that existing datasets contain many examples with captions of similar context but different *verbs*, the task can be solved by paying little attention to verbs. Instead, our goal is to encourage the video-language model to focus on verb reasoning. We do so by tasking a LLM to generate hard negative captions where only the verb(s) in the captions change. Second, we train with these additional negative captions. We find that naive training with additional data leads to imbalances affecting the resulting video-text feature space. We propose a simple but effective calibration mechanism to solve this.

Generating verb-focused hard negatives with PaLM. Given a caption T_i, we task PaLM to replace the verbs with other verbs that convey a different action, but still form a linguistically and semantically viable sentence (which may not be guaranteed with random verb replacements – see qualitative examples in Sec. B.4 of the appendix). For example, in the caption '*a man washes his face*', the verb '*washes*' should not be replaced with '*jumps*' or '*plays*'. The generated caption is then a negative match for the corresponding video V_i (albeit a *hard* negative, as the nouns and context remain the same). We experiment with different handcrafted prompts, and find our best performing prompt to be the following: *'In this task, you are given an input sentence. Your job is to tell me 10 output sentences with a different meaning by only changing the action verbs'*. We also add four input-output pair examples to the prompt, which increases the quality of PaLM's predictions (see Sec. A.3.2 of the appendix). We use one PaLM forward pass per caption T_i to generate ten verb-focused hard negatives for that caption (qualitative examples of the generated captions can be seen in Fig. 2). During training, we randomly sample N^{hard} generated captions for each pair (V_i, T_i) in the mini-batch, which we denote $\left(T_{i_k}^{\text{hard}}\right)_{k \in [1, N^{\text{hard}}]}$. Importantly, note that a $T_{i_k}^{\text{hard}}$ is a new generated text caption, or an *unpaired* data sample, meaning that it does not come with a corresponding matching ('positive') video.

Calibration. Interestingly, we observe that naively adding in negative captions into training with a contrastive loss leads to harmful feature space distortions, as some concepts are only seen in negative captions but never in positives. This is observed by careful analysis of downstream performance (see study in Tab. 3 and Tab. 4). We hence next describe a calibration mechanism to avoid such distortions: we first denote the vocabulary of all verb phrases in the original and generated captions as Ω. For each verb phrase ω (or 'concept') in Ω, we use S_ω to represent the number of times it appears in the captions of the original dataset and G_ω for the number of times it appears in the PaLM-generated captions. We then derive equations for R_ω (see Tab. 1), which we define as the ratio of the number of times a verb phrase ω is used as a negative versus as a positive during training, for different choices of the video-to-text contrastive loss (note L^{t2v} is unchanged).

Contrastive training with paired data (Baseline). We first note that *the ratio R_ω is independent of the verb phrase ω* in regular contrastive learning (paired data only). It simply depends on the batch size B, as S_w is cancelled from both the numerator and denominator. This means that the number of times a concept is used as a positive versus negative sample is the same regardless of the considered verb phrase. This naturally balances training, and is a great property of the contrastive framework.

Adding generated unpaired negative captions (HN). However, when training with unpaired captions, this ratio is proportional to G_ω / S_ω and therefore becomes *dependent* on the considered verb phrase ω. This can have significant consequences for the video-text feature representations. The model can learn to either ignore or always predict

Name	Video-to-text alignment loss	R_ω
Baseline	$-v_i^\top t_i/\sigma + \log \sum_{j=1}^{B} \exp(v_i^\top t_j/\sigma)$	$\frac{(B-1)S_\omega}{S_\omega}$ $\perp\!\!\!\perp \omega$
HN	$-v_i^\top t_i/\sigma + \log \left(\sum_{j=1}^{B} \exp(v_i^\top t_j/\sigma) + \sum_{j=1}^{B}\sum_{k=1}^{N^{\text{hard}}} \exp(v_i^\top t_{jk}^{\text{hard}}/\sigma)\right)$	$\frac{(B-1)S_\omega+BG_\omega}{S_\omega}$ $\propto B\frac{G_\omega}{S_\omega}$
Calibrated HN	$-v_i^\top t_i/\sigma + \log \left(\sum_{j=1}^{B} \exp(v_i^\top t_j/\sigma) + \sum_{k=1}^{N^{\text{hard}}} \exp(v_i^\top t_{i_k}^{\text{hard}}/\sigma)\right)$	$\frac{(B-1)S_\omega+G_\omega}{S_\omega}$ $\propto \frac{G_\omega}{S_\omega}$ with $G_\omega \approx S_\omega$

Table 1. **Different choices for video-to-text alignment** when training with additional hard negatives (HN). R_ω is the ratio of the number of times a given verb phrase ω is used as a negative versus the number of times it is used as a positive. We note that for the regular contrastive loss (**Baseline**), R_ω only depends on the batch size B, however when training with generated hard negatives (**HN**), it depends on the verb phrase ω. We minimise this effect using our proposed **Calibrated HN** loss, which we denote as L_i^{CHN}. See details in Section 3.2.1.

some concepts based on the average concept occurrences in positive or negative pairs during training.

Hard negatives with calibration (Calibrated HN). In order to make R_ω as ω-agnostic as possible, we introduce an ensemble of two techniques which we refer to as 'calibration'. First, we ignore the hard negative captions from the other elements of the batch (see row 3 in Tab. 1), which allows us to mitigate the influence of G_ω/S_ω by not amplifying it by the batch size B (equal to 256). Second, we filter the generated PaLM captions to have $G_\omega \approx S_\omega$. In practice, we discard some generations so that the number of times a verb phrase appears in the set of kept generations is equal to the number of times it is originally present in the dataset. We denote our video-to-text loss (text-to-video is unchanged) as L_i^{CHN} for calibrated hard negative training.

Video mining. An alternative to avoid imbalances due to the addition of negative captions would be to avoid training with unpaired data at all, by mining a matching video $V_{i_k}^{\text{hard}}$ for each generated caption $T_{i_k}^{\text{hard}}$. We attempt this via CLIP-based text-to-video retrieval in a large video database but found that finding a video matching a detailed, long caption is challenging, as such a precise video may not exist in a given corpus (see Sec. A.3.1 in the appendix for examples).

3.2.2 The verb phrase loss

In order to further encourage our model to focus on verbs, we introduce a contrastive 'verb phrase' loss. We use PaLM to extract the verb phrase T_i^{verb} in a caption T_i with the following prompt: '*In this task, you are given an input sentence. Your job is to output the action verb phrases.*' While multiple parts-of-speech (PoS) tagging tools exist, we use a LLM for the following reasons: (i) we would like to isolate verb phrases, which may correspond to single verbs or verb-noun pairs depending on the caption, (ii) LLMs deal better with ambiguous cases (see qualitative examples in Sec. B.5 of the appendix). We show the benefits experimentally via an ablation in Tab. 5. During training, we minimize the following loss:

$$L_i^{\text{verb-phrase}} = -v_i^\top t_i^{\text{verb}}/\sigma + \log \sum_{j=1}^{B} \exp(v_i^\top t_j^{\text{verb}}/\sigma)$$

where the negative verb phrase representations t_j^{verb} simply come from other captions in the batch. Note that we do not require the calibration mechanism described in Section 3.2.1 since all verb phrases T_i^{verb} have a positive video match V_i (i.e. the video aligned with T_i).

Overall, our verb-focused contrastive (VFC) pretraining optimizes the sum of three objectives:

$$L^{\text{VFC}} = \frac{1}{B}\sum_{i=1}^{B}\left(\lambda_1 L_i^{t2v} + \lambda_2 L_i^{\text{CHN}} + \lambda_3 L_i^{\text{verb-phrase}}\right)$$

with parameters λ_1, λ_2 and λ_3 weighting the contribution of the different terms. We learn the parameters of f and g via back-propagation.

3.3. Implementation details

Spoken Moments in Time (SMiT) pretraining dataset. The SMiT [49] training set consists of 481K pairs of 3 seconds video clips with corresponding captions. It is a subset of Moments in Time (MiT) [48]. Our work falls under the umbrella of transfer learning: we pretrain on SMiT and then use the resulting features to solve different downstream tasks in a zero-shot or fine-tuned manner. Pretraining is either done as in regular contrastive learning ('baseline') or with our VFC framework. We find that the baseline already performs competitively on our benchmarks, despite the relatively small size of SMiT compared to other datasets such as HowTo100M [46], due to the quality and diversity of the manually annotated captions. We encourage the community to consider SMiT as a powerful pretraining dataset.

PaLM. We use PaLM-540B [23] with beam size 4, output sequence length 512, and temperature of 0.7. The negative captions are generated in an autogressive way and are therefore of arbitrary length. We post-process them by removing text after any newline character and by filtering out candidates which contain the same verbs as the original caption.

Training details. Most hyper-parameters follow CLIP4CLIP [44]. We initialise our model with CLIP ViT/B-32 and train with VFC for 100 epochs with a batch size of 256, base learning rate of 1e-7, weight decay of 1e-2, temperature of 5e-3 and weights $\lambda_1 = 2$, $\lambda_2 = \lambda_3 = 1$ which we empirically find to work well in our experiments. Indeed, this balances the video-to-text and text-to-video loss terms. We also normalise each loss term by its value obtained from a random uniform prediction in order to have all loss terms in the same range (loss always equal to 1 for a random uniform prediction). We sample 32 frames

Method	Hard negatives	Verb$_H$	K-400
Baseline	$\varnothing$	69.9	55.6
w/o LLM			
	Random verb	73.6 (+3.7)	55.0 (-0.6)
	Antonym verb	72.4 (+2.5)	55.4 (-0.2)
w/ LLM			
	T5 [59]	75.1 (+5.2)	55.8 (+0.2)
Ours	PaLM [23]	78.0 (+8.1)	55.8 (+0.2)

Table 2. **Hard negatives generation.** We explore both LLM based and non LLM-based methods to obtain hard negative captions. Although PaLM LLM captions achieve the best performance, other LLMs (T5) achieve good results too. All methods are evaluated with calibration.

per video at 25fps, with a 2 frame stride. See Sec. C in the appendix for further implementation details and extensive evaluation protocols.

4. Experiments

We curate a suite of benchmarks from existing works to evaluate verb understanding which we present in Sec. 4.1. Then we ablate various components of our VFC framework in Sec. 4.2. Finally, we demonstrate improved performance on our diverse set of downstream tasks in Sec. 4.3, and compare to the state of the art.

4.1. Verb-Focused Benchmarks

MSR-VTT multiple choice (MC) is a benchmark of 10K videos of length 10–30 secs. We evaluate on the standard 3k split and on Verb$_H$ from [53]. In this setting, the task is to associate each video to the right caption among five choices. While the four wrong captions are randomly chosen from other videos in the standard 3k split, one of them is replaced by a *hard verb negative* in Verb$_H$ [53].

Video question answering on NEXT-QA The train (resp. val) split contains 3870 (resp. 570) videos with 32K (resp. 5k) questions. There are three types of questions: causal (C), temporal (T) and descriptive (D). We consider the standard setting as well as ATP$_{hard}$ [9], a subset automatically constructed with questions that are non-trivially solved with a single frame. ATP$_{hard}$ is designed to be a better benchmark for the model's true causal and temporal understanding which we believe is strongly related to verb reasoning.

Kinetics-400 is a video classification dataset with 400 human action classes. We report top-1, top-5 and their average classification accuracy. We follow [58] to evaluate classification in an open-set, zero-shot manner. This benchmark allows to assess transfer ability to *action* classification, which requires strong verb understanding (given actions are usually described with verb phrases).

SVO-probes dataset is a benchmark specifically designed to measure progress in verb understanding of image-text models [31]. It contains image–caption pairs with 421 different verbs. We simply replicate the image multiple times as input to our video model. We report Average Precision (AP) on the entire dataset as well as the verb-focused setting (details about our evaluation protocol are provided in Sec. C.4 of the appendix).

Method	R_ω	# HN	Verb$_H$	K-400
Baseline	$\perp\!\!\!\perp\ \omega$	0	69.9	55.6
w/o calibration	$\propto B\frac{G_\omega}{S_\omega}$	8.7M	80.5 (+10.6)	54.5 (-1.1)
w/ calibration	$\propto \frac{G_\omega}{S_\omega}, G_\omega \approx S_\omega$	0.9M	78.0 (+ 8.1)	55.8 (+0.2)

Table 3. **Importance of the calibration mechanism when training with hard negative captions.** The model trained without calibration suffers from a drop of performance on Kinetics.

	w/o calibration					w/ calibration				
	braiding hair	brushing hair	curling hair	dying hair	fixing hair	braiding hair	brushing hair	curling hair	dying hair	fixing hair
$R_\omega \propto$	37	12	78	53	27	1	1	1	1	1
braiding hair	38	14	1	1	5	47	2	5	2	7
brushing hair	2	51	0	0	2	4	41	3	1	7
curling hair	1	33	31	0	5	4	7	55	3	3
dying hair	0	12	1	36	4	1	1	4	45	6
fixing hair	6	23	1	1	9	6	15	7	3	12

Table 4. **Confusion matrix for the hair classes on Kinetics.** Without proper calibration, the verb phrase 'brushing hair' becomes highly attractive in the video-text feature space. This deteriorates the performance on all the 'hair' related classes. Our calibration mechanism alleviates this issue by making the ratio R_ω independent of verb phrases (see details in Sec. 3.2.1). More examples are shown in Sec. B.3 of the appendix.

4.2. Ablation Study

In this section, we analyze our different design choices. We report results when transferring the models on two of our benchmarks: MSR-VTT multi-choice verb split ('Verb$_H$') and Kinetics-400 video classification ('K-400'). We chose these two benchmarks as they have very different properties: the first involves captions, while the second involves action labels. We note that $N^{\text{hard}} = 1$ for all ablations unless otherwise specified.

Hard negative captions generation. In Tab. 2, we ablate the technique used to obtain additional negative captions: we compare two LLMs (T5 [59] and PaLM [23]) and two non LLM-based methods: (i) 'random verb': we replace verbs by random verbs from the UPenn XTag[1] verb corpus and (ii) 'antonym verb': we replace verbs with their antonyms, using the NLTK [8] package. We see in Tab. 2

[1] https://www.cis.upenn.edu/~xtag/

PaLM captions:	Verb$_H$	K-400
∅	69.9	55.6
Positive	69.3	55.4
Negative	**78.0**	**55.8**

Verb isolation:	Verb$_H$	K-400
∅	69.9	55.6
MiT labels	69.9	57.0
NLTK [8]	70.1	56.4
PaLM [23]	**70.3**	**57.6**

Table 5. (left): **Generating negative *versus* positive captions with PaLM.** (right): **Verb phrase isolation methods.**

Method	Hard negatives	Verb phrase	Verb$_H$	K-400
Baseline			69.9	55.6
	✓		78.0 (+8.1)	55.8 (+0.2)
		✓	70.3 (+0.4)	57.6 (+2.0)
VFC (Ours)	✓	✓	76.3 (+6.4)	58.5 (+2.9)

Table 6. **Combining hard negative and verb phrase loss** achieves 9.2% relative improvement on MSR-VTT MC (accuracy) and 5.2% relative improvement on Kinetics (top-1) compared to the baseline.

that 'random verb' and 'antonym verb' already give moderate performance gains on Verb$_H$ compared to the baseline. However, using LLM-based generations improves the results by a large margin compared to the non LLM-based methods. This is likely due to the fact that (i) random or antonym replacements often create non semantically or linguistically plausible negative captions; (ii) some verbs do not have antonyms in NLTK (see qualitative examples in Sec. B.4 of the appendix). Finally, we see in Tab. 2 that T5 generations work very well in our framework too, which demonstrates that our framework is LLM-agnostic and can be extended to other LLMs. We observe that the best performance is achieved using PaLM, with a substantial gain over the baseline on MSR multi-choice (+8.1%) and a moderate gain on Kinetics (+0.2%).

Hard negative captions: the importance of calibration. We demonstrate the effect of the calibration mechanism described in Section 3.2.1 for training with unpaired captions. Tab. 3 shows the performance of hard negative training with ('w/') *versus* without ('w/o') calibration. First, we observe that the performance boost on MSR-VTT compared to the baseline is slightly stronger without calibration than with calibration. We believe this is because calibrating the PaLM generations reduces their number. However, we see that training with hard negatives without calibration deteriorates a lot the performance on Kinetics (-2.0% compared to the baseline). We hypothesize that this is due to some verb phrases being seen only as repulsive in the video-text feature space, while others are seen equally as attractive and repulsive. We illustrate this in Tab. 4 by showing the confusion matrix for a subset of the Kinetics classes, along with the ratio R_ω (defined in Sec. 3.2.1) for each verb phrase. Intuitively, R_ω measures the 'attraction' (if low) and 'repulsion' (if high) of a verb phrase ω. The confusion matrix in Tab. 4 shows that the verb phrase 'brushing hair' becomes an attraction point in the absence of calibration. Indeed, the number of times the verb phrase 'brushing hair' is repulsive versus attractive is low ($R_{\text{brushing hair}} \approx 12$) compared to the other concepts such as for example 'curling hair' ($R_{\text{curling hair}} \approx 78$): we have $R_{\text{brushing hair}} << R_{\text{curling hair}}$. Hence, predictions for 'brushing hair' become dominant. This actually improves the performance for that class but deteriorates the performance on all the other classes related to 'hair'. We see in Tab. 4 that our calibration mechanism alleviates this effect by making the ratio R_ω independent of ω as in regular contrastive learning. Calibration allows us to improve performance over the baselines on both tasks with a single model.

Generating positive *versus* negative captions. In Tab. 5 (left), we investigate the impact of generating *positive* captions instead of negatives with PaLM. In this case, positives correspond to sentences where the verb in the original caption is changed to a synonym verb, but the remaining context is unchanged: PaLM therefore acts as a data augmentation generator for text (similar to [64, 97]). Details about the positive caption generation implementation are in Sec. C.5 of the appendix. We observe that using positive captions has a negative impact on the performance in our benchmarks, possibly because with positive captions the model becomes more *invariant* to different verbs.

Verb phrase loss. In Tab. 5 (right), we explore two alternatives for verb phrase extraction used in the verb phrase loss: (i) using human-annotated action labels for clips from the Moments in Time (MiT) dataset (these are available as SMiT data inherits from MiT [48]) and (ii) using a rule-based method (NLTK [8]) to isolate verbs. We observe in Tab. 5 that using PaLM to extract verb phrases from the caption outperforms both, probably because it extracts more fine-grained action information. Qualitative analysis of the verb phrases is shown in Sec. B.5 of the appendix.

Combining calibrated hard negatives and verb phrase loss. We show in Tab. 6 the complementarity between our two contributions: the calibrated hard negative training and the verb phrase loss. The former greatly improves performance on tasks requiring complex language understanding such as MC Verb$_H$. On the other hand, the verb phrase loss improves transfer to video classification by focusing particularly on the action label in the sentence. We see in Tab. 6 that combining both approaches during training results in a *single model* with excellent performance on both MSR-VTT MC and Kinetics zero-shot transfer. Indeed, compared to the baseline, VFC pretraining achieves 9.2% relative improvement on MSR-VTT MC and 5.2% relative improvement on Kinetics.

Number of hard negative captions. In Tab. 7, we experiment with increasing the maximum number of hard negative captions N^{hard} sampled per video in the batch. We find that setting this to 5 increases the performance on Verb$_H$ while

Method	N^{hard}	Verb$_H$	K-400
VFC (Ours)	1	76.3	58.5
VFC (Ours)	3	77.8	58.5
VFC (Ours)	5	**78.3**	**58.5**

Table 7. **Maximum number of hard negative captions.** We observe that increasing the maximum number of hard negative captions sampled per video increases the performance on Verb$_H$. We use $N^{\text{hard}} = 5$ in the remaining of the paper.

Method	Contrastive loss	Verb$_H$	K-400
Baseline	NCE	69.9	55.6
Baseline	HardNeg-NCE	72.0	56.4
VFC (Ours)	NCE	78.3	58.5
VFC (Ours)	HardNeg-NCE	**80.5**	**58.8**

Table 8. **Complementarity with other hard negative mining methods.** We observe that using the HardNeg-NCE loss, instead of standard NCE, gives the highest performance. We use HardNeg-NCE from now on. We note that for VFC we use $N^{\text{hard}} = 5$.

maintaining the performance on Kinetics. We use this setting going forward. We note that we do not try larger values as our maximum number of hard negatives per video after calibration is 5.

Complementarity with other hard negative mining methods. We investigate whether our VFC framework is complementary to existing approaches for hard negatives with the contrastive learning framework. Specifically, we reimplement the hard negative noise contrastive multimodal alignment loss from [57, 61], which is denoted as HardNeg-NCE. With this objective, difficult negative pairs (with higher similarity) are emphasised, and easier pairs are ignored. We use $\alpha = 1$ and $\beta = 0.1$ in the equations from [57]. We note that we only adapt L_i^{t2v} and L_i^{CHN} with HardNeg-NCE. Adapting $L_i^{\text{verb-phrase}}$ does not bring further improvements, so we omit this for simplicity. We observe in Tab. 8 that VFC is complementary to existing hard negative frameworks: using HardNeg-NCE instead of the standard NCE loss achieves the highest performance. We observe a large boost on Verb$_H$ [53], a benchmark that specifically involves hard negatives. We therefore adopt HardNeg-NCE in the remaining of this paper.

4.3. Comparisons to the State of the Art

We compare our VFC features to the state of the art on a diverse set of tasks requiring verb understanding. Note that we use the *same model* across different tasks, which is nontrivial in itself as the tasks cover a wide range of domains and evaluation protocols.

MSR-VTT MC results. We see in Tab. 9 that our verb-focused pretraining transfers well to the MSR-VTT multi-choice task, especially on the hard verb split (curated to assess exactly the task we are trying to solve). We even outperform concurrent InternVideo [75] while using a significantly smaller setting both in terms of architecture (InternVideo uses 2.8× more parameters and 12.4× more flops) and pretraining dataset size (they use 24× more data). We also note that our method does not degrade performance on other standard object-based tasks, such as text-to-video retrieval on MSR-VTT (results compared to the state of the art are shown in Sec. A.2 of the appendix).

NEXT-QA results. We show in Tab. 10 that our verb-focused pretraining gives a significant boost in both the standard and ATP$_{hard}$ setting introduced by [9]. We highlight the improved performance for the descriptive (and therefore more noun-focused) setting. To the best of our knowledge, we are the first work to report zero-shot results

Model	# params.	3k val.	Verb$_H$ [53]
ZERO-SHOT			
VideoCLIP [84]	–	73.9	-
CLIP [58]	151M	91.1	64.1
InternVideo [75]	≈ 460M	93.4	-
VFC (Ours)	164M	**95.1**	**80.5**
FINE-TUNED			
ClipBERT [38]	–	88.2	-
MMT [26]	–	92.4	71.3
VideoCLIP [84]	–	92.1	-
CLIP-straight [55]	151M	94.1	65.1
MMT [26] (CLIP features)	–	95.0	71.4
C4CL-mP [53]	151M	**96.2**	73.7
VFC (Ours)	164M	**96.2**	**85.2**

Table 9. **Multi-choice MSR-VTT.** We report accuracy on the 3k val and on the verb-focused Verb$_H$ [53] splits. While VFC improves the performance on both splits in a zero-shot setting, the gap with previous works is especially important on Verb$_H$ [53], a split measuring verb understanding. When available, we add model parameter counts.

Model	all	D	T	C	ATP$_{hard}$ [9] all	T	C
ZERO-SHOT							
CLIP [58]	43.9	57.0	38.1	43.6	23.0	21.8	23.8
VFC (Ours)	**51.5**	**64.1**	**45.4**	**51.6**	**31.4**	**30.0**	**32.2**
FINE-TUNED							
HGA‡ [33]	49.7	59.3	50.7	46.3	44.1	45.3	43.3
ATP [9]	49.2	58.9	46.7	48.3	20.8	22.6	19.6
Temp[ATP] [9]	51.5	65.0	49.3	48.6	37.6	36.5	38.4
TAATP† [83]	54.3	66.8	50.2	53.1	-	-	-
VGT [83]	55.0	64.1	**55.1**	52.3	-	-	-
VFC (Ours)	**58.6**	**72.8**	53.3	**57.6**	**39.3**	**38.3**	**39.9**

Table 10. **NEXT-QA video question answering.** We report accuracy. We consider either 'all' questions or only causal ('C'), temporal ('T') or descriptive ('D') questions. We also use ATP$_{hard}$ split [9]. VFC improves performance for both zero-shot and fine-tuning. †Temp[ATP]+ATP. ‡ Uses additional motion features.

Model	# param.	top-1	top-5	average
VAL-SET				
CLIP [58]	151M	48.9	75.8	62.4
ActionCLIP [74]	$\approx$ 164M	56.4	-	-
VFC (Ours)	164M	**59.4**	**85.3**	**72.4**
TEST-SET				
Flamingo-3B [3]	3B	45.2	66.8	56.0
Flamingo-80B [3]	80B	49.1	71.5	60.3
Flamingo-9B [3]	9B	49.7	71.5	60.6
CLIP [58]	151M	47.9	75.1	61.5
VFC (Ours)	164M	**58.8**	**84.5**	**71.7**

Table 11. **Zero-shot transfer to Kinetics-400.** We report top-1 accuracy, top-5 accuracy, and their average on the validation and test set, as well as the parameter counts of the different models.

Model	top-1	top-5
ZERO-SHOT		
CLIP [58]	59.7	83.9
VFC (Ours)	**70.2**	**92.5**
FINE-TUNED		
ER-ZSAR [13]	42.1	73.1
X-CLIP [52]	65.2	86.1
X-Florence [52]	68.8	88.4

Table 12. **Zero-shot transfer to Kinetics-600.** We report average top-1 and top-5 accuracies over three random 160-class splits, covering classes not in Kinetics-400 but within Kinetics-600. While [13, 52] fine-tune on Kinetics-400, we surpass their performance in zero-shot.

for NEXT-QA and our zero-shot numbers improve upon some previously published fine-tuning numbers. Finally, although HGA [33] performs worse than ours on the standard setting, it achieves a high accuracy of 44.1 on ATP$_{hard}$. Their high performance on ATP$_{hard}$ can be explained by the use of additional motion features, aiding in answering hard dynamics questions, as noted by [9]. The addition of extra motion features on the video side can be complementary to our verb-focused pretraining approach.

Zero-shot Kinetics-400 results. In Tab. 11 we see that our verb-focused features transfer very well to Kinetics video classification benchmark in a zero-shot setting, achieving state-of-the-art results. We achieve better results than Flamingo models [3] while using a significantly smaller model: relative improvement of 20% over Flamingo-80B model while using 489 $\times$ less parameters.

Zero-shot Kinetics-600 results. We evaluate our model on Kinetics-600 in Tab. 12 and follow the protocol in [13, 52]. Specifically, the subset of categories which are outside Kinetics-400, but within Kinetics-600 are used for evaluation. The evaluation is then run on a random sample of 160 categories from this subset. The final performance is averaged over three iterations. We observe that by evaluating our model in a zero-shot setting, we surpass the performance of works [13, 52] which fine-tune on Kinetics-400.

Method	all	Kinetics-verb
Baseline	55.6	52.1
VFC (Ours)	**58.8** (+3.2)	**57.1** (+5.0)

Table 13. **Zero-shot Kinetics-verb.** We report accuracy performance on our newly proposed Kinetics-verb split (from test split).

Model	AP	AP$_{\text{verb}}$
CLIP [58]	48.3	52.3
No-MRM-MMT [31]†	51.5	53.1
Baseline (Ours)	60.2	61.9
VFC (Ours)	**61.8**	**64.6**

Table 14. **Verb understanding on SVO-probes [31].** We report Average Precision (AP) on the entire dataset and on the verb setting. † Scores provided by authors and used to calculate AP.

Kinetics-verb. To further analyse the VFC framework's effect on action classification, we introduce the Kinetics-verb split. We isolate classes from the Kinetics-400 dataset that share a common noun with another class, but have a different verb (and therefore action). For example, distinguising between 'braiding hair', 'brushing hair' and 'curling hair' requires the model to focus on verb understanding as predictions cannot be inferred from the simple presence of hair in the frame. We use this rule to create a subset of 97 classes from the Kinetics-400 test set (see Sec. C.7 in the appendix) called 'Kinetics-verb'. We show in Tab. 13 that our VFC improves substantially over the baseline (+5%) on this split.

Assessing verb understanding on SVO-probes. In Tab. 14, we see that our VFC framework improves the performance on SVO-probes compared to the baseline (particularly in the verb setting), and outperforms prior work [31] with 21.7% relative improvement in the verb setting.

5. Conclusion

Video-language models based on CLIP have been shown to have limited verb understanding, relying extensively on nouns. We attempt to alleviate this problem with two technical contributions on the contrastive learning framework: first, we leverage LLMs to automatically generate hard negative captions focused on verbs; second, we introduce a verb phrase alignment loss. We validate our verb-focused pretraining by showing improved performance on a suite of benchmarks, chosen in particular to assess verb understanding. Our framework is general and could be employed for other video-language tasks, and further readily scales with the rapid progress in language modelling.

Acknowledgements. We would like to thank Ahmet Iscen, Anurag Arnab, Paul Hongsuck Seo, Antoine Yang, Shyamal Buch, Alex Salcianu for their precious help and discussions. We also thank Sagar Vaze for his invaluable support.

References

[1] Unaiza Ahsan, Rishi Madhok, and Irfan Essa. Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition. In *WACV*, 2019. 3

[2] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *CVPR*, 2016. 3

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Neurips*, 2022. 2, 9, 8

[4] Piyush Bagad, Makarand Tapaswi, and Cees G. M. Snoek. Test of time: Instilling video-language models with a sense of time. *arXiv preprint arXiv:2301.02074*, 2023. 3

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 1

[6] Nadine Behrmann, Mohsen Fayyaz, Juergen Gall, and Mehdi Noroozi. Long short view feature decomposition via contrastive video representation learning. In *ICCV*, 2021. 3

[7] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T. Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *CVPR*, 2020. 3

[8] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009. 3, 6, 7, 8, 13

[9] Shyamal Buch, Cristobal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "Video" in Video-Language Understanding. In *CVPR*, 2022. 2, 3, 6, 8, 9, 1

[10] Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. Locvtp: Video-text pre-training for temporal localization. In *ECCV*, 2022. 2, 3

[11] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2, 3, 13

[12] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *CVPR*, 2022. 2

[13] Shizhe Chen and Dong Huang. Elaborative rehearsal for zero-shot action recognition. In *ICCV*, 2021. 9

[14] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *CVPR*, 2020. 3

[15] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, 2022. 3

[16] Mostafa Dehghani, Alexey Gritsenko, Anurag Arnab, Matthias Minderer, and Yi Tay. Scenic: A jax library for computer vision research and beyond. In *CVPR*, 2022. 1

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, 2019. 2, 8, 13

[18] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jurgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large scale holistic video understanding. In *ECCV*, 2019. 3

[19] Michael Dorkenwald, Fanyi Xiao, Biagio Brattoli, Joseph Tighe, and Davide Modolo. Scvrl: Shuffled contrastive video representation learning. In *CVPRW*, 2022. 3

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 9

[21] Hazel Doughty, Ivan Laptev, Walterio Mayol-Cuevas, and Dima Damen. Action modifiers: Learning from adverbs in instructional videos. In *CVPR*, 2019. 3

[22] Hazel Doughty and Cees G. M. Snoek. How Do You Do It? Fine-Grained Action Understanding with Pseudo-Adverbs. In *CVPR*, 2022. 3

[23] Aakanksha Chowdhery et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 2, 3, 5, 6, 7, 13

[24] Zhenfang Chen et al. See, think, confirm: Interactive prompting between vision and language models for knowledge-based visual reasoning. *arXiv preprint arXiv:2301.05226*, 2023. 2

[25] Alex Falcon, Giuseppe Serra, and Oswald Lanz. A feature-space multimodal data augmentation technique for text-video retrieval. In *ACM*, 2022. 3

[26] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal Transformer for Video Retrieval. In *ECCV*, 2020. 8, 2

[27] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020. 2

[28] Deepti Ghadiyaram, Matt Feiszli, Du Tran, Xueting Yan, Heng Wang, and Dhruv Kumar Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *CVPR*, 2019. 2, 3

[29] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter N. Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 3

[30] Ben Harwood, Vijay Kumar B.G., Gustavo Carneiro, Ian Reid, and Tom Drummond. Smart mining for deep metric learning. In *ICCV*, 2017. 2

[31] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *ACL*, 2021. 1, 2, 6, 9, 8, 10

[32] De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and

Juan Carlos Niebles. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *CVPR*, 2018. 3
[33] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *AAAI*, 2020. 8, 9
[34] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *Neurips*, 2020. 2
[35] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr–modulated detection for end-to-end multi-modal understanding. *arXiv preprint arXiv:2104.12763*, 2021. 2
[36] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *AAAI*, 2019. 3
[37] Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022. 2, 3
[38] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learningvia sparse sampling. In *CVPR*, 2021. 8, 2
[39] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In *CVPR*, 2022. 2
[40] Hanwen Liang, Niamul Quader, Zhixiang Chi, Lizhe Chen, Peng Dai, Juwei Lu, and Yang Wang. Self-supervised spatiotemporal representation learning by exploiting video continuity. In *AAAI*, 2021. 3
[41] Xudong Lin, Fabio Petroni, Gedas Bertasius, Marcus Rohrbach, Shih-Fu Chang, and Lorenzo Torresani. Learning to recognize procedural activities with distant supervision. In *CVPR*, 2022. 2
[42] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *ECCV*, 2022. 1
[43] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Neurips*, 2019. 8
[44] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 1, 2, 3, 5, 7, 8, 9
[45] Ziyang Luo, Yadong Xi, Rongsheng Zhang, and Jing Ma. A frustratingly simple approach for end-to-end image captioning. *arXiv preprint arXiv:2201.12723*, 2022. 2
[46] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 5
[47] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and learn: Unsupervised learning using temporal order verification. In *ECCV*, 2016. 3
[48] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *TPAMI*, 2019. 5, 7, 6
[49] Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. Spoken moments: Learning joint audio-visual representations from video descriptions. In *CVPR*, 2021. 5
[50] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Manen Santiago, Sun Chen, and Cordelia Schmid. Learning audio video modalities from image captions. In *ECCV*, 2022. 2
[51] Arsha Nagrani, Chen Sun, David Ross, Rahul Sukthankar, Cordelia Schmid, and Andrew Zisserman. Speech2action: Cross-modal supervision for action recognition. In *CVPR*, 2020. 2
[52] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *ECCV*, 2022. 9
[53] Jae Sung Park, Sheng Shen, Ali Farhadi, Trevor Darrell, Yejin Choi, and Anna Rohrbach. Exposing the limits of video-text models through contrast sets. In *ACL*, 2022. 1, 2, 6, 8, 3, 4, 5, 10, 13
[54] Lyndsey C. Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard Scholkopf, and William T. Freeman. Seeing the arrow of time. In *CVPR*, 2014. 3
[55] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. A straightforward framework for video retrieval using CLIP. *arXiv preprint arXiv:2102.12443*, 2021. 8, 2
[56] Will Price and Dima Damen. Retro-actions: Learning 'close' by time-reversing 'open' videos. In *ICCVW*, 2019. 3
[57] Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training. *arXiv preprint arXiv:2301.02280*, 2023. 2, 8
[58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1, 3, 6, 8, 9, 2, 10
[59] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020. 2, 6, 13
[60] Adria Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Patraaucean, Florent Altché, Michal Valko, Jean-Bastien Grill, Aaron Oord, and Andrew Zisserman. Broaden your views for self-supervised video learning. *arXiv preprint arXiv:2021.00129*, 2021. 3
[61] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *ICLR*, 2021. 8
[62] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Chris Pal, Hugo Larochelle, Aaron Courville, and

Bernt Schiele. Movie description. *IJCV*, 2017. 1
[63] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling for video understanding. In *CVPR*, 2021. 3
[64] Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. Is a caption worth a thousand images? a controlled study for representation learning. *arXiv preprint arXiv:2207.07635*, 2022. 2, 7
[65] Konrad Schindler and Luc van Gool. Action snippets: How many frames does human action recognition require? In *CVPR*, 2008. 3
[66] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *CVPR*, 2022. 2
[67] Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. Only time can tell: Discovering temporal data for temporal modeling. In *WACV*, 2021. 2, 3
[68] Gunnar Sigurdsson, Olga Russakovsky, and Abhinav Gupta. What actions are needed for understanding human actions in videos? In *ICCV*, 2017. 3
[69] Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. Long-form video-language pretraining with multimodal temporal contrastive learning. In *Neurips*, 2022. 3
[70] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *Neurips*, 2021. 2
[71] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3
[72] Jue Wang, Gedas Bertasius, Du Tran, and Lorenzo Torresani. Long-short temporal contrastive learning of video transformers. In *CVPR*, 2022. 3
[73] Jiangliu Wang, Jianbo Jiao, and Yunhui Liu. Self-supervised video representation learning by pace prediction. In *ECCV*, 2020. 3
[74] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 9, 8
[75] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 8, 1, 2
[76] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. *arXiv preprint arXiv:2205.10747*, 2022. 2
[77] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *CVPR*, 2018. 3
[78] Michael Wray, G. Csurka, Diane Larlus, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *ICCV*, 2019. 3
[79] Michael Wray and Dima Damen. Learning visual actions using multiple verb-only labels. In *BMVC*, 2019. 3
[80] Chao-Yuan Wu, R. Manmatha, Alexander J. Smola, and Philipp Krähenbühl. Sampling matters in deep embedding learning. In *ICCV*, 2017. 2
[81] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krähenbühl, and Ross Girshick. Long-Term Feature Banks for Detailed Video Understanding. In *CVPR*, 2019. 3
[82] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021. 2
[83] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *ECCV*, 2022. 8
[84] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pretraining for zero-shot video-text understanding. In *EMNLP*, 2021. 2, 8
[85] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 1, 2
[86] Ran Xu, Caiming Xiong, Wei Chen, and Jason J. Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *AAAI*, 2015. 3
[87] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *CVPR*, 2022. 1
[88] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, 2021. 2
[89] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. In *Neurips*, 2022. 2
[90] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *ICCV*, 2021. 2, 3
[91] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *AAAI*, 2022. 2
[92] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. In *CVPR*, 2020. 3
[93] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*, 2023. 1, 2
[94] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Multimodal neural script knowledge through vision and language and sound. In *CVPR*, 2022. 2
[95] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022. 2
[96] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and

hierarchical modeling of video and text. In *ECCV*, 2018. 3

[97] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. *arXiv preprint arXiv:2212.04501*, 2022. 2, 7

[98] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, 2018. 3