ReGen: A good Generative zero-shot video classifier should be Rewarded

Adrian Bulat[1,2], Enrique Sanchez[1], Brais Martinez[1], Georgios Tzimiropoulos[1,3]
[1]Samsung AI Cambridge [2]Technical University of Iasi [3]Queen Mary University of London

Abstract

This paper sets out to solve the following problem: How can we turn a generative video captioning model into an open-world video/action classification model? Video captioning models can naturally produce open-ended free-form descriptions of a given video which, however, might not be discriminative enough for video/action recognition. Unfortunately, when fine-tuned to auto-regress the class names directly, video captioning models overfit the base classes losing their open-world zero-shot capabilities. To alleviate base class overfitting, in this work, we propose to use reinforcement learning to enforce the output of the video captioning model to be more class-level discriminative. Specifically, we propose ReGen, a novel reinforcement learning based framework with a three-fold objective and reward functions: (1) a class-level discrimination reward that enforces the generated caption to be correctly classified into the corresponding action class, (2) a CLIP reward that encourages the generated caption to continue to be descriptive of the input video (i.e. video-specific), and (3) a grammar reward that preserves the grammatical correctness of the caption. We show that ReGen can train a model to produce captions that are: discriminative, video-specific and grammatically correct. Importantly, when evaluated on standard benchmarks for zero- and few-shot action classification, ReGen significantly outperforms the previous state-of-the-art.

1. Introduction

Open-world or zero-shot video recognition is concerned with the problem of recognizing at test time, unseen during training, video/action classes. For example, during training, an open-world video recognition model might be trained to classify `dancing_blues` or `dancing_latin` but, during test time, it might be required to perform classification over new categories, not seen during training, like `dancing_tango`. Another application of open world-recognition is when a model is trained on dataset A with class taxonomy $\mathcal{C}_A$ and then it is applied directly (i.e. without re-training) on a different dataset B with a different (i.e. non- or partially-overlapping) class taxonomy $\mathcal{C}_B$.

A video captioning model uses (by construction) a generative language model, conditioned on features produced by a video backbone, to generate a human-interpretable free-form textual description of the video which, in principle, can be associated to any video/action class. Hence, a captioning model can be potentially used for open-world recognition. However, in practice, the generated captions are not discriminative enough for video/action recognition. In this paper, our goal is to turn a generative captioning model into a highly-accurate open-world video/action classification model which, at the same time, maintains its ability to generate video-specific grammatically correct captions.

To our knowledge, the only method that trains a generative captioning model for open-world video/action recognition is the recently proposed REST [8]. Therein, the authors first showed that directly fine-tuning a captioning model to auto-regress the action class names results in base class overfitting and severely hurts zero-shot generalizability. While REST addresses, to some extent, this problem by utilizing an unsupervised adaptation framework, it completely discards class information during training. As a result, the trained model might still not be very discriminative for open-world video classification.

Our main goal in this work is to address this important limitation of [8] by enabling the integration of class information into the training of a generative video captioning model. To this end, we propose *ReGen*, a newly introduced training framework based on *Re*inforcement Learning (RL) and 3 appropriate rewards to train a *Gen*erative model so that its output caption satisfies *3 key requirements*: (1) be class discriminative avoiding base overfitting, (2) maintain the video-specific granularity of the generated text, and (3) maintain the grammatical correctness of the generated text. Different to video captioning, ReGen does not use captions to train the model, only class label information. To this end, in this paper, **we make the following contributions**:

- To avoid base class overfitting, we avoid training with a standard language modelling loss and, instead, introduce RL and `CLS-R`, *a class discrimination reward computed from class names only*, that enforces the generated caption for a given video to be correctly classified into the corresponding ground truth class.
- As the optimization of `CLS-R` alone results in a model

whose output degenerates towards a generic class-specific caption, we introduce a CLIP-based reward, CLIP-R, that encourages the generated caption to continue to be descriptive of the video content.

- To ensure that the generated caption is grammatically-correct, we propose a grammar reward, GRAMMAR-R, that maintains the correctness of the produced caption.
- When evaluated on standard benchmarks for zero-shot and few-shot action classification, ReGen outperforms the previous state-of-the-art by a large margin. We also show very competitive results for zero-shot captioning.

2. Related work

Zero-shot Action Recognition: A number of methods aim to learn to align video representations with the word embeddings describing the action class names. The works of [62, 21, 21] learn to align features from frozen video networks to Word2Vec [38] class name embeddings. Other works propose to directly learn the video networks to produce features aligned with Word2Vec embeddings, representing the composition of scenes [7] or the class names [6]. In [37], an optimal transport assignment re-adjusts the prototype embeddings for the text classes according to the test video embeddings, creating a label space distribution at test time. The work of [16] proposes a method that separately models objects and their interactions. The work of [36] learns a Knowledge Graph to convert the word embeddings for objects/nouns into classification weights, and uses this to create the weights for unseen classes.

Recently, a number of discriminative methods based on contrastive learning have been proposed. X-CLIP [39] adapts CLIP with a video-specific cross-attention module producing an enhanced class embedding aligned with the video features. In [19] a transformer is proposed trained to predict nouns and actions from masked text by attending to the video features. The work of [30] proposes a joint encoder for the video and text embeddings and to express the embeddings of the unseen classes as a weighted combination of the seen ones. An inherent problem with discriminative methods is that they are trained in a fully supervised manner which is prone to base class overfitting.

As an alternative to discriminative methods, REST [8] introduces a new class of zero-shot models based on generative video captioning where the goal is to use the generated caption for video/action classification. To train such a model, REST proposes an unsupervised adaptation framework, based on retrieval-augmented self-training with pseudo-captions, which avoids base class overfitting. However, REST completely discards class label information which might be available during training.

ReGen combines the best of both worlds (discriminative and generative): it builds upon REST inheriting the advantages of a generative approach (e.g. fine-grained & human interpretable output, less base class overfitting) but also, through the proposed rewards for reinforcement learning, it enables the integration of class information for training of a significantly more discriminative (compared to REST) model for open-world video classification.

RL for image captioning has been primarily motivated by the mismatch between the cross entropy loss used to train an autoregressive model and the captioning metrics used for evaluation, and the gap induced by using ground-truth tokens at training time to predict the next word while the model uses its own predictions at test time. *Two fundamental differences with our work* are that (1) the goal of RL-based image captioning is to improve the quality of the generated caption while, in our work, we used RL to improve the discriminative properties of the caption for open-world video/action recognition, and (2) ReGen does not use captions for training but only the base classes' names.

MIXER [45] pioneered the use of REINFORCE [56] to learn a generative model in an annealing strategy whereby the influence of the rewards gradually overtakes that of the cross-entropy loss . Follow-up works proposed different options for the critic [59, 32] where the rewards were defined at a sentence-level instead of a word-level. The work of [46] proposes the self-critical sequence training as a form of REINFORCE that uses its own test-time inference algorithm to normalize the rewards. The work of [12] uses the transformer [50] and formulates the objective in a self-critic way where the reward is the CIDEr [51] score of the beam-searched [3] sequence with the baseline defined as the average reward. The success of [46, 3] settled the standard practice of firstly training a model using a cross-entropy loss and then fine-tuning it using RL [35, 24, 61, 27, 17], an approach also adopted in our work where the initial captioning model is trained using REST [8].

More recently, in [23], CLIP is proposed as a metric to evaluate captions. The work of [11] proposes to use such a metric to replace the CIDEr score reward for finetuning an image captioning model with RL. We also used a similar reward but our motivation is different: we use it to prevent our model's output from collapsing into a generic class-specific caption which is not descriptive of the input video.

3. Method

3.1. Goal & training setting

ReGen's goal: Our goal is to train a discriminative video captioning model whose output caption w can be used for classifying the given video v in terms of a set of $\mathcal{C}_N$ novel action/video classes, i.e. classes not seen during training. To this end, ReGen seeks to satisfy *three requirements*: (1) train a discriminative captioning model where the generated caption can be used for zero-shot video/action recognition,

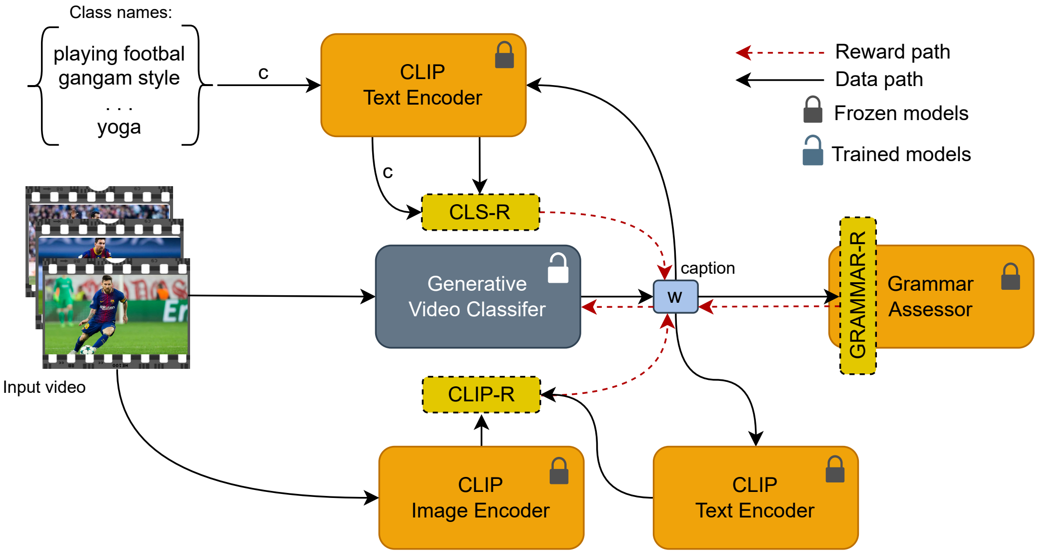


Figure 1: **ReGen overview:** Starting from a pre-trained generative video captioning model, ReGen fine-tunes it on a labelled video/action recognition dataset using RL with the goal of generating discriminative captions for open-world video/action recognition. The caption is tuned to be discriminative using a classification-based reward (`CLS-R`) which is a cross-entropy loss between CLIP text embeddings of the generated caption and the class names. Moreover, to prevent the model from producing generic class-specific captions which are not descriptive of the input video, a CLIP-based reward is proposed (`CLIP-R`). Finally, to ensure the grammatical correctness of the caption, we propose a grammar reward (`GRAMMAR-R`). Note that ReGen uses only class labels for RL-based fine-tuning (i.e. no captions are used for training at all).

i.e. *the caption should be discriminative*; (2) maintain or even enhance the video-specific granularity of the generated caption, i.e. *the caption should be video-specific*; (3) training should not compromise the grammatical correctness of the generated caption, i.e. *the caption should be grammatically correct*. In ReGen, each of these requirements is addressed with a dedicated loss as explained in Sec. 3.2. Fig. 1 shows an overview of our training framework.

Model architecture: Our model is a generative one $p_\theta(w|v)$ consisting of a video encoder $g_v(.)$ and an autoregressive text decoder $g_t(.)$. We used the same generative architecture for video captioning as in REST which, in turn, adapts recently proposed image-based auto-regressive models with minimal changes in order to accommodate the use of temporal information. In particular, our model is based on GIT [52], but we note that ReGen can be used to train other architectures, too (e.g. BLIP [29]). The vision encoder $g_v(.)$ is a ViT [15], the output of which is flattened and projected using a linear layer and a layer norm to match the dimensionality of the text decoder. When processing multiple frames, a temporal embedding is added to the vision features corresponding to each frame. The features are simply averaged across the time dimension thereafter. The text decoder $g_t(.)$ follows BERT [13] taking as input the concatenation of the flattened vision features and the text embeddings, separated by the `[BOS]` token. At test time, it auto-regressively generates a caption until `[EOS]` is reached. Finally, to use the generated caption for video classification, we compute a text embedding from it using CLIP's text encoder and, similarly, a text embedding for all class names in the given test dataset. Then, the predicted class is the one corresponding to the maximum inner product between the caption and the class name embeddings.

Training setting: As usual in open-world (i.e. zero-shot) recognition, the model is trained on an action/video recognition dataset of $\mathcal{C}_B$ base classes where $\mathcal{C}_B \bigcup \mathcal{C}_N = \emptyset$. The training dataset consists of N video-label pairs $\{v_i, q_i\}$ $i = 1, \ldots, N$. Note that we do not use ground truth captions for training (only the class names) but it is assumed that our model is already pre-trained to output a caption [1]. Specifically, our model has been pre-trained with REST [8] which adapts a generative image-based V&L model (GIT [52] or BLIP [29]) into a video captioner in an unsupervised manner i.e. *without using any class labels at all*. Under this setting, our training framework, ReGen, uses reinforcement learning to further train the initial model (pre-trained with REST) by enabling the integration of class label information to meet requirements (1)-(3) mentioned above.

[1] This makes our approach quite different to previous works which use reinforcement learning for video captioning.

3.2. ReGen

To meet the 3 key requirements introduced above, i.e. the generated caption should be (1) *discriminative*, (2) *video-specific*, and (3) *grammatically correct*), we introduce three respective rewards detailed in Sections 3.2.1, 3.2.2 and 3.2.3 which are used to train our model using reinforcement learning.

3.2.1 Classification Reward CLS-R

Training a video captioning model to directly auto-regress the class names with a standard language modelling loss (i.e. a cross entropy loss) leads to severe base class overfitting with the model completely losing its zero-shot generalization capabilities [8]. This is in contrast to the contrastive-based CLIP-adaptation approaches which tend to maintain decent zero-shot generalization properties [39].

To alleviate the aforementioned limitation and encourage class-level discrimination, we propose a classification-based reward (CLS-R). Given our generative captioning model $p_\theta(w|v)$, the closest class name to the generated caption w is found by measuring the cosine similarity between w and the class names in the embedding space of a pre-trained CLIP text encoder. Specifically, let $\mathbf{t} = \mathtt{CLIP}_T(w)$ be the embedding obtained by feeding w to the CLIP's text encoder. Moreover, let $\mathbf{t}_c = \mathtt{CLIP}_T(q_c), c = \{1, \dots, C\}$ be the CLIP text embedding corresponding to the name q_c of the c-th class (in practice, we used the text encoder corresponding to the ViT-B/16 variant of CLIP). Then, the probability over class labels is given by:

$$p(y|\mathbf{t}) = \frac{\exp(\mathtt{cos}(\mathbf{t}_y, \mathbf{t})/\tau)}{\sum_{c=1}^{C} \exp(\mathtt{cos}(\mathbf{t}_c, \mathbf{t})/\tau)}, \quad (1)$$

where τ is a temperature factor and $\mathtt{cos}$ the cosine similarity. For a given video v, we then define the classification reward CLS-R to be equal to the cross-entropy loss:

$$\mathtt{CLS\text{-}R}(w) = -\sum_{c=1}^{C} \log p(c|w) y_c. \quad (2)$$

The model $p_\theta(.)$ is then optimized using REINFORCE with a self-critique baseline [46]. Specifically, we approximate the gradient of the expected reward for the generated caption w by normalising the reward of the beam-searched caption w_b with that of the greedy decoded one w_g:

$$\nabla_\theta \mathbb{E}_{w \sim p_\theta(w|v)} \approx (r(w_b) - r(w_g)) \nabla_\theta p_\theta(w_b|v), \quad (3)$$

where $r(w) = \mathtt{CLS\text{-}R}(w)$.

Why CLS-R is good for open-world recognition? Firstly, training with CLS-R and RL enhances the model's discriminability. CLS-R reward evaluates the caption globally, measuring how well it is mapped to the correct class and

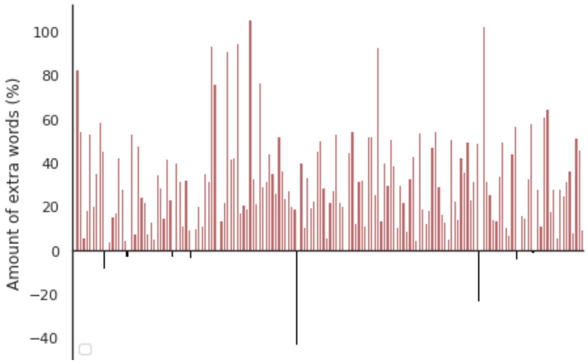
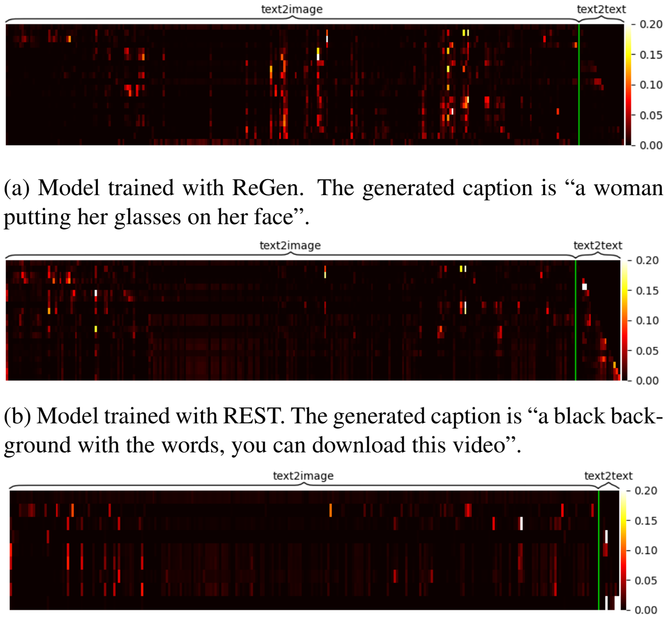


Figure 2: Change in word diversity per class between ReGen and REST [8]. Each bar represents a novel class. A red bar indicate a positive increase (measure in %), while a black bar a decrease. Notice that, for the vast majority of classes, ReGen significantly increases the word diversity.



(a) Model trained with ReGen. The generated caption is "a woman putting her glasses on her face".

(b) Model trained with REST. The generated caption is "a black background with the words, you can download this video".

(c) Model trained with directly auto-regressing the class names. The generated caption is "fixing hair".

Figure 3: Comparison of text-to-image and text-to-text attention maps of a model trained with (a) ReGen, (b) REST and (c) directly auto-regressing the class names. The input video's ground truth class is "adjusting glasses" which is a class not seen during training. Our approach in (a) attends mostly to the input video (the text-to-image attention part) as opposed to cases (b) or (c). The later concentrates its attention mostly on the text part (text-to-text region of the plot), a symptom of strong base class overfitting. Only ReGen's output correlates with the correct class. **Best viewed in colour.**

how distinguishable it is from the other potential classes. That is, the decision of whether a caption describes correctly a given class is taken holistically, based on the entire caption, and not locally as enforced by a fully supervised setting. Importantly, changing a single token will have a small effect on the overall supervised loss (as its value is the sum or average over the entire caption), but may change the global meaning entirely.

Secondly, training with `CLS-R` reward enhances the model's zero-shot generalizability. During each RL episode, we explore, in the neighbourhood of the current prediction, a set of trajectories, sampled stochastically with beam search, that go beyond the wording generated by a model trained with a language modelling loss. In this process, new words/text, previously unexplored during the next token prediction-based learning, emerge. By not enforcing a particular output (as for example with a standard cross-entropy loss), we allow the model to explore new words and sentence constructions, while avoiding the overfitting to a predefined set of classes or captions [2]. In Fig. 2, we show the relative increase in number of unique words per class used by a model trained with ReGen. Compared with the REST baseline, our predictions are generally between 20 and 100% more diverse, with the newly discovered words not learned by any explicit supervision.

As a result of training with `CLS-R`, a model trained with ReGen is relying more on the input video and less on the previously generated text to complete the caption generation process. This can be observed in Fig. 3 where the text-to-image and text-to-text attention maps of 3 models are compared. The output at the top figure is from a model trained with ReGen which clearly attends more to the video and less to the text to generate the caption. The output at the middle figure is from a model trained with REST which relies more on the text (observe the high score for some text-to-text attention values) thus "cheating" to some extent. Finally, the output at the bottom figure is from a model trained to directly auto-regress the class name which relies heavily on text and suffers from severe base class overfitting.

3.2.2 Video-specific reward CLIP-R

While training with the `CLS-R` reward encourages class-discrimination, the output caption may degenerate towards a generic textual description of the class i.e. stop being descriptive of each individual input video. For example, all videos belonging to *playing football* class may output the same caption, for example, *two persons playing football* despite the number of players changing from one sample to another. To address this problem, we propose to enforce that the video and the generated caption are similar in CLIP's joint embedding space. We define our video-caption `CLIP-R` reward as:

$$\texttt{CLIP-R}(v, w) = \max(\cos(\mathbf{f}, \mathbf{t}), 0), \tag{4}$$

where $\mathbf{f} = \frac{1}{F}\sum_{i=1}^{F} \texttt{CLIP}_I(v_i)$ and F the number of frames of v. The gradient of the expected reward is calculated by plugging Eq. 4 into Eq. 3.

Note that, very recently, [11] also proposed a clip-based reward with the intention of replacing CIDEr optimization for generating more accurate and fine-grained image captions. In contrast, our motivation is different: we use `CLIP-R` to prevent our model's output from collapsing into a degenerate class-specific caption.

3.2.3 Grammar Reward GRAMMAR-R

The previously proposed rewards may result in captions that are grammatically incorrect (e.g. incorrect word ordering, repeated word(s), etc.). Specifically, `CLS-R` is only sensitive to the discriminability of a caption with respect to the given classes and `CLIP-R` does not encourage grammatically coherent text as the CLIP text encoder is not sensitive to grammar or word ordering [58]. To address this limitation, we propose the following grammar rewards:

LM-based: Herein, the goal is to leverage the per-token log-likelihood of a pretrained LM as a proxy of assessing the quality of the generated caption by using its value as a reward. For example, as repeated or swapped words are unlikely to occur in natural language, they tend to have lower scores, guiding the model accordingly.

Given the sequence of tokens $\mathbf{U} = \{u_1, \ldots, u_{|U|}\}$ corresponding to a given caption w, we compute the log-likelihood score by replacing the j-th token with `[MASK]`: $\mathbf{U}_{/j} = \{u_1, \ldots, u_{j-1}, \texttt{[MASK]}, u_{j+1}, \ldots, u_{|U|}\}$ as:

$$\texttt{GRAMMAR-R}_{LM}(w) = \frac{1}{|U|}\sum_{j=1}^{|U|} \log L_\theta(u_j|\mathbf{U}_{/j}), \tag{5}$$

where L_θ is a pre-trained masked LM [13, 48]. Note that the score is normalized by its length to avoid biasing the predicted text towards shorter sentences. While a score could be computed in a similar manner in one go using auto-regressive models, since our text decoder uses a BERT tokenizer, the proposed approach allows us to provide a dense reward on a token-by-token basis simply by removing the sum from Eq. 5. We experimented with multiple models which we ablate in Sec. 5. Note that, depending on the data the LM was pretrained on, such a loss can also induce a style change to the generated caption.

Discriminator-based: Instead of relying on a pre-trained LM, herein, we train a grammar discriminator $d(.)$ to assess

[2] Overfitting to a set of captions occurs in REST where the model is trained for each video to generate a caption sampled from a small set of potential pseudo-captions.

whether a given caption is grammatically correct and coherent. The training data is created by perturbing the pseudo-captions from [8] with random errors: swapping, shuffling, repeating, inserting or removing tokens and groups of tokens. The discriminator is trained using a binary cross entropy loss, where 1 indicates a correct sample and 0 a wrong one. To allow for faster convergence, we initialize the model from CLIP's text encoder. The network is fine-tuned for 10 epochs using AdamW, a batch size of 256, and a learning rate of $1e-4$. In this case:

$$\texttt{GRAMMAR-R}_d(w) = d(w). \tag{6}$$

In both cases, the gradient of the expected reward is approximated by plugging Eq. 5 or Eq. 6 into Eq. 3. In practice, we employ only one of the two variants proposed above.

3.2.4 Total Reward

The total reward is: $r_{total} = \lambda_{CLS-R} \cdot \texttt{CLS-R} + \lambda_{CLIP-R} \cdot \texttt{CLIP-R} + \lambda_{GRAMMAR-R} \cdot \texttt{GRAMMAR-R}$, where $\lambda_{CLS-R}, \lambda_{CLIP-R}, \lambda_{GRAMMAR-R}$ are the corresponding reward weights.

4. Experimental setting

Datasets: Unless otherwise stated, we trained our models on the Kinetics-400 [25] dataset and evaluated them for zero-shot action recognition on the standard benchmarks of HMDB-51 [26] and UCF-101 [49]. Moreover, we evaluated on Kinetics-600 [9] using the (three) splits defined in [10]. Each split consists of 160 novel classes (not present in Kinetics-400) covering 220 classes in total across the 3 splits. To ensure no overlap, the classes were renamed from 600 to 620. We refer to this evaluation subset as Kinetics-220 in the zero-shot setting and as 620 in the generalized zero-shot one. We also performed few-shot recognition experiments on HMDB-51 and UCF-101. Finally, we conducted a zero-shot captioning experiment on VaTeX [55].

Models used and training setting: To show that ReGen's training framework is architecture agnostic, we used two architectures for the video captioning model, one based on BLIP [29] and one on GIT [52]. Following standard practices [46, 3], the initial captioning models were firstly pre-trained with a language modelling loss using REST [8] and then finetuned with RL using the proposed ReGen.

Training hyperparameters: We trained all of our models for 10 epochs using AdamW [34], with a learning rate of $1e-7$, a weight decay of 0.001 and a batch size of 8. Unless otherwise stated $\lambda_{CLS-R} = 1.0$, $\lambda_{CLIP-R} = 5.0$, $\lambda_{GRAMMAR-R} = 0.2$. For few-shot training, we followed the hyper parameters from [8]. We list all augmentations and hyper-parameters in the supplementary material. All of our models and training code were implemented using PyTorch [40].

5. Ablation studies

Effect of each reward function in ReGen: Herein we evaluate the effect of each reward function used in ReGen on the accuracy and behavior of the model. We report results for each component in Table 1 for zero-shot action recognition (HMDB-51, UCF-101 and Kinetics-220) and zero-shot video captioning (VaTeX). The later experiment is used here to assess the quality of the generated caption. As it can be observed, applying the proposed `CLS-R` reward alone results in models that exhibit strong zero-shot recognition ability (2nd row), but degraded text quality: The text is often incoherent with repetition of words or hallucinated class-relevant details that may not be present in the video. The addition of `CLIP-R` reward further boosts the discrimination ability (3rd row) with the model now learning to focus on the input video itself. However, as the CLIP text encoder is not sensitive to the word ordering, the quality of the generated caption remains low. This illustrates the necessity of the proposed `CLS-Grammar` reward which significantly boosts the generated caption's quality. It is worth noting that the combination of `CLS-R` with `CLS-Grammar` without including `CLIP-R` is still prone to generating captions that may not reflect the visual content, hence the lower CIDEr score in this case. Finally, the model trained using all 3 reward functions (last row) offers the best trade-off between classification accuracy and caption quality, showcasing the importance of using the combination of proposed rewards.

Effect of `GRAMMAR-R` reward used: In this section, we (a) compare the different variants of the proposed `GRAMMAR-R` reward (LM-based *vs.* discriminator-based; see Sec. 3.2.3), and (b) analyze the impact of the LM architecture used on the performance. As the results from Table 2 show, the discriminator-based grammar correctness reward produces sentences with a higher CIDEr score but of lower discriminability. This can be explained by the fact that it was trained on the pseudo-labels produced by REST, hence it tends to maintain the initial sentence structure and may penalize slightly new words. In contrast, the LM was trained on large corpora of text, and are less likely to penalize new tokens, but will generally change the style of the text. In terms of the LM architectures used for `GRAMMAR-R`, we tested the following BERT [13] variants and derivatives: DistillBERT-Base [48], BERT-Base, BERT-Large and DeBERTa-Base [22]. As shown in Table 3, all models perform similarly, with larger models having an edge.

Comparison with REST and effect of architecture: Herein, we report the results obtained by using two video captioning models for ReGen based on BLIP [29] and GIT [52] architectures. The first one is based on cross-attention between visual and text features, while the second one concatenates the vision and text tokens and performs self-attention. In both cases, we compare against the REST

CLS-R	CLIP-R	GRAMMAR-R	K220	HMDB-51	UCF-101	VaTeX
✗	✗	✗	29.6	49.9	71.6	39.1
✓	✗	✗	38.6	55.0	77.9	28.1
✓	✓	✗	38.9	55.5	78.5	29.3
✓	✗	✓	37.2	54.1	76.2	33.4
✓	✓	✓	38.2	55.1	76.4	40.5

Table 1: **Impact of each reward in ReGen** on zero-shot classification on HMDB-51, UCF-101 and Kinetics-220 (1-vs-620 setting) in terms of Top-1 (%) accuracy, and on zero-shot captioning on VaTeX in terms of CIDEr. The latter experiment reflects the quality of the generated caption.

GRAMMAR-R	K220	HDMB-51	UCF-101	VaTeX
LM	38.2	55.1	76.4	40.5
Discriminator	35.1	55.0	75.0	46.1

Table 2: **Impact of the `GRAMMAR-R` variant** on zero-shot classification on HMDB-51, UCF-101 and Kinetics-220 (1-vs-620 setting) in terms of Top-1 (%) accuracy, and on zero-shot captioning on VaTeX in terms of CIDEr. We compare the two versions, LM-based vs discriminator-based, proposed in Sec. 3.2.3

Model	K220	HDMB-51	UCF-101	VaTeX
DistillBERT-B [48]	37.4	53.6	74.4	38.2
BERT-B [13]	37.5	53.6	74.6	38.3
BERT-L [13]	38.2	55.1	76.4	40.5
DeBERTa-B [22]	37.6	55.3	76.7	39.2

Table 3: **Impact of the LM used for `GRAMMAR-R`** on zero-shot classification on HMDB-51, UCF-101 and Kinetics-220 (1-vs-620 setting) in terms of Top-1 (%) accuracy, and on zero-shot captioning on VaTeX in terms of CIDEr.

Method	Arch.	K220	HDMB-51	UCF-101
REST [8]	BLIP	29.51	49.7	69.1
ReGen (Ours)		37.6	54.5	75.1
REST [8]	GIT	29.6	49.9	71.6
ReGen (Ours)		38.2	55.1	76.4

Table 4: **Impact of the architecture used in ReGen** on zero-shot classification on HMDB-51, UCF-101 and Kinetics-220 (1-vs-620) in terms of Top-1 (%) accuracy.

baseline which is used to provide the initial captioning models. As the results from Table 4 show, ReGen offers large gains over REST for both architectures.

6. Comparison with state-of-the-art

In this section, we compare ReGen against the state-of-the-art on zero-shot action recognition, few-shot action recognition, and zero-shot video captioning. Unless otherwise stated, all methods reported (including ours) were trained on Kinetics-400 ensuring a fair comparison.

Zero-shot action recognition: Herein, we compare ReGen against the state-of-the-art for zero-shot action recognition on UCF-101, HMDB-51 and Kinetics-220. On UCF-101 and HMDB-51, as the results from Table 5 show, we outperform all prior methods by a large margin. In particular, on HMDB-51, we improve in absolute terms upon the previous best result of REST [8] by 5.4% while, on UCF-101, upon X-Florence [39] by 3.3%. On Kinetics-220, for the standard 1-vs-160 setting (*i.e.* classify in terms of 160 novel classes), we outperform REST by more than 10%, being second overall only to the recent discriminative method of [39]. As Table 7 shows, once all 620 class names are considered, *i.e.* once the base class names are included in the evaluation, (1-vs-620 setting; also known as generalized zero-shot setting), ReGen, outperforms X-CLIP [39] by more than 20%. This suggests that the superiority of X-CLIP on the 1-vs-160 setting can be considered "artificial" and that X-CLIP is prone to base class overfitting. For qualitative examples see Fig. 4 where we compare our approach with REST [8].

Method	HMDB-51	UCF-101
Discriminative approaches		
MTE [57]	19.7 ± 1.6	15.8 ± 1.3
ASR [54]	21.8 ± 0.9	24.4 ± 1.0
ZSECOC [42]	22.6 ± 1.2	15.1 ± 1.7
UR [62]	24.4 ± 1.6	17.5 ± 1.6
E2E [18]	32.7	48
TS-GCN [6]	23.2 ± 3.0	34.2 ± 3.1
ER-ZSAR [10]	35.3 ± 4.6	51.8 ± 2.9
CLIP [44]	46.2	73.0
MUFI [43]	31.0	60.9
ActionCLIP [53]	40.8 ± 5.4	58.3 ± 3.4
ClipBert [28]	21.4 ± 1.0	27.8 ± 0.8
Frozen [4]	27.8 ± 0.3	45.9 ± 1.3
ViSET-96 [14]	40.2	68.3
BridgeFormer [19]	37.7 ± 1.2	53.1 ± 1.4
AURL [41]	40.4	60.9
ResT_101 [30]	41.1 ± 3.7	58.7 ± 3.3
X-CLIP [39]	44.6 ± 5.2	72 ± 2.3
X-Florence [39]	48.4 ± 4.9	73.2 ± 4.2
Generative approaches		
REST [8]	49.7 ± 1.14	69.1± 0.62
ReGen (Ours)	**55.1 ± 0.4**	**76.4 ± 0.2**

Table 5: Zero-shot classification results on HMDB-51 and UCF-101 in terms of Top-1 (%) accuracy.

Method	HMDB-51				UCF-101			
	2	4	8	16	2	4	8	16
Discriminative approaches								
TSM [31]	17.5	20.9	18.4	31.0	25.3	47.0	64.4	61.0
TimeSformer [5]	19.6	40.6	49.4	55.4	48.5	75.6	83.7	89.4
Swin-B [33]	20.9	41.3	47.9	56.1	53.3	74.1	85.8	88.7
X-CLIP [39]	53.0	57.3	62.8	64.0	76.4	83.4	88.3	91.4
X-Florence [39]	51.6	57.8	64.1	64.2	84.0	88.5	92.5	94.8
Generative approaches								
REST [8]	54.0	59.1	62.1	64.0	88.2	90.2	92.6	93.5
ReGen (Ours)	**61.8**	**67.7**	**70.7**	**72.5**	**88.5**	**91.4**	**93.9**	**94.9**

Table 6: Few-shot classification results on HMDB-51 and UCF-101 in terms of Top-1 (%) accuracy.

Method	Top-1	Top-5
Discriminative approaches		
SJE [1]	22.3 ± 0.6	48.2 ± 0.4
ESZSL [47]	22.9 ± 1.2	48.3 ± 0.8
DEM [60]	23.6 ± 0.6	49.5 ± 0.4
GCN [20]	22.3 ± 0.6	49.7 ± 0.6
ER-ZSAR [10]	42.1 ± 1.4	73.1 ± 0.3
X-CLIP [39]	65.2 ± 0.4	86.1 ± 0.8
X-Florence [39]	**68.8 ± 0.9**	**88.4 ± 0.6**
Generative approaches		
REST [8]	51.7 ± 1.1	75.2 ± 0.4
ReGen (Ours)	**62.0 ± 0.8**	**83.8 ± 0.4**

Table 7: Zero-shot classification results on Kinetics-220 (1-vs-160 setting).

Method	Top-1	Top-5
Discriminative approaches		
X-CLIP [39]	14.76 ± 0.51	60.93 ± 0.25
Generative approaches		
REST [8]	29.51 ± 0.71	56.12 ± 0.37
ReGen (Ours)	**38.1 ± 0.5**	**66.8 ± 0.1**

Table 8: **Generalized** zero-shot classification results on Kinetics-220 (1-vs-620 setting).

Few-shot action recognition: We also adapted a model trained by ReGen for few-shot action recognition on UCF-101 and HMDB-51. Following the existing protocols, we train and test for both datasets using the 3 available splits. Table 6 shows that our approach significantly outperforms all prior works across different number of shots ({2,4,8,16}) setting a new state-of-the-art for both datasets, improving by 6.8-8.6% on HDMB-51 and by 0.3-1.4% on UCF-101.

Zero-shot captioning: Our main aim in this paper is to train a strong zero-shot action recognition model. However, as our approach produces a caption, we conducted an experiment to assess its performance for zero-shot video captioning, too. As the results from Table 9 show, our approach matches the considerably bigger Flamingo models [2] (Flamingo-80B is for example more than 160× bigger), trained using more than 2B images and 24M videos.

Method	CIDEr
Flamingo-3B [2]	40.1
Flamingo-8B [2]	39.5
Flamingo-80B [2]	46.7
REST (GIT-arch) [8]*	39.1
Ours (w. `LM`)	40.5
Ours (w. `Discriminator`)	46.1

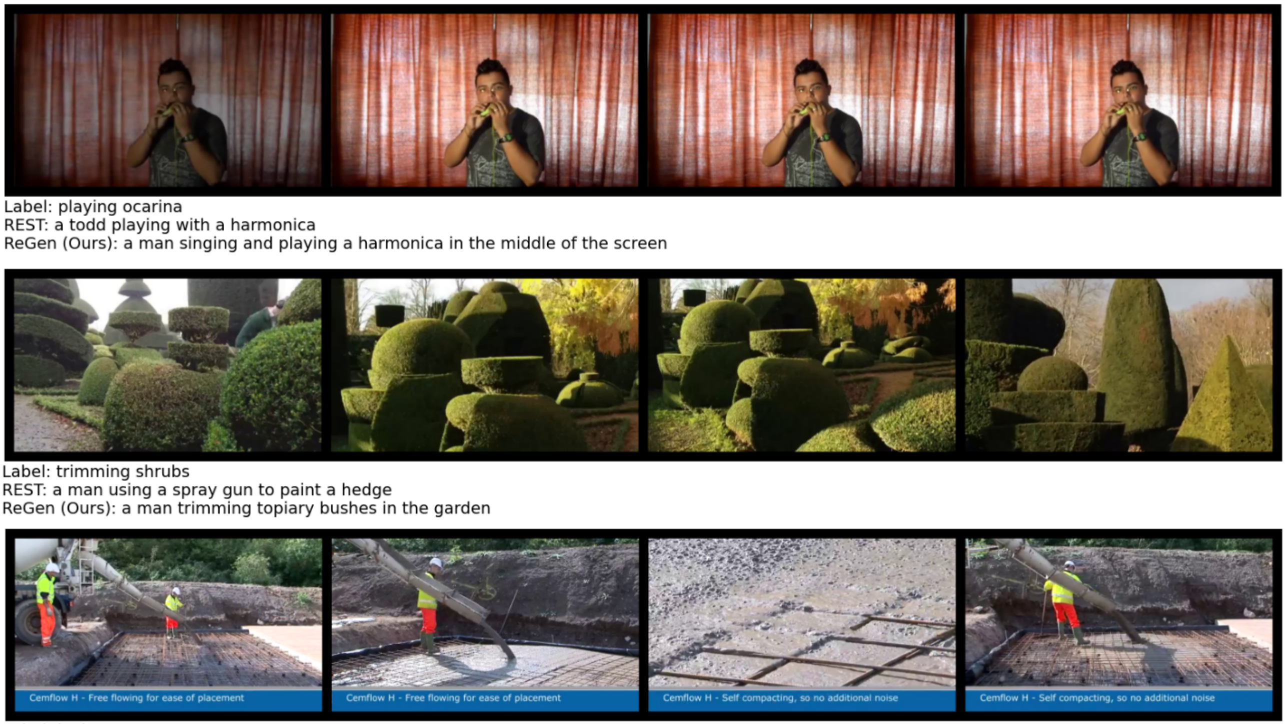
Table 9: Zero-shot video captioning results on VaTeX in terms of CIDEr score. * - re-implementation

7. Conclusions

We introduced, ReGen, a framework for training a discriminative captioning model for zero-shot video/action recognition. Our work, introduces a novel RL solution that alleviates base class overfitting using a 3-fold reward function: (a) a class discrimination reward, `CLS-R`, that enforces the generated caption to be correctly classified, (b) a CLIP-based reward, `CLIP-R` than encourages the caption to be video-specific, and (c) a grammar reward, `GRAMMAR-R`, that preserves the grammatical correctness of the caption. We show that the proposed RL solution is effective (and necessary) for both mitigating base-class overfitting and improving the discriminability of the generated caption. Moreover, ReGen sets a new state-of-the-art for both zero-shot and few-shot action recognition. Hence, we conclude that *a good generative zero-shot video classifier should be rewarded.*

References

[1] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-


Cemflow H - Free flowing for ease of placement
Cemflow H - Free flowing for ease of placement
Cemflow H - Self compacting, so no additional noise
Cemflow H - Self compacting, so no additional noise

Label: playing ocarina
REST: a todd playing with a harmonica
ReGen (Ours): a man singing and playing a harmonica in the middle of the screen

Label: trimming shrubs
REST: a man using a spray gun to paint a hedge
ReGen (Ours): a man trimming topiary bushes in the garden

Label: laying concrete
REST: a man pouring concrete into a pool
ReGen (Ours): a construction worker pouring concrete into the pit of a concrete foundation

Figure 4: Examples of captions produced by our approach and REST for a set of videos from Kinetics-220 (*i.e.* zero-shot setting).

grained image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 8

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances on Neural Information Processing Systems*, 2022. 8

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 6

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 7

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, 2021. 8

[6] Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. Rethinking zero-shot video classification: End-to-end training for realistic applications. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2, 7

[7] Carlo Bretti and Pascal Mettes. Zero-shot action recognition from diverse object-scene compositions. *British Machine Vision Conference*, 2021. 2

[8] Adrian Bulat, Enrique Sanchez, Brais Martinez, and Georgios Tzimiropoulos. REST: REtrieve & Self-Train for generative action recognition. *arXiv preprint arXiv:2209.15000*, 2022. 1, 2, 3, 4, 6, 7, 8

[9] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 6

[10] Shizhe Chen and Dong Huang. Elaborative rehearsal for zero-shot action recognition. In *IEEE International Conference on Computer Vision*, 2021. 6, 7, 8

[11] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with clip reward. *arXiv preprint arXiv:2205.13115*, 2022. 2, 5

[12] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3, 5, 6, 7

[14] Keval Doshi and Yasin Yilmaz. Zero-shot action recognition with transformer-based video semantic embedding. *arXiv preprint arXiv:2203.05156*, 2022. 7

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021. 3

[16] Valter Estevam, Rayson Laroca, Helio Pedrini, and David Menotti. Global semantic descriptors for zero-shot action recognition. *IEEE Signal Processing Letters*, 2022. 2

[17] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Injecting semantic concepts into end-to-end image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[18] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *AAAI Conference on Artificial Intelligence*, 2019. 7

[19] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2, 7

[20] Pallabi Ghosh, Nirat Saini, Larry S Davis, and Abhinav Shrivastava. All about knowledge graphs for actions. *arXiv preprint arXiv:2008.12432*, 2020. 8

[21] Meera Hahn, Andrew Silva, and James M Rehg. Action2vec: A crossmodal embedding approach to action learning. *arXiv preprint arXiv:1901.00484*, 2019. 2

[22] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020. 6, 7

[23] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 2

[24] Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *AAAI conference on artificial intelligence*, 2021. 2

[25] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6

[26] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: a large video database for human motion recognition. In *IEEE International Conference on Computer Vision*, 2011. 6

[27] Chia-Wen Kuo and Zsolt Kira. Beyond a pre-trained object detector: Cross-modal textual and visual context for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[28] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 7

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *International Conference on Machine Learning*, 2022. 3, 6

[30] Chung-Ching Lin, Kevin Lin, Lijuan Wang, Zicheng Liu, and Linjie Li. Cross-modal representation learning for zero-shot action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2, 7

[31] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *IEEE International Conference on Computer Vision*, 2019. 8

[32] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *IEEE International Conference on Computer Vision*, 2017. 2

[33] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 8

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[35] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. Dual-level collaborative transformer for image captioning. In *AAAI conference on artificial intelligence*, 2021. 2

[36] Zhekun Luo, Shalini Ghosh, Devin Guillory, Keizo Kato, Trevor Darrell, and Huijuan Xu. Disentangled action recognition with knowledge bases. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022. 2

[37] Pascal Mettes. Universal prototype transport for zero-shot action recognition and localization. *arXiv preprint arXiv:2203.03971*, 2022. 2

[38] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2

[39] 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. *European Conference on Computer Vision*, 2022. 2, 4, 7, 8

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances on Neural Information Processing Systems*, 2019. 6

[41] Shi Pu, Kaili Zhao, and Mao Zheng. Alignment-uniformity aware representation learning for zero-shot video classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 7

[42] Jie Qin, Li Liu, Ling Shao, Fumin Shen, Bingbing Ni, Jiaxin Chen, and Yunhong Wang. Zero-shot action recognition with error-correcting output codes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 7

[43] Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Xiao-Ping Zhang, Dong Wu, and Tao Mei. Boosting video representation learning with multi-faceted integration. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 7

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 7

[45] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015. 2

[46] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 4, 6

[47] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, 2015. 8

[48] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 5, 6, 7

[49] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances on Neural Information Processing Systems*, 2017. 2

[51] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[52] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 3, 6

[53] Mengmeng Wang, Jiazheng Xing, and Yong Liu. ActionCLIP: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 7

[54] Qian Wang and Ke Chen. Alternative semantic representations for zero-shot human action recognition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2017. 7

[55] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. VaTeX: A large-scale, high-quality multilingual dataset for video-and-language research. In *IEEE International Conference on Computer Vision*, 2019. 6

[56] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Reinforcement learning*, pages 5–32, 1992. 2

[57] Xun Xu, Timothy M Hospedales, and Shaogang Gong. Multi-task zero-shot action recognition with prioritised data augmentation. In *European Conference on Computer Vision*, 2016. 7

[58] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bag-of-words models, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022. 5

[59] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M Hospedales. Actor-critic sequence training for image captioning. *arXiv preprint arXiv:1706.09601*, 2017. 2

[60] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 8

[61] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[62] Yi Zhu, Yang Long, Yu Guan, Shawn Newsam, and Ling Shao. Towards universal representation for unseen action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 7