Efficient Vision-Language pre-training via domain-specific learning for human activities

Adrian Bulat* [1,2] **Yassine Ouali*** [1] **Ricardo Guerrero**[1] **Brais Martinez**[1]
Georgios Tzimiropoulos[1,3]

[1]Samsung AI Cambridge, [2]Technical University of Iasi, [3]Queen Mary University of London

Abstract

Current Vision-Language (VL) models owe their success to large-scale pre-training on web-collected data, which in turn requires high-capacity architectures and large compute resources for training. We posit that when the downstream tasks are known in advance, which is in practice common, the pretraining process can be aligned to the downstream domain, leading to more efficient and accurate models, while shortening the pretraining step. To this end, we introduce a domain-aligned pretraining strategy that, without additional data collection, improves the accuracy on a domain of interest, herein, that of human activities, while largely preserving the generalist knowledge. At the core of our approach stands a new LLM-based method that, provided with a simple set of concept seeds, produces a concept hierarchy with high coverage of the target domain. The concept hierarchy is used to filter a large-scale web-crawled dataset and, then, enhance the resulting instances with targeted synthetic labels. We study in depth how to train such approaches and their resulting behavior. We further show generalization to video-based data by introducing a fast adaptation approach for transitioning from a static (image) model to a dynamic one (i.e. with temporal modeling). On the domain of interest, our approach significantly outperforms models trained on up to $60\times$ more samples and between $10-100\times$ shorter training schedules for image retrieval, video retrieval and action recognition. Code will be released.

1 Introduction

Billion-scale vision-language pre-training on web collected image-text data (Radford et al., 2021; Jia et al., 2021; Yuan et al., 2021; Yu et al., 2022; Alayrac et al., 2022; Li et al., 2022a; Wang et al., 2022b) has significantly pushed the state-of-the-art for both uni-modal (e.g. action recognition) and multi-modal understanding (e.g. image captioning, VQA, etc.). The current trend in VL pre-training continues to scale up both the training datasets (Zhai et al., 2023) and model sizes (Sun et al., 2023), further increasing the already very large training- and test-time computational requirements. This is in part due to the generalist nature of such models. However, the domain of application is often known in practice. Ideally, we should use this prior knowledge to increase the training and inference time efficiency. This is the very goal of our work, as we propose a novel methodology to align the VL pre-training process to a given domain, without using additional data, and without catastrophically compromising the generalist capabilities. To the best of our knowledge, this is the very first work to systematically tackle and study such a setting. To illustrate these ideas and due to its relevance in many important tasks/applications, we focus on pre-training specialized to the domain of human activities.

With this in mind, we introduce a new LLM-based method that, provided with a simple set of concept seeds, iteratively produces a hierarchy of textual queries that provide high coverage of the concepts included in the target domain. These textual queries, alongside a set of predefined safety and quality filters, are used to retrieve relevant data from an internet-scale web-crawled dataset. We further enhance the retrieved instances using an image captioning model. This ensures the relevance of the textual captions, which is particularly important when relying on text-to-image retrieval. Empirically, we also show that training with domain-focused data naturally increases the frequency of hard negatives within a given batch, improving the training efficiency. This is a notable difference with other works within this area of VL model improvement via data enhancement, such as (Radenovic et al., 2023; Xu et al., 2023a,b) that focus on domain-invariant processes, using hardcoded rules

* - Denotes equal contribution

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 7978–8000
November 12-16, 2024

(e.g. based on WordNet, sentence complexity, etc.).

In summary, our contributions are:

- **HC-IT & HC-VL:** To increase the train-time efficiency, while at the same time improving the model's accuracy, we propose a *domain-specific* VL pre-training, herein tailored to the space of human activities. At the core of our approach stands the newly proposed LLM-based data filtering, whereby the LLM iteratively creates a hierarchy of textual queries providing extensive coverage of the semantic concepts in the domain of interest. The thousands of these activity-related queries are then used for data filtering through retrieval. As some captions may be incorrect, we supplement the original captions with synthetic ones. We coin the resulting subset Human-Centric Image-Text (HC-IT) dataset, and the model trained on these data as Human-Centric Vision-Language (HC-VL) model.
- **HC-VL+:** The image-based HC-VL model already produces state-of-the-art results across various datasets and settings. However, to further boost the model's performance, adding temporal modeling, we introduce a set of architectural changes that (a) maintain the zero-shot capabilities of the image model, (b) avoiding catastrophic forgetting, while being (c) train-time efficient.
- **Results:** Our approach, is significantly faster to train (10-100$\times$ fewer iterations) and requires up to 60$\times$ fewer samples (see Fig. 1). In terms of accuracy, our approach significantly outperforms on the domain of human activities, i.e. for action recognition (+6% on average), image retrieval (+7%) and video retrieval (+4%) models trained on up to 60$\times$ more samples (e.g. SigLIP (Zhai et al., 2023)), while largely retaining competitive performance on out-of-domain data (i.e. ImageNet).

2 Related work

Vision-Language pre-training has emerged as the foremost approach for representation learning and training of robust zero-shot models. Based on the training objective, methods can be broadly placed into two categories: generative (Wang et al., 2022b, 2021b; Bulat et al., 2023) and contrastive (Radford et al., 2021; Fürst et al., 2021; Yeh et al., 2022; Mu et al., 2021; Li et al., 2023b; Fu et al., 2021; Wang et al., 2022a), although, recently promising results

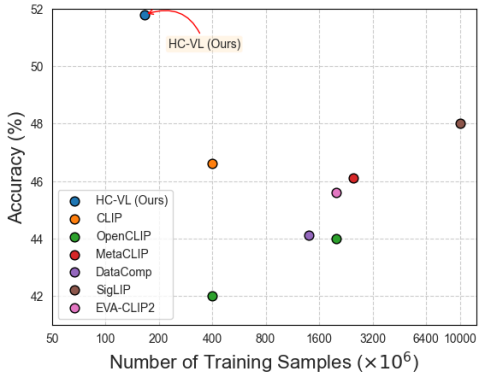

Figure 1: Our domain-specific model **HC-VL** is both more data-efficient and more accurate, outperforming models trained on 10B samples with only 167M. Accuracy aggregated over all 16 action recognition datasets.

were shown by approaches situated at their intersection (Yu et al., 2022; Li et al., 2022a, 2023a). From an architectural and training point of view, our approach closely follows CLIP (Radford et al., 2021) and hence it is part of the contrastive family of models. Such models are trained using a contrastive loss with the goal of learning a joint embedding space for the two modalities (i.e. vision and language). Following CLIP (Radford et al., 2021) subsequent works seek to either improve the loss function(Li et al., 2022a, 2023a; Yu et al., 2022; Yao et al., 2022; Fürst et al., 2021; Yeh et al., 2022; Mu et al., 2021; Li et al., 2023b; Fu et al., 2021; Wang et al., 2022a; Bulat et al., 2024) or improve and increase the model size (Alayrac et al., 2022; Zhai et al., 2022; Wang et al., 2022a) and/or the dataset (Alayrac et al., 2022; Jia et al., 2021; Pham et al., 2021; Yu et al., 2022) used. For example, DeCLIP (Li et al., 2022b) introduces multi-view and nearest-neighbor supervision, FILIP (Yao et al., 2022) applies the contrastive loss in a fine-grained manner while SigLIP (Zhai et al., 2023) replaces the contrastive loss with a sigmoid one. HiCLIP (Geng et al., 2023) introduces hierarchical attention, while (Fini et al., 2023) improves the architecture and the training scheduler. In contrast to the aforementioned works, we do not change the image-based architecture nor the training objective, focusing instead on domain-specific VL pre-training for obtaining highly discriminative and robust representations in a data-, training- and compute-efficient manner.

Dataset construction has very recently attracted the interest of the community that started to transition from huge noisy datasets (e.g.: LAION-400 (Schuhmann et al., 2021), LAION-5B (Schuh-

mann et al., 2022), ALIGN (Jia et al., 2021), Flamingo (Alayrac et al., 2022), SigLIP (Zhai et al., 2023)) to cleaner, higher quality ones (Xu et al., 2023b; Gadre et al., 2024). MetaCLIP (Xu et al., 2023b) attempts to recreate the collection process from the closed source dataset of (Radford et al., 2021) by querying based on Wikipedia articles, bigrams, and WordNet data combined with a filtering and balancing process. DiHT (Radenovic et al., 2023) introduces a rule-based system for retaining higher quality samples, while (Abbas et al., 2023) performs data deduplication. Building on prior works, DataComp (Gadre et al., 2024) provides a unified framework that provides rule-based recipes for constructing datasets ranging in size from 12.8M to 12.8B. However, none of these approaches consider the case of domain-specific training. Moreover, they require manual/hand-crafted seeding of the data filtering, as opposed to our LLM-based process. We significantly outperform all these methods while being up to an order of magnitude more data- and training-efficient.

3 Human-Centric Vision-Language model

In the section, we introduce HC-VL, a new VL model trained with domain priors that significantly improves the in-domain performance without notable out-domain degradations, while requiring (a) no additional data, (b) up to $60\times$ fewer training samples and (c) between $10-100\times$ fewer training iterations. This model builds around our newly introduced LLM-based data filtering and enhancement strategy, presented in Sec. 3.1. The model training itself, and the arch. changes made for fast adaptation to temporal data, are detailed in Sec. 3.2.

3.1 Human-Centric Image-Text dataset

LLM-based action taxonomy generation & filtering: Our domain-specific filtering strategy consists of two steps, the semi-automatic generation of a taxonomy of human activities, and the construction of language queries to query LAION-5B. More specifically, we first iteratively populate a semantic domain of actions by leveraging the pre-trained LLM GPT-3.5 accessed via the OpenAI API. To seed the process, we start by defining six broad categories: `physical`, `communication & cognitive`, `leisure`, `emotional`, `domestic/health & self-care`, and `creative & professional` activities. For each category, we iteratively prompt the LLM to create an initial exhaustive list of coarse-grained instances with the following strategies:
1) List seeding: we prompt the LLM to generate its most representative activities in each category.
2) Alphabetic listing: we prompt the model to generate human-centric activities in alphabetic order, starting with a different letter each time.
3) Synonym listing: for each activity, we prompt the model to generate a limited list of synonyms that describe or are related to it.

For each of the resulting entries, we further add fine-grained categories using the following strategies: *i) Verb composition*: given an activity verb, we prompt the LLM to generate a list of activities that can be composed of it (e.g. *making* $\rightarrow$ *making a cake, making a bed*). *ii) Sub-category listing*: Given an activity, we prompt the LLM to generate a list of related fine-grained activities (e.g. *playing football* $\rightarrow$ *playing American football, playing soccer, playing five-a-side, etc.*). We repeat the process if the generated activities can be further expanded.

To this, we add the activities obtained from Wikipedia's lists of human activities, hobbies, and sports. Finally, duplicates and near duplicates are automatically removed, and then manually filtered to remove unrelated keywords. The final list consists of $\sim$9.7K keywords. The distribution over the 6 categories is shown in Fig. 2(a). Fig. 2(b) shows that our keywords cover the target domain without spanning other undesired domains.

Next, for each keyword, we generate a set of textual queries by using both pre-defined templates (e.g. *a photo of a person [keyword]*) and by prompting GPT-3.5 to generate a set of descriptive and diverse phrases that represent real-world scenarios involving the specified activity. These phrases are then used to query the LAION-5B dataset by retrieving images with an image-text similarity above a threshold (e.g. 0.25) using the OpenAI ViT-L/14 model. The final retrieval yields 279M unique URL-text samples, out of which 256M were successfully downloaded.

Data denoising: Given that only $\sim 40\%$ of the captions are in English, we first start by automatically translating the non-English captions using the NLBB-200 model (Costa-jussà et al., 2022). Secondly, inspired by (Radenovic et al., 2023), we apply a caption complexity filter that removes samples that are not sufficiently complex. We employ spaCy (Honnibal et al., 2020) to build a sentence dependency tree, where the complexity is defined as the maximum number of edges at any particular

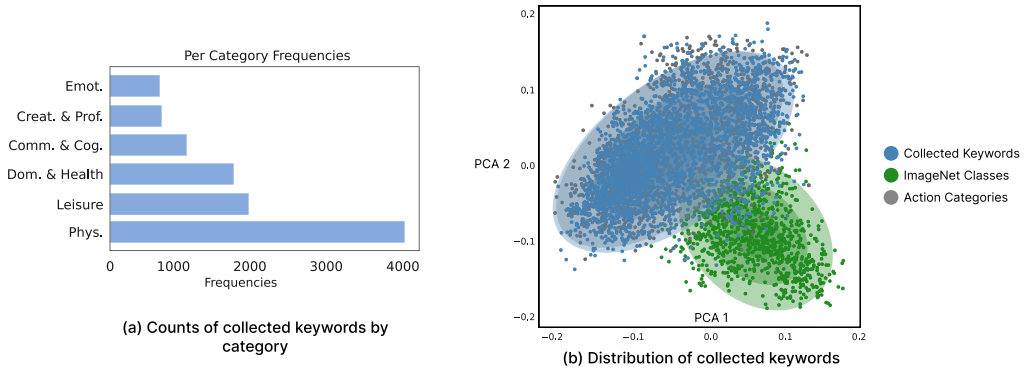

Figure 2: **Collected keywords and their distributions:** (a) shows the no. of keywords across the six pre-defined categories, and (b) shows a 2D PCA plot comparing embeddings of a random subset of 2K from the collected queries, ImageNet classes, and action categories of datasets used in Sec. 4.

node, and only consider samples with a complexity score $>= 2$. Additionally, we filter captions based on toxicity probability (Hanu and Unitary team, 2020), and images based on watermark and NSFW probabilities (Schuhmann et al., 2022), keeping samples with values < 0.9. The application of all these filters results in a dataset of 167M image-text samples from LAION-5B's English, multilingual, and no-language subsets.

Text enhancement by re-captioning: After filtering and especially translating, some samples may present semantic gaps (i.e. the text is irrelevant to the image) or even be incoherent. To alleviate this, we propose to leverage BLIP-2 (Li et al., 2023a) to enrich the crawled captions with an automatically-generated set of alternative captions. We use beam search (Vijayakumar et al., 2016) to generate 5 captions per image and adjust the similarity penalty to encourage the generation of diverse captions that cover different aspects of the image.

We notice however that while the captions produced are generally accurate, they tend to be somewhat generic and vague, often following a template (e.g. *a photo of* ...), likely a bias from BLIP-2's training dataset. Due to this, randomly sampling between the generated captions and the original ones degrades performance. We propose instead a CLIP score-based sampling strategy. Given a pre-trained CLIP model (herein, we use a ViT-B/16), we pre-compute the similarity score between the original caption and the corresponding image. At training time, we sample either a generated or an original caption conditioned on this score.

3.2 HC-VL and HC-VL+ models

HC-VL & image pre-training: Architecturally, our model is identical to CLIP (Radford et al., 2021; Ilharco et al., 2021) to allow direct and fair comparisons. Our HC-VL consists of an image encoder f_θ and a text encoder g_ϕ, instantiated as a ViT (Dosovitskiy et al., 2021) and a transformer (Vaswani et al., 2017) respectively. Similarly, we closely align our pre-training pipeline in terms of losses, augmentation, and initialization with CLIP (Radford et al., 2021), training our model with an image-text contrastive loss applied on the [CLS] token of the image encoder and the [EOS] of the text encoder. The main difference to CLIP is that we train our models on our HC-IT dataset. This has a crucial effect on the model's performance. Such result can be attributed to the properties of HC-IT: In addition to in-domain coverage shown in Fig. 2, our domain-specific training induces an implicit hard mining approach. In support of this, Fig. 3 shows the cosine similarities between image-text pairs of a randomly sampled batch sampled from HC-IT and LAION. Compared to LAION, our batch contains samples that are semantically closer, and hence harder for the model to differentiate. This is particularly noticeable in the high-score region (> 0.25) where our model is trained from thousands of such

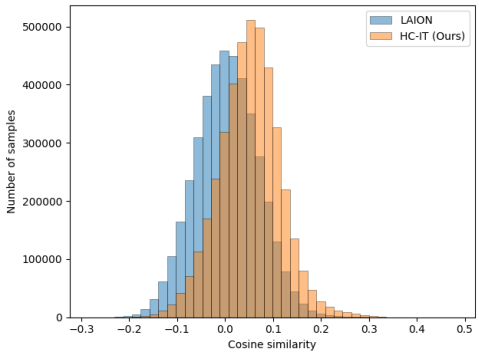

Figure 3: Cosine similarities between the image-text pairs of a randomly selected batch from LAION and HC-IT dataset computed using a pretrained CLIP (ViT-L/14) model. Notice that: (a) the sim. scores are higher overall for our dataset, and (b) a bigger number of pairs score higher than 0.25 in our case. This suggests that batches we form contain more hard negatives.

high-scoring negative pairs.

HC-VL+ by further pre-training on video: One option for further pre-training on video is to use the WebVid dataset (Bain et al., 2021). However, we found it's quality to be lower than our HC-IT dataset, leading to degraded performance. We instead propose to use the much smaller but higher quality Spoken Moments in Time (SMiT) (Monfort et al., 2021). To avoid overfitting and catastrophic forgetting, and inspired by (Wang et al., 2021a; Ni et al., 2022; Pan et al., 2022), we introduce minimal changes to our image HC-VL model to incorporate temporal information *while maintaining the zero-shot capabilities of our image model.*

Let $\mathbf{v}$ be a video with T frames $\{\mathbf{x}_i\}_{i=1:T}$. Let $\mathtt{A}_I(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathtt{softmax}(\frac{\mathbf{QK^T}}{\sqrt{L}})\mathbf{V}$ be the attention operation, where $\mathbf{K}$, $\mathbf{Q}$ and $\mathbf{V}$ are the output of projections of $\mathbf{W_q}$, $\mathbf{W_k}$ and $\mathbf{W_v}$ respectively, and L the number of tokens per frame. Our attention is computed as the addition of two attention branches. The first consists of per-frame spatial attentions $\mathtt{A}_I(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i)$. The second one computes the cross attention between the T class tokens and spatial tokens for frame i. More specifically, a learnable temporal embedding per frame is added to the [CLS] tokens $\mathbf{x}_i^{cls}$ and then projected using $\mathbf{W_k}$ and $\mathbf{W_v}$ to obtain $\mathbf{K}^c \in \mathbb{R}^{T\times d}$ and $\mathbf{V}^c \in \mathbb{R}^{T\times d}$. Then the cross attention $\mathtt{A}_I(\mathbf{Q}_i,\mathbf{K}^c,\mathbf{V}^c)$ is computed (i.e. spatial tokens of frame i attend to all class tokens). Both branches are finally combined as $\mathtt{A}_I(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i) + s\cdot\mathtt{A}_I(\mathbf{Q}_i,\mathbf{K}^c,\mathbf{V}^c)$, with s being a learnable scaling factor set initially to zero to add stability and avoid catastrophic forgetting and the second term helping diffuse temporal information through the [CLS] tokens into spatial tokens. At the end of the network, we add a temporal attention layer h_μ that performs temporal attention between the T [CLS] tokens $\mathbf{x}_i^{cls}$. The final video feature representation is $\frac{1}{T}\sum_i f_\theta(\mathbf{x}_i)$ + $h_\mu([f_\theta(\mathbf{x}_1),\ldots,f_\theta(\mathbf{x}_T)])$. Note that adding the pooled feature to the output of the temporal attention ensures that our model stays "close" to our image model. All parameters are frozen except the newly-introduced ones (s and h_μ). We coin such a model further pre-trained on video as **HC-VL+**.

Efficient pre-training: The large-scale nature of VL pre-training poses significant computational challenges. To make the training of larger models more feasible on our available hardware, we follow a two-step strategy: firstly, we train a model using larger patches for tokenization, which results in using fewer tokens. Then, we fine-tune it for 1/10 of the epochs using the target desired patch size. When initializing the model from the previous stage, in addition to the standard bilinear interpolation of the positional embeddings, we also propose to resize the kernels learned by the convolutional layer processing the patches. We found this to work well, and it is the strategy employed for training our ViT-B/16 variant, fine-tuning from a ViT-B/32.

4 Comparison with state-of-the-art

To showcase the effectiveness of the proposed domain-specific pre-training, we evaluate our models for image retrieval, video retrieval and action recognition as they are tasks well aligned with the domain of human activities. See the appendix for pre-training details and additional evaluations.

Specifically, we compare HC-VL and HC-VL+ with all the equivalently sized models available as part of the OpenCLIP repo (Ilharco et al., 2021), noting that not all methods offer both the B/32 and B/16 variants. Moreover, we emphasize that all models we compare with are trained using longer schedulers on a significantly larger number of samples, ranging from 400M for OpenCLIP (LAION-400) (Schuhmann et al., 2021) to 10B for SigLIP (Zhai et al., 2023) (see Fig. 1). In comparison, our models are trained only on 167M samples.

4.1 Zero-shot Image & video retrieval

Following (Zhai et al., 2023; Sun et al., 2023) we evaluate our models, HC-VL and HC-VL+, in a zero-shot manner on Flickr30k and MS-COCO for image retrieval, and respectively, MSRVTT (Xu et al., 2016a), MSVD (Chen and Dolan, 2011) and DiDemo (Anne Hendricks et al., 2017) for video retrieval. We note that under this setting (i.e. zero-shot), HC-VL and HC-VL+ are applied directly to the downstream tasks of image and video retrieval without any fully-supervised training. For image retrieval, we only report results for HC-VL, as HC-VL+ is a temporal model and operates on videos. For video retrieval, HC-VL is applied to video by means of simple temporal pooling (i.e. $\frac{1}{T}\sum_i f_\theta(\mathbf{x}_i)$) while HC-VL+ is applied as is.

For image retrieval, as the results from Table 1 show, for most cases, our model significantly outperforms all prior methods, despite some of them using improved architectures (e.g. (Sun et al., 2023)) and, in all cases, significantly more images (up to 10B (Zhai et al., 2023)) than our compar-

atively small 167M HC-IT dataset. Importantly, we outperform LAION-400/2B/5B models with our HC-IT (derived from LAION) dataset. As Table 2 shows, similar improvements are reported for video retrieval too, with our image (HC-VL) model matching or outperforming the other CLIP variants, and HC-VL+ showing significant further gains.

4.2 Zero-shot action recognition

To evaluate HC-VL&HC-VL+ on zero-shot action recognition, we firstly construct a new benchmark formed of 16 datasets: UCF-101 (Soomro et al., 2012), HMDB-51 (Kuehne et al., 2011), Daly (Weinzaepfel et al., 2016), Kinetics-400 (Kay et al., 2017), Kinetics-220 (Chen and Huang, 2021) (a subset of Kinetics-600 (Carreira et al., 2018) that includes only the classes not present in Kinetics-400), Kinetics-700 (Carreira et al., 2019), MiT (Monfort et al., 2019), HAA500 (Chung et al., 2021), HACS (Zhao et al., 2019), Hollywood2 (Marszalek et al., 2009), Olympic Sports (Niebles et al., 2010), UCF-50 (Reddy and Shah, 2013), UCF Sports (Soomro and Zamir, 2015), AVA (Gu et al., 2018), Something-Else (Materzynska et al., 2020) and Charades (Sigurdsson et al., 2018). For Kinetics-220, UCF-101, and HMDB-51, the results reported are the average across the 3 splits introduced by their authors. We use 8 frames for all datasets, but for Charades where we use 32 due to its longer videos. To the best of our knowledge, this is the most extensive zero-shot action recognition benchmark covering both multi- and single-label settings, with a high variability of the numbers of classes (up to 700), capturing conditions, length, and subdomains.

Under the zero-shot setting, HC-VL & HC-VL+ are applied directly to the downstream datasets without any supervised training, while HC-VL is applied to video by means of temporal pooling.

As the results from Tab. 3 show, *our models significantly outperform all CLIP variants,* emphasizing the importance of our domain-specific pre-training. This is more evident when considering the comparisons with OpenCLIP models which are trained on LAION, out of which we derived our HC-IT dataset. Another important conclusion is that our *ViT-B/32 models outperform CLIP ViT-B/16 models* by a large margin. Finally, it is worth noting that the impact of our additional video pre-training, resulting in our HC-VL+ model, is in many cases quite significant.

In addition to our newly constructed benchmark, following (Ni et al., 2022), Tab. 4 reports zero-shot action recognition results on UCF-101 and HMDB-51 where we compare with several state-of-the-art methods. We note that many of these results are not directly comparable to ours as most methods are trained on smaller datasets (mostly on Kinetics-400) and only a few of them are based on CLIP pre-training (i.e. (Wang et al., 2021a; Ni et al., 2022)). Despite this, as our results show, our models significantly outperform all other methods, setting a new state-of-the-art.

4.3 Few-shot action recognition

Finally, both HC-VL and HC-VL+ can be used to replace CLIP in all CLIP video adaptation methods which perform downstream fine-tuning (Wang et al., 2021a; Ju et al., 2022; Ni et al., 2022; Lin et al., 2022b). This is an important feature of our models, as they can be seamlessly combined with recent advancements in action recognition. Hence, aligning with the setting introduced in (Ni et al., 2022), we also report results for few-shot action recognition. We opt to fuse our model with the state-of-the-art architecture of (Ni et al., 2022). Integrating HC-VL with (Ni et al., 2022) is straightforward. To integrate HC-VL+, we simply insert within our model the following components: the video-specific prompt generator, the cross-frame interaction mechanism, and the multi-frame integration transformer. For the latter, we replace the cross-frame interaction mechanism with our temporal attention layer. Overall, we obtain two models, coined **X-HC-VL** and **X-HC-VL+**.

For the few-shot setting, we evaluate our X-HC-VL and X-HC-VL+ on 3 datasets: UCF-101, HMDB-51, and Kinetics-400 for 2, 4, 8 and 16-shot. As Tab. 5 shows, both of our variants, X-HC-VL and X-HC-VL+, outperform the previous state-of-the-art models X-CLIP and X-Florence (which benefits from pre-training on FlD-900M (Yuan et al., 2021)) by a large margin on all datasets.

5 Ablation studies

Effect of data quantity: Our models are trained on significantly fewer samples (167M) compared with the current state-of-the-art VL models, commonly trained on 400M (Schuhmann et al., 2021), 2B (Schuhmann et al., 2022) or even 6.6B (Pham et al., 2021) samples. Going one step further, herein, we explore the performance of our model in even lower data regimes. In Tab. 6, we report

Method	Flicrk30k		MS-COCO	
	T2I	I2T	T2I	I2T
ViT-B/32 architecture				
OpenCLIP (400M) (Ilharco et al., 2021)	59.7/90.3	78.1/96.6	34.2/70.6	52.3/84.3
OpenCLIP (2B) (Ilharco et al., 2021)	66.8/93.1	84.1/98.3	39.3/75.6	56.3/87.1
OpenCLIP (5B) (Ilharco et al., 2021)	64.5/91.7	82.7/97.8	37.8/73.5	53.5/86.4
CLIP (Radford et al., 2021)	58.8/90.0	78.9/98.2	30.4/66.9	50.1/83.5
MetaCLIP (Xu et al., 2023b)	65.2/92.7	80.8/97.3	38.1/74.3	55.2/86.5
CoCa (Yu et al., 2022)	63.4/91.4	81.6/97.4	36.2/71.8	54.6/85.6
DataComp (Gadre et al., 2024)	61.1/90.9	79.0/96.2	37.1/72.7	53.5/86.0
HC-VL (Ours)	**74.2/95.7**	**90.3/99.4**	**45.3/80.4**	**62.1/91.0**
ViT-B/16 architecture				
OpenCLIP (400M) (Ilharco et al., 2021)	65.7/93.0	83.5/98.5	38.3/73.9	55.4/86.9
OpenCLIP (2B) (Ilharco et al., 2021)	69.8/94.6	86.3/99.4	42.3/77.1	59.4/88.6
CLIP (Radford et al., 2021)	62.1/91.9	82.2/99.0	33.1/69.0	52.4/84.6
MetaCLIP (Xu et al., 2023b)	70.7/94.5	85.5/98.9	41.3/77.0	59.4/87.8
EVA-CLIP2 (Sun et al., 2023)	71.5/94.7	86.0/98.8	42.2/76.3	58.7/88.1
SigLIP (Zhai et al., 2023)	74.7/95.6	89.1/99.3	47.8/81.0	**65.7**/91.3
DataComp (Gadre et al., 2024)	67.6/93.0	85.1/98.4	40.2/75.6	57.4/88.3
HC-VL (Ours)	**77.8/97.4**	**92.6/99.9**	**48.6/83.1**	64.7/**92.4**

Table 1: **Zero-shot image retrieval results** in terms of R@1/R@10 retrieval accuracy on Flickr30k and MS-COCO.

Method	MSRVTT		MSVD		DiDemo	
	T2V	V2T	T2V	V2T	T2V	V2T
ViT-B/32 architecture						
OpenCLIP (400M) (Ilharco et al., 2021)	29.6/62.2	23.7/56.1	35.6/72.3	48.5/84.4	25.0/59.1	22.0/54.9
OpenCLIP (2B) (Ilharco et al., 2021)	34.9/67.8	28.1/60.6	40.5/77.3	56.4/88.0	27.8/65.1	24.8/59.3
OpenCLIP (5B) (Ilharco et al., 2021)	33.4/66.1	28.0/60.4	40.6/77.1	55.2/86.3	27.4/64.0	24.6/58.9
CLIP (Radford et al., 2021)	30.3/65.0	26.3/62.3	34.8/73.4	57.8/90.8	26.9/63.9	19.4/55.3
MetaCLIP (Xu et al., 2023b)	33.7/66.6	30.1/64.7	37.8/75.9	51.9/85.9	30.1/64.7	20.2/55.3
CoCa (Yu et al., 2022)	33.4/67.1	25.2/57.4	38.8/76.3	52.6/87.0	26.1/57.3	24.7/60.0
DataComp (Gadre et al., 2024)	31.5/63.9	24.2/54.5	39.2/76.5	54.6/88.9	26.9/60.4	23.3/56.8
HC-VL (Ours)	29.7/68.6	28.3/60.6	44.6/82.3	64.1/92.0	30.2/63.0	30.2/65.4
HC-VL+ (Ours)*	**39.0/75.7**	**36.9/72.3**	**50.1/85.6**	**67.8/94.3**	**35.4/70.0**	**32.1/66.8**
ViT-B/16 architecture						
OpenCLIP (400M) (Ilharco et al., 2021)	32.5/66.9	24.4/57.7	39.8/77.3	55.4/89.4	27.9/62.4	24.6/60.5
OpenCLIP (2B) (Ilharco et al., 2021)	36.7/69.7	27.7/58.3	41.5/77.8	55.3/90.3	31.2/65.9	27.1/61.5
CLIP (Radford et al., 2021)	33.4/65.6	30.5/64.6	39.0/76.2	62.6/92.9	30.7/63.5	23.3/54.4
MetaCLIP (Xu et al., 2023b)	36.0/68.2	29.6/61.1	43.5/81.1	62.1/91.7	30.6/65.3	26.5/60.1
EVA-CLIP2 (Sun et al., 2023)	35.1/69.7	27.3/58.3	44.2/81.7	61.0/91.7	35.2/67.7	30.1/63.9
SigLIP (Zhai et al., 2023)	34.6/67.0	30.9/63.2	47.0/82.1	64.6/94.1	30.7/65.3	27.0/60.5
DataComp (Gadre et al., 2024)	34.2/64.8	26.4/57.7	42.5/79.3	56.9/91.0	30.3/62.9	27.7/60.8
HC-VL (Ours)	36.0/69.9	30.0/62.1	45.6/84.3	66.1/94.3	36.1/70.0	31.2/66.2
HC-VL+ (Ours)*	**40.8/74.0**	**37.5/74.1**	**52.0/86.9**	**68.4/95.1**	**36.7/70.8**	**33.3/67.2**

Table 2: **Zero-shot video retrieval results** in terms of R@1/R@10 retrieval accuracy on MSRVTT, MSVD and DiDemo. * indicates temporal modeling.

results for zero-shot classification on UCF-101 and HMDB-51 for a ViT-B/32 HC-VL trained with 69M, 135M and 167M samples. In all cases, the model was trained for the same number of seen samples (i.e. 4B). We see that reducing the dataset size has a considerable impact on accuracy.

Effect of LLM-based domain construction: At the core of HC-IT is the proposed LLM-based domain construction and data filtering method. To further showcase its effect, we compare it with an equal-sized dataset of 167M image-text pairs sampled by: filtering based on CLIP's image-text

	Dataset	OpenCLIP (Ilharco et al., 2021) 400M	OpenCLIP (Ilharco et al., 2021) 2B	OpenCLIP (Ilharco et al., 2021) 5B	MetaCLIP (Xu et al., 2023b)	DataComp (Gadre et al., 2024)	Coca (Yu et al., 2022)	CLIP (Radford et al., 2021)	HC-VL (Ours)	HC-VL+ (Ours)
ViT-B/32	UCF-101	62.7 (87.1)	69.3 (92.1)	64.8 (91.1)	69.4 (91.1)	67.8 (89.2)	67.4 (90.8)	67.6 (91.3)	79.5 (96.9)	**81.6 (97.0)**
	HMDB-51	30.2 (55.6)	34.9 (63.0)	34.5 (63.2)	33.2 (62.0)	31.7 (61.7)	34.2 (64.2)	39.5 (67.3)	46.5 (74.2)	**51.0 (77.9)**
	Kinetics-400	40.3 (65.6)	47.6 (73.3)	45.9 (71.6)	46.9 (72.2)	47.2 (71.5)	45.4 (70.9)	48.0 (74.7)	55.6 (79.7)	**58.8 (82.6)**
	Kinetics-220	36.7 (59.6)	42.7 (68.4)	42.1 (68.6)	42.9 (69.0)	43.6 (68.2)	40.6 (66.2)	43.2 (69.9)	51.1 (77.1)	**53.2 (79.3)**
	Kinetics-700	29.4 (52.2)	35.0 (59.9)	33.8 (58.0)	35.7 (60.3)	35.5 (58.4)	33.1 (56.7)	36.5 (61.7)	42.5 (67.6)	**45.5 (71.4)**
	Daly	75.3 (**98.6**)	74.7 (96.6)	70.5 (97.3)	69.9 (95.9)	71.9 (96.6)	72.6 (97.2)	71.9 (97.9)	76.7 (**98.6**)	**81.5** (97.8)
	HAA500	30.0 (57.7)	37.1 (65.6)	34.8 (63.1)	34.9 (64.1)	36.6 (63.6)	33.7 (62.7)	37.4 (66.5)	**48.9** (77.7)	48.7 (**78.1**)
	HACS	57.3 (84.3)	64.9 (89.4)	63.9 (88.4)	64.7 (88.6)	64.6 (88.8)	63.6 (89.0)	64.4 (89.8)	75.6 (94.7)	**77.1 (95.8)**
	Hollywood2*	32.9	41.3	42.6	42.1	36.9	40.3	42.8	48.2	**54.5**
	Olympic Sports	46.3 (86.6)	50.7 (90.3)	48.5 (83.6)	46.3 (84.3)	44.8 (86.6)	44.8 (83.6)	47.8 (88.1)	50.0 (90.3)	**54.5 (94.0)**
	MiT	15.1 (32.4)	17.9 (37.5)	17.2 (36.4)	18.0 (37.9)	17.5 (36.2)	16.7 (35.3)	17.9 (37.7)	21.1 (43.6)	**22.8 (46.6)**
	AVA*	9.1	9.3	9.5	11.3	9.3	9.5	11.3	13.6	**17.1**
	Something-Else	8.6 (26.5)	10.9 (31.6)	10.3 (30.6)	10.7 (30.6)	10.4 (30.5)	9.7 (29.3)	10.0 (29.3)	11.0 (32.0)	**12.7/36.4**
	Charades*	14.3	17.8	17.1	16.9	16.8	17.6	19.1	22.0	**23.5**
	UCF-50	71.2 (89.8)	78.8 (96.5)	73.8 (93.3)	76.3 (93.1)	77.2 (94.1)	77.1 (94.3)	78.7 (95.2)	88.7 (**98.8**)	**90.2 (98.8)**
	UCF Sports	47.5 (83.6)	49.8 (83.6)	47.6 (85.3)	41.0 (85.3)	46.7 (85.2)	54.1 (83.6)	50.8 (88.5)	**55.8** (93.4)	55.1 (**93.5**)
	Average	37.9 (67.7)	42.7 (72.9)	41.1 (71.6)	41.3 (71.9)	41.1 (71.6)	41.3 (71.8)	42.9 (72.7)	49.2 (76.5)	**51.7 (80.7)**
	Dataset	OpenCLIP (Ilharco et al., 2021) 400M	OpenCLIP (Ilharco et al., 2021) 2B	MetaCLIP (Xu et al., 2023b)	DataComp (Gadre et al., 2024)	SigLIP (Zhai et al., 2023)	EVA-CLIP2 (Sun et al., 2023)	CLIP (Radford et al., 2021)	HC-VL (Ours)	HC-VL+ (Ours)
ViT-B/16	UCF-101	70.4 (91.5)	71.9 (92.4)	76.3 (95.3)	70.6 (91.0)	76.0 (94.1)	72.1 (93.0)	71.3 (93.9)	81.2 (97.6)	**84.0 (97.8)**
	HMDB-51	35.1 (63.5)	36.4 (65.2)	39.3 (68.8)	32.8 (63.4)	43.0 (68.2)	37.0 (63.3)	43.8 (70.6)	49.1 (76.3)	**52.5 (78.8)**
	Kinetics-400	47.2 (72.8)	49.3 (74.4)	53.5 (78.6)	51.1 (75.3)	54.9 (78.9)	52.2 (77.3)	53.5 (78.8)	59.5 (83.1)	**62.5 (85.6)**
	Kinetics-220	42.7 (67.3)	43.7 (69.7)	49.6 (75.1)	46.9 (71.8)	51.9 (77.3)	48.5 (73.4)	47.3 (73.4)	54.2 (80.7)	**57.6 (83.3)**
	Kinetics-700	35.2 (59.1)	36.4 (60.7)	41.1 (66.5)	39.0 (62.5)	42.8 (66.7)	40.7 (65.3)	41.1 (66.5)	46.1 (71.3)	**49.2 (75.4)**
	Daly	68.5 (97.3)	76.0 (96.6)	74.6 (97.3)	76.7 (95.9)	77.4 (99.3)	70.6 (98.6)	75.3 (98.6)	84.2 (99.3)	**87.7 (99.8)**
	HAA500	35.7 (64.8)	39.0 (69.6)	41.2 (72.3)	39.0 (67.6)	48.8 (77.4)	41.5 (70.6)	42.5 (70.9)	51.1 (81.0)	**53.6 (81.7)**
	HACS	64.3 (89.0)	68.0 (91.4)	70.5 (92.1)	68.3 (90.7)	72.7 (93.1)	68.9 (91.7)	69.9 (92.2)	77.7 (95.8)	**80.7 (96.6)**
	Hollywood2*	41.3	42.1	46.6	41.9	43.7	45.1	47.6	50.8	**57.0**
	Olympic Sports	45.5 (89.5)	47.8 (85.8)	45.5 (88.1)	44.8 (83.6)	52.3 (89.5)	48.5 (89.5)	50.8 (90.3)	54.4 (92.5)	**58.2 (94.8)**
	MiT	17.8 (37.1)	19.3 (39.2)	20.7 (42.6)	19.9 (39.5)	22.2 (43.9)	21.1 (42.8)	20.5 (42.4)	23.9 (47.8)	**25.6 (50.9)**
	AVA*	9.6	9.9	11.3	10.0	11.6	11.7	12.0	14.0	**17.8**
	Something-Else	10.1 (30.8)	11.6 (33.1)	10.6 (31.3)	13.0 (34.9)	**14.1 (38.3)**	12.6 (34.8)	10.8 (30.8)	12.0 (34.5)	13.5(36.6)
	Charades*	17.5	19.3	21.6	19.3	21.8	21.9	21.0	24.2	**24.5**
	UCF-50	78.5 (94.2)	80.1 (93.4)	83.7 (97.3)	79.2 (94.7)	82.0 (95.0)	78.6 (95.6)	81.5 (96.1)	90.8 (99.2)	**91.3 (99.3)**
	UCF Sports	52.5 (85.3)	52.4 (86.9)	50.8 (90.7)	52.5 (88.5)	52.4 (**93.4**)	52.4 (86.8)	**57.4** (90.1)	55.7 (91.8)	55.8 (88.5)
	Average	42.0 (72.5)	44.0 (73.7)	46.1 (76.6)	44.1 (73.8)	48.0 (78.0)	45.2 (75.6)	46.6 (76.5)	51.8 (80.8)	**54.5 (82.2)**

Table 3: **Zero-shot classification results** across a suite of action recognition datasets in terms of Top-1 (%) and Top-5 (%) accuracy (shown in parentheses). * - results reported in terms of mAP. HC-VL+ includes temporal modeling. Further comparisons against methods with temporal modeling are shown in Tab. 4.

Method	HMDB-51	UCF-101
ER-ZSAR (Chen and Huang, 2021)	35.3 ± 4.6	51.8 ± 2.9
MUFI (Qiu et al., 2021)	31.0	60.9
ActionCLIP (Wang et al., 2021a)	40.8 ± 5.4	58.3 ± 3.4
ClipBert (Lei et al., 2021)	21.4 ± 1.0	27.8 ± 0.8
Frozen (Bain et al., 2021)	27.8 ± 0.3	45.9 ± 1.3
ViSET-96 (Doshi and Yilmaz, 2022)	40.2	68.3
BridgeFormer (Ge et al., 2022)	37.7 ± 1.2	53.1 ± 1.4
CLIP (Radford et al., 2021)	43.8	70.6
AURL (Pu et al., 2022)	40.4	60.9
ResT_101 (Lin et al., 2022a)	41.1 ± 3.7	58.7 ± 3.3
X-CLIP (Ni et al., 2022)	44.6 ± 5.2	72 ± 2.3
X-Florence (Ni et al., 2022)	48.4 ± 4.9	73.2 ± 4.2
HC-VL (Ours)	49.1	81.2
HC-VL+ (Ours)	**52.5**	**84.0**

Table 4: **Zero-shot classification results** on HMDB-51 and UCF-101 in terms of Top-1 (%) accuracy.

similarity scores and simultaneously applying all the filtering steps described (e.g.: grammar, NSFW removal etc.), but without applying our proposed LLM-based approach. As the results from Table 7 show (1st and 2nd row), our approach significantly outperforms the CLIP filtering baseline, further highlighting the importance of the proposed LLM-based domain construction.

Out-of-domain generalization: An important aspect of our approach is its ability to concomitantly improve on the domain of interest while largely preserving the generalist abilities of the model. To showcase this, we devise a series of experiments measuring this, adding to the benchmark, as an out-domain dataset ImageNet (Deng et al., 2009) (Fig. 2 confirms its out-of-domain nature). Particularly, in Table 7, we report results by direct training on our 167M-large filtered subset, the result of applying the NLP-based approach presented in Sec. 3 (1st row); by selecting from LAION an equally sized subset of 167M, without using our approach (2nd row); using LAION-400 only (3rd row); by expanding our dataset with additional samples from LAION-400M (4th row), LAION-2B (5th row) and with the entirety of LAION-400M (6th row). Analyzing the results, we can make the following observations: (1) Adding extra data to our filtered set, from either dataset (i.e. LAION-400/2B), decreases the overall performance on in-domain data - this suggests that our selection process encourages the formation of natural hard negative pairs that help to drive the learning process, (2) Our approach largely preserver the out-of-domain performance

Arch.	Method	HMDB-51				UCF-101				Kinetics-400			
		2	4	8	16	2	4	8	16	2	4	8	16
ViT-B/32	X-CLIP	48.1	54.7	57.8	61.4	76.3	81.0	85.5	88.7	52.4	54.2	57.7	59.2
	X-OpenCLIP	39.8	49.3	53.8	58.2	76.9	81.2	86.2	89.5	52.0	53.9	57.6	59.0
	X-OpenCLIP (400M)	36.8	43.5	50.1	54.6	73.6	77.8	82.5	86.1	45.2	47.3	50.6	53.1
	X-HC-VL (Ours)	51.0	57.4	59.7	63.8	85.3	87.9	90.4	92.1	60.7	62.3	64.2	66.5
	X-HC-VL+ (Ours)	**53.1**	**58.7**	**60.7**	**64.5**	**86.2**	**88.8**	**91.3**	**93.0**	**61.6**	**63.0**	**64.9**	**67.0**
ViT-B/16	X-Florence* (Ni et al., 2022)	51.6	57.8	64.1	64.2	84.0	88.5	92.5	94.8	-	-	-	-
	X-CLIP (Ni et al., 2022)	53.0	57.3	62.8	64.0	76.4	83.4	88.3	91.4	59.2	60.8	62.2	63.4
	X-OpenCLIP	47.7	53.9	58.2	60.1	77.0	83.7	88.6	92.4	56.1	57.9	58.6	60.1
	X-OpenCLIP (400M)	40.1	47.7	53.7	57.5	73.3	80.1	83.5	87.9	53.0	55.5	57.1	58.2
	X-HC-VL (Ours)	57.8	62.1	64.6	66.7	88.6	90.8	93.1	95.2	64.9	66.0	67.9	69.9
	X-HC-VL+ (Ours)	**58.9**	**63.0**	**65.4**	**67.3**	**89.4**	**91.5**	**93.9**	**95.6**	**65.9**	**66.8**	**68.7**	**70.5**

Table 5: **Few-shot classification results** on HMDB-51, UCF-101 and Kinetics-400 in terms of Top-1 (%) accuracy for 2/4/8/16-shot. * - indicates results taken from (Ni et al., 2022).

num. samples	UCF-101 Top-1	UCF-101 Top-5	HMDB-51 Top-1	HMDB-51 Top-5
69M	71.7	92.8	41.4	69.5
135M	78.1	95.2	43.0	69.8
167M	79.4	96.4	45.4	71.6

Table 6: **Training dataset size vs accuracy**: Zero-shot results on UCF-101 and HMDB-51 with a ViT-B/32 for different numbers of image-text pairs, run for a fixed number of iterations.

when adjusting for the dataset size: 167M selected using our LLM-based approach vs selecting them without it, results in a similar performance on ImageNet (57.8 vs 56.2). This is perhaps somewhat surprising at first glance, given the distribution of the textual data from Fig. 2. However, in practice, the class names, although very rare, tend to be present at least once. This suffices in driving the model toward learning some concepts associated with them, that is in part due to the nature of the implicit hard negative mining that encourages the model to pay attention to finer-grained details (such as the object name). All in all, we show that our approach is faster to train, less data hungry, and capable of maximizing in-domain performance while largely retaining the out-of-domain one.

# Data	UCF-101	HMDB-51	K400	Imagenet
HC (Ours)	79.5	46.5	55.6	57.8
167M from LAION	67.8	32.4	47.2	56.2
LAION-400M	62.7	30.2	40.3	60.2
HC + 0.5× LAION-400M	77.0	40.7	54.4	62.5
HC + 0.1× LAION-2B	77.2	40.6	54.6	62.4
HC + LAION-400M	76.4	39.1	54.0	66.1

Table 7: **Out-of-domain generalization** and effect of the proposed LLM-based data construction.

6 Conclusions

In this work, we introduced a new take on VL pre-training, that aims to take advantage of in-domain priors without degrading the generalizability of the model, herein tested for the domain of human activities. To this end, we proposed a new LLM-based method for automatic data design which enables the construction of a vast human activity-specific dataset by retrieving, filtering, and re-captioning related samples from LAION, on which we train our VL model, HC-VL. Furthermore, we introduced a set of new architectural changes that allow for the fast addition of temporal modeling, without compromising the generalization capabilities of our image-based model. Our models achieve state-of-the-art performance across a large suite of datasets for zero-shot image recognition, video recognition and zero-shot action recognition. Notably, this is achieved without using additional data and while being up to $60\times$ more data efficient and $10-100\times$ faster to train due to requiring fewer iterations.

Limitations

Due to the nature of the automatic filtering process and the dataset size, manually checking for potential bias issues is not feasible. Moreover, from a technical standpoint the quality of the data, and as a consequence, of the model, will depend on that of the pre-trained models used to filter and augment the data (i.e. CLIP, BLIP). Different models may lead to different subsets being covered, with varying degrees of accuracy. As with all models trained on webly collected data, without manual filtering, we strongly recommend checking the models and the data carefully before deploying them. More generally, our approach is subject to the same considerations as the ones enunciated by the LAION-5B authors, and we encourage the readers to check (Schuhmann et al., 2022) for a more in-depth discussion.

References

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. 2023. SemDeDup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*.

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

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *IEEE International Conference on Computer Vision*.

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

Adrian Bulat, Yassine Ouali, and Georgios Tzimiropoulos. 2024. Fff: Fixing flawed foundations in contrastive pre-training results in very strong vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14172–14182.

Adrian Bulat, Enrique Sanchez, Brais Martinez, and Georgios Tzimiropoulos. 2023. Regen: A good generative zero-shot video classifier should be rewarded. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13523–13533.

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

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A short note on the Kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Annual meeting of the association for computational linguistics: human language technologies*.

Shizhe Chen and Dong Huang. 2021. Elaborative rehearsal for zero-shot action recognition. In *IEEE International Conference on Computer Vision*.

Jihoon Chung, Cheng-hsin Wuu, Hsuan-ru Yang, Yu-Wing Tai, and Chi-Keung Tang. 2021. HAA500: Human-centric atomic action dataset with curated videos. In *IEEE International Conference on Computer Vision*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

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

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

Enrico Fini, Pietro Astolfi, Adriana Romero-Soriano, Jakob Verbeek, and Michal Drozdzal. 2023. Improved baselines for vision-language pre-training. *arXiv preprint arXiv:2305.08675*.

Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. 2021. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*.

Andreas Fürst, Elisabeth Rumetshofer, Viet Tran, Hubert Ramsauer, Fei Tang, Johannes Lehner, David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto-Nemling, and Sepp Hochreiter. 2021. CLOOB: Modern Hopfield networks with InfoLOOB outperform CLIP. *arXiv preprint arXiv:2110.11316*.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2024. DataComp: In search of the next generation of multimodal datasets. *Advances on Neural Information Processing Systems*, 36.

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

Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. 2023. Hiclip: Contrastive language-image pretraining with hierarchy-aware attention. *arXiv preprint arXiv:2303.02995*.

Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. OpenCLIP.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*.

Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. 2022. Prompting visual-language models for efficient video understanding. *European Conference on Computer Vision*.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

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

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

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

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2022b. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *International Conference on Learning Representations*.

Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. 2023b. Scaling language-image pre-training via masking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

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

Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2022b. Frozen CLIP models are efficient video learners. *European Conference on Computer Vision*.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. 2009. Actions in context. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. 2020. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2017. Mixed precision training. *arXiv preprint arXiv:1710.03740*.

Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. 2019. Moments in time dataset: one million videos for event understanding. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, pages 1–8.

Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. 2021. Spoken moments: Learning joint audio-visual representations from video descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 2021. SLIP: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*.

Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. 2010. Modeling temporal structure of decomposable motion segments for activity classification. In *European Conference on Computer Vision*.

Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. 2022. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances on Neural Information Processing Systems*.

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

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. 2021. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*.

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

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

Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. 2023. Filtering, distillation, and hard negatives for vision-language pre-training. *arXiv preprint arXiv:2301.02280*.

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

Kishore K Reddy and Mubarak Shah. 2013. Recognizing 50 human action categories of web videos. *Machine vision and applications*, 24(5):971–981.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. 2018. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*.

Khurram Soomro and Amir R Zamir. 2015. Action recognition in realistic sports videos. In *Computer vision in sports*, pages 181–208. Springer.

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

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

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

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2022a. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*.

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

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer.

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

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021b. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. 2016. Human action localization with sparse spatial supervision. *arXiv preprint arXiv:1605.05197*.

Hu Xu, Saining Xie, Po-Yao Huang, Licheng Yu, Russell Howes, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2023a. CiT: Curation in training for effective vision-language data. *arXiv preprint arXiv:2301.02241*.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2023b. Demystifying clip data. *arXiv preprint arXiv:2309.16671*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016a. MSR-VTT: A large video description dataset for bridging video and language. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016b. MSR-VTT: A large video description dataset for bridging video and language. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2022. FILIP: Fine-grained interactive language-image pre-training. *International Conference on Learning Representations*.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. 2022. Decoupled contrastive learning. *European Conference on Computer Vision*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

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

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Hang Zhao, Zhicheng Yan, Lorenzo Torresani, and Antonio Torralba. 2019. HACS: Human action clips and segments dataset for recognition and temporal localization. *arXiv preprint arXiv:1712.09374*.

A Appendix

A.1 Additional ablation studies

Effect of BLIP2 re-captioning: Tab. 8 shows the effect of using additional BLIP-2 captions during pre-training. We evaluate the impact of the number of generated captions per sample and the impact of using the proposed sampling strategy based on the CLIP score between the original caption and the image. In the absence of our sampling strategy, the original caption is selected with a probability of 0.8. We see that lower values degrade performance. As the results show, using our proposed strategy along with re-captioning can boost the accuracy of the model by up to 2%.

# cap.	w. scores	UCF-101	HMDB-51
0	✗	79.4	44.4
1	✗	79.2	44.0
5	✗	79.4	45.1
5	✓	**79.5**	**46.5**

Table 8: **Effect of re-captioning:** Zero-shot results on UCF-101&HMDB-51 with ViT-B/32 when training with different automatic re-captioning variants.

A.2 Pre-training details

HC-VL pre-training details: Our procedure largely follows the training recipe of CLIP (Radford et al., 2021; Schuhmann et al., 2022), using AdamW (Loshchilov and Hutter, 2017) ($\beta_1 = 0.9$, $\beta_2 = 0.98$), a learning rate of $5e-4$ that is decayed using a cosine scheduler (Loshchilov and Hutter, 2016) after a ramp-up of 2,000 iterations, and a weight decay of 0.2. The global batch size is set to 32,800 and the model is trained with mixed precision (Micikevicius et al., 2017) for 132k iterations, seeing $\sim 4B$ samples, i.e. *significantly shorter than the typical CLIP scheduler* (Schuhmann et al., 2022). Unless otherwise stated, the image size is set to 224×224px and the text encoder context is set to 77. The augmentations applied during training match the ones used to train OpenCLIP (Schuhmann et al., 2022). Our ViT-B/32 variant is trained from scratch on HC-IT, introduced in Sec. 3. For the ViT-B/16 variant, we follow the *efficient pre-training* procedure described in Sec. 3.2 and initialize from the ViT-B/32 weights and then fine-tune for 12k iterations. The training is conducted on 32 A100 GPUs using PyTorch (Paszke et al., 2019) following the open-sourced implementation of CLIP (Ilharco et al., 2021).

HC-VL+ pre-training details: To train HC-VL+, we start from HC-VL (pre-trained on images), freeze all parameters except for the newly introduced ones and further pre-train the rest on SMiT (Monfort et al., 2021) dataset ($\sim$ 0.5M video-text pairs). The pre-training process follows the hyperparameters used in image pre-training, except for the batch size and training duration, which are set to 20, 480, and to 580 iterations, respectively. Note that *the video pre-training process is very fast*. For video data, we sample 8 frames uniformly at a resolution of 224×224px. We apply the following augmentations: random flipping (0.5), color jittering (0.8), random grayscaling (0.2) and random resizing and cropping.

A.3 Few-shot downstream fine-tuning details

To facilitate direct comparisons, for the results reported in the main paper, we aligned our setting with that of (Ni et al., 2022), using the same hyperparameters and number of shots (i.e. 2, 4, 8 and 16). For all experiments, we sample randomly from the training set K videos per class, which are then fixed for all experiments. The models are then fine-tuned using the hyperparameters listed in Table 10. As in (Ni et al., 2022), we test using a single view and 32 frames.

B Datasets

Table 12 lists and details the datasets used for evaluation in this work. Notice that the suite covers a wide range of tasks and dataset types.

C Additional results

C.1 Zero-shot evaluation of smaller and larger models

In the main manuscript, we conduct experiments using the ViT-B/32 and ViT-B/16 variants of our model. Herein, for completeness, to showcase that our approach scales to both smaller and larger models, we also report results using the ViT-S/32 and ViT-L/14. We note that for the latter, due to limited computational resources, we had to adapt our training schedule, reducing the number of iterations the models were trained for and/or using fewer patches, likely resulting in lower performance than with the full training setting. Despite this, our approach continues to outperform the equivalently sized OpenCLIP and CLIP models, trained with more data and more computational resources. We also note that, for ViT-S/32, no OpenCLIP and CLIP pre-trained

models are available online, hence we report results only for our models. Results are detailed in Table 11.

C.2 Additional results for fully supervised fine-tuning

Herein, in addition to the few-shot and zero-shot results reported in the main paper, for completeness, we also include fully supervised fine-tuning results.

Following (Ni et al., 2022), we conduct the fully supervised experiments on Kinetics-400 and Kinetics-600 datasets, using the entirety of the training and validation sets for training and testing, respectively. We note that we fully align our setting and hyperparameters with (Ni et al., 2022) to allow for a direct comparison. Specifically, during training, we randomly sample 8 frames using a sparse sampling strategy (Wang et al., 2016). Starting from our pre-trained models, HC-VL or HC-VL+, the networks are finetuned using the hyperparameters detailed in Table 10. Following (Ni et al., 2022), we use the multi-view inference with 3 spatial crops and 4 temporal clips.

As the results from Table 9 show, our model outperforms its direct competitors despite using significantly fewer training samples.

D Qualitative examples

Herein, we provide a few qualitative examples from the sub-sampled dataset.

Fig. 4 showcases a few randomly sampled keywords for each pre-defined category. The keywords cover a various range of activities and actions.

In Fig. 5 we show a few search queries alongside 3 retrieved samples. It can be observed, that generally, the search queries align well with the image-text pairs retrieved.

Fig. 6 shows a few randomly selected samples alongside their additional BLIP-generated captions. Notice that for cases where the translation fails or is incoherent, the generated captions can serve as a good alternative.

Table 9: **Fully-supervised classification results** on Kinetics-400 and Kinetics-600 in terms of Top-1 (%) accuracy. ViT-B/16 models were used for all variants.

Method	Pre-training iters	Number of samples	Kinetics-400		Kinetics-600	
			Top-1	Top-5	Top-1	Top-5
X-OpenCLIP (400M)	12B	400M	82.1	95.0	84.1	96.0
X-OpenCLIP	34B	2B	83.0	96.3	85.0	96.9
X-CLIP	-	400M	83.8	96.7	85.3	97.1
X-HC-VL (Ours)	4B	167M	**84.3**	**96.9**	**85.8**	**97.3**
X-HC-VL+ (Ours)	4B	167M + 0.5M	**84.5**	**97.0**	**86.0**	**97.4**

Table 10: The training hyperparameters for few-shot and fully supervised fine-tuning.

	Fully-sup.	Few-shot
Optimisation		
Optimizer	AdamW	
Optimizer betas	(0.9, 0.98)	
Batch size	256	64
Learning rate schedule	cosine	
Linear warmup epochs	5	
Base learning rate	8e-6	2e-6
Minimal learning rate	8e-8	2e-8
Epochs	30	50
Data augmentation		
RandomFlip	0.5	
MultiScaleCrop	(1, 0.875, 0.75, 0.66)	
ColorJitter	0.8	
GrayScale	0.2	
Label smoothing	0.1	
Mixup	0.8	
Cutmix	1.0	
Other regularisation		
Weight decay	0.001	

Table 11: **Zero-shot classification results** across a suite of action recognition datasets in terms of Top-1 (%) and Top-5 (%) accuracy (shown in parentheses). * - results reported in terms of mAP.

Arch.	Dataset	Method				
		OpenCLIP (400M) (Ilharco et al., 2021)	OpenCLIP (Ilharco et al., 2021)	CLIP (Radford et al., 2021)	**HC-VL** (Ours)	**HC-VL+** (Ours)
ViT-S/32	UCF-101	-	-	-	74.6 (94.2)	**76.6 (95.1)**
	HMDB-51	-	-	-	41.7 (68.5)	**45.2 (72.1)**
	Kinetics-400	-	-	-	49.3 (74.8)	**52.4 (77.7)**
	Kinetics-220	-	-	-	44.1 (71.0)	**46.3 (73.1)**
	Kinetics-700	-	-	-	36.5 (61.6)	**38.8 (64.0)**
	Daly	-	-	-	69.9 (97.3)	**73.8 (97.6)**
	HAA500	-	-	-	40.8 (70.5)	40.9 (**70.8**)
	HACS	-	-	-	68.4 (91.8)	**70.2 (92.9)**
	Hollywood2*	-	-	-	41.5	**45.7**
	Olympic Sports	-	-	-	50.0 (91.0)	**53.1 (93.2)**
	MiT	-	-	-	18.2 (38.4)	**20.0 (39.7)**
	AVA*	-	-	-	8.1	**10.0**
	Something-Else	-	-	-	7.0 (24.1)	**8.8 (27.6)**
	Charades*	-	-	-	16.1	**16.5**
	UCF-50	-	-	-	84.5 (97.2)	**86.0 (98.1)**
	UCF Sports	-	-	-	**52.5** (88.5)	52.3 (**90.5**)
	Average	-	-	-	44.0 (74.5)	**46.0 (76.3)**
ViT-L/14	UCF-101	74.1 (93.7)	77.8 (95.4)	80.3 (96.8)	84.5 (97.9)	**85.4 (98.0)**
	HMDB-51	34.5 (65.6)	37.8 (69.6)	48.0 (74.4)	49.3 (78.8)	**54.0 (80.1)**
	Kinetics-400	52.4 (76.2)	55.6 (79.0)	62.1 (84.7)	62.2 (85.1)	**65.7 (87.8)**
	Kinetics-220	47.6 (72.6)	52.6 (77.2)	57.1 (82.7)	57.1 (83.0)	**60.4 (85.3)**
	Kinetics-700	40.3 (64.3)	43.1 (67.5)	50.3 (75.4)	49.4 (74.3)	**52.9 (78.6)**
	Daly	76.0 (96.6)	81.5 (96.8)	84.9 (97.8)	84.3 (99.3)	**87.7 (99.8)**
	HAA500	42.4 (71.0)	45.8 (74.7)	49.2 (80.6)	54.3 (82.5)	**56.3 (84.0)**
	HACS	71.1 (93.2)	73.8 (93.8)	77.7 (95.8)	81.0 (96.9)	**82.6 (97.2)**
	Hollywood2*	46.4	44.6	49.7	54.1	**61.0**
	Olympic Sports	46.3 (89.6)	46.3 (90.3)	56.7 (91.8)	58.2 (92.1)	**58.2 (94.8)**
	MiT	19.3 (34.5)	22.1 (41.4)	22.3 (41.3)	25.8 (49.6)	**25.6 (50.9)**
	AVA*	13.0	13.2	15.0	15.7	**19.5**
	Something-Else	12.2 (34.6)	13.0 (35.0)	10.8 (35.3)	13.0 (36.2)	14.8(39.9)
	Charades*	20.6	23.8	24.9	27.1	**29.6**
	UCF-50	82.4 (97.1)	85.4 (97.1)	89.0 (98.4)	91.9 (99.5)	**92.5 (99.7)**
	UCF Sports	54.1 (91.8)	47.5 (83.6)	54.1 (91.8)	**55.8** (93.4)	54.2 (**96.7**)
	Average	45.8 (75.4)	47.7 (77.0)	52.1 (80.5)	54.0 (82.2)	**56.3 (84.1)**

Table 12: List of datasets used for evaluation.

Dataset	Num. classes	Num. samples	Descriptions
Action recognition datasets			
UCF-101 (Soomro et al., 2012)	101	13,320	Collected from the web, the dataset covers 5 action types: Human-Object Interaction, Body-Motion Only, Human-Human Interaction, Playing Musical Instruments and Sports. All clips have a fixed resolution of 320×240 pixel and a frame rate of 25 FPS. The mean clip length is 7.21 sec, ranging between 1.06sec to 71.04 sec.
HMDB-51 (Kuehne et al., 2011)	51	6849	Collected from YouTube, the dataset covers 5 action types: General facial actions, Facial actions with object manipulation, General body movements, Body movements with object interaction and Body movements for human interaction.
Kinetics 400 (Kay et al., 2017)	400	300K	Collected from YouTube, the dataset covers a broad range of classes including human-object interactions as well as human-human interactions. Each clip is roughly 10 sec long.
Kinetics 600 (Carreira et al., 2018)	600	500K	Extension of Kinetics 400. Most of the extra classes were sourced from Google's Knowledge Graph, in particular from the hobby list.
Kinetics 220 (Chen and Huang, 2021)	620	14K	Subset of Kinetics 600 that includes 220 classes not found in Kinetics 400, ensuring no overlap between the two.
Kinetics 700 (Carreira et al., 2019)	700	650K	Extension of Kinetics 600. The new classes are mostly sourced from other action recognition-related datasets, such as AVA (Gu et al., 2018).
Daly (Weinzaepfel et al., 2016)	10	500	Daly is an action recognition dataset with high-quality temporal and spatial annotation spanning 10 actions and 31 hours of YouTube videos. The dataset initially consisted of 500 videos, 366 of which are still available for download at the time of evaluation, and are divided into 220 train and 146 val videos. The reported results are on the val set.
HAA500 (Chung et al., 2021)	500	10K	HAA500 is a fine-grained action recognition dataset spanning 500 classes and 10K YouTube videos.

HACS (Zhao et al., 2019)	200	1.55M	HACS is a large-scale dataset with two versions, HACS Clips for action recognition and HACS Segments for temporal localization. This paper uses HACS Clips, which consists of 1.55M 2-second clips from YouTube spanning 200 actions. The 1.55M samples are divided into 1.5M training samples, 20,245 val samples and 20,293 test samples. At the time of evaluation, 17.3K out of the 20.2K val samples were still available for download and were subsequently used as our val split.
Hollywood2 (Marszalek et al., 2009)	12	2.5K	Hollywood2 is an action recognition dataset spanning 12 actions distributed over 2517 videos. The videos are divided into 1633 training samples and 884 test samples.
Hollywood2 (Marszalek et al., 2009)			The reported results are on this test set. Since this dataset is multi-label, we report the mean average precision (mAP) metric computed using scikit-learn's `metrics.average_precision_score` function instead of the Top-n accuracy scores.
Olympic Sports (Niebles et al., 2010)	16	783	Olympic Sports is an action recognition dataset of YouTube videos spanning 16 different sports. The 783 video samples are divided into 649 training samples and 134 val samples. The reported results are on the val set.
MiT (Monfort et al., 2019)	339	903K	Moments in Time is a large-scale action recognition dataset of $\sim$1M short video spanning 339 actions corresponding to dynamic events unfolding within 3 seconds. The dataset is divided into 802K training videos, 33.9K validation videos and 67.8K testing videos. The reported results are on the val set.
Something-Else (Materzynska et al., 2020)	174	180K	A compositional action recognition dataset annotated with object names (based on bounding boxes) and actions, allowing for compositional class names. The dataset consists of 168913 training samples and, respectively, 24777 samples for the validation set. The reported results are on the validation set.

Charades (Sigurdsson et al., 2018)	157	9848	Charades is a dataset consisting of 9848 videos of daily indoor activities, annotated with 157 action classes (41,104 labels) and with 46 object classes. Each video has been exhaustively annotated using consensus from 4 workers on the training set and from 8 workers on the test set. The evaluation follows the protocol described for the Hollywood2 dataset.
AVA (Gu et al., 2018)	80	1.58M	The AVA dataset consists of 430 15-min video clips annotated densely annotated 80 actions that are also localized in space and time, resulting in 1.58M action labels with multiple labels per person occurring frequently. As such, multiple actions, executed by different persons, often occur concomitantly. Hence, we follow the same protocol as for Hollywood2 to compute the mAP.
UCF-50 (Reddy and Shah, 2013)	50	6.6K	UCF-50 is an action recognition dataset spanning 50 actions of 6618 YouTube videos. Since no official splits are provided, we split the dataset into 4246 training videos and 2372 validation videos with disjoint groups. The reported results are on the val set.
UCF Sports (Soomro and Zamir, 2015)	13	150	UCF Sports is an action recognition and localization dataset of 150 videos originally spanning 10 actions. To make the task more challenging, we use an expanded 13-action version by considering different views of two actions, golfing and kicking. Since no official splits are provided, we generate the train and val split on a per-class basis, and use 61 videos as our val split, and the rest as the training set. The reported results are on the val set.
Image retrieval datasets			
Flicr30k (Young et al., 2014)	N/A	31K	Flickr30k is a dataset consisting of 31,783 images collected from Flickr, together with 5 reference sentences provided by human annotators. The dataset is a popular image retrieval and captioning dataset.

MS-COCO	N/A	328K	The MS COCO (Microsoft Common Objects in Context) dataset is a large-scale object detection, segmentation, key-point detection, and captioning dataset containing 328K images. In this work, we only used the annotations relevant for image retrieval, i.e. the provided captions, evaluating on the val set.
Video retrieval datasets			
MSR-VTT (Xu et al., 2016b)	N/A	6513	MSR-VTT (Microsoft Research Video to Text) is an open domain video captioning and retrieval dataset, consisting of 10,000 video clips from 20 categories (6,513 clips for training, 497 clips for validation, and 2,990 clips for testing).
MSVD (Chen and Dolan, 2011)	N/A	2K	The MSVD dataset consists of 120K sentences describing more than 2000 video samples. The videos were manually labeled by multiple annotators.
DiDemo (Anne Hendricks et al., 2017)	N/A	27K	The DiDeMo is a large-scale dataset for temporal localization of events in videos given natural language descriptions containing 26,892 moments.

Figure 4: **Taxonomy of keywords:** We show 30 randomly sampled keywords per each one of the six pre-defined categories.

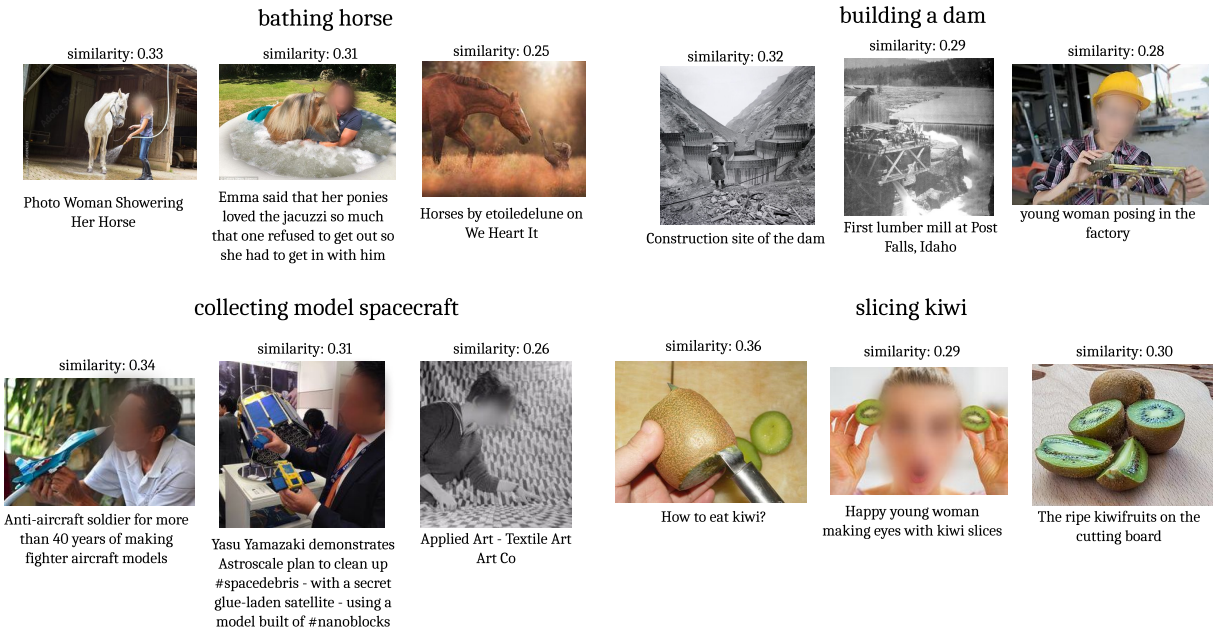

Figure 5: **Samples of queried images:** we show 3 images for 4 keywords, together with their text query-image cosine similarity, and either the original English caption or its English translation.

Figure 6: **Translated and BLIP2 generated captions:** (top-row) examples where the original caption is either good or well translated, and (bottom-row) examples with poor translation into English.