Supplementary material: Video-adverb retrieval with compositional adverb-action embeddings

Thomas Hummel[1]
thomas.hummel@uni-tuebingen.de
Otniel-Bogdan Mercea[1]
otniel-bogdan.mercea@uni-tuebingen.de
A. Sophia Koepke[1]
a-sophia.koepke@uni-tuebingen.de
Zeynep Akata[1,2]
zeynep.akata@uni-tuebingen.de

[1] University of Tübingen
Germany
[2] MPI for Intelligent Systems
Germany

A Dataset splits for unseen adverb-action compositions

In this section, we provide further details about our proposed dataset splits for unseen adverb-action compositions based on the ActivityNet Adverbs [3, 5] and MSR-VTT Adverbs [5, 17] datasets. In Table 1, we include information about the number of unlabelled samples (i.e. videos) and the number of unlabelled pairs (i.e. adverb-action compositions) in the dataset splits. The unlabelled samples are not used by ReGaDa, but we designed the splits so that we can fairly evaluate previous work [4] that uses unlabelled samples for training. The number of unlabelled samples and unlabelled pairs usually ranges from 30% to 50% of the total number of training samples and training pairs. This is significant, as methods like [4] use more training data than ReGaDa while performing significantly worse as observed in Table 6 in the main paper. We refer to the ActivityNet Adverbs and MSR-VTT Adverbs datasets as ActivityNet and MSR-VTT respectively.

In addition to the ActivityNet Adverbs and MSR-VTT Adverbs datasets, we use the VATEX Adverbs dataset [5, 15], and in particular the corresponding splits for unseen adverb-action compositions introduced in [4]. However, we use the same pre-extracted features as the current state-of-the-art work [10]. As some of the videos used in the split in [4] are not available anymore, it is not possible to extract S3D features for those. Hence, this resulted in fewer samples in the dataset, the number of training samples being reduced from 6921 to 6603, unlabelled samples from 3469 to 3317, and test samples from 3457 to 3293. In the following, we refer to the VATEX Adverbs dataset as VATEX.

Dataset	# train samples	# unlabelled samples	# test samples	# pairs train	# pairs unlabelled	# pairs test
VATEX	6603	3317	3293	319	168	316
MSR-VTT	987	306	454	225	114	225
ActivityNet	1490	634	848	635	537	543

Table 1: Statistics of our dataset splits for the retrieval of unseen adverb-action compositions on the MSR-VTT Adverbs and ActivityNet Adverbs datasets. Statistics are also provided for the VATEX Adverbs dataset for features from [10].

B Exploring the use of different word embeddings for unseen adverb-action compositions

Our REGADA framework composes adverb and action text embeddings in a shared embedding space. Specifically, we used a text model that was jointly trained with the S3D video model. In this section, we show results for different choices of word embeddings. Existing and widely-adopted word embeddings like GloVe [12], word2vec [9], and fastText [1] rely on unsupervised learning techniques to generate vector representations of words based on their co-occurrence statistics in a large corpus of text. Specifically, word2vec and GloVe focus on co-occurrences of words, whereas fastText uses co-occurrences of n-gram characters, which can be useful when dealing with rare words.

Model	VATEX	ActivityNet	MSR-VTT
CLIP [13]	54.5	55.1	57.0
Act. Mod. [4]	53.8	57.0	56.0
AC$_{\text{CLS}}$ [10]	54.3	55.1	53.7
AC$_{\text{REG}}$ [10]	54.9	53.9	59.0
REGADA	61.7	**58.4**	**61.0**
REGADA w2v	60.5	53.1	60.0
REGADA fastText	60.8	53.5	57.3
REGADA GloVe	58.0	54.0	57.7
REGADA GPT-3	**63.3**	53.5	60.3

Table 2: Effect of using different types of word embeddings in our REGADA framework on the performance for retrieving unseen action-adverb compositions on the VATEX, ActivityNet and MSR-VTT benchmarks. [4] uses pseudo-labelling.

Prior works on video-adverb retrieval leveraged GloVe embeddings of class labels [4, 5], while approaches in zero-shot learning commonly use word2vec or fastText embeddings as side information [6, 7, 8, 11, 16]. However, recent advances in language modelling have shown impressive progress on a variety of natural language processing tasks. For instance, large language models incorporate contextual information at the sentence level and beyond, which could result in more informative and accurate embeddings. To investigate their usefulness for our retrieval task, we extract word embeddings with GPT-3 [2] using the OpenAI API for the `text-embedding-ada-002` model. While word2vec, fastText, and GloVe provide 300-dimensional embeddings, GPT-3 embeddings have a much larger dimension of 1536. All text embeddings are projected to 400-dimensional vectors before being input into the text encoder. For CLIP [13], we extract visual CLIP features for each second of the video and CLIP text embeddings from the action-adverb labels (*e.g. cut slowly*). We then use the cosine similarity between temporally-averaged frame features and text embeddings for retrieval.

Table 2 shows that the choice of the text embedding results in significant performance changes, measured by the binary antonym classification accuracy. REGADA uses text embeddings jointly trained with the S3D video model like the other baselines (referred to as S3D embeddings in the following), and it is able to outperform all the baselines, as shown in the main paper. However, from Table 2 it can be observed that ReGaDa with S3D embeddings is outperformed by ReGaDa with GPT-3 embeddings on VATEX, leading to a performance of 63.3 compared to 61.7 for S3D embeddings. GPT-3 embeddings contain more contextual and fine-grained semantic information but suffer from a significant reduction in dimensions in the projection. We find that higher-dimensional text embeddings perform worse when training data is scarce (*e.g*. 53.5/60.3 for GPT-3 vs. 58.4/61.0 for S3D on ActivityNet/MSR-VTT), likely caused by a lack of training data to learn the down-projection. Overall, word2vec, fastText, and GloVe embeddings yield slightly worse results than S3D embeddings across datasets.

C Training without antonyms

In Table 3, we present the video-to-adverb and adverb-to-video retrieval performance when training without antonyms. This task was introduced in [10]. For the results in the main paper, ReGaDa is trained with antonyms as negative examples in its triplet loss. As it might not always be feasible to require adverb-action samples that are additionally annotated with an adverb-antonym, this scenario inspects the generalisation capabilities of ReGaDa to dataset settings with fewer constraints.

When training without adverb-antonyms, ReGaDa randomly samples an adverb as a negative sample which is not identical to the positive adverb sample. As there is no access to information about the adverb-antonym during evaluation, the Acc-A metric cannot be used in this context.

In Table 3 we can observe that ReGaDa outperforms all prior methods for this task across all datasets and metrics For example, on VATEX ReGaDa obtains a mAP W score of 0.292 compared to 0.283 for AC_{CLS}. Moreover, ReGaDa obtains a mAP M score of 0.136 which significantly outperforms AC_{CLS} with a score of 0.108.

	HowTo100M [5]		Adverbs in Recipes [10]		ActivityNet [4]		MSR-VTT [4]		VATEX [4]	
	mAP W	mAP M	mAP W	mAP M	mAP W	mAP M	mAP W	mAP M	mAP W	mAP M
Priors	0.446	0.354	0.491	0.263	0.217	0.159	0.308	0.152	0.216	0.086
S3D pre-trained	0.339	0.238	0.389	0.173	0.118	0.071	0.194	0.075	0.122	0.038
TIRG [14]	0.441	0.476	0.485	0.228	0.186	0.111	0.297	0.113	0.195	0.065
Act Mod [5]	0.408	0.352	0.508	0.249	0.187	0.127	0.233	0.134	0.144	0.060
$\text{AC}_{\text{CLS}}^{\dagger}$ [10]	0.562	0.420	0.606	0.289	0.130	0.096	0.305	0.131	0.283	0.108
$\text{AC}_{\text{REG}}^{\dagger}$ [10]	0.573	0.481	0.667	0.319	0.143	0.093	0.287	0.121	0.282	0.100
ReGaDa	**0.580**	**0.536**	**0.668**	**0.466**	**0.282**	**0.211**	**0.401**	**0.252**	**0.292**	**0.136**

Table 3: Results *without* antonyms during training for adverb-to-video retrieval (mAP W/M). Higher is better for all metrics. † refers to updated results provided by the authors of [10].

D Comparing ReGaDa with CLIP

In this section, we present additional video-adverb retrieval results with CLIP [13] in addition to the retrieval results for unseen compositions (see Table 2).

Similar to the experiment on unseen compositions (see Section B), we use the cosine similarity between temporally-averaged CLIP frame features and text embeddings for the retrieval with CLIP. Additionally, we examine the impact of replacing the S3D video/text embeddings of ReGaDa with CLIP embeddings ($\text{ReGaDa}_{\text{CLIP}}$).

In Table 4, we can observe that CLIP performs marginally better than the S3D pre-trained baseline. Using CLIP features in ReGAdA improves adverb retrieval (Acc-A) slightly on ActivityNet and VATEX. However, ReGAdA$_{\text{CLIP}}$ is worse than ReGAdA for video retrieval, likely caused by inferior visual features when extracting those only from a few video frames.

	ActivityNet			MSR-VTT			VATEX		
	mAP W	mAP M	Acc-A	mAP W	mAP M	Acc-A	mAP W	mAP M	Acc-A
S3D pre-tr.	0.118	0.070	0.560	0.194	0.075	0.603	0.122	0.038	0.586
CLIP [13]	0.120	0.067	0.611	0.206	0.084	0.677	0.129	0.039	0.644
ReGAdA$_{\text{CLIP}}$	0.201	0.151	**0.781**	0.352	0.142	0.784	0.247	0.098	**0.837**
ReGAdA	**0.239**	**0.175**	0.771	**0.378**	**0.228**	**0.786**	**0.290**	**0.113**	0.817

Table 4: Comparing ReGAdA with CLIP as a baseline, and when replacing ReGAdA's S3D video/text embeddings with CLIP embeddings (ReGAdA$_{\text{CLIP}}$).

E Seed experiments

In Table 5, we provide experimental results that test the robustness of our model with regard to the seeds used, as done in [10]. To compute these numbers, we use four seeds and compute the mean and the standard deviation over these runs. It can be observed that ReGAdA achieves a higher mean than the other baselines. Furthermore, the standard deviation with our model is relatively low.

	Adverbs in Recipes [10]		
	mAP W	mAP M	Acc-A
Act Mod	0.394 ± 0.023	0.140 ± 0.026	0.843 ± 0.013
MLP+Act Mod	0.407 ± 0.044	0.151 ± 0.033	0.842 ± 0.012
AC$_{\text{CLS}}{}^{\dagger}$	0.605 ± 0.001	0.287 ± 0.001	0.841 ± 0.000
AC$_{\text{REG}}{}^{\dagger}$	0.611 ± 0.002	0.239 ± 0.007	0.845 ± 0.001
ReGAdA	**0.699 ± 0.004**	**0.419 ± 0.012**	**0.876 ± 0.001**

Table 5: Performance of our ReGAdA framework on the Adverbs in Recipes dataset when using multiple random seeds. † refers to updated results provided by the authors of [10].

References

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. In *TACL*, 2017.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015.

[4] Hazel Doughty and Cees GM Snoek. How do you do it? Fine-grained action understanding with pseudo-adverbs. In *CVPR*, 2022.

[5] Hazel Doughty, Ivan Laptev, Walterio Mayol-Cuevas, and Dima Damen. Action Modifiers: Learning from Adverbs in Instructional Videos. In *CVPR*, 2020.

[6] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *CVPR*, 2021.

[7] Otniel-Bogdan Mercea, Thomas Hummel, A. Sophia Koepke, and Zeynep Akata. Temporal and cross-modal attention for audio-visual zero-shot learning. In *ECCV*, 2022.

[8] Otniel-Bogdan Mercea, Lukas Riesch, A. Sophia Koepke, and Zeynep Akata. Audio-visual generalised zero-shot learning with cross-modal attention and language. In *CVPR*, 2022.

[9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[10] Davide Moltisanti, Frank Keller, Hakan Bilen, and Laura Sevilla-Lara. Learning action changes by measuring verb-adverb textual relationships. In *CVPR*, 2023.

[11] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *CVPR*, 2021.

[12] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *EMNLP*, 2014.

[13] 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 *ICML*, 2021.

[14] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *CVPR*, 2019.

[15] 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 *ICCV*, 2019.

[16] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *CVPR*, 2016.

[17] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016.