Supplementary Materials
to Paper #4217
Learning Audio-Video Modalities from Image Captions

1 VideoCC3M dataset

In this section we provide some more details on the automatically mined clips that are part of the VideoCC3M dataset, including basic statistics, more qualitative examples, and a brief human study to assess the quality of the mined clips.

1.1 Dataset statistics

Table 1: **Dataset Statistics:** VideoCC3M is an order of magnitude larger than existing video-text datasets in the number of videos and captions. Rows highlighted in blue are large-scale, weakly annotated datasets. WVT uses titles and descriptions from YouTube videos, and HowTo100M has noisy text supervision from ASR. † Not publicly released.

dataset	domain	clips	# average clip length (s)	# captions	time (hr)	# pairs
MPII Cook [14]	cooking	44	600	6K	8	6K
TACos [12]	cooking	7K	360	18K	15.9	18K
DideMo [1]	flickr	27K	28	41K	87	41K
MSR-VTT [19]	open	10K	15	200K	40	200K
Charades [16]	home	10K	30	16K	82	16K
LSMDC15 [13]	movies	118K	4.8	118K	158	118K
YouCook II [22]	cooking	14K	316	14K	176	14K
ActivityNet [9]	action focused	100K	180	100K	849	100K
CMD [2]	movies	34K	132	34K	1.3K	34K
WebVid-2M	open	2.5M	18	2.5M	13K	2.5M
VideoCC3M	**open**	**6,323,992**	**10**	**974,247**	**17.5K**	**10,339,249**
WVT [17]†	action focused	70M	10	70M	194K	70M
HowTo100M [11]	instruction	136M	4	136M	134.5K	136M

We provide the total number of unique captions, video clips and pairs in Table 1 comparing VideoCC3M to other existing video and text datasets. Note that at 10M pairs, our dataset is much larger than manually annotated datasets but still much smaller than the large HowTo100M dataset. The full distribution of clips per caption is provided in Fig. 1, (note that the y-axis is on a log scale). Each caption is matched to a mean of 10.6 clips, with some captions matched to more than 10 clips. This is possible because, while we limit the clip mining to 10 clips per seed image, the original CC3M dataset has multiple seed images with the same caption, eg 'an image of digital art', leading to more than 10 mined clips for these captions. 96.6K out of 97K captions have less than 50 clips per caption. This added redundancy is an interesting feature of the data where visually similar clips share the same caption from the same seed image and visually distinct clips share that same caption from different seed images.

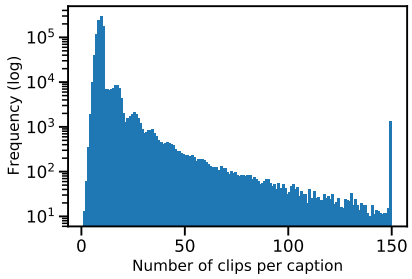

Fig. 1: **Distribution of clips per caption in VideoCC3M.** Frequency of samples (y-axis) is on a log-scale. Because the CC3M dataset has some seed images that share the same caption, one caption can have more than 10 mined clips. 96.6K out of 97K captions have less than 50 clips per caption. All samples with more than 150 clips per caption are grouped into a single bin.

1.2 Domains

We show the top 50 domains in Fig. 2 for both the VideoCC3M and the HowTo100M datasets, and group remaining samples into the 'Other' domain. This figure expands the analysis presented in Figure 3 (left) of the main paper. It is clear that the domains in VideoCC3M are more balanced, while HowTo100M videos are largely dominated by the 'Food' and 'Hobby' domains. This is unsurprising given that HowTo100M is limited to instructional videos.

1.3 Human study on quality

In order to quantiatively assess the quality of the mined clips in VideoCC3M, we also perform a quick manual assessment of 100 randomly sampled clips from the dataset. For each clip, we first annotate whether there is at least one frame in the clip matching the caption, and find that 91 out of 100 clips were labelled to have this property. We noticed that clips without a single frame matching the caption are often those where the seed image does not match the caption either, due to noise in the CC3M dataset. We then devise a simple quality score with the following scale of 3 values: 0 - not relevant, 1 - somewhat relevant, 2 - very relevant, to assess the degree to which the caption matches the retrieved sample. For examples of clips that are somewhat relevant, see Fig. 4. Over 100 samples, we get an average score of 1.51, with 9 samples having score 0, 31 having score 1 and 60 having score 2.

1.4 More qualitative examples

We show some more qualitative examples in Fig. 3. Note the diversity of retrieved samples, including an animated video of a tree on a white background. In Fig. 4, we also show some failure cases, where the clips are somewhat related to the captions but not perfectly.

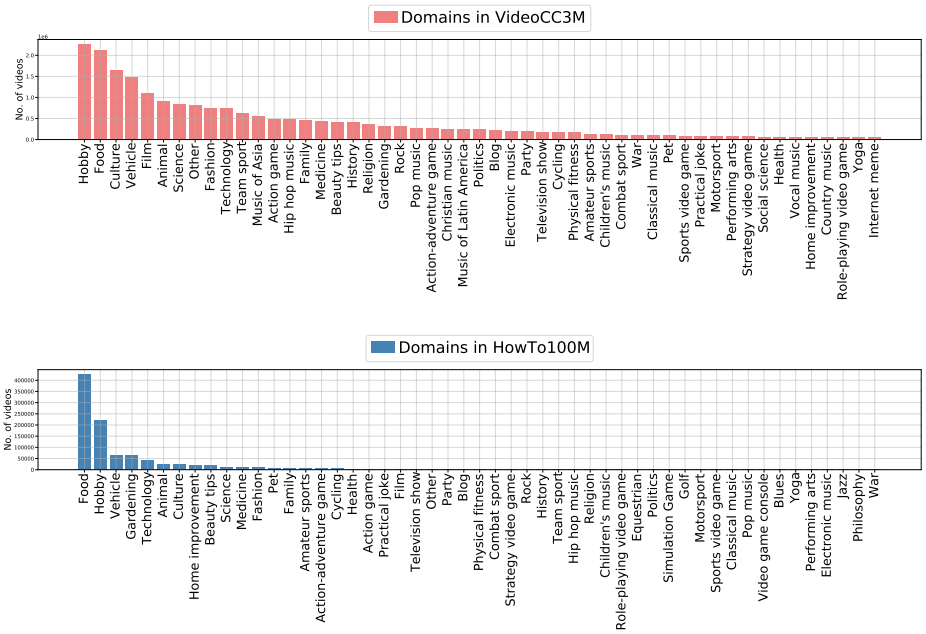

Fig. 2: **Domains in VideoCC3M vs HowTo100M.** We show the top 50 domains for each dataset and group remaining samples into 'Other'. Note how the domains in VideoCC are more balanced. Note HowTo100M has about 1M videos in the dataset.

1.5 Ablation on temporal length t

We show the effect of the length of the mined clips on zero-shot performance on the MSR-VTT dataset. Results are in Table 2. Although we know that video content diverges the further we are from the matched frame to the seed image, we find increasing the span actually increases performance up until 10 seconds. This is perhaps because videos tend to be correlated over time. Unrelated extra information could also act as a regularisation, wherein slight noise does not harm the results. We hence use clips of 10 seconds in all further experiments with VideoCC3M, but we note that future work will more intelligently determine the boundary of the mined clips.

$t(s)$	3	5	10	20	30
MSR-VTT (ZS)	16.4	17.1	18.9	18.8	18.8

Table 2: **Temporal Span t of the mined clips.** We report zero-shot R@1 performance on the MSR-VTT dataset.

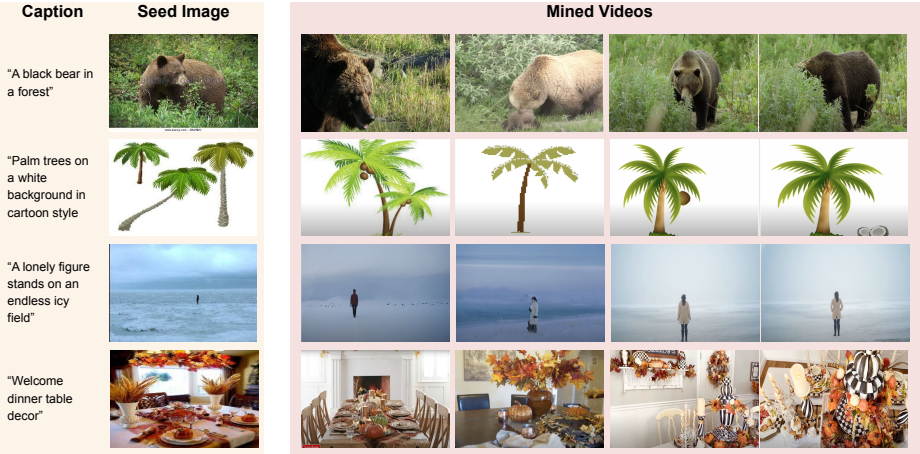

Fig. 3: **Examples of clips with captions that are mined automatically.** For each seed image, we show 3 'matched' clips obtained using our automatic video mining method. For the first 2 clips, we show only a single frame, but for the third clip we present 2 frames to show motion , either of the subjects in the video (first 3 rows - the bear, the coconut falling, the arms of the woman) or camera motion (last row). Note frames may have been cropped and resized for ease of visualisation.

2 VideoCC12M dataset

We ran our mining pipeline with an additional seed image captioning dataset called Conceptual Captions 12M [5] (CC12M). CC12M is the recently released extension of Conceptual Captions 3M [15] (CC3M). Note that while CC3M consists of higher quality captions [15], CC12M was created by relaxing the data collection pipeline used in CC3M, and hence the captions are far noisier. VideoCC3M consists of 10.3M clip-text pairs from 6.3M video clips and 970K unique captions, while VideoCC12M contains 48.0M clip-text pairs from 30.3M video clips and 5.7M unique captions. While we include results on VideoCC12M for completeness, we note that for most tasks VideoCC3M is sufficient for good performance with far less data.

2.1 Video Retrieval using VideoCC12M

We show results in Table 3. Pretraining on the VideoCC12M dataset provides a further boost to performance over the VideoCC3M pretraining, particularly for R@10 and R@5. This furthers the improvement over the state of the art, which was provided in Table 3 in the main paper. Our model trained on VideoCC12M achieves R@1 37.1 compared to FIT [3], which gets an R@1 of 32.5. Note FIT is pretrained on WebVid2M, COCO and CC3M.

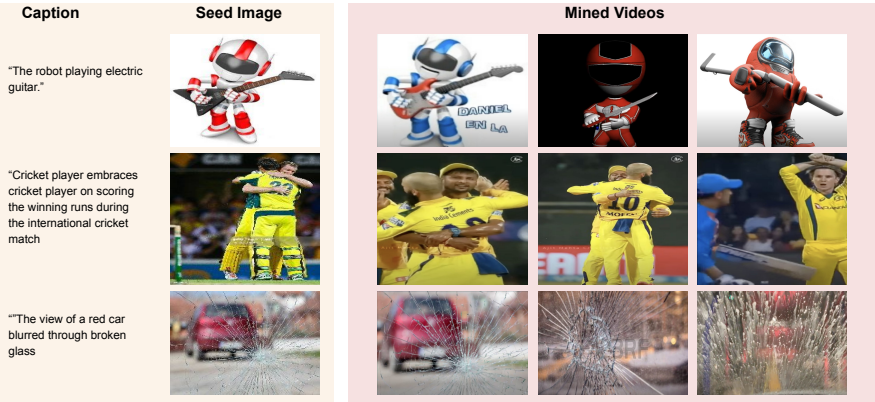

Fig. 4: **Failure Cases: examples of somewhat related clips with captions that are mined automatically.** For each seed image, we show 3 'matched' clips obtained using our automatic video mining method. Here we show failure cases, where the matched clips are somewhat relevant to the caption, but not entirely. For example, top row - in the last two clips the robot are holding a long object but it is not a guitar, second row - last clip contains cricketers but they are not hugging. Finally in the third row, note that the second clip has the broken glass but no red car, whereas the last clip has a red car but the glass is not broken, it is being washed in a car wash. Note that the original seed image of the red car is originally from a video, which we retrieve using our pipeline. Note frames may have been cropped and resized for ease of visualisation.

2.2 Video Captioning using VideoCC12M

Results for video captioning are provided in Table 4. With the additional data from VideoCC12M, we are on par with HowTo100M in the finetuning setting. For the zero-shot setting, training on VideoCC3M provides a substantial boost across all metrics, with a fraction of the training data, and also outperforms training on VideoCC12M in the zero-shot setting. This is interesting, and we hypothesise it is because the captions in CC3M are far cleaner than CC12M. We note the exact same trend was reported for zero-shot image captioning in the CC12M paper [5]. This suggests that zero-shot performance depends more on the transferred caption quality, and future work may improve transfer performance by cleaning up captions in larger data sets. This finding reinforces the theme that more data is not always better.

3 Implementation Details

In this section we provide more details about the inputs to the MBT video encoder. RGB frames for all datasets are extracted at 25 fps. For MSR-VTT we sample 32 RGB frames with stride 3 frames, while for AudioCaps we sample 8 RGB frames with a uniform stride of 56 frames. Audio for all datasets is sampled

Pretraining Data	Modality	# Caps	R@1	R@5	R@10
Finetuned					
-	V	-	31.2	60.7	71.1
HowTo100M [11]	V	130M	33.1	62.3	72.3
VideoCC3M	V	970K	35.0	63.1	75.1
VideoCC3M	A+V	970K	35.3	65.1	76.9
VideoCC12M	V	5.7M	36.9	66.5	75.6
VideoCC12M	A+V	5.7M	**37.1**	**67.5**	**77.6**
Zero-shot					
HowTo100M [11]	V	130M	8.6	16.9	25.8
VideoCC3M	V	970K	18.9	37.5	47.1
VideoCC3M	A+V	970K	19.4	39.5	50.3
VideoCC12M	V	5.7M	21.8	44.5	54.1
VideoCC12M	A+V	5.7M	**22.3**	**45.8**	**57.2**

Table 3: **Effect of pretraining data on text-video retrieval for the MSR-VTT dataset. # Caps:** Number of unique captions. Training on VideoCC provides much better performance than Howto100M, at a fraction of the dataset size, particularly for the zero-shot setting.

at 16kHz and converted to mono channel. Following MBT, we extract log mel spectrograms with a frequency dimension of 128, 25ms Hamming window and hop length 10ms. This gives us an input of size 128×100 for 1 second of audio. We sample 8 audio spectrograms for each video clip, and unlike MBT, we use a stride of 3 between spectrograms to cover 24 seconds of audio at a time. For the MSR-VTT data set examples missing audio, we feed in zeros as input.

4 Model architecture ablations

In this section we provide ablations on the stride of frames used in the video encoder as well as the initialisation of the audio encoder for the AudioCaps dataset.

4.1 Clip coverage

We use the stride of the sampled RGB frames to control the coverage of clips that are randomly sampled during training, and provide the results in Table 5. A randomly sampled 2 second clip from the video (stride=2) does much worse than using a stride of 14 (18s clip coverage). We find in general a greater clip coverage leads to better performance, indicating that the captions in MSR-VTT usually refer to concepts that either span the entire clip, or that may be missed by randomly sampling a 2s segment. This observation was also made by FIT [3]. Note that the numbers here are lower than our best model, as we use a batch size of 64 during training (compared to 256 used for our best model).

Method	PT	Modality	B-4	C	M
Finetuned					
ORG-TRL [21]	-	V	43.60	51	28.80
VNS-GRU [6]	-	V	45.30	53	29.90
UniVL [10]	HowTo100M	V+T	41.79	50	28.94
DECEMBERT [18]	HowTo100M	V	45.20	52	29.70
Ours	VideoCC3M	V	45.47	55	36.96
Ours	HowTo100M	V	**47.33**	55	37.11
Ours	VideoCC12M	V	47.21	**56**	**37.70**
Zero-shot					
Ours	HowTo100M	V	7.5	0.5	8.23
Ours	VideoCC3M	V	**13.23**	**8.24**	**11.34**
Ours	VideoCC12M	V	10.09	3.58	9.68

Table 4: **Results on the MSR-VTT dataset for video captioning.** Zero-shot results are obtained without any annotated video-text data. Modalities: **V:** RGB frames. **T:** ASR in videos.

Stride	Span (s)	R@1	R@5	R@10
2	2.56	24.1	53.5	66.2
6	7.68	24.2	53.7	66.1
10	12.80	24.8	55.1	67.8
14	17.92	**27.3**	**56.6**	**68.7**
18	23.04	26.9	**56.6**	68.5

Table 5: **Effect of stride on MSR-VTT performance, which affects the temporal span of a single clip.** All models are trained with RGB-only, using K400 initialisation, 32 input frames and a batch size of 64. At test time, we sample 4 equally spaced clips and average the similarity scores. Best performance is obtained with a stride of 14.

4.2 Audio encoder initialisation

We experiment with initialising the MBT backbone with ImageNet-21K and VGGSound weights, for the task of audio retrieval on the AudioCaps dataset. Results are in Table 6. Unlike the video initialisation ablation in Table 1 (left) of the main paper, we find that VGGSound initialisation provides a large improvement over Imagenet, and use this as a default for experiments on both the AudioCaps and Clotho datasets.

4.3 Societal Impact

We note that transformers are in general compute-heavy, which can have adverse environmental effects. We believe that releasing a dataset that is an order of magnitude smaller than HowTo100M, but provides better zero-shot generalisation, will lead to faster and cheaper language-video model innovation. Finally, our dataset may reflect biases present in videos online, as well as biases in the captions of the seed dataset. Existing biases may render models trained on this data unsuitable for certain applications. It is important to keep this in mind when deploying, analysing and building upon these models.

Init.	Modality	R@1	R@10
Scratch	A	19.1	64.7
ImageNet21K [8]	A	30.2	75.4
VGGSound [7]	A	**32.0**	**82.3**

Table 6: **Audio encoder initialisation on the AudioCaps dataset for text-audio retrieval.**

4.4 Fairness Analysis on the Data

We start with input data, CC3M [15], that has already tried to mitigate fairness issues. This data source has many fewer fairness issues than than website scraping efforts focusing on scale as evaluated in [4]. We made further efforts to mitigate fairness issues in the text domain, image domain, and video domain, by performing both automated and manual analysis. For automated analyses, we evaluated the text using NLP tools for toxicity and PII, while images and videos were reviewed for their likelihood of containing mature or offensive imagery. For manual analyses, we inspected thousands of caption-video pairs where the captions contained words that are sensitive or have been previously shown to have fairness disparities such as those listed in [4,20], and provided at least some further mitigation of extreme errors.

References

1. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: ICCV (2017) 1
2. Bain, M., Nagrani, A., Brown, A., Zisserman, A.: Condensed movies: Story based retrieval with contextual embeddings. In: ACCV (2020) 1
3. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. ICCV (2021) 4, 6
4. Birhane, A., Prabhu, V.U., Kahembwe, E.: Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv preprint arXiv:2110.01963 (2021) 8
5. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3558–3568 (2021) 4, 5
6. Chen, H., Li, J., Hu, X.: Delving deeper into the decoder for video captioning. In: ECAI (2020) 7
7. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 721–725. IEEE (2020) 8
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009) 8
9. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: ICCV (2017) 1
10. Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., Zhou, M.: UniVL: A unified video and language pre-training model for multimodal understanding and generation. arXiv e-prints (2020) 7
11. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In: ICCV (2019) 1, 6
12. Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., Pinkal, M.: Grounding action descriptions in videos. Transactions of the Association for Computational Linguistics **1**, 25–36 (2013) 1
13. Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., Schiele, B.: Movie description. International Journal of Computer Vision **123**(1), 94–120 (2017) 1
14. Rohrbach, M., Amin, S., Andriluka, M., Schiele, B.: A database for fine grained activity detection of cooking activities. In: CVPR (2012) 1
15. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018) 4, 8
16. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: ECCV (2016) 1
17. Stroud, J.C., Lu, Z., Sun, C., Deng, J., Sukthankar, R., Schmid, C., Ross, D.A.: Learning video representations from textual web supervision. arXiv preprint arXiv:2007.14937 (2020) 1
18. Tang, Z., Lei, J., Bansal, M.: Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In: NAACL (2021) 7
19. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: CVPR (2016) 1

20. Yang, K., Qinami, K., Fei-Fei, L., Deng, J., Russakovsky, O.: Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 547–558 (2020) 8
21. Zhang, Z., Shi, Y., Yuan, C., Li, B., Wang, P., Hu, W., Zha, Z.J.: Object relational graph with teacher-recommended learning for video captioning. In: CVPR (2020) 7
22. Zhou, L., Xu, C., Corso, J.: Towards automatic learning of procedures from web instructional videos. In: AAAI (2018) 1