Supplementary Material

A Additional Experiments

Performance on Spatial-Focused Datasets. We investigate its performance on the spatial-focused MSR-VTT and DiDeMo in Table 8. While our approach aims for motion representations it performs well on spatial tasks, surpassing Frozen, ALPRO, Lavender, and Singularity.

Table 8: MSR-VTT and Didemo Results. Our model performs well on spatial-focused datasets despite targeting motion.

	MSR-VTT				DiDeMo			
	R@1	R@5	R@10	Avg	R@1	R@5	R@10	Avg
Frozen [4]	31.0	59.5	70.5	53.7	31.0	59.8	72.4	54.4
ALPRO [40]	33.9	60.7	73.2	55.9	35.9	67.5	78.8	60.7
Lavender [44]	37.8	63.8	75.0	58.9	47.4	74.7	82.4	68.2
Singularity [37]	36.8	65.9	75.5	59.4	47.4	75.2	84.0	68.9
VindLU [10]	43.8	70.3	78.5	64.5	54.5	81.3	89.0	75.0
LocoMotion	39.3	69.8	78.2	62.3	51.2	76.5	84.9	70.9

Impact of Video Background. We assess the impact of using real WebVid [4] videos as the background to our generated motions. Specifically, we compare using a black background to a static frame of the video and the full video. From the results in Table 9 we conclude that using the original video, whether as a static frame or the full video is more successful than a blank black canvas. This demonstrates that our method is successful despite the sometimes unrealistic combination of objects and background scenes. While the black background is lower than using the video, it still obtains good results further showcasing the usefulness of our approach as motion-focused video-language representations can be learned without real videos. This could be particularly useful in specialized domains where internet-scale data is absent.

Table 9: Impact of Video Background. A background video is helpful, although it can be a static single frame.

	R@1	R@5	R@10	Avg
Black	54.0	89.7	96.6	80.1
Frame	58.6	94.3	99.4	84.1
Video	55.2	92.5	97.7	81.8

Scalability. Figure 9 ablates the scalability of our approach compared to the baseline. Our model scales much much quicker and with a sharper gradient.
Additional Video Language Models We perform a small-scale experiment with UMT [42] in Table 10. Our approach enables UMT to learn more effective

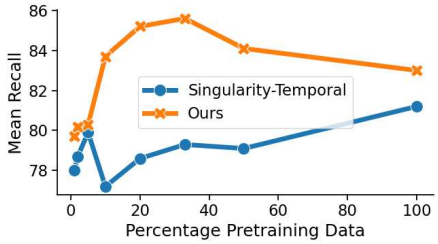

Fig. 9: Scalability. Our model scales much quick and with a sharper gradient than the baseline model.

Table 10: Our approach is effective with Unmasked Teacher [42]

	#Pre-train	SSv2-Template			SSv2-Label		
		R@1	R@5	R@10	R@1	R@5	R@10
UMT [A]	1.2M	79.3	100	100	49.1	77.0	85.1
+ ***LocoMotion***	1.2M	79.9	99.4	100	50.9	79.5	87.5

motion representations, outperforming WebVid trained UMT on SSv2-Template and SSv2-Label.

B Caption Content

In Figure 2 of the main paper we show radar plots displaying the average occurrences of different parts-of-speech per caption. For clarity, in Table 11 we display the raw numbers used to obtain these plots alongside the caption source and percentage of captions that the nouns can uniquely identify.

Table 11: Caption Content of common video-language datasets. Current video-language pretraining and downstream datasets have a spatial focus demonstrated by the average number of nouns each caption contains as well as the percentage of captions that can be uniquely identified by the nouns they contain.

	Source	Words per Caption					% Unique
Dataset	Caption	Noun	Adjective	Verb	Adverb	Adposition	Noun
WebVid [4]	Alt-text	6.9	1.8	1.2	0.2	2.1	57.8
HowTo100M [52]	ASR	1.5	0.4	0.7	0.1	0.0	14.9
YT-Temporal [86]	ASR	5.4	1.6	3.8	2.1	2.5	86.6
InternVid [78]	Generated	3.7	0.7	0.9	0.0	1.5	53.7
CMD [3]	Description	2.8	0.7	2.3	0.3	1.8	81.6
Charades [67]	Script	6.0	0.2	3.8	0.6	3.0	95.0
VATEX [77]	Manual	4.3	0.7	2.1	0.3	1.8	96.0
ActivityNet [7]	Manual	3.8	0.6	2.3	0.5	1.8	81.2
MSR-VTT [83]	Manual	3.3	0.5	1.3	0.1	1.2	78.4
DiDeMo [1]	Manual	2.6	0.6	1.1	0.3	1.2	53.1

C Additional Implementation Details

C.1 Describing Motions

As described in the main paper, we make each of the potential phrases equally likely. An object is considered *small* if it has a total area between 32×32 and 64×64, *big* if it has a total area between 96×96 and 128×128 and has no modifier when the total area is between 64×64 and 96×96. To be described as moving *quickly*, the average difference in the center location of the object between frames is > 7, to be described as moving *slowly* the difference is < 3, otherwise, no descriptor is used. The distance moved is considered *a lot* if the total distance moved is greater than 30% of the image width and *a little* if the distance is less than 10% of the image width and otherwise without a descriptor. If $|\theta_k| < 8$ *slightly* is used to describe the rotation amount, if $|\theta_k| > 16$ *significantly* is used, otherwise the rotation amount is not described.

C.2 VindLU

When combining our approach with VindLU [10] we use the same encoders and hyperparameters as in the original VindLU paper where possible. Specifically, the visual encoder $\mathcal{F}_v$ is BEiT [5] pre-trained on ImageNet-21K [13]. Additional temporal attention modules are randomly initialized. The text-encoder $\mathcal{F}_t$ uses the first 9 layers of $\text{BERT}_{\text{BASE}}$ [14], with the cross-modal encoder using the last three layers of the same $\text{BERT}_{\text{BASE}}$ model. Since VindLU has a single stage of pre-training, we use our videos v_{motion} with our motion caption $t_{paraphrase}$ as well as the original caption t. Pre-training uses 4 video frames with the model optimized for 10 epochs using AdamW with an initial learning rate of 1e-4 and a minimum learning rate of 1e-6. The batch size is 32. In fine-tuning and evaluation we use 12 frames.

D Potential Negative Impact and Responsibility to Human Subjects

This paper makes use of the WebVid dataset [4] following prior work [10, 37]. The WebVid dataset uses publicly available data that the people in the videos have consented to share online, however, the authors did not specify whether the data has been filtered for offensive content. By adding generated motions to videos and replacing the spatial-focused caption with motion descriptions, our proposed approach does reduce spatial bias and instead learns a motion-focused representation. However, the model may still learn some biases present in the original dataset.

E Limitations and Future Work

There are several open avenues for future work based on the limitations of this paper. First, while there are a huge number of different generated motions possible, the motion generation only makes use of linear motion. If future work were

able to describe the trajectories of non-linear motion accurately, the complexity and possible descriptions of the motion could be further increased. Future work could also investigate whether generating longer motions is useful for tasks requiring long-range motion understanding. Secondly, we create motions using masked objects and add these moving objects to randomly sampled videos. This keeps the pretraining simple however, the resulting videos are not realistic. It is worth exploring whether the video sampled for the background of the object's motion affects the success of the pretraining or whether generating motions in the 3D space leads to more realistic videos which can reduce the domain gap between generated pretraining data and real fine-tuning data. Another direction worth exploring is whether such motion-focused video-text pairs are valuable for other video-language tasks such as in the alignment stage of large-scale VLM training or in text-to-video generation.