Supplementary Material for HiVLP: Hierarchical Interactive Video-Language Pre-Training

1. Details of the Pre-Training Datasets

Table 1 shows the statistics of the pre-training datasets for HiVLP*. The datasets SBU, CC3M, CC12M and LAION are downloaded according to the URLs given by BLIP [4] and we can only download most of them. Also for the WebVid-2M dataset [1], we can only download a subset of it. When using CC3M to train HiVLP, we use the raw captions from [5] instead of the filtered captions [4] for fair comparison with FiT [1].

2. Video-to-Text Retrieval of DiDeMo

For the down-stream task of video-to-text retrieval, we also compare our HiVLP* with CAMoE [3], which use CLIP's image and text encoders as its encoders. Note that, we do not use any post-processing technologies such as DSL [3] or QB-Norm [2]. As shown in Table 2, HiVLP outperforms CAMoE by a large margin.

3. Qualitative results of Video Captioning

Figure 1 shows the qualitative examples of HiVLP. It can be seen that in the given videos, HiVLP not only has the ability to recognize the visual contents (e.g., horse), but also can correctly give details about the actions (e.g., cooking).

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021.

[2] Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5194–5205, 2022.

[3] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. In *AAAI*, 2022.

[4] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[5] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

	COCO	VG	SBU	CC3M	CC12M	LAION	WebVid-2M
# images/videos	113K	100K	12.7M			100M	2.4M
# captions	567K	769K	12.7M			100M	2.4M

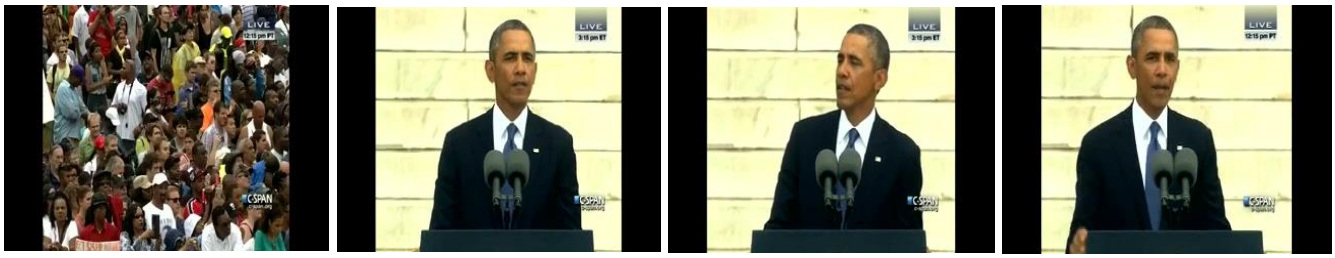
Table 1. Statistics of the pre-training datasets for HiVLP*.

Method	R@1↑	R@5↑	R@10↑
CAMoE[3]	45.5	71.2	-
HiVLP*	**48.5**	**74.2**	**81.4**

Table 2. Video-to-Text results on the DiDeMo dataset.

GT: a band is performing live music **HiVLP**: a band is performing a song

GT: a man is riding a horse **HiVLP**: a boy riding a show horse



GT: a man is giving a speech **HiVLP**: a man is giving a speech

GT: a person is playing a video game **HiVLP**: a man is playing a video game

GT: a person cooking in the kitchen **HiVLP**: a person is cooking

Figure 1. Examples of video captioning on the MSR-VTT dataset by HiVLP.