VicTR: Video-conditioned Text Representations for Activity Recognition - Supplementary Material

Kumara Kahatapitiya[1*] Anurag Arnab[2] Arsha Nagrani[2] Michael S. Ryoo[1,2]
[1]Stony Brook University [2]Google Research
kkahatapitiy@cs.stonybrook.edu

Details on auxiliary text classes: On Charades [6], we use 97 auxiliary classes: 43 objects, 15 places, 5 people-counts and 34 atomic-actions. People-count prompts are manually-selected, whereas the others are already annotated in the dataset. On Kinetics-400 [1], we use 88 auxiliary classes: 40 objects, 43 places and 5 people-counts. Atomic-actions on Kinetics-400 are too diverse to be categorized as a concise set, and thus omitted. On Kinetics-400, people-counts are similarly selected, and the others are generated by prompting ChatGPT3.5 with the set of 400 activity classes. The auxiliary vocabulary for each dataset is given below.

On Charades [6], we have the following:
`Objects`: *bag, bed, blanket, book, box, broom, chair, closet, cabinet, clothes, cup, glass, bottle, dish, door, doorknob, doorway, floor, food, groceries, hair, hands, laptop, light, medicine, mirror, paper, notebook, phone, camera, picture, pillow, refrigerator, sandwich, shelf, shoe, sofa, couch, table, television, towel, vacuum, window.*
`Places`: *basement, garage, pantry, recreation room, walk-in closet, laundry room, stairs, hallway, dining room, entryway, home office, bathroom, kitchen, bedroom, living room.*
`People`: *no people, one person, two people, three people, several people.*
`Atomic-actions`: *doing nothing, awakening, closing, cooking, dressing, drinking, eating, fixing, grasping, holding, laughing, lying, making, opening, photographing, playing, pouring, putting, running, sitting, smiling, sneezing, snuggling, standing, taking, talking, throwing, tidying, turning, undressing, walking, washing, watching, working.*

On Kinetics-400 [1], we have the following:
`Objects`: *bow and arrow, flowers, leaves or tree, computer, bed or baby crib, glass or bottle, dumbbell, treadmill or gym equipment, trampoline, mechanical bull or roller skates, bowling ball, cabinet or windows or dining table, sailboat or jet ski, fishing rod, cleaning supplies, grooming tools, pool, shoes, toilet, rope or ladder, barbecue grill or campfire, makeup tools, shovel, laundry or clothes, books or drawing materials, baseball, basketball or golf club, gymnastics mat, ice skates, dessert, fruits or vegetables, food items, fire extinguisher, hammer or meat grinder, musical instruments, board game, sporting equipment, gas pump, shopping cart, newspaper, animals, car, tractor or bicycle, rock climbing gear, electric sharpener or shredder.*
`Places`: *home, living room, dining room, bathroom, kitchen, bedroom, backyard or garden, staircase, hair salon, restaurant, outdoor, mountain or cliff, grass field, snow or ice, river or sea, sky, gym or fitness center, supermarket, foundary or workshop, forest, sports field, stadium, court or arena, massage palor, dance floor or stage, road or sidewalk, swimming pool, restaurant or bar, entrance or doorway, hospital or emergency room, bowling alley, building or skyscraper, theatre or auditorium, farm, recording studio or music room, news room, repair shop, garage, archery or shooting range, beach, underwater or sea bed, office or workspace, park, arcade or casino, school or classroom.*
`People`: *no people, one person, two people, three people, several people.*

On the selection of datasets: In literature, activity recognition is considered as the prominent video classification task. To understand the effectiveness of our *video-conditioned text* representations, we tackle a variety of activity recognition benchmarks. This includes few-shot and zero-shot activity recognition (on HMDB-51 [2], UCF-101 [7]), short-form recognition (on Kinetics-400 [1]) and long-form recognition (on Charades [6]). It is worth noting that Kinetics-400 usually contains single-person activities, whereas Charades includes multiple people and complex overlapping activities. Together, these provide a thorough spread of scenarios for both single-label and multi-label classification. Our evaluation setting is similar to many other prior work which evaluate on classification [3, 4, 9], yet extensive as it includes diverse contexts.

Compute requirement: Token-boosting increases the footprint of our model. However, our Video-Head is still lightweight, requiring minimal additional computations. In fact, it amounts for only 0.2% (0.5B) of total FLOPs in B/16 16-frame model (285B), and only 0.1% (0.6B) in

*Work done as a student researcher at Google.

Model	Rich text	HMDB-51	UCF-101
X-CLIP [4]	✗	44.6 ± 5.2	72.0 ± 2.3
VicTR (w/ CLIP Text emb.)	✗	43.9 ± 0.7	67.2 ± 0.7
VicTR	✗	51.0 ± 1.3	72.4 ± 0.3
VicTR (w/ CLIP Text emb.)	✓	43.9 ± 1.5	70.7 ± 0.3
VicTR	✓	52.1 ± 0.5	77.4 ± 0.2

Table A.1. **Impact of more-descriptive text:** We replace class labels in HMDB-51 [2] and UCF-101 [7] with rich class-descriptions generated by ChatGPT3.5. On zero-shot evaluation, our video-conditioned text embeddings benefit significantly-more from rich text inputs, compared to the CLIP [5] text embeddings.

L/14 8-frame model (656B). This is because of three reasons: (1) having fewer layers (*i.e*., 4 layers vs. 12/24 layers) and lightweight attention modules (*i.e*., temporal and cross-modal attention vs. spatial attention) compared to the image-VLM backbone [5], (2) processing significantly fewer tokens (*i.e*., only temporal and text-class tokens remain), and (3) doing text-conditioning only after the backbone (*i.e*., for the most part, all text embeddings go through shared computations). Ovrall, VicTR has a comparable footprint to prior work such as [3, 4, 9], providing a fair comparison (see respective GFLOPs in Table 3 and Table 4).

Other forms of semantic information: In our framework, we use a fixed vocabulary of auxiliary prompts as semantic inputs, that is specific to each dataset. Another way of providing semantic information is in the form of captions. If available, a detailed set of captions may provide better semantic supervision. However, they come with a significant cost, since they need to be annotated per-video. In contrast, our auxiliary prompts are freely-available and can be selected with only a minimal effort, as they are common for all videos in a dataset. Our model learns to highlight relevant information for a given video implicitly, via affinity weighting, without needing any ground-truth annotations.

Impact of more-descriptive text: By default, we use class labels with the standard CLIP [5] prompt template to generate text embeddings. However, if available, more-descriptive text such as human-annotated captions (expensive) or machine-generated descriptions (inexpensive) can provide richer information for our cross-modal attention, improving *video-conditioned text* representations. We validate this claim by replacing class-labels with rich class-descriptions from ChatGPT3.5 (Table A.1). On zero-shot evaluation, the relative gains from our text improve on both HMDB-51 [2] (+7.1% → +8.2%) and UCF-101 [7] (+5.2% → +6.7%), also raising the absolute performance.

Other reasoning tasks: The primary scope of this paper is on a broad spectrum of recognition tasks. Yet, it is also applicable to other reasoning tasks such as video VQA. In Table A.2, we evaluate VicTR on NExT-QA [11] under zero-shot settings, showing gains over comparable baselines

Model	Type	Params	NExT-QA
Random	-	-	20.0
CaKE-LM [8]	Enc-Dec	2.7B	34.9
InternVideo [10]	Enc-Dec	1.3B	49.1
SeViLA [13]	Enc-Dec	4.1B	63.6
Just-Ask [12]	Enc only	75M	38.4
X-CLIP [4]	Enc only	194M	43.8
VicTR (B/16)	Enc only	167M	45.5

Table A.2. **Video reasoning with VQA:** On NExT-QA [11] zero-shot evaluation, our model outperforms comparable baselines. Large-scale models with LLM decoders are de-emphasized.

with encoder-only designs (*i.e*., no LLM decoders). This validates that our model can readily be extended to other tasks with jointly-embedded video and text.

References

[1] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The Kinetics Human Action Video Dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[2] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In *ICCV*, pages 2556–2563. IEEE, 2011. 1, 2

[3] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen CLIP Models are Efficient Video Learners. *arXiv preprint arXiv:2208.03550*, 2022. 1, 2

[4] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding Language-Image Pretrained Models for General Video Recognition. In *ECCV*, pages 1–18. Springer, 2022. 1, 2

[5] 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*, pages 8748–8763. PMLR, 2021. 2

[6] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *ECCV*, pages 510–526. Springer, 2016. 1

[7] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv preprint arXiv:1212.0402*, 2012. 1, 2

[8] Hung-Ting Su, Yulei Niu, Xudong Lin, Winston H. Hsu, and Shih-Fu Chang. Language Models are Causal Knowledge Extractors for Zero-shot Video Question Answering. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4951–4960, 2023. 2

[9] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Action-

CLIP: A New Paradigm for Video Action Recognition. *arXiv preprint arXiv:2109.08472*, 2021. 1, 2

[10] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. InternVideo: General Video Foundation Models via Generative and Discriminative Learning. *arXiv preprint arXiv:2212.03191*, 2022. 2

[11] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 2

[12] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021. 2

[13] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36, 2024. 2