Supplementary Material: Audio-Enhanced Text-to-Video Retrieval using Text-Conditioned Feature Alignment

Sarah Ibrahimi[1*] Xiaohang Sun[2] Pichao Wang[2] Amanmeet Garg[2]
Ashutosh Sanan[2] Mohamed Omar[2*]
[1] University of Amsterdam [2] Amazon Prime Video
s.ibrahimi@uva.nl, {sunking, wpichao, amanmega, ashsanan}@amazon.com, mkamalmomd@gmail.com

1. Video-to-Text Retrieval Results

In our work, we focus on the task of text-to-video retrieval. We follow other works by also evaluating our model trained on video-to-text retrieval for MSR-VTT 9k, since competing methods have only provided video-to-text retrieval results on this data.

Method	R1 ↑	R5 ↑	R10 ↑	MdR ↓	MnR ↓
CLIP4Clip$_{meanP}$ [26]	43.1	70.5	81.2	2.0	12.4
X-Pool[16]	44.4	73.3	84.0	2.0	9.0
ECLIPSE$_{meanP}$† [22]	44.7	71.3	82.8	2.0	10.8
BridgeFormer*[12]	44.9	71.9	80.3	2.0	15.3
CAMoE [9]	45.1	72.4	83.1	2.0	10.0
TS2-Net [25]	45.3	74.1	83.7	2.0	9.2
X-CLIP [28]	46.8	73.3	84.0	2.0	9.1
TEFAL	**47.1**	**75.1**	**84.9**	2.0	**7.4**

Table T1. Video-to-Text Retrieval Results on MSR-VTT 9k split. All works use a CLIP ViT-B/32 backbone which is pre-trained on this Wikipedia-based image-text dataset.

2. Qualitative Results

In this section, we present additional examples on the MSR-VTT dataset [42] to highlight how audio provides complementary information to the video to achieve improved text-queried retrieval. The query words of sample 7152, visualized in Figure F1 is "a person is swimming in some white water rapids". While the video modality alone shows both the rapid water and the person, TEFAL w/o audio (video-only model) ranks the clip as the second matched retrieval. TEFAL, with the addition of the audio cue, correctly ranks the matched clip as the top retrieval. We notice that the presence of a person is confirmed by the voice in the latter part of the waveform (encircled in red), which clearly demonstrates that our model picks up complementary information from the audio modality. It is also observed that the first part of the clip is dominated by loud sound of streaming water but the water sound is greatly suppressed later in the clip though the continuous presence of water flowing in the video. This explicitly justifies building independent text-video and text-audio cross-modal attention blocks rather than aligning video and audio embeddings as it is in ECLIPSE [22], since the mandatory alignment between video and audio may introduce additional noise in the audiovisual feature.

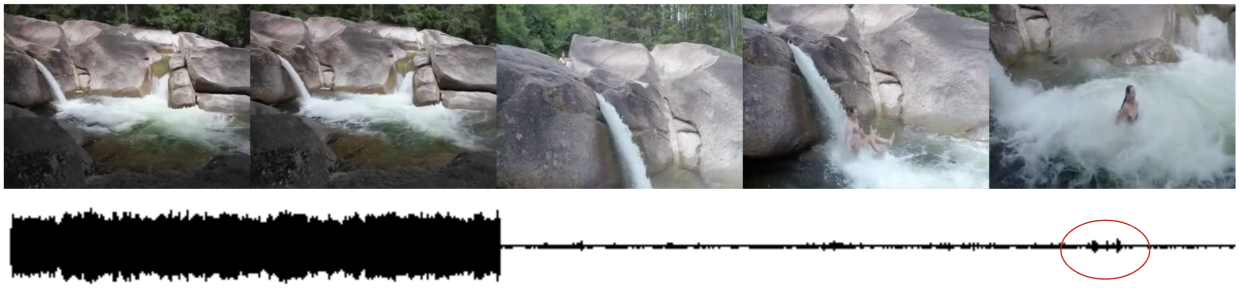
Additional examples are shown in Figure F2 to illustrate the correspondence between the text and audio modality that is otherwise missed between text and video. In the upper example, the girl talking can only be heard in the audio signal (sample 8827); in the middle example (sample 9233), the speech content is explicitly presented in the audio rather than video; and in the lower example, the word "oxiders" can only be matched in the audio from the man's talking (sample 9249).

3. Limitations

The main limitation of TEFAL is that without audio the method reduces to the text-video branch, and the performance is similar with XPool [16] (as indicated by Figure 1 in the main manuscript). If the missing audios are mostly in the train set, meta learning approaches could help lessen this issue [27].

*This work was done while Sarah Ibrahimi and Mohamed Omar were at Amazon Prime Video.

Rank #1 result in TEFAL, Rank #2 in TEFAL w/o audio

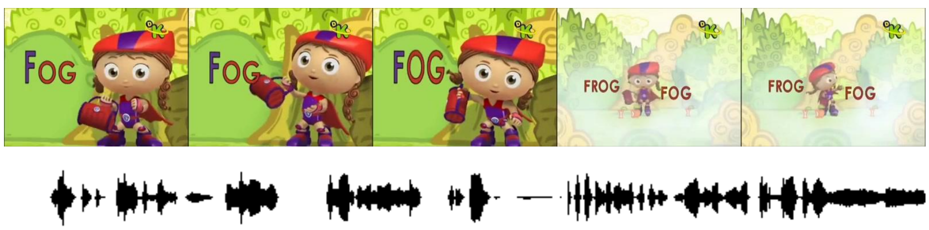


Query: a person is swimming in some white water rapids

Figure F1. In this figure an example is presented where a small sound has a large contribution to the final result. While TEFAL w/o audio is not able to select the correct video, TEFAL uses the audio to select the correct video as Rank 1

Rank #5 result in TEFAL, Rank #8 in TEFAL w/o audio

a)



Query: cartoon girl is talking

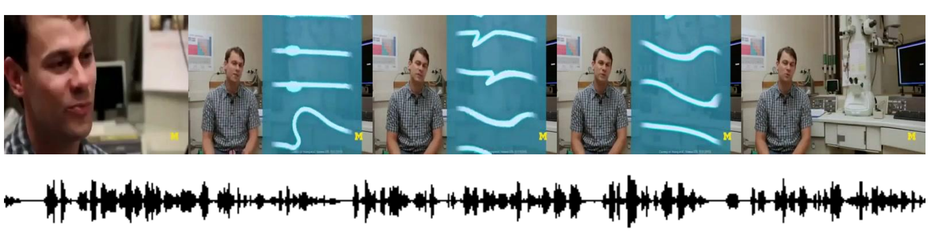
Rank #2 result in TEFAL, Rank #7 in TEFAL w/o audio

b)

Query: a man is giving a speech

Rank #1 result in TEFAL, Rank #7 in TEFAL w/o audio

c)



Query: a man is discussing oxiders in bulk form

Figure F2. This Figure shows three examples that illustrate the correspondence between the text and audio modality, that contains the verb "speaking", "talking" or a variation and specific words that correspond to the text query.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. VATT: transformers for multimodal self-supervised learning from raw video, audio and text. In *NeurIPS*, 2021.

[2] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. In *NeurIPS*, 2020.

[3] 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.

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *TPAMI*, 41(2), 2018.

[5] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38(8), 2022.

[6] Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *CVPR*, 2022.

[7] 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.

[8] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. *CVPR*, 2020.

[9] Xingyi Cheng, Hezheng Lin, Xiangyu Wu, F. Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021.

[10] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *ICCV*, 2021.

[11] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *TPAMI*, 44(8), 2021.

[12] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *CVPR*, 2022.

[13] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017.

[14] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. *NeurIPS*, 2020.

[15] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. In *Interspeech*, 2021.

[16] Satya Krishna Gorti, Noel Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *CVPR*, 2022.

[17] Wangli Hao, Zhaoxiang Zhang, and He Guan. Integrating both visual and audio cues for enhanced video caption. In *AAAI*, 2018.

[18] Fan Hu, Aozhu Chen, Ziyue Wang, Fangming Zhou, Jianfeng Dong, and Xirong Li. Lightweight attentional feature fusion: A new baseline for text-to-video retrieval. In *ECCV*, 2022.

[19] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, 2021.

[20] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *CVPR*, 2022.

[21] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *EMNLP*, 2020.

[22] Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. Eclipse: Efficient long-range video retrieval using sight and sound. In *ECCV*, 2022.

[23] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*, 2019.

[24] Yu Liu, Huai Chen, Lianghua Huang, Di Chen, Bin Wang, Pan Pan, and Lisheng Wang. Animating images to transfer clip for video-text retrieval. In *SIGIR*, 2022.

[25] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *ECCV*, 2022.

[26] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *Neurocomputing*, 508:293–304, 2022.

[27] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *AAAI*, 2021.

[28] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACM MM*, 2022.

[29] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a Text-Video Embedding from Imcomplete and Heterogeneous Data. *arXiv preprint arXiv:1804.02516*, 2018.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[31] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[32] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[33] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *CVPR*, 2015.

[34] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogério Feris, David Harwath,

James R. Glass, and Hilde Kuehne. Everything at once - multi-modal fusion transformer for video retrieval. In *CVPR*, 2022.

[35] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *ECCV*, 2016.

[36] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021.

[37] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[38] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. In *NeurIPS*, 2022.

[39] Jianren Wang, Zhaoyuan Fang, and Hang Zhao. Alignnet: A unifying approach to audio-visual alignment. In *WACV*, 2020.

[40] 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.

[41] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: global-local sequence alignment for text-video retrieval. In *CVPR*, 2021.

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

[43] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: a survey. *arXiv preprint arXiv:2206.06488*, 2022.

[44] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. MERLOT RESERVE: neural script knowledge through vision and language and sound. In *CVPR*, 2022.

[45] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *ECCV*, 2018.

[46] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. Centerclip: Token clustering for efficient text-video retrieval. In *SIGIR*, 2022.

[47] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, 2020.