In-Style: Bridging Text and Uncurated Videos with Style Transfer for Text-Video Retrieval Supplementary Material

Nina Shvetsova*[1,2,3] Anna Kukleva*[2] Bernt Schiele[2] Hilde Kuehne[1,3,4]
[1]Goethe University Frankfurt, [2]Max-Planck-Institute for Informatics, [3]University of Bonn, [4]MIT-IBM Watson AI Lab
{nshvetso,akukleva}@mpi-inf.mpg.de

In the supplementary material, we first elaborate on some discussions in Section A; further, we provide In-Style Method details in Section B and implementation details in Section C; and finally, we provide more qualitative evaluations in Section D and discuss limitations in Section E.

A. Additional Discussions

Text Query Style. Figure 1 and Figure 2 show the respective word clouds for the five datasets with and without stop words. In Table 1, we show five different text examples from the different datasets considered in this paper to further highlight differences of the styles.

Model Generalization. In the following, we discuss the generalization performance of the model trained on different text styles (Table 2 in the main paper.) When we consider models trained only with one text style (text queries that are only from one dataset) in the top half of Table 2, it shows that the mean retrieval performance is higher for the queries with the style of MSR-VTT, DiDeMo, or LSMDC datasets. Interestingly, MSVD-style text queries, which are similar to MSR-VTT queries in terms of sentence structure (Table 1), and usage of stop words (Figure 1 and Figure 2), show similar MSR-VTT retrieval performance compared to DiDeMo and LSMDC-style text queries, and moreover, lower performance on DiDeMo, and LSMDC datasets. We hypothesize that longer and more descriptive text is beneficial for model generalization (the MSVD dataset contains the shortest text descriptions of all datasets). In Table 2, we also report the number of shared video clips in generated pairs P_{gen} with different datasets text styles. Interestingly, P_{gen} based on the LSMDC text queries has the smallest number of pairs in general. Moreover, even though it has low overlap in videos with P_{gen} from all other datasets (0.18–0.24, see in Table 2b), LSMDC's P_{gen} shows one of the highest generalization to MSVD, DiDeMo, and MSR-VTT datasets. Analyzing examples in Table 1, we suggest that LSMDC captions are more concise and descriptive in terms of object-verb-subject details (who does what in a video). For example, in the 3rd MSR-VTT example, there is only a general description that people are fighting without specification of exact people and their actions, and in the 5th example, the caption only says that a man is playing an instrument, without specification of an instrument. At the same time, in LSMDC texts, the object-action-subject description is more detailed, such as in the first example, all consecutive actions are specified, or in the 2nd example, the exact hitting action ("he slaps") is specified. Therefore, we hypothesize, that a text style with more concise descriptions is better for model generalization. However, as Table 2 demonstrates, the best mean performance over all datasets is achieved while training with different target text styles.

*Equal contribution.

B. In-Style Method Details

Captioner. For the captioner, we follow BLIP [5] image captioner architecture, which we extend to video captioning as shown in Figure 3a. Namely, we encode m uniformly sampled frames (we use $m = 8$ for training and $m = 12$ for inference) from a video by the image transformer to obtain frame-wise tokens. Then, we feed this set of encoded tokens from all the frames into the cross-attention of an image-grounded text decoder. Therefore, the predicted text is conditioned on multiple video frames at once. The image-grounded text decoder predicts the next text token given an input of previous text tokens (where the "dec" token is concatenated to the beginning of the input sequence and denotes the start of the output). During the inference, the text tokens are generated one by one in an autoregressive manner.

Video-Text Dual Encoder. In the pseudo matching, filtering, and retrieval steps of our In-Style method, we use the video-text dual encoder model. We initialize the dual encoder model from image-text pre-trained CLIP [13] or BLIP [5] models, which we extend to video-text models. Specifically, we obtain a video embedding by averaging image representations of m uniformly sampled frames, as

Dataset	Examples
MSR-VTT (∼43 symbols in a text)	1) A bulldozer removes dirt 2) An infomercial with a pharmeceutical company talking about an epilepsy drug pending approval from the FDA 3) Extreme violence scenes with people fighting with each other 4) A woman is in front of a whiteboard talking about the numbers written on it 5) A man is playing an instrument
YouCook2 (∼39 symbols in a text)	1) Add some herb sprinkle and stir the meat 2) Bake the pizza on the grill 3) Pour butter into the wok 4) Peel an onion and chop into pieces 5) Add the tomato paste crushed tomatoes tomato puree and beef stock to the pan
DiDeMo (∼147 symbols in a text)	1) First time we see the dancers go down on one leg the men hit the ground with their sticks. They first start crouching and hitting the ground with the sticks 2) When the man puts his head down the guitar player is looking up. The guitarist is looking straight up. A man plays the guitar while looking up. The guitarist is looking straight up as he plays. 3) Red phone booth is visible a red phone booth is in the scene. A person walks in the middle of the camera. A red phone booth can be seen a red telephone booth is on the sidewalk. 4) The camera moves back to the left to the tree. White square exits frame left the camera pans back the way it came. Square area lines with stones comes into view the fence comes into view. 5) Fog moves in toward the ice skater. a woman spins around several times very fast. A woman pirouettes as she comes near the camera. Woman spins more than 5 times in a row.
MSVD (∼31 symbols in a text)	1) A lady is pouring raw strawberry juice into a bowl 2) A man is slicing the crust into a potato 3) A man lifts three sunflowers 4) A man is putting a pan into an oven 5) A boy rides around in circles on a tricycle
LSMDC (∼46 symbols in a text)	1) The dish is covered in saffron and spices. 2) He slaps SOMEONE again. 3) He and SOMEONE join forces to grab the cube, which's connected with several more wire. 4) In the race, a rider falls. 5) SOMEONE dashes to a clothes closet and ducks inside. The cup spins across the floor.
VATEX (∼71 symbols in a text)	1) Someone is demonstrating how to paint a metal sheet on a window 2) A person is cooking scallops in the pan over a fire place and then begins to pour them in some water 3) Two women make a video tutorial on how to bake cookies 4) A person is throwing garbage into the trash can and talking 5) A woman is outside and preparing an edible meal by inserting herbs into it, then placing them on the ground
Food.com (∼54 symbols in a text)	1) In a blender or food processor, puree the first 3 ingredients until smooth 2) Combine juice, remaining 2 tablespoons sugar and lemon juice 3) Place chicken on wire rack 4) Then toss in the artichokes and serve immediately 5) Cut beef into 3 - inch pieces
WikiHow (∼41 symbols in a text)	1) Do lunges in the park 2) Cut a large milk jug 3) Buy your favorite knit kit 4) Trim the lining of your sweater 5) If your child is in a fight, put his hands on the sand

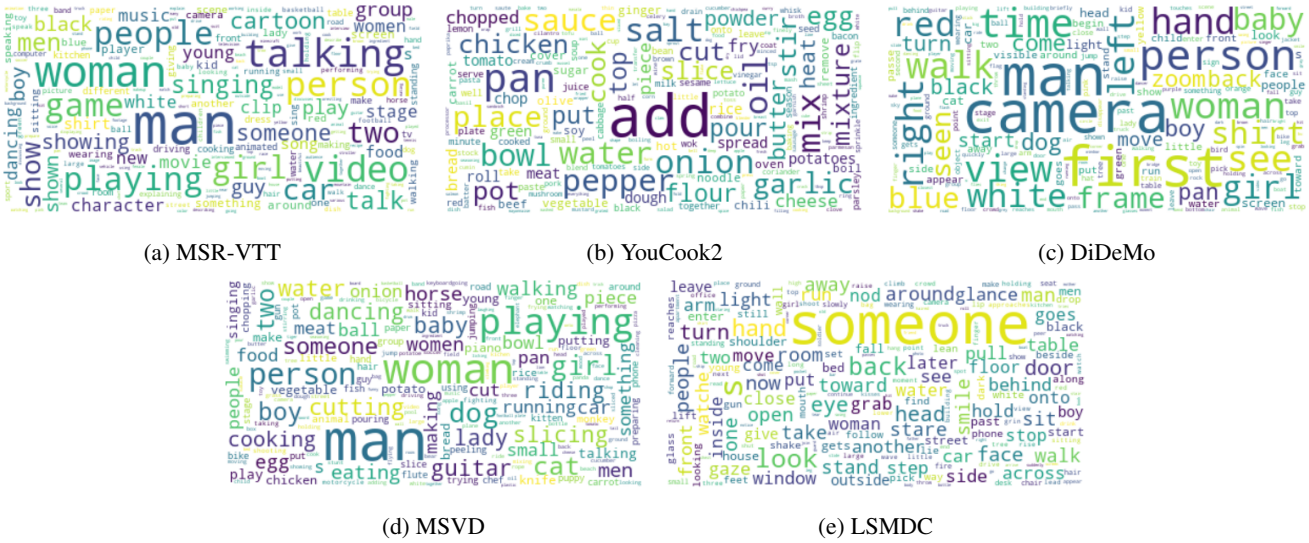
Table 1: Five random examples of text descriptions in different datasets. With the dataset name, we also report the median length of a text in the dataset.

shown in Figure 3b. By default, eight frames are used $m = 8$ during training and $m = 12$ during evaluation. But m is increased to 64 during evaluation on the fine-graned DiDeMo dataset as in [6, 15]. To obtain image representation, image patches are fed into ViT transformer [3], and the "cls" output token is later projected by a linear projection into common embedding space and further normalized. The text representation is obtained by projecting and normalizing an output "cls" token in the case of BLIP and an output "eot" token in the case of CLIP.

(a) MSR-VTT (b) YouCook2 (c) DiDeMo

(d) MSVD (e) LSMDC

Figure 1: Word clouds for text queries from different datasets (stop words are **included**).



(a) MSR-VTT (b) YouCook2 (c) DiDeMo

(d) MSVD (e) LSMDC

Figure 2: Word clouds for text queries from different datasets (stop words are **excluded**).

C. Implementation Details

EOA Model Details. In case of the EOA [14] architecture, we use the model variant with S3D visual backbone with a frozen S3D backbone pre-trained by Miech et al. [9] in the visual branch and a GoogleNews pre-trained Word2vec model [11] in the text branch. These backbones are fixed and not trained. As initialization, we use EOA [14] weights pre-trained with audio modality, but we discard the audio branch to consider text-video-only retrieval.

EOA Training Details. We follow [14] to train the EOA model. We compute one S3D feature per second [9] without any data augmentation; namely, we compute one feature per 16 frames, sampling frames with 16 fps and 224x224 resolution. We train with Adam [4] optimizer, no weight decay, and a batch size of 128. We use a temperature of 0.05 in the loss function.

BLIP Evaluation Details. We would like to highlight that while evaluating a model with a BLIP backbone, we use dual encoder architecture (unimodal encoders), not a cross-attention architecture (named an image-grounded text encoder in [5]), which performance on MSR-VTT dataset [16] was reported in the original paper [5]. Dual encoder models

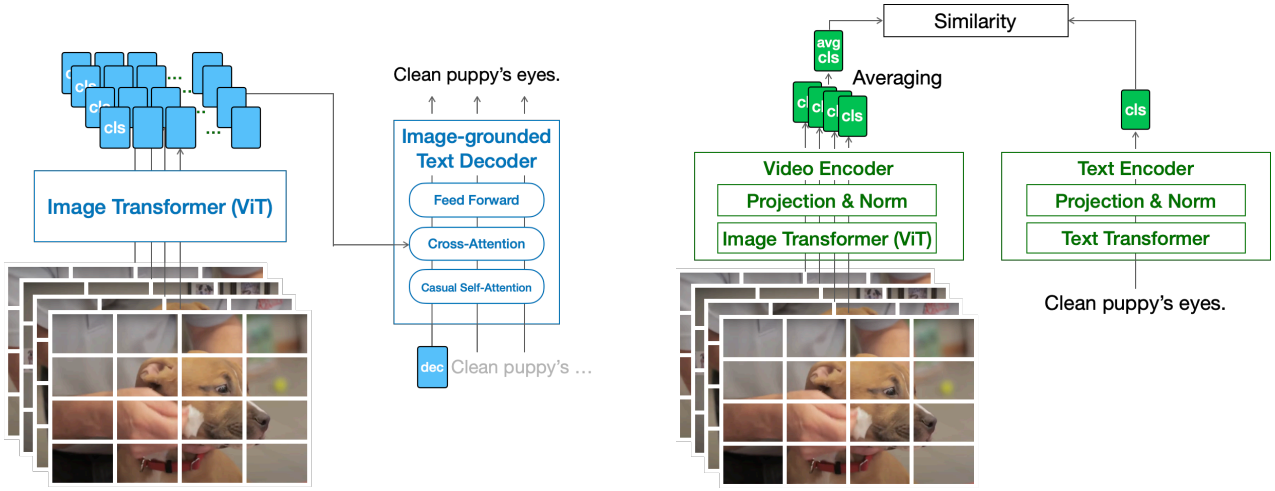
Dataset	MSR-VTT	YouCook2	DiDeMo	MSVD	LSMDC
MSR-VTT	495k	111k	160k	239k	87k
YouCook	111k	168k	65k	97k	40k
Didemo	160k	65k	280k	135k	57k
MSVD	239k	97k	135k	379k	75k
LSMDC	87k	40k	57k	75k	144k

(a) Number of shared video clips

Dataset	MSR-VTT	YouCook2	DiDeMo	MSVD	LSMDC
MSR-VTT	1	0.23	0.32	0.48	0.18
YouCook2	0.66	1	0.39	0.58	0.24
DiDeMo	0.57	0.23	1	0.48	0.2
MSVD	0.63	0.26	0.36	1	0.2
LSMDC	0.61	0.28	0.4	0.52	1

(b) Ratio of shared video clips per dataset

Table 2: Number/Ratio of shared video clips in the datasets' generated pairs P_{gen}.



(a) Captioner

(b) Video-Text Dual Encoder

Figure 3: Schematic visualization of (a) the video captioner architecture; and (b) the video-text dual encoder model.

independently encode videos and texts into common embedding space, allowing for fast retrieval among thousands of videos by pre-computing video embeddings and calculating the similarity between text and video embeddings with a dot product (cosine similarity) [8]. Cross-attention architectures compute similarity by propagating video and text together in the model with cross-attention layers, attending all words and all spatial-temporal video patches to each other. Cross-attention architecture significantly boosts retrieval performance compared to dual encoder models [8]; however, it demands enormous computational overhead in the inference phase, requiring propagating all videos paired with a given text query to compute similarity. In the original paper [5], BLIP performance on the MSR-VTT dataset was reported with the cross-attention model used to rerank 128 closest videos found by the dual encoder model. Since the majority of the methods [13, 14, 12, 10, 6, 1] leverage dual encoder architecture for video retrieval due to computational benefits, for comparison purpose we also base our method on dual encoder models.

D. Qualitative Results

Pseudo pairs P_{ps} and generated pairs P_{gen}. We provide additional qualitative results for pseudo pairs P_{ps} and generated pairs P_{gen} with the MSR-VTT dataset in Figure 4, the YouCook2 in Figure 5,the DiDeMo in Figure 6, the MSVD in Figure 7, and the LSMDC dataset in Figure 8. We observe on various datasets that generated captions capture content better than in initially obtained pseudo pairs after the matching step. It confirms our discussion of Table 4 (in the main paper), that generated pairs P_{gen} on average provide better improvement than pseudo pairs P_{ps}.

Text-video retrieval. We also demonstrate qualitative results of text-video retrieval on the MSR-VTT (Figure 9), YouCook2 (Figure 10), DiDeMo (Figure 11), MSVD (Figure 9), and LSMDC (Figure 13) datasets. We found that the proposed In-Style model retrieves more semantically similar videos to a given query compared to the zero-shot BLIP model.

E. Limitations

In this work, we rely on the pre-trained large image-language models such as CLIP [13] and BLIP [5]. We consider this as an advantage and disadvantage at the same time. On one side, we show how to adapt such models to the input style of text queries, whereas, on the other side, we inherit all the biases that such models include [2]. Moreover, CLIP has the property that it can read text from the images [7]; therefore, matching or filtering steps could suffer from that because some unrelated text (e.g. advertisement) appears on the frames (e.g. see Figure 4).

While our motivation is to avoid annotation costs for aligning text-video pairs, we still rely on a video collection; namely, we utilize a large-scale dataset of YouTube videos. However, we note that such kind of videos are easy to collect as they are available on YouTube and are not pre-processed; therefore, these videos include various types of potential noise from web as advertisements, camera motion, low quality of videos and other. Moreover, even though the distribution of our input text queries is not the same as the distribution of web support videos, we empirically observe that there is always some overlap between the distributions. Thence, we do not assume that our In-Style method will be helpful when input queries and test sets are from absolutely different domains like medicine (input text queries) and wildlife (test). As we show in our experiments, we obtain the most gain when the input text queries and test sets are from the same distribution.

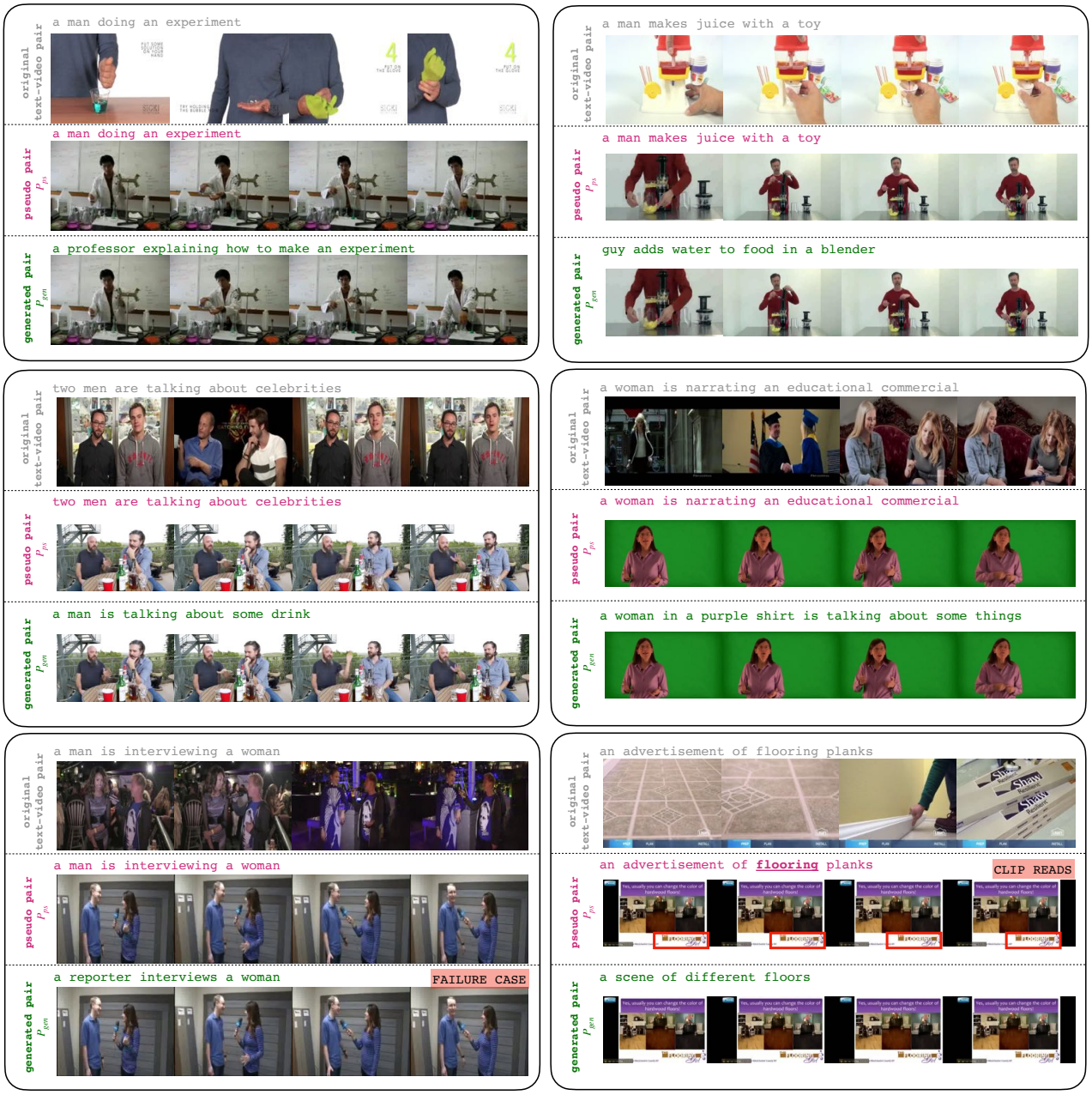


Figure 4: **Qualitative evaluation of P_{ps} and P_{gen} on the MSR-VTT.** First, a text query is matched with one of the videos (a pseudo pair P_{ps}), and then, after the style preservation step, for each video new caption is generated in the same style but with updated content (a generated pair P_{gen})

Figure 5: **Qualitative evaluation of P_{ps} and P_{gen} on the YouCook2.** First, a text query is matched with one of the videos (a pseudo pair P_{ps}), and then, after the style preservation step, for each video new caption is generated in the same style but with updated content (a generated pair P_{gen})

Figure 6: **Qualitative evaluation of P_{ps} and P_{gen} on the DiDeMo.** First, a text query is matched with one of the videos (a pseudo pair P_{ps}), and then, after the style preservation step, for each video new caption is generated in the same style but with updated content (a generated pair P_{gen})

Figure 7: **Qualitative evaluation of P_{ps} and P_{gen} on the MSVD.** First, a text query is matched with one of the videos (a pseudo pair P_{ps}), and then, after the style preservation step, for each video new caption is generated in the same style but with updated content (a generated pair P_{gen})

Figure 8: **Qualitative evaluation of P_{ps} and P_{gen} on the LSMDC.** First, a text query is matched with one of the videos (a pseudo pair P_{ps}), and then, after the style preservation step, for each video new caption is generated in the same style but with updated content (a generated pair P_{gen})

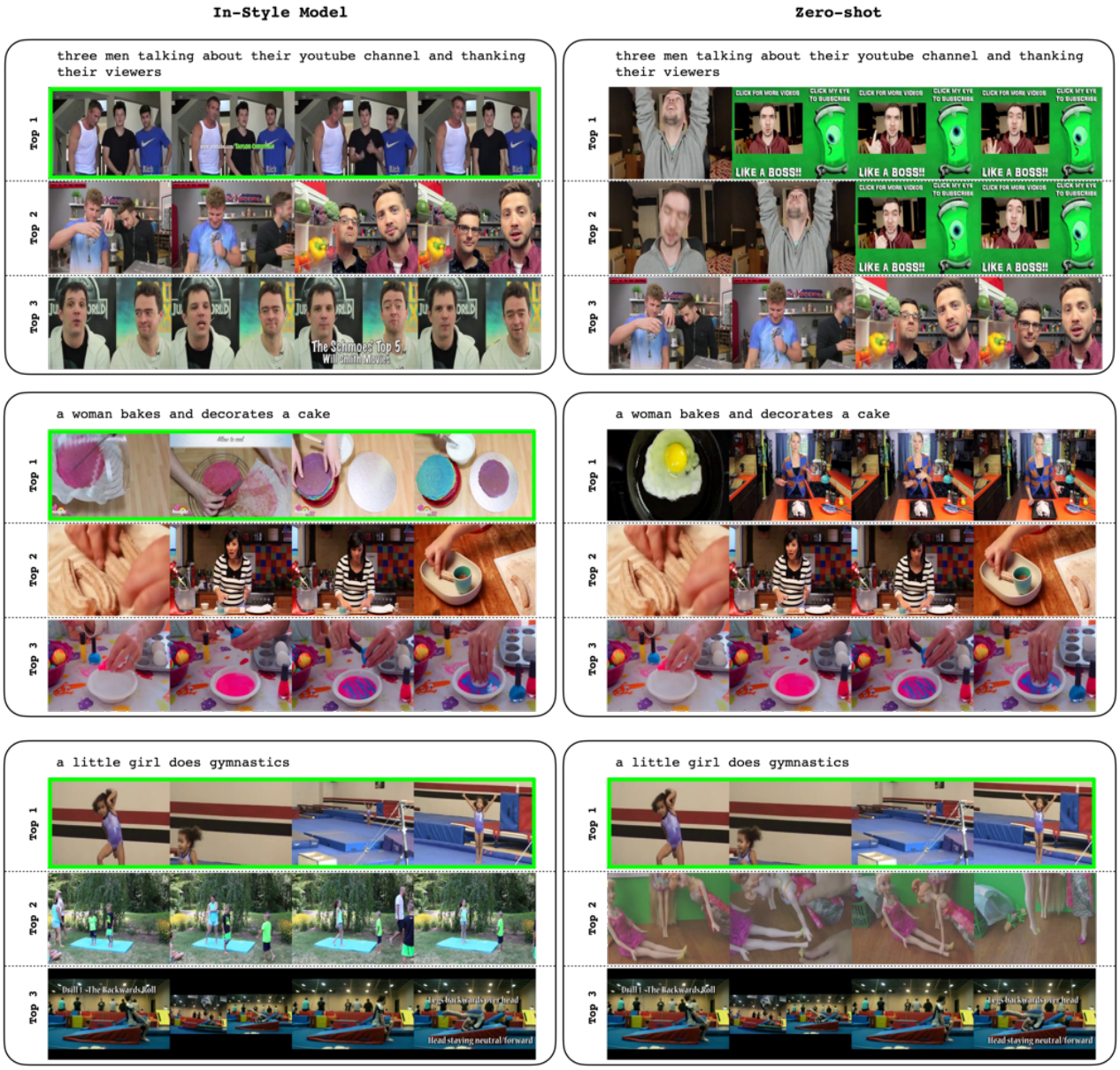


Figure 9: **Qualitative evaluation of text-video retrieval on the MSR-VTT.** Retrieval examples for the proposed In-Style Model and zero-shot BLIP model. Each box shows the top-3 retrieved videos for a given text query. The correct video is highlighted with a green color.

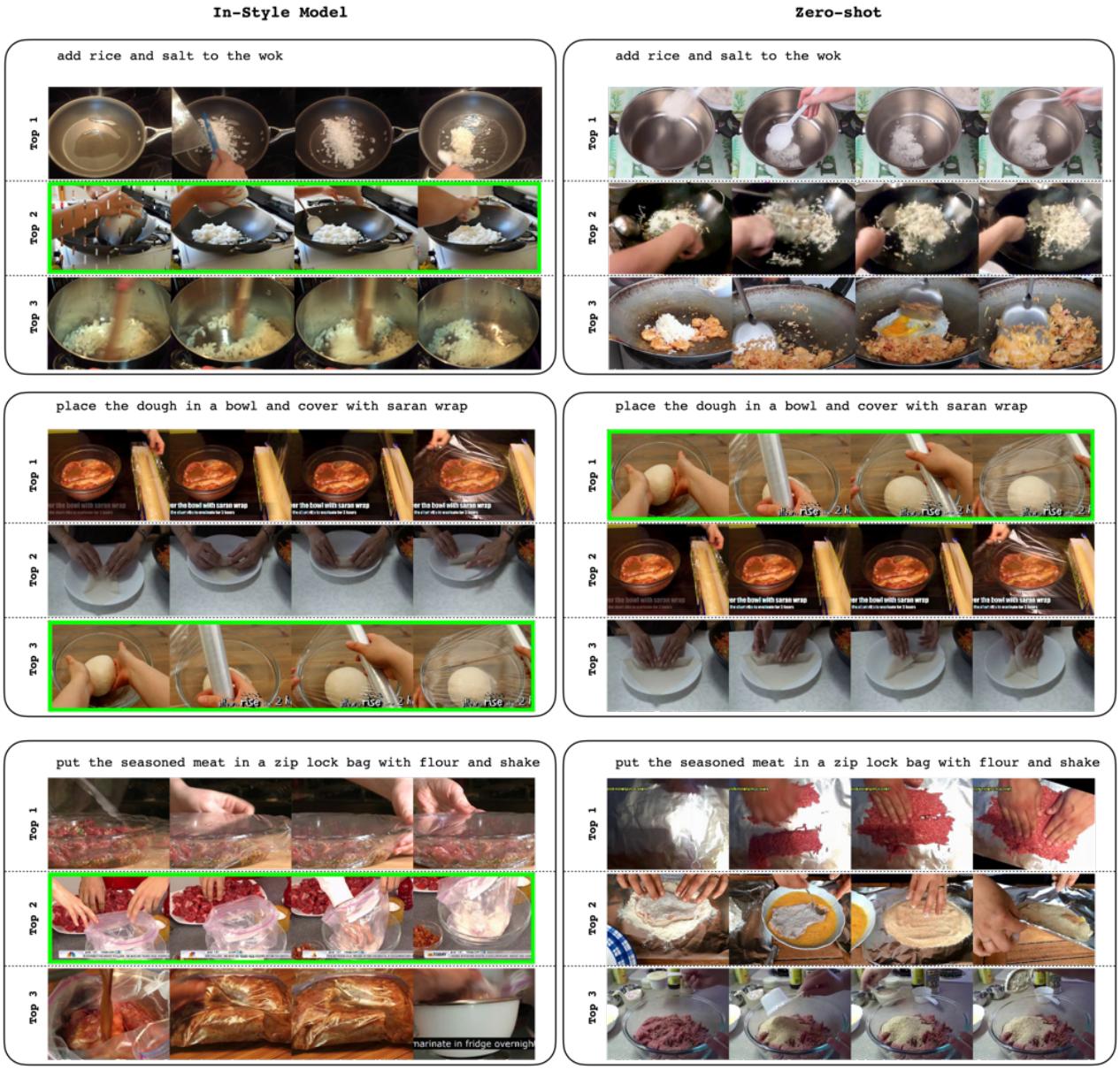


Figure 10: **Qualitative evaluation of text-video retrieval on the YouCook2.** Retrieval examples for the proposed In-Style Model and zero-shot BLIP model. Each box shows the top-3 retrieved videos for a given text query. The correct video is highlighted with a green color.

Figure 11: **Qualitative evaluation of text-video retrieval on the DiDeMo.** Retrieval examples for the proposed In-Style Model and zero-shot BLIP model. Each box shows the top-3 retrieved videos for a given text query. The correct video is highlighted with a green color.

Figure 12: **Qualitative evaluation of text-video retrieval on the MSVD.** Retrieval examples for the proposed In-Style Model and zero-shot BLIP model. Each box shows the top-3 retrieved videos for a given text query. The correct video is highlighted with a green color.

Figure 13: **Qualitative evaluation of text-video retrieval on the LSMDC.** Retrieval examples for the proposed In-Style Model and zero-shot BLIP model. Each box shows the top-3 retrieved videos for a given text query. The correct video is highlighted with a green color.

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. 4

[2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 5

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 3

[5] 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. 1, 3, 4, 5

[6] 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 and captioning. *Neurocomputing*, 508, 2022. 2, 4

[7] Joanna Materzyńska, Antonio Torralba, and David Bau. Disentangling visual and written concepts in clip. In *CVPR*, 2022. 5

[8] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *CVPR*, 2021. 4

[9] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 2020. 3

[10] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 4

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 3

[12] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *ICLR*, 2021. 4

[13] 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*, 2021. 1, 4, 5

[14] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multi-modal fusion transformer for video retrieval. In *CVPR*, 2022. 3, 4

[15] Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. Cap4video: What can auxiliary captions do for text-video retrieval? In *CVPR*, 2023. 2

[16] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 3