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

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

Abstract

Large-scale noisy web image-text datasets have been proven to be efficient for learning robust vision-language models. However, to transfer them to the task of video retrieval, models still need to be fine-tuned on hand-curated paired text-video data to adapt to the diverse styles of video descriptions. To address this problem without the need for hand-annotated pairs, we propose a new setting, text-video retrieval with uncurated & unpaired data, that uses only text queries together with uncurated web videos during training without any paired text-video data. To this end, we propose an approach, In-Style, that learns the style of the text queries and transfers it to uncurated web videos. Moreover, to improve generalization, we show that one model can be trained with multiple text styles. To this end, we introduce a multi-style contrastive training procedure, that improves the generalizability over several datasets simultaneously. We evaluate our model on retrieval performance over multiple datasets to demonstrate the advantages of our style transfer framework on the new task of uncurated & unpaired text-video retrieval and improve state-of-the-art performance on zero-shot text-video retrieval. [1]

1. Introduction

Vision-language retrieval refers to the task of retrieving an image or a video from a large data pool given a textual description of the content. Especially the field of text-image retrieval has seen remarkable progress, mainly spurred by the combination of image and text models trained on large-scale web collections [47, 31] of image-text pairs. While advances in video retrieval also rely on pre-trained image-language models, which serve for better task transfer, most systems still require a fine-tuning on downstream data. This requires hand-annotated text-video pairs, namely a trimmed segment of a larger video that is precisely described by the corresponding text pair, for the training and testing of each target downstream dataset. Collecting such aligned pairs of text and videos can be time and cost intensive, and particularly gathering videos that comply with national regulations and copyright can be a challenge. Also, in case of relying on free web content, some videos can become unavailable over time while the respective curated annotations stay available for download but do not have matching videos.

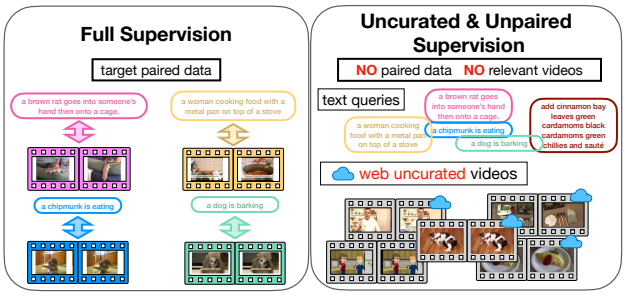

Figure 1: **Training data for supervised and uncurated & unpaired settings for text-video retrieval. Left**: standard supervised text-video retrieval, aligned and paired data is given for each target setting of the same distribution as target test set; **Right**: our uncurated & unpaired text-video retrieval setting. No paired data is available during training, only text queries, whereas to support training, we use uncurated web videos.

To address this problem, we propose a new setup, *text-video retrieval with uncurated & unpaired data*, assuming the availability of text queries only and without related videos during training (Figure 1). The setting is motivated by the fact that it can be considered easier to collect or generate text data, e.g. by producing topic-specific text queries, rather than providing a video to match a specific context. To allow the training of a text-video retrieval system based on the given text, we assume to have access to an uncurated video collection as the only source of available videos.

As different domains and datasets contain diverse styles of textual descriptions of videos, we propose a novel method, *In-Style*, to transfer the caption style of given text queries to uncurated web videos, which can be from a deviating distribution compared to the given text queries. To

*Equal contribution.
[1] github.com/ninatu/in_style

transfer the style of the text queries, we leverage large image-language models [31, 47] by creating pseudo pairs that correspond to the given text queries and videos from the uncurated collection by matching them in the shared embedding space [47]. Thus, we identify a subset of videos that have more similarity to the text queries than the rest of the videos. We then adopt an image-to-text captioning model (captioner) to mimic the style of our text queries by training with these pseudo pairs. The stylized captioner is now capable of producing relevant video descriptions in the desired style; therefore, we re-annotate the web videos with the captioner to obtain aligned paired data; we call them generated pairs. Finally, we show that generated pairs help to adapt models pre-trained on large-scale web data [31, 52] to the desired single or multiple styles of given text queries.

We evaluate our model on text-video retrieval over 5 benchmark datasets. Specifically, we demonstrate the advantages of the In-Style method on the new task of uncurated & unpaired text-video retrieval with image-language [31] and video-language [52] pre-trained backbones. We show the generalization of the proposed approach by training a single model for multiple datasets at once leading to an improved state-of-the-art zero-shot text-video retrieval performance.

We summarize our contributions in the following: (i) we introduce a new task of *text-video retrieval with uncurated & unpaired data* where during training, only text queries are available, whereas for standard text-video retrieval task, paired text-video data is used; (ii) we propose a novel method, In-Style, to transfer the style of text queries in an unsupervised way, showing that style is an important component for language-based retrieval tasks; therefore, we repurpose large pre-trained image-language models to generate pseudo-captions of the same style for uncurated web videos; (iii) we demonstrate the advantages of our In-Style method for the new task over 5 different datasets with individual models for each dataset as well as one generalized model and we achieve state-of-the-art performance on zero-shot text-video retrieval.

2. Related Work

Text-Video Retrieval. Text-video retrieval methods usually focus on learning modules that are able to capture relations between features from text and video modalities [63, 34, 17, 9, 12, 15, 57]. Currently, many approaches leverage pre-training on large-scale video-text [3, 41, 40] or image-text [29, 31] datasets with a further adaptation of the backbone to individually downstream datasets. In this context, ClipBERT [29] proposed sparse sampling instead of using dense full-length videos that allow lightweight training. However, foundation models [5] such as CLIP [47], combining the success of transformer architectures [14] using a contrastive objective [43] and being trained on large collections of text-image pairs from the web, providing a strong zero-shot [38, 46] baseline on downstream tasks that outperforms many previous methods. Therefore, more recent approaches focus on adapting text-image CLIP pre-trained models for text-video retrieval [20, 4, 16, 18, 35]. X-pool [20] introduces cross-modal attention to reason between text and frames of a video, TS2-Net [35] proposes dynamic adjustments over temporal and spatial token dimensions, which allows fine-tuning spatial model on video data without architecture changes. Another way to leverage foundation models is to enhance training data [58, 65]. Cap4Video [58] generates auxiliary captions for available curated training videos by using ZeroCap [54] that optimizes GPT-2 [48] text generation using a CLIP-based loss [47]. LaViLa [65] proposes to generate additional narrations for a dense coverage of long videos from the Ego4D dataset [11, 21] by fine-tuning a pre-trained large language model [48] on existing annotated text-video paired data. In contrast, we propose to exclude pre-annotated text-video paired data from the training and, relying on text descriptions only, generate text-video pairs leveraging uncurated web videos while transferring the style of original captions.

Large-scale Multimodal Pre-training. Representation learning [47, 31, 10, 23, 65, 6, 62] aims to obtain general representations that improve performance on downstream tasks such as retrieval [38, 46, 62, 31], classification [10, 65, 6], segmentation [6], question-answering [31, 62] and captioning [31, 62]. While some methods rely only on one modality such as images [6, 23] or text [48], there is also increasing interest in multi-modal representations [47, 30, 53, 32, 37, 52] which require multi-modal aligned pairs. However, the acquisition of human-annotated paired data is expensive; therefore, noisy web data [47, 41] allows to significantly scale such datasets. Many methods successfully utilize web image-text pairs [47, 25, 62], whereas uncurated video-text pairs are not only harder to collect but are also more prone to misalignments. Therefore, efforts are made to align ASR (automatic speech recognition) with video frames via contrastive learning [41, 40, 60, 64] or in an unsupervised way [22]. To overcome those issues, we propose to generate synthetic video descriptions with the desired caption style and train models on those captions instead of raw ASR text.

For contrastive-based vision-language representation learning methods, dual-encoder architectures are a common choice as it features two parallel branches for two modalities which are contrasted against each other to learn a joint embedding space [47, 53, 32, 37]. Recently, BLIP [31] and CoCa [62] propose a unified multi-task contrastive-generative framework that combines contrastive and captioning objectives. These methods rely on both, curated image-text and uncurated web image datasets, with BLIP additionally iteratively applying the generation and filtering

of synthetic captions. Compared to those works, we adopt pre-trained image-language models for uncurated & unpaired text-video retrieval by transferring the caption style directly on uncurated videos without any aligned data during training.

3. Uncurated & Unpaired Text-Video Retrieval

In this section, we introduce the proposed uncurated & unpaired text-video retrieval training setup. Typically, models for text-video retrieval are trained on *paired* text-video data. Given a set of pairs of captions t_i and corresponding videos v_i: $\{(t_i, v_i)\} \in D$, where D is a data distribution, the goal is to learn a similarity function $s(t_i, v_j)$ that calculates the similarity between the caption t_i and the video v_j. The training can be done from scratch, but typically pre-trained image-language [47, 31] or video-language models [37, 41] are fine-tuned on the target paired text-video data and then evaluated on the test set from the same distribution D [35, 16]. If the evaluation is performed on multiple datasets, the model is usually fine-tuned for each dataset individually.

In contrast, we propose *a text-video retrieval with uncurated & unpaired data*, where only target text queries are available during training without any videos. More precisely, given a set of text descriptions $\{t_i\}$ from data distribution D, we aim to learn useful information about the similarity $s(t_i, v_j)$ in D relying only on the clean textual descriptions. We further assume that a large set of freely accessible web videos $V' = \{v'_j\} \in D'$ without any paired text is available to support the training (such as videos of the HowTo100M dataset [41]). We note that the data distribution D' in the support video dataset can deviate from the distribution D.

Finally, to avoid to train different models individually for each target dataset, we further consider learning a *generalized* model that maintains the performance of individual models over a set of K datasets of different caption styles and coming from different data distributions $D_1, ..., D_K$.

4. In-Style Method

To address the task of uncurated & unpaired text-video retrieval, we aim to transfer the style of the text queries (the only available curated information) to an uncurated web video dataset. To this end, we rely on web-scale pre-trained image-language models as a supervisory signal and leverage them as a matching module and pre-trained captioning model that we adapt throughout the training process. The steps of the proposed In-Style method are shown in Figure 2. The first step is *Pseudo Matching*, described in Section 4.1, which matches the given text queries to the most relevant videos from the set of all uncurated web videos. The following *Style Transfer* step (Section 4.2) adapts the pre-trained captioning model (captioner) to the target text style by training it on the previously obtained pseudo pairs. The captioner is then used to generate new style-adapted captions for all available web videos, which are then filtered to avoid too noisy pairs; we refer to the resulting filtered web videos with style-adapted video descriptions as generated pairs. Finally, we adapt a pre-trained vision-language model for the task of text-video retrieval on the generated pairs (Section 4.3). Moreover, in Section 4.3, we propose the training of a generalized model on multiple styles of text queries at the same time and introduce a new contrastive objective, In-Style, that improves training on more than one text style at once.

4.1. Pseudo Matching

First, we obtain pseudo video-text pairs, with each pair containing one of the available text queries and the most relevant uncurated video from the web collection. For pseudo matching, we leverage image-language models such as CLIP [47] or BLIP [31] that excel in zero-shot retrieval performance [38]. Such models usually follow a dual-encoder architecture: encoders f_t and f_v projects text t and image x into a common multimodal embedding space. The similarity of text and image is computed as a cosine similarity in this common space: $sim(t, x) = \frac{f_t(t)^\top f_v(x)}{\|f_t(t)\| . f_v(x)\|}$. We use this metric to match the available text queries to the closest video.

Since available videos can vary in overall duration (for example, five or more minutes) and cover a lot of different actions, we divide all videos into non-overlapping clips of s-seconds. We denote $V' = \{v'_j\}$ as a set of all such video clips. Then we calculate a multimodal representation for each video clip v'_j as an average representation of m uniformly sampled frames of a video (see supplement). Using precomputed embeddings, we connect every caption t_i with a video v' with maximum similarity from available set of videos V', such as:

$$v_i^p = \underset{v'_j \in V'}{\arg\max}\ \text{sim}(t_i, v'_j). \tag{1}$$

To increase the diversity of matched videos, we don't allow multiple captions to match the same video clip; therefore, when video clip v_i^p is matched, we exclude it from V'. Thus, we obtain a set of pseudo text-video pairs $P_{ps} = \{(t_i, v_i^p)\}$. In Section 5, we show that this step allows us to introduce a weak supervision that may not find the exact match but provides a basis for further style transfer.

4.2. Style Transfer

We aim to transfer style of the given text queries to other, unrelated web videos by generating new captions with the desired style. Inspired by the ability of language models

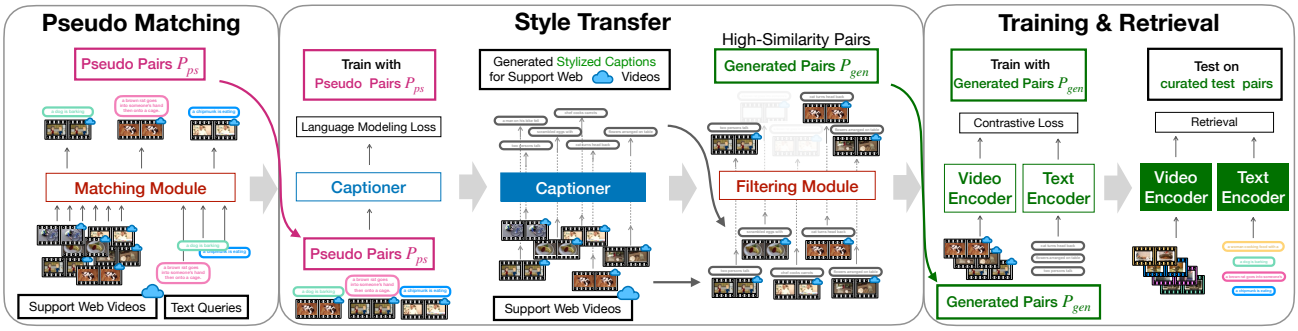

Figure 2: **The proposed In-Style method.** First, in the pseudo matching step, pseudo pairs P_{ps}, which consist of text queries and the most related web videos from the support set, are created. Style Transfer: captioner is tuned with the obtained pseudo pairs P_{ps} to adapt it to the style of the given text queries. Next, stylized new captions are generated for all videos in the support set and then filtered to avoid noisy captions; the resulting set of generated pairs P_{gen} contains web videos and aligned captions of the desired style. To complete the retrieval task, we adapt the dual video-text encoder model with the generated pairs P_{gen} and evaluate on curated paired text-video test sets.

conditioned on visual input [31] to generate plausible descriptions for diverse visual inputs, we propose to adapt the pre-trained image captioner g using the obtained set of noisy pseudo text-video pairs P_{ps}. By doing this, we adapt the captioner to both, the style of the captions as well as the style of the web videos. This allows us to generate new stylized captions P_{gen} for the full support set of videos V' using this captioner.

Captioner. More specifically, we follow the BLIP [31] captioner architecture, which we extend for video captioning by conditioning the model not only on a single image, but on a number of video frames. To this end, we apply the image encoder on each frame individually and inject a joint set of visual tokens into the text decoder model, which produces text in an autoregressive manner. We provide further details in the supplement. To train the captioner g on the pseudo text-video pairs P_{ps}, we utilize the common language model loss that optimizes cross-entropy loss between ground truth and predicted probabilities of the next token given a correct set of previous tokens in the sentence. Following BLIP, we also use label smoothing with parameter 0.1 while calculating cross-entropy.

Stylization of Captions. For each video $v'_i \in V'$, we generate a caption $t^g_i = g(v'_i)$ with a captioner g trained on pseudo pairs by using a nucleus sampling [24]. Nucleus sampling was shown to generate more diverse and detailed captions than a beam search [55, 31].

Filtering. As the captioner g is adapted on pseudo pairs and shifts the model closer to a vocabulary of given text queries $\mathcal{D}$, some of the generated captions t^g_i might be noisy and not descriptive for the web videos. Therefore, we further filter the generated pairs based on a similarity score $s(t^g_i, v'_i)$ utilizing the large pre-trained image-language dual encoders the same way as it was used for creating pseudo text-video pairs (Section 4.1). Leaving only pairs with similarity higher a threshold $s(t^g_j, v'_j) > th$, we obtain a paired set of web videos and stylized related captions $P_{gen} = \{(t^g_j, v'_j)\}$. In Section 5.7, we show that even a noisy set of pseudo pairs is enough to adapt a captioner for generating captions in a desired text style and that stylized captions combined with the following filtering provide a strong learning signal to boost the performance of retrieval in target distribution D.

4.3. Training and Retrieval

Single-Style Training. To allow for text-video retrieval based on the stylized captions and the paired video data, we train a dual-encoder architecture [31] on the set of generated pairs P_{gen} with the contrastive loss [43]. We show that P_{gen} provides better supervision than P_{ps} or even a combination P_{gen} + P_{ps}. Practically, we consider several pre-trained models: the image-text model BLIP [31], which we adapted for video as described in Section 4.1, as well as video-text model EAO [52], which is pre-trained on the HowTo100M dataset with ASR-video pairs, which serve as noisy supervision. Following previous works, we use symmetric contrastive loss, which brings together text t^g_i and video v_i from a text-video pair $(t^g_i, v_i) \in P_{gen}$ (a positive pair) in shared video-text embedding space, and contrasting them on video and text from different pairs (negatives), that are pushed apart:

$$L = -\frac{1}{2B}\sum_{i=1}^{B}\left(\log\frac{\exp(\frac{s(t^g_i,v_i)}{\tau})}{\sum_{j=1}^{B}\exp(\frac{s(t^g_i,v_j)}{\tau})} + \log\frac{\exp(\frac{s(v_i,t^g_i)}{\tau})}{\sum_{i=1}^{B}\exp(\frac{s(v_i,t^g_j)}{\tau})}\right), \tag{2}$$

where τ denotes a temperature parameter, and B is a number of pairs.

For the fine-tuning of the the BLIP model, we follow the original setup and utilize the extension of contrastive training with a momentum encoder and a queue that keeps more negatives, as well as soft labels. For the fine-tuning of the EAO model, we follow the respective setup without a momentum encoder or soft labels.

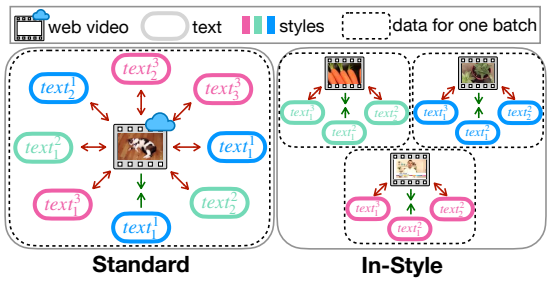

Figure 3: **Multi-dataset training. Left**: Standard contrastive training with multiple datasets. **Right**: Ours In-Style training procedure. Each batch consists of text queries that belong only to the same style. Note that we use only web videos from the support set; therefore, all videos are from the same distribution.

Multi-Style Training. Finally, we consider training a generalized model on multiple sources of text queries coming from different data distributions $D_1, ..., D_K$. Let's denote $P^1_{gen}, ..., P^N_{gen}$ set of generated pairs for the captions from $D_1, ..., D_N$ respectively. Here, different sources can have various styles that might highlight different aspects of videos in their captions (Table 3). As an example, captions in the YouCook2 dataset [66] are more "action"-oriented, e.g. "combine macaroni sauce and cheese" or "stir in crushed tomatos", while captions of the LSMDC dataset [49] are third-person descriptions, e.g. "Someone gazes at the beautiful animal" or "Someone chews the sweet'. In standard training [31], all different styles with their matching videos would be present in contrastive loss together, which can lead to a mixture of different visual topics and text styles, which are easy to separate and which might include only few hard negatives per sample. To avoid this possibly noisy setting, we propose to modify the training procedure and to select video-caption pairs with captions from the same data source for contrastive loss. Formally, during training, we iterate over generated sets of pair $P^1_{gen}, ..., P^N_{gen}$ sampling a minibatch $\{(t^g_i, v_i)\}^B_{i=1}$ from a single set $P_{gen} \in \{P^1_{gen}, ..., P^N_{gen}\}$ and calculating loss $L(\{(t^g_i, v_i)\}^B_{i=1})$ performing one optimization step with a minibatch (Figure 3). We note that for BLIP training we keep separate queues for each set P_{gen}.

We show in Section 5.4 that this setting can be beneficial for learning a generalized model. Our intuition is that text queries with the same style provide stronger negatives for the model, allowing the model to concentrate on the content of the captions rather than a style.

5. Experimental Evaluation

We evaluate the proposed uncurated & unpaired text-video retrieval approach on five popular benchmark datasets: MSR-VTT [61], YouCook2 [66], MSVD [8], LSMDC [49], and DiDeMo [2]. All datasets cover different styles of captions and videos, which includes YouTube and Flickr videos on various topics and video clips from movies. As a source of support videos, we use the large-scale web dataset HowTo100M [41]. We additionally test our model with text queries from the VATEX dataset [56] as well as with third-party text queries (not video captions), specifically with the recipe steps from Food.com dataset [39] and task descriptions from WikiHow dataset [27] datasets.

5.1. Dataset Details

MSR-VTT [61] contains in total 10k videos on various topics and 200K captions. More precisely, every 20 captions describe the same video in different words. We use split 9K+1K [17] in evaluation, resulting in 180K captions for training and 1K text-video for testing.

YouCook2 [66] is a dataset of 14K cooking instructional video clips, where each clip is annotated with a short cooking recipe step. Following[41, 52], we use a 10K+3.5K training-testing split, leveraging 10K captions for training.

MSVD [8] contains 2K video snippets, where each is associated with approximately 40 sentences. The standard split consists of 1200 videos for training, 100 for validation, and 670 for testing. The training set contains 48K captions.

LSMDC [49] is a collection of 202 movies sliced into 118K movie-clips with one description per clip with about 100K for training, while 7408 and 1000 text-video paired samples are used for validation and testing, respectively.

DiDeMo [2] is a fine-grained text-video dataset. 10K Flickr videos are paired with multiple detailed sentences (40K sentences in total). During training, we use the single sentences (33K captions), whereas for evaluation on the test set, we follow [3] and concatenate all the descriptions for video into one paragraph, acting as a video-paragraph retrieval task (we do not use ground truth time-stamp annotations).

VATEX [56] dataset contains 35K video clips with multiple annotated captions for a video, covering 600 different human activities. The training set contains 260K captions.

Food.com [39] is a text dataset that contains more than 230K recipe texts with over 2.2M recipe steps crawled from websites. We use recipe steps as text queries in our training.

WikiHow [27] is a large-scale text dataset using the online WikiHow knowledge base. The dataset contains more than 230K articles covering a variety of topics/tasks and descriptions of steps to solve these tasks. We use only headline steps as text quires, which gives us 1.7M captions.

HowTo100M [41] is a dataset of instructional videos that cover a large variety of topics. The dataset consists of more than 1M videos that were collected by querying on YouTube 23,000 different "how to" tasks. In our default setup, we use 8-second non-overlapping clips from a 100K random subset of the dataset (no more than 15 clips per video) as a support video dataset, resulting in ∼1.4M video clips.

Pre-trained Model	Method	Supervision	MSR-VTT				YouCook2				DiDeMo				MSVD				LSMDC				Mean			
			R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR
BLIP [31]	Zero-shot	none	34.1	60.2	70.6	3	6.0	16.2	23.1	70	28.2	52.0	62.7	5	38.8	64.8	74.0	2	14.5	29.3	36.4	32.5	24.3	44.5	53.4	22.5
	In-Style (ours)	only text	**36.2**	**61.8**	**71.9**	**3**	**8.6**	**21.6**	**30.0**	**37**	**32.1**	**61.9**	**71.2**	**3**	**44.8**	**72.5**	**81.2**	**2**	**16.1**	**33.6**	**39.7**	**25**	**27.6**	**50.3**	**58.8**	**14**
	GT fine-tuning	T-V pairs	42.9	69.7	78.9	2	12.6	32.0	43.6	15	40.2	70.6	79.3	2	48.1	76.6	85.0	2	23.8	41.1	50.9	10	33.5	58.0	67.5	6.2
EAO [52]	Zero-shot	none	9.9	24.0	32.6	28	19.8	42.9	55.1	8	6.6	19.0	26.8	42	18.0	40.4	52.3	9	3.6	8.5	13.0	177	11.6	27.0	36.0	52.8
	In-Style (ours)	only text	**16.4**	**35.8**	**48.9**	**10**	**20.3**	**46.4**	**58.8**	**7**	**13.2**	**31.6**	**44**	**15**	**23.4**	**50**	**62.4**	**5**	**4.9**	**12.3**	**16.7**	**94**	**15.64**	**35.22**	**46.16**	**26.2**
	GT fine-tuning	T-V pairs	22.8	47.8	60.3	6	26.7	55.9	68.6	4	19.2	43.1	54.4	8	25.1	53.6	65.7	5	8.9	21.2	29.4	40	20.5	44.3	55.7	12.6

Table 1: **Text-video retrieval with style transfer.** Comparison between upper bound, where the retrieval model trained with ground truth aligned text-video pairs (T-V pairs), zero-shot respective models (no style transfer or tuning) and our In-Style method, where we follow our new setting of *uncurated & unpaired text-video retrieval* for style transfer based only on input text queries.

Training Dataset	MSR-VTT				YouCook2				DiDeMo				MSVD				LSMDC				Mean			
	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR
MSR-VTT	36.2	61.8	71.9	3	7.6	18.8	25.9	62	29.0	54.5	65.4	4	43.3	70.7	79.9	2	15.2	28.5	35.3	31	26.3	46.9	55.7	20.4
YouCook2	31.5	55.5	64.4	4	8.6	21.6	30.0	**37**	25.1	53.9	65.2	4	41.1	67.3	76.8	2	14.2	28.8	36.9	30	24.1	45.4	54.7	15.4
Didemo	34.0	58.5	68.9	3	6.8	17.2	24.5	69	32.1	**61.9**	**71.2**	**3**	43.7	71.6	80.5	2	16.6	30.5	38.4	28	26.6	47.9	56.7	21
MSVD	36.0	59.4	69.5	3	6.4	16.4	23.6	70	27.0	54.9	65.0	4	**44.8**	72.5	81.2	**2**	14.5	27.4	34.8	32	25.7	46.1	54.8	22.2
LSMDC	33.9	60.3	69.9	3	7.1	18.1	25.6	68	31.7	59.9	69.1	3	44.6	71.7	80.0	2	16.1	**33.6**	39.7	**25**	26.6	48.7	56.8	20.2
Target dataset (mean over diagonal)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	27.5	**50.3**	58.8	**14**
All five datasets – standard training	36.4	62.1	71.8	3	**8.7**	21.4	29.4	44	31.4	62.5	71.2	3	44.7	72.9	81.5	2	16.3	31.9	39.5	25	27.5	50.2	58.7	15.4
All five datasets – In-Style (ours)	**36.7**	**61.9**	**72.3**	**3**	8.5	**21.8**	**30.4**	38.5	**32.6**	61.8	**71.2**	**3**	44.7	**73.1**	**82.0**	**2**	**16.6**	32.2	**39.8**	26	**27.8**	50.2	**59.1**	14.5

Table 2: **Generalization performance of different models over all datasets.** Mean denotes an average of R1, R5, R10, MR over 5 datasets, correspondingly. **Top**: the proposed In-Style method with the input text queries only from one respective training dataset. **Bottom:** training with 5 different text query styles. Comparison between standard multi-dataset training and proposed In-Style procedure.

5.2. Implementation Details

Model. We leverage the pre-trained dual-encoder CLIP (ViT-B/32) model [47] in the matching module and the filtering module. Captioner weights are initialized with BLIP (ViT-B/16) captioner [31] which is pre-trained on five different image-text datasets, including LAION [50] with 129M images. For retrieval, we consider two architectures: dual encoder image-text initialized with BLIP (ViT-B/16), and dual encoder video-text architecture initialized from EAO [52] pre-trained on HowTo100M with noisy ASR narrations. We follow [52] and use a model with a S3D [59] feature extractor and weights that were pre-trained with a video-text-audio triplet, but only utilize the video-text encoder and report all results without audio.

Training. For training, we uniformly sample $m = 8$ frames per video with a resolution of 224×224, augmented with RandAugment [13]. For the captioner training and BLIP-architecture retrieval model, we use AdamW optimizer [36] with a weight decay of 0.05 and a batch size of 128, and a learning rate $1.0e{-}05$ for captioner and $1.0e{-}06$ for retrieval. Following [52], for the EAO model, we used Adam optimizer [26] without weight decay. More training details can be found in the supplement.

Evaluation. For testing, we use $m = 64$ frames for the fine-grained DiDeMo dataset, and $m = 12$ for all others, following [38]. For text-video retrieval, we report standard recall metrics for R1, R5, R10, and the median rank (MR).

5.3. Text Query Style

We consider text style as a set of attributes and properties of the text shared across a text corpus. Such properties might be the usage of stop words, sentence construction, sentiment, text length, etc. To highlight those differences, we show three text examples from the different datasets in Table 3. The respective word clouds for these datasets with and without stop words can be found in Figure 2 in the supplement. It shows that the sentence structure and most frequent words change across datasets. For example, the YouCook2 test queries always start with an action verb, while in other datasets, the subject+verb+object structure is mostly used. While in the MSR-VTT dataset, frequent words are third person nouns like "man", "woman", "person", "people", the DiDeMo uses more words about camera position like "camera", "left", "right", "screen", "view", and the LSMDC mostly describes a subject as "someone". While the MSR-VTT and the MSVD datasets might look similar, Table 3 shows that sentences in the MSR-VTT are 1.5 times longer than in the MSVD on average. We consider such properties as style properties of the text.

5.4. Uncurated & Unpaired Text-Video Retrieval

Single Dataset Training. First, we demonstrate the efficiency of the proposed style transfer method in uncurated & unpaired text-video retrieval on five different downstream datasets in Table 1. We present results for the image-text pre-trained BLIP [31] model as well as for the video-text pre-trained EAO [52] model. We consider three evaluation scenarios: 1) zero-shot performance; 2) the perfor-

Dataset	Examples
MSR-VTT (∼43 symbols in a text)	1) The peoples are sharing their view on this car of different models 2) Someone is showing the ingredients for a dish they are going to make 3) A man is playing an instrument
YouCook2 (∼39 symbols in a text)	1) Combine macaroni sauce and cheese 2) Grate and cube potatoes 3) Stir in crushed tomatos
DiDeMo (∼147 symbols in a text)	1) A dog runs down a hill and stop behind a shrub. Dog sniffs and chews at patch of grass on rock. the dog approaches, then begins to sniff the cluster of plants first time hand is seen petting dog. 2) Only big screen is visible the camera first pans to the large screen. The view shifts from the basketball court to the fans in the seats across the stadium. Camera goes to the bigscreens the dancers are shown on the jumbotraun. 3) A bus stops. The bus stops at the end of the driveway. A kid is coming out of a school bus. School bus doors open.
MSVD (∼31 symbols in a text)	1) The cats are fighting 2) The lady sliced a vegetable 3) A man is eating a pizza
LSMDC (∼46 symbols in a text)	1) SOMEONE goes to the kitchen, wets a towel, comes back to the bed, kneels it, places the towel on SOMEONE's brow. 2) He slaps SOMEONE again. 3) SOMEONE moves off through the crowd.

Table 3: Three random examples of text descriptions in different datasets. With the dataset name, we also report the median length of a text in the dataset.

mance of our style transfer method in the text-video retrieval task with uncurated & unpaired data where only text queries are available during training; 3) training with the ground truth aligned text-video pairs, which can be considered as an upper bound for our task. It shows that the proposed In-Style method significantly outperforms zero-shot performance even without using any aligned training samples from the target distribution. This supports the hypothesis that the style of the text queries is an important component of text-video retrieval. Moreover, we observe that the gap between training with ground truth aligned pairs and the style transfer can be remarkably small, especially on the MSVD dataset, indicating the benefits with respect to a potential annotation cost reduction in the proposed setup.

Multi-Dataset Training. Second, we evaluate the proposed multi-dataset training procedure with the In-Style method in Table 2. Here, a minibatch is compiled from a single text source as shown in Figure 3. This is favorable compared to the standard training, where data points in a minibatch are randomly sampled from all data sources together. It shows that the proposed procedure leads to improved retrieval performance compared to individually trained models and better generalization across all datasets compared to the standard multi-dataset training. We attribute the performance increase compared to standard multi-dataset training to the fact that considering the captions of only the same style in contrastive loss provides a model with a cleaner learning signal with stronger text negative counterparts. As an example, "add sliced cucumber" in YouCook2 style would be a stronger negative in comparison to a correct "add sliced tomato" query than a "a person in a video puts sliced cucumber in a salad" in MSR-VTT style. More discussions of generalization can be found in the supplement.

5.5. Comparison with SOTA

We further compare the proposed method with zero-shot retrieval baselines in Table 4. We report the performance of BLIP and CLIP backbones trained with text queries from the VATEX dataset, thus text queries do not follow distribution of any of the test datasets. The closest counterpart to our model is Nagrani et al. method [42], which utilizes the pre-trained image-text CLIP backbone, which is further trained with the VideoCC3M dataset [42] – a video-text dataset collected by automatic transferring image captions from text-image CC3M dataset [51]. The conceptual difference between [42] and our method is that [42] proposes to transfer *image* captions from the image-caption dataset by pairing images to videos, while the proposed In-Style method adapts the model to the *video* captions. While noting that a direct comparison to different state-of-the-art methods is limited due to different pre-training datasets, it can be observed that the proposed In-Style method achieves the best results on four out of five datasets, underperforming only in YouCook2, which might benefit from HowTo100M pretraining. We additionally validate the statement that text queries can be used without any corresponding videos by using texts from WikiHow [27] and Food.com [39] datasets that contain descriptions of different actions/steps to solve tasks or cook meals. In Table 4, we show that style transfer from both datasets especially benefits YouCook2 retrieval performance that we attribute to the similarity in text styles (see the supplement). However, style transfer from the WikiHow, which is more diverse and covers a larger variety of topics, also improves the performance over the baselines on the DiDemo, MSVD, and LSMDC datasets.

5.6. Efficiency of Style Transfer

Training Pairs. In Table 5, we compare the performance of the models trained either with pseudo pairs P_{ps} or with generated pairs P_{gen}, or with a combination of them $P_{ps} + P_{gen}$. All setups boost the performance of text-video retrieval by a large margin compared to zero-shot

Method	Image-Text Datasets	Video-Text Datasets	MSR-VTT				YouCook2				DiDeMo				MSVD				LSMDC			
			R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR
HowTo100M [41]	-	HowTo100M	7.5	21.2	29.6	38	6.1	17.3	24.8	46	-	-	-	-	-	-	-	-	-	-	-	-
SupportSet [45]	-	HowTo100M	8.7	23.0	31.1	31	-	-	-	-	-	-	-	-	8.9	26.0	37.9	18	-	-	-	-
VATT [1]		HowTo100M+AS	-	-	29.7	49	-	-	45.5	13	-	-	-	-	-	-	-	-	-	-	-	-
EAO[§] [52]	-	HowTo100M	9.9	24.0	32.6	28	**19.8**	**42.9**	**55.1**	**8**	6.6	19.0	26.8	42	18.0	40.4	52.3	9	3.6	8.5	13.0	177
Nagrani et al. [42]	-	VideoCC3M	19.4	39.5	50.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Frozen in Time [3]	CC+COCO	WebVid-2M	24.7	46.9	57.2	7	-	-	-	-	21.1	46.0	56.2	7	-	-	-	-	-	-	-	-
CLIP-straight [46]	WIT	-	31.2	53.7	64.2	4	-	-	-	-	-	-	-	-	37.0	64.1	73.8	**2**	11.3	22.7	29.2	56.5
CLIP4CLIP [38]	WIT	HowTo100M	32.0	57.0	66.9	4	-	-	-	-	-	-	-	-	38.5	66.9	76.8	**2**	15.1	28.5	36.4	28
Nagrani et al. [42]	WIT	VideoCC3M	33.7	57.9	67.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BLIP[‖] [31]	CC+COCO+3more*	-	33.3	57.3	67.5	3.5	5.8	15.0	21.9	76	24.6	50.4	59.7	5.3	37.0	63.3	72.6	3	15.2	28.2	35.9	35
In-Style (ours) (CLIP)	WIT	HowTo100M[†]+VATEX[‡]	35.0	59.6	70.4	**3**	5.1	14.0	20.3	103	26.6	50.5	62.6	5	38.6	66.3	77.9	3	16.0	**31.6**	38.5	**26.5**
In-Style (ours) (BLIP)	CC+COCO+3more*	HowTo100M[†]+VATEX[‡]	**36.0**	**61.9**	**71.5**	**3**	6.8	16.7	24.5	63	**29.4**	**59.2**	**68.6**	**3.5**	**44.9**	**72.7**	**81.1**	**2**	16.4	30.1	38.7	28
In-Style (ours) (BLIP)	CC+COCO+3more*	HowTo100M[†]+WikiHow	34.2	59.6	69.0	**3**	7.3	19.2	27.1	46	29.7	56.2	67.4	4	42.8	70.2	79.1	**2**	**17.0**	30.8	**39.6**	27
In-Style (ours) (BLIP)	CC+COCO+3more*	HowTo100M[†]+Food.com	32.8	54.9	65.8	4	7.2	19.8	27.9	47	25.7	52.8	63.1	5	39.5	64.9	74.9	**2**	14.5	28.9	37.2	30.5
In-Style (ours) (BLIP)	CC+COCO+3more*	HowTo100M[†]+Target[‡]	36.2	61.8	71.9	3	8.6	21.6	30.0	37	32.1	61.9	71.2	3	44.8	72.5	81.2	2	16.1	33.6	39.7	25
In-Style (ours) (EAO)	-	HowTo100M+Target[‡]	16.4	35.8	48.9	10	20.3	46.4	58.8	7	13.2	31.6	44.0	15	23.4	50.0	62.4	5	4.9	12.3	16.7	94

Table 4: **Zero-shot comparison with other methods. Top**: zero-shot retrieval with methods pre-trained on video-language or/and images-language web or/and curated datasets which exclude target datasets during training. For our In-Style method, the VATEX dataset is used as a source of text queries. **Bottom**: uncurated & unpaired text-video retrieval with text queries from the respective target datasets for comparison purposes. Note that this setting is not zero-shot. † denotes that only videos were used (without paired text) and ‡ – only text (without videos). [§]For EAO, performance with S3D backbone is reported. [‖]For BLIP, the performance of dual encoder architecture is reported (not image-grounded text encoder). *CC [7]+COCO [33]+VG [28]+SBU [44] +LAION [50]. AS denotes AudioSet [19].

Training Data	MSR-VTT				YouCook2				DiDeMo				MSVD				LSMDC				Average			
	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR
— (zero-shot)	34.1	60.2	70.6	3	6.0	16.2	23.1	70	28.2	52.0	62.7	5	38.8	64.8	74.0	2	14.5	29.3	36.4	32.5	24.3	44.5	53.3	22.5
Pseudo pairs P_{ps}	35.0	61.4	70.9	3	7.5	19.6	28.9	43	**33.1**	59.8	71.2	**3**	44.3	72.4	81.0	2	16.8	32.7	**40.4**	**25**	27.3	49.2	58.4	15.2
Generated pairs P_{gen}	**36.2**	**61.8**	**71.9**	**3**	8.6	21.6	**30.0**	**37**	32.1	**61.9**	**71.2**	**3**	**44.8**	**72.5**	**81.2**	**2**	16.1	**33.6**	39.7	**25**	**27.6**	**50.3**	**58.8**	**14.0**
Combined P_{ps} + P_{gen}	36.0	61.3	71.5	3	**8.9**	**21.8**	29.8	**37**	32.6	61.8	70.2	3	44.4	72.2	80.8	2	**17.1**	32.4	**40.4**	26	27.8	49.9	58.5	14.2

Table 5: **Different types of training pairs for text-video retrieval step.** We evaluate text-video retrieval with pseudo pairs P_{ps} only, with generated pairs P_{gen} only, and the combination of both P_{ps} + P_{gen}.

text-video retrieval. The generated pairs P_{gen} achieve a better performance than pseudo pairs P_{ps} on all datasets except LSMDC, whereas a combination of $P_{ps}+P_{gen}$ does not improve performance on average. We note that the number of pairs in P_{gen} is significantly larger than in P_{ps} (Table 7b) for all datasets except LSMDC (a dataset of movies, which might contain a larger domain shift to YouTube videos compared to other datasets). We assume that in this case P_{gen} contains better-aligned pairs since each generated text description is conditioned on the corresponding video, while in P_{ps} a fixed set of descriptions is matched (see examples in Figure 4) explaining the performance drop with $P_{ps} + P_{gen}$.

Style Transfer. In Table 6, we consider how much the text style transfer in the generated pairs P_{gen} influences the retrieval performance. For this, we considered three sets of P_{gen} for the training retrieval model: 1) P_{gen} generated with zero-shot BLIP captioner; 2) In-Style P_{gen} generated with captioner trained on P_{ps} with text queries from a different non-target dataset (we used the VATEX dataset); 3) In-Style P_{gen} with a captioner trained on P_{ps} with text queries from the target dataset. We observe that training the model with generated text-video pairs (from uncurated web videos from the HowTo100M dataset) by a zero-shot image-pretrained captioner already improves the performance in all video retrieval datasets. We attribute this to the content and style adaptation of the image-language model to the specific appearances in the videos. However, such models tend to generate "static" descriptions that do not involve actions. Thus, text queries from non-target video datasets, namely the VATEX dataset, improve the retrieval performance further. Yet, we notice that YouCook2 does not benefit from the VATEX text queries as from the zero-shot generated captions. Finally, using training text queries from the target dataset excels on the considered benchmarks.

5.7. Ablation Study

Matching Method. To obtain generated pairs, we train the captioner with pseudo pairs that were created by a matching module. In Table 7c, we consider two options for the matching module: image-text pre-trained dual encoders from BLIP [31] and CLIP [47], as well as the "Random" option where text queries are simply matched with the random videos. We report the text-video retrieval performance of our final model using the given option of the matching module. We observe that matching module based on CLIP leads to better performance. We attribute that to the robustness of CLIP to the noisy web data as it was trained on large-scale web image-text pairs, whereas BLIP utilizes ad-

Training data	MSR-VTT				YouCook2				DiDeMo				MSVD				LSMDC				Average			
	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR	R1	R5	R10	MR
— (zero-shot)	34.1	60.2	70.6	**3**	6.0	16.2	23.1	70	28.2	52.0	62.7	5	38.8	64.8	74.0	**2**	14.5	29.3	36.4	32	24.3	44.5	53.3	22.5
P_{gen} with zero-shot captioner	**36.3**	61.6	71.8	**3**	7.1	18.4	25.6	65	28.7	56.3	65.0	4	43.8	71.2	80.1	**2**	16.0	29.2	37.7	30	26.3	47.3	56.1	20.8
In-Style P_{gen} (non-target)	36.0	61.9	71.5	**3**	6.8	16.7	24.5	63	29.4	59.2	68.6	3.5	44.9	**72.7**	81.1	**2**	**16.4**	30.1	38.7	28	26.7	48.1	56.9	19.9
In-Style P_{gen} (target)	36.2	**61.8**	**71.9**	**3**	**8.6**	**21.6**	**30.0**	**37**	**32.1**	**61.9**	**71.2**	**3**	**44.8**	72.5	**81.2**	**2**	16.1	**33.6**	**39.7**	**25**	**27.6**	**50.3**	**58.8**	**14**

Table 6: **Source of generated pairs P_{gen} for text-video retrieval.** Comparison between zero-shot BLIP (no adaption of retrieval model), zero-shot BLIP captioner, and adapted BLIP captioner with our In-Style method with either text queries from VATEX (non-target) or text queries from the target datasets.

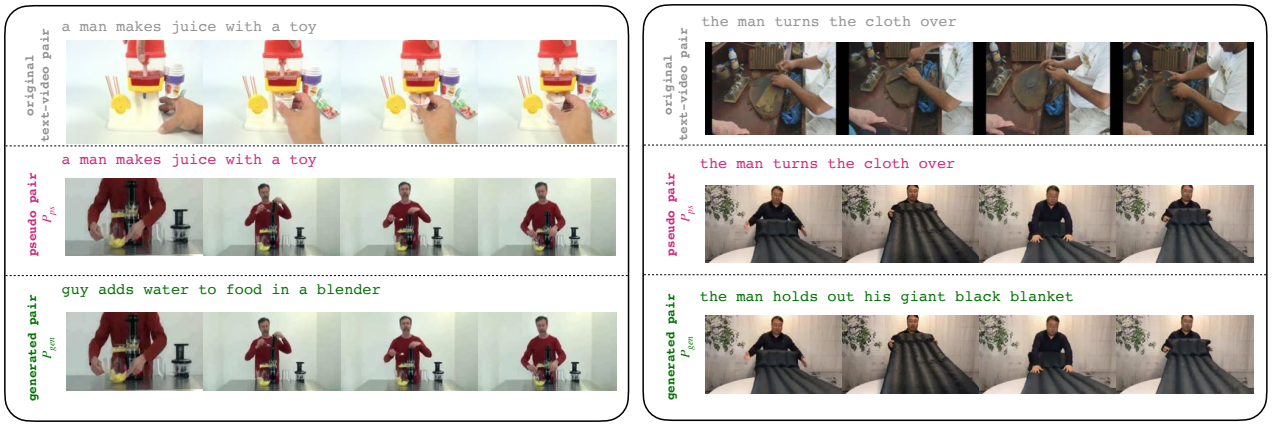

Figure 4: **Qualitative evaluation of P_{ps} and P_{gen} on the MSR-VTT (left) and DiDeMo (right) datasets.** First, a text query is matched with one of the videos (a pseudo pair P_{ps}), and then, after the style preservation step, for each video new caption is generated in the same style but with updated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

Filt. Thr.	R1	R5	R10	MR
0.26	43.9	71.8	80.8	**2**
0.27	44.2	72.2	80.9	**2**
0.28	44.8	**72.5**	**81.2**	**2**
0.29	45.0	72.3	80.9	**2**
0.30	**45.1**	72	80.8	**2**

(a) Filtering threshold

Dataset	#Pseudo Pairs	#Generated Pairs
MSR-VTT	180k	495k
YouCook	10k	168k
Didemo	33k	280k
MSVD	48k	379k
LSMDC	101k	144k

(b) Number of P_{ps} and P_{gen}

Matching	R1	R5	R10	MR
Random	39.1	66.3	75.8	**2**
BLIP	44.1	71.4	80.0	**2**
CLIP	**44.8**	**72.5**	**81.2**	**2**

(c) Matching method

Training pairs	B@4	ROUGE	CIDEr
– (zero-shot)	0.305	0.519	0.610
Pseudo pairs	0.559	0.628	1.059
GT pairs	0.659	0.680	1.296

(d) Captioning performance

Table 7: Ablations of our In-Style method on the MSVD.

ditional filtering to reduce the noise in the training.

Filtering Threshold. In Table 7a, we consider the effect of filtering on the quality of the generated pairs P_{gen}. We find threshold $th = 0.28$ works the best, indicating that filtering is an important step for our style transfer framework.

Captioning Performance Finally, we evaluate the captioning performance of the captioner trained with pseudo pairs P_{ps} with the standard NLP metrics BLEU@4, ROUGE and CIDEr. Table 7d demonstrates that the captioner trained with pseudo pairs almost doubles the zero-shot captioner performance, significantly reducing the gap to the training with ground truth supervision.

6. Conclusion

In this work, we address a new task of *text-video retrieval with uncurated & unpaired data*, where during training only text queries are available. Motivated by the fact that different domains imply diverse styles of video descriptions, we introduced the In-Style method that preserves the style of the given input queries and transfers it to the support set of unrelated web videos, creating aligned text-video pairs with the style of input. Utilization of obtained text-video pairs as supervision leads to a significant performance boost in text-video retrieval. Moreover, we show the performance generalization of a single model that we train with multiple styles simultaneously, proposing a training procedure for multi-dataset training. We evaluate the proposed model over multiple datasets and show the advantages of the In-Style method on the task of uncurated & unpaired text-video retrieval and achieve new state-of-the-art results for zero-shot text-video retrieval.

Acknowledgements

We would like to thank Stephan Alaniz for his invaluable help in this work. Nina Shvetsova is supported by German Federal Ministry of Education and Research (BMBF) project STCL - 01IS22067.

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. 8

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. 2017. 5

[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. 2, 5, 8

[4] Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *CVPR*, 2022. 2

[5] 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. 2

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2

[7] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 8

[8] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 2011. 5

[9] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *CVPR*, 2020. 2

[10] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 2

[11] Ego4D Consortium et al. Egocentric live 4d perception (ego4d) database: A large-scale first-person video database, supporting research in multi-modal machine perception for daily life activity. 2

[12] 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. 2

[13] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshops*, 2020. 6

[14] 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

[15] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. Mdmmt: Multidomain multimodal transformer for video retrieval. In *CVPR*, 2021. 2

[16] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 2, 3

[17] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020. 2, 5

[18] Zijian Gao, Jingyu Liu, Sheng Chen, Dedan Chang, Hao Zhang, and Jinwei Yuan. Clip2tv: An empirical study on transformer-based methods for video-text retrieval. *arXiv preprint arXiv:2111.05610*, 2021. 2

[19] Jort F Gemmeke, Daniel PW 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. 8

[20] Satya Krishna Gorti, Noël 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. 2

[21] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 2

[22] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *CVPR*, 2022. 2

[23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 2

[24] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *ICLR*, 2020. 4

[25] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 2

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[27] Mahnaz Koupaee and William Yang Wang. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*, 2018. 5, 7

[28] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*. 8

[29] 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. 2

[30] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, 2020. 2

[31] 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, 2, 3, 4, 5, 6, 8

[32] 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. 2

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 8

[34] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. 2019. 2

[35] 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. 2, 3

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[37] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 2, 3

[38] 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, 3, 6, 8

[39] Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. Generating personalized recipes from historical user preferences. In *EMNLP*, 2019. 5, 7

[40] 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. 2

[41] 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. 2, 3, 5, 8

[42] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In *ECCV*, 2022. 7, 8

[43] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 4

[44] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *NeurIPS*, 24, 2011. 8

[45] 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. 8

[46] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. A straightforward framework for video retrieval using clip. In *Pattern Recognition: 13th Mexican Conference*, 2021. 2, 8

[47] 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, 2, 3, 6, 8

[48] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. 2019. 2

[49] Anna Rohrbach, Marcus Rohrbach, and Bernt Schiele. The long-short story of movie description. In *GCPR*. 5

[50] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 6, 8

[51] 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. 7

[52] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multimodal fusion transformer for video retrieval. In *CVPR*, 2022. 2, 4, 5, 6, 8

[53] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020. 2

[54] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *CVPR*, 2022. 2

[55] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. In *AAAI*, 2015. 4

[56] 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. 5

[57] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: global-local sequence alignment for text-video retrieval. In *CVPR*, 2021. 2

[58] 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

[59] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018. 6

[60] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP*, 2021. 2

[61] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 5

[62] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[63] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, 2018. 2

[64] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *NeurIPS*, 2021. 2

[65] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *arXiv preprint arXiv:2212.04501*, 2022. 2

[66] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 5