ConTra: (Con)text (Tra)nsformer for Cross-Modal Video Retrieval

Adriano Fragomeni Michael Wray Dima Damen

Department of Computer Science, University of Bristol, UK

Abstract. In this paper, we re-examine the task of cross-modal clip-sentence retrieval, where the clip is part of a longer untrimmed video. When the clip is short or visually ambiguous, knowledge of its local temporal context (i.e. surrounding video segments) can be used to improve the retrieval performance. We propose **Con**text **Tra**nsformer (ConTra); an encoder architecture that models the interaction between a video clip and its local temporal context in order to enhance its embedded representations. Importantly, we supervise the context transformer using contrastive losses in the cross-modal embedding space.
We explore context transformers for video and text modalities. Results consistently demonstrate improved performance on three datasets: YouCook2, EPIC-KITCHENS and a clip-sentence version of ActivityNet Captions. Exhaustive ablation studies and context analysis show the efficacy of the proposed method.

1 Introduction

Millions of hours of video are being uploaded to online platforms every day. Leveraging this wealth of visual knowledge relies on methods that can understand the video, whilst also allowing for videos to be searchable, e.g. via language. Methods can query the entire video [67, 49, 22] or the individual segments, or clips, that make up a video [14, 41]. In this work, we focus on the latter problem of clip-sentence retrieval, specifically from long untrimmed videos. This is particularly beneficial to retrieve all instances of the same step (e.g. folding dough or jacking up a car) from videos of various procedures.

In Fig. 1, we compare current clip-sentence retrieval approaches (e.g. [46, 58, 11, 7, 39, 3]) to our proposed context transformer. We leverage local temporal context clues, readily available in long videos, to improve retrieval performance. Local sequences of actions often use similar objects or include actions towards the same goal, which can enrich the embedded clip representation.

We emphasise the importance of learnt **local temporal clip context**, that is the *few* clips surrounding (i.e. before and after) the clip to be embedded. Our model, ConTra, learns to attend to relevant neighbouring clips by using a transformer encoder, differing from previous works which learn context over frames [20] or globally across the entire video [22] (see Video-Paragraph Retrieval, Fig. 1 top). We supervise ConTra with cross-modal contrastive losses and a proposed neighbouring loss that ensures the embedding is distinct across overlapping contexts.

Our contributions are summarised as follows: (i) we explore the task of cross-modal clip-sentence retrieval when using local context in clip, text or in both modalities simultaneously (ii) we propose ConTra, a transformer based architecture that learns to attend

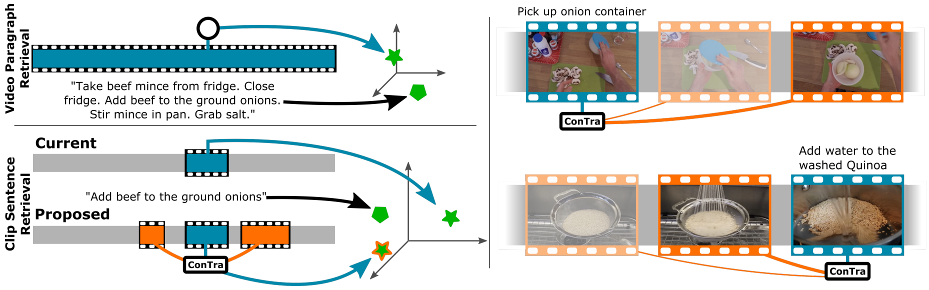
Video Paragraph Retrieval
"Take beef mince from fridge. Close fridge. Add beef to the ground onions. Stir mince in pan. Grab salt."
Current
Clip Sentence Retrieval
"Add beef to the ground onions"
Proposed
ConTra
Pick up onion container
ConTra
Add water to the washed Quinoa
ConTra

Fig. 1: **Left**: We compare video-paragraph retrieval (top) to current and proposed clip-sentence retrieval (bottom) in long videos. In ConTra, we propose to attend to local context of neighbouring clips. **Right**: Examples where ConTra can enrich the clip representation from next/previous clips, observing the onion (top) or that the quinoa has already been washed (bottom). Line thickness/brightness represents attention weights.

to local temporal context, supervised by a multi-term loss that is able to distinguish consecutive clips by the introduction of a neighbouring contrastive loss (iii) we demonstrate the added value of local context by conducting detailed experiments on three datasets.

2 Related works

In this section, we split video-text retrieval works into those which primarily focus on either Clip-Sentence or Video-Paragraph retrieval before presenting works that use temporal context for other video understanding tasks.

Clip-Sentence Retrieval: Most works primarily rely on two-stream (i.e. dual) approaches [3, 32, 40, 42, 58, 60, 46, 23, 70, 65], using multiple text embeddings [9, 61, 11, 15, 43], video experts [36, 37, 41], or audio [2, 7, 58, 1]. Recently, single stream cross-modal encoders have also been used [39, 72, 52, 38, 64], improving inter-modality modelling at the cost of increased computational complexity. In ConTra, we use a dual stream model with separate branches for the visual and the textual components.

Temporal modelling of frames *within a clip* is a common avenue for retrieval approaches [20, 58, 72]. Gabeur et al. [20] use multiple video experts with a multi-modal transformer to better capture the temporal relationships between modalities. Wang et al. [58] learn an alignment between words and frame features alongside the clip-sentence alignment. ActBert [72] also models alignment between clip and word-level features using self-supervised learning. Bain et al. [3] adapt a ViT [16] model, trained with a curriculum learning schedule, to gradually attend to more frames within each clip. MIL-NCE [40] alleviates noise within the automated captions, matching clips to neighbouring sentences. However, the learned representation does not go beyond the clip extent. VideoCLIP [65] creates positive clips by sampling both the centre-point (within narration timestamp) and the clip's duration to better align clips and sentences, foregoing the reliance on explicit start/end times. In our work, we go beyond temporal modelling of the clip itself to using local context outside the clip.

Other works improve modelling of the textual representation [15, 61, 46]. Patrick et al. [46] introduce a generative task of cross-instance captioning to alleviate false negatives. They create a support set of relevant captions and learn to reconstruct a sample's text representation as a weighted combination of a support-set of video representations from the batch. However, whilst they use information from other sentences, there is no notion of those which are temporally related. Instead, we propose to explore relationships between neighbouring sentences using local context within the same video.

Video-paragraph Retrieval: Another retrieval task is video-paragraph retrieval [67, 49, 60, 36, 11, 58, 51, 34, 22], where videos and paragraphs describing the full videos are embedded in their entirety. There are two main approaches: using hierarchical representations between the paragraph/video and constituent sentences/clips [67, 36, 22] or jointly modelling the entire video/paragraph with a cross-modal transformer [51, 34].

COOT [22] models the interactions between levels of granularity for each modality by using a hierarchical transformer. The video and paragraph embeddings are obtained via a combination of their clips and sentences. ClipBERT [34] inputs both text and video to a single transformer encoder after employing sparse sampling, where only a single or a few sampled clips are used at each training step. In this work, we focus on clip-sentence retrieval, but take inspiration from video-paragraph works in how they relate clips within a video. Importantly, we focus on local context, which is applicable to long videos with hundreds of clips.

Temporal Context for Video Understanding: We also build on works that successfully used *local temporal context* for other video understanding tasks such as: action recognition [30, 6, 62, 68]; action anticipation [48, 19]; object detection and tracking [4, 5]; moment localisation [69]; and Content-Based Retrieval [50]. Bertasius and Torresani [5] use local context for mask propagation to better segment and track occluded objects. Kazakos et al. [30] use the context of neighbouring clips for action recognition using a tranformer encoder along with a language model to ensure the predicted sequence of actions is realistic. Feichtenhofer et al. [62] allow for modelling features beyond the short clips. They use a non-local attention block and find that using context from up to 60 seconds can help recognise actions. Shao et al. [50] use a self-attention mechanism to model long-term dependencies for content-based video retrieval. They use a supervised contrastive learning method that performs automatic hard negative mining and utilises a memory bank to increase the capacity of negative samples.

To the best of our knowledge, ours is the first work to explore using neighbouring clips as context for cross-modal clip-sentence retrieval in long untrimmed videos.

3 Context Transformer (ConTra)

We first explicitly define the task of clip-sentence retrieval in Sec. 3.1 before extending the definition to incorporate *local* clip context, in untrimmed videos. We then present our clip context embedding in Sec. 3.2 where we provide details of our architecture followed by the training losses in Sec. 3.3. We then extend ConTra to context in both modalities in Sec. 3.4. An overview of our approach can be seen in Fig. 2.

3.1 Definitions

We begin with a set of untrimmed videos, $v_i \in V$. These are broken down further into ordered clips, $c_{ij} \in v_i$, each with a corresponding sentence/caption, t_{ij}, describing the

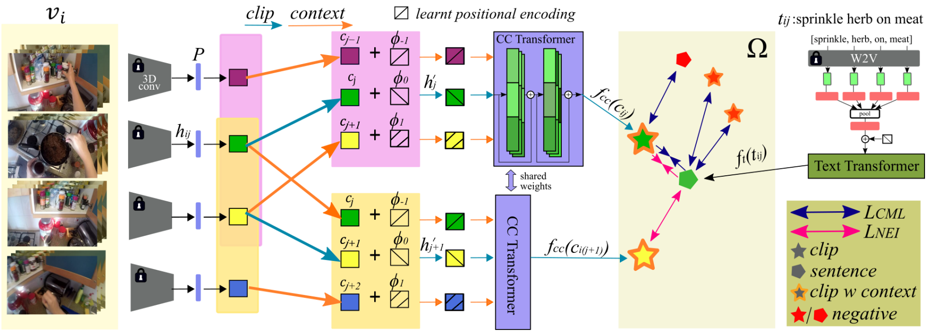

Fig. 2: **Overview of ConTra:** Given a video v_i split into clips c_{ij}, we encode clips into features h_{ij}, projected by P and tagged with a learnt position encoding ϕ *relative* to the centre clip. A Clip Context (CC) encoder learns an enriched representation of the centre clip (cyan arrow) attending to context clips (orange arrows). The embedding space Ω is learnt with cross-modal (CML) and neighbouring (NEI) losses. NEI pushes overlapping contexts further apart—shown for $f_{cc}(c_{ij})$ and $f_{cc}(c_{i(j+1)})$.

action within the clip. Querying by the sentence t_{ij} aims to retrieve the corresponding clip c_{ij} and vice versa for cross-modal retrieval.

Learning a dual stream retrieval model focuses on learning two projection functions, $f_c : c \longrightarrow \Omega \subseteq \mathbb{R}^d$ and $f_t : t \longrightarrow \Omega \subseteq \mathbb{R}^d$, which project the video/text modalities respectively into a common d-dimensional embedding space, Ω, where c_{ij} and t_{ij} are close. The weights of these embedding functions can be collectively trained using contrastive-based losses including a triplet loss [56, 55, 54, 47] or noise-contrastive estimation [24, 29, 40].

Instead of using the clip solely, we wish to utilise its *local* temporal context to enrich the embedded representation. We define the temporal Clip Context (CC) using m, around the clip c_{ij}, as follows:

$$CC_m(c_{ij}) = (c_{i(j-m)}, \cdots, c_{ij}, \cdots, c_{i(j+m)}) \tag{1}$$

There are $2m$ clips that are temporally adjacent to c_{ij} in the same video v_i. Note that the length of the adjacent clips governed by m differs per dataset and video. Importantly, we still aim to retrieve the corresponding sentence t_{ij} for the centre clip c_{ij}, but utilise the untrimmed nature of the video around the clip to enrich c_{ij}'s representation.

In Table 1, we differentiate between existing tasks in Sec. 2 and our proposed settings. Note that in Video-Paragraph retrieval, models cannot be used to retrieve individual clips. Different from the standard Clip-Sentence setting, we utilise neighbouring clips to enrich the clip representation, the sentence representation or both. Next, we describe our Clip Context Transformer.

Task	Clip?	Video	Text
Video-Paragraph	×	all clips	all sentences
Clip-Sentence			
No Context	✓	clip	sentence
Clip Context	✓	clip+context	sentence
Text Context	✓	clip	sent.+context
Context in Both	✓	clip+context	sent.+context

Table 1: Comparison of tasks with/without using context in video and text.

3.2 Clip Context Transformer

We learn the embedding function $f_{cc} : CC_m(c) \longrightarrow \Omega$, using the local clip context in Eq. 1. We consider each clip as a linear projection of its features $P(h_j)$. We drop the video index i here for simplicity. We learn $2m + 1$ distinct positional embeddings, $(\phi_{-m}, \cdots, \phi_0, \cdots, \phi_m)$, that are added such that $h'_{j+\alpha} = P(h_{j+\alpha}) + \phi_{0+\alpha}$, where $-m \leq \alpha \leq m$ and ϕ_0 is the positional embedding of the centre clip. Note that the positional embeddings emphasise the order of the clip within the context window rather than the full video, thus reflecting the *relative* position of the context to the centre clip c_{ij}, and are identical across contexts. We showcase this on two neighbouring clips in Fig. 2. We form the input to the encoder transformer as:

$$H' = [h'_{j-m}, \cdots, h'_j, \cdots, h'_{j+m}] \tag{2}$$

From H', we aim to learn the embedding of the centre clip. We use a multi-headed attention block [53] with the standard self-attention heads and residual connections. The output of the r^{th} attention head is thus computed as,

$$A_r = \sigma\left(\frac{(\boldsymbol{\theta_r^Q H'})(\boldsymbol{\theta_r^K H'})^\top}{\sqrt{d}}\right)(\boldsymbol{\theta_r^W H'}) \tag{3}$$

where $\boldsymbol{\theta_r^Q}$, $\boldsymbol{\theta_r^K}$, $\boldsymbol{\theta_r^W} \in \mathbb{R}^{(2m+1)\times d/R}$ are learnable projection matrices. The output of the multi-head attention is then calculated as the concatenation of all R heads:

$$A = [A_1, \ldots, A_R] + H' \tag{4}$$

For the clip embedding, we focus on the output from A corresponding to the centre clip j, such that:

$$f_{cc}(c_{ij}) = g(A_j) + A_j \tag{5}$$

where g is one or more linear layers with ReLU activations, along with another residual connection. Note that the size of f_{cc} is d, independent of the context length m. f_{cc} can be extended with further multi-head attention layers. We discuss how we train the ConTra model next.

3.3 Training ConTra

Cross-Modal Loss. For both training and inference, we calculate the cosine similarity $s(c_{ij}, t_{kl})$ between the embeddings of the context-enriched clip $f_{cc}(c_{ij})$ and a sentence $f_t(t_{kl})$. Cross-modal losses are regularly used in retrieval works such as the triplet loss [22, 42, 9, 36] and the Noise-Contrastive Estimation (NCE) loss [7, 1, 64, 35]. We use NCE as our cross-modal loss (L_{CML}) [24, 29]:

$$L_{CML} = \frac{1}{|B|} \sum_{(c_{ij},t_{ij})\in B} -\log\left(\frac{e^{s(c_{ij},t_{ij})/\tau}}{e^{s(c_{ij},t_{ij})/\tau} + \sum\limits_{(c',t')\sim\mathcal{N}'} e^{s(c',t')/\tau}}\right) \tag{6}$$

where B is a set of corresponding clip-captions pairs, i.e. (c_{ij}, t_{ij}) and τ is the temperature parameter. We construct the negative set $\mathcal{N}'$ in each case from the batch by combining $(c_{ij}, t_{lk})_{ij\neq lk}$ as well as $(c_{lk}, t_{ij})_{ij\neq lk}$, considering negatives for both clip and sentence across elements in the batch.

Uniformity Loss. The uniformity loss is less regularly used, but was proposed in [57] and used in [10, 17] works, for image retrieval. It ensures that the embedded representations preserve maximal information, i.e. feature vectors are distributed uniformly on the unit hypersphere. We use the uniformity loss (L_{UNI}) such as:

$$L_{UNI} = \log \left(\frac{1}{|B|} \sum_{u,u' \in U \times U} e^{-2\|u-u'\|_2^2} \right) \tag{7}$$

where $U = \{c_1, t_1, ..., c_B, t_B\}$, are all the clips and sentences in the batch. This loss term is applied to all the clips and sentences in a batch.

Neighbouring Loss. We additionally propose a neighbour-contrasting loss (L_{NEI}) to ensure that the embeddings of context items are well discriminated. Indeed, one of the challenges of introducing local temporal context is the overlap between contexts of neighbouring clips. Consider two neighbouring clips in the same video, say c_{ij} and $c_{i(j+1)}$ (see Fig. 2), the context windows $CC(c_{ij})$ and $CC(c_{i(j+1)})$ share $2m$ clips. While the positional encoding of the clips differ, distinguishing between the embedded neighbouring clips can be challenging. This can be considered as a special case of hard negative mining, as in [18, 27], however our usage of it, where only neighbouring clips are considered as negatives is novel.

Accordingly, we define the L_{NEI} using the NCE loss:

$$L_{NEI} = \frac{1}{|B|} \sum_{(c_{ij},t_{ij}) \in B} - \log \left(\frac{e^{s(c_{ij},t_{ij})/\tau}}{e^{s(c_{ij},t_{ij})/\tau} + e^{s(c_{i(j+\alpha)},t_{ij})/\tau}} \right) \tag{8}$$

where α is randomly sampled from $[-m, m]$ subject to $t_{ij} \neq t_{i(j+\alpha)}$. We thus randomly sample a neighbouring clip, avoiding neighbours where the sentences are matching (e.g. the sentence, “mix ingredients” might be repeated in consecutive clips).

In practice, the neighbouring loss is calculated by having another batch of sampled neighbouring clips of size B. We use a single negative neighbour per clip to keep the batch size to B regardless of the length m, though we do ablate differing numbers of sampled negatives in Sec. 4.2.

We optimize our ConTra model by minimizing the overall loss function L:

$$L = L_{CML} + L_{NEI} + L_{UNI} \tag{9}$$

We keep the weights between the three losses the same in all experiments and datasets showcasing that we outperform other approaches without hyperparameter tuning. In supplementary, we report results when tuning the weights, to ablate these.

Once the model is trained, it can be used for both sentence-to-clip and clip-to-sentence retrieval. The clip is enriched with the context, whether used in the gallery set (in sentence-to-clip) or in the query (in clip-to-sentence). When performing sentence-to-clip retrieval, our query consists of only one sentence as usually done in other approaches, and is thus comparable to these. During inference, the gallery of clips is always given, and thus all approaches have access to the same information.

3.4 Multi-modal Context

In previous sections (Sec. 3.1–3.3), we motivated our approach by focusing on local clip context—i.e. context in the visual modality. However, ConTra could similarly be applied to the local context of the text modality. As an example, given steps of a recipe, these can be utilised to build a Text Context (TC), such that:

$$TC_m(t_{ij}) = (t_{i(j-m)}, \cdots, t_{ij}, \cdots, t_{i(j+m)}) \tag{10}$$

To give an example for clip-to-sentence retrieval, a single clip is used as the query, but the gallery is constructed of captions that have had their representation enriched via sentences of neighbouring clips (e.g. “Add the mince to the pan” has attended to “Take the beef mince out of its wrapper” and “Fry until the mince is browned”). This contextual text knowledge could come from video narrations or steps in a recipe.

We also assess the utilisation of context in both modalities. This setup assumes access to local context in both clip and sentence. L_{NEI} is thus applied to both neighbouring clip contexts and text contexts, using one negative for each case. The architecture for both f_{cc} and f_{tc} are identical, but are learned as two separate embedding functions with unique weights and positional embeddings[1].

4 Results

We first present our experimental settings and the choice of untrimmed datasets in Sec. 4.1. We then focus on clip context results including comparison with state-of-the-art (SOTA) methods in Sec. 4.2 before exploring text context and context in both modalities in Sec. 4.3. Finally, we discuss limitations and avenues for future work.

4.1 Experimental Settings

Datasets. Video datasets commonly used for cross-modal retrieval can be split into two groups: trimmed and untrimmed. In trimmed datasets, such as MSRVTT [66], MSVD [8] and VATEX [59], the full video is considered as a single clip and thus no context can be utilised. In Table 2, we compare the untrimmed datasets for their size and the number of clips per video. Datasets with 1-2 clips per video on average limit the opportunity to explore long or local temporal context. While we include QuerYD [44] in the table, this dataset does not allow for context to be explored as clips from the same video are split between the train and test sets. We choose to evaluate our method on three untrimmed datasets, whose average number of clips/video is greater than 3. We describe the notion of context in each:

YouCook2 [71] contains YouTube cooking videos. On average, training videos contain 7.75 clips, each associated with a sentence. The dataset has been evaluated for clip-sentence retrieval [23, 52, 40, 70] as well as video-paragraph retrieval [22]. We focus on clip-sentence retrieval, utilising the local context, which represents previous/follow-up steps in a recipe. Given YouCook2’s popularity, we use it for all ablation experiments.

[1] We experimented with sharing these embeddings but similar to previous approaches [28], this performed worse, see supplementary.

	#clips		**#videos**		**#clips per video**	
Datasets	Train	Test	Train	Test	Train	Test
Charades-STA [21]	5,657	1,596	5338	1334	1.60	1.20
DiDeMo [27]	21,648	2,650	8511	1037	2.21	2.21
QuerYD* [44]	9,118	1,956	1,283	794	8.3	
YouCook2 [71]	10,337	3,492	1,333	457	7.75	7.64
ActivityNet CS [33]	37,421	17505	10,009	4,917	3.74	3.56
EPIC-KITCHENS-100 [12]	67,217	9,668	495	138	135.79	70.06

Table 2: Comparing untrimmed video datasets by size and number of clips per video. *QuerYD videos are split across train and test so we report overall clips/video.

ActivityNet Captions [33] consists of annotated YouTube videos from ActivityNet [26]. The dataset has only been evaluated for video-paragraph retrieval [22, 67, 34] where all clips in the same video are concatenated, and all corresponding captions are also concatenated to form the paragraph. Instead, we consider the *val_1* split, and introduce an **ActivityNet Clip-Sentence (CS)** variant using all the individual clips and their corresponding captions/sentences. We emphasise that this evaluation *cannot* be compared to published results on video-paragraph retrieval and instead evaluate two methods to act as baselines for comparison.

EPIC-KITCHENS-100 [12] offers a unique opportunity to explore context in significantly longer untrimmed videos. On average, there are 135.8 clips per video of kitchen-based actions, shot from an egocentric perspective. We use the train/test splits for the multi-instance retrieval benchmark, but evaluate on the single-instance retrieval task using the set of unique narrations.

Evaluation Metrics. We report the retrieval performance, for both clip-to-sentence and sentence-to-clip tasks, using the two standard metrics of: Recall at $K = \{1, 5, 10\}$ (R@K) and median rank (MR). We also report the sum of cross-modal R@K as RSum to demonstrate overall performance. Where figures are plotted, tables of exact results are given in supplementary.

Visual Features. To be comparable to prior work, we use the same features as recent methods per dataset. For YouCook2, we use the S3D backbone provided by [40] pre-trained on [42] , extracting 1024-d features. We uniformly sample 32 frames from each clip with a 224x224 resolution. For ActivityNet CS, we use frame features provided by [67]. These frame features are combined into clip features using a single transformer, trained with shared weights across clips, as proposed in [22], obtaining 384-d features. Note that this transformer is trained for clip-sentence alignment, using the code from [22], without the global context and contextual transformer. For EPIC-KITCHENS-100, we use the publicly available 3072-d features from [31].

Text Features. For YouCook2 and EPIC-KITCHENS-100, we take a maximum of 16 words without removing stopwords from each sentence and we extract 2048-d feature vectors using the text branch in [40] pre-trained on [42]. This consists of a linear layer with a ReLU activation applied independently to each word embedding followed by max pooling and a randomly initialised linear layer to reduce dimensionality. We fine-tune the text branch layer, to accommodate missing vocabulary[2]. For ActivityNet CS,

[2] We add 174 and 104 missing words from the model in [42] for YouCook2 and EPIC-KITCHENS-100 respectively.

we feed the sentences into a pretrained BERT-Base Uncased model [13] and use the per-token outputs of the last 2 layers to train a sentence transformer, as in [22], and obtain 384-d text features.

Architecture Details. The number of layers and heads in the ConTra encoder differs depending on the dataset size. For the small-scaled YouCook2, we use 1 layer and $R = 2$ heads to avoid overfitting. For the larger two datasets we use 2 layers and $R = 8$ heads. The inner dimension of the transformers is 2048-d. The learnt positional encoding matches the feature dimension: 512-d for YouCook2 and EPIC-KITCHENS-100, and 384-d for ActivityNet CS, initialised from $\mathcal{N}(0, 0.001)$. Due to the small dimension of the features of ActivityNet CS, we remove the linear projection P from our architecture. We apply dropout of 0.3 at h_j and h'_j.

Implementation Details. We use the Adam optimizer with a starting learning rate of 1×10^{-4} and decrease it linearly after a warmup period of 1300 iterations. The size of the batch is fixed to 512. The temperature τ in Eq. 6 and 8 is set to 0.07 as in [$25, 45,$ 63], and the dimension of the common embedding space Ω is set to 512 for YouCook2 and EPIC-KITCHENS-100, and 384 for ActivityNet CS. We ablate these values in supplementary. If the clip does not have sufficient temporal context (i.e. is at the start/end of the video), we pad the input by duplicating the first/last clip to obtain a fixed-length context. *Our code is available at https://github.com/adrianofragomeni/ConTra*

4.2 Clip Context Results

Context Length Analysis. We analyse the effect of clip context length by varying m from its no-context baseline, $m = 0$, to $m = 5$. Results are presented in Fig. 3. In all three datasets, the largest improvement is obtained comparing $m = 0$, i.e. no context, to $m = 1$, i.e. introducing the smallest context, where the RSum increases by 8.5, 17.6 and 23.2 for Youcook2, ActivityNet CS, and EPIC-KITCHENS-100, respectively. This highlights that neighbouring clips are able to improve the retrieval performance. Moreover, every $m > 0$ outperforms $m = 0$ on all datasets. We obtain the best performance on YouCook2 at $m = 3$. ActivityNet CS also obtained best performance when using $m = 3$, and EPIC-KITCHENS-100 when $m = 4$. Although ConTra introduces a new hyperparameter, m, Fig. 3 shows that RSum saturates when $m \geq 3$ across all datasets. Using a larger context does not further improve the performance. We show the attention weights learned by the multi-headed attention layers in Fig. 4 averaged over all videos, per layer (left), and for specific examples (right). YouCook2 focuses more on the later clips due to the recipes being more recognisable when ingredients are brought together. The attention weights for EPIC-KITCHENS-100 and ActivityNet CS are higher for earlier clips, with ActivityNet's first layer and EPIC-KITCHENS-100's second layer attending to past clips. EPIC-KITCHENS-100 specifically has higher attention weights

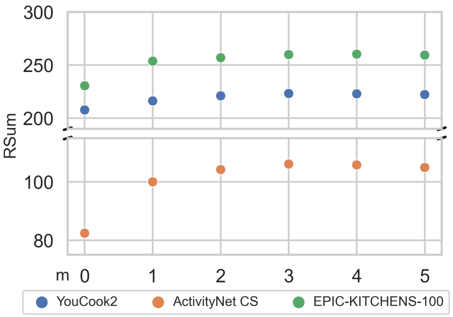

Fig. 3: Analysis of clip context (CC) with differing m across YouCook2, ActivityNet CS and EPIC-KITCHENS-100.

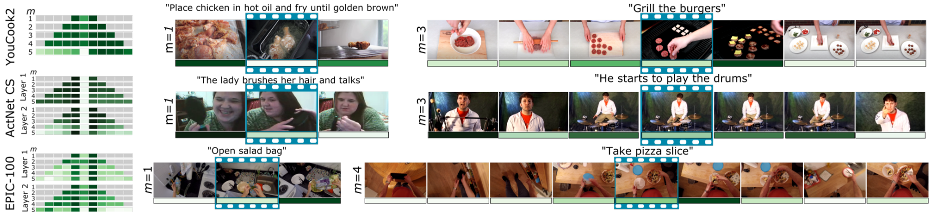
YouCook2
m
1
2
3
4
5
ActNet CS
Layer 1
Layer 2
EPIC-100
"Place chicken in hot oil and fry until golden brown"
m=1
"Grill the burgers"
m=3
"The lady brushes her hair and talks"
m=1
"He starts to play the drums"
m=3
"Open salad bag"
m=1
"Take pizza slice"
m=4

Fig. 4: Left: average attention weights over videos as m changes, per dataset and layer. Right: qualitative examples with clip attention.

on directly neighbouring clips. From the examples in Fig. 4 (right), ConTra uses local context to discriminate objects which may be occluded, such as chicken in YouCook2, the contents of the salad bag in EPIC-KITCHENS-100, or the brush in ActivityNet CS.

In Fig. 5, we analyse how individual words are affected by using context. On a word-by-word basis we find all captions that contain a given word and count the number of times the rank of those captions improved/worsened after adding context. E.g., the word 'pan' in YouCook2 is present in 373 captions, in which 182 captions improve their rank with context while 126 captions worsen their rank resulting in a delta of 56. In YouCook2, 'salt', 'sauce', and 'oil' all see a large improvement when using context, likely due to easily being obscured by their containers. In comparison, for EPIC-KITCHENS-100, verbs benefit the most from context—these actions tend to be very short so surrounding context can help discriminate them.

Overall, these results showcase that using local Clip Context (CC) enhances the clip embedding representation and consistently results in a boost in performance.

Comparison with State of the Art. The most commonly used untrimmed dataset in prior work is YouCook2. For fair comparison, we split the SOTA methods on this dataset into blocks according to the pre-training and fine-tuning datasets: (i) training only on YouCook2, (ii) training only on other large-scale datasets, (iii) pre-training on large-scale datasets then fine-tuning on YouCook2; note that this is where ConTra lies and (iv) additionally, pre-training with proxy tasks on large-scale datasets. Table 3 compares ConTra with the SOTA on YouCook2.

Overall, ConTra outperforms all directly-comparable SOTA works [42, 23, 22]. ConTra outperforms COOT [22] that trains for video-paragraph retrieval on the full video. Distinct from COOT [22], we only use clip context, i.e. single sentences in training and inference, and local context.

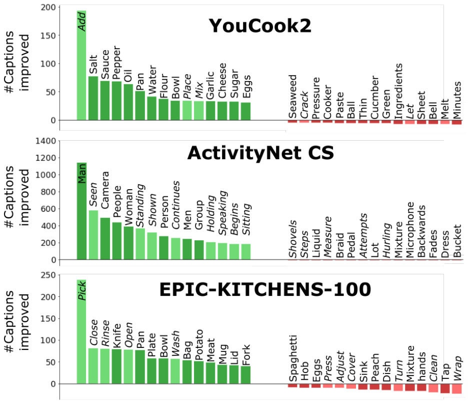

Fig. 5: 15 most improved/hindered words per dataset when context is used. Verbs are lighter/italicised. Best viewed in colour.

	Method	PX	FT	R@1	R@5	R@10	MR	RSum
(i)	HGLMM [32]	×	✓	4.6	14.3	21.6	75	40.5
	UniVL (FT-joint) [38]	×	✓	7.7	23.9	34.7	21	66.3
(ii)	ActBert [72]	✓	×	9.6	26.7	38.0	19	74.3
	MMV FAC [2]	×	×	11.7	33.4	45.4	13	90.5
	VATT-MBS [1]	×	×	-	-	45.5	13	-
	MCN [7]	×	×	18.1	35.5	45.2	-	98.8
	MIL-NCE [40]	×	×	15.1	38.0	51.2	10	104.3
(iii)	HowTo100M [42]	×	✓	8.2	24.5	35.3	24	68.0
	GRU+SSA [23]	×	✓	10.9	28.4	-	-	-
	COOT [22]	×	✓	16.7	40.2	52.3	9	109.2
	MIL-NCE (from [70])	×	✓	15.8	40.3	54.1	8	110.2
	ConTra (ours)	×	✓	16.7	42.1	55.2	8	114.0
(iv)	CUPID [70]	✓	✓	17.7	43.2	57.1	7	117.9
	DeCEMBERT [52]	✓	✓	17.0	43.8	59.8	9	120.6
	UniVL (FT-joint) [38]	✓	✓	22.2	52.2	66.2	5	140.6
	VLM [64]	✓	✓	27.0	56.9	69.4	4	153.3
	VideoCLIP [65]	✓	✓	32.2	62.6	75.0	-	169.8

Table 3: Sentence-to-Clip comparison with SOTA on YouCook2 test set. PX: pre-train end-to-end with proxy tasks. FT: fine-Tuning on YouCook2. $-$: unreported results.

			Sentence-to-Clip				**Clip-to-Sentence**				
	Method	FT	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR	RSum
(ii)	MIL-NCE [40]*	×	2.4	6.8	10.0	460	2.1	6.2	9.2	543	36.7
(iii)	HowTo100M [42]*	✓	3.8	12.5	18.9	68	3.6	11.4	17.3	78	67.5
	COOT [22] $-g^*$	✓	3.7	11.9	18.6	67	3.7	12.0	18.8	64	68.7
	ConTra (ours)	✓†	5.9	18.4	27.6	38	6.4	19.3	28.5	37	106.1
	COOT [22]*	✓	6.2	18.8	28.4	33	6.3	19.0	28.4	32	107.0

Table 4: Comparison on ActivityNet CS for Clip-Sentence Retrieval. *Our reproduced results. †: Transformer fine-tuned first—clip/sentence features match those from COOT [22] $-g$.

			Sentence-to-Clip				**Clip-to-Sentence**				
	Method	FT	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR	RSum
(ii)	MIL-NCE [40]*	×	3.2	10.4	15.4	188	2.1	7.8	12.3	194	51.2
(iii)	JPoSE [61]*	✓	2.5	7.5	11.6	13	4.4	17.4	27.2	17	70.6
	ConTra (ours)	✓	22.2	43.4	53.4	9	28.2	52.0	61.1	5	260.3

Table 5: Comparison with baseline on EPIC-KITCHENS-100. *Our reproduced results.

The last block of Table 3 includes works that are not directly comparable to ConTra, as these models are pre-trained *end-to-end* on HowTo100M with additional proxy tasks, e.g. masked language modelling, whereas ConTra is initialised randomly. Although DeCEMBERT [52] is not directly comparable, ConTra is less complex with $9.5M$ parameters compared to DeCEMBERT's $115.0M$ and our results are only marginally lower.

To the best of our knowledge, no prior work has evaluated on ActivityNet CS for clip-sentence retrieval. For comparison, we evaluate [40, 42] on ActivityNet CS for clip-sentence retrieval using public code. We also run the code from COOT [22], trained on video-paragraph retrieval, and obtain their results on clip-sentence retrieval. We then replace the text input with sentence-level representa-

	Sentence-to-Clip				Clip-to-Sentence				
Loss	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR	RSum
L_{NEI}	6.4	18.3	27.3	39	4.3	15.5	23.6	43	95.4
L_{CML}	15.7	39.9	53.4	9	14.5	38.5	52.3	**9**	214.3
L_{CML}+$L_{HardMining}$	15.7	39.8	53.5	9	14.2	39.0	51.9	10	214.1
L_{CML}+L_{NEI}	16.2	41.4	54.1	9	**14.8**	39.3	52.6	**9**	218.4
L_{CML}+L_{NEI}+L_{UNI}	**16.7**	**42.1**	**55.2**	**8**	**14.8**	**40.5**	**53.9**	**9**	**223.2**

Table 6: Ablation of loss function terms: Neighbouring Loss (L_{NEI}), Cross Modal Loss (L_{CML}), Hard Triplet Mining ($L_{HardMining}$), and Uniformity Loss (L_{UNI}).

tions, and remove their global alignment to produce the COOT$-g$ baseline reported above. Table 4 shows that ConTra outperforms MIL-NCE [40], HowTo100M [42] and COOT$-g$ by a considerable margin. Our RSum is only marginally lower than COOT, where global context is considered for both modalities during training.

Note that methods that train for global context cannot be used for datasets with hundreds of clips per video, like EPIC-KITCHENS-100. In Table 5, we compare ConTra to JPoSE [61] and our reproduced results of MIL-NCE [40] on EPIC-KITCHENS-100, outperforming on all metrics by a large margin. We cannot train or evaluate COOT on EPIC-KITCHENS-100 which has 136 clips per video on average. Additionally, clips in EPIC-KITCHENS-100 are significantly shorter, increasing the benefits of attending to local context.

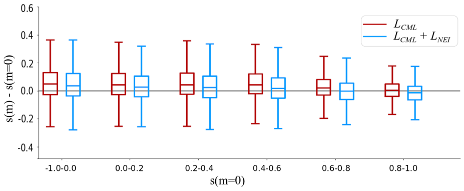

Fig. 6: Comparison between similarities to neighbouring clips, $s(f_{cc}(j + 1), f_s(j))$, with and without using L_{NEI}. Without L_{NEI} ConTra gives higher similarities to neighbouring clips.

Ablation studies. We ablate ConTra on YouCook2. Ablations on the other two datasets are in supplementary.

Loss Function. In Table 6, we first test the formulation of the two losses L_{NEI} and L_{CML} individually, using the NCE loss [64, 1, 7, 35]. On its own, L_{NEI} learns with limited variety of only neighbouring clips as negatives.

Then, we compare L_{NEI} to the standard hard mining approach proposed in [18]. L_{NEI} consistently outperforms hard mining. Our proposed loss L, with its 3 terms, performs the best, improving RSum by 4.8 when adding the uniformity loss L_{UNI} which allows preserving maximal information and so obtains better embeddings.

We further demonstrate the benefits of L_{NEI} in Fig. 6. We bin the neighbouring clips $j \pm 1$ based on their similarity to the sentence j along the x-axis. We then calculate the difference between this similarity with and without context, and provide the average and extent of these differences in a box plot over all datasets. When this difference, on the y-axis, is > 0, the context transformer would have increased the similarity between the neighbouring clip and the sentence. Without L_{NEI}, the similarity is increased further, particularly for clips and sentences with low cosine similarity, depicted on the x-axis.

Number of negatives neighbouring clips. Table 7 shows how the performance changes when we consider more than one negative for our neighbouring loss L_{NEI}. Increasing the negatives from 1 to 2 improves the retrieval results marginally. RSum remains the same when increasing the number of negatives further to 3. We keep the number of negatives equal to 1 in all the experiments.

	Sentence-to-Clip				**Clip-to-Sentence**				
#Negatives	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR	RSum
1	16.7	42.1	55.2	**8**	14.8	40.5	**53.9**	**9**	223.2
2	16.9	42.2	**55.7**	**8**	**15.6**	40.5	**53.9**	**9**	**224.8**
3	**17.3**	**42.4**	55.6	**8**	14.9	**40.7**	**53.9**	**9**	**224.8**

Table 7: Analysis of the performance varying the number of negative in L_{NEI}.

Aggregate context. As explained in Sec. 3.2, we select the middle output of the transformer encoder as our clip embedding. In order to justify this design choice, we compare to other aggregation approaches. These are of two types based on where local context is aggregated, i.e the visual features h_j or the outputs of the clip transformer encoder f_{cc}. Moreover we experimented with two aggregation techniques, Maximum and Average. Table 8 shows that aggregating features has a poor performance. Moreover, using the middle output outperforms the other two aggregation techniques, as the model is enriching the embedding of the anchor clip from its contextual neighbours.

		Sentence-to-Clip				**Clip-to-Sentence**				
h_j	f_{cc}	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR	RSum
Avg	-	4.9	16.3	26.3	40	4.3	16.0	25.8	41	93.6
Max	-	3.1	13.1	22.0	46	3.4	13.4	21.7	49	76.7
-	Avg	15.6	40.1	53.9	9	14.0	38.6	51.6	10	213.8
-	Max	**16.9**	41.7	54.8	**8**	14.7	40.1	53.7	**9**	221.9
-	Mid	16.7	**42.1**	**55.2**	**8**	**14.8**	**40.5**	**53.9**	**9**	**223.2**

Table 8: Comparing aggregation approaches.

4.3 Results of Modality Context

In Fig. 7 and Fig. 8 we provide comparable analysis as m increases from no context ($m = 0$) up to $m = 5$, for text context (Fig. 7) and context in both modalities (Fig. 8). Results consistently demonstrate context to be helpful in both cases, $m = 1$ outperforms $m = 0$ by a large margin in every case. For some datasets, e.g. ActivityNet CS, performance saturates and drops slightly for $m > 3$. For long videos, e.g. EPIC-KITCHENS-100, performance continues to improve with larger context.

In Fig. 9 we show qualitative examples from models trained with clip context and context in both modalities. Clip context (left) shows additional visual context helps. For example, in row 1, the previous clip includes potatoes before being mashed—a more recognisable shape compared to their mashed state. When using context in both modalities (right), these benefits are combined, leading to the model discriminating between difficult examples in which very similar clips are described.

Limitations. We find that local context is certainly beneficial for Clip-Sentence retrieval, but also acknowledge there are cases in which it is detrimental. An example of this is in Fig. 9 (row 3, left—"add water"), context drops the rank of the correct clip from 4 to 15. Studying the correct video, we note that neighbouring clips are not always related to the main action, i.e. —"turn on the cooker". The enriched clip representation

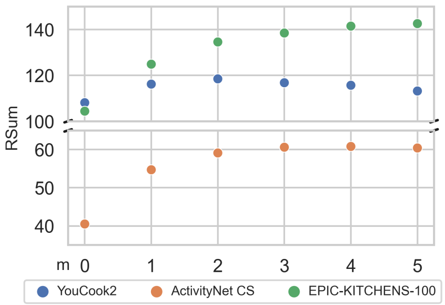

Fig. 7: Analysis of temporal text context (TC), reporting RSum in S2C.

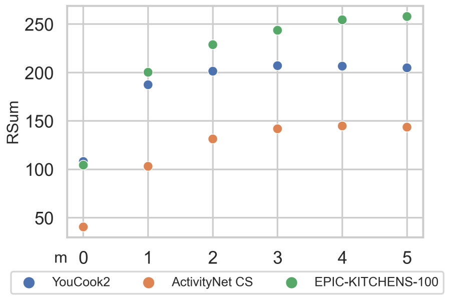

Fig. 8: Analysis of temporal both context (BC), reporting RSum in S2C.

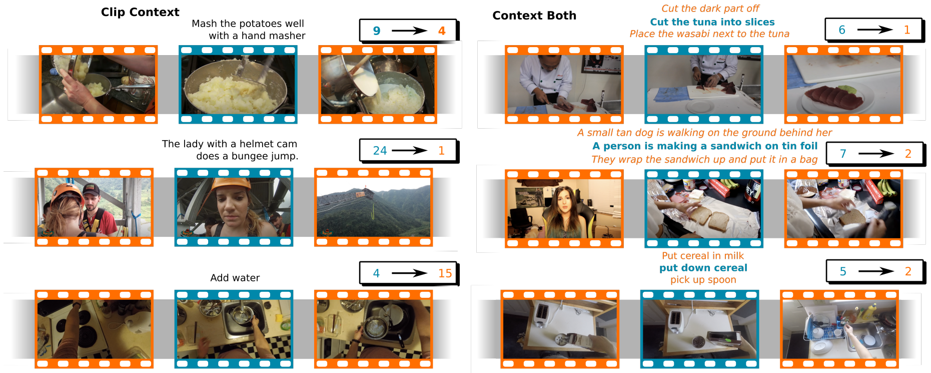
Clip Context
Mash the potatoes well with a hand masher
9 → 4
The lady with a helmet cam does a bungee jump.
24 → 1
Add water
4 → 15
Context Both
Cut the dark part off
Cut the tuna into slices
Place the wasabi next to the tuna
6 → 1
A small tan dog is walking on the ground behind her
A person is making a sandwich on tin foil
They wrap the sandwich up and put it in a bag
7 → 2
Put cereal in milk
put down cereal
pick up spoon
5 → 2

Fig. 9: Qualitative clip-to-sentence results for clip context (left) and context in both modalities (right) from 3 datasets: YouCook2 (top), ActivityNet CS (middle), and EPIC-KITCHENS-100 (bottom). The change in rank of retrieved video from no context (cyan) to using context (orange), e.g. from rank 9 to rank 4 when using context.

is thus less similar to the query sentence. In Fig 5, we also show specific words harmed by clip context.

5 Conclusions

In this work, we introduce the notion of local temporal context, for clip-sentence cross-modal retrieval in long videos. We propose an attention-based deep encoder, which we term Context Transformer (ConTra), that is trained using contrastive losses from the embedding space. We demonstrate the impact of ConTra on individual modalities as well as both modalities in cross-modal retrieval. We ablate our method to further show the benefit of each component. Our results indicate, both qualitatively and by comparing to other approaches, that local context in retrieval decreases the ambiguity in clip-sentence retrieval on three video datasets.

Acknowledgements This work used public dataset and was supported by EPSRC UMPIRE (EP/T004991/1) and Visual AI (EP/T028572/1).

References

1. Akbari, H., Yuan, L., Qian, R., Chuang, W., Chang, S., Cui, Y., Gong, B.: VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. In: Conference on Neural Information Processing Systems (NeurIPS) (2021)
2. Alayrac, J., Recasens, A., Schneider, R., Arandjelovic, R., Ramapuram, J., Fauw, J.D., Smaira, L., Dieleman, S., Zisserman, A.: Self-supervised multimodal versatile networks. In: Conference on Neural Information Processing Systems (NeurIPS) (2020)
3. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: International Conference on Computer Vision (ICCV) (2021)
4. Beery, S., Wu, G., Rathod, V., Votel, R., Huang, J.: Context R-CNN: long term temporal context for per-camera object detection. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
5. Bertasius, G., Torresani, L.: Classifying, segmenting, and tracking object instances in video with mask propagation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
6. Cartas, A., Radeva, P., Dimiccoli, M.: Modeling long-term interactions to enhance action recognition. In: International Conference on Pattern Recognition (ICPR) (2021)
7. Chen, B., Rouditchenko, A., Duarte, K., Kuehne, H., Thomas, S., Boggust, A.W., Panda, R., Kingsbury, B., Feris, R.S., Harwath, D., Glass, J.R., Picheny, M., Chang, S.: Multimodal clustering networks for self-supervised learning from unlabeled videos. In: International Conference on Computer Vision (ICCV) (2021)
8. Chen, D.L., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Association for Computational Linguistics (ACL/IJCNLP) (2011)
9. Chen, S., Zhao, Y., Jin, Q., Wu, Q.: Fine-grained video-text retrieval with hierarchical graph reasoning. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
10. Chun, S., Oh, S.J., de Rezende, R.S., Kalantidis, Y., Larlus, D.: Probabilistic embeddings for cross-modal retrieval. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
11. Croitoru, I., Bogolin, S.V., Leordeanu, M., Jin, H., Zisserman, A., Albanie, S., Liu, Y.: TeachText: Crossmodal generalized distillation for text-video retrieval. In: International Conference on Computer Vision (ICCV) (2021)
12. Damen, D., Doughty, H., Farinella, G.M., , Furnari, A., Ma, J., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. International Journal of Computer Vision (IJCV) (2021)
13. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT) (2019)
14. Dong, J., Li, X., Snoek, C.G.: Word2visualvec: Image and video to sentence matching by visual feature prediction. CoRR, abs/1604.06838 (2016)
15. Dong, J., Li, X., Xu, C., Yang, X., Yang, G., Wang, X., Wang, M.: Dual encoding for video retrieval by text. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2021)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations, (ICLR) (2021)

17. El-Nouby, A., Neverova, N., Laptev, I., Jégou, H.: Training vision transformers for image retrieval. CoRR (2021)
18. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: improving visual-semantic embeddings with hard negatives. In: British Machine Vision Conference (BMVC) (2018)
19. Furnari, A., Farinella, G.M.: What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In: International Conference on Computer Vision (ICCV) (2019)
20. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: European Conference on Computer Vision (ECCV) (2020)
21. Gao, J., Sun, C., Yang, Z., Nevatia, R.: TALL: temporal activity localization via language query. In: International Conference on Computer Vision (ICCV) (2017)
22. Ging, S., Zolfaghari, M., Pirsiavash, H., Brox, T.: COOT: cooperative hierarchical transformer for video-text representation learning. In: Conference on Neural Information Processing Systems (NeurIPS) (2020)
23. Guo, X., Guo, X., Lu, Y.: Ssan: Separable self-attention network for video representation learning. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
24. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. Journal of Machine Learning Research (2012)
25. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B.: Momentum contrast for unsupervised visual representation learning. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
26. Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
27. Hendricks, L.A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.C.: Localizing moments in video with temporal language. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2018)
28. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: International Conference on Machine Learning (ICML) (2021)
29. Józefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., Wu, Y.: Exploring the limits of language modeling. CoRR, abs/1602.02410 (2016)
30. Kazakos, E., Huh, J., Nagrani, A., Zisserman, A., Damen, D.: With a little help from my temporal context: Multimodal egocentric action recognition. In: British Machine Vision Conference (BMVC) (2021)
31. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: International Conference on Computer Vision (ICCV) (2019)
32. Klein, B., Lev, G., Sadeh, G., Wolf, L.: Associating neural word embeddings with deep image representations using fisher vectors. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
33. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos. In: International Conference on Computer Vision (ICCV) (2017)
34. Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T.L., Bansal, M., Liu, J.: Less is more: Clipbert for video-and-language learning via sparse sampling. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
35. Liu, S., Fan, H., Qian, S., Chen, Y., Ding, W., Wang, Z.: Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)

36. Liu, Y., Albanie, S., Nagrani, A., Zisserman, A.: Use what you have: Video retrieval using representations from collaborative experts. In: British Machine Vision Conference (BMVC) (2019)
37. Liu, Y., Chen, Q., Albanie, S.: Adaptive cross-modal prototypes for cross-domain visual-language retrieval. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
38. Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Chen, X., Zhou, M.: Univilm: A unified video and language pre-training model for multimodal understanding and generation. CoRR, abs/2002.06353 (2020)
39. Miech, A., Alayrac, J., Laptev, I., Sivic, J., Zisserman, A.: Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
40. Miech, A., Alayrac, J., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
41. Miech, A., Laptev, I., Sivic, J.: Learning a text-video embedding from incomplete and heterogeneous data. CoRR, abs/1804.02516 (2018)
42. Miech, A., Zhukov, D., Alayrac, J., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: International Conference on Computer Vision (ICCV) (2019)
43. Mithun, N.C., Li, J., Metze, F., Roy-Chowdhury, A.K.: Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In: International Conference on Multimedia Retrieval (ICMR) (2018)
44. Oncescu, A., Henriques, J.F., Liu, Y., Zisserman, A., Albanie, S.: QUERYD: A video dataset with high-quality text and audio narrations. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2021)
45. Patrick, M., Asano, Y.M., Kuznetsova, P., Fong, R., Henriques, J.F., Zweig, G., Vedaldi, A.: Multi-modal self-supervision from generalized data transformations. In: International Conference on Computer Vision (ICCV) (2021)
46. Patrick, M., Huang, P., Asano, Y.M., Metze, F., Hauptmann, A.G., Henriques, J.F., Vedaldi, A.: Support-set bottlenecks for video-text representation learning. In: International Conference on Learning Representations (ICLR) (2021)
47. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
48. Sener, F., Singhania, D., Yao, A.: Temporal aggregate representations for long-range video understanding. In: European Conference on Computer Vision (ECCV) (2020)
49. Shao, D., Xiong, Y., Zhao, Y., Huang, Q., Qiao, Y., Lin, D.: Find and focus: Retrieve and localize video events with natural language queries. In: European Conference on Computer Vision (ECCV) (2018)
50. Shao, J., Wen, X., Zhao, B., Xue, X.: Temporal context aggregation for video retrieval with contrastive learning. In: Winter Conference on Applications of Computer Vision (WACV) (2021)
51. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: VideoBERT: A joint model for video and language representation learning. In: International Conference on Computer Vision (ICCV) (2019)
52. Tang, Z., Lei, J., Bansal, M.: Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (2021)

53. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Conference on Neural Information Processing Systems (NeurIPS) (2017)
54. Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y.: Learning fine-grained image similarity with deep ranking. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
55. Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2018)
56. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
57. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning (ICML) (2020)
58. Wang, X., Zhu, L., Yang, Y.: T2VLAD: global-local sequence alignment for text-video retrieval. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
59. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y., Wang, W.Y.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: International Conference on Computer Vision (ICCV) (2019)
60. Wei, J., Xu, X., Yang, Y., Ji, Y., Wang, Z., Shen, H.T.: Universal weighting metric learning for cross-modal matching. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
61. Wray, M., Csurka, G., Larlus, D., Damen, D.: Fine-grained action retrieval through multiple parts-of-speech embeddings. In: International Conference on Computer Vision (ICCV) (2019)
62. Wu, C., Feichtenhofer, C., Fan, H., He, K., Krähenbühl, P., Girshick, R.B.: Long-term feature banks for detailed video understanding. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
63. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
64. Xu, H., Ghosh, G., Huang, P., Arora, P., Aminzadeh, M., Feichtenhofer, C., Metze, F., Zettlemoyer, L.: VLM: task-agnostic video-language model pre-training for video understanding. In: Association for Computational Linguistics (ACL/IJCNLP) (2021)
65. Xu, H., Ghosh, G., Huang, P.Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., Feichtenhofer, C.: VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2021)
66. Xu, J., Mei, T., Yao, T., Rui, Y.: MSR-VTT: A large video description dataset for bridging video and language. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
67. Zhang, B., Hu, H., Sha, F.: Cross-modal and hierarchical modeling of video and text. In: European Conference on Computer Vision (ECCV) (2018)
68. Zhang, C., Gupta, A., Zisserman, A.: Temporal query networks for fine-grained video understanding. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
69. Zhang, Z., Han, X., Song, X., Yan, Y., Nie, L.: Multi-modal interaction graph convolutional network for temporal language localization in videos. IEEE Trans. Image Process. (2021)
70. Zhou, L., Liu, J., Cheng, Y., Gan, Z., Zhang, L.: CUPID: Adaptive curation of pre-training data for video-and-language representation learning. CoRR, abs/2104.00285 (2021)
71. Zhou, L., Xu, C., Corso, J.J.: Towards automatic learning of procedures from web instructional videos. In: Conference on Artificial Intelligence (AAAI) (2018)

72. Zhu, L., Yang, Y.: ActBERT: Learning global-local video-text representations. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020)