VTC: Improving Video-Text Retrieval with User Comments

Laura Hanu[1], James Thewlis[1], Yuki M. Asano[2], and Christian Rupprecht[3]

[1] Unitary Ltd.
[2] University of Amsterdam
[3] University of Oxford

Abstract. Multi-modal retrieval is an important problem for many applications, such as recommendation and search. Current benchmarks and even datasets are often manually constructed and consist of mostly clean samples where all modalities are well-correlated with the content. Thus, current video-text retrieval literature largely focuses on video titles or audio transcripts, while ignoring user comments, since users often tend to discuss topics only vaguely related to the video. Despite the ubiquity of user comments online, there is currently no multi-modal representation learning datasets that includes comments. In this paper, we a) introduce a new dataset of videos, titles and comments; b) present an attention-based mechanism that allows the model to learn from sometimes irrelevant data such as comments; c) show that by using comments, our method is able to learn better, more contextualised, representations for image, video and audio representations. Project page: https://unitaryai.github.io/vtc-paper.

1 Introduction

Training large scale multi-modal models from paired visual/text data from the web has seen great success in video understanding and retrieval. However, typically only the caption (i.e. title or "alt text") is used, ignoring potentially relevant text present on the web page such as user comments.

We explore how to leverage comments for the task of video-text retrieval. We consider how comments can be seen as an extra modality, yet with the peculiar characteristics that they are neither inherently derived from the video (as text from speech or OCR would be), nor are they merely extra captions which can be used in place of the title. This results in two different, yet equally interesting research questions: "Can we use comments to augment and adapt our title representations?" and "Can we use them to adapt our video features?" We address both of these in this paper.

A challenge is that comments may often be only tangentially related to the contents of the video (e.g. "cool video!"), or may be relevant but non-distinctive ("cute cat!" applies to many videos). Yet, since comments often discuss contextual details lacking from the title or video themselves, we hypothesize that correctly

leveraging this signal can improve retrieval, endowing either the query or target features with extra context.

Other modalities can also exhibit this behavior, for example, many current works that learn from audio-visual correspondence [2,1,41] leverage clean datasets such as Kinetics [7] or VGGSound [8] to learn meaningful correspondence between videos and sound, whereas online videos tend to have for example background music that replaces the actual sounds happening in the video, or images overlaid with sounds.

In this paper, we propose a method that can take advantage of this auxiliary context provided by comments while simultaneously filtering it for meaningful information. Most current models enforce a strict correlation between the different input modalities under the assumption that all are informative of the content. The main intuition of our work is that when training a model on partially unrelated data, we need to introduce a mechanism that allows the model to discount auxiliary data when it is not helpful for the task.

To this end, we build a model with a hierarchical attention structure. Current representation learning models that are based on transformer architectures already exploit the idea of an attention mechanism to model the correlation between different *parts* of an input signal. For example in text understanding, the attention mechanism is applied per word, allowing the model to understand the structure of natural language. Even though in principle one could use the same scheme to model the importance of different text inputs on a per-word basis, we find that this makes it difficult to learn the individual importance of inputs. Moreover, due to the computational complexity of current transformers (squared with sequence length) this approach would only work for a small number of comments. We thus add a second layer of attention *per processed input* that allows the model to assess the amount of information at a higher level of features. With quantitative and qualitative experiments, we find that this mechanism aligns well with the intuition that some inputs are very relevant to the problem and others can be disregarded.

To the best of our knowledge, there is no large scale dataset that contains videos, titles, and user comments. Thus, to advance the field of representation learning, we introduce "VTC" (Videos, Titles and Comments), a new dataset of 339k videos and titles with an average of 14 comments per video with which we train and evaluate our representations. A more detailed summary of the dataset statistics can be found in the Appendix. In our experiments we show that we can indeed learn meaningful information from user comments for three different modalities: audio, images, and video and that representations learned can generalize to other datasets. Additionally, we show that the model can correctly identify whether auxiliary information is informative of the content of a video or not.

The ability to incorporate auxiliary contextual information also opens up possibilities for useful applications. In the video retrieval setting, our method can be used to iteratively refine a text descriptor with new inputs as shown in Figure 1, allowing incremental searching. In the zero-shot video classification

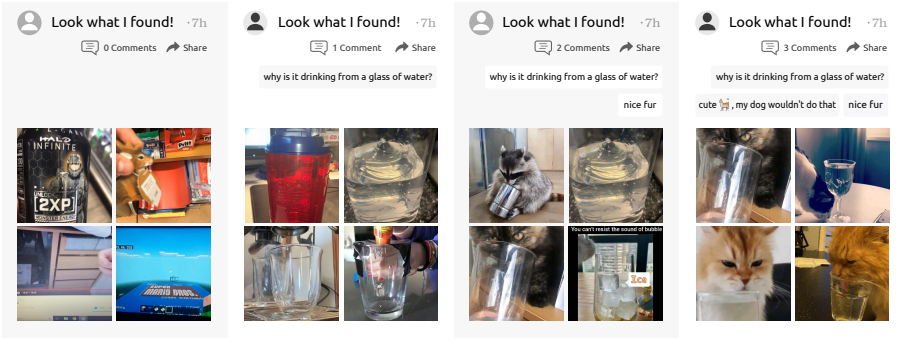


Fig. 1: Video retrieval from title and comments. We show the top 4 videos retrieved for the ambiguous title *"Look what I found!"*. From left to right, we progressively add more comments which our model uses to refine the results.

setting (*i.e.*, "retrieving" the correct class description prompt), the prediction for an ambiguous video can be steered towards the correct class using surrounding text from a webpage or user hints.

Overall this paper has three main contributions: 1) We quantify the value of the *comments* modality for video-text learning. 2) For this, we propose, train, and evaluate on a new dataset VTC of videos, titles, and comments. And 3) we introduce a new hierarchical attention method that learns how to identify relevant auxiliary information, and that can learn representations that even generalize to other datasets.

2 Related Work

In this work, we focus on multi-modal learning with a particular focus on learning video-text encoders for retrieval by proposing a novel, multi-modal adaptation module.

Video-text Pretraining. Originating from the NLP domain, where the transformer architectures has been a key ingredient and subject to optimization in a multitude of ways [58,14,46,47,30,11,29,50,27,28,33,15], it has recently found applications in the vision-language domain. For example, recent works have leveraged transformers to learn generalizeable image [12,51,9], multi-modal image-text [34,36,56,53,31,10] or video-multilingual text [21] representations. A few works [55,54,63,37] combine visual and text modalities as inputs to a BERT model to simultaneously learn semantic video and text representations. For representation learning, the availability of large-scale datasets such as HowTo100M [40] has enabled more effective pretraining of video-text representations for multiple downstream tasks. More recently, [44] show that adding a generative objective to contrastive pretraining can yield gains in video-text downstream tasks. Based on the CLIP model [45], which works well even without finetuning for some retrieval

tasks [38], [4] train video-text CLIP-initialized models by gradually scaling up video training from image training and a custom dataset. While we also start with a CLIP initialization as in [38], the focus of our paper lies in developing a novel method for leveraging user comments, a modality that has previously been overlooked as a valuable source of information in the text-video retrieval literature. We also note that there has been a surge in recent vision-text pretrained models inspired by CLIP [4,42,62,32]. As we show in the experiments section, our method is agnostic to the pretraining method employed and generalizes beyond CLIP. There are many existing video-text datasets [59,61,49,20,25] but these do not include comments.

Multi-modal domain adaptation. While residual adapters for domain adaptation have been explored for uni-modal models such as CNNs, e.g. in [48], there are no works that translate this concept to the multi-modal domain, where cross-modal learning dominates [1,2,41].

Vision-text Pretraining. While there is a wealth of image-text datasets that provide images with captions, such as OpenImages [26], ConceptualCaptions [52], or COCO [35], the recent state of the art methods train on large-scale weakly-supervised datasets that are obtained from image descriptions from for example Reddit (RedCaps [13]) or YFCC [57].

Comment Datasets. To the best of our knowledge, there is only one vision dataset which does include user-comments, the LiveBot dataset [39]. However this dataset, which contains under 3000 videos, is constructed for artificial comment generation and uses the less common "video barrage" (*i.e.* time-synchronous) type of comments. Despite this, we evaluate our method on this dataset and also find performance gains for video-text representation learning when using comments. In the context of learning from comments, there is little prior work. While the work of [17] is somewhat related, as it uses comments and reactions to posts to refine harm predictions on a social media site, we are the first to demonstrate that user comments can be used as a complementary modality when learning video-text representations.

3 VTC Dataset

We collect a dataset "VTC" of videos along with their titles and comment threads from social news site reddit.com, using their provided API. The videos are collected and used in a manner compatible with national regulations on usage of data for research. Unlike most curated video datasets, this data is more representative of the types of videos shared "in the wild", containing a large proportion of videogames, screenshots and memes.

Using a classifier trained on a small amount of labelled data, we estimate that videogame footage makes up 25% of examples, other screenshots, memes and comics make up 24%, live action footage is 49% and artistic styled content (such as drawn animation) is 2%. The average video length is 33s.

From 1 million raw videos collected, we perform deduplication and filtering, ending up with a training set of 461k videos. For the experiment in Table 8 on training without faces we do further filtering to remove faces, finding that about 65% of videos contain a face. To compensate for the decrease in quantity of training data we gather extra non-face-containing videos, ending up with 339k videos. For the evaluation results we use a test set consisting of 5000 videos with at least three comments each.

For each example in the dataset we obtain: The `title` of the post, a high quality `preview image`, which is generated automatically by Reddit, typically 640 pixels wide and corresponding to the middle frame of the video, the `video` itself, downloaded in low quality and resized to have height 320 pixels for storage reasons, and up to 500 randomly selected `comments` per post.

For all the image-based experiments, we use the high quality preview image, whereas for the video experiments in Table 7 we use the extracted video frames. To fairly compare video and image models given the lower video resolution and quality, in Table 7 the "1 frame" case corresponds to the first frame from the video rather than the high quality preview image.

Deduplication We use the GPU implementation of the FAISS similarity search toolkit [22] to efficiently deduplicate the dataset by indexing the video thumbnail embeddings obtained from a ImageNet pretrained ResNet18 [19]. These indices are then used to discard video entries with a high similarity to other posts.

Safety and Privacy Additionally, we remove toxic text content (such as slurs and hate speech) from titles and comments using the detoxify library [18]. Table 1 show the prevalence of content that has been removed this way.

Detoxify label	% titles	% comments
toxicity	2.32	5.62
severe toxicity	0.00	0.00
obscenity	1.23	3.73
identity attack	0.00	0.00
insult	0.82	1.95
threat	0.05	0.07
sexually explicit	0.09	0.22

Table 1: Prevalence of toxic text before filtering. We report the proportion of posts, titles, and comments that are flagged as having potentially offensive content by the open-source library Detoxify. We use a threshold of 0.9

It is crucial that a dataset is well-conceived and potential risks are thought-out before release. We take two steps to ensure safety and usefulness of our proposed dataset. First, for the releasing the dataset we further filter the dataset to exclude videos that contain faces using the automatic face-detection filtering process from PASS [3]. In our experiments we show that this does not lead to a significant change in performance. Second, we provide a *Datasheet* [16] for the proposed dataset which can be found in the supplementary material. This dataset will be released for research use together with the paper.

4 Methods

In this section we will first recap the mechanism behind current contrastive, multi-modal representation learning methods that rely on clean data. We will then introduce our Context Adapter Module that allows learning from the auxiliary modality through an attention mechanism. Finally, we will describe how we can extend an existing backbone for images to videos and audio, to be able to leverage large, pretrained models.

4.1 Background

In multi-modal representation learning we are given a dataset $\mathcal{X}$ of N samples $x_i \in \mathcal{X}, i \in \{1, \dots N\}$ that individually consists of different signals. Most previous work focuses on two modalities and we will—for now—also adhere to this standard to simplify the notation. This means that each input sample $x_i = (v_i, t_i)$ is a pair of—in our case—a visual input $v_i \in \mathcal{V}$ and its associated text, often the title, $t_i \in \mathcal{T}, 1 \leq i \leq N$.

The goal is now to learn mappings $f_v : \mathcal{V} \mapsto \mathcal{Y}, f_v(v_i) = \phi_{v,i}$ and $f_t : \mathcal{T} \mapsto \mathcal{Y}, f_t(t_i) = \phi_{t,i}$ from each of the modalities to a d-dimensional, joint embedding space $\mathcal{Y} = \mathbb{R}^d$. Recent methods, such as [45], learn the mapping (in their case from images and their captions) to the embedding space with a double contrastive loss over a mini-batch $\mathcal{B} \subset \mathcal{X}$ using an affinity matrix A computed between all pairs of samples in the batch:

$$A_{ij} = \left\langle \frac{\phi_{v,i}}{\sqrt{\tau}\|\phi_{v,i}\|}, \frac{\phi_{t,j}}{\sqrt{\tau}\|\phi_{t,j}\|} \right\rangle \tag{1}$$

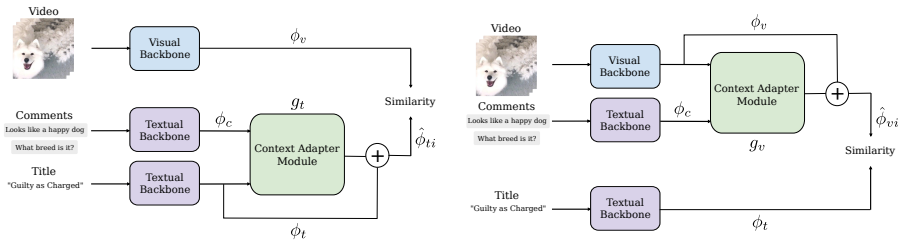
An entry A_{ij} measures the similarity between the embeddings $\phi_v(v_i)$ and $\phi_t(t_i)$ via cosine similarity that is scaled by a temperature parameter τ. The idea is now to maximize the similarity between the embeddings from the *same* sample, *i.e.* the diagonal of A and minimize all non-diagonal entries. This can be achieved efficiently using a double-contrastive formulation that operates across columns and rows of A,

$$\mathcal{L}(A) = \frac{1}{2} \sum_{i=1}^{|\mathcal{B}|} \frac{A_{ii}}{\log \sum_{j=1}^{|\mathcal{B}|} \exp A_{ij}} + \frac{A_{ii}}{\log \sum_{j=1}^{|\mathcal{B}|} \exp A_{ji}}. \tag{2}$$

This formulation has the neat effect that it accomplishes maximizing the diagonal entries and minimizing all other entries of A in one self-balancing formulation. However, it makes the critical assumption that both modalities are equally informative of each other. In the case of sometimes irrelevant data, or when one modality has much less information content than the other (*e.g.* "*nice video!*"), this assumption does not hold and training with this objective will result in a very volatile learning objective and thus a sub-optimal joint embedding.

In the next section we will introduce our Context Adapter Module that is able to deal with this type of inputs by allowing it to discount information when it is not relevant for the context.

4.2 Context Adapter Module



(a) Adapting the text branch.

(b) Adapting the visual branch.

Fig. 2: Method Overview. We introduce a context adapter module that uses inputs of the auxiliary modality to adapt the embedding of another branch. With this module the model is able to accept or discount information.

In order to capture and filter the relevant information from the comments, we propose a transformer-based Context Adapter Module (CAM) which operates in a residual fashion, additively adapting either the visual or text branch of CLIP with contextual information obtained from the comments (see Figure 2). Formally, we are now adding another modality—the comments—to the input which extends it to $x_i = (v_i, t_i, c_{i,1}, \ldots, c_{i,M})$ with $c_{i,k} \in \mathcal{T}$. To reduce clutter in the notation, we have defined a fixed number of comments M for each sample. Since both title and comments share the same modality (*i.e.* text), we can leverage the same encoder to transform comments to embeddings $f_t(c_{ik}) = \phi_{c,ik}$.

As we expect the comments to be sometimes unrelated, our Context Adapter Module needs a mechanism to discount off-topic comments and update the primary modality $\phi_v(v_i)$ or $\phi_t(t_i)$, steering it in the most informative direction.

We introduce this mechanism as a function of both the primary modality and the comment embeddings $\phi_{c,ik}$, as we want to compare the informativeness of all these inputs at a high level. To this end, we design adapter modules g_v and g_t that extract information from the comments in the form of a residual:

$$\hat{\phi}_{oi} = \phi_{oi} + g_o(\phi_{oi},\, \phi_{c,i,1},\, \ldots,\, \phi_{c,i,M})\,,\; o \in \{v, t\} \tag{3}$$

With the adapted embeddings $\hat{\phi}_{vi}$ and $\hat{\phi}_{ti}$ we recompute the affinity matrix (now $\hat{A}$) (Eq. 1) and use it for the loss $\mathcal{L}(\hat{A})$. This design has several advantages. On one hand, extracting "only" a residual from the auxiliary inputs c_{ik} means that the model is easily able to ignore them by predicting $g(\cdot) = 0$. On the other hand, this effectively allows us to skip the adapter module when we evaluate without comments, while still learning the joint embedding from richer data.

In practice, we implement g as a small transformer architecture. Rather than operating on tokenised words, this transformer operates on embeddings (ϕ_{vi} and ϕ_{ti}) themselves, taking as input the encoded feature from the branch to be adapted, along with comment features $\phi_{c,ik}$. By treating embeddings as tokens

in their own right, we allow the embeddings to attend to each other and learn what combinations of the inputs should be used to update the original feature through the residual connection.

Additionally, to avoid bleeding information between the two modalities through the Context Adapter Module, during training, we only adapt either the video embedding with g_v or the text embedding with g_t. If we would use both adapters simultaneously, there is a trivial solution that minimizes the loss $\mathcal{L}$: when the adapters learn to remove the original embedding through the residual, *e.g.* $g_o(\phi_{oi}, \{\phi_{c,i,k},\}) = -\phi_{oi} + \phi_{c,i,1}$ both adapted embeddings become the same $\hat{\phi}_{vi} = \hat{\phi}_{ti}$ which trivially maximizes their similarity, thus preventing the model to learn a meaningful modality alignment. To prevent the model from learning a transformation of the embedding space through the residual, we train only one adapter at a time. We also randomly skip the adapter entirely with probability 0.5, which ensures that the un-adapted features are still meaningful in isolation, and the adapter can be bypassed at evaluation time if comments are not available.

4.3 Video

To leverage the capacity of large pre-trained computer vision models, we adopt the architecture by [45] as our backbone models f_v and f_t. While this transformer was trained on a huge volume of image-text data, it cannot be applied directly to videos since it is built for images and has no temporal extent. To take advantage of the temporal information present in video data, we use the Divided Space-Time attention mechanism recently introduced in the TimeSformer architecture [5]. We modify the image transformer architecture by adding patchwise self-attention across 8 frames in time to each of the 12 residual attention blocks, followed in each case by a zero-initialised linear layer. We also add a learned temporal position embedding which is summed to the input and again zero-initialised. The initialisation is transparent, such that when loading pretrained weights trained from images, at initialization time, the modifications do not affect the inference of the model. During training, the model can then gradually activate the additional temporal components to learn from the temporal information of a video. Full details on the architecture are provided in the Appendix.

4.4 Audio

To further compare the effect of the newly proposed comments modality with another common modality besides text, we also conduct experiments using audio. For this we utilize the audio-encoder from GDT [43] that was pretrained on a large video-audio dataset. The audio-encoder works on 2s audio segments converted into a spectrogram, please see the Appendix for further details.

5 Experiments

This section has two main objectives. The first is to show how the additional modality of user comments can be used to improve multi-modal representation

learning. Second, the experiments show how our new dataset VTC can be used to learn video, audio, image and text representations.

Implementation details. We use CLIP [45] (ViT-B/32 checkpoint unless otherwise mentioned) as the initialisation for the backbone. Our concrete implementation of the CAM g is a 2-layer transformer, consisting of two residual multihead self-attention blocks. The input consists of $M+1$ input embeddings (for the M comments and title/video embedding ϕ_{oi}) having 512 dimensions each. Each block performs 8-head self-attention on the inputs, followed by two linear layers with output size 2048 and 512 respectively. LayerNorm normalisation is used, along with GELU activation following the first linear layer. From the $M+1$ outputs of the transformer, we then normalize, take the mean and renormalize. We use the Adam [24] optimizer with a learning rate of 1×10^{-6} when training the entire model on its own or with the adapter. All implementation and architecture details can be found in the Appendix.

We report the standard Recall@N metrics as a percentage (often abbreviated R@N), giving the proportion of results where the ground truth is ranked in the top N. We show both Text-Video-Retrieval (TVR) and Video-Text-Retrieval (TVR). Unless otherwise mentioned we use 5 comments for evaluation.

5.1 Additional Datasets

LiveBot Dataset. Prior work on building a dataset with videos and comments is LiveBot [39], which consists of 2361 videos and 895,929 comments, obtained from Chinese social network Bilibili. This differs a lot from our setting, since the comments in question are made while the video is being streamed live and associated with certain timecodes, and comments and titles are in Chinese rather than English. Nevertheless, in order to evaluate how well our method works for this sort of data, we use automatic translation to translate the titles and five comments for the 100 videos in the LiveBot test set, which we call LiveBotEN and show in Table 7. Due to duplicate video and missing split metadata in the original LiveBot release, we follow the split used in [60].

KineticsComments. As an additional video dataset with comments, we construct a dataset based on Kinetics-700 [23,6], for which we download the videos along with associated YouTube metadata including title, description and comments. We translate non-English titles and descriptions into English using a commercial translation API. We use the title as the primary text modality, and for auxiliary context we use comments. We construct a test set, consisting of videos from the Kinetics test set for which we have at least 3 comments, giving a set of 6292 videos which we use to evaluate our method in Table 7.

5.2 Evaluating the Context Adapter Module

In this section we evaluate our Context Adapter Module on the above described datasets with comments.

Table 2: Adaption Mechanisms. Comparing different ways in incorporate auxiliary information: adapting the title with 5 comments

Method	TVR R@1	TVR R@10	VTR R@1	VTR R@10
no comments (zero-shot)	11.1	26.0	11.1	25.3
no comments (fine-tuned)	15.5	34.9	14.4	33.4
averaging (zero-shot)	7.3	22.7	6.9	20.0
averaging (fine-tuned)	16.6	42.3	18.1	43.3
ours	**18.4**	**43.2**	**18.6**	**44.0**

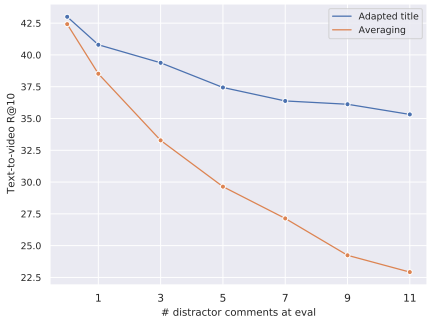


Fig. 3: Influence of Distractor Comments. We gradually add irrelevant distractor comments during evaluation. The context adapter module is able to deal with irrelevant information much better than baseline, showing that it has learned to down-weigh uninformative content

Context Adapter Module. To verify that the Context Adapter Module is indeed able to learn better representations from the comment modality, we compare it to several baselines in Table 2. The most trivial baseline is to ignore any comments and to train simply with image-title pairs. This results in the lowest performance, showing that there is valuable information in the comment data. Another baseline consists of averaging the features from the titles with the features of the comments, which is a direct way to incorporate the comments. We make these baselines stronger by fine-tuning the backbone during training which does result in a performance improvement. Finally, a baseline where all text is concatenated would be interesting to evaluate, however due to memory/text-length limitations concatenating more than 2-3 comments is intractable with current encoder architectures.

Finally, our context adapter module is able to improve over all baselines. We hypothesize that this comes from the ability of the adapter module to ignore irrelevant comments. To test this, we perform an experiment where we add random irrelevant distractor comments (during evaluation only) and measure the impact of distractors on the performance.

The results of this experiment can be seen in Figure 3, where we gradually increase the number of distractors and evaluate retrieval performance. The averaging baseline is strongly affected by this "misinformation" whereas the context adapter module has implicitly learned to ignore irrelevant information during training. Note that there is no explicit supervision for this during training and the model has to learn this behavior directly form the data. As the backbone

Table 3: Backend Fine-tuning. Effect of fine-tuning the encoders

Method	TVR R@1	TVR R@10	VTR R@1	VTR R@10
no fine-tuning	11.1	26.0	11.1	25.3
fine-tuning	15.5	34.9	14.4	33.4

Table 4: Encoder Backbones. Comparing different pre-trained encoders. We keep the encoders frozen and just train the CAM. Showing Recall @ 10, retrieving image from text+comments

Backbone	No Comments	5 comments	20 comments
FiT [4]	8.8	12.0	12.8
SLIP (ViT-B) [42]	9.3	10.2	11.6
CLIP (ResNet50) [45]	22.7	27.4	27.9
CLIP (ViT-B/32) [45]	25.3	32.3	34.1
CLIP (ViT-L/14) [45]	32.9	42.0	44.1

is trained also for the averaging baseline, both methods can learn to ignore generally uninformative content ("look what I found") but the context adapter module can learn to exploit the context of the title with relation to the comments through the attention mechanism.

Comparing Encoders. As in all current multi-modal approaches, the architecture and pre-training of the visual/text/audio encoder is important. In Table 3 we show that fine-tuning the (in this case CLIP [45]) encoder does improve the performance by a significant margin. This shows that even though the encoder was trained on an extremely large image/text dataset, there is a domain gap with VTC (videos and comments) that can be bridged by fine-tuning.

In Table 4 we compare different model types of CLIP [45] with other current models: SLIP [42] and FiT [4]. Naturally, larger architectures perform better, in line with ResNet50 falling behind ViT based encoders for CLIP. Comparing to CLIP, FiT and SLIP have been trained on roughly two orders of magnitude smaller datasets (400M image-text pairs for CLIP) resulting in decreased image and text understand capabilities.

Additionally, we find that adding comments improves the performance of *all* encoders. Adding more comments consistently improves the performance again, however with diminishing returns. We further investigate this behavior in Figure 4, where we vary the number of comments during training and evaluation time. All models benefit from using comments compared to not using comments. Interestingly, training with one comment seems to be insufficient to learn how to extract additional information when there is more than one available.

Different Modalities. The intuition behind the context adapter module is that it allows to adapt information in a feature with potentially unreliable auxiliary data. As described earlier, the comments can be used to either adapt the information in the image or in the title. In Table 5 we compare these two options and

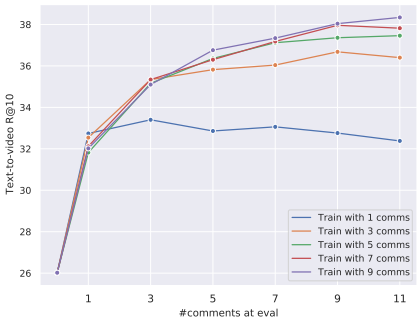


Fig. 4: Varying Number of Comments. We show the influence of varying the number of comments during training and testing time. All variants benefit from using comments. Training with a single comment is not enough to learn a stable filtering behavior

Table 5: Adaption Modality. Comparing different ways to incorporate auxiliary information: adapting the title or image with comments

Method	TVR R@1	TVR R@10	VTR R@1	VTR R@10
none	15.5	34.9	14.4	33.4
title	18.0	43.2	18.7	43.9
image	28.2	51.2	25.1	49.9

find that adapting the image results in a larger performance improvement than adapting the title. This can be explained by the modality gap between visual and textual information. When adapting the image information with the text from the comments, the context adapter module can learn to close the information gap between text and image much more effectively than when adapting the tile with the text from the comments. However, we find that in the context of retrieval and multi-modal representation learning a more realistic (and challenging) scenario is posed when the title is adapted (as for example seen in Figure 1).

Table 6: Combining Modalities. We show that our method is robust to different combinations of modalities, both at train and at test time

		Text → Video		Video → Text	
training	**inference**	**R@1**	**R@10**	**R@1**	**R@10**
CLIP	img+title	11.1	26.0	11.1	25.3
img+title	img+title	15.5	34.9	14.4	33.4
img+title+cmts	img+title	15.5	34.5	14.4	33.3
img+title+cmts	img+title+cmts	18.0	43.2	18.7	43.9
img+title+cmts+audio	img+title	15.4	34.0	14.3	32.9
img+title+cmts+audio	img+title+audio	15.8	36.9	12.2	30.4
img+title+cmts+audio	img+title+cmts+audio	19.6	45.6	20.6	47.2

Another benefit of the context adapter module is that, not only can it deal with a variable number of comments during inference, but it also allows for

evaluation without any comments. Table 6 shows that training with comments does not have an impact on the performance of the model in a setting where no comments are available at test time. This means that the learned model is flexible and can be used in both settings directly and without any changes.

The idea of learning from potentially unreliable auxiliary data extends beyond the use of comments and in Table 6 we perform additional experiments using the audio information in the videos. In many current video datasets the quality of the audio varies drastically. For example, some videos replace the natural audio with music, removing any aural clues about the content of the video. Similar to comments, including audio in the context adapter module during training allows the model to identify irrelevant audio information. This results in a virtually unchanged performance when no audio is available during test time, but further improves the final performance when considering all four modalities.

Video Data. In this section we evaluate the impact of using videos instead of single frames in combination with also adding comments. For video evaluation we take the 8 initial frames with a stride of 16. In Table 7 we find that on all datasets adding comments boosts the retrieval performance for both video-to-text (VTR) and text-to-video (TVR) significantly, confirming the value of the modality. It is important to note, that all models were trained only on VTC and the improvements translate directly to KineticsComments and LiveBotEN. In most cases, the improvement gained from adding comments is considerably larger than the information gained by incorporating temporal information. This is an additional data point for the importance of the comment modality. While VTC test set does not benefit largely from video training itself, using videos during training still improves the performance on the other datasets.

Table 7: Video results. Experiments using video frames. Trained adapting the video branch with comments, with either one or eight frames from the video. Showing Recall@10

		VTC		**KineticsComms**		**LiveBotEN**	
inference	**#frames**	**VTR**	**TVR**	**VTR**	**TVR**	**VTR**	**TVR**
video	1	28.9	28.3	48.8	46.9	48.0	49.0
video+comments	1	40.8	41.0	61.1	59.2	64.0	64.0
mean-pooling	8	19.3	24.2	54.1	49.8	69.0	66.0
video	8	28.9	27.6	56.9	55.8	70.0	72.0
video+comments	8	41.5	41.9	68.0	66.1	69.0	80.0

Privacy. Finally, we perform an experiment on the effect of removing all videos from the dataset that contain a face. Table 8 shows that even though this reduces the size of the training set, the performance is not negatively affected. The evaluation is performed on the same test set. We can even see a small increase in

Table 8: Privacy – Removing Faces. Effect of removing all videos that contain a face from the dataset. The evaluation is performed on the same test set (that contains faces). The difference in performance is marginal

Method	TVR R@1	TVR R@10	VTR R@1	VTR R@10
with faces	18.0	43.2	18.7	43.9
without faces	18.2	44.0	18.1	45.0

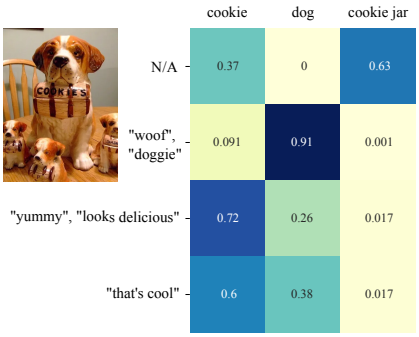


Fig. 5: Failure Case. A heatmap showing the similarities between the image adapted with different comments (rows), and captions (columns). The adapter can steer away the embedding from the right association "cookie jar" depending on the comment input. This means that adversarial comments could affect the performance of the model

performance, that could potentially be attributed to a more balanced training set, as videos of humans tend to dominate the dataset before the face removal.

6 Discussion and Conclusions

Limitations. We find that the context adapter can be led to override the information in a title if we adversarially craft comments that all point to different content. Qualitative examples of this can be seen in Figure 5. The model without comments, correctly associates the image with a cookie (jar), however when adding a comment about a "dog" the model prefers the dog label over cookie.

Conclusion. We have presented VTC, a new dataset with videos, titles, and comments and a context adapter module, which is able to extract information from auxiliary input sources for learning a joint, multi-modal embedding. The dataset fills a gap in current vision-text datasets as it includes comments that potentially provide additional information about the content. In our experiments, we are able to show that learning from comments improves video-text retrieval when adapting the representation with user comments. Moreover, the context adapter module is able to identify whether an auxiliary input is relevant to the content in the other modalities or not. This mechanism could, for example, be used to filter datasets for meaningful auxiliary content.

Acknowledgements. This project is supported by Innovate UK (project 71653) on behalf of UK Research and Innovation (UKRI). Y.M.A. and C.R. were also supported by an AWS Machine Learning Research Award (MLRA). We also thank Sasha Haco from Unitary for her support for this project.

References

1. Alwassel, H., Korbar, B., Mahajan, D., Torresani, L., Ghanem, B., Tran, D.: Self-supervised learning by cross-modal audio-video clustering. In: NeurIPS (2020)
2. Asano, Y.M., Patrick, M., Rupprecht, C., Vedaldi, A.: Labelling unlabelled videos from scratch with multi-modal self-supervision. In: NeurIPS (2020)
3. Asano, Y.M., Rupprecht, C., Zisserman, A., Vedaldi, A.: Pass: An imagenet replacement for self-supervised pretraining without humans. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (2021)
4. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. arXiv preprint arXiv:2104.00650 (2021)
5. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? arXiv preprint arXiv:2102.05095 (2021)
6. Carreira, J., Noland, E., Hillier, C., Zisserman, A.: A short note on the kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987 (2019)
7. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
8. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 721–725. IEEE (2020)
9. Chen, M., Radford, A., Child, R., Wu, J., Jun, H.: Generative pretraining from pixels. In: ICML (2020)
10. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Learning universal image-text representations. arXiv preprint arXiv:1909.11740 (2019)
11. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: ELECTRA: Pre-training text encoders as discriminators rather than generators. In: ICLR (2020)
12. Desai, K., Johnson, J.: Virtex: Learning visual representations from textual annotations. arXiv preprint arXiv:2006.06666 (2020)
13. Desai, K., Kaul, G., Aysola, Z., Johnson, J.: Redcaps: Web-curated image-text data created by the people, for the people. arXiv preprint arXiv:2111.11431 (2021)
14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: ACL (2019)
15. Fang, H., Xiong, P., Xu, L., Chen, Y.: Clip2video: Mastering video-text retrieval via image clip. arXiv preprint arXiv:2106.11097 (2021)
16. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Iii, H.D., Crawford, K.: Datasheets for datasets. Communications of the ACM **64**(12), 86–92 (2021)
17. Halevy, A., Ferrer, C.C., Ma, H., Ozertem, U., Pantel, P., Saeidi, M., Silvestri, F., Stoyanov, V.: Preserving integrity in online social networks. arXiv preprint arXiv:2009.10311 (2020)
18. Hanu, L., Unitary team: Detoxify. Github. https://github.com/unitaryai/detoxify (2020)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
20. Hendricks, L.A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with temporal language. In: Empirical Methods in Natural Language Processing (EMNLP) (2018)

21. Huang, P.Y., Patrick, M., Hu, J., Neubig, G., Metze, F., Hauptmann, A.: Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In: Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL) (June 2021)
22. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus. arXiv preprint arXiv:1702.08734 (2017)
23. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
25. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: CVPR (2017)
26. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4. International Journal of Computer Vision **128**(7), 1956–1981 (2020)
27. Lei, C., Luo, S., Liu, Y., He, W., Wang, J., Wang, G., Tang, H., Miao, C., Li, H.: Understanding chinese video and language via contrastive multimodal pre-training. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2567–2576 (2021)
28. 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. CVPR (2021)
29. Lewis, M., Ghazvininejad, M., Ghosh, G., Aghajanyan, A., Wang, S., Zettlemoyer, L.: Pre-training via paraphrasing. arXiv preprint arXiv:2006.15020 (2020)
30. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: ACL (2020)
31. Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D., Zhou, M.: Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In: AAAI (2020)
32. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation (2022)
33. Li, L., Chen, Y.C., Cheng, Y., Gan, Z., Yu, L., Liu, J.: Hero: Hierarchical encoder for video+ language omni-representation pre-training. EMNLP (2020)
34. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
35. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
36. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: NeurIps (2019)
37. Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., Zhou, M.: UniVL: A unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353 (2020)
38. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval. arXiv preprint arXiv:2104.08860 (2021)
39. Ma, S., Cui, L., Dai, D., Wei, F., Sun, X.: Livebot: Generating live video comments based on visual and textual contexts. In: AAAI 2019 (2019)
40. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: ICCV (2019)

41. Morgado, P., Vasconcelos, N., Misra, I.: Audio-visual instance discrimination with cross-modal agreement. arXiv preprint arXiv:2004.12943 (2020)
42. Mu, N., Kirillov, A., Wagner, D., Xie, S.: Slip: Self-supervision meets language-image pre-training. arXiv preprint arXiv:2112.12750 (2021)
43. Patrick, M., Asano, Y.M., Kuznetsova, P., Fong, R., Henriques, J.F., Zweig, G., Vedaldi, A.: Multi-modal self-supervision from generalized data transformations (2021)
44. Patrick, M., Huang, P.Y., Asano, Y., Metze, F., Hauptmann, A., Henriques, J., Vedaldi, A.: Support-set bottlenecks for video-text representation learning. arXiv preprint arXiv:2010.02824 (2020)
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
46. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (2019)
47. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683 (2019)
48. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Learning multiple visual domains with residual adapters. In: NeurIPS (2017)
49. Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., Schiele, B.: Movie description. International Journal of Computer Vision (2017)
50. Ruan, L., Jin, Q.: Survey: Transformer based video-language pre-training. AI Open (2022)
51. Sariyildiz, M.B., Perez, J., Larlus, D.: Learning visual representations with caption annotations. In: ECCV. pp. 153–170. Springer (2020)
52. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565 (2018)
53. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: VL-BERT: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530 (2019)
54. Sun, C., Baradel, F., Murphy, K., Schmid, C.: Learning video representations using contrastive bidirectional transformer. arXiv preprint arXiv:1906.05743 (2019)
55. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: ICCV (2019)
56. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. In: EMNLP (2019)
57. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)
58. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
59. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence – video to text. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
60. Wu, H., Jones, G.J., Pitie, F.: Response to livebot: Generating live video comments based on visual and textual contexts. arXiv preprint arXiv:2006.03022 (2020)
61. Xu, J., Mei, T., Yao, T., Rui, Y.: MSR-VTT: A large video description dataset for bridging video and language. In: CVPR (2016)

62. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: FILIP: Fine-grained interactive language-image pre-training. In: International Conference on Learning Representations (2022)
63. Zhu, L., Yang, Y.: Actbert: Learning global-local video-text representations. In: CVPR (2020)