

Unaligned Video-Text Pre-training using Iterative Alignment

Arka Sadhu^{1*} Licheng Yu² Animesh Sinha² Yu Chen² Ram Nevatia¹ Ning Zhang²
¹University of Southern California ²Meta AI

{asadhu,nevatia}@usc.edu {lichengyu,animeshsinha,hugochen,ningzhang}@meta.com

Abstract

Existing state-of-art vision-language models follow the widely-used recipe of pre-training on a large corpus of image-text pairs followed by fine-tuning on one or more downstream tasks. Similar methods have also been shown to be successful in video-language tasks. However, such pre-training schemes are inherently restricted by the availability of large-volume of high-quality paired video captions, often only found in particular video domains such as stock footage or instructional videos. To address this limitation, we explore utilizing unaligned vision and text corpora with two distinct advantages: (i) access to orders of magnitude more unaligned data (ii) such unaligned data can be obtained for diverse domains. We show that our proposed iterative alignment method to perform alignment between vision and language modalities in the pre-training step can significantly improve downstream task performance compared to no pre-training setup. Experiments on multiple diverse video-language benchmarks validate the effectiveness of our approach.

1. Introduction

Vision-Language Pre-training (VLP) has become the de facto method for tackling most vision-language tasks. In this paradigm, vision-language models typically based on transformers [47] are pre-trained on large-scale image-text corpus [18, 26, 33, 40] and then fine-tuned on downstream tasks such as QA [1], retrieval [53], or captioning [26].

Recent works, such as CLIP [34], use a lightweight text-encoder to scale VLP in the image-text domain to very large corpora in the order of hundreds of millions of image-text pairs. Visual features learned by such models have proven to be very powerful [41]. They can be used as an initialization for the vision-backbone for VLP on an image-text corpus with a powerful text encoder [10, 24].

Despite the abundance of video data, extensions to the video-text domain [50] have yet to show similar improve-

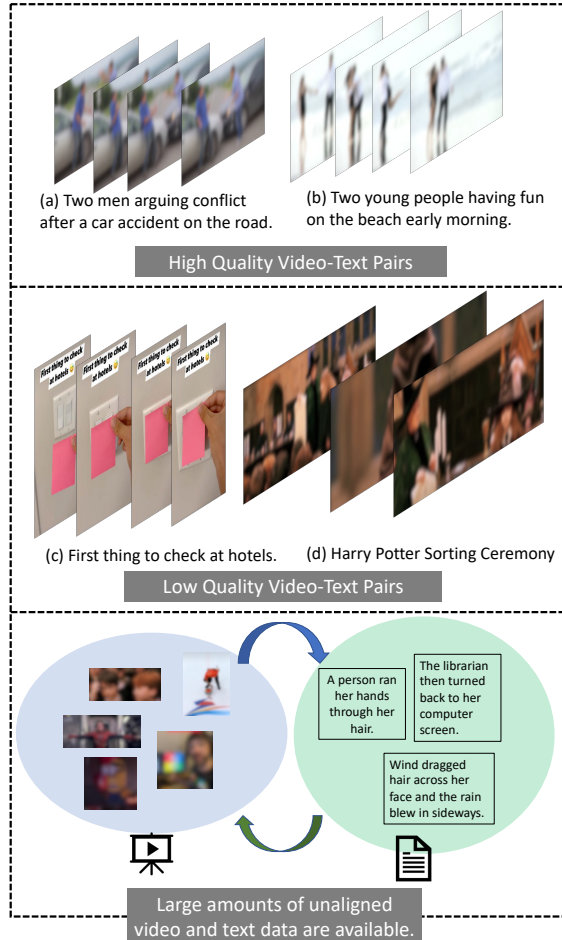


Figure 1. (a,b): Video-Text Pre-Training requires high quality video-text pairs such as those obtained from stock footage. (c,d): Video-text pairs from other domains, such as short-form user generated videos or movie-clips, are often uncorrelated and cannot be used for pre-training. We propose the task of Unaligned Video-Text Pre-Training (UVTP), where we leverage existing unimodal vision and text corpora to pre-train vision-language models.

ments in video-text tasks. We posit that a key reason for this discrepancy is that highly correlated paired video-text data for videos is available in limited domains such as

*Arka was an intern at Meta AI.

stock-footage [4] or instructional videos [32]. To bridge this gap, previous work [11, 30] initialize the video backbone using weights from image-text pre-trained networks such as CLIP. Such initialization leads to significant improvements in downstream tasks, especially retrieval based tasks. However, further investigation [21] attributes these improvements to purely better image understanding model rather than any progress in video understanding.

In this work, we investigate a different approach; we observe that while paired video-text is limited to specific domains, unaligned videos and text data are available in vast amounts in both modalities. Successfully exploiting such unaligned data has two-fold advantages: first, orders of magnitude more unimodal data can be obtained from the web compared to paired data, and second such unaligned data is not restricted to any particular domain. Inspired by previous works in unsupervised machine translations [19, 20] and unaligned image-text pre-training [25, 57], we propose the task of Unaligned Video-Text Pre-Training (UVTP) *i.e.* pre-training on unaligned video-text data. We illustrate this fundamental idea in Figure 1.

There are two primary considerations for pre-training on unaligned video-text data. First, the method should be scalable. As a result, naively extending previous unaligned image-text pre-training methods like U-VisualBERT [25] and UVLP [57], which use object detection features [39, 56], is not an option. This is because computing object features for each frame is bulky, making the storage of such features and the data-loading process cumbersome.

Second, we need an initial alignment between the video and text modality. Unlike unsupervised machine-translation tasks [19, 20], where we have access to human-curated dictionary mapping between words and phrases for various languages, we don't have any equivalent mapping for videos and text. This problem is exacerbated by the fact that mapping between the two modalities is not one-to-one but many-to-one mapping. Thus, without any initialization of the alignment, learning from unaligned vision text corpus becomes an ill-posed problem. For instance, naively pre-training on random pairs of image and text will not lead to any meaningful improvements in downstream tasks.

To address the first issue, we focus on methods to learn directly from raw video frames. In particular, we extend existing image-text transformer, such as METER [10], to handle video frames by duplicating the vision-backbone. To tackle the second issue, we consider two cases: (i) a fully unaligned setting where we don't have access to any paired vision-data and (ii) a semi-supervised setting where we have access to limited paired vision-text data. For the first case, we exploit co-occurrence heuristics - if the same objects appear in a video and in a sentence, the two are likely to be correlated. Similar strategy is used in previous unaligned image-text pre-training, such as UVLP [57]. For the second

case, we utilize the existing paired data to pre-train a model and then use it to retrieve the best matching text for a given video. We detail our model design in Section 3.1 and the two settings in Section 3.2.

While a given alignment between the two modalities would allow pre-training, the efficacy of the pre-trained model is constrained by the initial alignment. We motivate this by showing that pre-training on a more aligned video-text corpus leads to improved downstream performance. To mitigate this issue, we propose to utilize the already pre-trained model to re-align the two modalities after a fixed number of training iterations. The updated pairs from this new alignment is used in the pre-training process till convergence. We describe our proposed method, *Iterative Alignment*, in Section 3.3.

To study pre-training on unaligned video-text corpora, we follow previous work [25, 57] by first creating a shuffled dataset from existing image-text (CC3M [40], SBU [33], COCO [26], VG [18]) and video-text (WebVid-2M [4]) corpora. In other words, we artificially remove the correspondence between the text and image/video instances. Such a simulated testbed allows performing fine-grained ablative studies and guarantees existence of close text match for every image/video. We then extend this to a realistic setting by considering keeping the same images/videos but using text from a completely different corpora such as BookCorpus [59].

Experiments on multiple downstream tasks such as video-QA [49], video retrieval [2, 51], action localization [45] and segmentation [60] show the benefits of using additional unaligned video text corpora. We further find *Iterative Alignment* improves downstream performance compared to having a fixed alignment.

Our main contributions are (i) a systematic study of unaligned pre-training on video-text data (ii) a method to iteratively refine alignment to drive better downstream performance (iii) detailed ablative study and benchmarking to guide future work.

2. Related Works

Image-Language Pre-Training is a heavily studied topic. Inspired by the success of masked-language-modeling (MLM) for large-scale language pre-training [8, 27, 35], earlier works such as LXMERT [44], ViLBERT [29], VL-BERT [43], UNITER [6] extended the success of BERT to image-language domain using pre-extracted object features. These works utilized existing paired image-text datasets such as COCO [26], VG [18], CC3M [40], SBU [33]. More recent works such as ViLT [17], ALBEF [24], METER [10] opt for vision-language transformers which can directly learn in an end-to-end fashion. These models are supported by vision-transformers [5, 9, 46] for their visual backbone. In this work, we follow recent trend and

opt for end-to-end learning. This is particularly important in video-language domain where saving and loading bulky object features can become a bottleneck.

Video-Language Pre-Training has been supported by the availability of large-scale video-text corpora such as instructional videos from HowTo100M [32] and stock footage from WebVid-2M [4]. While other large video-text corpora such as those based on movie clips [3, 14] exists, the correlation between the visual scene and the paired text is often poor making them unsuitable for large-scale pre-training. Prior work can be broadly classified based on its visual backbone: space-time encoders [4, 23, 31] or shared image-encoders [21, 22, 54, 55]. Our work follows the second approach. We extend existing image-language framework to incorporate video frames.

Very Large-Scale Image-Language Pre-Training have recently gained immense popularity. Two prominent works include CLIP [34] and ALIGN [16]. These models perform contrastive-learning using a simple dual-encoder image-language transformer model but scale it to hundreds of millions or billions of images. The resulting vision modules can be used to initialize visual backbones in an image-language transformer which leads to dramatic improvement in downstream image-language [41]. However, similar improvements have yet to be observed in video-language domain. The best known example is VideoClip [50] which exhibits zero-shot capabilities but is outperformed by CLIP-based models on multiple retrieval tasks [11, 30, 52].

Unaligned and Semi-Supervised Video-Language Training is unexplored but there are related works in image-language domain. U-VisualBERT [25] first introduced the task of learning from unaligned image-text corpora and used a shared transformer for both masked language modeling and masked image modeling with the object tags being the implicit bridge between the two domains. UVLP [57] expands on this idea and observes that better alignment in pre-training leads to better downstream performance. Our work extends the idea to videos but differs in two key ways. First, both UVLP and U-VisualBERT used object detection features which cannot be extended to videos due to scalability. Second, both consider a fixed alignment case and only depend on heuristics to align the two domains whereas we propose iterative alignment to improve the alignment. We further explore the semi-supervised setting where we are provided a fraction of aligned data but also have access to vast amounts of non-aligned data which is a more realistic setting.

3. Method

We first briefly describe our model design (Section 3.1) to allow video-text pre-training. We then formally detail the task of unaligned video-text pre-training (Section 3.2) followed by description of our proposed *Iterative*

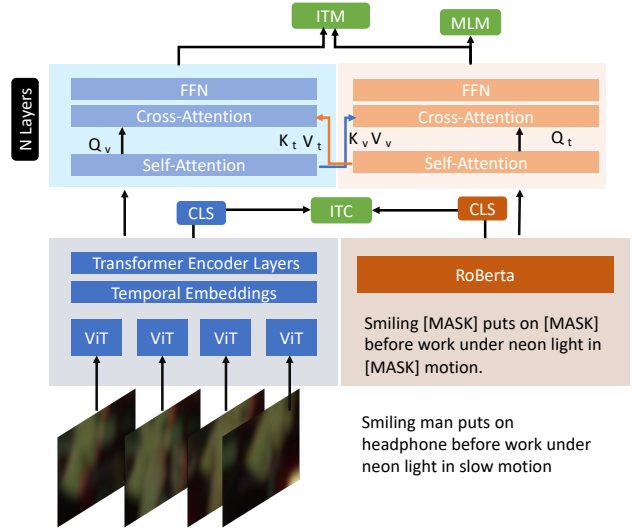


Figure 2. A schematic of our proposed VMET which extends image-text transformer (METER [10]) to handle videos. The vision-backbone duplicates the ViT block (shared weights) and processes the frames separately. Temporal embeddings are added to the ViT outputs followed by 2-layers of transformer [47] encoders. The text backbone is kept as is. Three losses are used in pre-training: masked-language modeling (MLM), image-text matching (ITM) and image-text contrastive loss (ITC).

Alignment to obtain richer video-text pairs.

3.1. Model Design

End-to-end image-language transformers such as ALBEF [24], METER [10] support learning directly from raw image-frames. We propose an extension of such transformers to allow processing both images and videos with small overhead. An alternative way would be to instead use object features from a detection model like VinVL [56] such as in ActBert [58]. We opt for end-to-end learning due to bulkiness of the pre-extracted features which leads to both storage and loading bottlenecks with more frames.

In this work, we focus on METER model but note similar extensions follow for other end-to-end vision-language transformers. For videos, we uniformly sample $k=4$ frames; images are considered single frame videos. We process each frame through the image-backbone (ViT [9]), and then add the temporal position embeddings followed by 2 transformer encoder layers. The text processing part is kept intact. We also add a contrastive head similar to that used in CLIP [34] using the first token ([CLS] token) of vision and text backbones respectively. The model is trained with Masked-Language-Modeling (MLM), Image-Text Matching (ITM) and Image-Text Contrastive Loss (ITC) respectively. We provide additional model implementation details in the supplementary. The resulting model dubbed VMET

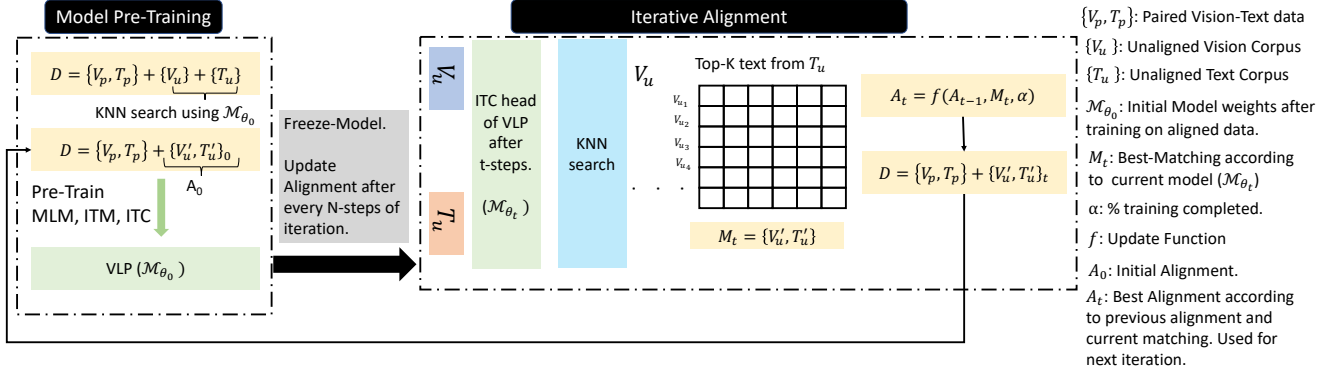


Figure 3. A schematic of our proposed Iterative Alignment. We are provided a VLP model pre-trained on aligned data \mathcal{M}_{θ_0} . The aligned data could be heuristically generated as for fully-unaligned or have been already provided in semi-supervised case. Using \mathcal{M}_{θ_0} we first obtain an initial alignment A_0 . We pre-train the VLP with the given alignment for N steps then use the current model \mathcal{M}_{θ_t} to obtain a new matching M_t . We then obtain the new alignment A_t considering the previous alignment A_{t-1} , the new matching M_t , percentage of training completed α and the update function f . The updated alignment is now used to pre-train the model. The process is repeated every N steps.

is illustrated in Figure 2.

3.2. Task: Unaligned Video Text Pre-Training

The status-quo in VLP is to pre-train a vision-language model on paired vision-text data. However, obtaining highly correlated paired video-text is only possible in certain domains such as stock videos [4] or instructional videos [32]. Given the large amounts of video and text data in their respective domains, we define the task of Unaligned Video-Text Pre-Training in being able to exploit the additional data for pre-training and subsequently improve downstream performance.

Formally, let D denote the entire video and text data available to use. Let $D_p = \{V_p, T_p\}$ denote the subset of paired video-text data and $\{V_u\}$ and $\{T_u\}$ denote the set of unaligned video and text data respectively. An implicit assumption is that set of unaligned video and text corpora is significantly larger than the paired corpus *i.e.* $\min(|V_u|, |T_u|) \gg \min(|V_p|, |T_p|)$. Thus, $D = \{V_p, T_p\} + \{V_u\} + \{T_u\}$. Then, the UVTP task requires pretrained models trained on D to outperform those trained on just D_p . The main criteria would be better performance on set of chosen downstream tasks.

With the above notation, we introduce two particularly important settings. (i) **Fully unaligned** setting *i.e.* $D_p = \phi$ or $D = \{V_u\} + \{T_u\}$ (ii) **Semi-Supervised** setting *i.e.* $D_p \neq \phi$.

Fully unaligned. Here, we don't have access to any paired vision-text data. Learning directly from unimodal data has immense potential for improving downstream tasks on very diverse domains. Unfortunately, the problem is ill-defined by construction as we don't have existing mapping between the two modalities. To circumvent this issue, we utilize an object co-occurrence heuristic *i.e.* similar ob-

jects appearing in a video and text would make them correlated. This is similar to the weak alignment used in previous work [25, 57]. Essentially, we use the labels from object detection system as a bridge between the modalities.

Specifically, we use an object detection system [37, 39] to detect the objects across multiple frames. Given the set of unique objects in the video, we create an object string by simply concatenating all the unique objects in their order of appearance. Using this object string as a query, we retrieve the topK ($K = 1$) matching texts from the text corpus using a sentence-similarity model such as SentenceBert [38]. We can then pre-train VMET on the above correspondence to obtain a retrieval model.

Semi-Supervised. Here, we have access to some aligned video-text data. Different from the "Fully unaligned" case, this is a more realistic case as there are existing paired video-text datasets. Recall that our implicit assumption is the amount of unaligned data is significantly larger than the paired data. Since we have existing paired data, instead of using heuristics we can instead pre-train on the paired data to obtain a retrieval model. This retrieval model can then be used as a bridge between the two modalities. Similar to previous case, we can obtain topK text-matches for every unpaired video instance. We keep the already paired video-text instances as is.

3.3. Iterative Alignment

For both fully unaligned and semi-supervised case, we have retrieval model. For the former, the model is trained on matches obtained via heuristics, while for the latter the model is trained on paired ground-truth data. Denote this retrieval model as \mathcal{M}_{θ_0} . Suppose this alignment obtained using this retrieval model be called $A_0 = \{V'_u, T'_u\}_0$. We

Algorithm 1 Pseudo-Code for Iterative Alignment

Require: Data $D=\{V_p, T_p\} + \{V_u\} + \{T_u\}$ **Require:** Alignment Update Stride N **Require:** Update Function f **Require:** Model \mathcal{M}_θ initialized as \mathcal{M}_{θ_0} ;1: Initial Alignment $A_0 = \{V'_u, T'_u\}_0$;2: % Training completed $\alpha \leftarrow 0$ 3: **while** $\alpha < 1$ **do**4: PreTrain \mathcal{M}_θ with MLM, ITM, ITC for N steps.5: Freeze \mathcal{M}_θ . Process $\{V_u\}$ and $\{T_u\}$ 6: KNN on ITC output $M_t = \{V'_u, T'_u\}$ 7: New Alignment $A_t = \{V'_u, T'_u\}_t = f(A_{t-1}, M_t, \alpha)$ 8: Update $D_t = \{V_p, T_p\} + \{V'_u, T'_u\}_t$ 9: Increment α 10: **end while**

AD	1	0.9	0.75	0.5	0.25	0	NA
Acc	42.01	40.74	41.66	41.10	40.31	34.10	39.35

Table 1. Fine-tune accuracy on MSRVT-QA for VMET pre-trained with different amounts of aligned data in WebVid-2M [4]. AD: Fraction of aligned data. NA: Model is fine-tuned without any pre-training.

use V'_u instead of V_u to denote not all vision instances have match. We note that we can directly pre-train VMET on this initial alignment A_0 . We call this Fixed Alignment since the pairing remains the same throughout the training. However, the initial alignment so obtained could be sub-optimal.

In Table 1 we note the performance of VMET pre-trained on WebVid-2M [4] with different amounts of aligned data. To simulate this, we randomly pair $(1-AD)$ amount of video-text data while keeping the remaining AD amount from the original dataset. As can be observed, if the entire alignment is corrupted ($AD=0$), the model is unable to learn anything and in fact performs worse than when no pre-training was involved. We further note that there is a direct correlation between higher percentage of aligned data and improved downstream task.

Motivated from the above experiment, if we can train VMET with improved alignment A' , it would lead to better downstream performance. Suppose that the model pre-trained on the initial aligned data is “good enough”. This model can itself act as a retrieval system and produce a matching M between the two corpora. We can utilize the initial alignment A_0 along with the new matching M to provide an updated alignment. We can then repeat this process after every N steps till convergence. This leads us to Iterative Alignment for which the pseudo-code is provided in Alg 1. A schematic of the process is provided in Figure 3.

We note two particular parameters used to obtain the new

alignment A_t (Line 7): α the percentage of training completed and f the update function. We condition the update on α as we want to prioritize the existing alignment for initial epochs and the predicted alignment M_t by the model in later epochs. The update function f can be any decay function, in practice we use a linear decay. For linear decay, we have $A_t = (1 - \alpha)A_{t-1} + \alpha M_t$.

In practice, we cannot store all the matches since the number of instances in either modality can be in the order of millions. To approximate it, we compute Top-50 text-matches for each video and assume matching score of 0 for the remaining sentences. We update the scores for the union of text-matches from the previous iteration (A_{t-1}) and current matches (M_t) weighted by percentage of training completed α . Once the scores for the union is computed, we sort them in a descending order and keep only the top-50 matches, and continue the process.

Fixed Alignment. We note that while Iterative Alignment is applicable to both fully unaligned and semi-supervised setting there is a small distinction when it comes to fixed alignment *i.e.* $A_t=A_0$. For the fully-unaligned case, we train our model on the entire unaligned corpus which is paired using the heuristic ($D = \{V'_u, T'_u\}$). Thus, fixed alignment is largely equivalent to training with simply additional epochs with the same paired sets. However, for semi-supervised case our model was trained only on the paired data ($D_p = \{V_p, T_p\}$). Fixed alignment on semi-supervised case would include $D = D_p + \{V'_u, T'_u\}_0$. Thus, fixed alignment is different from additional epochs as the training set itself has expanded and now includes the unaligned data as well.

4. Experiments

We describe the datasets used for our experiments (Section 4.1) followed by experimental setup about the baselines and implementation details (Section 4.2). We then detail our results in Section 4.3.

4.1. Datasets

Pre-Training Datasets. For image-text datasets we consider CC3M [40], SBU [33], COCO [26] and VG [18], and for video-text corpus we use WebVid-2M [4].

Shuffled Dataset. To simulate the unaligned setting we explore training on a shuffled dataset similar to previous work [57]. Specifically, we use all the images, videos and text from the above pre-training dataset but remove the correspondence of the specific vision and text pairs. For the captions, we remove close duplicates which reduces the text corpus from $12M$ to around $9M$ captions. In the semi-supervised case, we consider $X\%$ aligned data setting where X denotes the percentage of paired vision-text data with unique image/video instances. When creating such data, we sample $X\%$ from each of the corresponding

datasets and concatenate them. We focus on 10% aligned setting for the remainder of the paper and provide more $X\%$ cases in supplementary.

BookCorpus Dataset. To simulate a real-world setting, we use the images and videos from existing corpora but use texts obtained from BookCorpus [59]. We pre-process the BookCorpus data to remove sentences with no mentions of objects, and remove duplicate or very similar sentences. The resulting text from book corpus has around $12M$ captions which is of very similar size as that of the combined image+video dataset. A detailed processing pipeline can be found in the supplementary.

Downstream Tasks. For downstream tasks, we primarily consider MSRVTQA [49] for video question answering and MSRVT for video retrieval [51]. For video retrieval we train on the $7K$ videos. We also consider diverse downstream tasks such as DiDeMo [2] for paragraph retrieval, COIN [45] for action segmentation and CrossTask [60] for action localization.

4.2. Experiment Setup

Compared Models. We consider NP-VMET case where the VMET is directly fine-tuned on the target downstream dataset. Here NP denotes No Pre-training. This serves as a lower-bound for almost every dataset.

For the fully unaligned setting, we first use heuristics to align the datasets and pre-train over it. We then introduce two models (i) UA-VMET + FA where the initial alignment is kept fixed (ii) UA-VMET + IA where the alignment is iteratively updated. Here, UA denotes unaligned setting, FA denotes fixed alignment and IA denotes iterative alignment.

For semi-supervised setting, we first pre-trained on the available paired data. We call this model SS-VMET. Following the previous case, we again introduce two models (i) SS-VMET + FA and (ii) SS-VMET + IA denoting the fixed and iterative alignment variants. Here, SS denotes the semi-supervised setting.

For both fully unaligned and semi-supervised setting, we additionally consider a fixed alignment variant where the initial alignment is provided by CLIP [34]. We note that CLIP is already trained on very large corpus defeating the purpose of UVTP, but the provided alignment serves as an upper-bound *i.e.* the best achievable performance with state-of-art retrieval system. To retrieve videos using CLIP, we simply use the second-frame of the video as the candidate (among the four uniformly sampled frames). We refer to these baselines as UA-VMET + CFA and SS-VMET + CFA respectively. For the latter, CLIP is used on only the 90% unpaired data with 10% ground-truth data intact. Here, CFA denotes CLIP with fixed alignment.

Finally, we also provide a fully-supervised model FS-VMET. As such, the overarching goal of UVTP task is to bridge the gap between NP-VMET and FS-VMET.

We also compare with other video-text models such FROZEN [4], ALPRO [23], All-In-One [48].

Metrics. We use standard metrics for downstream datasets such as Acc@1 for MSRVTQA, Recall@X for Retrieval on MSRVT and DiDeMo. For COIN, we use FrameAcc@1 for frame-wise classification and for CrossTask we use Recall for the action segmentation.

We also consider CLIP-Score [12] as an intermediate metric which can be directly computed using the pre-trained model over the unaligned training dataset. To compute this, we use the pre-trained model to perform an alignment and then evaluate the CLIP-similarity score between the retrieved sentence and the image/video (for videos we consider second frame). It should be noted that (i) A higher CLIP-Score doesn't automatically lead to higher downstream performance, particularly for retrieval benchmarks. (ii) CLIP-Score is only used as a metric is not a part of the pre-training process such as early-stopping or any learning rate schedule.

Implementation Details. We follow similar hyper-parameters as in METER [10]. METER by default uses a batch size of 4096, due to resource constraints of training on videos, we instead use a batch size of 1024, but we keep the number steps to be $100k$ for pre-training. We use AdamW [28] with differential learning rates for the co-attention layers and unimodal layers with $1e^{-4}$ and $1e^{-5}$ respectively during pre-training. We use deepspeed [36] and use "DeepSpeed ZeRO Stage 2" with 16-bit precision. For both cases, the initial 10% of the training is warm-up followed by a linear decay. We use 224×224 as the image dimension for pre-training in all cases and don't use any other augmentation. For visual backbone, we use ViT-B/32 weights trained on Imagenet [7] unless otherwise specified.

During pre-training, for each batch we perform MLM, ITM and ITC. For MLM we use mask ratio of 15%, for ITM we compare each image/video with 15 other negatives, for ITC we consider the entire batch of 1024 instances. To create a particular batch, we randomly sample an image/video instance and then if it contains more than one corresponding text, we select one at random.

To obtain the heuristic alignment in the fully-unaligned setting, we apply VinVL object detector [56] on 4 uniformly sampled frames. We then create an "object-string" by concatenating the unique objects in their order of appearance. We then use Sentence-Bert [38] model (all-mpnet-base-v2 model [42]) to obtain the closest text-match for each instance of image/video.

In the semi-supervised setting, we first pre-train on the $X\%$ of the aligned dataset. Different from usual pre-training scheme where we train for $100k$ steps, we instead train for 40 epochs which is roughly equivalent to $30k$ steps. Other hyper-parameters such as learning rate schedules, and batch size are kept the same.

	AD	CS	QA-Acc	t2i@1	t2i@5	t2i@10	i2t@1	i2t@5	i2t@10
NP-VMET	0.00		39.35	11.70	31.40	43.00	15.40	39.00	51.00
UA-VMET + FA	0.00	17.23	40.10	12.20	33.60	45.10	17.77	41.12	52.40
UA-VMET + IA	0.00	22.1	40.40	14.82	36.55	52.70	18.00	42.30	55.60
SS-VMET	0.10	23.9	39.90	14.12	34.58	46.55	17.50	42.01	53.10
SS-VMET + FA	0.10	24.48	40.01	14.28	35.56	48.57	17.89	42.83	53.31
SS-VMET + IA	0.10	25.2	41.10	15.78	39.95	51.91	19.90	43.30	54.70
UA-VMET + CFA	0.00	33.31	40.50	17.20	42.30	54.70	21.70	45.30	55.60
SS-VMET + CFA	0.10	32.98	41.83	20.10	48.90	58.10	24.77	46.71	58.19
FS-VMET	1.00	29.93	42.50	27.20	51.20	64.30	26.50	50.70	62.20
FROZEN [4]	1.00	-	-	31	59.5	70.5	-	-	-
ALPRO [23]	1.00	-	42.1	33.9	60.7	73.2	-	-	-
All-In-One [48]	1.00	-	44.3	34.4	65.4	75.8	-	-	-

Table 2. Results on **Shuffled Dataset**. AD: Amount of Aligned Data, CS: CLIP-Score, QA-Acc: Accuracy@1 on MSRVT-QA, t2i@X: Text-to-Video Retrieval Recall on MSRVT, i2t@X: Video-to-Text Retrieval Recall on MSRVT. See Section 4.2 for model information.

Given the trained model on initial paired data (\mathcal{M}_{θ_0} in Section 3.3), we perform *Iterative Alignment* by training for additional 10 epochs, but re-use the same hyper-parameters as for pre-training. In the semi-supervised setting, we re-use the existing paired data along with the newly aligned data at each iteration. We set N such that it is approximately 20% of the epoch. For batch size of 1024, and total of $6M$ images/videos, we set $N = 1200$.

4.3. Results and Discussions

In Table 2 we report the results of fully unaligned and semi-supervised setting by pre-training on the shuffled dataset. The fine-tuning datasets are based on video-question answering on MSRVT-QA [49] and video-retrieval on MSRVT [51].

Fully unaligned results on Shuffled Dataset. We first observe that for both fixed alignment (UA-VMET + FA) and iterative alignment (UA-VMET + IA), the performance on both QA and Retrieval task improves compared to no pre-training setup (NP-VMET). We also note a stark improvement in CLIP-Score for the iterative alignment case over 5 points compared to fixed alignment; similar improvements are reflected in the corresponding downstream tasks. We attribute this drastic improvement to the fact that the initial alignment which was based on heuristic was poor. The above results highlight two things (i) additional unaligned data always helps compared to performing no pre-training (ii) iterative alignment significantly helps for fully unaligned setting since the initial alignment was based on naive heuristics.

Semi-Supervised results on Shuffled Dataset. For the semi-supervised setting we compare three models. We find that even training on just 10% of the ground-truth data

(SS-VMET) can be very helpful in improving the downstream performance. Further, the obtained CLIP-Score of this model already surpasses UA-VMET model suggesting the importance of good quality paired data. For the next two models, we start from the trained SS-VMET model. In the fixed alignment case (SS-VMET + FA), we perform the alignment on the remaining 90% of the data only once and pre-train over the existing 10% ground-truth and 90% aligned data. In the iterative alignment case (SS-VMET + IA), the 10% ground-truth data is kept as is, but remaining 90% alignment is updated. Comparing the two cases, we find the relative improvement of the CLIP-Score to be small but significant and consistent, which is also found in downstream tasks. We conclude that (i) both fixed and iterative alignment help which is expected since we are again training on additional data (ii) the relative improvement of performing iterative alignment over the semi-supervised case to be smaller compared to fully unaligned case, likely because the initial alignment was significantly stronger (iii) fixed alignment on semi-supervised performs worse than iterative alignment in unaligned case suggesting the importance of updating the alignment.

CLIP-Aligned baselines for Shuffled Dataset. We report UA-VMET + CFA and SS-VMET + CFA, two strong baselines where the alignment on the unpaired data is obtained from CLIP [34] model. Recall that our model VMET is kept as previous models with the visual backbone initialized from ViT trained on ImageNet [7]. We find that directly using CLIP alignment far outperforms SS-VMET + IA. This is expected because the quality of retrieved examples from CLIP model is significantly better than that VMET which is trained on a smaller set.

Comparison to fully-aligned on Shuffled Dataset.

	%AD	CS	QA-Acc	t2i@1	t2i@5	t2i@10	i2t@1	i2t@5	i2t@10
NP-VMET	0.00		39.35	11.70	31.40	43.00	15.40	39.00	51.00
UA-VMET + FA	0.00	15.79	39.52	11.71	31.26	43.19	16.81	40.91	52.40
UA-VMET + IA	0.00	18.68	40.37	13.16	35.30	50.36	17.05	42.92	53.15
SS-VMET + FA	0.10	23.9	39.90	14.12	34.58	46.55	17.50	42.01	53.10
SS-VMET + IA	0.10	25.1	41.05	15.90	40.10	51.60	21.05	44.50	55.10
UA-VMET + CFA	0.00	35.27	41.59	18.50	46.71	58.93	22.53	45.41	57.64

Table 3. Results on **BookCorpus Dataset**. AD: Amount of Aligned Data, CS: CLIP-Score, QA-Acc: Accuracy@1 on MSRVT-QA, t2i@X: Text-to-Video Retrieval Recall on MSRVT, i2t@X: Video-to-Text Retrieval Recall on MSRVT. See Section 4.2 for model information.

	Didemo			COIN	CrossTask
	t2i@1	t2i@5	t2i@10	Frame Acc	Recall
NP-VMET	7.61	24.79	37.84	55.21	32.69
SS-VMET	28.10	55.32	65.49	55.83	34.54
SS-VMET + IA	30.71	57.12	69.74	57.73	35.65
FS-VMET	35.10	65.72	75.93	60.15	37.28
ALPRO	35.90	67.50	78.80	-	-
VideoClip	-	-	-	68.70	47.30

Table 4. Results on DiDeMo, COIN and CrossTask.

While using additional unaligned data in both semi-supervised and fully unaligned setting is helpful and is further improved with *Iterative Alignment*, there is a large gap with the fully aligned setting (FS-VMET). The gap is particularly large for retrieval benchmark where better quality paired data is quintessential for training.

Comparison to VLP Models on Shuffled Dataset. We compare our fully supervised model (FS-VMET) with other baselines such as FROZEN [4], ALPRO [23] and All-In-One [48]. Our model obtains slightly higher performance on QA, but lower performance on retrieval. We attribute this gap to the architectural difference such as our use space-time encoders used in FROZEN and ALPRO.

Results on BookCorpus Dataset. In Table 3 we compare our model VMET pre-trained using images and videos from Shuffled Dataset but the text used is instead obtained from BookCorpus [59]. This is a significantly challenging as well as realistic setting where both vision corpus and text corpus are different. We note very similar trends as with **Shuffled Dataset** with *Iterative Alignment* improving over *Fixed Alignment* in both fully-unaligned and semi-supervised cases. The consistent trends highlights why using additional data even if the corpora are unaligned can be useful for pre-training. Interestingly, we find the raw CLIP-Score as well as downstream performance for fully unaligned setting (UA-VMET + IA) worse than in Shuffled Dataset but improved result for semi-supervised set-

ting (SS-VMET + IA). We also report results where our model VMET is directly trained on the fixed alignment provided by CLIP (UA-VMET + CFA). Similar to previous case, model trained on CLIP alignment outperforms semi-supervised models.

Additional Down-stream Datasets In Table 4 we report results on three downstream tasks namely Didemo [2], COIN [45] and CrossTask [60]. We compare to two baselines namely ALPRO [23] and VideoClip [50]. We note that a fair comparison with VideoClip is expensive since it is trained on a large HowTo100M [32] dataset. Further, HowTo100M, COIN and CrossTask lie in the instructional videos domain which could provide the model with unfair advantage. We note consistent improvements in the semi-supervised setting suggesting the gains observed in the pre-training process are generalizable to a number of video-language downstream tasks.

Visualizations. We provide visualizations for the retrieved examples during both fixed and iterative alignment. We also provide visualizations on the down-stream tasks in the supplementary material.

5. Conclusion

In this work, we explore the task of Unaligned Video-Text Pre-Training (UVTP) which involves leveraging large amounts of unpaired video and text data to pre-train video-text models and thereby improve downstream performance. We consider two cases: a fully unaligned setup where we don't have access to any paired data and a semi-supervised setup where we have access to limited paired data. We address these setup by aligning the two modalities via object tags and pre-training a retrieval system. We further show that resulting pre-trained models can be utilized to iteratively refine the alignment between the two modalities. We evaluate our method in both synthetic setup where alignment is artificially removed as well as in realistic setup where the video and text corpora are distinct, and show the benefits of using unaligned data.

Appendix

The supplementary section contains additional details:

1. Model specification such as model hyper-parameters (Section A).
2. Dataset Information and BookCorpus Processing (Section B)
3. Visualizations for the retrieved data (Section C)
4. Additional Ablative study for semi-supervised setting with different amounts of data (Section D)

A. Model Implementation Details

As noted before, our model is based on METER implementation [10]. We report the hyper-parameters used for pre-training in our work in Table 5. This is the default set of hyper-parameters. We train our models on single node with 8 x A100 GPUs which accommodates 40GB size. To optimize for efficiency, we utilize DeepSpeed [36] for pre-training, specifically “DeepSpeed ZeRO Stage 2” setting which saves us lot of GPU memory ($\sim 30\%$).

We pre-train our models for 100k steps for completely aligned as well as fully unaligned settings. For Iterative Alignment part, we train for 10 epochs which is roughly equivalent to 100k steps. For semi-supervised setting, we constrain to max of 40 epochs or 100k steps whichever is earlier to avoid over-fitting on small datasets.

We use three pre-training losses namely, masekd language modeling (MLM), image-text matching (ITM) and image-text contrastive loss (ITC). MLM remains identical to METER implementation with the mask ratio at 15%. For ITM, we compare the positive image-text pair with 15 other image-text pairs. To create such pairs, we randomly sample 15 other text from the dataset. For ITC, we use the contrastive loss on the entire mini-batch. We use the contrastive loss implementation provided in CLIP [15] which introduces a scaling parameter.

B. Dataset Information

We note the pre-training dataset sizes in Table 6. For both images and videos, we resize with shorter side to 256 and during training time use center crop of 224×224 . For videos, we encode each video at 30 fps with CRF value of 23 to make the video storage manageable and not be a bottleneck during data-loading.

Shuffled Dataset. For shuffled dataset, we simply colate images and videos from the above pre-training datasets and the texts but remove the correspondence. Since we rely on KNN search (top-50) to obtain closest correspondence between the image/video and the text it is necessary to remove duplicates. Otherwise, the top-50 matches for an image/video could contain multiple texts which are essentially

Vision-Backbone	ViT-B/32
Language-Backbone	Roberta-Base
Additional Vision Layers	2L, 12H
Effective Batch Size	1024
TopLR	$1e^{-4}$
Bottom LR	$1e^{-5}$
Warmup steps	10k
Max Steps	100k
LR Schedule	Poly Decay with Linear Warmup
Precision	FP16 (DeepSpeed stage 2)
Image Dim	224×224
MLM Mask Ratio	15%
ITM # comparisons	15 negatives
ITC # comparisons	1024

Table 5. Model Hyper-parameters for usual pre-training. L, H denote number of layers and heads in the multi-head attention. MLM, ITM, ITC denote the three pre-training losses.

	COCO	VG	SBU	CC3M	WV2M
# Images	113k	108k	875k	2.8M	2.4M
# Paired Texts	565k	5.5M	875k	2.8M	2.4M
# Texts	565k	5.5M	1M	3.3M	2.5M

Table 6. Number of Texts denote the total number of annotations (image/video link and text) provided in the dataset. However, not all images/videos are available. We denote the number of images/videos we could download and use for pre-training in the first row, and the corresponding paired texts in the second row.

the same. For shuffled dataset, we simply remove obvious duplicate *i.e.* same lower case after removing white-spaces. This brings the total number of text from 12M to 9M texts.

BookCorpus Dataset For bookcorpus, simply reading all the sentences as separate text provides us with approximately 86M texts. Unlike the Shuffled Dataset case, many of the sentences from the BookCorpus dataset have (i) proper nouns such as names of person like “Charlie” (ii) sentences with no occurrence of objects or actions (iii) non-visual or abstract verbs such as “envy”. To save computation time in the KNN search, we pre-process the dataset and aim to have similar number of texts as in Shuffled Dataset.

We use Spacy [13] to first obtain lemmatized verbs and object names from each sentence, and remove all sentences where no objects other than “person” are found. We further replace the names of people by using NER-tagging to obtain “PER” tokens and replace them with a phrase “person”. Finally, we filter out sentences which only contain non-visual verbs such as “belong”, “possess”, “know”, “realize”. These are also known as “Stative Verbs”. At the end of the filtering process, we are left with around 30M texts. We then prune out close duplicate sentences, which









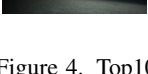

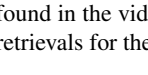
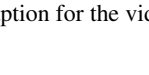


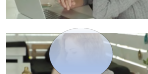
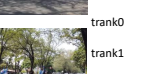

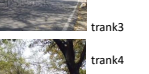



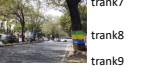
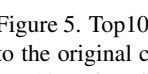
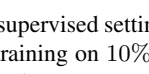
	Object String	light finger person shirt hand door book arm table man ring		Object String	neck shirt rack hand necklace eye woman bag hat arm apple ear shelf hair doll nose face
	GT Caption	Closeup footage of young man taking passport and cash from safe in hotel		GT Caption	Woman choosing yarn for knitting in shop for hobby and handmade. 4k uhd.
	trank0	Wireless keyboard. female hands typing on keyboard. white keyboard. computer. internet. technology. social network media. businesswoman. red nails. manicure fingers typing keyboard. wooden table. uhd		trank0	Christmas ornaments baubles bauble glass ball balls xmas ornament
	trank1	Necklace on the neck. finger in mouth		trank1	Handicraft items on the shelf
	trank2	Light of the desk lamp illuminates the dried flowers, a key, an old letter, writing-pad and inkwell with a pen. memory, family archive		trank2	Variety of clothes hanging on rack in boutique
	trank3	Wrist watch. macro shot.		trank3	Toys jewelry handmade sweets
	trank4	Men accessories		trank4	Christmas presents close up. christmas decoration
	trank5	Spiral clock track of time. antique clock dial close-up. vintage pocket watch.		trank5	Many different toys on the christmas tree in the living room
	trank6	Set of 8 hand touchscreen gestures, showing the uses of computer touchscreen, tablet or trackpad. married man hand. tablet. luma matte.		trank6	Clothes hanging and lying on the shelves in the store
	trank7	A macro shot of a watch on a hand.		trank7	Hanging a brown fedora hat and a black coat on an old fashioned coat rack with antique wallpaper.
	trank8	Closeup man hand pointing button clicking keyboard working notebook laptop wooden table public workplace mockup green screen empty monitor internet education website chroma key chromakey network media		trank8	New year, wedding, birthday, anniversary gifts, souvenirs in the box
	trank9	Hand shape black background screen finger catch touch hands shadow shade gesture human sign arm dark icon skin		trank9	Wedding decor. rack focus. vases of flowers on the table outdoor

Figure 4. Top10 Retrieved Texts for given videos from WebVid-2M for fully unaligned case. Object String denotes the unique objects found in the video in order of appearance. GT Caption denote the original caption for the video. Finally, “trank0-9” denote the top10 text retrievals for the video. Visible Faces are blurred.

	GT Caption	Young businesswoman with laptop in outdoor cafe 4k		GT Caption	Bandung, west java / indonesia - august 25, 2019: established shot of busy traffic at w. r supratman street
	Initial Caption	Lifestyle woman eating sushi in a hotel room in the evening after work.		Initial Caption	Flock of pigeons eating bread outdoors in the city street. lot of pigeons eat food on the street. feeding pigeons on the sidewalk in the park. thousands of pigeons crowd on sidewalk.
	trank0	Woman hand use laptop mouse and drinking coffee, stock footage		trank0	Time lapse of a suburban boulevard street from ground level with blurred traffic passing by during the afternoon.
	trank1	Business, people, paperwork and technology concept - businesswoman with laptop computer and papers at office.		trank1	People walking on a sidewalk with palm trees.
	trank2	Happy young freelancer lying on the floor and working on a laptop.		trank2	A local market located next to bodhi gaya sell almost everything from agricultural product to souvenir.
	trank3	Business teamwork drinking coffee working on laptop colleagues office interior.		trank3	Hanoi, vietnam - feb 20, 2017: people and vehicles at the morning , busy intersection locating next to hoan kiem lake in old quarter of Hanoi.
	trank4	1970s: woman quickly types on typewriter while reading paper. office with men and women at work.		trank4	Taiping, malaysia - 17 october 2017: walking along malaysia traditional oriental shop with five foot sidewalk.
	trank5	Woman working as lawyer with old clients or financial consultant talking to elderly customers. senior husband and wife meeting health insurance broker or retirement plan agent in office.		trank5	Ho chi minh city, vietnam 2016: overview the center of ho chi minh city from sai gon river at night very beautiful . ho chi minh city is the biggest city in Vietnam.
	trank6	Young woman using laptop and drinking cocktail in café.		trank6	Ho chi minh city, Vietnam - 13 feb, 2018: ben binh dong (binh dong harbour) in lunar new year with flower boats along side the river, sai gon, Vietnam.
	trank7	Young beautiful woman working on laptop in a city port with yachts on background.		trank7	Bangkok - october 10,2014: congestion on the road on rush hour time in Thailand.
	trank8	Side view of young woman looking at catalog on the couch.		trank8	New york city, new york - october 7: slow moving traffic in downtown establishing of new york city, new york on october 7, 2017.
	trank9	Closeup of young woman signing a contract to become mlm partner. cosmetics consulting, direct sale and multi level marketing concept.		trank9	Hanoi, vietnam, january 2020: city roundtrip in a open bus

Figure 5. Top10 Retrieved Texts for given videos from WebVid-2M for semi-supervised setting (10% semi-supervised). GT Caption refers to the original caption. Initial Caption refers to the Top-1 caption after pre-training on 10% aligned data. Finally, “trank0-9” denote the Top10 retrieved texts at the end of Iterative Alignment. Visible faces are blurred.

we define as when the order of the lemmatized objects and actions are exactly the same. After such filtering, we are left around 12M captions which we use for all bookcorpus related training.

C. Visualizations

We visualize the retrieval outputs for fully unaligned case in Figure 4. Given a video, we first obtain an Object String. We use this Object String as a sentence and use Sentence-Bert [38] to retrieve Top-10 text from the Shuffled Dataset. As expected, given that this a heuristic the retrieved texts are not close to the Ground-Truth caption. For instance, the text matches with “manicure fingers” due to “finger” object appearing in the original video.

In Figure 5, we visualize the retrieval results for semi-supervised case. Given some aligned video-text data (10% of Shuffled Dataset in this case), we first pre-train VMET on this data using MLM, ITM and ITC. Then, we use ITC

head of this model to perform retrieval. The top-1 retrieval is denoted by “Initial Caption”. While the Initial Caption is not very close to the Ground-Truth caption, we still find some relevance. For instance, in the left case, it was able to relate to the concept of “woman” and “working”. Similarly, in second case, it found “sidewalk”. At then end of pre-training via Iterative Alignment the retrievals are significantly improved.

D. Ablative Study semi-supervised setting

Ablation on Aligned Data. In Table 7, we provide results of pre-training in the semi-supervised setting with varying amounts of aligned data. For pre-training we use the Shuffled Dataset, and for fine-tuning we use MSRVTT-QA with QA-Acc@1 as our metric. In each case, we first train VMET on the amount of aligned data available. In second column (FT), we directly fine-tune the pre-trained model. In the third and fourth column, we use the pre-

AD	FT	FA + FT	IA + FT
0.00	39.35	-	-
0.01	36.13	36.92	36.91
0.05	38.59	39.94	40.72
0.10	39.90	40.01	41.10
0.25	40.53	41.32	41.64
0.50	40.56	41.51	41.91
0.75	41.22	41.82	41.94
0.90	42.04	42.10	42.14
1.00	42.50	42.50	42.50

Table 7. Ablative study by pre-training VMET on different amounts of Aligned Data available in Shuffled Dataset. The downstream task is MSRVTQ and the metric is Acc@1. Models are first pre-trained on the available aligned data. FT: Directly fine-tune pre-trained model. FA + FT: Apply fixed alignment on the entire Shuffled Dataset, and then fine-tune. IA+FT: Apply iterative alignment then fine-tune.

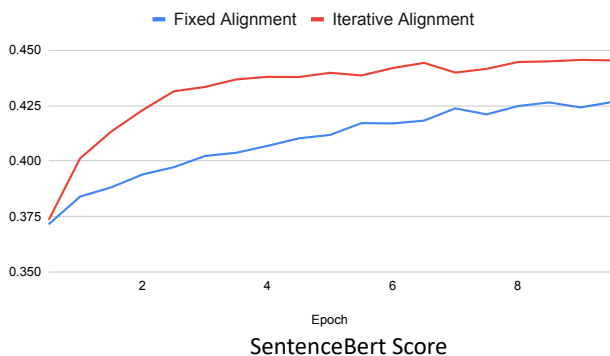


Figure 6. Sentence-Bert Matching Score between ground-truth and top-1 prediction. We use the Shuffled Dataset and the 10% semi-supervised setting.

trained model to obtain either a fixed alignment (FA + FT) or Iterative Alignment (IA+FT). The first row denotes no pre-training case, hence we report only fine-tuning results. The last row denotes fully aligned case. Since we have the entire alignment, all three columns are exactly the same.

We observe that in the low aligned data regime, pre-training leads to worse performance than not pre-training. We attribute this to over-fitting of the model to the small amount of data. In the higher aligned data regime, we find additional aligned data leads to improvement although the improvement margins are lower because the additional amount of data for alignment used is small.

Ablation of Fixed vs Iterative Alignment. In Figure 6, we plot the Sentence-Bert matching score between the ground-truth matching text and the top-1 prediction of the model. We note that we have access to the ground-truth

since we are working on Shuffled Dataset, thus the same metric cannot be plotted for BookCorpus dataset. We sample data around 20k images/videos for every 0.25 of an epoch. As can be observed, our Iterative Alignment improves the SentenceBert score compared to the Fixed Alignment setting.

References

- [1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4–31, 2015. [1](#)
- [2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. [2](#), [6](#), [8](#)
- [3] Max Bain, Arsha Nagrani, A. Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings. *ArXiv*, abs/2005.04208, 2020. [3](#)
- [4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)
- [5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. [2](#)
- [6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. [2](#)
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. [6](#), [7](#)
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. [2](#)
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. [2](#), [3](#)
- [10] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18145–18155, 2022. [1](#), [2](#), [3](#), [6](#), [9](#)
- [11] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. [2](#), [3](#)
- [12] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. [6](#)
- [13] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. [9](#)
- [14] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiase Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *The European Conference on Computer Vision (ECCV)*, 2020. [3](#)
- [15] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below. [9](#)
- [16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. [3](#)
- [17] Wonjae Kim, Bokyoung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. [2](#)
- [18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. [1](#), [2](#), [5](#)
- [19] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*, 2018. [2](#)
- [20] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *EMNLP*, 2018. [2](#)
- [21] Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *ArXiv*, abs/2206.03428, 2022. [2](#), [3](#)
- [22] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7327–7337, 2021. [3](#)
- [23] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963, 2022. [3](#), [6](#), [7](#), [8](#)
- [24] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. [1](#), [2](#), [3](#)
- [25] Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. Unsupervised vision-and-language pre-training without parallel images and captions. In *NAACL*, 2021. [2](#), [3](#), [4](#)

- [26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 2, 5
- [27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2
- [28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6
- [29] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 2
- [30] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 2, 3
- [31] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9876–9886, 2020. 3
- [32] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 2, 3, 4, 8
- [33] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 1, 2, 5
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 3, 6, 7
- [35] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 2
- [36] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020. 6, 9
- [37] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 4
- [38] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. 4, 6, 10
- [39] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015. 2, 4
- [40] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. 1, 2, 5
- [41] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *ArXiv*, abs/2107.06383, 2022. 1, 3
- [42] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020. 6
- [43] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vi-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020. 2
- [44] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. 2
- [45] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 2, 6, 8
- [46] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 2
- [47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 3
- [48] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Ge Yuying, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*, 2022. 6, 7, 8
- [49] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017. 2, 6, 7
- [50] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 1, 3, 8

- [51] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 2, 6, 7
- [52] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Rui Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *ArXiv*, abs/2209.06430, 2022. 3
- [53] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 1
- [54] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16354–16366, 2022. 3
- [55] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021. 3
- [56] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*, 2021. 2, 3, 6
- [57] Mingyang Zhou, Licheng Yu, Amanpreet Singh, Mengjiao MJ Wang, Zhou Yu, and Ning Zhang. Unsupervised vision-and-language pretraining via retrieval-based multi-granular alignment. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16464–16473, 2022. 2, 3, 4, 5
- [58] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8743–8752, 2020. 3
- [59] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015. 2, 6, 8
- [60] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545, 2019. 2, 6, 8