Aligning Step-by-Step Instructional Diagrams to Video Demonstrations

Jiahao Zhang[1,*] Anoop Cherian[2] Yanbin Liu[1]
Yizhak Ben-Shabat[1,3,†] Cristian Rodriguez[4] Stephen Gould[1,‡]
[1]The Australian National University, [2]Mitsubishi Electric Research Labs
[3]Technion Israel Institute of Technology, [4]The Australian Institute for Machine Learning
[1]{first.last}@anu.edu.au [2]cherian@merl.com [3]sitzikbs@gmail.com [4]crodriguezop@gmail.com
https://davidzhang73.github.io/en/publication/zhang-cvpr-2023/

Abstract

Multimodal alignment facilitates the retrieval of instances from one modality when queried using another. In this paper, we consider a novel setting where such an alignment is between (i) instruction steps that are depicted as assembly diagrams (commonly seen in Ikea assembly manuals) and (ii) segments from in-the-wild videos; these videos comprising an enactment of the assembly actions in the real world. We introduce a supervised contrastive learning approach that learns to align videos with the subtle details of assembly diagrams, guided by a set of novel losses. To study this problem and evaluate the effectiveness of our method, we introduce a new dataset: IAW—for Ikea assembly in the wild—consisting of 183 hours of videos from diverse furniture assembly collections and nearly 8,300 illustrations from their associated instruction manuals and annotated for their ground truth alignments. We define two tasks on this dataset: First, nearest neighbor retrieval between video segments and illustrations, and, second, alignment of instruction steps and the segments for each video. Extensive experiments on IAW demonstrate superior performance of our approach against alternatives.

1. Introduction

The rise of *Do-It-Yourself* (DIY) videos on the web has made it possible even for an unskilled person (or a skilled robot) to imitate and follow instructions to complete complex real world tasks [4, 23, 31]. One such task that is often cumbersome to infer from instruction descriptions yet easy to imitate from a video is the task of assembling furniture from its parts. Often times the instruction steps involved in such a task are depicted in pictorial form, so that they are comprehensible beyond the boundaries of language (e.g., Ikea assembly manuals). However, such instructional diagrams can sometimes be ambiguous, unclear, or may not match the furniture parts at disposal due to product variability. Having access to video sequences that demonstrate the precise assembly process could be very useful in such cases.

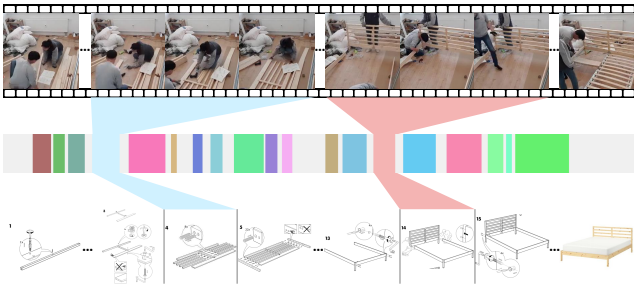


Figure 1. An illustration of video-diagram alignment between a YouTube video (top) He0pCeCTJQM and an Ikea furniture manual (bottom) s49069795.

Unfortunately, most DIY videos on the web are created by amateurs and often involve content that is not necessarily related to the task at hand. For example, such videos may include commentary about the furniture being assembled, or personal assembly preferences that are not captured in the instruction manual. Further, there could be large collections of videos on the web that demonstrate the assembly process for the same furniture but in diverse assembly settings; watching them could consume significant time from the assembly process. Thus, it is important to have a mechanism that can effectively align relevant video segments against the instructions steps illustrated in a manual.

In this paper, we consider this novel task as a multimodal alignment problem [25, 27], specifically for aligning in-the-wild web videos of furniture assembly and the respective diagrams in the instruction manuals as shown in Fig. 1. In contrast to prior approaches for such multimodal alignment, which usually uses audio, visual, and language modalities,

*Supported by an ANU-MERL PhD scholarship agreement.
†Supported by Marie Sklodowska-Curie grant agreement No. 893465.
‡Supported by an ARC Future Fellowship No. FT200100421.

our task of aligning images with video sequences brings in several unique challenges. First, instructional diagrams can be significantly more abstract compared to text and audio descriptions. Second, illustrations of the assembly process can vary subtly from step-to-step (e.g., a rectangle placed on another rectangle could mean placing a furniture part on top of another). Third, the assembly actions, while depicted in a form that is easy for humans to understand, can be incomprehensible for a machine. And last, there need not be common standard or visual language followed when creating such manuals (e.g., a furniture piece could be represented as a rectangle based on its aspect ratio, or could be marked with an identifier, such as a part number). These issues make automated reasoning of instruction manuals against their video enactments extremely challenging.

In order to tackle the above challenges, we propose a novel contrastive learning framework for aligning videos and instructional diagrams, which better suits the specifics of our task. We utilize two important priors—a video only needs to align with its own manual and adjacent steps in a manual share common semantics—that we encode as terms in our loss function with multimodal features computed from video and image encoder networks.

To study the task in a realistic setting, we introduce a new dataset as part of this paper, dubbed IAW for Ikea assembly in the wild. Our dataset consists of nearly 8,300 illustrative diagrams from 420 unique furniture types scraped from the web and 1,005 videos capturing real-world furniture assembly in a variety of settings. We used the Amazon Mechanical Turk to obtain ground truth alignments of the videos to their instruction manuals. The videos involve significant camera motions, diverse viewpoints, changes in lighting conditions, human poses, assembly actions, and tool use. Such in-the-wild videos offer a compelling setting for studying our alignment task within its full generality and brings with it a novel research direction for exploring the multimodal alignment problem with exciting real-world applications, e.g., robotic imitation learning, guiding human assembly, etc.

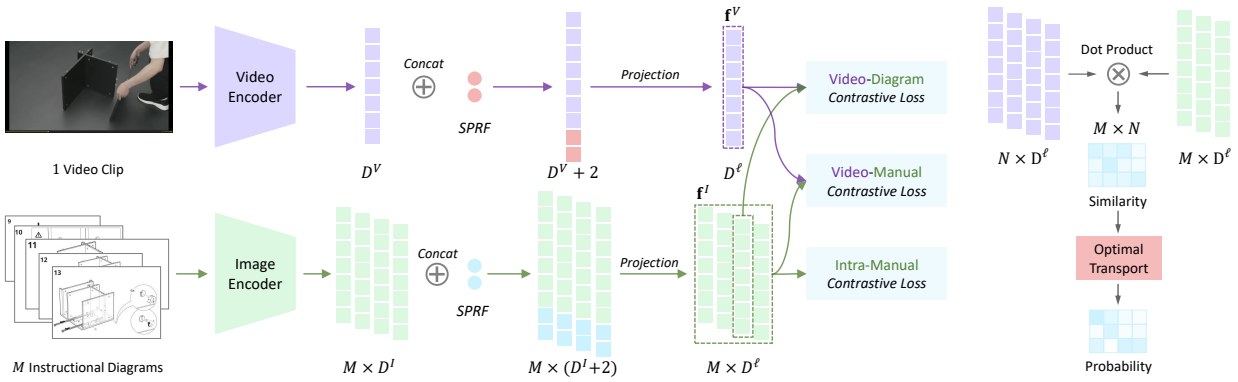
To evaluate the performance of our learned alignment, we propose two tasks on our dataset: (i) nearest neighbor retrieval between videos and instructional diagrams, and (ii) alignment of the set of instruction steps from the manual to clips from an associated video sequence. Our experimental results show that our proposed approach leads to promising results against a compelling alternative, CLIP [27], demonstrating 9.68% improvement on the retrieval task and 12% improvement on the video-to-diagram alignment task.

2. Related Work

Assembly and Instructional Datasets. Multimodal video datasets (e.g., [2, 10, 28, 32, 37, 44]) bridge the gap between video and other modalities such as the narratives from the video or instruction texts. Among them, EPIC Kitchens [10] and YouCook2 [44] align each video clip with the cooking procedure narratives. Our dataset is more closely related to IKEA ASM [2] and IKEA-FA [32], which demonstrate furniture assembly instructions. There are some other datasets focusing on converting assembly manuals to more comprehensible formats. LEGO [37] demonstrates how to obtain an executable plan from the assembly manuals while Shao et al. [28] parses furniture assembly instructions into 3D models based on their manuals. Unlike all of the above datasets, the proposed IAW dataset aims to achieve the novel multimodal task of aligning in-the-wild web videos with step-by-step instructional diagrams.

Multimodal Alignment. The classic work of Everingham et al. [13] focuses on aligning subtitle-transcript with person IDs in videos. Later works [12, 30] started aligning video segments with text story-lines. Recently, different approaches have been proposed for the text-video retrieval task, e.g., extracting fine-grained text features [18, 39], augmenting with more modalities [15, 38], and contrastive text-video learning [3, 8, 22, 25]. Among them, Han et al. [19] tackles alignment between assembly videos and text manuals. However, due to the modality distinctions between text and image, these methods cannot be directly adopted to solve our problem. Apart from the video modality, sketch images are similar to our instructional diagrams in the sense that both are black-white, text-free, and highly iconic abstract images. Sketch-based video retrieval [9] aims to retrieve specific video clips given a sketch image or sequence. A recent related work to ours is Xu et al. [41], which extracts image features from both sketches and motion vector images, and optical flow from video clips. They apply a triplet loss and a relation module on these extracted multimodal features to train the model. However, their motion vector sketch is ad-hoc to specific video types, such as sports. Compared with Xu et al. [41], our method is more general and supports two tasks from both video-to-diagram and diagram-to-video directions.

Contrastive Learning. Contrastive learning was first introduced for self-supervised representation learning [6, 7, 16, 20, 26]. Then, the idea was naturally adapted for multimodal learning tasks, such as text-image alignment [27] and text-video retrieval [25]. CLIP [27] designs a contrastive pre-training approach by predicting the correct pairs between images and their captions. Since CLIP verifies the effectiveness of cross-modality contrastive learning, recent works [1, 35, 42] have incorporated it into the related models, facilitating the cross-modality video retrieval task. Different from existing works, our alignment problem not only requires good contrastive between video clips and instructional diagrams but also entails distinguishing the subtle details in step-by-step manuals. This motivates us to design three task-specific contrastive losses.



(a) Training stage. (b) Inference stage.

Figure 2. Schematic of our overall architecture. During training we extract features from each input video clip and set of instructional diagrams, respectively, using pre-trained encoders. We concatenate these with sinusoidal progress rate features (SPRF) introduced in Sec. 3.1 and project into the same D^ℓ dimensionality space. The matched video clip and instructional diagram feature pairs are used for Video-Diagram Contrastive Loss, the video clip feature and M instructional diagram features are fed into Video-Manual Contrastive Loss, and M instructional diagram features themselves are used in Intra-Manual Contrastive Loss introduced in Sec. 3.2. During inference all video features from N sequential video clips, and all M instructional diagram features from the corresponding manual are computed. We then form a similarity matrix and apply optimal transport (OT) introduced in Sec. 3.3 to produce the final alignment probabilities.

3. Video-Instruction Alignment

We formulate the task of aligning video segments to instructional diagrams as a variant of video-to-image matching. Here the idea is to retrieve the image from a candidate set that most closely depicts the activity occurring in the short video clips and vice versa. Importantly, since different instructional videos will involve different numbers of steps the candidate set will necessarily have variable cardinality (unlike, say, multi-class classification tasks).

Formally, given a set of N video clips $\{V_i\}_{i=1}^N$ and a set of M instructional diagrams $\{I_j\}_{j=1}^M$, our goal is to train a model to predict the correspondence between diagrams and clips. A standard approach for addressing this problem is to learn a joint embedding space for videos and diagrams such that matching video-diagram pairs map near to each other in the embedding space. Let $\mathbf{f}_i^{\mathrm{V}}$ and $\mathbf{f}_j^{\mathrm{I}}$ denote the feature embedding for the i-th video clip and j-th instructional diagram, respectively, and let f_{sim} be some similarity measure. Then, once the embedding space is learned we can use the model to predict the index of the instructional diagram corresponding to a given video clip V as

$$j^\star = \underset{j=1,\ldots,M}{\operatorname{argmax}} \; f_{\text{sim}}(\mathbf{f}^{\mathrm{V}}, \mathbf{f}_j^{\mathrm{I}}). \tag{1}$$

Likewise, we can find the video segment that most closely matches a given instructional diagram I as

$$i^\star = \underset{i=1,\ldots,N}{\operatorname{argmax}} \; f_{\text{sim}}(\mathbf{f}_i^{\mathrm{V}}, \mathbf{f}^{\mathrm{I}}). \tag{2}$$

This can be generalized to top-k retrieval. Last, we can enforce matching constraints, such as through optimal transport or dynamic time warping if order information is available, to jointly match all clips in a video to all steps in an instruction manual. Fig. 2 depicts the overall model.

In this work, we use cosine similarity for f_{sim}. The embedding vectors $\mathbf{f}_i^{\mathrm{V}}$ and $\mathbf{f}_j^{\mathrm{I}}$ are computed using video and image encoders trained under a contrastive loss and optionally augmented with temporal features such as we now describe.

3.1. Sinusoidal Progress Rate Feature

Instruction manuals contain an ordered sequence of steps that is typically, although not always, followed during the assembly process. However, the time needed to perform each step varies greatly depending on complexity of the step and experience of the assembler. This suggests a weak correlation between (proportional) timestamps in the video and progress through the assembly process. We can make use of this prior by including temporal ordering information in the video and diagram feature representations.

Given a video clip V sampled from a full video of length t_{duration} seconds, with start time t_{start} and end time t_{end}, we define the video progress rate r^{V} of that video clip as

$$r^{\mathrm{V}} = (t_{\text{start}} + t_{\text{end}})/2t_{\text{duration}} \tag{3}$$

and the instructional diagram progress rate r^{I} for the j-th step from a manual with M total steps is simply $r^{\mathrm{I}} = j/M$. Because we are using a cosine similarity function f_{sim}, we map the progress feature onto a half circle so that high similarity score coincides with when they align. The final sinusoidal progress rate feature (SPRF) is then

$$(\sin(\pi r^V), \cos(\pi r^V)) \tag{4}$$

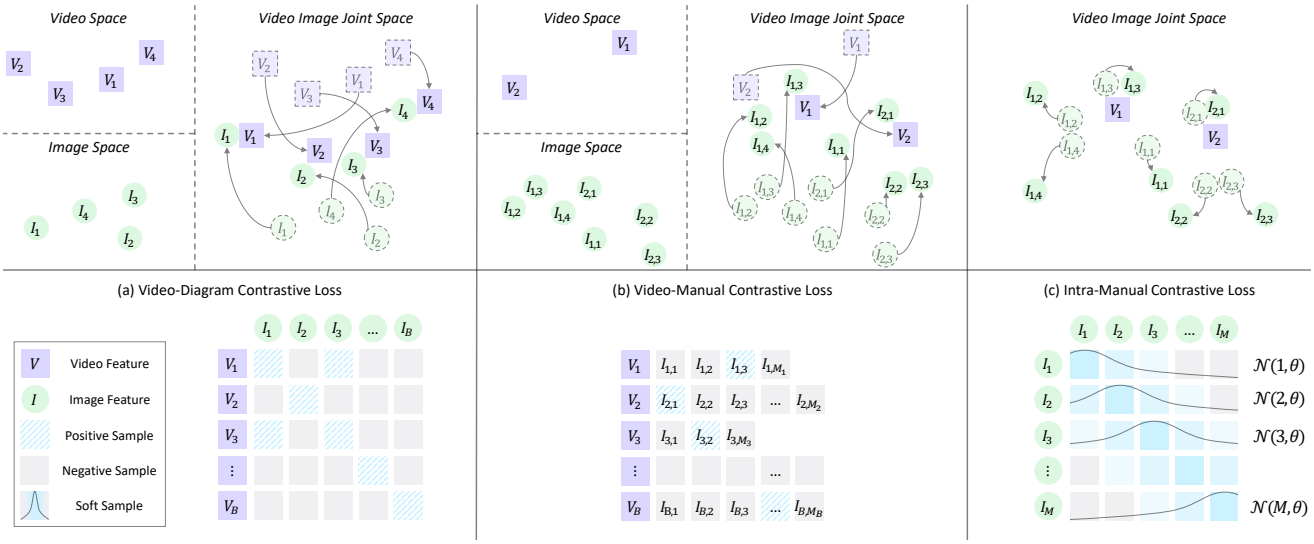


Figure 3. Visualization of our three losses described in Sec. 3.2. The intent is depicted in the first row and the batch formation in the second row. Loss (a) tries to pair video and image up across the entire dataset. Loss (b) only matches video clips and images corresponding to the same manual. And loss (c) push images from the same manual apart from each other for better feature discrimination.

for video and similarly for the instructional diagram, which we append to the feature embeddings extracted from the respective encoders (see Fig. 2). Before and after the concatenation, the features are L2 normalized to alleviate side-effect due to fluctuation of numerical value scale. Two fully connected layers then project each modality feature into the same dimensional space to form representations $\mathbf{f}^I$ and $\mathbf{f}^V$ for further similarity comparison.

3.2. Training Losses

Starting with pre-trained video and image encoders we finetune our model using variants of contrastive learning, which has recently been made popular for cross-modal matching by models like CLIP [27]. In this setting mini-batches are constructed by sampling video clip-instructional diagram pairs (V_i, I_i) to optimize an infoNCE loss [17, 26] where pairs (V_i, I_i) are considered positive and (V_i, I_j), $i \neq j$ are considered negative. Here we sample randomly from all videos and instruction manuals in the training data.

Formally, for mini-batch containing B pairs, define

$$p_{ij}^{V2I} = \frac{\exp(f_{\text{sim}}(\mathbf{f}_i^V, \mathbf{f}_j^I)/\tau)}{\sum_{b=1}^{B} \exp(f_{\text{sim}}(\mathbf{f}_i^V, \mathbf{f}_b^I)/\tau)} \tag{5}$$

$$p_{ji}^{I2V} = \frac{\exp(f_{\text{sim}}(\mathbf{f}_i^V, \mathbf{f}_j^I)/\tau)}{\sum_{b=1}^{B} \exp(f_{\text{sim}}(\mathbf{f}_b^V, \mathbf{f}_j^I)/\tau)} \tag{6}$$

to be the probability of matching video V_i to image I_j and the probability of matching image I_j to video V_i, respectively. Here τ is a temperature parameter that controls the bias towards difficult examples [34]. Standard contrastive learning then minimizes

$$\mathcal{L}_{\text{infoNCE}} = -\frac{1}{2B}\left(\sum_{i=1}^{B} \log p_{ii}^{V2I} + \sum_{j=1}^{B} \log p_{jj}^{I2V}\right). \tag{7}$$

We note that this vanilla version of contrastive learning does not consider situations where there may be many-to-one matches between pairs. Specifically, in our application multiple video clips may map to the same step. We introduce a specialized loss to deal with this scenario.

Video-Diagram Contrastive Loss (Fig. 3(a)). Contrastive learning frameworks benefit from large batch sizes [27]. However, as batch size increases there is a greater chance that we sample multiple videos matching to the same diagram within the batch, which violates the assumptions of the infoNCE loss. To better handle these cases we build on the work of [36] that introduces a Kullback-Leibler (KL) divergence loss between predicted and ground truth distributions, $\boldsymbol{p}$ and $\boldsymbol{q}$, respectively. However, rather than KL-divergence, we prefer Jensen-Shannon (JS) divergence, which we find improves training stability.

Let $\boldsymbol{p}^{V2I}$ and $\boldsymbol{p}^{I2V}$ be vectors containing all video-to-diagram and diagram-to-video probabilities introduced in Eqs. 5 and 6, respectively. Similarly, let $\boldsymbol{q}^{V2I}$ and $\boldsymbol{q}^{I2V}$ be the corresponding ground truth alignment distributions. Then our video-diagram contrastive loss is defined as

$$\mathcal{L}^{\text{VI}} = \frac{1}{2}\left(D_{JS}(\boldsymbol{p}^{V2I} \| \boldsymbol{q}^{V2I}) + D_{JS}(\boldsymbol{p}^{I2V} \| \boldsymbol{q}^{I2V})\right) \tag{8}$$

where D_{JS} is the Jensen-Shannon divergence.

Video-Manual Contrastive Loss (Fig. 3(b)). The above losses align video and diagram pairs globally across the entire training dataset. However, for our task we know that a given video clip only needs to match against one of the steps in its corresponding instruction manual, not other manuals. Hence, we can perform a more task-specific discrimination by exploiting this prior information in the model. To do so we modify our procedure for constructing a mini-batch to first sample a video clip V_i and then include all instructional diagrams $\{I_1, \ldots, I_{M_i}\}$ from the video's corresponding manual. One of these diagrams will be the ground truth positive match for the clip. We then employ a classification loss based on cross entropy (CE) as

$$\mathcal{L}^{VM} = \sum_{i=1}^{B} \frac{M_i}{\sum_{b=1}^{B} M_b} CE(\boldsymbol{p}_i^{V2I}, \boldsymbol{p}_i^{gt}) \tag{9}$$

where M_i indicates the length of the manual corresponding to the i-th video. Here $\boldsymbol{p}_i^{V2I} \subseteq (p_{ij}^{V2I})_{j=1}^{B}$ is a subvector of probabilities for matching video V_i to all diagrams I_j from the corresponding manual and $\boldsymbol{p}_i^{gt}$ is the associated one-hot ground truth encoding. We weight each term in the loss by $\frac{M_i}{\sum_{b=1}^{B} M_b}$ to give more emphasis to more difficult assemblies, assumed to be the ones containing more steps.

Intra-Manual Contrastive Loss (Fig. 3(c)). The previous losses only consider contrasting embeddings between videos and diagrams. However, most furniture assembly tasks involve a progressive process where the visual similarity between successive steps is large. Indeed, the main component of the assembly is often introduced early in the assembly process and dominates the instructional diagram. This makes it challenging to distinguish between steps. To encourage diagrams from the same manual to be spread out in embedding space, so that they are more easily distinguished, we introduce an intra-manual contrastive loss.

Similar to the video-to-diagram and diagram-to-video matching probabilities defined above, let

$$p_{jk}^{I2I} = \frac{\exp(f_{\text{sim}}(\mathbf{f}_j^I, \mathbf{f}_k^I)/\tau)}{\sum_{m=1}^{M} \exp(f_{\text{sim}}(\mathbf{f}_j^I, \mathbf{f}_m^I)/\tau)} \tag{10}$$

be the probability of matching diagram I_j to diagram I_k from the same manual according to our similarity metric. Then we define our intra-manual contrastive loss as

$$\mathcal{L}^{M} = \sum_{j=1}^{B} \frac{M_j}{\sum_{b=1}^{B} M_b} D_{JS}\left(\boldsymbol{p}_j^{I2I} \| \mathcal{N}(j, \theta)\right) \tag{11}$$

where $\boldsymbol{p}_j^{I2I}$ is the softmax normalized diagram-to-diagram probability vector associated with diagram I_j, and $\mathcal{N}(j, \theta)$ is a univariate Gaussian distribution with mean j, learnable variance θ and discretized on support $\{1, \ldots, M_j\}$. This encourages distances in diagram embedding space to correspond to distances between steps in the manual. We use a normal distribution instead of a delta distribution as a relaxation since nearby negative diagrams are still likely to share some semantics.

3.3. Set Matching

Our model is very general. Given a single video clip we can retrieve the most likely diagram showing the assembly step and given a single diagram we can retrieve a set of best matching video clips. To align an entire video (sequence of clips) to an entire instruction manual, we can add approximate one-to-one matching priors or temporal constraints, through optimal transport (OT) or dynamic time warping (DTW), respectively. As we will see in our experiments, the absence of temporal order constraints in OT slightly outperforms DTW due to occasional out-of-order execution of assembly steps or strong false matches.

To apply either method we first extract features $\mathbf{f}_i^V$ for an entire video $\{V_i\}_{i=1}^{N}$ and $\mathbf{f}_j^I$ for all instructional diagrams in the corresponding manual $\{I_j\}_{j=1}^{M}$. Denote by s_{ij} the similarity $f_{\text{sim}}(\mathbf{f}_i^V, \mathbf{f}_j^I)$ between video clip V_i and diagram I_j. Let $\overline{s} = \max_{i,j} s_{ij}$ and $\underline{s} = \min_{ij} s_{ij}$. We then construct a cost matrix $C \in \mathbb{R}^{N \times M}$ with entries

$$C_{ij} = \frac{s_{ij}^{\alpha} - \underline{s}^{\alpha}}{\overline{s}^{\alpha} - \underline{s}^{\alpha}}. \tag{12}$$

Here $\alpha > 1$ accentuates the similarity differences and the normalization by $\overline{s}^{\alpha} - \underline{s}^{\alpha}$ restricts the range of C_{ij} to $[0, 1]$. The optimal transportation plan $T^{\star}$ obtained by solving the entropy regularized optimal transport problem,

$$\begin{aligned} \text{minimize} \quad & \textstyle\sum_{i=1}^{N} \sum_{j=1}^{M} T_{ij} C_{ij} - \epsilon H(T) \\ \text{subject to} \quad & \textstyle\sum_{i=1}^{M} T_{ij} = \frac{1}{N}, \text{ for } j = 1, \ldots, N \\ & \textstyle\sum_{j=1}^{N} T_{ij} = \frac{1}{M}, \text{ for } i = 1, \ldots, M, \end{aligned} \tag{13}$$

gives the joint probability of matching videos and diagrams. It can be found efficiently by applying the Sinkhorn-Knopp algorithm [29] to the optimization problem defined above.

In a similar fashion, we can use DTW to find the optimal path through the cost matrix to give the most likely matching subject to the ordering constraint that later video clips cannot match to earlier instructional diagrams and vice versa. More formally, if video clip V_i matches to diagram I_j then clip V_{i+1} cannot match to diagram $I_{j'}$ with $j' < j$ and diagram I_{j+1} cannot match to video clip $V_{i'}$ with $i' < i$.

4. Ikea Assembly in the Wild Dataset (IAW)

In order to study the problem of understanding instructional videos, we collected a large well-labeled dataset called the Ikea assembly in-the-wild (IAW) dataset with annotations obtained using Amazon Mechanical Turk and a publicly available in-browser video annotation tool Vidat [43]. The IAW dataset contains 420 Ikea furniture pieces from 14 common categories, e.g., sofa, bed,

wardrobe, table, etc. Each piece of furniture comes with one or more user instruction manuals, which are first divided into pages and then further divided into independent steps cropped from each page (some pages contain more than one step and some pages do not contain instructions). There are 8,568 pages and 8,263 steps overall, on average 20.4 pages and 19.7 steps for each piece of furniture. We crawled YouTube to find videos corresponding to these instruction manuals and as such the conditions in the videos are diverse on many aspects, e.g., duration, resolution, first- or third-person view, camera pose, background environment, number of assemblers, etc. The IAW dataset contains 1,005 raw videos with a length of around 183 hours in total. Among them, approximately 114 hours of content are labeled as 15,649 actions to match the corresponding step in the corresponding manual.

The dataset is split into a train, validation, and test set (with 30,876 segments, 6,871 segments and 11,103 segments, respectively) by using a greedy algorithm to balance the distribution with respect to all attributes including viewpoint, indoor or not, camera motion and number of assemblers involved, and it is guaranteed that all video in both validation and testing sets are unseen in the training set.

5. Experiments

We evaluate our model on the IAW dataset for tasks of finding the best instructional diagram for a given video clip (video-to-diagram retrieval) and finding the top-k video clips corresponding to a given diagram (diagram-to-video retrieval). We consider two settings: independent retrieval where we are given one query video (resp. diagram) at a time, and set retrieval where we are given an entire video and corresponding instruction manual. The latter is the alignment problem and allows us to use structured inference methods such as optimal transport (OT) and dynamic time warping (DTW).

Preprocessing. We re-sample all videos to 30fps and then sub-sample into 10s segments to align with common practice of action recognition tasks. Video clips of duration 2.13s (64 frames) are used as input to the video encoder as shown in Fig. 4. The short side of each video frame is down-sampled to 224, maintaining aspect ratio. The long side of each instructional diagram is down-sampled to 224, also maintaining aspect ratio and padding the short side with white pixels. Random resize crops are used as data augmentation for videos, and random resize crops, horizontal flips and rotations are applied to instructional diagrams.

Architecture. We choose ResNet-50 [21] pretrained on ImageNet [11] as our backbone image encoder, and a ResNet-50 based Kinetics 400 [5] pretrained Slowfast-8x8 [14] for video encoding. For each 64-frame clip, 8 frames are uniformly sub-sampled for the slow path, and 32 frames for the fast path. We remove the classification heads from these

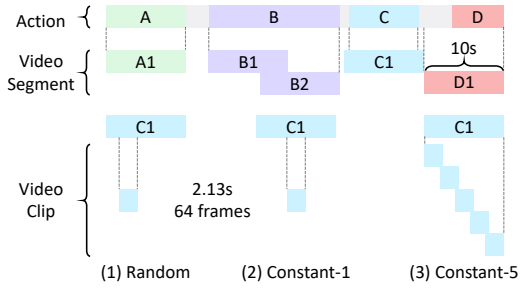


Figure 4. Demonstration of the video clip sampler for a task with four steps, A, B, C and D. Each step is sub-sampled to multiple 10s video segments (*e.g.* B $\rightarrow$ B1 and B2) with back padding. As input to the video encoder, each video segment is further divided into 2.13s video clips. We randomly sample one clips for training (1), choose a constant single clip for validation (2), and average over five clips for testing (3).

two backbone models and freeze the entire video encoder, but only first three layers of the image encoder allowing the later layers to be finetuned. The dimensionality of both video and diagram features is set to 1024.

Training Details. A dedicated learnable temperature parameter τ is assigned to each loss and initialized to 0.07 following [40]. The variance σ in intra-manual contrastive loss is initialized to 1 to represent a standard normal distribution. We use AdamW [24] as the optimizer with learning rate 5×10^{-4} and weight decay 5×10^{-3}. All models are trained for 20 epochs with 128 video clips per batch (and number of instructional diagram depending on the losses being used as described in Sec. 3). We select the model from the epoch with highest top-1 accuracy on the validation set for reporting test set results. It takes approximately 20 hours on a single Nvidia A100 GPU 80GB per experiment. During alignment testing, similarity scores are aggregated into a single $N \times M$ matrix for each video and corresponding instruction manual and optimal transport applied with hyper-parameters $\epsilon = 4$ and $\alpha = 7$.

Evaluation metrics. We report average top-1 accuracy and average index error (AIE) for the video-to-diagram retrieval task on the test set. AIE is useful for characterizing errors since predicting a step near to the ground truth is better than predicting one that is far away. It is defined as,

$$\text{AIE} = \frac{1}{N} \sum_{i=1}^{N} |j_i^\star - j_i^{\text{gt}}| \tag{14}$$

where $j_i^\star$ is the predicted diagram index for the i-th video and j_i^{gt} is its true index. Since a single instructional diagram can correspond to multiple video clips we adopt recall@1, recall@3 and area under the ROC curve as metrics for the diagram-to-video task. Unless otherwise stated we report

Table 1. Results comparing model alternatives. Performance on cropped *step* diagrams is denoted by S and entire *pages* from the manual by P. For fair comparison, the backbone for encoders is kept the same and only the loss and post-processing are varied. COSSIM uses cosine similarity loss and CLIP uses infoNCE loss on paired features. † AUROC values below 0.5 come from the fact that not every step or page diagram has a corresponding video segment in the test set, i.e., some queries have no positives.

Method	Video to diagram retrieval				Diagram to video retrieval					
	Top1 Acc.%↑		AIE↓		R@1↑		R@3↑		AUROC↑	
	S	P	S	P	S	P	S	P	S	P
Random	5.664	5.107	9.334	8.131	6.576	3.393	19.90	10.16	0.375	0.244
COSSIM	11.89	11.06	4.360	4.368	12.43	6.780	32.90	20.93	0.561	0.336
CLIP	19.61	19.05	4.274	4.180	16.94	10.25	38.67	23.45	0.590	0.373
Ours	28.62	34.55	3.734	2.928	22.30	16.48	45.00	32.20	0.617	0.390†
w/o SPRF	21.73	27.08	6.018	4.485	16.90	13.17	36.07	26.70	0.558	0.357
w/ DTW	31.45	36.20	**3.382**	**2.752**	23.20	17.32	32.45	17.55	0.467	0.310
w/ OT	**31.61**	**36.71**	3.458	2.816	**26.62**	**18.28**	**49.11**	**32.28**	**0.626**	**0.401**

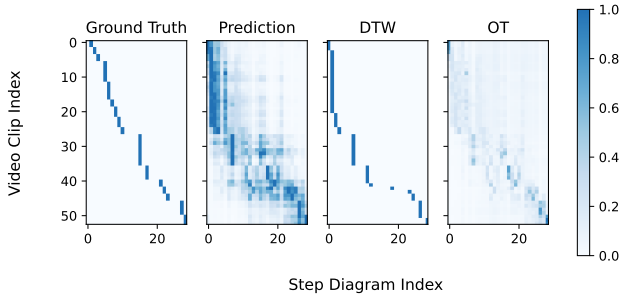


Figure 5. An example of post processing with DTW and OT with furniture 30341149 and video dzLNgz861Hk.

results at the video segment level. Here we sample consecutive 64-frame video clips (2.13s) from each 10s video segment and average the features from the video encoder.

5.1. Main Results

Our main results are reported in Tab. 1. We compare to two baseline methods, COSSIM and CLIP, which use a cosine similarity loss and infoNCE loss only on paired features. We report results on four variants of our approach using all three losses described in Sec. 3. The first ("Ours") includes the sinusoidal progress rate features (SPRF) capturing temporal information. Note that we also experimented with more standard positions encoding methods popular with transformers [33] but found these to produce inferior results (omitted here for brevity). The second variant ("w/o SPRF") shows results with this feature removed. The last two variants use dynamic time warping (DTW) and optimal transport (OT) in the alignment setting, i.e., with access to complete videos and instruction manuals.

Observe that our method significantly outperforms the baseline approaches on both video-to-diagram retrieval and diagram-to-video retrieval. This is largely due to our SPRF feature but also thanks to the improved loss functions (we provide a complete ablation analysis below). Further improvement in performance can be gained by post processing with DTW or OT. Interestingly, OT does slightly better than DTW indicating the the ordering constraint imposed by DTW is too restrictive for this task. See Fig. 5 for an example alignment.

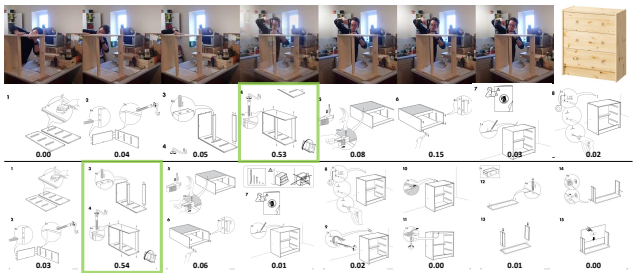
Qualitative results are shown in Fig. 6. We show one correct alignment and one incorrect alignment for matching to assembly steps. Notice the high degree of similarity between steps in the assembly process, which makes this an extremely challenging task. Further examples are included on the project website for this paper.

5.2. Effect of Losses

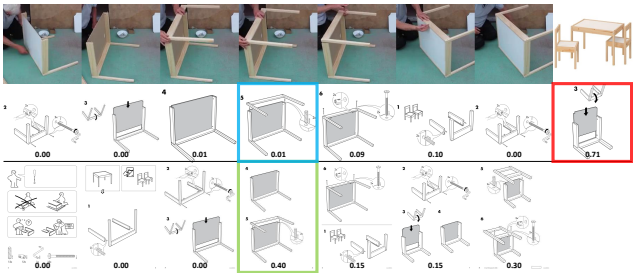
Our work introduces three novel loss terms for the task of aligning videos to step-by-step instructions. We now analyze the effectiveness of each loss by evaluating our model trained using different combinations. The results are summarized in Tab. 2. We can draw several conclusions from these results First, our video-diagram contrastive loss (A) slightly outperforms the standard infoNCE loss used by CLIP. This confirms our intuition that infoNCE is adversely affected by the many-to-one matchings between video clips and instructional diagrams albeit only slightly.

Second, the video-manual contrastive loss (B) gives the greatest boost in performance over the baseline approaches and once used gain little benefit from the video-diagram contrastive loss (A). The intra-manual contrastive loss (C) combined with the other losses slightly improves results.

Last, including losses on page diagrams even when evaluating on step diagrams improves results (but not vice versa). We hypothesize that this is because page diagrams provide a regularizing effect on learning since it is easier to match against pages than individual steps.



(a) Successful alignment between YouTube video moq_A1o3ZKw and Ikea furniture manual 60356219.



(b) A failed alignment between YouTube video d6sbVuHV0bc and Ikea furniture manual 10178413.

Figure 6. Qualitative results. Rows show frames from a single video clip; step instructional diagrams (subset shown); and page instructional diagrams. Prediction is highlighted by a green box for correct or a red box for incorrect; ground truth is then highlighted with a blue box.

Table 2. Ablation analysis on different loss combinations reported without post processing. Batches have 128 video clips (plus diagrams as required) except for those marked with † where we double the number of video clips to 256 (plus a single matching diagram). See Sec. 3.

Exp.	Loss A		Loss B		Loss C		Video to diagram retrieval				Diagram to video retrieval					
							Top1 Acc.%↑		AIE↓		R@1↑		R@3↑		AUROC↑	
	S	P	S	P	S	P	S	P	S	P	S	P	S	P	S	P
CosSim†							11.89	11.06	4.360	4.368	12.43	6.780	32.90	20.93	0.561	0.336
CLIP†							19.61	19.05	4.274	4.180	16.94	10.25	38.67	23.45	0.590	0.373
A1†	✓						20.87	15.28	3.991	4.635	17.92	8.650	41.00	22.17	0.592	0.358
A2†		✓					19.02	19.49	4.086	3.979	17.42	10.57	38.99	24.69	0.577	0.373
A3†	✓	✓					20.58	19.34	4.036	4.090	17.08	10.13	39.89	24.64	0.583	0.371
B1			✓				27.20	20.74	3.842	4.160	20.93	10.52	44.35	25.29	0.622	0.376
B2				✓			24.40	**35.07**	**3.672**	**2.883**	19.66	16.75	42.52	33.08	0.613	0.396
B3			✓	✓			28.20	34.59	3.789	2.991	21.02	16.64	44.43	31.93	0.618	0.393
C1	✓		✓				27.54	19.36	3.992	4.438	21.15	10.12	44.06	24.13	0.619	0.374
C2		✓		✓			24.50	34.91	3.702	2.998	19.79	**17.26**	42.62	**33.36**	0.612	**0.399**
C3	✓	✓	✓	✓			28.43	33.72	3.779	3.120	21.41	16.32	**45.06**	32.49	0.617	0.396
D1			✓	✓	✓	✓	**28.62**	34.55	3.734	2.928	**22.30**	16.48	45.00	32.20	0.617	0.390
D2	✓	✓	✓	✓	✓	✓	28.26	34.94	3.761	3.048	21.47	16.49	44.66	32.32	**0.620**	0.392

6. Conclusion

In this paper, we investigated the problem of aligning instructional videos with a high-level schematic representation of the task, depicted by abstract instructional diagrams showing the steps in the process. We proposed a method based on contrastive learning to align video and diagram features using three novel losses designed specifically for this task. Our focus is on Ikea furniture assembly where alignment is done between in-the-wild videos and the corresponding official assembly manuals. To this end, we also collected a dataset of 183 hours of in-the-wild assembly videos and nearly 8,300 diagrams. Two tasks are designed on this dataset to evaluate the performance of our method: (i) a nearest neighbor retrieval task between video clips and instructional diagrams, (ii) alignment of the instruction diagrams to their corresponding assembly video clips. On both tasks, experimental results show that our proposed sinusoidal progress rate feature and optimal transport modules lead to better temporal alignment and each one of the proposed losses enables the model to learn better representations, compared with compelling alternatives that do not take into account the unique nature of the problem.

Our work suggests several directions for future work. First, it would be interesting to consider including additional modalities such as video narrations into our framework. Second, extending the task to unsupervised or weakly supervised settings would overcome our current limitation of requiring ground truth alignments for learning. Last, an ambitious long-term goal is to develop applications, built on our alignment model, that automatically monitor and guide a user through an assembly process or facilitate robot-human collaboration on instructional tasks.

References

[1] 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

[2] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 847–859, 2021. 2

[3] Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5194–5205, 2022. 2

[4] Alessandro Bonardi, Stephen James, and Andrew J Davison. Learning one-shot imitation from humans without humans. *IEEE Robotics and Automation Letters*, 5(2):3533–3539, 2020. 1

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 2

[8] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021. 2

[9] John P Collomosse, Graham McNeill, and Yu Qian. Storyboard sketches for content based video retrieval. In *2009 IEEE 12th International Conference on Computer Vision*, pages 245–252. IEEE, 2009. 2

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 2

[11] 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

[12] Pelin Dogan, Boyang Li, Leonid Sigal, and Markus Gross. A neural multi-sequence alignment technique (neumatch). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8749–8758, 2018. 2

[13] Mark Everingham, Josef Sivic, and Andrew Zisserman. Hello! my name is... buffy"–automatic naming of characters in tv video. In *BMVC*, volume 2, page 6, 2006. 2

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 6

[15] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pages 214–229. Springer, 2020. 2

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2

[17] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 4

[18] Ning Han, Jingjing Chen, Guangyi Xiao, Hao Zhang, Yawen Zeng, and Hao Chen. Fine-grained cross-modal alignment network for text-video retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3826–3834, 2021. 2

[19] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2906–2916, 2022. 2

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[22] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11915–11925, 2021. 2

[23] YuXuan Liu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Imitation from observation: Learning to imitate behaviors from raw video via context translation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1118–1125. IEEE, 2018. 1

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[25] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 1, 2

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

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 4

[28] Tianjia Shao, Dongping Li, Yuliang Rong, Changxi Zheng, and Kun Zhou. Dynamic furniture modeling through assembly instructions. *ACM Transactions on Graphics*, 35(6), 2016. 2

[29] Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967. 5

[30] Makarand Tapaswi, Martin Bauml, and Rainer Stiefelhagen. Book2movie: Aligning video scenes with book chapters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1827–1835, 2015. 2

[31] Faraz Torabi, Garrett Warnell, and Peter Stone. Recent advances in imitation learning from observation. *arXiv preprint arXiv:1905.13566*, 2019. 1

[32] Sam Toyer, Anoop Cherian, Tengda Han, and Stephen Gould. Human pose forecasting via deep markov models. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2017. 2

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 7

[34] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2021. 4

[35] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. *arXiv preprint arXiv:2209.07526*, 2022. 2

[36] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 4

[37] Ruocheng Wang, Yunzhi Zhang, Jiayuan Mao, Chin-Yi Cheng, and Jiajun Wu. Translating a visual lego manual to a machine-executable plan. *arXiv preprint arXiv:2207.12572*, 2022. 2

[38] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5079–5088, 2021. 2

[39] Peng Wu, Xiangteng He, Mingqian Tang, Yiliang Lv, and Jing Liu. Hanet: Hierarchical alignment networks for video-text retrieval. In *Proceedings of the 29th ACM international conference on Multimedia*, pages 3518–3527, 2021. 2

[40] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 6

[41] Peng Xu, Kun Liu, Tao Xiang, Timothy M Hospedales, Zhanyu Ma, Jun Guo, and Yi-Zhe Song. Fine-grained instance-level sketch-based video retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):1995–2007, 2020. 2

[42] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022. 2

[43] Jiahao Zhang, Stephen Gould, and Itzik Ben-Shabat. Vidat—ANU CVML video annotation tool. https://github.com/anucvml/vidat, 2020. 5

[44] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2