Learning to Ground Instructional Articles in Videos through Narrations

Effrosyni Mavroudi[*], Triantafyllos Afouras[*], Lorenzo Torresani
Meta AI
{emavroudi, afourast, torresani}@meta.com

Abstract

In this paper we present an approach for localizing steps of procedural activities in narrated how-to videos. To deal with the scarcity of labeled data at scale, we source the step descriptions from a language knowledge base (wikiHow) containing instructional articles for a large variety of procedural tasks. Without any form of manual supervision, our model learns to temporally ground the steps of procedural articles in how-to videos by matching three modalities: frames, narrations, and step descriptions. Specifically, our method aligns steps to video by fusing information from two distinct pathways: i) direct *alignment of step descriptions to frames, ii)* indirect *alignment obtained by composing steps-to-narrations with narrations-to-video correspondences. Notably, our approach performs global temporal grounding of all steps in an article at once by exploiting order information, and is trained with step pseudo-labels which are iteratively refined and aggressively filtered. In order to validate our model we introduce a new benchmark – HT-Step – obtained by manually annotating a 124-hour subset of HowTo100M[1] with steps sourced from wikiHow articles. Experiments on this benchmark as well as zero-shot evaluations on CrossTask demonstrate that our multi-modality alignment yields dramatic gains over several baselines and prior works. Finally, we show that our inner module for matching narration-to-video outperforms by a large margin the state of the art on the HTM-Align narration-video alignment benchmark.*

1. Introduction

Instructional videos have emerged as a popular means for people to learn new skills and improve their abilities in executing complex procedural activities, such as cooking a recipe, performing home improvements, or fixing things. In addition to being useful teaching materials for humans, how-to videos are a promising medium for learning by machines, as they provide revealing visual demonstrations of complex activities and show elaborate human-object interactions in a variety of domains. Motivated by this observation, in this work we look at the task of temporally localizing the steps of procedural activities in instructional videos. This problem is foundational to the broader goal of human-procedure understanding and advances on this task promise to enable breakthrough applications, such as AI-powered skill coaching and human-to-robot imitation learning.

Prior work has tackled procedural step localization by leveraging either (a) fully-annotated datasets where the task

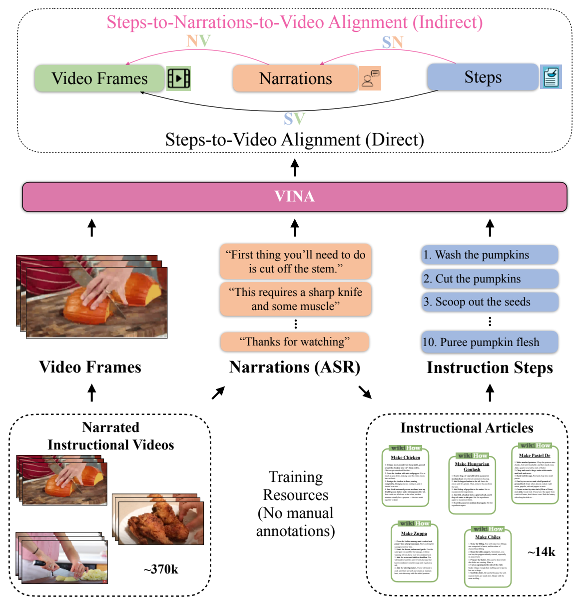

Figure 1: Our proposed Video, Instructions, and Narrations Aligner (VINA) learns to simultaneously ground narrations and instruction steps in how-to videos from an uncurated set of narrated videos and a separate knowledge base of instructional articles, *without any manual annotations*. This is contrary to prior work that learns how to align a video with a *single* sequence of sentences by leveraging *ground-truth pairs* of video-text sequences, e.g., a video and its narrations [22], or a video and an annotated, ordered list of steps demonstrated in it [16].

[*]Equal contribution
[1]A test server is accessible at https://eval.ai/web/challenges/challenge-page/2082.

shown in the video is given (*video-level labeling*) and manually annotated temporal segments are provided for each step (*segment-level labeling*) [59] or (b) weakly-annotated training sets where the task and the order in which the steps appear in the video is given [79]. However, due to the inherent manual cost involved in collecting step annotations, these works have relied on datasets that are small-scale both in the number of tasks (e.g., at most few hundreds [79]) and in the number of video samples (e.g., 12k videos [59]). These limitations affect both the generality and the complexity of the models that can be trained on these benchmarks. In this paper, we therefore pose the following question: *can we leverage large-scale, unlabeled video datasets to train a model that can ground procedural steps in how-to videos*?

To answer this question, we propose a novel training framework for weakly-supervised step grounding that utilizes two freely available sources of information: (a) instructional articles which define ordered lists of steps for a wide variety of tasks (e.g., from wikiHow) and (b) narrations which provide instance-specific rich commentaries of the execution of the task in the video, e.g., from ASR transcriptions. Our work treats the former as an abstraction of the latter and uses the video-specific narrations to support the grounding of the article steps. Specifically, *during training*, our method leverages narrations as an auxiliary signal to (i) identify the task shown in the video, (ii) temporally ground the article steps that are visually-demonstrated and (iii) filter out steps that are not executed in the given instance. To further motivate this mechanism, let us look at the example in Figure 1. The narrations help disambiguate the task (*make a pumpkin puree*), enabling the automatic retrieval of relevant instructional articles for the video. Furthermore, the narrations can be matched to steps described in the articles to roughly localize the steps that are represented in the video. In this example, the timestamp of "First thing you'll need to do is cut off the stem" provides a loose temporal prior for the matching step "Cut the pumpkins." On the other hand, steps that do not have any matching narrations (e.g., "Wash the pumpkins") are unlikely to be represented in the video and thus can be rejected. Based on this intuition, we propose a procedure that learns to align steps to video by fusing information from two pathways. The first is an *indirect* pathway inferring step-frame alignments by composing step-to-narration assignments with narration-to-frame correspondences. The second is a *direct* pathway that learns associations between step descriptions and frames by leveraging information from all videos having steps in common.

In our experiments we demonstrate that our multi-modality alignment leads to significant performance gains over several baselines, including single-pathway temporal grounding, as well adaptations of prior works to our problem. *During inference*, the direct pathway can be used by itself to temporally ground steps in absence of transcribed narrations. When narrations are available at test time, our method improves further the accuracy of temporal grounding by fusing the inference outputs of the two pathways.

To summarize, our work makes the following contributions: 1) we learn to align steps to frames in how-to videos, using only weak supervision in the form of noisy ASR narrations and instructional articles; 2) we propose a novel approach for joint dense temporal grounding of instructional steps and video narrations; 3) we introduce a new benchmark for evaluating instructional step grounding which we will make available to the community; 4) we demonstrate state-of the art results on multiple benchmarks for both step as well as narration grounding.

2. Related Work

Procedural step recognition. Going beyond classifying a trimmed segment video to a procedural step [9, 32, 41, 76], there has been active research in untrimmed video segmentation [10, 59, 72, 77], step anticipation [40, 47, 48], unsupervised step discovery [18, 19], procedure planning [6, 12, 74] and step description [68, 72]. In this work we focus on parsing untrimmed procedural videos into a sequence of step segments. Prior approaches can be categorized into two groups: *action localization-based* methods [3, 29, 50, 59] use fixed taxonomies of step classes and are not able to recognize unseen steps, while *grounding-based methods* [2, 5, 21], including ours, operate in an open-world setting, treating the problem as video-text alignment. Grounding-based approaches can be further subdivided into single step grounding, where single steps are queried over the whole video[27, 57] and dense temporal grounding methods [14, 22] where the objective is to jointly ground a sequence of steps or whole article into the video. Chen et al. [14] use video-level instructional step labels for (weak) supervision of a model that grounds instructional articles to videos. Their approach attempts to localize steps without using any narration information during training. Instead, we argue that narrations provide a much richer source of supervision for training step grounding models, while essentially coming for free.

Weakly-supervised approaches Existing methods also vary by the level of supervision used during training. One option is leveraging fully-annotated datasets with known temporal segments for each step [10, 26, 29, 50, 53, 60, 71, 80]. Other works use weakly-annotated training sets where the task and the order in which the steps appear in the video are known [7, 8, 11, 16, 20, 28, 45, 79], or only the task and potential steps are known [44], or only loose association between video and instructional articles is given [14, 17]. More related to our work are approaches that are trained using weak supervision coming for free from instructional video narrations [1, 20, 36, 49, 75]. Narrations and their timestamps have been used to discover procedural steps in videos and then align them with visual clusters [1, 51], to obtain temporal constraints for the location of steps in videos [20, 79], and to roughly localized steps by leveraging their textual similarity with ASR narrations [36]. In this work we utilize narrations, and in particular the estimated video-narration alignment, via multi-task learning and complementary inference pathways [73] for training a step grounding model.

Instructional articles have been recently used for learning strong video representations [32, 40] by replacing noisy ASR narrations with steps mined from instructional articles using their semantic similarity. Although we also associate steps from wikiHow articles to video frames through the use of narrations, our work differs from prior work in several aspects: we align steps to video by a global procedure that takes into account all ordered steps in the article (inspired by dense temporal grounding methods [5]) and temporally grounds them in the whole video, instead of matching individual video clips to an orderless collection of steps; our step grounding uses *video* in addition to steps and narrations while the method proposed in Lin et al. [32] relies purely on text-matching narrations to step descriptions; finally, the works differ in objectives with our aim being step grounding in long how-to videos rather than learning video-clip representations. In summary, our model is trained by combining supervision signals from video narrations and instructional articles, and can be applied for aligning procedural steps to videos with and without narrations.

Video-Text alignment The availability of large-scale video-text datasets such as HowTo100M has prompted many works on joint video-language embedding training [10, 24, 32, 33, 37, 46, 52, 70]. A form of contrastive loss is often adopted for bringing together the representations of the two modalities [4, 34, 35, 37, 42, 43, 66, 69], while masked objectives are also gaining popularity [15, 23, 31, 35, 56, 57, 58, 61, 78]. Some works perform end-to-end representation learning [37, 38], while others freeze representation and focus on longer-term temporal modelling, which aims to capture context [66]. Recently Han *et al.* investigated directly aligning contextualized narration representations to video frames [22]. We build our method off of this approach – we note however that our objective is complementary: rather than aligning a video's narrations as an end-goal, we use this functionality to ground a set of independent steps sourced from instructional articles; in that process we show that the synergy that develops while training jointly on the two tasks results in improved performance for both.

3. Narration-Aided Step Grounding

We first present our architecture for joint narration and step grounding (Sec. 3.2), followed by learning objectives (Sec. 3.3.1) and pseudo-labeling strategy (Sec. 3.3.3); we discuss inferring the video task in (Section 3.3.2).

3.1. Problem Formulation

Let $(\mathcal{V}, \mathcal{N})$ be a video-narration pair, consisting of T video frames and a sequence of N narrations. Also, let $\mathcal{S}$ be an ordered list of S steps from an instructional article for a candidate task τ. Our objective is to ground each step of $\mathcal{S}$ to the video, conditioned on the other steps and the ASR transcript[2]. In particular, the desired output of our model is an alignment matrix $Y^{SV} \in \{0,1\}^{S\times T}$, where $Y_{st} = 1$ only if frame t is depicting the s-th step of task τ, and zero otherwise. Note that some steps might not be represented in the video.

3.2. Joint Narration and Article Step Grounder

As shown in Figure 2, our proposed VINA model follows the popular paradigm of leveraging Transformers for modeling multimodal interactions [62, 67].

Unimodal Encoders. Before feeding the video, narrations and article steps to our model, we preprocess them to extract a sequence of tokens. Given a video-narration pair ($\mathcal{V}$, $\mathcal{N}$) we extract visual features, $X^v \in \mathbb{R}^{T\times D_v}$ and narration features $X^n \in \mathbb{R}^{N\times D_n}$ using standard backbone networks (e.g., a frozen S3D [65] network for visual features, and pooled Word2Vec [39] embeddings for narration features). Similarly, we encode the sequence of steps in a sequence of features $X^s \in \mathbb{R}^{S\times D_s}$. The features of each modality m are embedded into a common embedding space of dimensionality D using a Unimodal Encoder that consists of a modality-specific MLP network, and then learnable, modality-specific positional embeddings P^m are added to them:

$$H^m = MLP(X^m;\theta_m) + P^m, \tag{1}$$

where $m \in \{v, n, s\}$ denotes the modality.

Multimodal Encoder. The outputs of the Unimodal Encoders are concatenated into a sequence of tokens: $H = [H^v; H^n; H^s] \in \mathbb{R}^{(T+N+S)\times D}$ and fed to the Multimodal Encoder, which is a standard Transformer with multiple layers of multi-head self-attention:

$$Z = \text{Transformer}(H) \in \mathbb{R}^{(T+N+S)\times D}. \tag{2}$$

The contextualized embeddings $Z = [Z^v; Z^n; Z^s]$ computed by the Multimodal Encoder capture interactions within each modality (e.g., temporal relationships within the video and context among steps of an article) and across modalities. We can then compute cosine similarity matrices between all pairs of modalities: narrations-to-video $A^{NV} \in \mathbb{R}^{N\times T}$, steps-to-video $A^{SV} \in \mathbb{R}^{S\times T}$,

[2]ASR transcripts are assumed to be always available for training and optionally during inference.

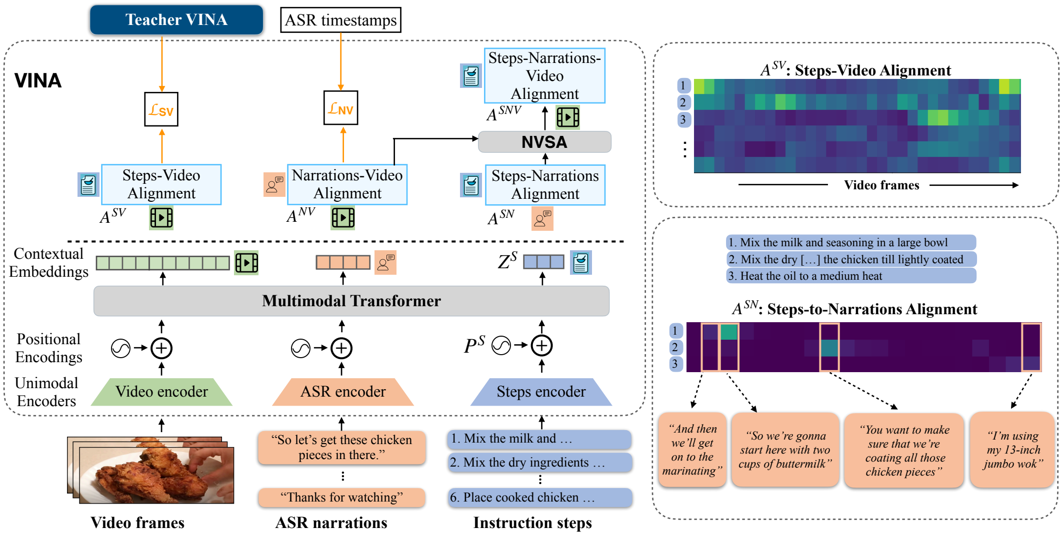

Figure 2: (*left*) Schematic illustration of our system. First, it extracts representations for each modality (video, ASR narrations, task steps) with three *Unimodal Encoders*. These representations are fed to a Transformer-based *Multimodal Encoder* for capturing interactions among video frames as well as among the textual modalities and the video. The contextualized embeddings for video frames, narrations and steps are used to compute correspondences between *all pairs of modalities*. Step grounding is achieved by fusing the output of two pathways: a *direct* pathway aligning steps to video (A^{SV}) and an *indirect* pathway that composes steps-to-narration (A^{SN}) with narration-to-video (A^{NV}) alignments to produce a second step-to-video alignment (A^{SNV}). We train our model using a teacher-student strategy with iteratively refined and filtered step pseudo-labels. (*right*) Qualitative examples of our learned steps-to-video alignment and steps-to-narrations alignment matrices for a video snippet.

and steps-to-narrations $A^{SN} \in \mathbb{R}^{S\times N}$. For example, the narrations-to-video similarity matrix A^{NV} is obtained by simply computing the cosine similarity between each frame embedding and each narration embedding: $A^{NV}_{nt} = {\mathbf{z}^n_n}^\top \mathbf{z}^v_t / (||\mathbf{z}^n_n||\ ||\mathbf{z}^v_t||)$.

Narration-aided Step Grounding. A straightforward inference path for temporally grounding the steps in the video is directly through the A^{SV} similarity matrix, which captures the similarity of each video frame with each instructional step. However, this alignment does not explicitly take into account the narrations of the video (only implicitly, through the Multimodal Transformer). We observe that an alternative way to ground steps in a video is to first identify narrations in the ASR transcript that are relevant to the step and then exploit the similarity of those narrations with video frames to get a loose prior over the step location. This is computed by combining the information captured in the steps-to-narrations and narrations-to-video alignment matrices A^{SN} and A^{NV}:

$$A^{SNV} = \tilde{A}^{SN} A^{NV} \in \mathbb{R}^{S\times T}, \tag{3}$$

where $\tilde{A}^{SN}$ is the predicted steps-to-narrations alignment matrix A^{SN} after being normalized with a softmax function with temperature ξ: $\tilde{A}^{SN}_{sn} = \frac{\exp{(A_{sn}/\xi)}}{\sum_{j=1}^{N}\exp(A_{sj}/\xi)}$.

The resulting A^{SV} and A^{SNV} alignment matrices provide two complementary inference paths to align steps to video frames. The mutual agreement between the direct steps-to-video alignment provided by A^{SV} and indirect, narration-based steps-to-video alignment provided by A^{SNV} can be used to better ground steps. Intuitively, if a frame is both very similar to a step in the joint embedding space learned by the Multimodal Transformer, and also very similar to a narration that is relevant to the step, then it is more likely to be indeed relevant to the step. Hence, we fuse the A^{SV} and A^{SNV} alignment matrices to a matrix $A^F = (A^{SV} + A^{SNV})/2$.

3.3. Weakly-Supervised Training from Narrated Instructional Videos

Next, we discuss how to supervise the VINA model in order to learn steps-to-video alignment and narrations-to-video alignment. We first present the training objective assuming that the ground-truth temporal segments for each narration and step in the video are given. Then we describe our approach for obtaining automatic pseudo-labels for the temporal segments.

3.3.1 Learning on Labeled Data

Let $\mathcal{B} = \{\mathcal{V}_i, \mathcal{N}_i, \mathcal{S}_i, Y_i^{NV}, Y_i^{SV}\}_{i=1}^{B}$ denote a set of training tuples, each comprising a video-narration pair, an ordered list of relevant task steps, and the target video-narrations and video-steps alignment matrices, we train the VINA model by optimising the following objective:

$$\mathcal{L} = \frac{1}{B}\Big[\sum_{i=1}^{B} \lambda_{NV}\mathcal{H}(Y_i^{NV}, A_i^{NV}) + \lambda_{SV}\mathcal{H}(Y_i^{SV}, A_i^{SV})\Big], \tag{4}$$

where $\mathcal{H}(\cdot,\cdot)$ is the modified InfoNCE loss used by [22] for aligning video with text using noisy ground-truth temporal segments:

$$\mathcal{H}(Y, A) = -\frac{1}{K}\sum_{k=1}^{K} \log \frac{\sum_t Y_{t,k} \exp{(A_{t,k}/\eta)}}{\sum_t \exp{(A_{t,k}/\eta)}}, \tag{5}$$

where η is a temperature constant. Note that although we do not explicitly supervise the steps-narrations alignment A^{SN}, meaningful alignments emerge during training due to the joint grounding of narrations and steps to the same video samples, as seen in Figure 2. Note that although we do not directly supervise the steps-to-narrations alignment, our model is able to learn meaningful correspondences, which go beyond simple pairwise textual matching.

3.3.2 Pairing Videos with Articles

We assume access to a set of instructional articles $\mathcal{A} = \{\mathcal{S}_j, \tau_j\}_{j=1}^{W}$, where τ_j denotes the article title and $\mathcal{S}_j$ the associated set of steps. To assign a set of steps to a given video from our training set $\mathcal{B}$ we need to associate it with an article from $\mathcal{A}$. If our video dataset provides metadata (e.g., a task id for every video), then this can be used to obtain the association – although there is no guarantee that this will result in the best article-match for the video (see discussion in supplementary materials for more analysis). If such metadata is not available, we can predict a task id, using the similarity between the narration and the titles of the available articles. To that end we use an off-the-shelf language model (e.g. MPNet [55]) to compute semantic embeddings of the ASR captions of every video and the title of each article $\tau_j \in \mathcal{W}$. For every video $\mathcal{V}_i$ we then calculate the semantic similarity between all the N captions in $\mathcal{N}_i$ and all task titles $\tau_j \in \mathcal{W}$, and assign N votes; the vote of every caption goes to the task that best matches it. Finally the video is assigned the task with the most votes. Alternatively, in order to obtain multiple sets of steps for a video, we rank the tasks by the number of votes.

3.3.3 Narration-Aided Pseudo-Labeling

Once a task τ_j has been associated with a video, we have access to a list of steps $\mathcal{S}_j$ from the article of the task. However, whether these steps appear in the video and their temporal location remain unknown. Inspired by self-labeling approaches from the SSL literature [30, 54, 64], we follow a teacher-student approach where a teacher version of our models generates pseudo-labeled temporal segments for training the student. For every step represented by a row in the learned steps-to-video alignment matrix we obtain a pseudo-ground truth segment by finding the maximal activation (peak) and expanding a temporal segment on both sides until the activation falls below an adaptive threshold ζ (e.g., 70% of the peak). To avoid training with unreliable pseudo-labels, we filter out pseudo-labels with low confidence: if the peak activation is below a fixed threshold γ, the alignment of that step is treated as unreliable for pseudo-labeling, and is altogether ignored.

Training curriculum. For the first E_b epochs we perform burn-in training of the student model on fixed pseudo-labels generated by feeding the video and the list of steps $\mathcal{S}_j$ to TAN [22], an off-the-shelf model pre-trained on the task of video-text alignment. Afterwards, we switch to using pseudo-labels generated from the teacher, where the teacher is initialized by duplicating the burn-in student model and then updated every ν epochs. During both stages, we utilize the original ASR timestamps for supervising the video-to-narrations alignment.

4. Experiments

4.1. Datasets and Metrics

We train our models on narrated videos from the HowTo100M dataset by leveraging the dataset release of wikiHow instructional articles [25], without using any form of manual annotations. In order to evaluate the effectiveness of our method, we evaluate: step grounding on HT-Step (a new benchmark, described below), narration alignment on HTM-Align [22], and zero-shot step localization on CrossTask [79]. A statistics summary of all the datasets used for training and evaluation is provided in Table 1.

HowTo100M (Training). The HowTo100M dataset [38] contains instructional videos from YouTube. Following Han *et al.* [22], we use the Food & Entertainment subset containing approximately $370K$ videos, where each video is complemented by the "sentencified" ASR transcription of its audio narration.

wikiHow (Training). We train using 14,541 cooking tasks from the wikiHow-Dataset [25]. For each task, we generate an ordered list of steps by extracting the step headlines.

CrossTask (Evaluation). We use this established instructional video benchmark for *zero-shot* grounding, i.e., by

Dataset	Step Annot.	# Videos	# Activities	# Steps	# Segments
HowTo100M [38]	✗	1.2M	25k	-	-
HTM-Align [22]	✗	80	80	-	-
CrossTask [79]	✓	4.8k (2.8k)	83 (18)	133	20.9k
HT-Step (val)	✓	600	120	1,204	3,441
HT-Step (test)	✓	600	120	1,242	3,631
wikiHow		-	14k	100k	-

Table 1: Summary statistics for the datasets in our work. For CrossTask, the statistics for primary activities only are shown in parentheses.

directly evaluating on CrossTask our model learned from HowTo100M. Following common practices, we use two evaluation protocols: the first one – *step localization* – aims at predicting a single timestamp for each occurring step in videos from 18 primary tasks [79]. Performance is evaluated by computing the recall (denoted as Avg. R@1) of the most confident prediction for each task and averaging the results over all query steps in a video, where R@1 measures whether the predicted timestamp for a step falls within the ground truth boundaries. We report average results over 20 random sets of 1850 videos [79]. The second task – *article grounding* – requires predicting temporal segments for each step of an instructional article describing the task represented in the video. We use the mapping between CrossTask and *simplified* wikiHow article steps provided in Chen et al. [14] and report results on 2407 videos of 15 primary tasks obtained excluding three primary tasks following the protocol of [14] (see supplementary materials for details). Performance for this task is measured with Recall@K at different IoU thresholds [14].

HTM-Align (Evaluation). This benchmark is used to evaluate our model on narration grounding. It contains 80 videos where the ASR transcriptions have been manually aligned temporally with the video. We report the R@1 metric [22], which evaluates whether the model can correctly localize the narrations that are alignable with the video.

HT-Step (Evaluation). To evaluate the effectiveness of our model in grounding article steps, we introduce an evaluation benchmark consisting of 1200 HowTo100M narrated videos spanning a total of 216 unique tasks, with each video manually annotated with temporal segments for each occurring step. For each video, annotators were provided with the task name (e.g., Make Pumpkin Puree) and candidate recipe steps from the corresponding wikiHow article. We refer the reader to supplementary materials for details about the data annotation. We split the annotated videos into a validation and a test set, each containing 600 videos, with 5 videos per task. We ensure that our validation set does not contain videos from HTM-Align.

4.2. Implementation Details

As video encoder we adopt the S3D [65] backbone pretrained with the MIL-NCE objective on HowTo100M [37]. Following previous work [22, 66], we keep this module frozen and use it to extract clip-level features (one feature per second for video decoded at 16 fps). For extracting context-aware features for each sentence (step or narration), we follow the Bag-of-word (BoW) approach based on Word2Vec embeddings [39]. Our methods hyperparameters were selected on the HT-Step validation set and are: $\lambda_{SV} = \lambda_{NV} = 1$, temperatures $\eta, \xi = 0.07$, and threshold $\gamma = 0.65$. We train our model for 12 epochs, with 3 epochs burn-in and then we update the teacher every 3 epochs. Pseudo-labels are obtained based on the steps-to-video alignment matrix. To obtain temporal segment detections from the step-to-video alignment output of our model (e.g. for evaluating on the CrossTask article grounding setting) we use a simple 1D blob detector [63]. Unless otherwise specified, we use the fused alignment matrix for step grounding when narrations are available during inference time. More details are included in supplementary materials.

Method	Train. Inp.	HT-Step ↑R@1		HTM-Align ↑R@1
		w/o nar.	w/ nar.	
CLIP (ViT-B/32) [43]	-	-	-	23.4
MIL-NCE [37]	N	<u>30.7</u>	-	34.2
TAN (Joint+Dual, S2) [22]	N	-	-	<u>49.4</u>
TAN* (Joint, S1, LC) [22]	N	31.2	-	47.1
TAN* (Joint, S1, PE+LC) [22]	N	7.9	-	63.0
Ours	N+S	**35.6** $\pm$ 0.4	**37.4** $\pm$ 0.4	**66.5** $\pm$ 0.9

Table 2: **Comparison with state-of-the-art methods for step and narration grounding.** We report results on the HT-Step and HTM-Align test sets, respectively. TAN* refers to our improved baselines of [22]. S1 and S2 refer to the training stages followed in [22]. PE denotes the addition of positional encoding to the output of the narration encoder. LC denotes long context, *i.e.*, our improved TAN* baseline using 1024 seconds of context as opposed to 64 for TAN. Previous best results are shown underlined. Our VINA results are reported after 5 random runs. VINA clearly outperforms all previous work – as well as our improved TAN baselines – by large margins on both narration alignment and step grounding.

4.3. Results

4.3.1 Comparison with the State of the Art

Weakly-Supervised Narration and Step Grounding. Table 2 compares the step and narration grounding performance of our method with recent state-of-the-art video-text alignment methods trained on HowTo100M using ASR narrations: MIL-NCE [37] and TAN [22]. When using them for narration alignment, we feed them with ASR as input.

But we also evaluate them as strong baselines for zero-shot step grounding by feeding them with the sequence of steps as input. Our model achieves 66.5% R@1 on narration alignment on HTM-Align, leading to an absolute improvement of 17.1% over the previously reported state-of-the-art (49.4%). Notably, on HTM-Align our method surpasses *TAN* (Joint, S1, PE, LC)* which is a new version of TAN [22] implemented by us and much stronger in video-narration alignment. Our re-implementation uses positional encodings for ASR narrations, is trained on long-form videos (up to 17 minutes) only with original ASR timestamps, while TAN was trained on 1 min video-clips with refined narration timestamps and used a fusion of two models during inference (Joint+Dual). Our method also outperforms all baselines for step grounding on HT-Step even when seeing only steps during inference, while being trained with (video, narrations, steps) triplets. It also outperforms TAN* (Joint, S1, LC), which is a second re-implementation of TAN designed for maximum performance on the task of step grounding. Additionally, VINA is able to use ASR transcripts of videos during test time, if available, to further boost the performance.

Method	↑Avg. R@1 (%)
Supervised	
TempCLR [70]	52.5
Zero-Shot	
Zhukov [79]	22.4
HT100M [38]	33.6
VideoCLIP [66]	33.9
MCN [13]	35.1
DWSA [51]	35.3
MIL-NCE [37]	40.5
VT-TWINS* [24]	40.7
UniVL [33]	42.0
Ours w/o nar.	**44.1**
Ours w/ nar.	**44.8**

Table 3: **Comparison with state-of-the-art methods for zero-shot action step localization on the CrossTask dataset**. The performance of the state-of-the-art fully-supervised method (TempCLR [70]) is reported as an upper-bound to the zero-shot approaches. * denotes results reported on different test splits, and hence not directly comparable with the rest. Our model outperforms all previous works by a clear margin (2.1% absolute improvement over the previous best result on the standard split). When providing narrations as additional inputs during inference (only text, not the timings), we obtain a further 0.7% boost.

Step localization on CrossTask. In Table 3 we compare our model against the state-of-the art in step localization on the CrossTask benchmark. Our approach sets a new state-of-the-art for zero-shot step localization on this challenging benchmark. Importantly, most approaches are evaluated on this dataset by feeding their predicted steps-to-frames alignment matrix to a dynamic programming algorithm which finds the optimal assignment of each step with exactly one short clip *assuming a canonical, fixed ordering of steps for each task*. In contrast, our method, which is naturally aware of context and ordering by densely grounding steps, can outperform prior results without imposing any constraints during inference.

Model	↑R@50(IOU)			↑R@100(IOU)		
	0.1	0.3	0.5	0.1	0.3	0.5
MIL-NCE-max [37]	33.5	12.0	4.9	39.7	14.3	5.9
MIL-NCE-avg [37]	42.9	24.3	12.9	56.8	32.1	17.0
WSAG [14]	40.1	23.1	10.1	54.3	31.3	14.0
Ours	**87.1**	**59.0**	**30.0**	**90.6**	**61.1**	**30.9**

Table 4: **Comparison with state-of-the-art approaches for article grounding on the CrossTask dataset.**

Article grounding on CrossTask. VINA is robust to the type of language in which task steps are described. It can handle both atomic phrases (as demonstrated by our results on step localization on CrossTask), but also rich, natural language step descriptions, as evidenced by performance on HT-Step. To further demonstrate this, we compare against the state-of-the art on the article grounding task of CrossTask in Table 4. Our model outperforms all previous works by a large margin. We emphasize the performance improvement we obtain compared to WSAG, which highlights the importance of exploiting the narration information for training.

4.3.2 Ablation Studies

We perform ablations to assess the impact of the various design choices in our method by measuring step grounding performance on the HT-Step validation set and video-narration alignment performance on HTM-Align.

Effect of weak supervision from instructional articles. Row 3 in Table 5 shows the step grounding results obtained from an instance of our model that includes only the direct video-step alignment pathway and that is trained just on wikiHow steps (without narrations) using the fixed step pseudo-labels from TAN* [22] without any form of iterative pseudo-labeling (row 1). Remarkably, this variant improves by 3.3% over the step-grounding performance of TAN*. When we let this variant update the step pseudo-labels (row 4), the recall improves further (5.1% over TAN*). These results provide evidence of the strong benefits of utilizing instructional articles for the learning of step grounding.

Method	Train. Inp.	Iter. Pseudo.	HT-Step ↑R@1		HTM-Align
			w/o nar.	w/ nar.	
Baseline/Initial Step Pseudo-labels					
(1) TAN Joint S1	N		30.7	-	47.1
Single-Task Training					
(2) Ours	N		-	-	63.2
(3) Ours	S		34.0	-	-
(4) Ours	S	✓	35.8	-	-
Multi-Task Training					
(5) Ours	N+S		34.3	36.1	64.8
(6) Ours	N+S	✓	**36.9**	**39.1**	**67.0**

Table 5: **Ablation of main components of our framework**. We study the contribution of (a) multi-task training for narration and step grounding, (b) iterative step pseudo-labeling (*Iter. Pseudo*), and (c) narration-aware step grounding (*w/ nar.*). We report results on the HT-Step val set for STG and HTM-Align for NG. We compare training only with narrations (N), only with wikiHow steps (S), and training with narrations-steps sequence pairs (N+S). We also compare the performance with and without providing narrations during inference.

Effect of multimodal training and inference. Training our model with multi-modal textual inputs (steps and narrations), we observe an improvement of 1.6% in narration grounding (row 5 of Table 5) compared to its single-task counterpart (row 2). However the gain in step grounding is marginal when seeing only video frames during inference (*w/o nar.*, 34% in row 3 vs 34.3% in row 5). Our conjecture is that the missing modality (narrations) leads to some drop in performance. Providing both steps and narrations during inference leads to a stronger step grounding performance, which surpasses the TAN* baseline by 5.4% (30.7 → 36.1).

Effect of iterative pseudo-labeling. By comparing row 5 to row 6 of Table 5 we observe a clear boost in performance on both step grounding and narration alignment. This is a clear indication of the gains produced by iteratively refining the pseudo-labels using our model as a teacher during training.

Alignment	S → V	S → N	N → V	HT-Step ↑R@1
S → V	learned	-	-	34.3
S → N → V	-	learned	learned	30.5
S → N → V	-	learned	ASR	27.9
S → N → V	-	MPNet [55]	ASR	19.0
Fused	learned	learned	learned	**36.1**

Table 6: **Impact of the alignment matrix used during inference with narrations**. The same model is used for all results (corresponding to row 5 in Table 5).

Impact of pathways during inference. In Table 6 we study the effects of using different pathways and alignment information during inference. All results are produced

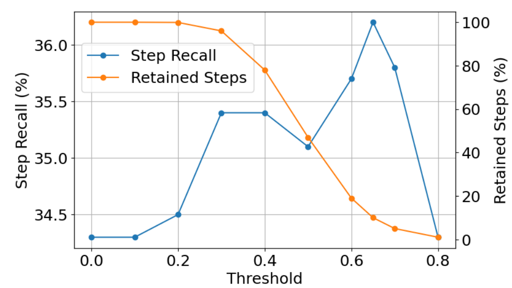

Figure 3: Step grounding performance and minimum retained step ratio as a function of the filtering threshold. Aggressive unreliable pseudo-label filtering with high confidence thresholds (large maximum step discard ratio) leads to improved performance.

from the same model trained for joint narration and step grounding with fixed pseudo-labels from TAN (row 5 in Table 5). Grounding steps using the indirect steps-to-video alignment only lags by 3.8% behind the direct steps-to-video alignment that directly computes the similarity between steps and video frames (30.5% vs 34.3%). Their fusion outperforms their individual grounding performance. This suggests that they capture complementary information. We also explore substituting our learned steps-to-narrations alignment with an alignment computed with an off-the-shelf language model. This significantly degrades performance (19.0) showing that our joint steps and narrations grounding model learns relationships between steps and narrations that go beyond textual similarity between pairs of sentences. Similarly, substituting our learned narrations-to-video alignment with an alignment based on the original ASR timestamps reduces performance by 2.6% (see additional details in supplementary materials).

Impact of pseudo-label filtering threshold. In Figure 3 we study the effect of the filtering threshold used in the iterative pseudo-labeling stage. We can observe that using aggressive filtering (i.e., high thresholds translating to a high maximum percentage of pseudo-labels that are discarded) is key to observing gains from iterative pseudo-labeling compared to training with fixed pseudo-labels from TAN. Intuitively, a large percentage of steps described in wikiHow articles are not represented in the given instructional video due to task mismatch, variations in recipe execution, and some steps being optional. Therefore, starting with a small subset of reliable pseudo-labels can facilitate step grounding.

Task selection. In Table 7 we investigate different strategies to select the wikiHow articles during training. This selection determines the set of steps to be grounded. We evaluate two strategies for video-task association from narrations and compare them with using the task id provided in

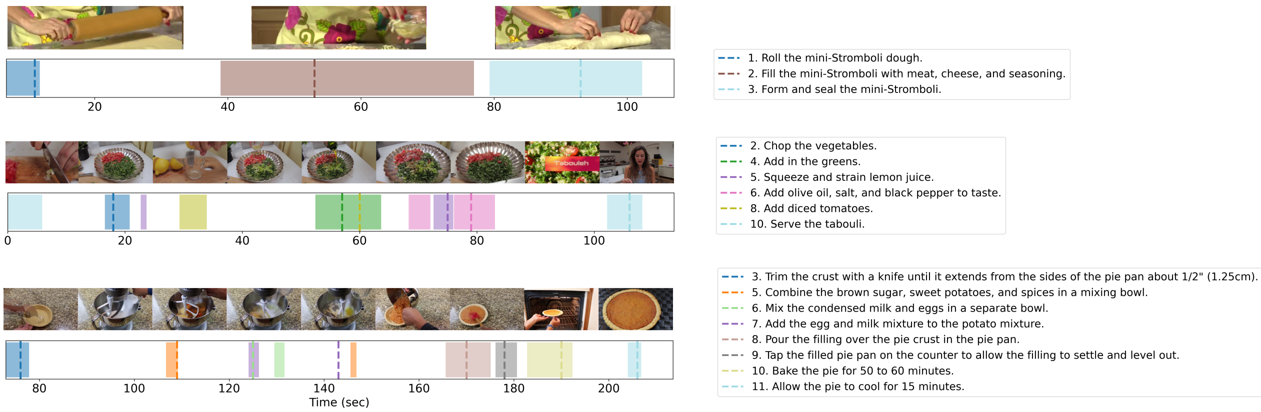

Figure 4: Visualization of sample frames per step, ground-truth segments (*colored bars*) and a single predicted timestamp per step (*dashed lines*) for the groundable steps of three HT-step validation videos demonstrating (a) a *Mini-Stromboli* recipe, (b) a *Make Tabouli* recipe, and (c) a *Bake a Sweet Potato Pie* recipe. Best viewed in color.

Task ID selection	**HT-Step ↑R@1**
HT100M metadata	34.3
Top-1 prediction	34.7
Random / top-5 pred	34.3

Table 7: **Sensitivity to task id selection.** We assess how the performance of our method changes when using different strategies to associate videos with articles. We experiment with using the task ids available from HT100M, as well as the the two predictive strategies presented in Section 3.3.2. We conclude that our method is robust to the task selection, and the task labels are not necessary for training.

the HowTo100M metadata for each video[3]. We see that our automatic selection approaches yield results on par with or even slightly better than those based on metadata.

4.3.3 Qualitative Results

Qualitative grounding results obtained with the proposed VINA model are presented in Figure 4. These examples showcase that our model can successfully retrieve complex and fine-grained steps, such as combining multiple ingredients in a mixing bowl or tapping a filled pie pan for the *Bake a Sweet Potato Pie* recipe. Failure cases include discriminating between steps that involve the same ingredients, such as mixing the milk and eggs versus adding the mixture to the existing mix.

5. Conclusion

We have presented a method for learning how to temporally ground sequences of steps in instructional videos, without any manual supervision. Our proposed method exploits the weak supervision naturally provided in such videos through their narrations, and solves for joint alignment of narrations and steps, while fusing two complementary pathways for step-to-video alignment. We demonstrated strong quantitative performance, surpassing the state-of-the-art on multiple benchmarks for both narration and step grounding.

Acknowledgements. We thank Huiyu Wang, Yale Song, Mandy Toh, and Tengda Han for helpful discussions.

References

[1] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4575–4583, 2016. 2, 3

[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

[3] Kumar Ashutosh, Santhosh Kumar Ramakrishnan, Triantafyllos Afouras, and Kristen Grauman. Video-mined task graphs for keystep recognition in instructional videos. *arXiv preprint arXiv:2307.08763*, 2023. 2

[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. 3

[5] Peijun Bao, Qian Zheng, and Yadong Mu. Dense events grounding in video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 920–928, 2021. 2, 3

[6] Jing Bi, Jiebo Luo, and Chenliang Xu. Procedure planning

[3] During inference, metadata task ids are used in all of our HowTo100M experiments in order to evaluate against the ground-truth step annotations.

in instructional videos via contextual modeling and model-based policy learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15611–15620, 2021. 2

[7] Piotr Bojanowski, Rémi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Weakly supervised action labeling in videos under ordering constraints. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 628–643. Springer, 2014. 2

[8] Piotr Bojanowski, Rémi Lajugie, Edouard Grave, Francis Bach, Ivan Laptev, Jean Ponce, and Cordelia Schmid. Weakly-supervised alignment of video with text. In *Proceedings of the IEEE international conference on computer vision*, pages 4462–4470, 2015. 2

[9] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10615–10624, 2020. 2

[10] Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. Locvtp: Video-text pre-training for temporal localization. In *European Conference on Computer Vision*, 2022. 2, 3

[11] C. Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3541–3550, 2019. 2

[12] Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *European Conference on Computer Vision*, pages 334–350. Springer, 2020. 2

[13] Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, Rameswar Panda, Brian Kingsbury, Rogerio Feris, David Harwath, et al. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8012–8021, 2021. 7

[14] Long Chen, Yulei Niu, Brian Chen, Xudong Lin, Guangxing Han, Christopher Thomas, Hammad Ayyubi, Heng Ji, and Shih-Fu Chang. Weakly-supervised temporal article grounding. In *Conference on Empirical Methods in Natural Language Processing*, pages 9402–9413, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. 2, 6, 7

[15] 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. 3

[16] Nikita Dvornik, Isma Hadji, Konstantinos G. Derpanis, Animesh Garg, and Allan D. Jepson. Drop-dtw: Aligning common signal between sequences while dropping outliers. *ArXiv*, abs/2108.11996, 2021. 1, 2

[17] Nikita Dvornik, Isma Hadji, Hai Pham, Dhaivat Bhatt, Brais Martinez, Afsaneh Fazly, and Allan D Jepson. Flow graph to video grounding for weakly-supervised multi-step localization. In *European Conference on Computer Vision*, pages 319–335. Springer, 2022. 2

[18] Ehsan Elhamifar and Dat Huynh. Self-supervised multi-task procedure learning from instructional videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 557–573. Springer, 2020. 2

[19] Ehsan Elhamifar and Zwe Naing. Unsupervised procedure learning via joint dynamic summarization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6341–6350, 2019. 2

[20] Daniel Fried, Jean-Baptiste Alayrac, Phil Blunsom, Chris Dyer, Stephen Clark, and Aida Nematzadeh. Learning to segment actions from observation and narration. In *Annual Meeting of the Association for Computational Linguistics*, pages 2569–2588, Online, July 2020. Association for Computational Linguistics. 2, 3

[21] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. *Advances in neural information processing systems*, 33:22605–22618, 2020. 2

[22] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2906–2916, June 2022. 1, 2, 3, 5, 6, 7

[23] Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2485–2494, 2019. 3

[24] Dohwan Ko, Joonmyung Choi, Juyeon Ko, Shinyeong Noh, Kyoung-Woon On, Eun-Sol Kim, and Hyunwoo J Kim. Video-text representation learning via differentiable weak temporal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5016–5025, 2022. 3, 7

[25] Mahnaz Koupaee and William Yang Wang. Wikihow: A large scale text summarization dataset. *ArXiv*, abs/1810.09305, 2018. 5

[26] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014. 2

[27] Hilde Kuehne, Juergen Gall, and Thomas Serre. An end-to-end generative framework for video segmentation and recognition. In *Proc. IEEE Winter Applications of Computer Vision Conference (WACV 16)*, Lake Placid, Mar 2016. 2

[28] Hilde Kuehne, Alexander Richard, and Juergen Gall. Weakly supervised learning of actions from transcripts. *Computer Vision and Image Understanding*, 163:78–89, 2017. 2

[29] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks: A unified approach to action segmentation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 47–54. Springer,

2016. 2
[30] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 5
[31] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. HERO: Hierarchical encoder for Video+Language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, Online, Nov. 2020. Association for Computational Linguistics. 3
[32] Xudong Lin, Fabio Petroni, Gedas Bertasius, Marcus Rohrbach, Shih-Fu Chang, and Lorenzo Torresani. Learning to recognize procedural activities with distant supervision. *arXiv preprint arXiv:2201.10990*, 2022. 2, 3
[33] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 3, 7
[34] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-CLIP:: End-to-end multi-grained contrastive learning for video-text retrieval. *arXiv preprint arXiv:2207.07285*, 2022. 3
[35] Yue Ma, Tianyu Yang, Yin Shan, and Xiu Li. Simvtp: Simple video text pre-training with masked autoencoders. *arXiv preprint arXiv:2212.03490*, 2022. 3
[36] Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nicholas Johnston, Andrew Rabinovich, and Kevin Murphy. What's cookin'? interpreting cooking videos using text, speech and vision. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 143–152, Denver, Colorado, May–June 2015. Association for Computational Linguistics. 2, 3
[37] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *CVPR*, 2020. 3, 6, 7
[38] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 3, 5, 6, 7
[39] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. 3, 6
[40] Medhini Narasimhan, Arsha Nagrani, Chen Sun, Michael Rubinstein, Trevor Darrell, Anna Rohrbach, and Cordelia Schmid. Tl;dw? summarizing instructional videos with task relevance and cross-modal saliency. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 540–557, Cham, 2022. Springer Nature Switzerland. 2, 3
[41] Medhini Narasimhan, Licheng Yu, Sean Bell, Ning Zhang, and Trevor Darrell. Learning and verification of task structure in instructional videos. *arXiv preprint arXiv:2303.13519*, 2023. 2
[42] Yookoon Park, Mahmoud Azab, Seungwhan Moon, Bo Xiong, Florian Metze, Gourab Kundu, and Kirmani Ahmed. Normalized contrastive learning for text-video retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 248–260, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. 3
[43] 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. 3, 6
[44] Alexander Richard, Hilde Kuehne, and Juergen Gall. Action sets: Weakly supervised action segmentation without ordering constraints. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5987–5996, 2018. 2
[45] Alexander Richard, Hilde Kuehne, Ahsan Iqbal, and Juergen Gall. Neuralnetwork-viterbi: A framework for weakly supervised video learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7386–7395, 2018. 2
[46] Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, et al. Avlnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199*, 2020. 3
[47] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 2
[48] Fadime Sener and Angela Yao. Zero-shot anticipation for instructional activities. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 862–871, 2019. 2
[49] Ozan Sener, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Unsupervised semantic parsing of video collections. In *Proceedings of the IEEE International conference on Computer Vision*, pages 4480–4488, 2015. 2
[50] Yuhan Shen and Ehsan Elhamifar. Semi-weakly-supervised learning of complex actions from instructional task videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3344–3354, 2022. 2
[51] Yuhan Shen, Lu Wang, and Ehsan Elhamifar. Learning to segment actions from visual and language instructions via differentiable weak sequence alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10156–10165, 2021. 3, 7
[52] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20020–20029, 2022. 3

[53] Bharat Singh, Tim K Marks, Michael Jones, Oncel Tuzel, and Ming Shao. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1961–1970, 2016. 2

[54] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 5

[55] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*, 2020. 5, 8

[56] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020. 3

[57] Chen Sun, Austin Myers, Carl Vondrick, Kevin P. Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7463–7472, 2019. 2, 3

[58] 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. 3

[59] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[60] Yansong Tang, Jiwen Lu, and Jie Zhou. Comprehensive instructional video analysis: The coin dataset and performance evaluation. *IEEE transactions on pattern analysis and machine intelligence*, 43(9):3138–3153, 2020. 2

[61] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022. 3

[62] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics. 3

[63] Huiyu Wang, Mitesh Kumar Singh, and Lorenzo Torresani. Ego-only: Egocentric action detection without exocentric pretraining, 2023. 6

[64] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4248–4257, 2022. 5

[65] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *European Conference on Computer Vision*, 2018. 3, 6

[66] 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 (EMNLP)*, Online, Nov. 2021. Association for Computational Linguistics. 3, 6, 7

[67] Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal learning with transformers: A survey, 2022. 3

[68] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10714–10726, 2023. 2

[69] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11542–11552, 2021. 3

[70] Yuncong Yang, Jiawei Ma, Shiyuan Huang, Long Chen, Xudong Lin, Guangxing Han, and Shih-Fu Chang. Temporal alignment representation with contrastive learning. *arXiv preprint arXiv:2212.13738*, 2022. 3, 7

[71] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. In *The British Machine Vision Conference (BMVC)*, 2021. 2

[72] Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23056–23065, 2023. 2

[73] Amir R Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J Guibas. Robust learning through cross-task consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11197–11206, 2020. 3

[74] He Zhao, Isma Hadji, Nikita Dvornik, Konstantinos G Derpanis, Richard P Wildes, and Allan D Jepson. P3iv: Probabilistic procedure planning from instructional videos with weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2938–2948, 2022. 2

[75] Yiwu Zhong, Licheng Yu, Yang Bai, Shangwen Li, Xueting Yan, and Yin Li. Learning procedure-aware video representation from instructional videos and their narrations. *arXiv preprint arXiv:2303.17839*, 2023. 2

[76] Honglu Zhou, Roberto Martín-Martín, Mubbasir Kapadia, Silvio Savarese, and Juan Carlos Niebles. Procedure-aware pretraining for instructional video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10727–10738, 2023. 2

[77] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[78] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. pages 8743–8752, 06 2020. 3
[79] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3537–3545, 2019. 2, 3, 5, 6, 7
[80] Andrea Ziani, Zicong Fan, Muhammed Kocabas, Sammy Christen, and Otmar Hilliges. Tempclr: Reconstructing hands via time-coherent contrastive learning. In *International Conference on 3D Vision (3DV)*, 2022. 2