Supplementary material for: HT-Step: Aligning Instructional Articles with How-To Videos

Anonymous Author(s)
Affiliation
Address
`email`

This supplementary material provides more comprehensive statistics about the dataset (Section C), architecture and implementation details for the baselines discussed (Section E), extra details about the evaluation protocol (Section G), a more detailed overview of the activity and step taxonomy (Section D), a breakdown of model performance by activity (Section F), and a datasheet containing detailed documentation about HT-Step (Section I).

A Supplementary .zip contents

This supplementary material contains the following files:

- **taxonomy.csv:** CSV file with the full step taxonomy in. For every step we list its corresponding activity (and its variation if applicable), headline and paragraph.
- **annotation_sample.json:** Sample annotations for 100 videos from our training split. Refer to section I.2 in our Datasheet for detailed explanation of the annotations format.
- **step_examples.mp4:**Video with step examples from various activities and videos in our dataset.
- **1_gt_annots_-2vbgoZc-RI.mp4, 2_gt_annots_4Xifp6umIZc.mp4, 3_gt_annots_-2maV1TTL5U.mp4:** Example videos with visualization of step annotations.

B Video examples

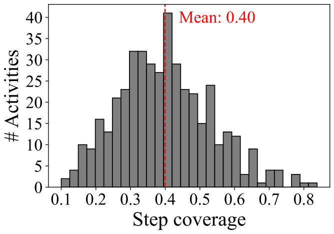
We have included three videos from the HT-Step training set with overlaid step annotations. In the beginning of each video, we indicate its activity, variation and the article steps.

- Videos `-2vbgoZc-RI` and `4Xifp6umIZc`: These are videos showing two variations of the same activity: *Make Eggless Cookies*. FM indicates full match, while PM partial match. Note that there are non-groundable steps.
- Video `-2maV1TTL5U`: This is a video showing how to *Make Bollilos*. Note that there are non-groundable steps, like *9. Fold each roll into thirds*. Also, the video showcases an example of a composite step that is partially groundable in each temporal segment. The step *Punch down and knead the dough* has multiple relevant temporal segments. One of them just shows the sub-step of punching down the dough, while a later one shows the kneading.

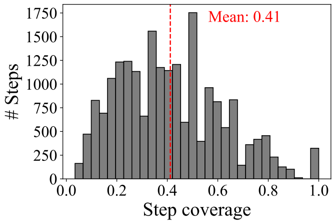
Submitted to the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks. Do not distribute.

C Extra statistics

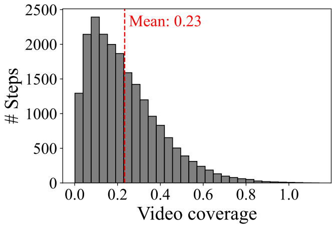
Figures 1a-1i contain additional statistics about HT-Step.



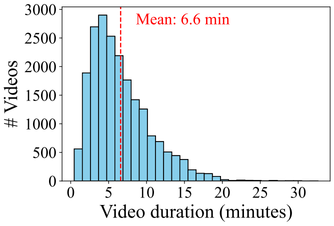
(a) **Distribution of step coverage across activities.**



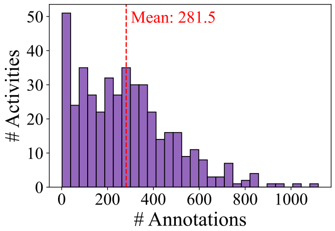
(b) **Distribution of step coverage across videos.**



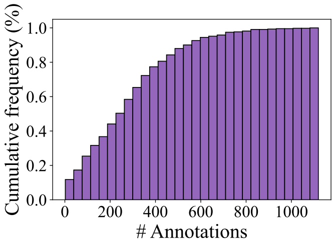
(c) **Distribution of annotated video durations.**



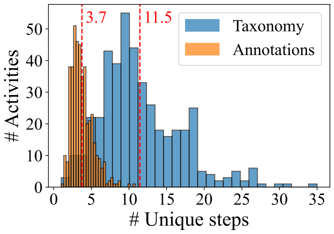
(d) **Distribution of step coverage across activities.**



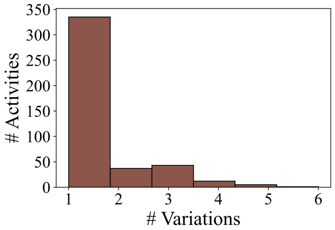
(e) **Distribution of step coverage across videos.**



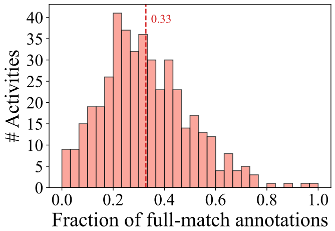
(f) **Distribution of annotated video durations.**



(g) **Distribution of unique steps annotated per video vs steps in taxonomy**



(h) **Distribution of the number of variations appearing in HT-step across activities.**



(i) **Distribution of annotated video durations.**

D Taxonomy

We provide the full step taxonomy in `taxonomy.csv`. Every step includes its corresponding activity, headline and paragraph, as well as variation information.

E Baselines implementation details

We train all the baselines on top of the same feature sequences, extracted from frozen backbones.

TimeSformer (TS) features. The TimeSformer features are extracted using the public model[1] of [7] pre-trained with distant supervision on HowTo100M. We obtain 1 feature per second, by resampling the video at 8 fps and extracting features with a stride of 8 frames. The feature dimensionality is 768.

[1] `https://dl.fbaipublicfiles.com/video-distant-supervision/TimeSformer_divST_8x32_224_HowTo100M_pretrained.pth`

S3D features. The S3D features are extracted using the published model[2] of [9] pre-trained with MIL-NCE [9]. We obtain 1 feature vector per second, by resampling the video at 16 fps and extracting features with a stride of 16 frames. The feature dimensionality is 1024.

ActionFormer. We use the official ActionFormer implementation[3] provided by the authors[13]. We set the number of classes to 4958 *i.e.*, one detection output for each step in the taxonomy. We set the max sequence length to 512 and train for 20 epochs with a batch size of 16, using the AdamW optimiser with cosine learning rate schedule, base learning rate of $1e-4$ and 5 warm-up epochs on a single Nvidia Tesla V100 GPU with 32GB of memory.

ActionFormer-T. ActionFormer-T is trained with the exact same hyper-parameters, loss and labels as ActionFormer. We extract the text representations, using the MPNet implementation provided by the `sentence_transformers` library[4]. The text embeddings are frozen, so no gradients are backpropagated into the pre-trained langauge model. A single linear layer is used to project between the $768-$dimensional text embeddings and the $512-$dimensional video embeddings. Step text descriptions from different sources (e.g. headline, paragraph, activity, see also Table 3 of the main paper) are combined by simple concatenation at the text level, *i.e.* the combined sentences including all three have the form “Activity: Headline. Paragraph".

UMT. For UMT, we use the authors' official code [5]. We train models with learning rate $1e-3$ and batch size 64 and train for 200 epochs. We use only the unimodal encoder for video (we do not use the audio encoder, or the cross-attention modules). All remaining hyperparameters follow the configuration for the QVHighlights task provided by the authors.

MT+BCE. The input to our temporal article grounding baseline is a temporal sequence of visual features extracted with a sliding window (using either the TimeSformer or S3D backbones as explained above), and a sequence of step sentences (consisting of the activity name and the article step headlines). We base our model on the VINA [8] architecture by removing the additional narrations modality, i.e., we do not use the narration unimodal encoders, positional encodings and the alignments of steps to narrations or narrations to video. We use the TAN[6] codebase for our implementation. All of the architecture hyperparameters (e.g., number of Multimodal Transformer layers, embedding dimensions etc.) are adopted from VINA. The only difference is the maximum length of the input video which we increase to 1200 seconds to account for the longer videos in the HT-Step training set.

To obtain temporal segment predictions for each article step from the Multimodal Transformer outputs, we: (1) compute the normalized dot product between each step contextual embedding and each video clip contextual embedding. This results in a $T \times S$ alignment matrix, where T is the number of timesteps and S is the number of steps. (2) We pass these similarities through a sigmoid activation (with temperature 0.07) to obtain a confidence score about whether each timestep t is aligned with step s, (3) we post-process the temporal sequence of confidence scores for each step with an 1D blob detection routine to obtain temporal segments at multiple scales. In particular, we apply Laplacian of Gaussian filters at 13 scales, covering Gaussian standard deviations from 1 to 480 [12].

The model is trained with binary cross-entropy loss applied at each temporal timestep and for each article step. We train our model for 9 epochs using the same optimizer, learning rate and batch size as VINA [8].

Adding paragraph information: For our ablations in Table 3 of the main paper, we added paragraph information to the MT+BCE model simply by interleaving step headlines with step paragraph

[2] https://github.com/antoine77340/S3D_HowTo100M

[3] https://github.com/happyharrycn/actionformer_release

[4] https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2

[5] https://github.com/TencentARC/UMT/tree/main

[6] https://github.com/TengdaHan/TemporalAlignNet

sentences. In other words, we tokenize the article into sentences (with a maximum of 28 sentences per step) and we encode and feed that sequence of sentences to the Multimodal Transformer. We use the same positional encoding for all sentences associated with the same article step. In order to obtain a single contextual embedding for each step, we max pool the embeddings of the headline and paragraph sentences of that step.

Model weights initialization: We train all variants of MT+BCE from scratch, except for the model in the last row of Table 2, which was trained after initializing the unimodal encoders, positional encodings and Multimodal Transformer weights using a VINA model pretrained on the HTM370k [3] subset of HowTo100M using pseudo-labels for wikiHow steps (and no ASR narrations) [8]. For this experiment, we adopt the same maximum video length as VINA (1024 seconds).

F Per-activity predictions

In Table 1 we show the per-activity AP breakdown of the performance of the two best models. We show the 25 highest and 25 lowest scoring activities, ranked by the performance of the ActionFormer detection model. Note that activities that are challenging for the fixed taxonomy, detection model (such as Cook Pork Tenderloin for which the AP is 0.62%) are handled better by the temporal grounding model (achieving 28.9% AP for Cook Pork Tenderloin). For this particular example of *Cook Pork Tenderloin*, this can be explained since this activity has only 4 examples in the training set. Therefore, the detection model does not have enough training samples to learn a good representation for the steps of this activity. On the other hand, the temporal article grounding model, that has been initialized with a model trained with weak-supervision on a much larger dataset (HTM370k) can perform better in this few-shot scenario. Another interesting observation is that for some activities the detection-based model outperforms language-based grounding.

G Evaluation protocol details

G.1 Article-grounding AP metric

Approaches in our proposed temporal article grounding benchmark are evaluated using Article Grounding mean Average Precision (AGrd. mAP) over temporal IoU thresholds from 0.3 to 0.7 with a step size set to 1 (as in existing benchmarks [4]), and using three fixed tIoU thresholds at 0.3, 0.5 and 0.7. As explained in the main paper, our proposed metric computes an AP per activity (which might be associated with multiple articles if is has variations) by treating all article steps associated with that activity as class-agnostic text queries (similar to the temporal grounding Average Precision introduced in [6]). The per-activity AP is only computed on videos demonstrating each particular activity. The final article-grounding mAP is computed by averaging the per-activity APs. Our mAP-based metric is more suitable for the temporal article grounding task than existing recall-based metrics for grounding [1, 14] which ignore non-groundable steps, or frame-wise metrics for step detection [11], which ignore the temporal extent of each segment.

G.2 Breakdown of article-grounding mAP per match type (*full* vs *partial*)

In Table 4 of our main paper, we report article-grounding mAP computed per step match type (*full* vs *partial*). The mAP for full matches was computed separately for step queries that only have fully-matching temporal segments (or no matching segments) in their corresponding video. Step queries that have both full and partial matches in the same video were ignored from the computation of the mAP on full matches. Furthermore, APs are only computed for activities that have ground-truth step queries with full matches and averaged over those. Overall, the mAP for full matches was computed based on 78 activities, with 1176 ground-truth instances. The mAP for partial matches was computed in a corresponding manner, covering 79 activities and 2375 ground-truth instances.

Table 1: **Breakdown of AP performance per activity on the seen test set (S1).** We show the 25 highest and 25 lowest scoring activities, ranked by the performance of the ActionFormer model.

	Model	
	ActionFormer	MT+BCE(VINA)
Activity		
Make Lunch Box Oatmeal Cookies	55.02	66.65
Make Chicken Liver Pate	51.94	45.56
Deep Fry a Turkey	51.56	43.02
Make Tomato Pie	47.97	38.43
Make Buttermilk Fried Chicken	45.53	39.40
Bake a Sweet Potato Pie	43.62	25.70
Make Scotch Eggs	42.90	41.82
Make Pecan Crusted Blackened Catfish	42.37	26.54
Make Vegetable Paniyaram	42.30	39.91
Cook Arepas	42.16	43.75
Prepare Mexican Chilaquiles	41.15	40.09
Make Chiles Rellenos	40.82	35.48
Make Beef Stroganoff	40.59	27.03
Clarify Butter	40.38	39.71
Make Toad in the Hole	38.56	43.81
Make Focaccia	38.54	39.11
Clean Flounder	37.95	15.60
Make Chicken Piccata	37.77	44.94
Brine, Truss, and Roast a Turkey	36.15	49.03
Grill Bacon	35.89	55.26
Make Eggplant Pasta Sauce	35.31	30.93
Make Mofongo	35.05	43.49
Make Saltimbocca	34.84	34.21
Cook Brussels Sprouts with Chestnuts	34.52	42.86
Make Beignets	34.34	40.71
. . .		
Make White Chili	15.79	11.41
Make Fairy Cakes with Self Raising Flour	15.51	36.88
Make Chicken Cacciatore	15.42	28.50
Make Grilled Artichokes	15.19	27.35
Make Healthier Fish Sticks	14.98	23.45
Bake a Queen Elizabeth Cake	14.91	27.57
Make Coconut Rice	14.82	31.91
Make Hostess Twinkies	14.54	24.67
Cook Cube Steak	13.20	34.79
Make Bannock	12.75	17.94
Make Mango Chutney	12.74	11.38
Make Overnight Caramel Pecan Rolls	12.47	18.48
Make Vegan Ceviche	11.96	4.85
Cook Black Eyed Peas	11.44	19.02
Make a Cheese Crisp	9.32	14.16
Cook Bacon in the Microwave	9.14	40.50
Make Italian Ice	9.01	19.73
Make Quick and Easy Sausage Rolls	8.27	16.67
Braai Steak	8.08	9.27
Make a Hearty Stew	7.73	31.03
Make Mediterranean Vegetable Cheese Pie	6.19	16.37
Make Hungarian Goulash	6.00	18.43
Make Bacon Toffee	3.92	8.15
Make Blueberry Strudel	1.96	9.29
Cook Pork Tenderloin	0.62	28.94
mAP	**25.4**	**29.8**

H Training, validation, and test splits

We have included details about the training, validation and test splits in Section 3.2 of the main paper. Here we add some comments.

Seen val/test set (S1). We note that these sets are balanced, each containing 600 videos in total, 5 videos per each of 120 activities, with an overlap between validation and test amounting to 63 activities. Labels are released for the val set, while labels for the seen test set are withheld and a fair evaluation protocol on this set is supported via a test server that will be made available to the community.

Unseen val/test set (S2). Note that the headlines or paragraphs of some steps in the unseen val/test sets may be very similar to steps of the activities included in the training set, due to the compositionality of recipes. For example, the unseen activity of *Make Poutine* contains the step "Add the garlic and shallot" which is similar to steps such as "Add the garlic and cook for 30 seconds" from the seen activity *Make a Hearty Stew* and "add the garlic slices and cook for 1 minute." from *Make Tumbet*. Evaluation on the unseen test set will be made possible through the test server.

I Datasheet for HT-Step

In this section we provide a detailed documentation about our dataset, following the format introduced by Gebru et al. [2]. Our HT-Step dataset provides a new set of annotations for a subset of *existing* videos from the HowTo100M dataset. The annotations depend on the HowTo100M [10] videos (no new videos were collected or recorded) and wikiHow articles from the wikiHow dataset [5]. In all our responses below, the term “data” specifically refers to the *annotations*, not the HowTo100M videos or the wikiHow articles associated with HT-Step, unless otherwise noted.

I.1 Motivation

a) **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

HT-Step was created to support research in procedural video understanding. It provides a collection of segment-level step annotations that greatly surpasses existing labeled datasets in this area along multiple axes: scale, number of activities, and richness of natural language step descriptions.

b) **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization?**

To maintain the anonymity of this submission, we will provide these details upon publication.

c) **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

To maintain the anonymity of this submission, we will provide these details upon publication.

I.2 Composition

a) **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance is a set of temporal segment annotations, denoting relevant temporal segments, for a specific article step on a particular video in HowTo100M [10] We refer the reader to Listing 1 for an example of the annotation format as well as to the sample json file that we provide `annotation_sample.json`, containing the annotations on 100 training videos.

b) **How many instances are there in total (of each type, if appropriate)?** There are 116k segment-level annotations.

Listing 1: Example annotation in a JSON format

```
{
  "-2maV1TTL5U": [
    {
      "segment": [
        103.73678,
        118.19032
      ],
      "step_label": "Proof the yeast.",
      "partial": "Full Match",
      "activity": "Make Bolillos",
      "step_index": 0,
      "variation_index": "1",
      "global_step_index": 3167
    },
    {
      "segment": [
        124.4,
        146.6
      ],
      "step_label": "Add most of the remaining dough ingredients.",
      "partial": "Full Match",
      "activity": "Make Bolillos",
      "step_index": 1,
      "variation_index": "2",
      "global_step_index": 3168
    },
    ...
  ]
  "-2vbgoZc-RI": [
    ...
  ]
}
```

c) **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

HT-Step covers a small and focused subset of HowTo100M. It amounts to approximately 1.7% of the total videos and it covers 433 cooking activities. The cooking domain was selected as it represents a large part of HowTo100M (approximately a third), contains relatively low-complexity tasks which can be annotated by non-experts that are often procedural in nature. The 433 tasks represent cooking activities for which HowTo100M contains at least 70 videos. Some manual filtering was done to keep only procedural activities. The final list of annotated videos in HT-Step was determined directly by the annotators – videos that were not related to the corresponding activity were rejected.

d) **What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.**

The annotation instances are organized per video, *i.e.* each video entry contains a list of annotations. Each annotation contains a temporal segment, the step index within the activity, the global step index (within the taxonomy), the variation index, an partial or full match indicator and the activity name. For an example, please see Listing 1.

e) **Is there a label or target associated with each instance?** If so, please provide a description. There are various labels associated with each instance. We provide the definition of each field below.

- `segment`: Time interval, represented as start and end timestamps in seconds.
- `activity`: The video's activity.
- `step_label`: The step headline text (as appearing in the corresponding wikiHow article).
- `variation_index`: The index of the activity variation, if any (listed in the taxonomy).
- `step_index`: The (local) index of the step within the activity.
- `global_step_index`: The global index of the step within the whole taxonomy.
- `partial`: One of "Full Match" or "Partial Match", indicating whether the annotation is full or partial.

f) **Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No information was removed intentionally, the annotations are provided as marked by the annotators.

g) **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.

This dataset does not provide any metadata about relationships between individual instances, besides the grouping of videos by activity.

h) **Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.

We provide data splits for training, validation, and testing (including seen and unseen val/test splits). Full details about the splits are given in Section 3.2 of the main paper. We will release the full annotations for the training and seen validation splits and withholding the test (seen/unseen) splits. To facilitate evaluation, we will set up and maintain a test server on EvalAI. Participants can upload their results to the server, where they will be evaluated automatically.

i) **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.

As we describe in Section 3.2 of the main paper, we followed a strict QA process to ensure the quality of the annotations. Full QA of the dataset was too costly, therefore only a fraction (13%) has been fully reviewed. Despite this effort, as human annotators can be prone to mistakes, it is possible that there are noisy annotations. To contain noise to a minimum for evaluation, we created the two test sets exclusively from annotations on videos that had were QA reviewed.

j) **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

HT-Step is a new set of annotations on existing videos of HowTo100M. These videos are publicly available through YouTube. There is no guarantee that the videos will always remain online and accessible, but this is a known issue with YouTube-mined datasets. The majority of the videos are available for free, under the policies of YouTube and may have individual licenses.

HT-Step will be released under the CC BY NC SA licence, which is can be found at https://creativecommons.org/licenses/by-nc-sa/2.0/. There is no dependency on wikiHow as the text for the steps is included and self-contained within the dataset's taxonomy and individual annotations.

k) **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non- public communications)?** If so, please provide a description.

No information contained in HT-Step is considered confidential.

l) **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.

N/A.

m) **Does the dataset identify any subpopulations (e.g., by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

n) **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** If so, please describe how.

It should not be possible to identify individuals from the labels provided in HT-Step. All the annotation text is taken from a fixed taxonomy of 4,958 steps, that were gathered from an independent source (wikiHow). Therefore there is no way to inject information specific to individual videos through them.

o) **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** If so, please provide a description.

N/A.

I.3 Collection Process

a) **How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Annotators directly watched each video and were also given a full list of the article steps (headlines and paragraphs) that they were asked to temporally localize in the video.

b) **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** How were these mechanisms or procedures validated?

We used an internal tool developed for web-based annotation. The same tool has been used for the annotation of other public datasets. To maintain the anonymity of this submission, we will provide more details upon publication.

c) **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

See our response in D.2.c.

d) **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

The collection was done by a third party ventor and annotators were compensated by contract. The vendor was involved with multiple annotation projects involved with the vendor, and thus their exact compensation is not available to us.

e) **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?** If not, please describe the timeframe in which the data associated with the instances was created.

The annotations were collected over approximately 2 months. This timeframe does not match the creation timeframes of the videos. These vary across the videos and are, due to the nature of HowTo100M, which was mined from YouTube, not easy to determine.

f) **Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Yes. This annotation project went through a rigorous internal privacy and ethical compliance review process. To maintain the anonymity of this submission, we will provide more details upon publication.

g) **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

We did not collect or record new videos as part of HT-Step annotation. We did collect annotations on these videos from annotators managed by a third-party vendor as discussed above. Annotations were collected a web-based, internal annotation tool.

h) **Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

i) **Did the individuals in question consent to the collection and use of their data?** If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Yes, see above.

j) **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

No.

k) **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

I.4 Processing, cleaning, labeling

a) **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remaining questions in this section.

We did not directly parse wikiHow articles but used the parsed version provided by https://github.com/mahnazkoupaee/WikiHow-Dataset. This corpus contains metadata that we used to determine which sets of steps are grouped into variations (i.e. they are organized as methods). The variations were obtained automatically, but we needed to correct a small number of them manually. The two most common causes for manual intervention were that either i) alternative variations of an activity were not flagged as such but listed as integral steps, or ii) subgroups of procedural steps were wrongly listed as alternative methods.

b) **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** If so, please provide a link or other access point to the "raw" data.

No. The final annotations format contains all the information collected by the process, except for the time of annotation and annotator ids.

c) **Is the software that was used to preprocess/clean/label the data available?** If so, please provide a link or other access point.

N/A.

I.5 Uses

a) **Has the dataset been used for any tasks already?** If so, please provide a description.

The full HT-Step dataset is not public yet, so no other papers have used it. As explained in the main paper, we form the seen validation and test split (S1) from [8], which introduces this small subset of the dataset and uses it for evaluating weakly-supervised temporal article grounding models.

b) **Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.

N/A.

c) **What (other) tasks could the dataset be used for?**

Potential uses of the dataset include training procedural activity models, temporal grounding and detection, step recognition and anticipation, and mining task graphs.

d) **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

N/A.

e) **Are there tasks for which the dataset should not be used?** If so, please provide a description.

N/A.

I.6 Distribution

a) **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.

Yes. The dataset will be made publicly available.

b) **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?

The dataset will be distributed through a dedicated website and an official test server will be set up and maintained on https://eval.ai. To maintain the anonymity of this submission, we will provide more details upon publication.

c) **When will the dataset be distributed?**

We will release HT-Step shortly after the decision is announced.

d) **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

HT-Step will be distributed under the CC BY NC SA licence, which can be found at https://creativecommons.org/licenses/by-nc-sa/2.0/.

e) **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

f) **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

I.7 Maintenance

a) **Who will be supporting/hosting/maintaining the dataset?**

We will support, host, and maintain the dataset. To maintain the anonymity of this submission, we will provide concrete details upon publication.

b) **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

To maintain the anonymity of this submission, we will provide this information upon publication.

c) **Is there an erratum?** If so, please provide a link or other access point.

Not for the time being.

d) **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

If any labeling corrections or new labels become available in the future, we will update the dataset, by providing a new version and clear documentation of the changes through the website.

e) **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.

No.

f) **Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

Yes.

g) **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

No, we currently do not envision a process for this.

References

[1] 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. 4

[2] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, nov 2021. 7

[3] 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. 4

[4] Haroon Idrees, Amir R. Zamir, Yu-Gang Jiang, Alexander Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017. 4

[5] Mahnaz Koupaee and William Yang Wang. Wikihow: A large scale text summarization dataset. *ArXiv*, abs/1810.09305, 2018. wikiHow dataset license available at: `https://github.com/mahnazkoupaee/WikiHow-Dataset`. 7

[6] Jie Lei, Tamara L. Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries. In *Neural Information Processing Systems*, 2021. 4

[7] 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

[8] Effrosyni Mavroudi, Triantafyllos Afouras, and Lorenzo Torresani. Learning to ground instructional articles in videos through narrations. *arXiv preprint arXiv:2306.03802*, 2023. 3, 4, 11

[9] 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

[10] 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. HT100M dataset license available at: `https://github.com/antoine77340/howto100m/blob/master/LICENSE`. 7

[11] 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. 4

[12] Huiyu Wang, Mitesh Kumar Singh, and Lorenzo Torresani. Ego-only: Egocentric action detection without exocentric pretraining, 2023. 3

[13] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, volume 13664 of *LNCS*, pages 492–510, 2022. 3

[14] 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. 4