Supplementary material for: Learning to Ground Instructional Articles in Videos through Narrations

This Appendix provides: additional details (annotation procedure, statistics) about the HT-Step dataset that we introduced for evaluating models on step grounding (Section A), additional details for the rest of the datasets that were used for training/evaluation (Section B), implementation details (Section C), qualitative results for step grounding on HT-Step (Section D), additional ablation studies (Section E), and additional details about the evaluation of our models on HTM-Align (Section F).

A. HTM-Step Dataset

In this section we provide details about the creation of the HT-Step benchmark that we used for evaluating our models. This benchmark was designed to provide a high-quality set of step-annotated instructional videos for a plethora of tasks, described in rich, structured language instead of atomic phrases.

Annotation setup. We used videos from the HowTo100M dataset; each one of those videos contains a task id label that corresponds to a wikiHow article. This association enabled us to obtain a set of potential step descriptions for every video, directly from the corresponding wikiHow article. We note that this association is noisy, e.g. the video might show a variation of a specific recipe, where some of the steps in the article often do not appear at all, appear partially, are executed in different order, or are repeated multiple times.

Annotation instructions. For each video, annotators were provided with the task name (e.g., Make Pumpkin Puree) and the recipe steps from the corresponding wikiHow article. The annotators where asked to watch the whole video and first decide whether it is relevant to the given task – i.e. if at least some of the given steps were visually demonstrated and the task's end goal was the same (e.g. a specific recipe) – or reject it otherwise. When a video was deemed relevant, annotators were asked to mark all instances of the provided steps with a temporal window. We note that WikiHow articles often contain several variations/methods for completing a given task. For tasks where this was the case, the annotators were asked to select the set of steps corresponding to the variation that best fits every video and only use those steps for annotating the entire video.

QA process. To ensure the quality of the annotations, we followed a rigorous multi-stage Quality Assurance (QA) process: In the first stage, the videos were annotated by a single annotator. These initial annotations were then reviewed by more experienced annotators, who either approved all the annotations on a given video (meaning all the marked steps were correct and no steps were missing) or marked it for redoing with specific comments on which annotations needed fixing and in what way. At the last stage of the QA process, the annotations that were marked as incorrect were redone by third, independent annotators.

Statistics. We provide per-activity statistics for the annotations in Table 1. The metrics used, *i.e.* number of unique steps, step and video coverage, are given to provide an understanding of how the number of steps varies between different tasks and how the steps of a task may appear partially in the HowTo100M videos.

Validation and test (val/test) split. Overall during the full annotation process, approximately 35% of the videos were rejected as irrelevant to the given tasks. We split the remaning, annotated videos into a validation and a test set, each containing 600 videos, with 5 videos per task. We ensured that our validation set does not contain videos from HTM-Align. In total 87 human annotators manually annotated 1200 videos over 177 tasks: 120 in the validation and 120 in the test set, with 5 videos per task, i.e. with 63 tasks overlapping between the two sets.

B. Datasets Details

HowTo100M (Training). HowTo100M contains over 1M unique instructional videos, spanning over 24k activities including cooking, DIY, arts and crafts, gardening, personal care, fitness and more. Each instructional video is complemented by the ASR transcription of it's audio, which usually contains the real time narration/commentary of the instructor during the activity. We use the "senticified" version of the ASR sentences provided by Han *et al.* [2]. Following Han *et al.* [2] we also train only using the Food & Enteratainment subset, which includes a subset of approximately $370k$ videos.

wikiHow (Training). We train using 14,541 cooking tasks

Task	# steps	step coverage	video coverage
Make Zucchini Pancakes	4.0	0.83	0.37
Make a Hearty Stew	3.5	0.82	0.12
Make Beef and Broccoli	3.1	0.78	0.24
Make Coconut Popsicles	3.8	0.76	0.28
Make Yorkshire Pudding	5.3	0.76	0.11
Cook Spaghetti alla Carbonara	4.6	0.73	0.39
Make Vegan Pesto	2.2	0.73	0.15
Make Corn Fritters	6.4	0.72	0.28
Make Buttermilk Fried Chicken	4.2	0.70	0.44
Make a Shrimp Po Boy Sandwich	4.2	0.70	0.27
⋮	⋮	⋮	⋮
Cook Prime Rib	2.6	0.19	0.19
Cure Bacon	2.2	0.18	0.11
Make Dim Sum	4.6	0.18	0.15
Make Vegan Ceviche	2.8	0.17	0.08
Make Lobster Bisque	3.6	0.17	0.28
Make Giblet Gravy	2.8	0.16	0.23
Make Pickled Eggs	4.4	0.16	0.19
Pickle Onions	1.6	0.15	0.12
Cook Rib Eye Roast	2.0	0.12	0.28
Make Pap	2.0	0.12	0.21
Average	**4.0**	**0.42**	**0.24**

Table 1: Statistics of the annotations used to create the HT-Step benchmark. The metrics are computed per task (for 177 tasks in total), averaged over all the annotated videos for a given task. **# steps** denotes the average number of unique steps annotated per video, per activity; **step coverage** denotes the fraction of a task's steps that have been found and annotated in every video; **video coverage** denotes the fraction of the video's duration that is covered by step annotations; Rows are sorted by step coverage; only the 10 tasks with the highest and lowest step coverage are shown here for brevity.

from the wikiHow-Dataset [3]. For each task, we generate an ordered list of steps by extracting the step headlines. The HowTo100M dataset was curated using a semi-automatic pipeline that involved searching YouTube with queries based on the titles of wikiHow articles. Consequently there is an almost complete overlap in activities between the two corpora, which makes wikiHow a natural choice for mining step-level articles to associate with instructions in HowTo100M videos. In the context of this paper we used the wikiHow-Dataset [3] to collect the articles for 14,541 cooking tasks.

CrossTask (Evaluation). We use this established instructional video benchmark for *zero-shot* grounding, i.e., by directly evaluating on CrossTask our model learned from HowTo100M. The Crosstask dataset [10]. is an established benchmark for temporal localization of steps in instructional videos. It consists of 4800 videos from 83 activities, which are divided into 18 primary (14 related to cooking and 4 to DIY car repairs and shelf assembly) and 65 related activities. The videos in the primary activities are annotated with step annotations in the form of temporal segments from a predefined taxonomy of 133 steps. Those steps tend to be atomic, *e.g.* for activity "Make Taco Salad" the available steps are "add onion", "add taco", "add lettuce", "add meat", "add tomato", "add cheese", "stir", and "add tortilla". 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 [10]. 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 [10]. 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. [1] and report results on 2407 videos of 15 primary tasks obtained excluding three primary tasks following the protocol of [1]. Performance for this task is measured with Recall@K at different IoU thresholds [1].

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. In the main submission, we report the R@1 metric [2], which evaluates whether the model can correctly localize the narrations that are alignable with the video. In Section F we also evaluate our model in terms of its capability to decide whether a narration is visually groundable in the video or not using the ROC-AUC metric [2]. AUC denotes the area the ROC curve of the alignment task, and measures the ability of the model to correctly predict whether a given step is alignable within a video or not.

C. Implementation Details

As video encoder we adopt the S3D [8] backbone pretrained with the MIL-NCE objective on HowTo100M [5]. Following previous work [2, 9], 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 [6]. These embeddings are initialized based on MIL-NCE Word2Vec and are fine-tuned during training.

The hyperparameters of the model compared with state-of-the-art methods in Tables 1,2,3 of the main submission were selected based on R@1 performance on the HT-Step validation set and are: $\lambda_{SV} = \lambda_{NV} = 1$, temperatures $\eta, \xi = 0.07$, and pseudo-label filtering threshold $\gamma = 0.65$. We train our model for 12 epochs, with 3 epochs burn-in training with step pseudo-labels generated by TAN, and then we update the teacher VINA every 3 epochs. We use

the AdamW [4] optimizer, having an initial learning rate of $2e-4$ decayed with a cosine learning schedule. Our batch size is 32 videos, with maximum length of 1024 seconds.

Pseudo-labels are obtained based on the steps-to-video alignment matrix and are generated (before filtering) as follows: for each step we find the timestep with maximum similarity with the step and then extend a temporal segment to the left and right of that peak as long as the similarity score does not follow below 0.7 of the peak height. Pseudo-labels whose peak score falls below the filtering threshold γ are not used for training.

The rest of hyperparameters were selected based on TAN [2]. The multimodal encoder is a pre-norm multi-layer transformer which consists of 6 layers of self-attention, with 8 heads and has hidden dimension $D = 512$. A learnable positional encoding of size $D = 512$ is used to inject temporal information to each frame/narration/step token.

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 or for the qualitative video included in this supplementary) we use a simple 1D blob detector [7]. Unless otherwise specified, we use the fused alignment matrix for step grounding when narrations are available during inference time.

Our model is trained on 8 GPUs (Tesla V100-SXM2-32GB) and training lasts approximately 10-12 hours. All models were implemented in Python using Pytorch and are based on the PySlowFast (https://github.com/facebookresearch/SlowFast) and TAN (https://github.com/TengdaHan/TemporalAlignNet) open-source codebases. For ablation studies, we choose the best checkpoint for each configuration based on performance on HT-Step validation set and report its test split performance.

D. Qualitative Results

In this section, we provide qualitative results for the ground-truth steps-to-video alignment and predicted alignments by our improved baseline that serves as the initial teacher model (TAN*), and our model (using the direct steps-to-video alignment without narrations) or the fusion with the indirect steps-to-video alignment (with narrations). From these qualitative results (Figure 1), we observe that our VINA model can correctly temporally localize visually groundable steps, despite being trained only with noisy pairs of narrated videos and instructional steps. Predicted alignments tend to also be less noisy than TAN*, showcasing the effectiveness of training a video-language alignment model with distant supervision from WikiHow articles. Our model can also leverage ASR transcripts (without any temporal information regarding when the instructor uttered each narration) to further improve its results (Figure 2).

E. Extra Ablations

Architecture ablations. In Table 2 we study the design of the unimodal encoder used to embed steps before they are fed to our Multimodal Transformer. Overall, using positional embeddings capturing the ordering of steps in a task, and using modality-specific projection MLPs leads to a slightly better performance in step grounding (w/o narration input). Narration grounding seems to benefit from using a shared text encoder, possibly because this facilitates knowledge transfer from the WikiHow steps.

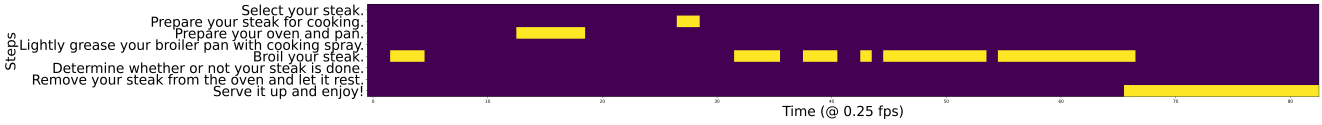
PE	Sep. MLP	HT-Step ↑R@1		HTM-Align
		w/o nar.	w/ nar.	
	✓	33.5	34.0	65.8
		34.0	34.9	65.9
✓		33.8	34.4	**67.0**
✓	✓	**34.3**	**36.1**	64.8

Table 2: **Ablation study on architecture design**. We study the contribution of positional encodings for steps (*PE*) and of specialized text projection layers for wikiHow article steps (*Sep. MLP*). All models are trained for joint narration and step grounding with fixed pseudo-labels from TAN and evaluated on HT-Step val split (last row corresponds to row 5 in Table 4 of the main text).

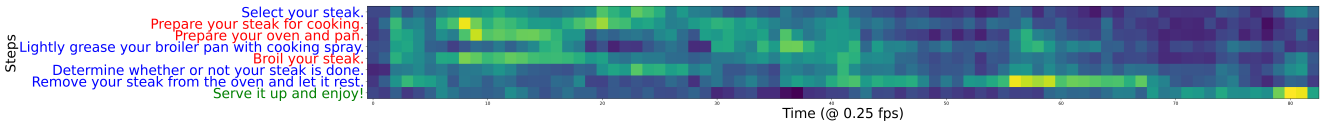
F. Experimental Setup on HTM-Align

As explained in the official code repository of TAN [2] (https://github.com/TengdaHan/TemporalAlignNet/tree/main/htm_align), the results reported for HTM-Align are obtained with a text moving window of 1 minute, i.e., for each 1-minute temporal segment only ASR captions whose original time-stamps fall within a 3-min window centered around this temporal segment are considered for grounding. Instead, for all our reported results (for TAN* and VINA) we operate in the more challenging setup where an ASR caption can be grounded in any timestep of the original video (there is no knowledge about the original ASR timestamps during inference). Under this more challenging setup, our model outperforms TAN both in narration retrieval, as measured by Recall@1 (66.5% vs 49.4%, as seen in Table 1 of the main submission).

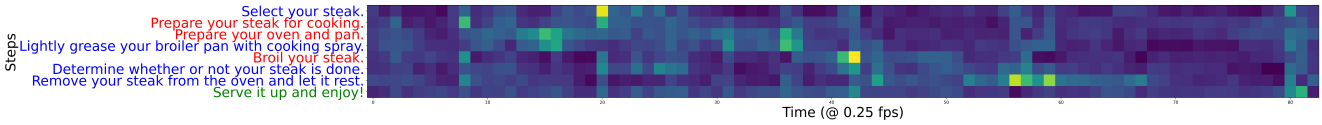
Our model also performs comparably with TAN in step alignability prediction, as measured by ROC-AUC (76% vs 75.1%). Note that our model does not have dedicated alignability head for predicting whether a narration exists or not in the video as TAN [2]. Instead, we simply obtain an alignability score by using the maximum cosine similarity score over time, where cosine similarities of each narration with each video frame are computed based on the outputs of the unimodal encoders.



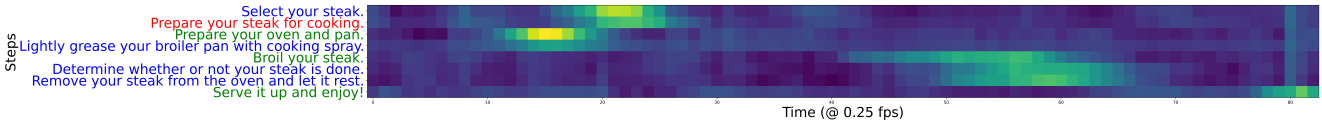
(a) Ground-truth step grounding annotations.



(b) Predicted alignment by TAN*.



(c) Predicted alignment by VINA w/o narrations.



(d) VINA w/ narrations.

Figure 1: Qualitative results on a validation video of the HT-Step dataset (`VIQYQkA3mNU`) demonstrating how to *Broil Steak*. Steps that are not visually groundable in the video are highlighted in blue, steps that are correctly retrieved by each model are highlighted in green, while steps that are not retrieved are shown in red. Figure best viewed zoomed in and in color.

G. Limitations and Ethical Concerns

From the qualitative results, we observe that due to the losses used during training, which do not explicitly penalize wrong temporal extent (as long as the predicted heatmap has a peak within the target temporal window), grounded temporal segments tend to be short. This is especially prominent when using the direct steps-to-videos alignment that is explicitly supervised (second to last row of the predicted alignment figures). Furthermore, our training objective does not utilize negative examples, e.g. steps that are not visually groundable, to suppress detections. This can lead to confident detections for missing steps. Another limitation of our approach (similar to previous approaches that operate on the same pre-extracted visual features) is that our performance is limited by the quality of the extracted visual representations. In future work we plan to train our model on instructional videos from additional, non-cooking domains, as well as explore alternative training objectives for accurate temporal segment prediction that also leverage the paragraph information of article steps in addition to the headlines. Regarding ethical concerns, public instructional video datasets and public knowledge base datasets may have gender, age, geographical and cultural bias.

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. 2

[2] 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

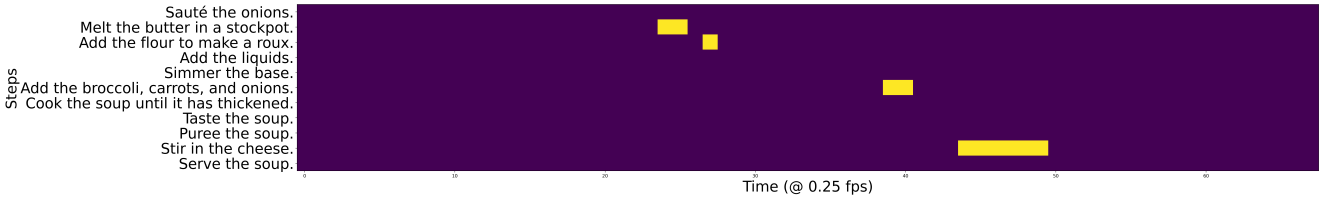
[3] Mahnaz Koupaee and William Yang Wang. Wikihow: A large scale text summarization dataset. *ArXiv*, abs/1810.09305, 2018. 2

[4] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 3

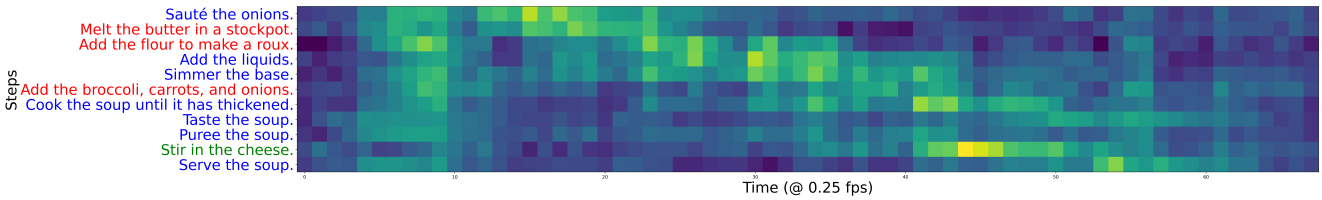
[5] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *CVPR*, 2020. 2

[6] 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. 2

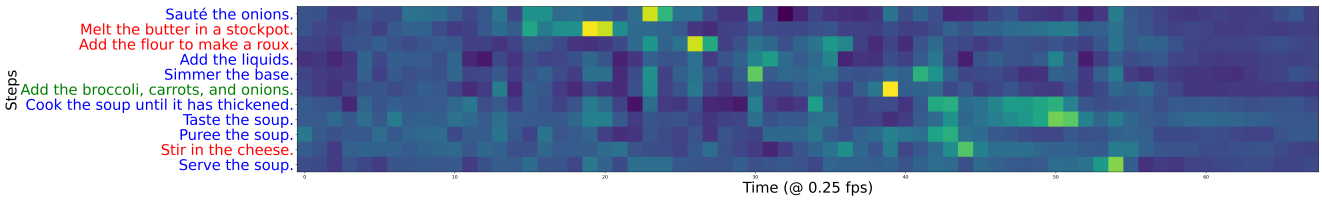
[7] Huiyu Wang, Mitesh Kumar Singh, and Lorenzo Torresani.



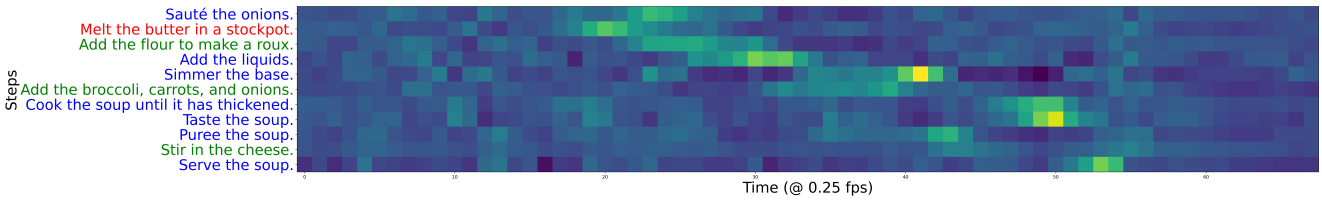
(a) Ground-truth step grounding annotations.



(b) Predicted alignment by TAN*.



(c) Predicted alignment by VINA w/o narrations.



(d) VINA w/ narrations.

Figure 2: Qualitative results on a validation video of the HT-Step dataset (`0dHofx1lqAg`) demonstrating how to *Make Broccoli Cheese Soup*. Steps that are not visually groundable in the video are highlighted in blue, steps that are correctly retrieved by each model are highlighted in green, while steps that are not retrieved are shown in red. Figure best viewed zoomed in and in color.

Ego-only: Egocentric action detection without exocentric pretraining, 2023. 3

[8] 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. 2

[9] 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. 2

[10] 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