

Non-Sequential Graph Script Induction via Multimedia Grounding

Yu Zhou¹, Sha Li², Manling Li², Xudong Lin³, Shih-Fu Chang³, Mohit Bansal⁴, Heng Ji²

¹ University of California, Los Angeles ² University of Illinois Urbana-Champaign

³ Columbia University ⁴ University of North Carolina at Chapel Hill

yu.zhou@ucla.edu, {shal2, manling2, hengji}@illinois.edu

mbansal@cs.unc.edu, {xudong.lin, shih.fu.chang}@columbia.edu

Abstract

Online resources such as wikiHow compile a wide range of scripts for performing everyday tasks, which can assist models in learning to reason about procedures.¹ However, the scripts are always presented in a linear manner, which does not reflect the flexibility displayed by people executing tasks in real life. For example, in the CrossTask Dataset, 64.5% of consecutive step pairs are also observed in the reverse order, suggesting their ordering is not fixed. In addition, each step has an average of 2.56 frequent² next steps, demonstrating "branching". In this paper, we propose a new challenging task of non-sequential graph script induction, aiming to capture *optional* and *interchangeable* steps in procedural planning. To automate the induction of such graph scripts for given tasks, we propose to take advantage of loosely aligned videos of people performing the tasks. In particular, we design a multimodal framework to ground procedural videos to wikiHow textual steps and thus transform each video into an observed step path on the latent ground truth graph script. This key transformation enables us to train a script knowledge model capable of both generating explicit graph scripts for learnt tasks and predicting future steps given a partial step sequence. Our best model outperforms the strongest pure text/vision baselines by 17.52% absolute gains on $F_1@3$ for next step prediction and 13.8% absolute gains on $Acc@1$ for partial sequence completion. Human evaluation shows our model outperforming the wikiHow linear baseline by 48.76% absolute gains in capturing sequential and non-sequential step relations.

1 Introduction

A script consists of typical actions that are performed to complete a given task. Online re-

¹Our data and code are publicly available for research purposes at <https://github.com/bryanzhou008/Multimodal-Graph-Script-Learning/>

²Occurred in more than 10 videos.

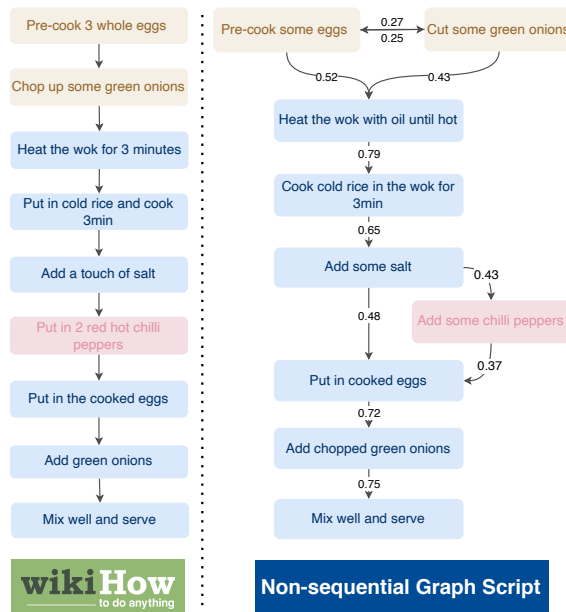


Figure 1: Example of wikiHow linear script of the procedural task *make egg fried rice* compared to an ideal example of our non-sequential graph script consisting of *optional* and *interchangeable* steps.

sources such as wikiHow³ provide a wide variety of community-edited scripts for everyday tasks (Fig.1). Such a large library of linear scripts can serve as a starting point for learning goal-step knowledge (Zhang et al., 2020; Yang et al., 2021b). However, as the saying goes, “all roads lead to Rome”. There is usually more than one way to achieve any given goal. Practically speaking, users should be presented with multiple alternative step sequences so that they can pick the most suitable route according to their unique situations and preferences. Robots and virtual assistants also stand to gain the crucial abilities of global planning optimization and on-the-spot improvisation from alternative step paths.

In particular, we observe that two types of steps

³www.wikiHow.com

are overlooked by linear scripts: *optional steps* and *interchangeable steps*. Optional steps such as Add some chili peppers can be skipped based on the users' preference or item availability. Interchangeable steps such as Pre-cook some eggs and Cut some green onions can be performed in either order without affecting the overall task completion. After accounting for these two step types, the original linear script is converted into a 'non-sequential graph script', as shown in Fig.1 (right).

Previous efforts like Proscript (Sakaguchi et al., 2021) obtained non-linear graph scripts via crowdsourcing, which is not scalable. In this work, we automate the process of transforming a linear text script into a non-linear graph script by grounding into visual observations (videos) of people executing the task. If we observe that people often skip a certain step, then it is natural to denote that step as optional. Similarly, if people tend to swap the ordering of a group of steps, these steps are likely interchangeable. Since wikiHow does not contain such empirical observations, we align wikiHow scripts with procedural video datasets such as Crosstask (Zhukov et al., 2019) and Howto100M (Miech et al., 2019) (see Fig.2).

To map a video to a sequence of wikiHow steps, we perform alignment on both task-level and step-level. On the task level, we use a title matching algorithm based on Sentence-BERT similarity to select videos and wikiHow documents for the task. Then, we propose an effective pre-processing strategy (simplification + deduplication) to create the wikiHow step library. At the step level, we consider two situations based on whether the video has been segmented into steps. When manual segmentation is provided, we directly map video annotations to the wikiHow step library. Otherwise, we first segment the video into clips based on ASR sentence groups (Fig.2), and then map them to wikiHow steps using a fault tolerant grounding strategy (§3.1) that is robust to inaccurate ASR sentence boundaries. When grounding is complete, we obtain the set of observed step sequences for each task.

Next, to obtain the desired graph script from the observed step sequences, we use auto-regressive seq2seq models (Sutskever et al., 2014) to learn the distribution of valid paths (step sequences) along the graph (§3.2). As opposed to directly training a graph generation model, our path generation learning format is better aligned with existing procedural

video data and also takes advantage of pretrained seq2seq models to improve generalization across tasks. Since the cross-entropy loss used for training auto-regressive models focuses on penalizing local "one-step" errors (the errors in predicting each single step), we further introduce a Path-level Constraint Loss to reduce global inconsistencies of the entire path. To generate hard negative contrastive-paths that fail to complete the task, we manipulate the video-grounded positive paths through *global reordering*, *shuffling*, and *re-sampling* (§3.2).

After training, our model is able to produce complete paths given input step libraries from various domains, including but not limited to: cooking, car maintenance, and handcrafting, etc. To automatically generate explicit graph scripts, we implement step-level constraint beam-decoding to sample multiple generated step sequences and record a step-adjacency matrix for constructing the final graph script.

For downstream evaluation, we adapt the existing CrossTask dataset (Zhukov et al., 2019) to set up two new evaluation sub-tasks: *Next Step Prediction* and *Partial Sequence Completion*. Compared against top-performing test/video only baselines, our best model achieves 17.52% absolute gains in overall $F_1@3$ for next step prediction and 13.8% absolute gains on Accuracy@1 for partial sequence completion. Moreover, we use MTurk to perform *Human Evaluation* on the correctness and expressiveness of our auto-generated graph scripts. Results show our model can correctly capture optional, interchangeable and sequential step relationships with up to 82.69% overall accuracy.

Key contributions of this paper include:

- We introduce an automatic method for converting sequential text scripts into non-sequential graph scripts by aligning / grounding textual scripts to video datasets.
- We propose a path generation model capable of learning from video-grounded step sequences with Path-Level Constraint Loss.
- Experiments show our non-sequential path generation model to be more effective than existing text/vision baselines in next step prediction and partial sequence completion.
- Human evaluation of generated graph scripts demonstrates our non-sequential graph scripts to be more accurate and expressive in capturing step-relationships.

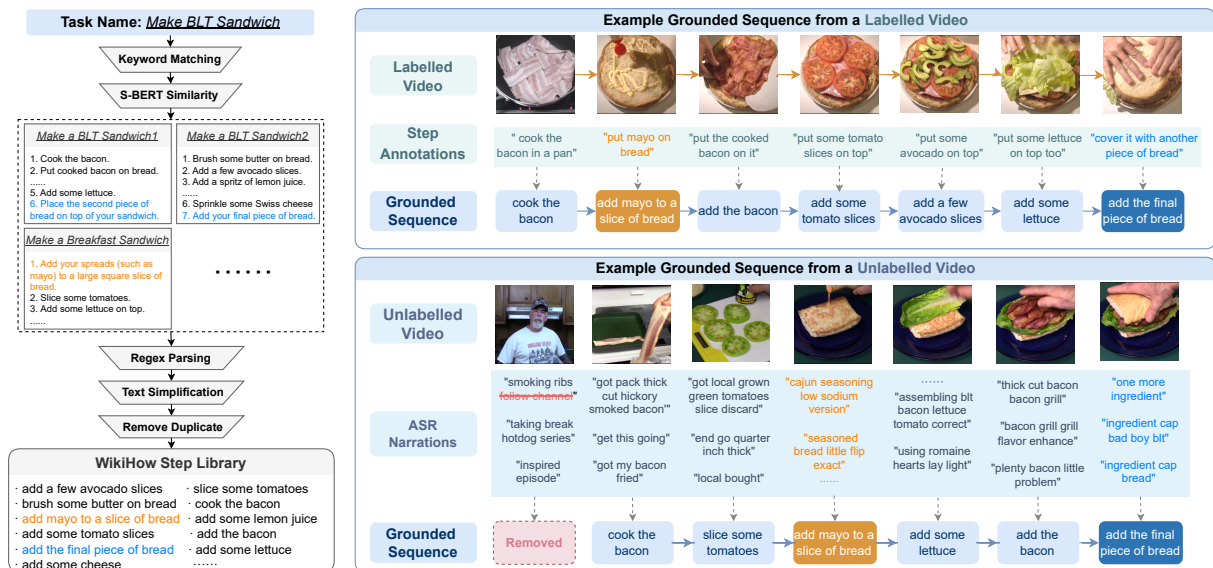


Figure 2: Example of grounding procedural videos of the Making BLT Sandwich task to wikiHow steps. We create a wikiHow step library through task-level matching and step pre-processing, and then ground video step annotations/asr-narrations to textual steps from the wikiHow step library.

2 Task Formulation

In this paper, we propose a new challenge of graph script induction for procedural tasks: Given a procedural task \mathcal{T} represented by a task name, our goal is to induce a graph script for the task using the steps in the linear script. In particular, the graph script should capture the following relations between steps: (1) *sequential* $\langle s_i \rightarrow s_j \rangle$ where two steps should be completed sequentially; (2) *interchangeable* $\langle s_i \leftrightarrow s_j \rangle$ where two steps can be completed in either order or at the same time; (3) *optional* $\langle s_i \rightarrow s_k, s_i \rightarrow s_j \rightarrow s_k \rangle$ where a step can be optionally added between other steps.

To achieve this goal, we assume that we have access to a large repository of textual scripts (wikiHow) and a set of videos that record people carrying out the tasks.⁴ The videos might have step-level annotations or accompanying narration which we can convert into text using ASR tools.

3 Methodology

To learn a graph script induction model, we first ground the video dataset to textual steps on both task-level and step-level (Fig. 2). After grounding, each video can be seen as a valid step sequence sampled from the ground truth graph script. Then, we use such grounded step sequences to train our graph script model and enhance model learning

⁴Or a large repository of videos from which we can find matching videos using retrieval.

by introducing a Path-Level Constraint Loss over carefully designed contrastive step sequences.

3.1 Video to Script Grounding

For each video, we first perform task-level alignment to find the top- m most relevant wikiHow documents and then step-level alignment to ground the video to specific wikiHow steps. We consider the following two cases based on whether the video dataset includes step-level annotation:

Labelled Video Datasets: Labelled video datasets like Crosstask (Zhukov et al., 2019) contain procedural videos grouped by human-annotated task names. In addition, the videos are labelled with temporal step segmentation and relatively accurate step annotations in the form of short imperative English sentences. The example video in Fig.2 for task: *"Make BLT Sandwich"* is annotated with steps: "cook the bacon in a pan", "put mayo on bread", etc.

At the task level, we first use keyword matching to quickly find all relevant wikiHow documents whose title contains $\geq 85\%$ of keywords in the task name. For example in Fig. 2, the task name: *"Make BLT Sandwich"* is matched to wikiHow documents: *"Make a BLT Sandwich1"*, *"Make a Breakfast Sandwich"*, etc. After we retrieve a list of relevant wikiHow documents, they are further ranked by cosine similarity between Sentence-BERT embeddings of document title and

the task name. Finally, the steps of the top m wikiHow documents are selected to form the initial wikiHow step candidate pool.

In step-level grounding, we first record Sentence-BERT Similarity scores between each video step annotation and all processed steps in the wikiHow step library. Then, we do greedy matching between video step annotations and wikiHow steps with priority given to higher scoring pairs. Here we keep video steps with best score $\geq k_1^5$, while lower scoring video steps are considered ungroundable. When all videos have been grounded, unused steps from the wikiHow step library are removed.

Unlabelled Video Datasets: Although we achieve high grounding quality for annotated video datasets, step-level annotation is quite costly and often not available for a wide range of tasks that we are interested in. A more practical scenario is when we have a large repository of videos like Howto100M from which we can retrieve videos corresponding to the target task. Task-level alignment for Howto100M is different from that of annotated video datasets due to questionable video grouping. In Howto100M, videos for each task are selected purely based on Youtube search ranking. This ranking often prioritizes popular videos that have low correlation to the task at hand. To ensure high video-task correlation, we re-select Howto100M videos for each task based on BERT-Similarity between video title and the task name (only videos with similarity score $\geq k_2$ are selected).

Step-level alignment also becomes much more challenging as we must rely on video ASR transcriptions without human step-level annotations. ASR narrations usually comprise of short partial sentence pieces without strict temporal step boundary labels (Fig.2). In addition, since Howto100M videos are collected from Youtube, some ASR narrations contain task-irrelevant information such as subscription requests (Fig.2). To address these challenges, we use a more fault tolerant grounding strategy shown in Fig.2: First, we remove all sentence pieces containing Youtube stop words including “subscribe”, “channel”, “sponsor”, etc. Then, we expand each ASR sentence piece by concatenating it with surrounding pieces until the length of the resulting piece exceeds 10 words⁶. Finally, we

⁵Hyperparameters in the grounding section are empirically selected based on qualitative evaluation over a small subset.

⁶This parameter is borrowed from (Lin et al., 2022) which

ground each resulting ASR step to wikiHow steps with a higher match threshold k_3 .

Processing the wikiHow Step Library: High quality step-level alignment demands the wikiHow Step Library used for grounding to contain clean, non-overlapping steps that are homogeneous in format and granularity to the video step annotations. Since the vanilla wikiHow dataset (Koupae and Wang, 2018) does not meet these criteria, we perform a series of pre-processing before step-level alignment:

1. First, we put the steps in the initial wikiHow step library through a series of regex-based parsing to standardise stylistic elements like capitalization, punctuation and bracket/parentheses usage.
2. Then, we use a seq2seq text simplification model (Maddela et al., 2021) to reduce granularity in wikiHow steps which are often more fine-grained than video step annotations.
3. Finally, we deduplicate the wikiHow Step Library by enforcing a minimum weighted Levenshtein distance of 0.1 between any two steps and removing overly similar duplicate steps.

3.2 Model Training

Graph Script Learning Inspired by (Bojchevski et al., 2018), we transform the graph script learning problem into a path learning problem by treating steps as nodes and temporal relationships between the steps as directed edges (edges point to future step). For each procedural task \mathcal{T} , the wikiHow step library of task-relevant steps $\mathcal{W}_{\mathcal{T}}$ generated in §3.1 represents the set of nodes used to construct the latent ground-truth graph script. In §3.1, we grounded each procedural video to a wikiHow step sequence. These step sequences can be regarded as observed step node paths that lead to successful completion of \mathcal{T} . In this formulation, learning the latent graph script for a task can be regarded as learning the weights of valid paths through $\mathcal{W}_{\mathcal{T}}$.

For our basic architecture, we train a BART-base model (Lewis et al., 2019) to generate complete step sequences given a wikiHow step library. As illustrated in Fig.3, for each task \mathcal{T} , we first shuffle the corresponding wikiHow step library to remove any pre-existing step ordering. Then, we concatenate the shuffled step library with a special separator uses the same length threshold

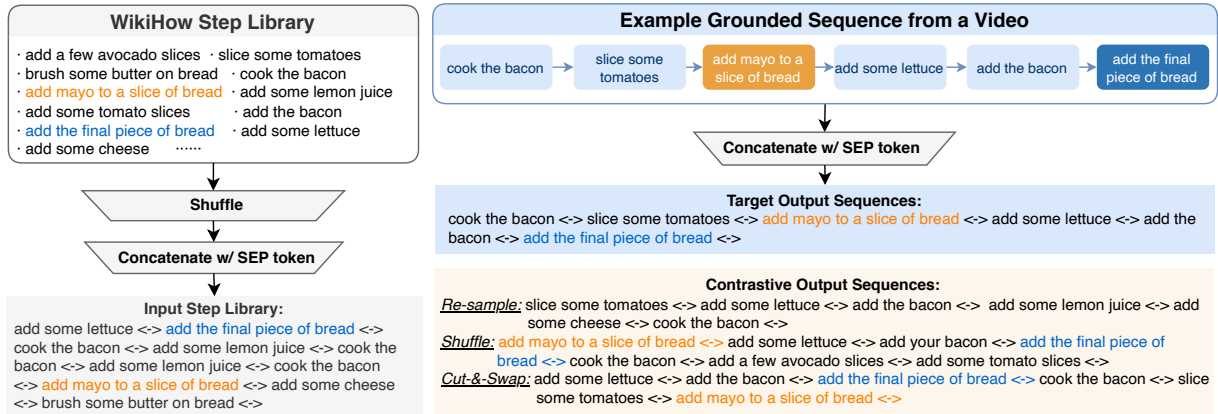


Figure 3: **Example input/output sequence in model training.** We create the input sequence by shuffling and concatenating the wikiHow step library. We use the concatenated grounded sequence as the target output (positive example) and its permuted/resampled versions as the contrastive output (negative example).

rator token⁷ appended to the end of every step to indicate step boundary. The resulting sequence is used as the input sequence for all training data regarding \mathcal{T} . For each target output, we first collect all grounded step sequences of videos completing \mathcal{T} . Similar to input sequences, steps in the output are also appended with the same separator token and concatenated. Finally, each processed video-grounded step sequence is used individually as a target output sequence for our model.

Path-Level Constraint Besides being able to generate valid step sequences that lead to successful task completion, we also enable our model to differentiate valid step sequences from invalid ones that fail to complete the task. We accomplish this by introducing a Path-Level Constraint in the form of a contrastive loss. For each positive step sequence, we generate n negative contrastive sequences using the following 3 methods (Fig.3):

1. *Re-sample*: randomly re-sample a step sequence of the same length from the wikiHow step library. Both step selection and step ordering are wrong.
2. *Shuffle*: shuffle the sequence until no longer valid. Step selection is preserved, but local/global step ordering are wrong.
3. *Cut & Swap*: cut the sequence at a random position and swap the latter part to the front. Step selection and local step ordering are preserved, but global step ordering is wrong.

To maximize the model’s learning potential, we follow the paradigm of curriculum learning (Bengio

⁷We define the separator token as <->.

et al., 2009) when introducing contrastive examples: we start with contrastive sequences generated via *Re-sample* because they are most dissimilar from valid sequences. As training progresses, we shift toward *Shuffled* and *Cut & Swap* by gradually increasing the probability of sampling from those contrastive sequence groups.

Inspired by (Saha et al., 2022), we use the last layer of the decoder in BART as the representation of each token in the sequence and obtain the sequence representation by averaging over the constituent token representations. Let the hidden representations of our generated sequence $\mathbf{s}^{(g)}$, true grounded sequence $\mathbf{s}^{(p)}$ and negative contrastive sequence $\mathbf{s}^{(n)}$ be denoted by $\mathbf{z}^{(g)}$, $\mathbf{z}^{(p)}$ and $\mathbf{z}^{(n)}$, respectively. Let $\mathcal{Z} = \{\mathbf{z}^{(p)}\} \cup \{\mathbf{z}_i^{(n)}\}_{i=1}^M$, with M as the number of negative contrastive sequences. Hence, we define our Path-level Contrastive Loss:⁸

$$\mathcal{L}_{PC} = -\log \frac{\exp[\text{sim}(\mathbf{z}^{(g)}, \mathbf{z}^{(p)})/\tau]}{\sum_{\mathbf{z}^{(i)} \in \mathcal{Z}} \exp[\text{sim}(\mathbf{z}^{(g)}, \mathbf{z}^{(i)})/\tau]}, \quad (1)$$

where the temperature τ is a hyperparameter and sim denotes cosine similarity. Finally, our overall loss combines the Path-level Contrastive Loss with the Cross-Entropy Loss of seq2seq models:

$$\mathcal{L}_{CE} = \sum_i -\log P(\mathbf{s}_i^{(p)} | \mathbf{s}_{<i}^{(p)}, \mathcal{W}_{\mathcal{T}}), \quad (2)$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{CE} + \alpha \mathcal{L}_{PC}, \quad (3)$$

⁸Based on the InfoNCE Contrastive Loss(van den Oord et al., 2018)

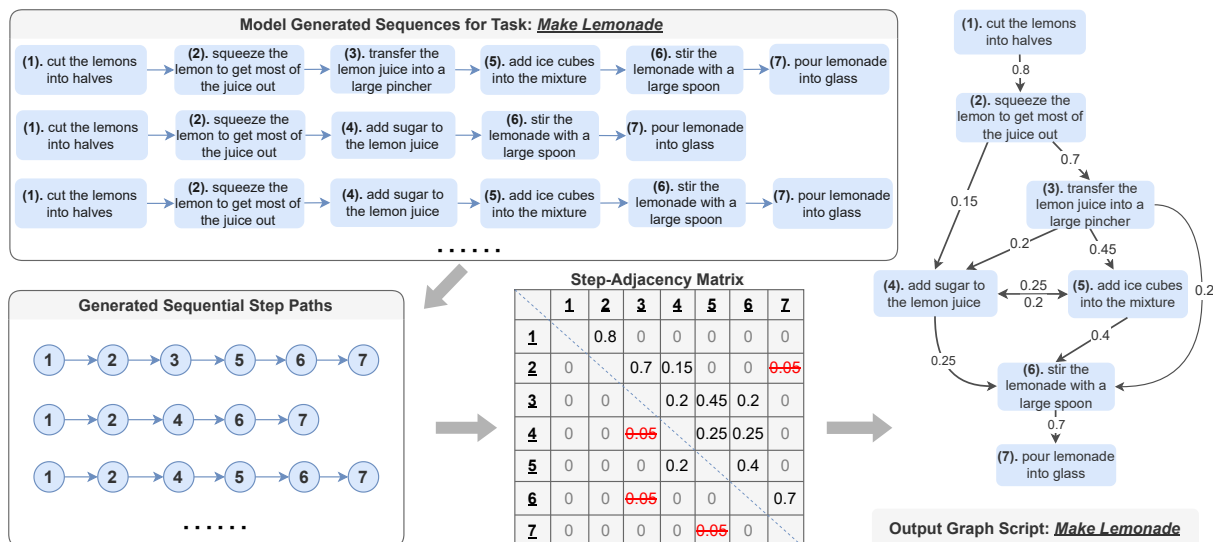


Figure 4: **Example of Graph Script Generation.** To decode a graph from our generator, we first ask the model to generate alternative step sequences via beam-decoding and record them in an step-adjacency matrix, which is then be used to reconstruct the non-sequential graph script (with low-frequency edges removed).

where α is a hyperparameter and $\mathcal{W}_{\mathcal{T}}$ denotes the task-specific wikiHow step library.

3.3 Graph Script Generation

In §3.2, we transformed the graph script learning problem into a path learning problem by treating procedural step relationships as edges between nodes and abstracting the latent ground truth graph as the collection of paths through node-set $\mathcal{W}_{\mathcal{T}}$ that lead to successful task completion. After our model has learnt the latent ground truth graph scripts for a set of tasks, we use it to reconstruct explicit graph scripts through the following procedure:

For each task \mathcal{T} , we use $\mathcal{W}_{\mathcal{T}}$ as model input and have the model generate output step sequences consisting only of steps within $\mathcal{W}_{\mathcal{T}}$. We enforce this by implementing Step-constrained Beam Search, an extension of Constrained Beam Search (De Cao et al., 2021), where the model is only allowed to generate valid next words that lead to entities stemming from a fixed prefix trie \mathcal{P} . Here, we construct $\mathcal{P}_{\mathcal{T}}$ containing all steps in $\mathcal{W}_{\mathcal{T}}$ and ask the model to repeatedly decode from $\mathcal{P}_{\mathcal{T}}$ to generate step sequences. After each step is fully generated, the model is given the choice to end generation by producing the end-of-sentence (eos) token or continue decoding the next step by producing a token from the root of $\mathcal{P}_{\mathcal{T}}$. After generating the predicted step sequences, we break them down and record the edges in an graph adjacency matrix between all generated step nodes. The low-frequency

edges representing unlikely paths are removed to improve graph script confidence. Finally, we reconstruct the output graph script from the graph adjacency matrix. An example of this process on the task "*Make Lemonade*" is detailed in Fig.4.

4 Experiments

To evaluate our non-sequential graph script induction model, we propose 3 new downstream tasks:

1. **Graph Script Generation:** for each task \mathcal{T} , the system is asked to produce a 2-dimensional probabilistic graph script similar to Fig.1 that captures the step relationships introduced in section 1). The model is scored based on human evaluation of its generated graph scripts.
2. **Next Step Prediction:** given a partial step sequence $S_p = (s_1 \rightarrow \dots \rightarrow s_{t-1})$, the model is asked to predict the top-k most likely choices for the next step s_t from $\mathcal{W}_{\mathcal{T}}$. For each partial step sequence, there can be a variable number of correct next steps.
3. **Partial Sequence Completion:** given a partial step sequence $S_p = (s_1 \rightarrow \dots \rightarrow s_{t-1})$, the model is asked to produce a sequence $S = (s_1 \rightarrow \dots \rightarrow s_n)$ using steps from $\mathcal{W}_{\mathcal{T}}$ that completes the task \mathcal{T} . This task is particularly challenging because the model is asked to predict a variable-length step sequence that best completes the task at hand.

Model	HT100M	Next Step Prediction				Partial Sequence Completion		
		Acc@1 ↑	Acc@3 ↑	Rec@3 ↑	F ₁ @3 ↑	Acc@1 ↑	Edit Dist. ↓	Normalized Edit Dist. ↓
TimeSformer+DS	✗	59.91	60.82	52.98	43.83	-	-	-
Random	✗	31.34	50.32	28.84	38.04	1.20	2.398	.6935
wikiHow Linear	✗	44.05	59.51	54.02	42.14	11.74	1.872	.6061
ReBART	✗	49.07	58.00	61.39	44.38	18.28	1.802	.4411
Direct NSP (Grounding)	✗	68.89	63.02	79.01	53.85	-	-	-
Direct PSC (Grounding)	✗	-	-	-	-	29.17	1.214	.4118
Ours (Grounding)	✗	75.59	67.50	83.17	58.29	20.12	1.639	.4296
Ours (Grounding)	✓	70.97	74.68	74.14	61.52	29.34	1.193	.4093
Ours (Grounding + PLC)	✗	75.49	71.89	72.51	58.48	26.70	1.228	.4267
Ours (Grounding + PLC)	✓	76.09	73.72	78.22	61.90	32.08	1.123	.3849

Table 1: **Automatic Evaluation Results on Next Step Prediction and Partial Sequence Completion.** Here “HT100M” denotes whether the model is pre-trained on the Howto100M dataset with temporal order information. “Normalized Edit Dist.” represents the average Levenstein distance normalized by sequence length. “Grounding” denotes whether the model used our grounded video sequences for training. “PLC” represents Path-Level Constraint.

4.1 Baselines

Baseline: TimeSformer+DS. TimeSformer (Bertasius et al., 2021) trained with unsupervised distant supervision (Lin et al., 2022) provides the state-of-the-art step-level video representation for pure-video-based step forecasting. We fine-tuned the model on CrossTask videos before testing.

Baseline: wikiHow Linear. This model is trained on all wikiHow linear step sequences selected during title-matching (§3.1). To ensure fairness in comparison, the training sequences undergo the same step processing as that of the non-sequential model. For each training sequence, the model takes the complete wikiHow step library as input and one linear sequence from the selected wikiHow documents as target output.

Baseline: ReBART. ReBART (Chowdhury et al., 2021) is the state-of-the-art sentence re-ordering method that uses a text-to-marker generation format. Numbered markers are inserted before each step in the training data, and the target output step sequence is translated into corresponding marker sequences.

Ablation Study: Direct Next Step Prediction & Direct Partial Sequence Completion. These two task-specific models are included as variants of our model (§3.2) where the input training sequence is a partial start sequence and the target output sequence is just the next step (for next step prediction) or the remaining sequence (for partial sequence completion). The training data for these two models are also constructed from our grounded video step sequences (§3.1).

4.2 Automatic Evaluation

Evaluation Dataset Inspired by (Chen et al., 2022), we build our evaluation dataset on top of the existing CrossTask Dataset (Zhukov et al., 2019) and reuse their manual temporal step annotations. Using procedures in §3.1, we ground annotated CrossTask videos (Fig.2) to sentence-simplified wikiHow Steps. Afterwards, we randomly select 40% of grounded step sequences to form the training set. Remaining sequences form the test set.

For each grounded step sequence $S = (s_1 \rightarrow \dots \rightarrow s_n)$ in the test set, we split after all steps ($s_t | t \in [1, n-1]$) to produce partial start sequences $S_p = (s_1 \rightarrow \dots \rightarrow s_t)$. For next step prediction, the correct output corresponding to S_p is the next step s_{t+1} ; while for partial sequence completion, the correct output corresponding to S_p is the remaining sequence ($s_{t+1} \rightarrow \dots \rightarrow s_n$). In the case where multiple grounded step sequences share the same partial start sequence S_p but have different next step / remaining steps, the input sequence S_p would have multiple correct answers for next step prediction / partial sequence completion.

Next Step Prediction As shown in Table 1, our models trained using video-to-text grounded step sequences outperform other baselines trained with wikiHow linear step sequences by 15% ~ 20% absolute gains in all next step prediction metrics. This shows the advantage of our video-grounded step sequences over wikiHow linear sequences in improving the model’s ability to predict next steps. Comparing our models trained on complete step sequences against models trained directly on next

Relation Type	Linear		Ours	
	#/task	Acc	#/task	Acc
Sequential	10.56	35.79	12.50	88.02
Optional	1.40	19.23	2.44	65.91
Interchangeable	0.44	37.50	1.44	88.46
Overall	12.40	33.93	16.38	82.69

Table 2: Human Evaluation results by step-relation type.

Task Category	Linear		Ours	
	#/task	Acc	#/task	Acc
Cooking	12.1	35.16	16.2	81.07
Household	12.5	28.33	16.0	75.00
Car Maintenance	15.0	36.67	17.5	88.89

Table 3: Human Evaluation results by task category. #/task denotes average number of relations per task.

step prediction without whole script knowledge, we see a large performance gap. This shows the importance of learning whole-script knowledge for next step prediction. When predicting top-3 most likely next steps, models pretrained on Howto100M significantly outperform models w/o pretraining. This can be attributed to the pretrained models having better knowledge of sequence "branching" from observing more diverse task executions.

Partial Sequence Completion Our best performing models trained using video-to-text grounded step sequences typically achieves over 13% absolute gains on Accuracy@1 and over 14% relative gains on normalized edit distance against other baselines trained using wikiHow linear step sequences, showing grounded videos step sequences can boost models' ability in partial sequence completion. When comparing models trained with the Path-Level Constraint (Sec.3.2) to otherwise identical models trained without such constraint, we see significant gains across all metrics. This demonstrates the effectiveness of our Path-Level Constraint in teaching the model to produce valid step sequences while avoiding their invalid counterparts. We also observe a performance gain for models pretrained on Howto100M vs the same models w/o such pretraining. This result combined with similar results in next step prediction shows that pretraining on a large unlabelled procedural video dataset can improve the model's ability to learn scripts for other tasks.

4.3 Human Evaluation

Using the graph construction method in §3.3, we generate two graph scripts for each procedural task in CrossTask using the wikiHow Linear baseline (§4.1) and our non-sequential graph script induction model. To evaluate the correctness and expressiveness of generated graph scripts, we design T/F questions regarding *sequential*, *optional*, and *interchangeable* relations. For optional and interchangeable step relationships indicated by the graph script, we ask annotators whether the relationship is appropriate. For other steps in the connected graph script, we ask annotators whether their previous and subsequent steps are sequentially appropriate.

Table 2 and table 3 show our model achieves 46.68% ~ 52.23% absolute gains in Accuracy across all relation types and task categories. In addition, our model is able to accurately capture 74% more optional steps and 227% more interchangeable step pairs in generated graph scripts.

5 Related Work

Text-based Script Induction Temporal relations have always been the core of script (schema) related tasks, which can either be learned from data or human annotation. When human-written scripts are available, previous works have typically assumed that the human-provided ordering of steps is the only correct order (Jung et al., 2010; Ostermann et al., 2017; Nguyen et al., 2017; Lyu et al., 2021; Sakaguchi et al., 2021). Another line of work has attempted to learn event ordering from data alone, either by assuming that the events follow narrative order (Chambers and Jurafsky, 2008, 2009; Jans et al., 2012; Rudinger et al., 2015; Ahrendt and Demberg, 2016; Wang et al., 2017) or by using an event-event temporal relation classifier to predict the true ordering of events (Li et al., 2020, 2021). Our work is distinct from both paradigms as we use human-written scripts as a basis and learn the event ordering from observed sequences in videos.

Video-based Script Induction Existing efforts that utilize visual information in script induction can be mainly classified into implicit script knowledge models and explicit sequential script induction models. Some previous efforts have focused on training models with implicit script knowledge that can make step-level predictions based on textual (Yang et al., 2021c), visual (Sener and Yao, 2018; Lin et al., 2022; Zhou et al., 2023),

or multimedia (Zellers et al., 2021; Wu et al., 2022; Wang et al., 2022) input. Other models aim to produce explicit sequential graph scripts that only capture procedural relations between steps (Salvador et al., 2018; Yang et al., 2021a). Another line of works use multimedia information to generate explicit graph scripts that model only pre-conditional/dependency relationships between events (Logeswaran et al., 2023) and sub-events (Jang et al., 2023). Ours is the first work to generate explicit non-sequential graph scripts that capture rich procedural, optional, and interchangeable relations through multimedia learning.

6 Conclusions and Future Work

We propose the new task of Non-sequential Graph Script Induction to capture optional and interchangeable steps in procedural tasks. Instead of relying on the script annotation, we automatically induce graph scripts by grounding procedural videos to a wikiHow textual step library. We transform the graph generation problem to a path generation problem that can better align with video observations, and train a seq2seq model using our grounded step sequences while imposing path-level constraints via a contrastive loss. Experiments demonstrate our model’s superiority on downstream tasks including next step prediction and partial sequence completion. Human evaluation confirms our model’s ability to generate graph scripts that correctly capture optional and interchangeable steps. Future work will focus on incorporating more video supervision signals such as enriching steps from videos and adding the repeatable steps.

7 Limitations

7.1 Representation of Repeatable Steps

Our current approach is not able to capture repeatable steps due to data source constraints from our video datasets. The video datasets we use in this work, namely Howto100M and CrossTask, are both constructed from Youtube videos. At the end of many Youtube instructional videos, there is a brief recap of the whole task, where many steps are displayed for a second time. Since CrossTask was originally proposed for step segmentation, the step annotations capture all video references to task-related steps, including the brief mentions at the end of the videos that are not actually part of task execution. Similarly, Howto100M videos ASR

pieces near the end of the video would also capture the voiceover going through such step references.

Therefore, to ensure the grounded video step sequence only contains steps included in the execution of the task, we simply removed all repeated steps in the grounded step sequence and only kept the first occurrence. However in this process, we also removed valid repeats of the same step. For example, if the step Add some salt was executed twice at different stages of the task. We leave this area of improvement for future works.

7.2 Enrichment of Steps from Video

In our current model, all the steps in the wikiHow step library are processed steps from related wikiHow documents. However, it has been shown that textual sources can be prone to reporting bias, occasionally ignore task-relevant information that is present only in the vision modality (Chen et al., 2021).

Continuous frames from video data can capture details that text descriptions do not explicitly mention. If the model is able to make use of such vision-exclusive details and learn patterns from them, its overall ability can be improved. The challenge in utilizing such underlying visual information is to differentiate task-relevant video steps from their task-irrelevant counterparts. This area has not been covered by our current graph script induction pipeline, we hope to provide comprehensive solutions in future work.

8 Ethics and Broader Impact

8.1 Datasets

In this work, we used publicly available text data from the wikiHow Dataset (<https://github.com/mahnazkoupae/wikiHow-Dataset>) Creative Commons License (CC-BY-NC-SA), which is under the *Attribution-Noncommercial-Share Alike 3.0 Creative Commons License* which allows us to use the dataset for non-commercial purposes. For video data, we used the publicly available CrossTask Dataset (<https://github.com/DmZhukov/CrossTask>) under the BSD 3-Clause "New" or "Revised" License and the Howto100M Dataset (<https://www.di.ens.fr/willow/research/howto100m/>) under the Apache License 2.0. Both licenses allow us to use the datasets for non-commercial purposes.

The datasets we use consist of non-offensive instructional and procedural videos / text scripts about everyday tasks. Our usage of the datasets only concerns the task related information and does not violate privacy.

8.2 Human Evaluation

As detailed in §4.3, we conduct human evaluation for our generated graph scripts in this paper via Amazon Mechanical Turk(<https://www.mturk.com/>). All annotators involved in the human evaluation are voluntary participants and receive a fair wage. All annotators were instructed of the task nature and consent to complete the annotation via an online consent form. We have applied for IRB exemption and the request was approved.

8.3 Model Usage

Our graph script induction framework is not intended to be used for any activity related to any human subjects. Instead, it should only be used for generating graph scripts regarding everyday tasks that benefit people’s learning and understanding. It may also be used for predicting/instructing future step/steps to facilitate completion of relevant tasks. Note that our graph script induction framework is intended for wikiHow visual tasks and might not be applicable for other scenarios.

Acknowledgement

Many thanks to Prof. Mark Yatskar, Prof. Chris Callison-Burch, Prof. Long Chen, and Prof. Juanzi Li for helpful discussions and insightful feedback. We would also like to thank the anonymous reviewers for their constructive suggestions. This research is based upon work supported by U.S. DARPA KAIROS Program No. FA8750-19-2-1004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Simon Ahrendt and Vera Demberg. 2016. Improving event prediction by representing script participants. In *North American Chapter of the Association for Computational Linguistics*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. *Curriculum learning*. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML ’09*, page 41–48, New York, NY, USA. Association for Computing Machinery.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*.

Aleksandar Bojchevski, Oleksandr Shchur, Daniel Zügner, and Stephan Günnemann. 2018. Netgan: Generating graphs via random walks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 609–618.

Nathanael Chambers and Dan Jurafsky. 2008. *Unsupervised learning of narrative event chains*. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Annual Meeting of the Association for Computational Linguistics*.

Brian Chen, Xudong Lin, Christopher Thomas, Manling Li, Shoya Yoshida, Lovish Chum, Heng Ji, and Shih-Fu Chang. 2021. Joint multimedia event extraction from video and article. In *Conference on Empirical Methods in Natural Language Processing*.

Long Chen, Yulei Niu, Brian Chen, Xudong Lin, Guangxing Han, Christopher Thomas, Hammad Ayyubi, Heng Ji, and Shih-Fu Chang. 2022. Weakly-supervised temporal article grounding. In *Empirical Methods in Natural Language Processing (EMNLP), 2022*.

Somnath Basu Roy Chowdhury, Faeze Brahman, and Snigdha Chaturvedi. 2021. *Reformulating sentence ordering as conditional text generation*. *CoRR*, abs/2104.07064.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. *Autoregressive entity retrieval*. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Y. Jang, Sungryull Sohn, Lajanugen Logeswaran, Tiange Luo, Moontae Lee, and Ho Hin Lee. 2023. Multimodal subtask graph generation from instructional videos. *ArXiv*, abs/2302.08672.

Bram Jans, Steven Bethard, Ivan Vulic, and Marie-Francine Moens. 2012. Skip n-grams and ranking functions for predicting script events. In *Conference of the European Chapter of the Association for Computational Linguistics*.

- Yuchul Jung, Jihee Ryu, Kyung min Kim, and Sung-Hyon Myaeng. 2010. Automatic construction of a large-scale situation ontology by mining how-to instructions from the web. *J. Web Semant.*, 8:110–124.
- Mahnaz Koupaee and William Yang Wang. 2018. [Wikihow: A large scale text summarization dataset](#). *CoRR*, abs/1810.09305.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics*.
- Manling Li, Sha Li, Zhenhailong Wang, Lifu Huang, Kyunghyun Cho, Heng Ji, Jiawei Han, and Clare R. Voss. 2021. The future is not one-dimensional: Complex event schema induction by graph modeling for event prediction. In *Conference on Empirical Methods in Natural Language Processing*.
- Manling Li, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, Jonathan May, Nathanael Chambers, and Clare R. Voss. 2020. Connecting the dots: Event graph schema induction with path language modeling. In *Conference on Empirical Methods in Natural Language Processing*.
- Xudong Lin, Fabio Petroni, Gedas Bertasius, Marcus Rohrbach, Shih-Fu Chang, and Lorenzo Torresani. 2022. Learning to recognize procedural activities with distant supervision. *arXiv preprint arXiv:2201.10990*.
- Lajanugen Logeswaran, Sungryull Sohn, Y. Jang, Moon-tae Lee, and Ho Hin Lee. 2023. Unsupervised task graph generation from instructional video transcripts. *ArXiv*, abs/2302.09173.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- Qing Lyu, Li Zhang, and Chris Callison-Burch. 2021. Goal-oriented script construction. *ArXiv*, abs/2107.13189.
- Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*.
- Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2630–2640.
- Dai Quoc Nguyen, Dat Quoc Nguyen, Cuong Xuan Chu, Stefan Thater, and Manfred Pinkal. 2017. Sequence to sequence learning for event prediction. In *IJCNLP*.
- Simon Ostermann, Michael Roth, Stefan Thater, and Manfred Pinkal. 2017. Aligning script events with narrative texts. *ArXiv*, abs/1710.05709.
- Rachel Rudinger, Pushpendre Rastogi, Francis Ferraro, and Benjamin Van Durme. 2015. Script induction as language modeling. In *Conference on Empirical Methods in Natural Language Processing*.
- Swarnadeep Saha, Prateek Yadav, and Mohit Bansal. 2022. Explanation graph generation via pre-trained language models: An empirical study with contrastive learning. In *ACL*.
- Keisuke Sakaguchi, Chandra Bhagavatula, Ronan Le Bras, Niket Tandon, Peter Clark, and Yejin Choi. 2021. [proScript: Partially ordered scripts generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2138–2149, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Amaia Salvador, Michal Drozdal, Xavier Giró i Nieto, and Adriana Romero. 2018. Inverse cooking: Recipe generation from food images. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10445–10454.
- Fadime Sener and Angela Yao. 2018. Zero-shot anticipation for instructional activities. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 862–871.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. [Sequence to sequence learning with neural networks](#). *CoRR*, abs/1409.3215.
- Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. [Representation learning with contrastive predictive coding](#). *CoRR*, abs/1807.03748.
- Qingyun Wang, Manling Li, Hou Pong Chan, Lifu Huang, J. Hockenmaier, Girish V. Chowdhary, and Heng Ji. 2022. Multimedia generative script learning for task planning. *ArXiv*, abs/2208.12306.
- Zhongqing Wang, Yue Zhang, and Ching-Yun Chang. 2017. Integrating order information and event relation for script event prediction. In *Conference on Empirical Methods in Natural Language Processing*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.
- Te-Lin Wu, Alex Spangher, Pegah Alipoormolabashi, Marjorie Freedman, Ralph Weischedel, and Nanyun Peng. 2022. Understanding multimodal procedural knowledge by sequencing multimodal instructional manuals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4525–4542.

- Yue Yang, Joongwon Kim, Artemis Panagopoulou, Mark Yatskar, and Chris Callison-Burch. 2021a. Induce, edit, retrieve: Language grounded multimodal schema for instructional video retrieval. *ArXiv*, abs/2111.09276.
- Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. 2021b. [Visual goal-step inference using wikiHow](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2167–2179, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yue Yang, Artemis Panagopoulou, QING LYU, Li Zhang, Mark Yatskar, and Chris Callison-Burch. 2021c. [Visual goal-step inference using wikihow](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. In *Neural Information Processing Systems*.
- Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020. [Reasoning about goals, steps, and temporal ordering with WikiHow](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4630–4639, Online. Association for Computational Linguistics.
- Honglu Zhou, Roberto Mart'in-Mart'in, Mubbasir Kapadia, Silvio Savarese, and Juan Carlos Niebles. 2023. Procedure-aware pretraining for instructional video understanding. *ArXiv*, abs/2303.18230.
- Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David F. Fouhey, Ivan Laptev, and Josef Sivic. 2019. Cross-task weakly supervised learning from instructional videos. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3532–3540.

A Appendix

A.1 Grounding Details

The following hyper-parameters used in the grounding section are determined empirically. In video to text grounding, for each video, we find the top-10 most relevant wikiHow documents. For keyword matching at the task level, we first select wikiHow documents whose title contains $\geq 85\%$ of keywords in the task name. This is to avoid calculating Sentence-BERT similarity between the task name and all wikiHow document titles. If this does not yield ≥ 10 documents, we relax the threshold to 75%. For title matching and step matching, the Sentence-BERT similarity thresholds are determined empirically by qualitative evaluation over a small subset of 150 examples. For labelled videos, the step-level grounding similarity threshold k_1 is 0.35. For unlabelled videos, the task-level grounding similarity threshold k_2 is 0.75 and the step-level grounding similarity threshold k_3 is 0.40.

A.2 Training Details

For our models and baselines, we mainly use the BART-base model (140M Parameters) from the Huggingface Framework (Wolf et al., 2019)⁹. We normalize all input and target output sentences into lower case and remove special non-English characters. For training, we use the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 2×10^{-5} and 1000 warm-up steps. We use max input and output sequence length of 1024 for training and testing.

For the InfoNCE contrastive loss, we set the temperature $\tau = 0.1$. To implement curriculum learning. We reset the probability of sampling from different contrastive sequence groups every 5 epochs. At first we only use 're-sampled' contrastive sequence, then in every 5 epochs we transfer 20% probability to sampling from the 'shuffled' contrastive sequences. After 25 epochs, we starting the same shift from 'shuffled' contrastive sequences to 'cut & swapped' contrastive sequences.

We use NVIDIA V-100 GPUs with 16GB RAM and full precision. Due to GPU RAM limitation, we use gradient accumulation with equivalent batch size of 32. Training our basic model on the CrossTask training set takes approximately 5 hours while training our contrastive model with

⁹<https://huggingface.co/docs/transformers/index>

Path-Level Constraint will take 20 hours. Pre-training our model on Howto100M grounded sequence takes approximately 3 days.

A.3 Inference Details

During graph generation, for Step-constrained Beam Search, we use a beam number of 40 to sample steps sequences for producing the graph script. Afterwards, we filter out low-frequency edges in the adjacency graph with edge weight ≤ 0.175 (or in this case occurrence ≤ 7). The remaining edges are used to construct the final graph script.

A.4 Human Evaluation Details

In our human evaluation of model-generated graph scripts, three types of questions are asked regarding corresponding types of step relationships as displayed in the generated graph script:

1. *Optional*: Do you think step (a) is optional when completing this task?
2. *Interchangeable*: Do you think the steps (b) and (c) are interchangeable (can be executed in either order) when completing this task?
3. *Sequential*: Do you think the previous and/or subsequent steps for step (d) are reasonable when completing this task?

To make questions more direct and objective for the annotators, each question only focuses on a small portion of steps in the generated graph script. For example, given the output graph script for the task "Make Strawberry Cake" as shown in Fig.5, the annotator would be asked the following questions (partial):

1. *Optional*: Make Strawberry Cake: Do you think the step "cut the strawberries" is optional when completing this task?
2. *Interchangeable*: Make Strawberry Cake: Do you think the steps "add sugar to the mixture" and "whisk the mixture" are interchangeable (can be executed in either order) when completing this task?
3. *Sequential*: Make Strawberry Cake: Do you think the previous and/or subsequent steps for step "add flour to the mixing bowl" are reasonable when completing this task?

We used 8 human annotators while each annotator answered (on average) 65 questions. Each question is assigned to ≥ 2 annotators with 72.31% inter-annotator agreement. An example screenshot of the annotation interface is shown in Fig.6.

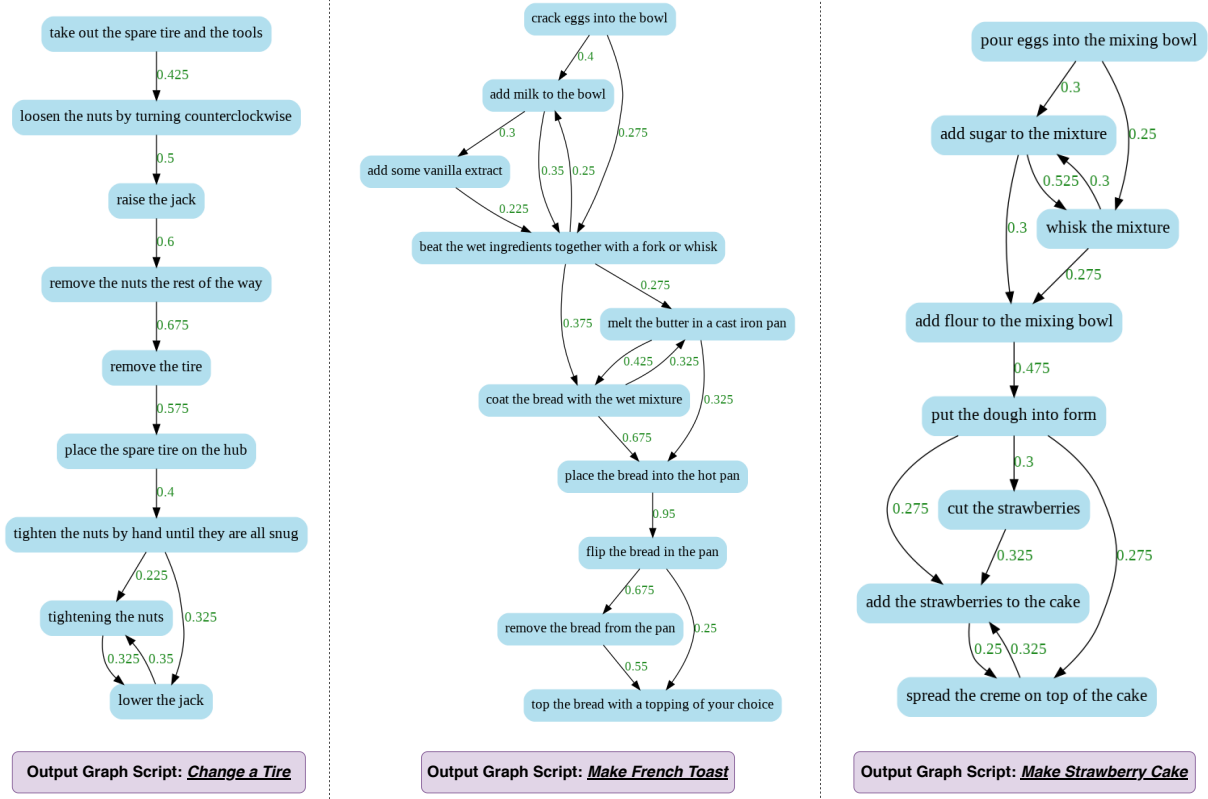


Figure 5: Example outputs of our non-sequential graph script induction model used in Human Evaluation

Instructions Shortcuts Make French Toast: In this graph script, do you think the previous and/or subsequent steps for step 'beat the wet ingredients together with a fork or whisk' are reasonable when completing this task?

Select an option	
Yes	1
No	2

Figure 6: A screenshot from our annotation task on Mechanical Turk.

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
7
- A2. Did you discuss any potential risks of your work?
8
- A3. Do the abstract and introduction summarize the paper’s main claims?
1
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

3

- B1. Did you cite the creators of artifacts you used?
1
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
8
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
8
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
8
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
This information is not publicly available.
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix

C Did you run computational experiments?

4

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

Appendix

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

4, Appendix

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

Appendix

D Did you use human annotators (e.g., crowdworkers) or research with human participants?

4

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

Appendix

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

8

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

8

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

No response.

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

The data sources we used did not reveal this information.