Multi-modal Action Chain Abductive Reasoning

Mengze Li[1], **Tianbao Wang**[1], **Jiahe Xu**[1], **Kairong Han**[1], **Shengyu Zhang**[1], **Zhou Zhao**[1*], **Jiaxu Miao**[1*], **Wenqiao Zhang**[1*], **Shiliang Pu**[4], **Fei Wu**[2,3]
[1]Zhejiang University [2]Shanghai Institute for Advanced Study of Zhejiang University
[3]Shanghai AI Laboratory [4]Hikvision Research Institute

Abstract

Abductive Reasoning, has long been considered to be at the core ability of humans, which enables us to infer the most plausible explanation of incomplete known phenomena in daily life. However, such critical reasoning capability is rarely investigated for contemporary AI systems under such limited observations. To facilitate this research community, this paper sheds new light on *Abductive Reasoning* by studying a new vision-language task, **M**ulti-modal **A**ction chain abductive **R**easoning (**MAR**), together with a large-scale *Abductive Reasoning* dataset: Given an incomplete set of language described events, MAR aims to imagine the most plausible event by spatio-temporal grounding in past video and then infer the hypothesis of subsequent action chain that can best explain the language premise. To solve this task, we propose a strong baseline model that realizes MAR from two perspectives: (i) we first introduce the transformer, which learns to encode the observation to imagine the plausible event with explicitly interpretable event grounding in the video based on the commonsense knowledge recognition ability. (ii) To complete the assumption of a follow-up action chain, we design a novel symbolic module that can complete strict derivation of the progressive action chain layer by layer. We conducted extensive experiments on the proposed dataset, and the experimental study shows that the proposed model significantly outperforms existing video-language models in terms of effectiveness on our newly created MAR dataset. Our dataset is available [1].

First author.
mengzeli@zju.edu.cn

*Corresponding author.
zhaozhou@zju.edu.cn
jiaxu.miao@yahoo
wenqiaozhang@zju.edu.cn
[1]https://github.com/BirdFly16/TO-MAR

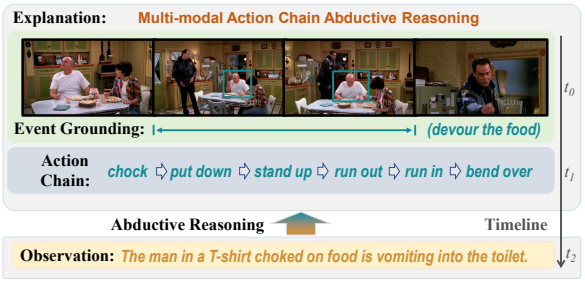

Figure 1: A diagram of the MAR task.

1 Introduction

Abductive Reasoning typically begins with an incomplete observation or several observations and then proceeds to the likeliest possible explanation for the set (Du et al., 2021; Peirce, 1974). Given an event observation ($\mathcal{O}$), humans can find some related information in the recollection and easily trace the complete process with strong reasoning ability as a hypothesis ($\mathcal{H}$) to explain the observation. For instance, when we observe the $\mathcal{O}$: "The man in a T-shirt chocked on food is vomiting into the toilet" and remember that he used to devour food in the kitchen, we could infer the complete event chain about that man as hypothesis $\mathcal{H}$: the man devoured the food in the kitchen $\rightarrow$ he was choking $\rightarrow$ he put down the food $\rightarrow$ he left the kitchen in a hurry $\rightarrow$ he ran into the bathroom $\rightarrow$ he bent over the toilet. This ability enables us to perform better than machines in high-level reasoning and would be the most precious capacity for modern AI. Therefore, it is important to enhance such *Abductive Reasoning* capacities of AI models, *i.e.*, complete process as explanation.

Motivated by the aforementioned *Abductive Reasoning* scenario, we present a novel vision-language task, called **M**ulti-modal **A**ction chain abductive **R**easoning (**MAR**), which is illustrated in Figure 1. Specifically, given a set of language-described observations, the MAR task targets to precisely localize the target event in the past video

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4617–4628
July 9-14, 2023

(visual recollection simulation of human) about the language-described person and rigorously reason out the subsequent action chain (subsequent events inference), to explain the observation. Different from the previous *Abductive Reasoning* task (Bhagavatula et al., 2019) focusing on the unimodal and partial reasoning, our new task has the following characteristics: **(i)** MAR needs to locate the target event from the complex video information to explain the textual observation; **(ii)** MAR requires rigorous recovery of the complete action chain.

These characteristics introduce two challenges to the MAR task: **(1) Heterogeneous Information Alignment.** To realize the event grounding, aligning the cross-modal information is necessary. However, unlike the highly concise language description, the videos in real scenes usually contain complex and redundant information, including multiple people with different appearances, actions, scenes, etc. Only a small amount of information in videos aligns with the text-described observation. Precisely extracting information from the complicated video information to align is difficult but necessary for the AI system. **(2) Action Chain Reasoning.** Rigorous action chain reasoning is an interlocking and progressive process. If one step of reasoning is wrong, the correctness of subsequent steps cannot be guaranteed. Therefore, for action chain reasoning, it is highly required to correctly learn the logical relationship between actions and correctly select from multiple next-step actions in each step of reasoning.

We contribute a carefully annotated large-scale dataset, TO-MAR, based on our collected data to facilitate the challenges solved for the MAR task. It contains $14,201$ cross-modal examples based on the videos manually collected from the TV show and the existing dataset (Sigurdsson et al., 2016). To address the MAR task challenges, these examples have targeted manual annotations: **(i) Commonsense Knowledge Annotation for Assisted Alignment.** We provide the full annotation of commonsense knowledge for every textual observation related person in the large-scale videos, including the character's appearance, clothing, actions, sentiment, etc. **(ii) Rigorous Annotation for Action Chains.** Expert annotators with strong logical ability are asked to annotate the language-described observations and the action chains, ensuring the accuracy and rigor of logical annotations.

Based on the constructed dataset, we propose an end-to-end **N**eural-symb**O**lic model **V**ia commonsense knowledg**E** for multi-moda**L** action chain *Abductive Reasoning* (**NOVEL**). There are two key targeted designs: **(i) Knowledge-guided Alignment.** We adopt the multi-task learning paradigm to synchronize the recognition learning of commonsense knowledge. Based on such knowledge recognition ability, our NOVEL can minimize the interference of the inferred event and past video, thereby more easily learning to generate explicit event grounding conditioned on textual observations. **(ii) Graph-aware Symbolic Reasoning.** Motivated by the powerful reasoning ability of the symbolic network (Yi et al., 2018), we design the targeted symbolic reasoning module based on the traditional graph theory. Specifically, we store the learned action association graph in the training process. During inference, we determine the intermediate action chain between the textual observation and the grounded video event with Dijkstra's algorithm (Dijkstra, 1959). Our contributions are three-fold:

- We introduce a new task, **M**ulti-modal **A**ction chain abductive **R**easoning (**MAR**), which includes two sub-parts: target event grounding and sequential action chain reasoning.
- A carefully collected large-scale dataset, TO-MAR, is provided, in which the complete observation-explanation pairs are accurately and rigorously annotated. In addition, a variety of commonsense knowledge that can aid in training is annotated in detail.
- An end-to-end neural-symbolic model named NOVEL is proposed for MAR with knowledge-guided alignment and graph-based symbolic reasoning. Extensive experiments demonstrate the model design rationality.

2 Related Work

Multi-modal Spatio-temporal Grounding. Our MAR task is related to the multi-modal spatio-temporal grounding task, which aims to detect target visual information described by the sentence from the video. It is an important task in the visual understanding domain (Miao et al., 2023, 2021; Zhang et al., 2019, 2022b,a). For video grounding research direction, most researchers focus their research on temporal grounding task (Yang et al., 2021; Xiao et al., 2021). However, spatio-temporal

grounding and spatial grounding (Li et al., 2022a; Yang et al., 2022; Jin et al.; Su et al., 2021; Li et al., 2022b, 2023) have received less attention. (Zhang et al., 2020) uses the graph neural network to model the spatio-temporal relationship between objects to align text descriptions for object localization. In addition, to evaluating the model performance, this paper proposes a complete large-scale dataset. (Su et al., 2021) designs an end-to-end multi-modal grounding model based on the transformer. It outperforms all previous models without pre-training. Later, (Yang et al., 2022) makes a targeted design to fit the pre-trained parameters and achieves a great improvement in accuracy.

Neural-symbolic Reasoning. Compared with the pure neural network (Li et al., 2020b; Wu et al., 2022; Li et al., 2020a; Wu et al., 2020; Miao et al., 2022), neural-symbolic models have stronger inference and perception capabilities. (Yi et al., 2018) is an earlier paper exploring this direction. It stitches symbolic models behind the multi-modal neural network to reason on the information the network perceives. In this neat way, the model achieves excellent results. (Li et al., 2020c) combines the laws of physics with deep learning to make models capable of fitting complex physical processes. In this way, the model can effectively predict the motion trends of objects in the physical world. Similarly, (Ding et al., 2021) uses physical laws such as collision to design a symbolic model to process the information perceived by the neural network, which can effectively predict the future motion of objects such as balls or sliders. (Greff et al., 2019) applies neural network and symbolic model to high-level and low-level visual relation detection, respectively, and achieves good performance through the cooperation of the two.

Abductive Reasoning. There is a limited amount of existing research work on abductive reasoning AI systems (Du et al., 2021). Previous abductive reasoning tasks (Bhagavatula et al., 2019; Liang et al., 2022) require AI systems to provide a unimodal and partial explanation, which may lack some key information.

3 Dataset Description

To advance research on *Abductive Reasoning*, we propose the **M**ulti-modal **A**ction chain abductive **R**easoning task (**MAR**) and contribute a large-scale **T**ext-vide**O** dataset for the **MAR** task (**TO-MAR**). To complete the MAR task, the AI system needs to reason out the complete process (the video event and the subsequent action chain) to explain the observed event described by the textual observation $\mathcal{O}$. In detail, the target event $\mathcal{E}_t$ is grounded in the video $\mathcal{V}$ by localizing the temporal boundary $\mathcal{T}$ and the target person bounding boxes $\mathcal{B}$ in the event $\mathcal{E}_t$. After that, the action chain $\mathcal{A} = \{\mathcal{A}_i\}_{i=1}^{N_\mathcal{A}}$ following the target event $\mathcal{E}_t$ is inferred, where the $N_\mathcal{A}$ is the number of actions in chain $\mathcal{A}$.

Table 1: Statistics of commonsense knowledge types for the TO-MAR dataset.

Category	Subcategory
Appearance	Gender, Hair Length, Age
Clothing	Length of Lower-body Clothing, Type of Lower-body Clothing, Type of Upper-body Clothing, Sleeve Length, 3 Other Outfits, 9 Colors of Upper-body Clothing, 9 Colors of Lower-body Clothing
Action	Intransitive Verb, Transitive Verb, Object
Sentiment	None
Scene	None

3.1 Dataset Preparation

We collect and annotate the proposed TO-MAR dataset based on the above MAR's definition. To increase the variety of data, a two-source dataset collection is conducted for labeling: (1) We select lifestyle videos from the Charades dataset (Sigurdsson et al., 2016). These examples contain diverse people and rich activities. (2) The TV show videos are selected from 92 well-known American dramas, such as The Big Bang Theory, Grey's Anatomy, etc. Notably, the number of people is relatively limited compared to the first source, but the causal links among events are clearer.

3.2 Dataset Annotation

The TO-MAR dataset contains commonsense knowledge annotations and MAR task annotations. More details are in the appendix.

Commonsense Knowledge Annotation. We annotate the commonsense knowledge that is related to the person conditioned on the observations for event grounding, including the appearance and clothing of key video characters. We also annotated the characters' actions, sentiments and located scenes in each frame. Each category contains several critical subcategories (*e.g.*, Appearance: Gender, Hair Length, Age.) recognized

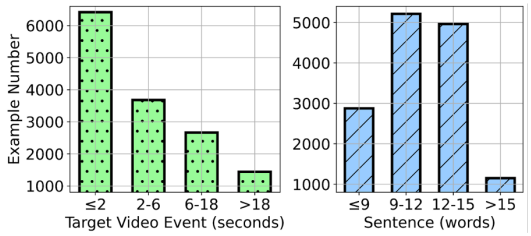

Figure 2: Statistics of the example number with different target video event lengths, sentence lengths, and action chain lengths.

by labeling experts to ensure the MAR model the relationship between the event and observations.

MAR Task Annotation. The MAR task annotation for the video $\mathcal{V}$ consists of a language-described observation $\mathcal{O}$ and its explanation, including the spatio-temporal markers (the bounding boxes $\mathcal{B}$ and the temporal boundary $\mathcal{T}$) of the target video event $\mathcal{E}_t$ and the subsequent action chain $\mathcal{A}$. In order to ensure the quality of the annotations, we employ experts with related annotation experience from leading AI research institutions. The annotation process contains two steps: **Step 1: Annotating.** The annotators annotate the natural language description $\mathcal{O}$ of the observation referred to the commonsense knowledge and the follow-up video content for the TV show video clips. The corresponding target video event $\mathcal{E}_t$ and the subsequent action chain $\mathcal{A}$ annotations are also recorded. **Step 2: Verification.** The validators carefully validate the annotated examples. If the annotation is not agreed upon by the validators, it is relabeled or dropped.

Dataset Features and Statistics. To further introduce the TO-MAR dataset, we analyze the data distribution: **(1) Dataset Split.** We separate the TO-MAR dataset into $train/val/test$ sets with $12,527/426/1,248$ labeled examples. The videos of the three sets do not overlap. More detailed statistics are shown in Figure 2. **(2) Diversity.** To include various causal relationships, our dataset contains rich activities (cooking, work, etc.) and various scenes (family, hospital, etc.). **(3) Large-Scale.** The TO-MAR dataset consists of $14,201$ examples, which proves a testbed for the evaluation of the MAR task models.

4 Method

MAR Task Formulation. Given the language description of the observed event $\mathcal{O}$ and the video $\mathcal{V}$ consisting of $N_{\mathcal{E}}$ past events $\mathcal{E} = \{\mathcal{E}_i\}_{i=1}^{N_{\mathcal{E}}}$, the **M**ulti-modal **A**ction chain abductive **R**easoning task (**MAR**) aims to explain the observation $\mathcal{O}$ by localizing the target event (the t-th event) $\mathcal{E}_t$ in the video $\mathcal{V}$, and supplementing the intermediate action chain $\mathcal{A} = \{\mathcal{A}_i\}_{i=1}^{N_{\mathcal{A}}}$ between the target event $\mathcal{E}_t$ and observation $\mathcal{O}$. All actions in the action chain $\mathcal{A}$ are chosen from the action set $\mathcal{S} = \{\mathcal{S}_i\}_{i=1}^{N_{\mathcal{S}}}$. Meanwhile, the $\mathbf{N}_{\mathcal{A}}$ and the $\mathbf{N}_{\mathcal{S}}$ are the action number in the action chain $\mathcal{A}$ and the action set $\mathcal{S}$, respectively. To formulate the model training process, we define $\mathcal{M}$ as the trained model initialized with parameter Θ. Then, the training optimization function can be expressed as:

$$\mathcal{M}((\mathcal{O},\mathcal{V}),(\mathcal{E},\mathcal{S});\Theta) = \max_{\Theta} \epsilon(\xi(\mathcal{E},\mathcal{S}),\delta(\mathcal{O},\mathcal{V};\Theta)). \quad (1)$$

In it, the function $\xi(.)$ outputs the ground truth, which contains: **(1)** the temporal boundary $\mathcal{T}$ of the target video event $\mathcal{E}_t$ and the target person bounding boxes $\mathcal{B}$ in the event $\mathcal{E}_t$; **(2)** the category of each action in the action chain $\mathcal{A}$. The $\delta(.)$ outputs the prediction, and the Θ is the learnable parameter. The function $\epsilon(.)$ calculates the consistency of $\xi(.)$ and $\delta(.)$.

Model Pipeline. As shown in the Figure 3, after extracting features from the observation $\mathcal{O}$ and the frames of the video $\mathcal{V}$ with RoBERTa (Liu et al., 2019) and Resnet101 (He et al., 2016), our NOVEL $\mathcal{M}$ mainly contains two parts to process the features. **(1)** First, the multi-modal features is reasoned by the transformer model (Vaswani et al., 2017). Based on the recognition ability of the commonsense knowledge (*e.g.*, the character's actions, appearance, etc.), the model focuses on the key video information aligned with the textual observation $\mathcal{O}$, and learns to infer the temporal boundary $\mathcal{T}$ and the target bounding boxes $\mathcal{B}$ of the target event $\mathcal{E}_t$ in the video. **(2)** Second, the symbolic reasoning part maintains a relation memory module and stores the learned action relation in it at the training step. In the inference phase, the action graph is constructed based on the action relations. We use Dijkstra's algorithm to find the connected path on the action graph so as to infer the action chain $\mathcal{A}$ between the observation $\mathcal{O}$ and the target video event $\mathcal{E}_t$.

4.1 Knowledge-guided Alignment

We follow the general training and prediction protocol of cross-modal transformer applied in other video grounding methods (Kamath et al., 2021;

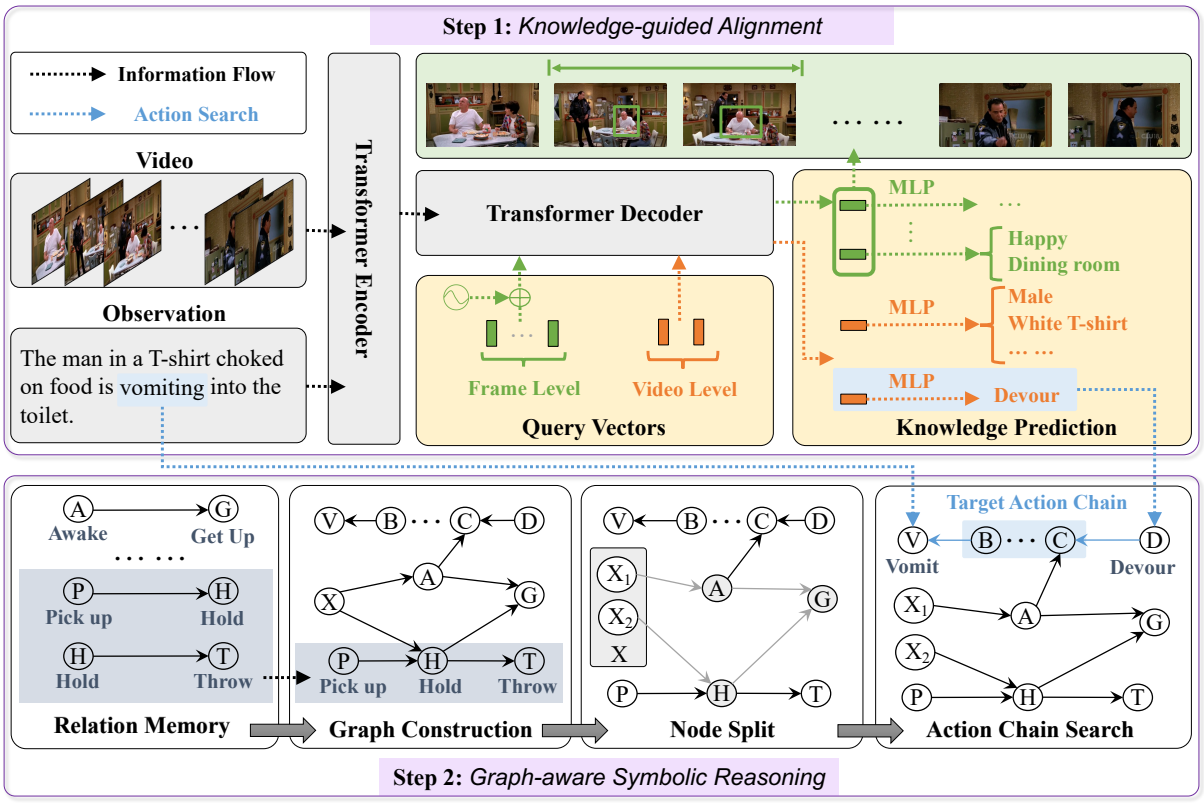

Figure 3: The diagram of our proposed NOVEL for MAR task. (1) The transformer-based neural network learns to ground the target video event $\mathcal{E}_t$ and predicts the language-described person's target action, based on the commonsense knowledge recognition ability. (2) We construct the action graph from the learned action relation memory and reason out the connection path between the start node (the video action) and the end node (the observation action) with Dijkstra's algorithm.

Yang et al., 2022). The transformer decoder generates the features for all video frames. For each frame, we predict the bounding boxes and whether it is the temporal start or end of the target event $\mathcal{E}_t$.

However, there is complex information in the input video $\mathcal{V}$. We need to guide the model to focus on the video information aligned with the observation $\mathcal{O}$ during the training process, using the prediction learning for the commonsense knowledge (*e.g.*, human action, appearance, etc.) of the target character in the video. Specifically, following previous transformer-based models (Yang et al., 2022; Carion et al., 2020), several query vectors are defined: frame level vectors $\mathbf{Q}_r = \{\mathbf{q}_r^i\}_{i=1}^{N_r}$ and video level vectors $\mathbf{Q}_v = \{\mathbf{q}_v^i\}_{i=1}^{2}$. The N_r is the frame number in the video $\mathcal{V}$. With these query vectors, we apply transformer decoder $\mathcal{D}$ to analyze multi-modal features $\mathbf{F}$ fused by the transformer encoder:

$$[\mathbf{F}_r, \mathbf{F}_v] = \mathcal{D}([\mathbf{Q}_r, \mathbf{Q}_v], \mathbf{F}), \tag{2}$$

where [.] represents the feature concatenated.

Next, we predict the commonsense knowledge of the language-described character using these output frame level features $\mathbf{F}_r = \{\mathbf{f}_r^i\}_{i=1}^{N_r}$ and video level features $\mathbf{F}_v = \{\mathbf{f}_v^i\}_{i=1}^{2}$. The scene where the character is located and the character's sentiment change over time. Thus, we predict them frame by frame using **M**ulti**L**ayer **P**erceptron (**MLP**). Viewing the i-th frame as an example, the prediction process is represented as:

$$\mathbf{p}_{sc}^i = \text{softmax}(MLP_{sc}(\mathbf{f}_r^i)), \tag{3}$$

$$\mathbf{p}_{se}^i = \text{softmax}(MLP_{se}(\mathbf{f}_r^i)). \tag{4}$$

In them, the MLP_{sc} and MLP_{se} are the MLP applied to predict the probabilities of all scene and sentiment classes ($\mathbf{p}_{sc}^i$ and $\mathbf{p}_{se}^i$). In addition, we apply the video level features $\mathbf{F}_v$ to predict the appearance and clothing of the target character described by the language sentence $\mathcal{O}$, and the character's action in the target video event $\mathcal{E}_t$:

$$\mathbf{p}_{ap} = \text{softmax}(MLP_{ap}(\mathbf{f}_v^1)) \tag{5}$$

$$\mathbf{p}_{cl} = \text{softmax}(MLP_{cl}(\mathbf{f}_v^1)), \tag{6}$$

$$\mathbf{p}_{ac} = \text{softmax}(MLP_{ac}(\mathbf{f}_v^2)). \quad (7)$$

In them, the MLP_{ap}, MLP_{cl} and MLP_{ac} are the MLP applied to predict the probabilities of all appearance, clothing, and action classes ($\mathbf{p}_{ap}$, $\mathbf{p}_{cl}$, and $\mathbf{p}_{ac}$), respectively. We employ the cross-entropy loss function to train the prediction of commonsense knowledge.

4.2 Graph-aware Symbolic Reasoning

The prediction for the action chain $\mathcal{A}$ requires rigorous layer-by-layer logical reasoning ability. Considering the symbolic network's reasoning ability (Yi et al., 2018; Li et al., 2020c), we design a traditional graph theory based symbolic module for searching targeted nodes (actions).

In the training phase, we divide the action chain annotation labeled for each training example into several single-step action mappings and store them in the action relation memory module. During the process, new actions are continuously introduced. We initialize the prototype feature $\mathbf{f}_p^i$ for the newly added action category with index i. In addition, at each step of the relation memory update, we construct the action graph based on the action relations. There may be N_d different connected paths between two nodes, which may result in the predicted action chain $\mathcal{A}$ not unique. To address the problem, we replace the starting node $\mathbf{f}_p^s$ with N_d different nodes $\{\mathbf{f}_p^{s_j}\}_{j=1}^{N_d}$ and view them as the starting of each path.

During the inference process, the action node described by the textual observation $\mathcal{O}$ is detected from the action graph. Using the traditional graph theory algorithm, Dijkstra (Dijkstra, 1959), we find all paths ending at this node. Assuming that there are N_s nodes on these paths, we view them as the candidate starting nodes and calculate the probability of being selected $\mathbf{p}_{sn} = \{p_{sn}^i\}_{i=1}^{N_s}$, using the transformer decoder predicted feature $\mathbf{f}_v^2$:

$$\mathbf{p}_{sn} = \text{softmax}([\mathcal{S}(\mathbf{f}_p^1, \mathbf{f}_v^2), \mathcal{S}(\mathbf{f}_p^2, \mathbf{f}_v^2), ..., \mathcal{S}(\mathbf{f}_p^{N_s}, \mathbf{f}_v^2)]), \quad (8)$$

where the $\mathcal{S}(.)$ is the similarity calculation function. The node with max probability is chosen out from the candidate starting nodes. With the starting and ending nodes, the actions corresponding to the middle nodes between them constitute the intermediate action chain $\mathcal{A}$.

5 Experiments

We evaluate the effectiveness of the proposed NOVEL on TO-MAR dataset, followed by a discussion of NOVEL's property with controlled studies.

Table 2: Compared with baselines on TO-MAR.

Methods	m_vIoU	vIoU@0.3	vIoU@0.5	ACC
STVGBert	8.5	10.1	3.6	-
IT-OS	13.4	16.5	9.7	-
TubeDETR	15.7	18.7	9.9	-
Cycle_C	-	-	-	58.5
FUTR	-	-	-	62.1
NOVEL	**18.2**	**24.4**	**13.9**	**72.0**

5.1 MAR Experiments

Implement Detail. Our model is implemented based on the PyTorch framework, which is trained on a Linux server. The implementation of the transformer part is based on the TubeDETR (Yang et al., 2022). For the training data, we randomly rotate and resize the input frames. In addition, random horizontal flips and size cropping are applied during the video frame preprocessing. For the validation and test data, we only normalize and randomly resize each frame. In the training process, the batch size is 1 and the random seed is 42. The learning rate is set to 0.00005 and the weight decay is 0.0001. All experimental environments are deployed in Hikvision (https://www.hikvision.com/en/).

Evaluation Metrics. Following the evaluation protocols of the spatio-temporal grounding (Su et al., 2021), we adopt **m_vIoU** and **vIoU@R** to evaluate the model performance. The vIoU is calculated by $\frac{1}{|S_p \cup S_{gt}|} \sum_{n \in S_p \cap S_{gt}} IoU(\hat{b}_t, b_t)$. In it, the S_p and the S_{gt} are the frame sets in the predicted and ground truth tubes, respectively. The $\hat{b}_t$ and the b_t are the predicted and ground truth bounding boxes of the frame t. The vIoU@R is the ratio of samples whose vIoU>R. The m_vIoU is the mean vIoU of all samples. In addition, we adopt the action chain accuracy (**ACC**) to evaluate the model performance for the action chain prediction.

Baselines. Existing methods of other tasks cannot be transferred directly to our MAR task. Thus, we extend several SOTA multi-modal and reasoning models as the baselines to compare. In detail, for a comprehensive comparison, we consider: (1) multi-modal video grounding methods, including TubeDETR (Yang et al., 2022), IT-OS (Li et al., 2022a), and STVGBert (Su et al., 2021); (2) action chain prediction methods, Cycle_C (Farha et al., 2020) and FUTR (Gong et al., 2022).

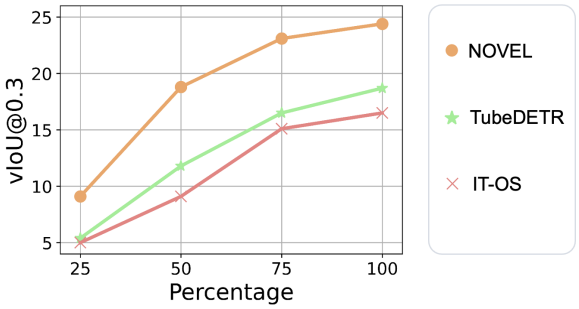

Figure 4: Compared with the baselines under different proportions of the training set.

Table 3: Ablation study on the TO-MAR dataset. * represents that the ablation model's grounding part and action chain reasoning part are trained separately.

Methods	m_vIoU	vIoU@0.3	vIoU@0.5	ACC
Base*	16.4	18.4	10.8	62.1
(+Δ_{kno})*	18.2	24.4	13.9	62.1
Base	12.8	15.9	8.7	59.7
+Δ_{kno}	16.1	19.8	11.8	60.9
+Δ_{sym}	16.4	18.4	10.8	67.6
+Δ_{kno}+Δ_{sym} (NOVEL)	**18.2**	**24.4**	**13.9**	**72.0**

Performance Comparison. We compare our NOVEL model with the baselines on the TO-MAR dataset for the MAR task. The experiment results are shown in Table 2. From it, we can observe that our NOVEL model performs better than all previous methods. Specifically, compared to the previous state-of-the-art, TubeDETR, the NOVEL significantly improves the target event grounding (vIoU@0.3) from 18.7 to 24.4. In addition, the NOVEL model improves the action chain prediction from 62.1 to 72.0 compared with the best performance baseline, FUTR. We attribute the performance improvement of our model to commonsense knowledge-driven perception design. It helps the model focus on the correct visual semantics aligned with the textual observation. The graph symbol model rigorously describes the logical relationship between actions, and Dijkstra's algorithm accurately reasons out the action chain.

Comparison using Different Training Data Volumes. We are interested in how the NOVEL model performance varies with the amount of training data. To this end, we randomly select different proportions of examples from the training set and compare our NOVEL model with several state-of-the-arts trained on them. The experimental results are shown in Figure 4. From it, we can observe that the NOVEL model performance is always the best under different data volumes. Based on the commonsense knowledge recognition ability, the NOVEL model can eliminate the interference of irrelevant information on training, so that the model can learn the target video event grounding more effectively. Even if the training set is small, the NOVEL model still has higher accuracy.

Ablation Study. To fully evaluate the NOVEL model's effectiveness, we need to understand how different components contribute. The new architectures are constructed by removing several components from the NOVEL. The investigated building blocks include the knowledge-guided alignment and the graph-aware symbolic reasoning module. For convenience, we use Δ_{kno} and Δ_{sym} to represent these key components. After removing, the state-of-the-art model, FUTR, compensates for the lack of action chain prediction functionality.

The experiment results are shown in Table 3. From it, we can observe the following: **(1)** We evaluate our ablation grounding models and the FUTR model training together or not (labeled without or with * in the table). Notably, when training separately, the Δ_{kno} is bound to the ablation grounding models, to prove its contribution to the NOVEL's module. When training together, the pure neural network models fail to effectively capture the correlation between the target video event grounding and the action chain prediction. Even in the training process, the learning of the two interferes with each other, which leads to a loss of precision. **(2)** The NOVEL model performs better after adding each building block. It reveals the reasonable design of these key modules. **(3)** When the knowledge-guided alignment module and the graph-aware symbolic module are used together, the performance of the action chain reasoning is better. The results demonstrate the knowledge-guided alignment design aids in the extraction of knowledge useful for reasoning, and the symbolic reasoning module can complete accurate reasoning on this basis.

Case Study. A case study is conducted to demonstrate the NOVEL's capability in visuals. In detail, two examples are sampled from our TO-MAR dataset. A comparison of the NOVEL model with the state-of-the-arts, TubeDETR and FUTR, is necessary to fully demonstrate model performance on these examples. The experiment results are visualized in Figure 5. From the figure, we can find that our NOVEL model predicts the target video

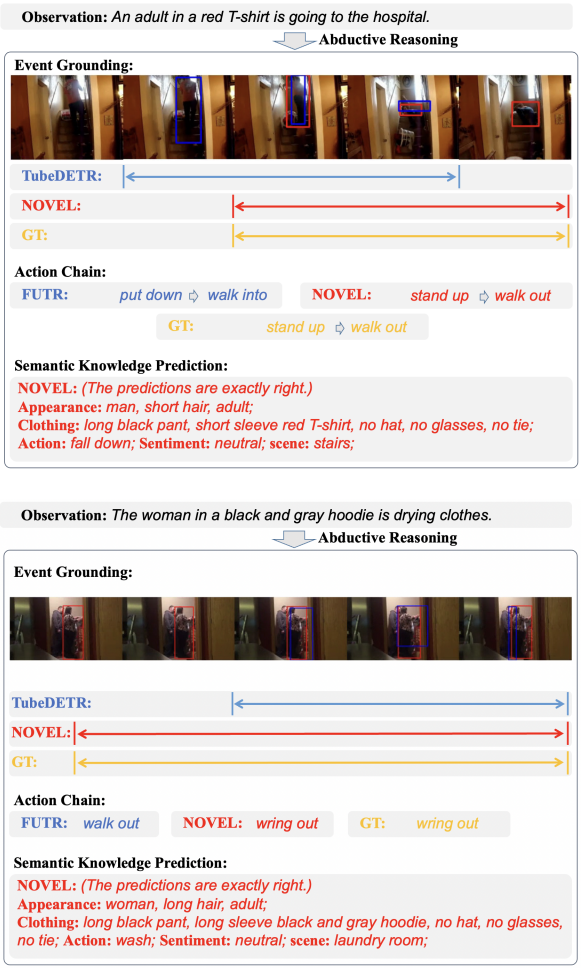

Figure 5: Two examples of the NOVEL model and the baselines (TubeDETR and FUTR) predictions. "**GT**" represents the **G**round **T**ruth.

event and the action chain accurately. In addition, its commonsense knowledge prediction is also correct. In contrast, the baseline predictions are not as satisfactory. This intuitively reflects that our neural-symbolic model, NOVEL, is reasonably designed for the multi-modal action chain *Abductive Reasoning* task.

5.2 Spatio-Temporal Grounding Experiments

Based on our proposed TO-MAR dataset, the models suitable for another similar language-vision understanding task, **M**ulti-modal **S**patio-**T**emporal **G**rounding (**MSTG**), can be evaluated. This task aims to detect the spatio-temporal tube described by the concise language sentence from the complex video content (Su et al., 2021). We further evaluate the effectiveness of the knowledge-guided alignment in our NOVEL model on this task.

Dataset. $9,143$ labels for this task based on the TO-MAR dataset are annotated by annotators. Each label contains a natural language query, video clip temporal boundaries, and target object bounding boxes. We split all examples into $7,845/277/1,021$ (train/val/test) without overlap and name this dataset as **TO-MSTG**. We describe the annotation method in detail in the appendix. In the future, we will further expand the data scale.

Table 4: Compared with baselines on TO-MSTG.

Methods	m_vIoU	vIoU@0.3	vIoU@0.5
STVGBert	4.7	3.8	1.0
IT-OS	8.0	9.0	3.1
TubeDETR	8.9	9.5	4.5
NOVEL	**11.7**	**12.0**	**7.0**

Performance Comparison. The experiment results are shown in Table 4. From it, we can observe that our NOVEL model performs best compared with the other three baselines. Specifically, it improves the accuracy (m_vIoU/vIoU@0.3/vIoU@0.5) from $8.9/9.5/4.5$ to $11.7/12.0/7.0$. This again demonstrates the power of commonsense knowledge guidance for the heterogeneous information alignment problem. In addition, the knowledge-guided alignment design generalizes effectively to different AI tasks, where heterogeneous alignment problem exists.

6 Conclusion

In this paper, we propose a new task, multi-modal action chain abductive reasoning, to promote the development of the abductive field. This cross-modal task targets to reason out a more complete explanation (explanation event grounding and sequential action chain inference) than the previous abductive reasoning tasks. Furthermore, we propose a large-scale dataset (TO-MAR) and a neural-symbolic model via commonsense knowledge (NOVEL) for our new task as a strong baseline. Extensive experiments on the TO-MAR dataset and the TO-MSTG dataset demonstrate the effectiveness of our NOVEL model.

Limitations

This work is currently limited to the action chain as the abstract summary of the complete explanation for the given limited observation. In the future, we will further upgrade this task, *e.g.*, considering the progressive textual descriptions as the complete explanation. We hope our work can advance the

reasoning AI system research community.

7 Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 62037001, the National Key RD Program of China under Grant No.2022ZD0162000, the National Natural Science Foundation of China under Grant No. 62222211 and Grant No.61836002. In addition, our research is funded by the Starry Night Science Fund at Shanghai Institute for Advanced Study (Zhejiang University).

References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

EW Dijkstra. 1959. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271.

Mingyu Ding, Zhenfang Chen, Tao Du, Ping Luo, Josh Tenenbaum, and Chuang Gan. 2021. Dynamic visual reasoning by learning differentiable physics models from video and language. *Advances In Neural Information Processing Systems*, 34:887–899.

Li Du, Xiao Ding, Ting Liu, and Bing Qin. 2021. Learning event graph knowledge for abductive reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5181–5190. Association for Computational Linguistics.

Yazan Abu Farha, Qiuhong Ke, Bernt Schiele, and Juergen Gall. 2020. Long-term anticipation of activities with cycle consistency. *arXiv preprint arXiv:2009.01142*.

Dayoung Gong, Joonseok Lee, Manjin Kim, Seong Jong Ha, and Minsu Cho. 2022. Future transformer for long-term action anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3052–3061.

Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. 2019. Multi-object representation learning with iterative variational inference. In *International Conference on Machine Learning*, pages 2424–2433. PMLR.

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

Yang Jin, Zehuan Yuan, MU Yadong, et al. Embracing consistency: A one-stage approach for spatio-temporal video grounding. In *Advances in Neural Information Processing Systems*.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790.

Mengze Li, Ming Kong, Kun Kuang, Qiang Zhu, and Fei Wu. 2020a. Multi-task attribute-fusion model for fine-grained image recognition. In *Optoelectronic Imaging and Multimedia Technology VII*, volume 11550, pages 114–123. SPIE.

Mengze Li, Kun Kuang, Qiang Zhu, Xiaohong Chen, Qing Guo, and Fei Wu. 2020b. Ib-m: A flexible framework to align an interpretable model and a black-box model. In *2020 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 643–649. IEEE.

Mengze Li, Han Wang, Wenqiao Zhang, Jiaxu Miao, Wei Ji, Zhou Zhao, Shengyu Zhang, and Fei Wu. 2023. Winner: Weakly-supervised hierarchical decomposition and alignment for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Mengze Li, Tianbao Wang, Haoyu Zhang, Shengyu Zhang, Zhou Zhao, Jiaxu Miao, Wenqiao Zhang, Wenming Tan, Jin Wang, Peng Wang, et al. 2022a. End-to-end modeling via information tree for one-shot natural language spatial video grounding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8707–8717.

Mengze Li, Tianbao Wang, Haoyu Zhang, Shengyu Zhang, Zhou Zhao, Wenqiao Zhang, Jiaxu Miao, Shiliang Pu, and Fei Wu. 2022b. Hero: Hierarchical spatio-temporal reasoning with contrastive action correspondence for end-to-end video object grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3801–3810.

Yunzhu Li, Toru Lin, Kexin Yi, Daniel Bear, Daniel Yamins, Jiajun Wu, Joshua Tenenbaum, and Antonio Torralba. 2020c. Visual grounding of learned physical models. In *International conference on machine learning*, pages 5927–5936. PMLR.

Chen Liang, Wenguan Wang, Tianfei Zhou, and Yi Yang. 2022. Visual abductive reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15565–15575.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. 2022. Large-scale video panoptic segmentation in the wild: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21033–21043.

Jiaxu Miao, Yunchao Wei, Xiaohan Wang, and Yi Yang. 2023. Temporal pixel-level semantic understanding through the vspw dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. 2021. Vspw: A large-scale dataset for video scene parsing in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4133–4143.

Charles Sanders Peirce. 1974. *Collected papers of charles sanders peirce*, volume 5. Harvard University Press.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer.

Rui Su, Qian Yu, and Dong Xu. 2021. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1533–1542.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-biased court's view generation with causality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 763–780.

Yiquan Wu, Yifei Liu, Weiming Lu, Yating Zhang, Jun Feng, Changlong Sun, Fei Wu, and Kun Kuang. 2022. Towards interactivity and interpretability: A rationale-based legal judgment prediction framework. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4787–4799.

Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. 2021. Boundary proposal network for two-stage natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2986–2994.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16442–16453.

Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1–10.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31.

Wenqiao Zhang, Jiannan Guo, Mengze Li, Haochen Shi, Shengyu Zhang, Juncheng Li, Siliang Tang, and Yueting Zhuang. 2022a. Boss: Bottom-up cross-modal semantic composition with hybrid counterfactual training for robust content-based image retrieval. *arXiv preprint arXiv:2207.04211*.

Wenqiao Zhang, Siliang Tang, Yanpeng Cao, Shiliang Pu, Fei Wu, and Yueting Zhuang. 2019. Frame augmented alternating attention network for video question answering. *IEEE Transactions on Multimedia*, 22(4):1032–1041.

Wenqiao Zhang, Lei Zhu, James Hallinan, Shengyu Zhang, Andrew Makmur, Qingpeng Cai, and Beng Chin Ooi. 2022b. Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20666–20676.

Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. 2020. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10668–10677.

ACL 2023 Responsible NLP Checklist

A For every submission:

☐ A1. Did you describe the limitations of your work?
Not applicable. Left blank.

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
section 1

☐ A4. Have you used AI writing assistants when working on this paper?
Not applicable. Left blank.

B ☐ Did you use or create scientific artifacts?

Not applicable. Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ 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)?
Not applicable. Left blank.

☐ 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?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ 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.
Section 3

C ☑ Did you run computational experiments?

Section 5

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. Left blank.

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?
Not applicable. Left blank.

☐ 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?
Not applicable. Left blank.

☑ 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.)?
Section 5

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 3

☐ 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.?
Not applicable. Left blank.

☐ 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)?
Not applicable. Left blank.

☐ 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?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.