Combining Tradition with Modernness: Exploring Event Representations in Vision-and-Language Models for Visual Goal–Step Inference

Chong Shen and **Carina Silberer**
Institute for Natural Language Processing, University of Stuttgart, Germany
{chong.shen,carina.silberer}
@ims.uni-stuttgart.de

Abstract

Procedural knowledge understanding underlies the ability to infer goal–step relations. The task of Visual Goal–Step Inference addresses this ability in the multimodal domain. It requires the identification of images that depict the steps necessary to accomplish a textually expressed goal. The best existing methods encode texts and images either with independent encoders, or with object-level multimodal encoders using blackbox transformers. This stands in contrast to early, linguistically inspired methods for event representations, which focus on capturing the most crucial information, namely actions and participants, to learn stereotypical event sequences and hence procedural knowledge. In this work, we study various methods and their effects on procedural knowledge understanding of injecting the early shallow event representations to nowadays multimodal deep learning-based models. We find that the early, linguistically inspired methods for representing event knowledge do contribute to understand procedures in combination with modern vision-and-language models. This supports further exploration of more complex event structures in combination with large language models.[1]

1 Introduction

Procedural Knowledge Understanding (PKU) implies reasoning about how to complete a task or achieve a goal (Mujtaba and Mahapatra, 2019). While previous works focus on plain texts (Yang and Nyberg, 2015; Zhou et al., 2019; Zhang et al., 2020a,b; Lyu et al., 2021; Sun et al., 2022), recent studies extend the task to the visual–linguistic domain. They ground procedural everyday tasks in the visual world, as a step towards situated procedural understanding in the real world.

Yang et al. (2021) propose a novel PKU task that utilizes both textual and visual information by selecting an image conditioned on a sentence which describes a high-level goal (illustrated in Figure 1, cf. Section 2.2). Their experimental results show that there is still a large gap to human performance on this task. While Yang et al. (2021) represent goal descriptions by their neural embeddings, earlier approaches to representing procedural knowledge or stereotypical event sequences (i.e., goals and steps; cf. scripts, Shank and Abelson, 1977), in contrast, focus on capturing the most essential information of events, namely the actions and their main participants (Balasubramanian et al., 2013; Pichotta and Mooney, 2014, inter alia).

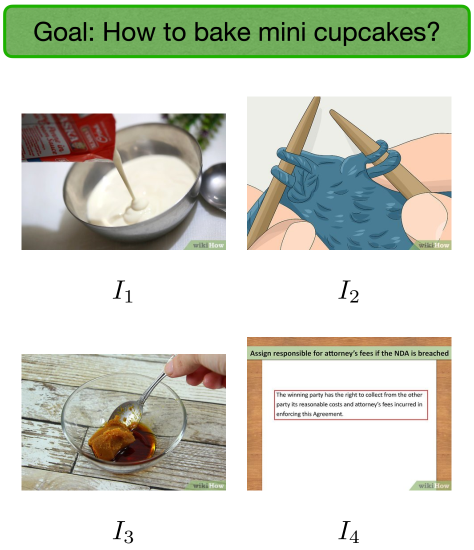

Goal: How to bake mini cupcakes?
I_1
I_2
Assign responsible for attorney's fees if the NDA is breached
The winning party has the right to collect from the other party its reasonable costs and attorney's fees incurred in enforcing this Agreement.
wikiHow
I_3
I_4

Figure 1: An example of the VGSI task. For the given goal G, image I_1 (*Combine the milk and cream before adding everything to the large bowl*) should be selected since it depicts a step S that leads to accomplishing G.

In this work, we explore different ways to inject these linguistically inspired representations to the recent powerful deep learning approaches, and study their contribution to multimodal PKU. Specifically, we investigate the relational event representation (Balasubramanian et al., 2013) and the multi-argument event representation (Pichotta

[1] The code is available at https://github.com/st143575/Exploring-Event-In-VGSI.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics - Student Research Workshop, pages 254–265
July 10-12, 2023 ©2023 Association for Computational Linguistics

and Mooney, 2014, 2016) due to their simple but condensed structure holding the most crucial information such as the action and the main participants in the main clause. We also evaluate different approaches to encode and inject such event knowledge to the model used by Yang et al. (2021), while also taking the contextual information into account. We conduct our experiments from three perspectives. First, we explore two approaches for event knowledge injection: (1) EVENT replaces the sentence describing the event by the two aforementioned event representations; (2) SENTENCE+EVENT appends the two types of event representations to the sentence describing that event. Second, we compare the embeddings extracted from different layers of the text encoder based on the finding of Jawahar et al. (2019) and Vulić et al. (2020), namely that lexical, syntactic and semantic information tend to be captured by the first, middle and last couple of layers, respectively. And third, we study the contribution of contextualised embeddings to represent the event and its participants compared to local embeddings.

The main contributions of this paper are: (1) comparison between two approaches for linguistically-inspired event knowledge injection for the task of multimodal procedural knowledge learning; (2) comparison of three levels of linguistic information in the text embedding; (3) investigation of local and contextualised event embeddings; (4) assessment of different abstract representations for the implicit subject of instructional texts.

We find that appending the multi-argument event representation to the input sentence with the <|startoftext|> token as the implicit subject, and taking the average of the last 4 hidden layers of CLIP's text encoder is the best way to encode and inject event knowledge to a deep learning model. Specifically, first encoding the full sentence and then extracting and averaging the word-level embeddings of the components of the event representation can use the contextual information in the sentence outside the event itself.

2 Related Work

2.1 Event Definitions and Representations

The concept *event* can be defined in various ways. In early works, an event is either defined as a verb (Katz and Arosio, 2001), or an expression that have implicit time dimension and is either a verb or a noun phrase (Schilder and Habel, 2001), or a proposition consisting of the subject and the predicate (Filatova and Hovy, 2001). Pustejovsky et al. (2005) define an event as a predicate describing a state or a circumstance in which something holds true. Li et al. (2021) define an event as the occurrence of an action causing a state change, which is performed by some participant(s) in a particular manner. For instance, image I_3 in Figure 1 illustrates the event of *A person beating together butter and sugar with a mixer*.

Later studies on *script learning* (Zhang, 2022) extend the definition of the event by its surrounding components in the text. Chambers and Jurafsky (2008) represent an event as a *(verb, dependency)*-pair extracted from narrative texts using a dependency parser. Balasubramanian et al. (2013) generate event schemata from news articles using *(subject, verb, object)*-pairs as the event representation. Pichotta and Mooney (2014, 2016) represent events as *(subject, verb, object, preposition)* tuples that model the interactions between entities in a script.

In contrast, recent works focus on extracting events with more complex structures and richer information from contexts. Yu et al. (2022) design a BERT-based framework for building event extractors in a weak supervised manner. Chen et al. (2021) train a multimodal Transformer (Vaswani et al., 2017) to jointly extract events from videos and texts. Wei et al. (2023) propose a framework for zero-shot event extraction using a sibling model to InstructGPT (Ouyang et al., 2022). Knowledge graphs (Hogan et al., 2021) have been widely used to extract events from multimodal data and represent events in a more complex structure (Li et al., 2020, 2022). We adopt the relational event representation of Balasubramanian et al. (2013) and the multi-argument event representation of Pichotta and Mooney (2014) for our experiments due to the low performance of recent event extractors on the dataset used for our experiments.

2.2 Procedural Knowledge Understanding

A *procedure* is a compound event that can be broken down into multiple events (Zhang, 2022). It consists of a goal and a sequence of steps towards accomplishing that goal. Procedural knowledge understanding (PKU) is the task of learning the relations between the goal and the steps. Various approaches have been proposed to understanding procedures using event knowledge. Tandon et al.

(2020) use entity tracking to generate state changes from procedural text. Zhang et al. (2020b) learn goal–step relations and step–step temporal relations in procedural texts and introduce a 4-way multiple choice task for goal–step inference. Yang et al. (2021) extend it to the multimodal domain and learn goal–step relations from texts and images. Lyu et al. (2021) generate the sequence of steps conditioned on a given goal. Zhou et al. (2022) discover the hierarchical structure in procedural knowledge using action linking. Based on the work of Yang et al. (2021), we investigate different ways to encode and inject classical event knowledge to recent deep learning models.

Goal–Step–Inference (Zhang et al., 2020b) is the task of reasoning about goal–step relations from instructional texts. Given a goal sentence and four candidate step descriptions, a model should choose the step that leads to the goal. The main challenge of this task is that it requires to understand both, the actions of goals and steps and their relations. Yang et al. (2021) extend the task to the multimodal domain through the *Visual Goal–Step Inference* task, in which steps are described by images. They attempt to overcome the challenge by matching the goal sentence and the step image. However, they still observe a significant gap between model and human performance. Our work seeks to bridge this gap with multiple approaches by combining state-of-the-art neural models with early linguistically motivated event representations (see above).

2.3 Vision-and-Language Models

In recent years, Vision-and-Language (V&L) models have made tremendous progress on a wide range of multimodal tasks, such as visual commonsense reasoning (Lu et al., 2019), image–text retrieval (Chen et al., 2020), text-to-image and image-to-text generation (Rombach et al., 2022; Li et al., 2023). One strand of models are *fusion encoders* which learn a fused representation of images and texts. For example, LXMERT (Tan and Bansal, 2019) uses attention (Vaswani et al., 2017) to learn intra-modal and cross-modal relationships while training a language encoder, an object relationship encoder and a cross-modality encoder. Although the model learns the alignment between images, objects and words in sentences via the object-level pretraining objectives, it does not understand the relations between the objects and the action. Another line of works propose *dual encoders* which learn separate encodings of images and language. A prominent example is CLIP (Radford et al., 2021), which uses a contrastive objective to train a text encoder (GPT-2, Radford et al., 2019) and an image encoder (e.g., ViT Dosovitskiy et al., 2020). CLIP achieves state-of-the-art performance across multiple tasks. Different from LXMERT, CLIP is trained to match an image as a whole to a text description. We use this advantage and extract image-grounded sentence embeddings using CLIP's text encoder. Since CLIP applies a subtoken-level tokenization, the outputs of its text encoder are embeddings for the subtokens in the input sentence. Although it is a common practice to use the embedding of the classification token as the overall sentence embedding, this approach has been shown to be suboptimal (Vulić et al., 2020). We conduct experiments to find the optimal sentence representation.

3 VGSI: Visual Goal–Step Inference Task

Task Definition. Yang et al. (2021) define VGSI as a 4-way multiple choice problem. As shown in the example in Figure 1, given a textual *goal* G and four images I_i, $i \in \{1, 2, 3, 4\}$ representing four candidate *steps*, the task is to select the image that represents a correct step towards accomplishing G.

In this paper, we additionally explore a stricter definition of VGSI, where the task is to select the respective correct image of *all* steps that are necessary to reach the goal G.

3.1 Methods

3.1.1 Event Representations

To obtain event representations from goal and step sentences, we first extract the *subject*, *verbal predicate*, *direct object* and *prepositional phrase* from the sentences using a dependency parser (Dozat and Manning, 2016)[2].

Implicit Subject Representation. Due to the nature of the dataset of procedural instructions, textual goals and steps are usually imperative sentences, and as a consequence, the subject is left off. To encode the subject, we conduct experiments to compare event representations with no explicitly mentioned subject to those which express the subject (1) by the token *person*, or (2) by the special *<|startoftext|>* token of the CLIP tokenizer. Since the *<|startoftext|>* token added by

[2] We use SuPar available at `https://github.com/yzhangcs/parser`

us is always between the *<|startoftext|>* token of the CLIP tokenizer and the verbal predicate, its embedding is supposed to capture syntactic information from these two surrounding tokens via the attention mechanism (i.e. the information about the position of the subject of a sentence). To verify this hypothesis, we conduct two groups of probing experiments using the most common and the least common token in the input text as the pseudo-subject, respectively (see Section 5.1). We find that sentences with the *<|startoftext|>* token as the pseudo-subject lead to the best result.

Event Representations. The event representation is an essential component of our task. As introduced in Section 2, we represent events in the goal and step sentences using two types of representations: (1) the relational event representation (Balasubramanian et al., 2013) which is a *(subject, verb, object)* tuple, and (2) the multi-argument event representation (Pichotta and Mooney, 2014) which is a *(subject, verb, object, prepositional phrase)* tuple. Table 1 shows examples of all representations we explore. In the case that the object or prepositional phrase is absent, we represent it by a *[PAD]* token, e.g., (<|*startoftext*|>, pour, sauce, *[PAD]*).

Local vs. Contextualised Event. To assess the effectiveness of event representations, we deliberately use non-contextualised embeddings to disentangle the *subj–pred–obj(–pp)* information from the overall sentence. In detail, the components of the event representations are concatenated to form a sentence, which is then encoded by the CLIP text encoder (i.e. GPT-2). For instance, the event (<|*startoftext*|>, pour, sauce) is turned into the input <|*startoftext*|> *pour sauce*. We compare this encoding method to one that uses contextualised embeddings: We first encode the whole sentence and extract all word embeddings. If the tokenizer split a word into subtokens, we mean-pool their corresponding embeddings. Then, we mean-pool the word embeddings which are part of the components of the event representations. For example, the word embeddings in the object phrase *into container or jug* are averaged to a single vector. Note that for both local and contextualised approaches, the CLIP tokenizer automatically adds a *<|startoftext|>* and an *<|endoftext|>* token to the start and the end of the input, respectively. We remove these two special tokens after the encoding, such that only the embedding of the *<|startoftext|>* as the implicit subject is averaged with other words. We evaluate the text embeddings obtained from three groups of layers of CLIP.[3] The visual embeddings, in turn, are the last hidden state of the CLIP image encoder (i.e. ViT).[4]

text	Pour the soy or tamari sauce into a suitable small mixing container or jug.
event$_{rel}$	(*<|startoftext|>*, pour, sauce)
event$_{mult}$	(*<|startoftext|>*, pour, sauce, into container or jug)

Table 1: Example of the relational and multi-argument event representation.

3.1.2 Triplet Network for Goal–Step Inference

We use Triplet Network (Hoffer and Ailon, 2015) in all our experiments and use the cosine similarity as the similarity metric.

Training. The triplet network for training is implemented as a three-branch network with a text module and an image module, where the two branches of the image module share the same parameters. The input is a triplet *(G+S*, I_{pos}, I_{neg}*)*, where *G+S* is the embedding of the concatenated goal–step sentence, I_{pos} is the embedding of the positive image, I_{neg} is the embedding of a negative image (see Section 4.3). The model learns a cross-modal embedding space by minimizing the distance between *G+S* and I_{pos}, while maximizing the distance between *G+S* and I_{neg}. Different from Yang et al. (2021) which use *G* as the textual input for training, we use *G+S* because *S* share common information with *I* and serves as a bridge between *G* and *I*. Thus, *G+S* could help the model to better understand the relation between *G* and *I*.

Inference. During inference, we follow the input format of Yang et al. (2021), i.e. the textual input is the goal alone. The model takes each pair (G, I_i), $i \in \{1, 2, 3, 4\}$ from a test data point $(G, [I_1, I_2, I_3, I_4])$ as input. By computing the similarity between G and I_i, the model predicts the correct step image $\hat{I}$ as that with the highest simi-

[3] Based on (Vulić et al., 2020)'s findings, we do not use the embedding of the classification token, cf. Sect. 2.

[4] We use `clip-vit-large-patch14` from HuggingFace available at `https://huggingface.co/openai/clip-vit-large-patch14`

Experiment group	Embed size	#params	Input format	Event injection
SENTENCE	768 (text) 1024 (image)	3,936,256	goal+step (train) goal (test)	s
EVENT	768 (text) 1024 (image)	3,936,256	goal+step (train) goal (test)	e
SENTENCE+EVENT	1536 (text) 1024 (image)	4,722,688	goal+step (train) goal (test)	s+e

Table 2: Embedding size, number of parameters, input formats to the text encoder and event injection approaches of different experiment groups: concatenation of goal and step headline (goal+step), goal only (goal); sentence only (s), event only (e), sentence+event (s+e).

larity as follows:

$$\hat{I} = \arg\max_{I_i} cos(G, I_i) \tag{1}$$

4 Experiments

4.1 Data

We conduct our experiments on **wikiHow-VGSI** (Yang et al., 2021),[5] a dataset for multimodal goal-oriented PKU collected from the English wikiHow[6]. The dataset contains articles of instructions to complete tasks across a wide range of daily-life topics, including health, home and garden, education, recipes etc. Each article contains a *goal G* in the form of a "How to"-sentence and a set of *methods* (e.g., "How to bake mini cupcakes", Figure 1). Each method comprises a list of *steps*. Each step has a *step headline S* which is an imperative sentence describing that step, and an image *I* corresponding to that step (e.g., I_1 and S in Fig. 1). To describe a goal and its steps, we use the *goal G* and the *step headline S* and its associated image *I*, respectively.

We lowercase all the texts in the dataset, and use the special token *<|startoftext|>* to represent the subject in all sentences (i.e., *pseudo-subject*). Specifically, *<|startoftext|>* substitutes *How to* in all goals and is prepended to all step headlines. Since we found some issues in the dataset, such as duplicates or non-English text, we removed 3 goals and 56 step headlines. Details to our filtering procedure are given in Appendix 9.1. As a result, the dataset used for our experiments contains $53,186$ goals, $772,221$ step headlines and $772,277$ step images.

4.2 Models

We assess the benefit of the two approaches for the event knowledge injection (relational and multi-argument representations, see Sect. 3.1.1) when being used as the only representation of the goal *G* and step *S* during training (EVENT), or when being used as additional information to the full sentences (SENTENCE+EVENT). We compare them against only using the full sentence (SENTENCE), which is also employed by Yang et al. (2021). Table 2 gives an overview of the different inputs and the corresponding hyperparameters of the models.

Jawahar et al. (2019) observed that the embeddings obtained from different layers of BERT tend to be dominated by different levels of linguistic information: surface (i.e. lexical) information in bottom layers, syntactic information in middle layers and semantic information in top layers. Thus, we examine sentence embeddings of three linguistic levels in each of these experiment groups: (1) FIRST4 averages the outputs of the first 4 layers of CLIP's text encoder; (2) MIDDLE4 averages the outputs of the 5-th to the 8-th layers of the encoder; (3) LAST4 averages the outputs of the last 4 layers.

4.2.1 EVENT

In this group of experiments, the goal and step sentences are replaced by the event representations extracted from them. For example, the sentence in Table 1 is replaced by <|*startoftext*|> *pour sauce* for the relational event representation and by <|*startoftext*|> *pour sauce into container or jug* for the multi-argument event representation.

4.2.2 SENTENCE+EVENT

In this group of experiments, the event representations are appended to the goal and step sentences. For example, the aforementioned sentence is converted to <|*startoftext*|> *pour the soy or tamari*

[5] https://github.com/YueYANG1996/wikiHow-VGSI

[6] https://www.wikihow.com/Main-Page

sauce into a suitable small mixing container or jug. <|*startoftext*|> *pour sauce.* for the relational event representation, and <|*startoftext*|> *pour the soy or tamari sauce into a suitable small mixing container or jug.* <|*startoftext*|> *pour sauce into container or jug.* for the multi-argument event representation.

4.2.3 SENTENCE

While event representations have been found valuable in earlier, linguistically motivated research on procedural texts (see Section 2), it stands the question whether they fully provide the crucial information for learning procedural knowledge. Hence, we also compare against a model that takes the encoded full sentence describing the goal or the goal+step as textual input, i.e. the model learns the task-relevant features from the full goal sentence or the step headline.

4.3 Training Procedure

We apply the random sampling strategy of Yang et al. (2021) to select negative step images. For each data point, we randomly select three different articles and take a random image from each article as the negative step image. We leave the experiments with other sampling methods used in Yang et al. (2021) to future work.

We initialize the weights using He-uniform with ReLU non-linearity. All models are trained for 200 epochs with batch size 1024 and a learning rate of 1e-5 with early stopping. In each experiment group, the model is trained and evaluated five times. We implemented the models in Keras with Tensorflow 2.0 and trained them on a single RTX A6000.

4.4 Evaluation Measures

We evaluate our models with two settings. The first one, which we call **weak**, follows the original task definition by Yang et al. (2021), where a data point in the test set is considered correctly predicted, if one step towards the goal given by that data point is correctly selected. To better fit the concept of procedural knowledge, we also apply a **strict** setting, in which a data point is correctly predicted, if all the steps required to achieve the goal given by the data point are correctly selected. We report the mean accuracy obtained by the five individual training and testing runs, as well as the corresponding standard deviation.

5 Results

Tables 3 and 5 give the most important results. The full results can be found in Appendix 9.2.

5.1 Event-based Representations

Table 3 shows the performance of the models with the *<|startoftext|>* token as pseudo-subject, using different event representations containing different levels of linguistic knowledge. The last two rows list the results of the best model and the human evaluation in Yang et al. (2021).

As expected, by comparing the $\text{EVENT}_{rel,*}$ and $\text{EVENT}_{mult,*}$ groups (i.e., <[2],[3]>, <[7],[8]>, <[12],[13]>), we observe that the multi-argument event representation outperforms the relational event representation.

Linguistic Level Embedding. To find out which level of linguistic knowledge is most suitable for the task, we compare the following three groups of results in Table 3: <[2],[7],[12]>, <[3],[8],[13]> and <[5],[10],[15]>. On average, the LAST4 groups achieve the highest accuracy, while the FIRST4 groups perform the worst. The performance gap between FIRST4 and the other two groups is considerably larger than that between MIDDLE4 and LAST4. This indicates that both semantic and syntactic information play important roles in the task, while lexical information is far less important than syntactic and semantic information.

Event Knowledge Injection. The results of <[3],[5]>, <[8],[10]>, and <[13],[15]> in Table 3 show that SENTENCE+EVENT results in higher accuracy than EVENT. This reveals the advantage of attaching event knowledge to the sentence over using only the event knowledge. It also implies that the sentence could provide additional information to the event, which could help models better understand procedural knowledge.

Local vs. Contextualised Embeddings. By comparing the results of local and contextualised event embeddings in Table 3, we observe a significant improvement of the performance in the latter group. On average, the accuracy with contextualised embeddings is 3.71% and 13.73% higher than that with the local ones in the **weak** setting and in the **strict** setting, respectively. This verifies the observation in the last paragraph that sentences provide additional, useful information.

Models	Local Event		Contextualised Event	
	weak	**strict**	**weak**	**strict**
[2] Event$_{rel,first4}$	68.9±0.3	9.9±0.3	71.6±0.4	12.2±0.3
[3] Event$_{mult,first4}$	75.8±0.4	15.3±0.5	77.0±0.1	15.9±0.2
[5] Sentence+Event$_{mult,first4}$	80.9±0.8	19.3±1.3	81.0±0.1	19.6±0.3
[7] Event$_{rel,middle4}$	70.3±0.2	11.1±0.2	74.9±0	14.9±0.1
[8] Event$_{mult,middle4}$	76.9±0.6	16.9±0.9	79.9±0	19.1±0.4
[10] Sentence+Event$_{mult,middle4}$	82.4±0.1	22.1±0.3	82.8±0.9	22.4±1.5
[12] Event$_{rel,last4}$	69.1±0.3	11.5±0.4	75.9±0	16.7±0.1
[13] Event$_{mult,last4}$	77.3±0.4	18.8±0.4	80.8±0	21.2±0.2
[15] Sentence+Event$_{mult,last4}$	81.1±0.7	21.5±0.8	**84.7±0**	**26.4±0.2**
[16] Event$_{mult,last4,+1layer}$	76.6±0.3	17.9±0.2	80.5±0	20.7±0
Triplet Net (BERT) (Yang et al., 2021)†	72.8	-	72.8	-
Human (Yang et al., 2021)	84.5	-	84.5	-

Table 3: Accuracy (%) of experiments using different event representations encoded by different layers of the CLIP text encoder. The implicit subject is represented by *<|startoftext|>* (**sot+sent**). †Results adopted from the authors, they are not directly comparable.

Implicit(/Pseudo-)Subject	**weak**	**strict**
sot+sent	**82.7**	**22.3**
person+sent	80.3	19.9
-+sent	79.4	19.4
sot	24.2	0.11
most-frequent+sent	79.8	20.3
least-frequent+sent	68.6	10.4

Table 4: Accuracy (%) of Sentence experiments using different implicit (top) / pseudo (bottom) subjects: *<|startoftext|>*+sentence (sot+sent), *person*+sentence (person+sent), sentence without subject (-+sent), *<|startoftext|>* only (sot).

Implicit Subject Abstract Representation. The sentences in the dataset either begin with *How to*, or they do not have an explicit subject. Thus, we assess the contribution of different abstract representations for the implicit subject of the sentences. Table 4 (top) shows the performance of the Sentence$_{middle4}$ models with four abstract representations as the subject. The results show that *<|startoftext|>* is the most powerful abstract representation for the subject. However, we observe a significant performance degradation when using this token separately as the representation of the whole sentence (i.e. *sot* in Table 4). In this case, the embedding of *<|startoftext|>* is derived from the last hidden state of CLIP's text encoder. A possible reason could be that the *<|startoftext|>* token is always located between the verbal predicate and the *<|startoftext|>* token added by CLIP's tokenizer which indicates the start of the sentence. Hence, its embedding may capture syntactic information about the subject's position in the sentence from these contextual tokens via the attention mechanism. To verify this hypothesis, we conduct two groups of probing experiments for the syntactic information in the *<|startoftext|>* token. We evaluate the Sentence$_{middle4}$ model by taking the most and the least frequent token in the dataset ("." and "50.0", respectively) as a pseudo-subject of the input text, as we assume them to be generally less informative for the sentences. We observe a considerable performance drop with the least frequent token (see Table 4, bottom), indicating that *<|startoftext|>* indeed gives the model valuable cues about the subject position in a sentence.

5.2 Event-Enhanced Sentences

Table 5 compares the performance of using sentence-only embeddings with using event-enhanced sentence embeddings. As a result, Sentence+Event outperforms Sentence with contextualised event embeddings when using the average of the last 4 hidden layers of the CLIP text encoder. The groups using the first 4 and middle 4 layers achieve comparable performance. Moreover, the best model (i.e., [15]) reaches the human upper

Models	Local Event		Contextualised Event	
	weak	**strict**	**weak**	**strict**
[1] SENTENCE$_{first4}$	81.6±0.1	20.1±0.1	81.2±0.0	19.7±0.2
[5] SENTENCE+EVENT$_{mult,first4}$	80.9±0.8	19.3±1.3	81.0±0.1	19.6±0.3
[6] SENTENCE$_{middle4}$	82.7±0.4	22.3±0.5	82.7±1.1	22.2±1.7
[10] SENTENCE+EVENT$_{mult,middle4}$	82.4±0.1	22.1±0.3	82.8±0.9	22.4±1.5
[11] SENTENCE$_{last4}$	82.1±0.4	22.3±0.7	84.6±0.1	26.0±0.2
[15] SENTENCE+EVENT$_{mult,last4}$	81.1±0.7	21.5±0.8	**84.7±0.0**	**26.4±0.2**
Triplet Net (BERT) (Yang et al., 2021)†	72.8	-	72.8	-
Human (Yang et al., 2021)	84.5	-	84.5	-

Table 5: Accuracy (%) and standard deviation of the experiments using different event representations encoded by different layers of the CLIP text encoder.

bound, demonstrating the necessity of applying the strict evaluation setting.

5.3 Disentangle the Influence of Model Sizes and Embeddings

Since the models in the SENTENCE+EVENT group have more trainable parameters due to the concatenation of sentence- and event embeddings, the performance gain could attribute either to the number of parameters or to the embeddings. To disentangle the influence of these two factors, we conduct an experiment based on EVENT$_{mult,last4}$, with the text module of the triplet network being extended by an additional dense layer. This increases the number of trainable parameters of the model to $4,750,973$, which is comparable with the most effective SENTENCE+EVENT$_{mult,last4}$ models. The results of [16] in Table 3 show that there is no considerable change in performance from [13] and [15], indicating that the performance gain is due to attaching the event representation to the sentence.

6 Qualitative Analysis

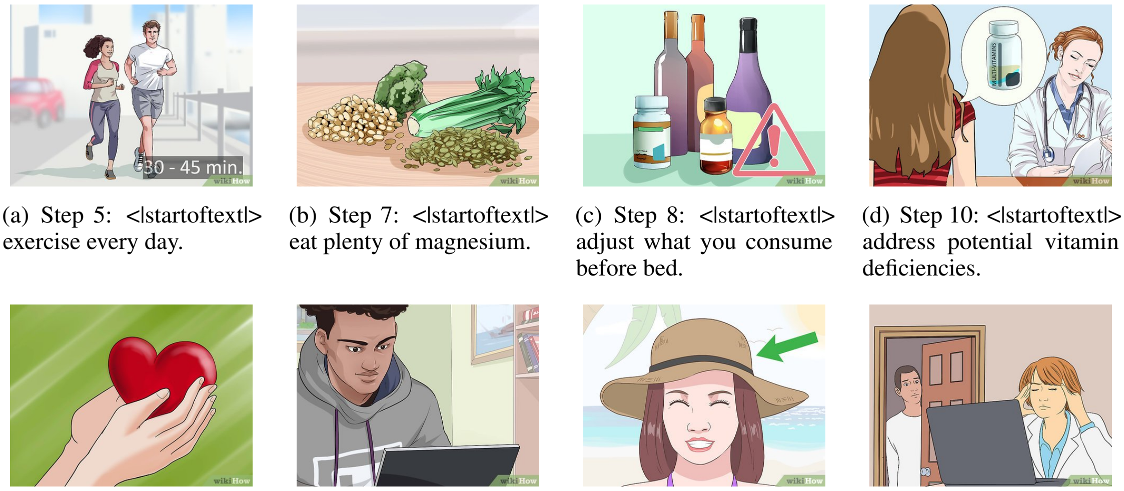
We provide a qualitative analysis on the semantic gap between the ground-truth and the predicted images. Figure 2 shows part of an example of the model's predictions for the goal *How to stop twitching in your sleep?* In this example, four out of ten steps are incorrectly predicted.

For Step 5, the textual input for training is <|*startoftext*|> *stop twitching in your sleep.* <|*startoftext*|> *exercise every day.* The model selects Image (e) which depicts a hand holding a heart. The model may associate "twitching" with the heart in the image, but fails to infer the relation between "twitching" and the jogging people in the correct image (a). Thus, the model may not learn causal relationships between the goal and the step image, such as "Jogging can improve people's health condition and thus stop twitching in the sleep".

For Step 7 with the textual input <|*startoftext*|> *stop twitching in your sleep.* <|*startoftext*|> *eat plenty of magnesium.*, the model selects Image (f) illustrating a person sitting at a laptop. Possible reasons could be: (1) The action "eat" is usually performed by humans, but the correct image only describes some food, which the model misses to associate with "eat"; and (2) The phrase "plenty of magnesium" may mislead the model to select the wrong image with a laptop, which is associated more with magnesium than vegetables. Hence, the model may only learn knowledge about simple, superficial properties of the objects in images, and may lack more complex commonsense knowledge about the relations between objects, such as "Laptop is not edible" or "Human cannot take magnesium by eating laptops".

For Step 8, the input is <|*startoftext*|> *stop twitching in your sleep.* <|*startoftext*|> *adjust what you consume before bed.* The model selects the image showing a lady with a hat being pointed to by an arrow. This again indicated that the model's decision heavily relies on the verb. Furthermore, it also suggests that the model has limited capability of identifying the affordances of the objects in the image and associating them with the goal.

For Step 10 with the input <|*startoftext*|> *stop twitching in your sleep.* <|*startoftext*|> *address potential vitamin deficiencies.*, the model again



(a) Step 5: <|startoftext|> exercise every day.

(b) Step 7: <|startoftext|> eat plenty of magnesium.

(c) Step 8: <|startoftext|> adjust what you consume before bed.

(d) Step 10: <|startoftext|> address potential vitamin deficiencies.

(e) Step 5: <|startoftext|> be gentle.

(f) Step 7: <|startoftext|> search online for job postings.

(g) Step 8: <|startoftext|> put on a sun hat to protect your hair and keep you cool.

(h) Step 10: <|startoftext|> start to learn about and change any patterns in your daily life that may act as triggers or contribute to your loved one's destructive behavior.

Figure 2: Ground-truth (top) and model's false predictions (bottom) for Steps 5, 7, 8, 10. Goal: *How to stop twitching in your sleep?*

seems to not capture causal relationships such as "Vitamin deficiency can lead to twitching in sleep", but to base its inference on shallow object features such as "A man opens the door and wakes the sleeping woman up".

In conclusion, our observations indicate that the model's decision highly depends on shallow features in the image and their alignment to the verbs and nouns in the text, while its effectiveness is impaired by its limited understanding of deeper semantics and causal relationships between the goal and the step images.

7 Conclusions

In this paper, we investigate two linguistically-inspired event knowledge injection approaches for the Visual Goal–Step Inference (VGSI) task. We experimentally compare three levels of linguistic information in the text embedding produced by state-of-the-art neural deep learning models. Furthermore, we also compare event embeddings which encode only the information of the event components themselves with contextualised event embeddings which include information about the overall sentence syntactically not belonging to the arguments forming an event representation itself. Last but not least, we assess different representations for the implicit subject of instructional sentences. We find that the early, linguistically inspired methods for representing event knowledge do contribute to understand procedures in combination with modern V&L models.

8 Limitations

We explore early, very simple structured event representations. Recent works in visual–linguistic semantic representations which use richer representations comprising predicate–argument structures and event types and argument roles, the general graph-based approaches, as well as scene graphs, are left for future work. Furthermore, the wikiHow articles may reflect the bias of their human authors.

Acknowledgements

We would like to thank Professor Parisa Kordjamshidi for her valuable feedback in the Pre-Submission Mentorship Program. We are also grateful to the anonymous reviewers for their detailed comments on our work. We would further like to thank Yue Yang for meaningful discussions about the VGSI task.

References

Niranjan Balasubramanian, Stephen Soderland, Oren Etzioni, et al. 2013. Generating coherent event schemas at scale. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1721–1731.

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

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. *arXiv preprint arXiv:2109.12776*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

Elena Filatova and Eduard Hovy. 2001. Assigning timestamps to event-clauses. In *Proceedings of the ACL 2001 Workshop on Temporal and Spatial Information Processing*.

Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*, pages 84–92. Springer.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. 2021. Knowledge graphs. *ACM Computing Surveys (CSUR)*, 54(4):1–37.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Graham Katz and Fabrizio Arosio. 2001. The annotation of temporal information in natural language sentences. In *Proceedings of the ACL 2001 Workshop on Temporal and Spatial Information Processing*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. 2022. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16420–16429.

Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, et al. 2020. Gaia: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 77–86.

Qian Li, Jianxin Li, Jiawei Sheng, Shiyao Cui, Jia Wu, Yiming Hei, Hao Peng, Shu Guo, Lihong Wang, Amin Beheshti, et al. 2021. A compact survey on event extraction: Approaches and applications. *arXiv preprint arXiv:2107.02126*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Qing Lyu, Li Zhang, and Chris Callison-Burch. 2021. Goal-oriented script construction. *arXiv preprint arXiv:2107.13189*.

Dena Mujtaba and Nihar Mahapatra. 2019. Recent trends in natural language understanding for procedural knowledge. In *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 420–424. IEEE.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Karl Pichotta and Raymond Mooney. 2014. Statistical script learning with multi-argument events. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 220–229.

Karl Pichotta and Raymond Mooney. 2016. Learning statistical scripts with lstm recurrent neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

James Pustejovsky, Robert Ingria, Roser Sauri, José M Castaño, Jessica Littman, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Inderjeet Mani. 2005. The specification language timeml.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Frank Schilder and Christopher Habel. 2001. From temporal expressions to temporal information: Semantic tagging of news messages. In *Proceedings of the ACL 2001 workshop on temporal and spatial information processing*.

Roger Shank and Robert Abelson. 1977. Scripts, plans, goals and understanding.

Chenkai Sun, Tie Xu, ChengXiang Zhai, and Heng ji. 2022. Incorporating task-specific concept knowledge into script learning.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Niket Tandon, Keisuke Sakaguchi, Bhavana Dalvi Mishra, Dheeraj Rajagopal, Peter Clark, Michal Guerquin, Kyle Richardson, and Eduard Hovy. 2020. A dataset for tracking entities in open domain procedural text. *arXiv preprint arXiv:2011.08092*.

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.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. *arXiv preprint arXiv:2010.05731*.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. 2021. Visual goal-step inference using wikihow. *arXiv preprint arXiv:2104.05845*.

Zi Yang and Eric Nyberg. 2015. Leveraging procedural knowledge for task-oriented search. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 513–522.

Pengfei Yu, Zixuan Zhang, Clare Voss, Jonathan May, and Heng Ji. 2022. Building an event extractor with only a few examples. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 102–109.

Li Zhang. 2022. Reasoning about procedures with natural language processing: A tutorial. *arXiv preprint arXiv:2205.07455*.

Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020a. Intent detection with wikihow. *arXiv preprint arXiv:2009.05781*.

Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020b. Reasoning about goals, steps, and temporal ordering with wikihow. *arXiv preprint arXiv:2009.07690*.

Shuyan Zhou, Li Zhang, Yue Yang, Qing Lyu, Pengcheng Yin, Chris Callison-Burch, and Graham Neubig. 2022. Show me more details: Discovering hierarchies of procedures from semi-structured web data. *arXiv preprint arXiv:2203.07264*.

Yilun Zhou, Julie A Shah, and Steven Schockaert. 2019. Learning household task knowledge from wikihow descriptions. *arXiv preprint arXiv:1909.06414*.

9 Appendix

9.1 Data Preprocessing and Cleaning

1. We remove the goals with file-ID *385799* and *5323060*, as they contain non-English words.

2. Two data points share the same file-ID *411540*, each refers to the goal *How to keep healthy family relationships* and *How to keep relationships healthy within your family*. The first data point is automatically removed when building a mapping from file-IDs to goals.

3. We remove the step headlines with step-IDs *1926747_3_0*, *2191502_0_0* and *985548_2_0*, since they contain only a dot (.) and cannot be parsed by the dependency parser.

9.2 Full Table of the Results

As a supplement to Table 3 and Table 5, Table 6 shows the results of all experiment groups.

Experiments	Local Event		Contextualised Event	
	weak	**strict**	**weak**	**strict**
[1] SENTENCE$_{first4}$	81.6±0.1	20.1±0.1	81.2±0	19.7±0.2
[2] EVENT$_{rel,first4}$	68.9±0.3	9.9±0.3	71.6±0.4	12.2±0.3
[3] EVENT$_{mult,first4}$	75.8±0.4	15.3±0.5	77.0±0.1	15.9±0.2
[4] SENTENCE+EVENT$_{rel,first4}$	79.9±0.3	17.9±0.7	80.4±0.1	18.6±0.1
[5] SENTENCE+EVENT$_{mult,first4}$	80.9±0.8	19.3±1.3	81.0±0.1	19.6±0.3
[6] SENTENCE$_{middle4}$	82.7±0.4	22.3±0.5	82.7±1.1	22.2±1.7
[7] EVENT$_{rel,middle4}$	70.3±0.2	11.1±0.2	74.9±0	14.9±0.1
[8] EVENT$_{mult,middle4}$	76.9±0.6	16.9±0.9	79.9±0	19.1±0.4
[9] SENTENCE+EVENT$_{rel,middle4}$	81.8±0.3	21.2±0.3	81.8±1.1	20.4±1.8
[10] SENTENCE+EVENT$_{mult,middle4}$	82.4±0.1	22.1±0.3	82.8±0.9	22.4±1.5
[11] SENTENCE$_{last4}$	82.1±0.4	22.3±0.7	84.6±0.1	26.0±0.2
[12] EVENT$_{rel,last4}$	69.1±0.3	11.5±0.4	75.9±0	16.7±0.1
[13] EVENT$_{mult,last4}$	77.3±0.4	18.8±0.4	80.8±0	21.2±0.2
[14] SENTENCE+EVENT$_{rel,last4}$	80.3±0.6	20.2±1.0	84.1±0.4	25.2±1.1
[15] SENTENCE+EVENT$_{mult,last4}$	81.1±0.7	21.5±0.8	**84.7±0**	**26.4±0.2**
[16] EVENT$_{mult,last4,+1layer}$	76.6±0.3	17.9±0.2	80.5±0	20.7±0
Triplet Net (BERT) (Yang et al., 2021)†	72.8	-	72.8	-
Human (Yang et al., 2021)	84.5	-	84.5	-

Table 6: Accuracy (%) of experiments using different event representations encoded by different layers of the CLIP text encoder (full table).