Situation Recognition for Holistic Video Understanding

Thesis submitted in partial fulfillment
of the requirements for the degree of

Masters of Science
in
Computer Science and Engineering by Research

by

Zeeshan Khan
2021701029
zeeshan.khan@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
May, 2023

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled “Situation Recognition for Holistic Video Understanding” by Zeeshan Khan, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Prof. C V Jawahar

Date

Adviser: Prof. Makarand Tapaswi

To, my family and friends

Acknowledgments

I would like to express my deepest gratitude to my advisors, Prof. C V Jawahar, Prof. Makarand Tapaswi, for their constant guidance and support throughout my Master's journey. I am very grateful for their time and effort. Prof. Jawahar always emphasized on the best practices to be followed that not only helped me grow and mature as a researcher but as a human too. His visionary and discerning approach towards problem formulation and solution taught me the ability to extract the full potential out of any problem and contribute with the highest standards. Prof. Makarand with his brilliant ideas, invaluable insights and all round perfection, showed me what it takes to be a high performing researcher. His constant support and a cool way of advising at every level of my masters journey, not only kept me motivated throughout but also pushed my limits without pain. The teachings I learnt from my advisors is the most invaluable gift I will always cherish.

I would like to thank my CVIT colleagues- Ravi, Madhav, Darshan, Pranav, Rudrabha, Aditya, Bipasha, Soumya, Seshadri, and Shubham who made this journey enjoyable and fulfilling. I would also like to thank my juniors- Shivanshu, Varun and Subrata, for giving me the honour to boss around ;) I am also indebted to the staff specially Rohitha for helping with all the administrative tasks.

I would like to thank my friends, both within and outside of the academic community, for their encouragement, motivation, and support.

Finally, I would like to express my gratitude to my parents and sister for their selfless support and affection.

Abstract

Video is a complex modality consisting of multiple events, complex action, humans, objects and their interactions densely entangled over time. Understanding videos has been the core and one of the most challenging problem in computer vision and machine learning. What makes it even harder is the lack of structured formulation of the task specially when long videos are considered consisting of multiple events and diverse scenes. Prior works in video understanding have tried to address the problem only in a sparse and a uni-dimensional way, for example action recognition, spatio-temporal grounding, question answering and, free form captioning. However it requires holistic understanding to fully capture all the events, actions, and relations between all the entities, and represent any natural scene with the highest detail in the most faithful way. It requires answering several questions such as *who is doing what to whom, with what, how, why, and where*.

Recently, Video Situation Recognition (VidSitu) through semantic role labeling is framed as a task for structured prediction of multiple events, their relationships, and actions and various verb-role pairs attached to descriptive entities. This is one of the most dense video understanding task posing several challenges in identifying, disambiguating, and co-referencing entities across multiple verb-role pairs, but also faces some challenges of evaluation due to the free form captions for representing the roles. In this work, we propose the addition of spatio-temporal grounding as an essential component of the structured prediction task in a weakly supervised setting, without requiring ground truth bounding boxes. Since evaluating free-form captions can be difficult and imprecise this not only improves the current formulation and the evaluation setup, but also improves the interpretability of the models decision, because grounding allows us to visualise where the model is looking while generating a caption.

To this end we present a novel three stage Transformer model, *VideoWhisperer*, that is empowered to make joint predictions. In stage one, we learn contextualised embeddings for video features in parallel with key objects that appear in the video clips to enable fine-grained spatio-temporal reasoning. The second stage sees verb-role queries attend and pool information from object embeddings, localising *answers* to questions posed about the action. The final stage generates these answers as captions to describe each verb-role pair present in the video. Our model operates on a group of events (clips) simultaneously and predicts verbs, verb-role pairs, their nouns, and their grounding on-the-fly. When evaluated on a grounding-augmented version of the VidSitu dataset, we observe a large improvement in entity captioning accuracy, as well as the ability to localize verb-roles without grounding annotations at training time.

Contents

Chapter Page

List of Figures

List of Tables

Chapter 1

Introduction

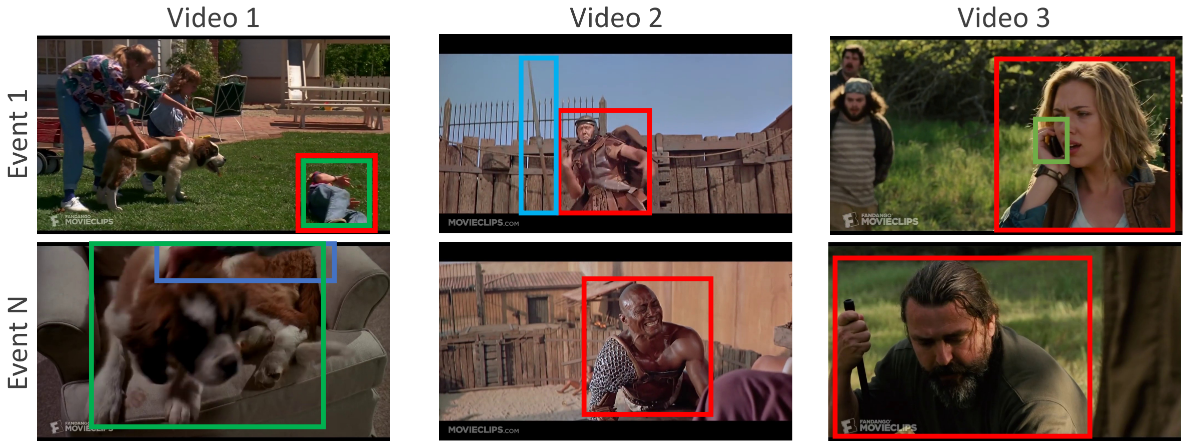
Note: this work [12] is published at NeurIPS 2022 and the text in the paper is repeated.

At the end of *The Dark Knight*, we see a short intense sequence that involves Harvey Dent toss a coin while holding a gun followed by sudden action. Holistic understanding of such a video sequence, especially one that involves multiple people, requires predicting more than the action label (*what* verb). For example, we may wish to answer questions such as *who* performed the action (agent), *why* they are doing it (purpose / goal), *how* are they doing it (manner), *where* are they doing it (location), and even *what happens after* (multi-event understanding).

While humans are able to perceive the situation and are good at answering such questions, many works often focus on building tools for doing single tasks, *e.g* predicting actions [8] or detecting objects [2, 4] or image/video captioning [21, 32]. We are interested in assessing how some of these advances can be combined for a holistic understanding of video clips. In particular, we foresee that successful systems could have wide-ranging applications from teaching embodied agents to understand and interact with the world [46], to video retrieval [22] and question-answering [43], and even fine-grained movie understanding [11, 38].

A recent and audacious step towards this goal is the work by Sadhu *et al* [31]. They propose Video Situation Recognition (VidSitu), a structured prediction task over five short clips consisting of three sub-problems: (i) recognizing the salient actions in the short clips; (ii) predicting roles and their entities that are part of this action; and (iii) modelling simple event relations such as enable or cause. Similar to the predecessor image situation recognition (imSitu [45]), VidSitu is annotated using Semantic Role Labelling (SRL) [25]. A video (say 10s) is divided into multiple small events ($\sim$2s) and each event is associated with a salient action verb (*e.g hit*). Each verb has a fixed set of roles or arguments, *e.g agent-Arg0*, *patient-Arg1*, *tool-Arg2*, *location-ArgM(Location)*, *manner-ArgM(manner)*, *etc*, and each role is annotated with a free form text caption, *e.g agent: Blonde Woman*, as illustrated in Fig. 1.1.

Grounded VidSitu. VidSitu poses various challenges: long-tailed distribution of both verbs and text phrases, disambiguating the roles, overcoming semantic role-noun pair sparsity, and co-referencing of entities in the entire video. Moreover, there is ambiguity in text phrases that refer to the same unique entity (*e.g* “man in white shirt” or “man with brown hair” A model may fail to understand which attributes



Event 1, Verb: ROLL	
Arg0 (Roller)	Boy in striped shirt
Arg1 (Thing rolled)	Himself
ArgM (Direction)	Back and forth
Arg Scene	Backyard

Event N, Verb: RUB	
Arg0 (Rubber)	Person in blue shirt
Arg1 (Thing rubbed)	Dog
Arg2 (Surface)	Hand
Arg Scene	Backyard

Event 1, Verb: HIT	
Arg0 (Hitter)	Man in armor
Arg1 (Thing hit)	Bald man
Arg2 (Instrument)	Spear
Arg Scene	Arena

Event N, Verb: WINCE	
Arg0 (Wincer)	Bald man
Arg Scene	Arena

Event 1, Verb: LIFT	
Arg0 (Elevator)	Blonde woman
Arg1 (Thing lift)	Her phone
ArgM (Direction)	Up
ArgM (Manner)	Quickly
Arg Scene	An open field

Event N, Verb: TALK	
Arg0 (Talker)	The kneeling man
Arg1 (Hearer)	Blonde woman
Arg2 (Manner)	Confidently
Arg Scene	An open field

Figure 1.1 Overview of GVSR: Given a video consisting of multiple events, GVSR requires recognising the action verbs, their corresponding roles, and localising them in the spatio-temporal domain. This is a challenging task as it requires to disambiguate between several roles that the same entity may take in different events, *e.g* in Video 2 the *bald man* is a *patient* in event 1, but an *agent* in event N. Moreover, the entities present in multiple events are co-referenced in all such events. Colored arguments are grounded in the image with bounding boxes (figure best seen in colour).

are important and may bias towards a specific caption (or pattern like shirt color), given the long-tailed distribution. This is exacerbated when multiple entities (*e.g agent* and *patient*) have similar attributes and the model predicts the same caption for them (see Fig. 1.1). To remove biases of the captioning module and gauge the model's ability to identify the role, we propose *Grounded Video Situation Recognition* (GVSR) - an extension of the VidSitu task to include spatio-temporal grounding. In addition to predicting the captions for the role-entity pairs, we now expect the structured output to contain spatio-temporal localization, currently posed as a weakly-supervised task, *i.e*we don't require any ground truth bounding box supervision during training. The supervision comes only from the ground truth semantic role captions.

Joint structured prediction. Previous works [31, 42] modeled the VidSitu tasks separately, *e.g* the ground-truth verb is fed to the SRL task. This setup does *not* allow for situation recognition on a new video clip without manual intervention. Instead, in this work, we focus on solving three tasks jointly: (i) verb classification; (ii) SRL; and (iii) Grounding for SRL. We ignore the original event relation prediction task in this work, as this can be performed later in a decoupled manner similar to [31].

We propose *VideoWhisperer*, a new three-stage transformer architecture that enables video understanding at a global level through self-attention across all video clips, and generates predictions for the above three tasks at an event level through localised event-role representations. In the first stage, we use a Transformer encoder to align and contextualise 2D object features in addition to event-level video features. These rich features are essential for grounded situation recognition, and are used to predict both the verb-role pairs and entities. In the second stage, a Transformer decoder models the role as a query, and applies cross-attention to find the best elements from the contextualised object features, also enabling visual grounding. Finally, in stage three, we generate the captions for each role entity. The three-stage network disentangles the three tasks and allows for end-to-end training.

To evaluate the proposed framework for multi-event multi-role grounding we annotate the validation set with ground truth bounding boxes, and propose a new IoU based metric.

1.1 Contributions

The contributions of the thesis are as follows:

- We present a new framework that combines grounding with semantic role labeling (SRL) for end-to-end Grounded Video Situation Recognition (GVSR).
- We formulate the new grounded SRL task, which includes localization of each role entity across the entire video spanning multiple events in a single shot.
- We release the grounding annotations and also include them in the evaluation benchmark for GVSR.
- We design a new three-stage Transformer based model- *VideoWhisperer* for joint verb prediction, semantic-role labelling through caption generation, and weakly-supervised grounding of visual entities.
- We propose role prediction and use role queries contextualised by video embeddings for SRL, circumventing the requirement of ground-truth verbs or roles, enabling end-to-end GVSR.
- We propose to combine object features with video features and highlight multiple advantages enabling weakly-supervised grounding and improving the quality of SRL captions leading to a 22 points jump in CIDEr score in comparison to a video-only baseline.

- Finally, we present extensive ablation experiments to analyze our model. Our model achieves the state-of-the-art results on the VidSitu benchmark.

1.2 Organization of Thesis

The rest of the thesis is organized as follows.

- In Chapter 2, we discuss the related works in image and video situation recognition literature. We also discuss the limitations and challenges in the existing frameworks.
- In Chapter 3, we formulate the GVSR task and describe the *VideoWhisperer* model in detail. Then we discuss the training and inference strategy.
- In Chapter 4, we discuss the dataset, metrics and, the evaluation setup. Then we explain the 2 frameworks emerging from the model 1) Framework 1: Using ground truth roles for grounded SRL. And 2) Framework 2: End-to-end prediction of Verbs, SRL and, grounding. We show all the experiments and ablation studies for both the frameworks in detail.
- Chapter 5 presents the concluding thoughts and future works.

Chapter 2

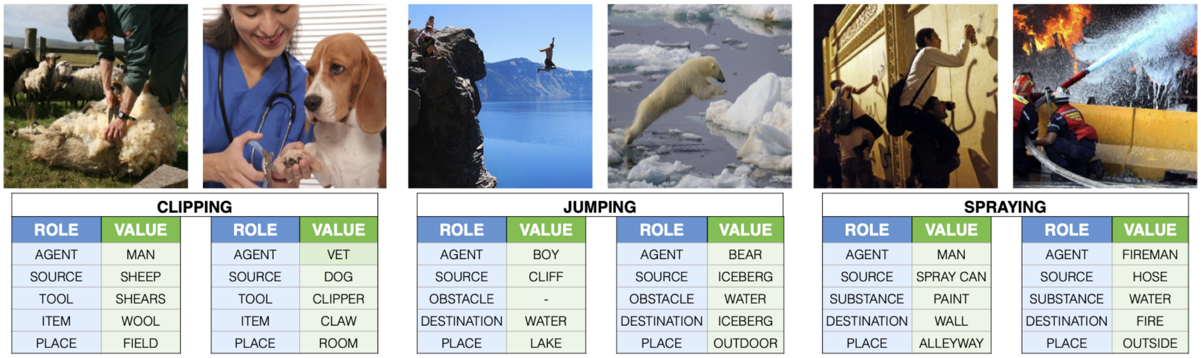
Related Works

2.1 Image Situation Recognition

Situation Recognition in images was first proposed by [10] where they created datasets to understand actions along with localisation of objects and people. Another line of work, imSitu [45] proposed situation recognition via semantic role labelling by leveraging linguistic frameworks, FrameNet [3] and WordNet [23] to formalize situations in the form of verb-role-noun triplets. Recently, grounding has been incorporated with image situation recognition [27] to add a level of understanding for the predicted SRL. Situation recognition requires global understanding of the entire scene, where the verbs, roles and nouns interact with each other to predict a coherent output. Therefore several approaches used CRF [45], LSTMs [27] and Graph neural networks [16] to model the global dependencies among verb and roles. Recently various Transformer [36] based methods have been proposed that claim large performance improvements [6, 7, 40]. In the next subsections we will look into this task more closely and discuss several State-of-the-art approaches mentioned above.

2.1.1 Visual Semantic Role Labeling for Image Understanding

This work introduces situation recognition [45], the problem of producing a structured and a concise summary of the situation happening in an image that includes: (1) the main activity (e.g., clipping), (2) the participating actors, objects, entities, and locations (e.g., man, shears, sheep, wool, and field) and (3) the roles these participants play in the activity (e.g., the man is clipping, the shears are his tool, the wool is being clipped from the sheep, and the clipping is in a field). To define the verbs and roles, the authors use FrameNet, a verb and role lexicon devel- oped by linguists, to define a large space of possible situations and collect a dataset containing over 500 activities, 1,700 roles, 11,000 objects, 125,000 images, and 200,000 unique situations. The authors also introduce a structured prediction baseline using Conditional Random Fields and show that, situation-driven prediction of objects and activities outperforms independent object and activity recognition. CRF was used to model dependencies between verb-role-noun pairs. In particular, a neural network was trained in an end-to-end fashion to both, predict the unary potentials for verbs and nouns, and to perform inference in the CRF. While their model captured the dependency between the verb and role-noun pairs, dependencies between the roles were not modeled explicitly.



CLIPPING

ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	VET
SOURCE	SHEEP	SOURCE	DOG
TOOL	SHEARS	TOOL	CLIPPER
ITEM	WOOL	ITEM	CLAW
PLACE	FIELD	PLACE	ROOM

JUMPING

ROLE	VALUE	ROLE	VALUE
AGENT	BOY	AGENT	BEAR
SOURCE	CLIFF	SOURCE	ICEBERG
OBSTACLE	-	OBSTACLE	WATER
DESTINATION	WATER	DESTINATION	ICEBERG
PLACE	LAKE	PLACE	OUTDOOR

SPRAYING

ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	FIREMAN
SOURCE	SPRAY CAN	SOURCE	HOSE
SUBSTANCE	PAINT	SUBSTANCE	WATER
DESTINATION	WALL	DESTINATION	FIRE
PLACE	ALLEYWAY	PLACE	OUTSIDE

Figure 2.1 [45] Multiple images depicting situations where actors, objects, substances, and locations play different roles in an activity. Below each image is a semantic role frame that describes the situation: the left columns (blue) list verb or action specific roles that are predefined from FrameNet- a broad coverage verb lexicon, while the right columns (green) list the values of each role from ImageNet classes. Three different activities are shown, for 6 different scenes, which highlights that even in the same activity visual properties vary widely between role values (e.g., clipping a dog's nails looks very different from clipping a sheep's wool).

2.1.2 Situation Recognition with Graph Neural Networks

This work addressed the problem of recognizing situations in images and propose a model based on Graph Neural Networks that allows to efficiently capture joint dependencies between roles using neural networks defined on a graph. [16] Their approach is able to propagate information between verbs and roles together and significantly outperforms the previous CRF and RNN based baselines. The authors stressed that the information of the verb and roles heavily depends on each other and hence they propose to model these dependencies through a graph $G = (A, B)$. Where the nodes in graph $a \in A$ are of two types of verb or role, and take unique values of V or N , respectively. Since each image in the dataset is associated with one unique verb, every graph has a single verb node. Edges in the graphs $b = (a', a)$ encode dependencies between role-role or verb- role pairs, and can be directed or undirected. Fig.2.1.2 shows an example of such a graph where verb and role nodes are connected to each other. They design a Gated Graph Neural Networks (GGNNs) that learns the representation of a graph, which is then used to predict node or graph-level output. Each node of a GGNN is associated with a hidden state vector that is updated in a recurrent fashion. At each time step, the hidden state of a node is updated based on its history and incoming messages from its neighbors. These updates are applied simultaneously to all nodes in the graph at each propagation step. The hidden states after T propagation steps are used to predict the output. They adopt the GGNN framework to recognize situations in images. Each image is associated with one verb v that corresponds to a semantic role frame f with a set of roles E_f which is predefined. Then they instantiate a graph G_f for each image that consists of one verb node, and E_f (number of roles associated with the frame) role nodes. To capture the dependency between roles to the full extent, they propose creating undirected edges between all pairs of roles. This approach led to state-of-the-art results and was the first one to establish that situation recognition requires dense and fully connected processing of verbs and roles, which led to the rise of several Transformer based models for situation recognition [6, 7, 40].

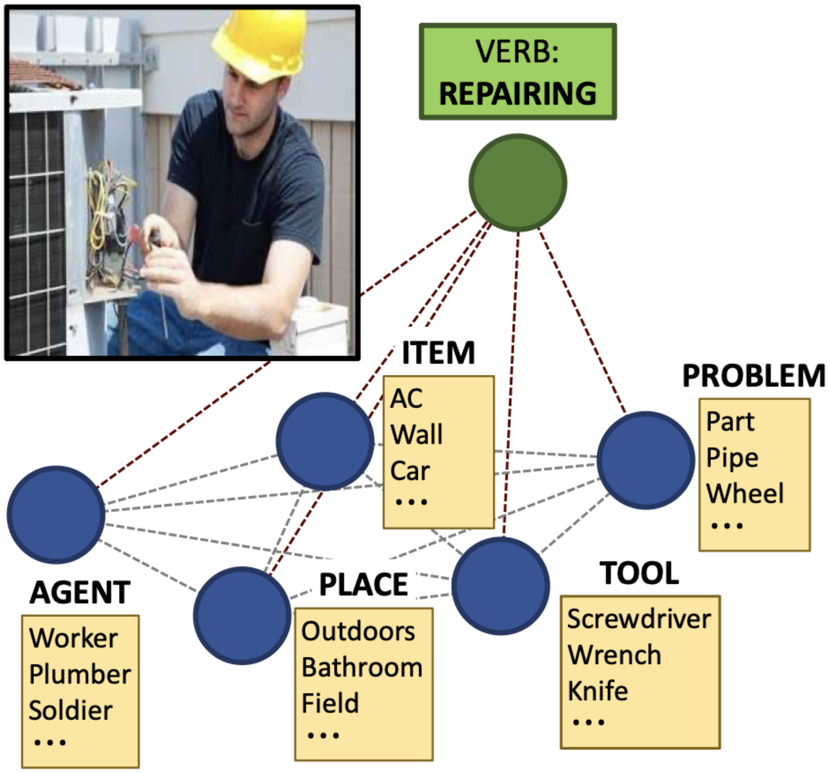


Figure 2.2 [16] Understanding an image involves more than just predicting the most salient action. We need to know who is performing this action, what tools he may be using, what is the location etc. All these information are dependent on each other, and coherently understanding situations requires to model all the verb, roles, and the corresponding nouns together. The figure shows a glimpse of the model that uses a graph to model dependencies between the verb and its roles.

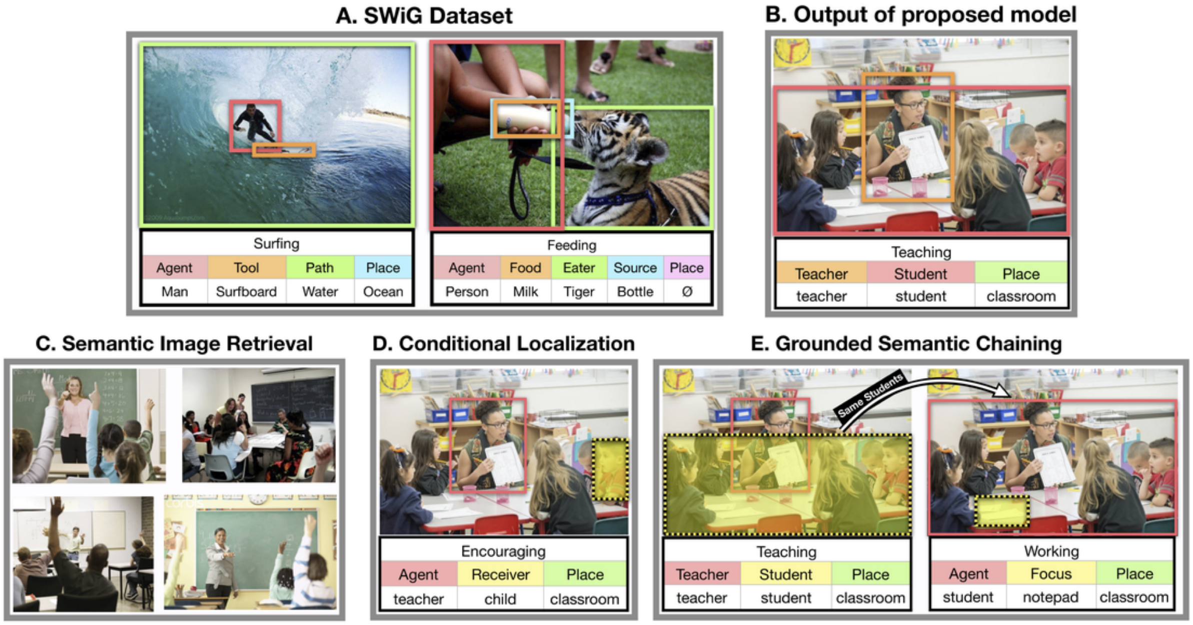

A. SWiG Dataset
Surfing
Agent Tool Path Place
Man Surfboard Water Ocean
Feeding
Agent Food Eater Source Place
Person Milk Tiger Bottle Ø
B. Output of proposed model
Teaching
Teacher Student Place
teacher student classroom
C. Semantic Image Retrieval
D. Conditional Localization
Encouraging
Agent Receiver Place
teacher child classroom
E. Grounded Semantic Chaining
Same Students
Teaching
Teacher Student Place
teacher student classroom
Working
Agent Focus Place
student notepad classroom

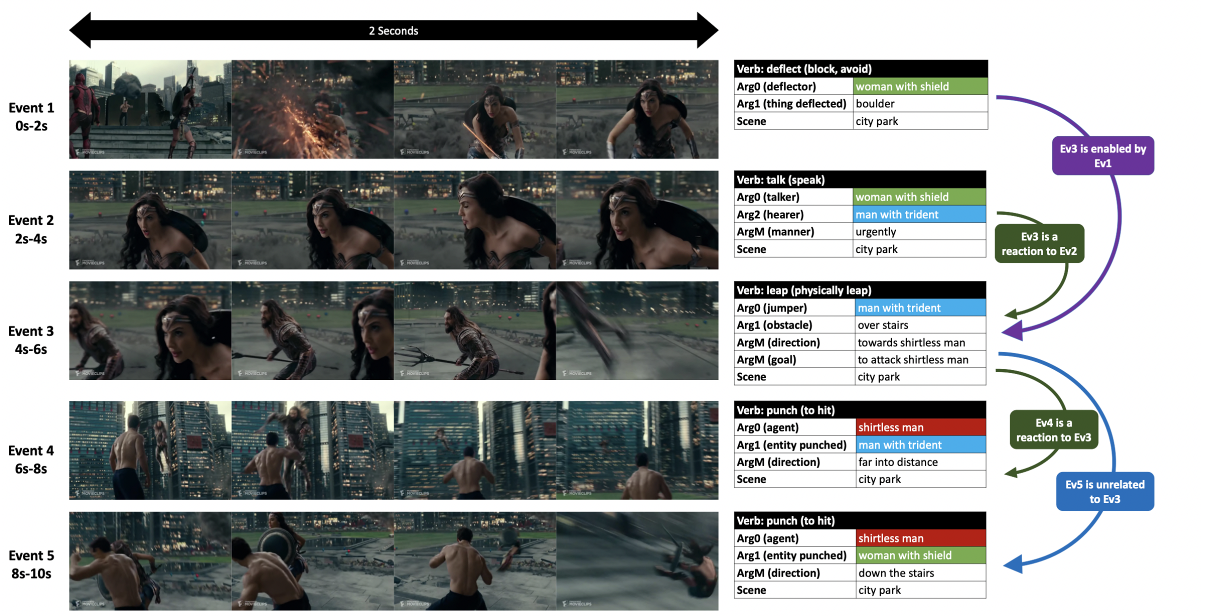
Figure 2.3 [27] A Two examples from our dataset: semantic frames describe primary activities and relevant entities. Groundings are bounding-boxes colored to match roles. B Output of our model (dev set image). C Top-4 nearest neighbors to B using model predictions. Beyond visual similarity, these images are clearly semantically similar. D Output of the conditional model: given a bounding-box (yellow-dashed), predicts a relevant frame. E Example of grounded semantic chaining: given query boxes the model can chain situations together. E.g. the teacher teaches students so they may work on a project

2.1.3 Grounded Image Situation Recognition

This work introduced Grounded Situation Recognition(GSR) [27], is an extension of the previously defined image situation recognition task where in addition to producing structured semantic summaries of images describing: the primary activity, entities engaged in the activity with their roles (e.g. agent, tool), it also grounds the corresponding role entities using bounding box prediction. To study this new task the authors create the Situations With Groundings (SWiG) dataset which adds 278,336 bounding-box groundings to the 11,538 entity classes in the imSitu [45] dataset. They propose a Joint Situation Localizer that jointly predicts situations and groundings, this allows for a role's noun and grounding to be conditioned on the nouns and groundings of previous roles and the verb. This end-to-end training outperforms independent training. The authors use RNN without fusion approach for joint prediction of situations and grounding. The embeddings are extracted from ResNet-50, and are used for verb prediction, then an LSTM sequentially predicts the noun and the bounding boxes for each role in the frame. Moreover the authors show initial findings on three future directions- conditional querying, visual chaining, and grounded semantic aware image retrieval as shown in Figure 2.1.3.

2.2 Video Situation Recognition

2.2.1 Visual Semantic Role Labeling for Video Understanding



Verb: deflect (block, avoid)	
Arg0 (deflector)	woman with shield
Arg1 (thing deflected)	boulder
Scene	city park

Verb: talk (speak)	
Arg0 (talker)	woman with shield
Arg2 (hearer)	man with trident
ArgM (manner)	urgently
Scene	city park

Verb: leap (physically leap)	
Arg0 (jumper)	man with trident
Arg1 (obstacle)	over stairs
ArgM (direction)	towards shirtless man
ArgM (goal)	to attack shirtless man
Scene	city park

Verb: punch (to hit)	
Arg0 (agent)	shirtless man
Arg1 (entity punched)	man with trident
ArgM (direction)	far into distance
Scene	city park

Verb: punch (to hit)	
Arg0 (agent)	shirtless man
Arg1 (entity punched)	woman with shield
ArgM (direction)	down the stairs
Scene	city park

Figure 2.4 [31] A sample video and annotation from VidSitu. The figure shows a 10-second video annotated with 5 events, one for each 2-second interval. Each event consists of a verb (like "deflect") and its arguments (like Arg0 (deflector) and Arg1 (thing deflected)). Entities that participate in multiple events within a clip are co-referenced across all such events (marked using the same color). Finally, all events are related to the central event (Event 3).

Recently, imSitu was extended to videos as VidSitu[31], a large scale video dataset based on short movie clips spanning multiple events. Compared to image situation recognition, VidSRL not only requires understanding the action and the entities involved in a single frame, but also needs to coherently understand the entire video while predicting event-level verb-SRLs and co-referencing the entities participating across events. The videos are represented as a set of related events, wherein each event consists of a verb and multiple entities that fulfill various roles relevant to that event. VidSitu consists of 29K

10-second movie clips richly annotated with a verb and semantic-roles every 2 seconds. Entities are coreferenced across events within a movie clip and events are connected to each other via event-event relations. Clips in VidSitu are drawn from a large collection of movies (3K) and have been chosen to be both complex (4.2 unique verbs within a video) as well as diverse (200 verbs have more than 100 annotations each). The authors proposed to use standard video backbones for feature extraction followed by multiple but separate Transformers to model all the tasks individually, using ground-truth of previous the task to model the next. For feature extraction and verb prediction per event, they use SlowFast [8] pretrained on Kinetics [5] and then finetuned on the VidSitu videos. For Semantic role labeling, a transformer encoder is used to contextualise all the event representations and an autoregressive decoder starting with a ground truth verb is used to predict all the captions for each semantic role one by one.

2.2.2 Hierarchical Self-supervised Representation Learning for Movie Understanding

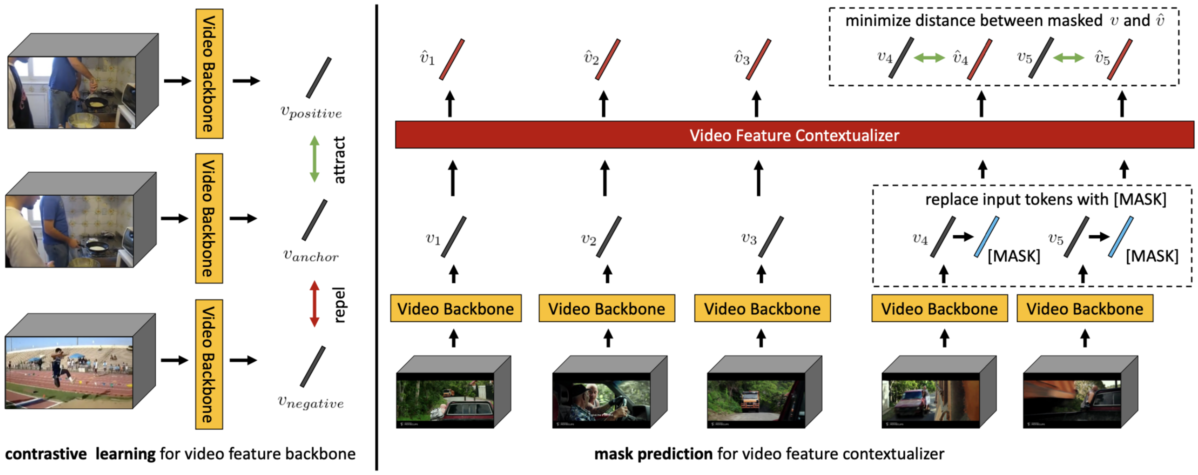


Figure 2.5 [42] Overview of the hierarchical pretraining methods. The left shows how contrastive learning is used to pretrain the video feature backbone — features v_{anchor} and $v_{positive}$ produced from two clips of the same video are pulled together to each other, whereas the feature $v_{negative}$, computed from a clip sampled from another video, is pushed away. Whereas the right shows how to pretrain the transformer feature contextualizer using mask prediction — in this sequence of 5 tokens, input tokens v_2 and v_3 are masked out and fed to the contextualizer, and then forward through to get the outputs $\hat{v}_i$. Then the learning objective is set to minimize the distance between the output tokens $(\hat{v}_2, \hat{v}_3)$ and the masked-out input tokens (v_2, v_3).

This work proposed a self-supervised video representation learning approach for movie understanding [42]. The authors proposed a novel hierarchical pretraining strategy that separately pretrains each level of the hierarchical situation recognition model proposed by [31]. The low-level video feature extractor is trained using contrastive learning. Specifically they use InfoNCE objective, with the goal of pulling together the representations of two clips sampled from the same video, while pushing apart those clips that are sampled from different videos. The high level transformer video contextualising encoder is pretrained using event mask prediction tasks resulting in large performance improvements on SRL. Our goals are different with respect to these previous works in video situation recognition, as we propose to learn and predict all three tasks simultaneously. To achieve this, we predict verb-role pairs on the fly and design a new role query contextualised by video embeddings to model SRL. This eliminates the need for ground-truth verbs and enables end-to-end situation recognition in videos. We also propose to learn contextualised object and video features enabling weakly-supervised grounding for SRL, which was not supported by the previous works.

2.3 Video Understanding

Video understanding is a broad area of research, dominantly involving tasks like action recognition [5, 8, 9, 33, 39, 41], localisation [18, 19], object grounding [30, 44], question answering [35, 47], video captioning [29], and spatio-temporal detection [9, 34]. These tasks involve visual temporal understanding in a sparse uni-dimensional way. Where as, dense video understanding tasks like captioning [49] provides fine-grained details but in an unstructured way, which makes it difficult to evaluate and take decision upon. In contrast, GVSR involves a hierarchy of tasks, coming together to provide a fixed structure, enabling dense situation recognition. The proposed task requires global video understanding through event level predictions and fine-grained details to recognise all the entities involved, the roles they play, and simultaneously ground them. Note that our work on grounding is different from classical spatio-temporal video grounding [48, 44] or referring expressions based segmentation [13] as they require a text query as input. In our case, both the text and the bounding box (grounding) are predicted jointly by the model.

Chapter 3

Grounded Video Situation Recognition

3.1 Problem Formulation

Preliminaries Given a video V consisting of several short events $\mathcal{E} = \{e_i\}$, the complete situation in V, is characterised by 3 tasks. (i) Verb classification, requires predicting the action label v_i associated with each event e_i; (ii) Semantic role labelling (SRL), involves guessing the nouns (captions) $\mathcal{C}_i = \{C_{ik}\}$ for various roles $\mathcal{R}_i = \{r | r \in \mathcal{P}(v_i) \forall r \in \mathcal{R}\}$ associated with the verb v_i. $\mathcal{P}$ is a mapping function from verbs to a set of roles based on VidSitu (extended PropBank [25]) and $\mathcal{R}$ is the set of all roles); and (iii) Spatio-temporal Grounding of each visual role-noun prediction C_{ij} is formulated as selecting one among several bounding box proposals $\mathcal{B}$ obtained from a pretrained object detector applied over subsampled keyframes of the video. We evaluate this against ground-truth annotations done at a keyframe level. No ground truth bounding boxes are used during training.

3.2 VideoWhisperer

We now present the details of our three stage Transformer model, *VideoWhisperer*. A visual overview is presented in Fig. 3.1. For brevity, we request the reader to refer to [36] for now popular details of self- and cross-attention layers used in Transformer encoders and decoders.

3.2.1 Contextualised Video and Object Features (Stage 1)

GVSR is a challenging task, that requires to coherently model spatio-temporal information to understand the salient action, determine the semantic role-noun pairs involved with the action, and simultaneously localise them. Different from previous works that operate only on event level video features, we propose to model both the event and object level features simultaneously. We use a pretrained video backbone ϕ_{vid} to extract event level video embeddings $\mathbf{x}_i^e = \phi_{\text{vid}}(e_i)$.

For representing objects, we subsample frames $\mathcal{F} = \{f_t\}_{t=1}^{T}$ from the entire video V. We use a pretrained object detector ϕ_{obj} and extract top M object proposals from every frame. The box locations

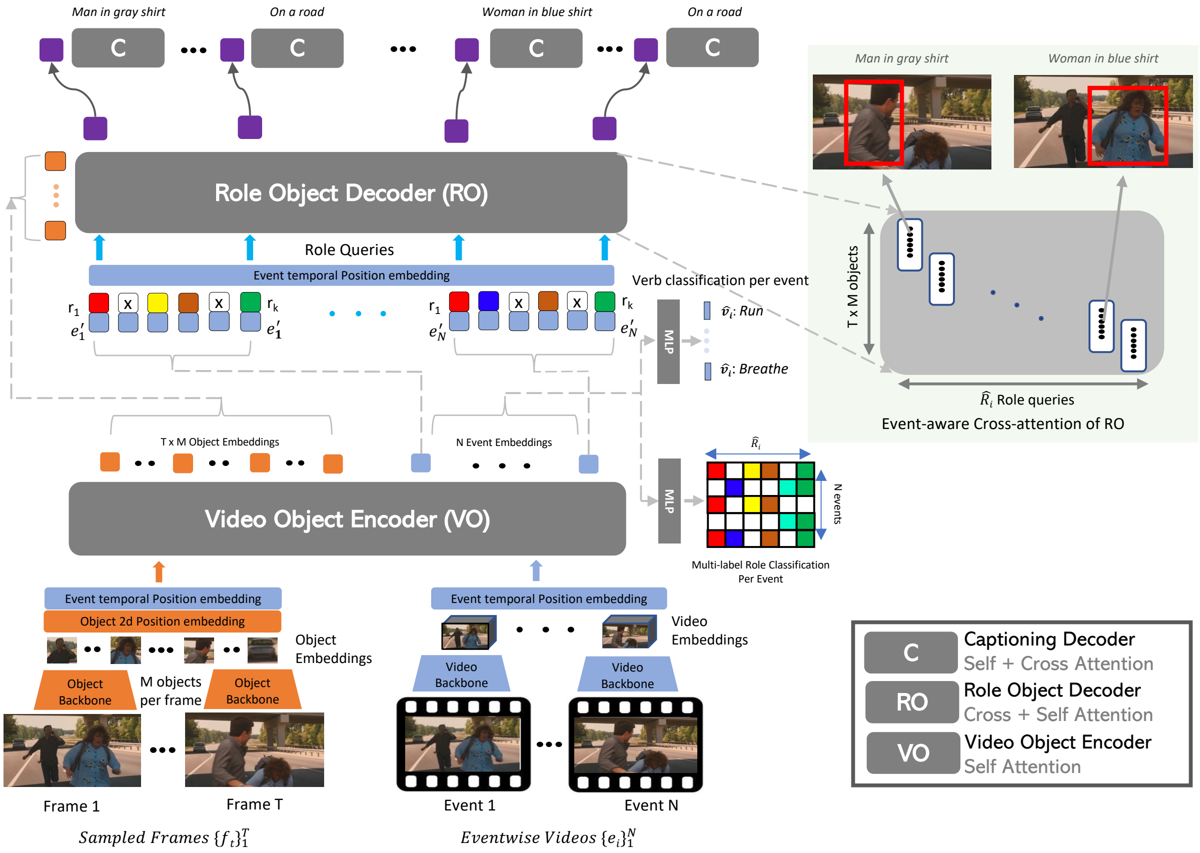


Figure 3.1 VideoWhisperer: We present a new 3-stage Transformer for GVSR. Stage-1 learns the contextualised object and event embeddings through a video-object Transformer encoder (VO), that is used to predict the verb-role pairs for each event. Stage-2 models all the predicted roles by creating role queries contextualised by event embeddings, and attends to all the object proposals through a role-object Transformer decoder (RO) to find the best entity that represents a role. The output embeddings are fed to captioning Transformer decoder (C) to generate captions for each role. Transformer RO's cross-attention ranks all the object proposals enabling localization for each role.

(along with timestamp) and corresponding features are

$$\mathcal{B} = \{b_{mt}\}, m = [1, \ldots, M], t = [1, \ldots, T], \quad \text{and} \quad \{\mathbf{x}^o_{mt}\}_{m=1}^M = \phi_{\text{obj}}(f_t) \quad \text{respectively.} \tag{3.1}$$

The subset of frames associated with an event e_i are computed based on the event's timestamps,

$$\mathcal{F}_i = \{f_t | e_i^{\text{start}} \leq t \leq e_i^{\text{end}}\}\,. \tag{3.2}$$

Specifically, at a sampling rate of 1fps, video V of 10s, and events e_i of 2s each, we associate 3 frames with each event such that the border frames are shared. We can extend this association to all object proposals based on the frame in which they appear and denote this as $\mathcal{B}_i$.

3.2.1.1 Video-Object Transformer Encoder (VO)

Since the object and video embeddings come from different spaces, we align and contextualise them with a Transformer encoder [36]. Event-level position embeddings PE_i and TE type embeddings are added to both representations, event $\mathbf{x}_i^e$ and object $\mathbf{x}_{mt}^o$ $(t \in \mathcal{F}_i)$. In addition, 2D object position embeddings PE_{mt} and frame number embedding FE_t are added to object embeddings $\mathbf{x}_{mt}^o$. Together, they help capture spatio-temporal information. The object and video tokens are passed through multiple self-attention layers to produce contextualised event and object embeddings:

$$[\ldots, \mathbf{o}'_{mt}, \ldots, \mathbf{e}'_i, \ldots] = \text{Transformer}_{\text{VO}}\left([\ldots, \mathbf{x}_{mt}^o + \text{PE}_i + \text{PE}_{mt} + \text{FE}_t + \text{TE}_o, \ldots,\right. \\ \left. \mathbf{x}_i^e + \text{PE}_i + \text{TE}_v, \ldots\right]. \quad (3.3)$$

3.2.1.2 Verb and role classification

Each contextualised event embedding $\mathbf{e}'_i$ is not only empowered to combine information across neighboring events but also focus on key objects that may be relevant. We predict the action label for each event by passing them through a 1-hidden layer MLP,

$$\hat{v}_i = \text{MLP}_e(\mathbf{e}'_i)\,. \quad (3.4)$$

Each verb is associated with a fixed set of roles based on the mapping $\mathcal{P}(\cdot)$. This prior information is required to model the SRL task. Previous works [31, 42] use ground-truth verbs to model SRL and predict both the roles and their corresponding entities. While this setup allows for task specific modelling, it is not practical in the context of end-to-end video situation recognition. To enable GVSR, we predict the relevant roles for each event circumventing the need for ground-truth verbs and mapped roles. Again, we exploit the contextualised event embeddings and pass them through a role-prediction MLP and perform *multi-label* role classification. Essentially, we estimate the roles associated with an event as

$$\hat{\mathcal{R}}_i = \{r | \sigma(\text{MLP}_r(\mathbf{e}'_i)) > \theta_{\text{role}}\}\,, \quad (3.5)$$

where $\sigma(\cdot)$ is the sigmoid function and θ_{role} is a common threshold across all roles (typically set to 0.5). Armed with verb and role predictions, $\hat{v}_i$ and $\hat{\mathcal{R}}_i$, we now look at localising the role-noun pairs and generating the SRL captions.

3.2.2 Semantic Role Labelling with Grounding (Stage 2, 3)

A major challenge in SRL is to disambiguate roles, as the same object (person) may take on different roles in the longer video V. For example, if two people are conversing, the *agent* and *patient* roles will switch between *speaker* and *listener* over the course of the video. Another challenge is to generate descriptive and distinctive captions for each role such that they refer to a specific entity. We propose to use learnable role embeddings $\{\mathbf{r}_{ik}\}_{k=1}^{|\mathcal{R}_i|}$ which are capable of learning distinctive role representations. As mentioned earlier, roles such as *agent*, *patient*, *tool*, *location*, *manner*, *etc*. ask further questions about the salient action.

3.2.2.1 Creating role queries

Each role gets updated by the verb. For example, for an action *jump*, the *agent* would be referred to as the *jumper*. We strengthen the role embeddings by adding the contextualised event embeddings to each role, instead of encoding ground-truth verbs. This eliminates the dependency on the ground-truth verb-role pairs, and enables end-to-end GVSR. Similar to the first stage (VO), we also add event-level temporal positional embeddings to obtain role *query* vectors

$$\mathbf{q}_{ik} = \mathbf{r}_k + \mathbf{e}'_i + \mathrm{PE}_i \, . \tag{3.6}$$

Depending on the setting, k can span all roles $\mathcal{R}$, ground-truth roles $\mathcal{R}_i$ or predicted roles $\hat{\mathcal{R}}_i$.

3.2.2.2 Role-Object Transformer Decoder (RO)

It is hard to achieve rich captions while using features learned for action recognition. Different from prior works [31, 42], we use fine-grained object level representations instead of relying on event-based video features. We now describe the stage two of our VideoWhisperer model, the Transformer decoder for SRL. Our Transformer decoder uses semantic roles as queries and object proposal representations as keys and values. Through the cross-attention layer, the event-aware role query attends to contextualised object embeddings and finds the best objects that represent each role.

We incorporate an event-based attention mask, that limits the roles corresponding to an event to search for objects localised in the same event, while masking out objects from other events. Cross-attention captures local event-level role-object interactions while the self-attention captures the global video level understanding allowing event roles to share information with each other.

We formulate event-aware cross-attention as follows. We first define the query, key, and value tokens fed to the cross-attention layer as

$$\mathbf{q}'_{ik} = W_Q \mathbf{q}_{ik}, \quad \mathbf{k}'_{mt} = W_K \mathbf{o}'_{mt}, \quad \text{and} \quad \mathbf{v}'_{mt} = W_V \mathbf{o}'_{mt} \, . \tag{3.7}$$

Here, $W_{[Q|K|V]}$ are learnable linear layers.

Next, we apply a mask while computing cross-attention to obtain contextualised role embeddings as

$$\mathbf{r}'_{ik} = \sum_{mt} \alpha_{mt} \mathbf{v}'_{mt}, \quad \text{where} \quad \alpha_{mt} = \mathrm{softmax}_{mt}(\langle \mathbf{q}'_{ik}, \mathbf{k}'_{mt} \rangle \cdot \mathbb{1}(f_t \in \mathcal{F}_i)) \, , \tag{3.8}$$

where $\langle \cdot, \cdot \rangle$ is an inner product and $1(\cdot)$ is an indicator function with value 1 when true and $-\infty$ otherwise to ensure that the cross-attention is applied only to the boxes $\mathcal{B}_i$, whose frames f_t appear within the same event e_i.

After multiple layers of cross- and self-attention, the role query extracts objects that best represent the entities for each role.

$$[\ldots, \mathbf{z}_{ik}, \ldots] = \text{Transformer}_{\text{RO}}([\ldots, \mathbf{q}_{ik}, \ldots; \ldots, \mathbf{o}'_{mt}, \ldots]) \,. \tag{3.9}$$

3.2.2.3 Captioning Transformer Decoder (C)

The final stage of our model is a caption generation module. Specifically, we use another Transformer decoder [36] whose input context is the output role embedding $\mathbf{z}_{ik}$ from the previous stage and unroll predictions in an autoregressive manner.

$$\hat{C}_{ik} = \text{Transformer}_{\text{C}}(\mathbf{z}_{ik}) \,. \tag{3.10}$$

The role-object decoder in stage 2 shares all the necessary information through self-attention, and allows us to generate the captions for all the roles in parallel; while [31, 42] generate captions sequentially , *i.e* for a given event, the caption for role k is decoded only after the caption for role $k-1$. This makes VideoWhisperer efficient with a wall-clock runtime of 0.4s for inference on a 10s video, while the baseline [31] requires 0.94 seconds.

3.2.2.4 Grounded Semantic Role Labelling

The entire model is designed in a way to naturally provide SRL with grounding in a weakly-supervised way, without the need for ground-truth bounding boxes during training. Cross-attention through the Transformer decoder RO scores and ranks all the objects based on the role-object relevance at every layer. We extract the cross-attention scores α_{mt} for each role k and event e_i from the final layer of Transformer$_{\text{RO}}$, and identify the highest scoring box and the corresponding timestep as

$$\hat{b}^*_m, \hat{b}^*_t = \arg\max_{m,t} \alpha_{mt} \,. \tag{3.11}$$

3.3 Training and Inference

Training. VideoWhisperer can be trained in an end-to-end fashion, with three losses. The first two losses, CrossEntropy and BinaryCrossEntropy, are tapped from the contextualis ed event embeddings and primarily impact the Video-Object Transformer encoder

$$L_i^{\text{verb}} = CE(\hat{v}_i, v_i) \quad \text{and} \quad L_i^{\text{role}} = \sum_{r \in \mathcal{R}_i} BCE(r \in \hat{\mathcal{R}}_i, r \in \mathcal{R}_i) \,. \tag{3.12}$$

The final component is derived from the ground-truth captions and helps produce meaningful SRL outputs. This is also the source of weak supervision for the grounding task,

$$L_{ik}^{\text{caption}} = \sum_{w} CE(\hat{C}_{ik}^{w}, C_{ik}^{w})\,, \tag{3.13}$$

where the loss is applied in an autoregressive manner to each predicted word w. The combined loss for any training video V is given by

$$\mathcal{L} = \sum_{i} L_{i}^{\text{verb}} + \sum_{i} L_{i}^{\text{role}} + \sum_{ik} L_{ik}^{\text{caption}}\,. \tag{3.14}$$

Inference. At test time, we split the video V into similar events e_i and predict verbs $\hat{v}_i$ and roles $\hat{\mathcal{R}}_i$ for the same. Here, we have two options: (i) we can use the predicted verb and obtain the corresponding roles using a ground-truth mapping between verbs and roles $\mathcal{P}(\hat{v}_i)$, or (ii) only predict captions for the predicted roles $\hat{\mathcal{R}}_i$. We show the impact of these design choices through experiments.

Chapter 4

Experiments

We evaluate our model in two main settings. (i) **Framework 1:** This setup mimics VidSitu [31], where tasks are evaluated separately. We primarily focus on (a) Verb prediction, (b) SRL and (c) Grounded SRL. This setting uses ground-truth verb-role pairs for modelling (b) and (c). (ii) **Framework 2:** End-to-end GVSR, where all the three tasks are modelled together without using ground truth verb-roles.

4.1 Dataset and Metrics

Dataset. We evaluate our model on the VidSitu [31] dataset that consists of 29k videos (23.6k train, 1.3k val, and others in task-specific test sets) collected from a diverse set of 3k movies. All videos are truncated to 10 seconds, have 5 events spanning 2 seconds each and are tagged with verb and SRL annotations. There are a total of 1560 verb classes and each verb is associated with a fixed set of roles among 11 possible options, however not all are used for evaluation due to noisy annotations (we follow the protocol by [31]). For each role the corresponding value is a free-form caption.

Metrics. For verb prediction, we report Acc@K, *i.e* event accuracy considering 10 ground-truth verbs and top-K model predictions and Macro-Averaged Verb Recall@K. For SRL we report CIDEr [37], CIDEr-Vb: Macro-averaged across verbs, CIDEr-Arg: Macro-averaged across roles, LEA [24], and ROUGE-L [17]. For more details on the metrics pleas refer to [31].

4.2 Implementation details.

We implement our model in Pytorch [26]. We extract event (video) features from a pretrained Slow-Fast model [8] for video representation (provided by [31]). For object features, we use a FasterRCNN model [28] provided by BUTD [2] pretrained on the Visual Genome dataset [15]. We sample frames at 1 fps from a 10 second video, resulting in $T = 11$ frames. We extract top $M = 15$ boxes from each frame, resulting in 165 objects per video.

All the three Transformers have the same configurations - they have 3 layers with 8 attention heads, and hidden dimension 1024. We use the tokenizer and vocabulary provided by VidSitu [31] which uses byte pair encoding. We have 5 types of learnable embeddings: (i) event position embeddings PE$_i$ with 5 positions corresponding to each event in time; (ii) object localization 2D spatial embedding; and (iii) role embeddings, for each of the 11 roles. The verb classification MLP has a single hidden layer of 2048 d and produces an output across all 1560 verbs. (iv) Object and Video type embeddings, TE$_o$ and TE$_v$ respectively. (V) Frame number embedding FE$_t$ from frame 1 to 11.

The role classification MLP also has a single hidden layer of 1024 d and produces output in a multi-label setup for all the 11 roles mentioned above. We threshold role prediction scores with $\theta_{\text{role}} = 0.5$.

We use the Adam optimizer [14] with a learning rate of 10^{-4} to train the whole model end-to-end. As we use pretrained features, we train our model on a single RTX-2080 GPU, batch size of 16.

4.3 Grounding SRL: Annotation and Evaluation

As free form captions and their evaluation can be ambiguous, we propose to simultaneously ground each correct role in the spatio-temporal domain. To evaluate grounding performance, we obtain annotations on the validation set. We select the same $T = 11$ frames that are fed to our model sampled at 1fps. For each frame, we ask annotators to see if the visual roles (*agent*, *patient*, *instrument*), can be identified by drawing a bounding box around them using the CVAT tool [1] (see Appendix 4.10 for a thorough discussion). For each event i and role k, we consider all valid boxes and create a dictionary of annotations $\mathcal{G}_{ik}$ with keys as frame number and value as bounding box. During prediction, for each role $r \in \hat{\mathcal{R}}_i$, we extract the highest scoring bounding box as in Eq. 3.11. The Intersection-over-Union (IoU) metric for an event consists of two terms. The first checks if the selected frame appears in the ground-truth dictionary, while the second compares if the predicted box has an overlap greater than θ with the ground-truth annotation,

$$\text{IoU@}\theta = \frac{1}{|\mathcal{R}_i|} \sum_{k=1}^{|\mathcal{R}_i|} 1[\hat{b}_t^* \in \mathcal{G}_{ik}] \cdot 1[\text{IoU}(\hat{b}_m^*, \mathcal{G}_{ik}[t]) > \theta] \,. \tag{4.1}$$

4.4 Framework 1: Using ground truth roles for Grounded SRL

We evaluate our model in two main settings. This setup mimics VidSitu [31], where tasks are evaluated separately, and ground truth from the previous task are used to model the next task. We primarily focus on (a) Verb prediction, (b) SRL and (c) Grounded SRL, and only require ground truth roles for each ground truth verb to model the SRL and grounded SRL task.

4.4.1 Grounded SRL Ablations

We analyze the impact of architecture choices, role query embeddings, and applying a mask in the cross-attention of the role-object decoder. All ablations in this section assume access to the ground-truth verb or roles as this allows us to analyze the effect of various design choices. Similar to [42] we observe

Table 4.1 Architecture ablations. All the models use event-aware cross-attention. + indicates stages of the model. V: Video encoder, VO: Video-Object encoder, VOR: Video-Object-Role encoder, RV: Role-Video decoder, RO: Role-Object decoder, and C: Captioning Transformer.

#	Architecture	Query Emb.	CIDEr	IoU@0.3	IoU@0.5
1	RV + C	Role + GT-verb	47.91 ± 0.53	-	-
2	RO + C	Role + GT-verb	**70.48** ± 1.09	0.14 ± 0.01	0.06 ± 0.003
3	VOR + C	Role + Event	67.4 ± 0.81	0.22 ± 0.00	0.09 ± 0.002
4	V + RO + C	Role + Event	69.15 ± 0.62	0.23 ± 0.03	0.09 ± 0.01
5	VO + RO + C	Role + Event	68.54 ± 0.48	**0.29** ± 0.013	**0.12** ± 0.01

large variance across runs, therefore we report the average accuracy and the standard deviation over 3 runs for all the ablation experiments and 10 runs for the proposed model (VO+RO+C).

4.4.1.1 Architecture design

We present SRL and grounding results in Table 4.1. Rows 1 and 2 use a two-stage Transformer decoder (ignoring the bottom video-object encoder). As there is no event embedding $\mathbf{e}'_i$, role queries are augmented with ground-truth verb embedding. Using role-object pairs (RO) is critical for good performance on captioning as compared to role-video (RV), CIDEr 70.48 vs. 47.91. Moreover, using objects enables weakly-supervised grounding. Row 3 represents a simple Transformer encoder that uses self-attention to model all the video events, objects, and roles (VOR) jointly. As before, role-object attention scores are used to predict grounding. Incorporating videos and objects together improves the grounding performance. We switch from a two-stage to a three-stage model between rows 1, 2, 3 vs. 4 and 5. Rows 2 vs. 5 illustrates the impact of including the video-object encoder. We see a significant improvement in grounding performance 0.14 to 0.29 for IoU@0.3 and 0.06 to 0.12 for IoU@0.5 without significantly affecting captioning performance. Similarly, rows 4 vs. 5 demonstrate the impact of contextualizing object embeddings by events. In particular, using contextualised object representations $\mathbf{o}'_{mt}$ seems to help as compared against base features $\mathbf{x}^o_{mt}$.

4.4.1.2 Role query embeddings design

Prior works in situation recognition [7, 31, 40] use verb embeddings to identify entities from both images or videos. In this ablation, we show that instead of learning verb embeddings that only capture the uni-dimensional meaning of a verb and ignore the entities involved, event (or video) embeddings remember details and are suitable for SRL. In fact, Table 4.2 (architecture: VO + RO + C) row 2 vs. 3

Table 4.2 Comparing role query embeddings.

#	Query Emb.	CIDEr	IoU@0.3	IoU@0.5
1	Role only	68.61 ± 0.61	0.27 ± 0.011	0.11 ± 0.009
2	Role + GT-verb	68.71 ± 1.06	0.25 ± 0.02	0.10 ± 0.01
3	Role + Event	68.54 ± 0.48	**0.29** ± 0.013	**0.12** ± 0.01

show that event embeddings are comparable and slightly better than GT-verb embeddings when evaluated on SRL and Grounding respectively, eliminating the need for GT verbs. Somewhat surprisingly, we see that the role embeddings alone perform quite well. We believe this may be due to role embeddings (i) capture the generic meaning like *agent* and *patient* and can generate the correct entities irrespective of the action information; and (ii) the role query attends to object features which are contextualised by video information, so the objects may carry some action information with them.

4.4.1.3 Masked cross-Attention in RO decoder

We use masking in event-aware cross-attention to ensure that the roles of an event attend to objects coming from the same event. As seen in Table 4.3 (model: VO + RO + C, query is role + event embedding), this reduces the object pool to search from and improves both the SRL and Grounding

Table 4.3 Impact of masking in RO decoder.

Mask	CIDEr	IoU@0.3	IoU@0.5
No	67.02 ± 0.51	0.25 ± 0.02	0.10 ± 0.012
Yes	**68.54** ± 0.48	**0.29** ± 0.013	**0.12** ± 0.01

4.4.2 Grounded SRL SoTA comparison on the validation set

In Table 4.4, we compare our results against VidSitu [31] and a concurrent work that uses far better features [42]. We reproduce results for VidSitu [31] by teacher-forcing the ground-truth role pairs to make a fair comparison while results for work [42] are as reported in their paper. Nevertheless, we achieve state-of-the-art performance with a 22 points gain in CIDEr score over [31] and a 8 point gain over [42], while using features from [31]. Moreover, our model allows grounding, something not afforded by the previous approaches.

Table 4.4 SoTA comparison, results for SRL and grounding with GT verb and role pairs.

Method	CIDEr	C-Vb	C-Arg	R-L	Lea	IoU@0.3	IoU@0.5
SlowFast+TxE+TxD [31]	46.01	56.37	43.58	43.04	**50.89**	-	-
Slow-D+TxE+TxD [42]	60.34 ± 0.75	69.12 ± 1.43	53.87 ± 0.97	43.77 ± 0.38	46.77 ± 0.61	-	-
VideoWhisperer (Ours)	**68.54** ± 0.48	**77.48** ± 1.52	**61.55** ± 0.79	**45.70** ± 0.30	47.54 ± 0.55	**0.29** ± 0.013	**0.12** ± 0.01
Human Level	84.85	91.7	80.15	39.77	70.33	-	-

4.4.3 SRL SoTA comparison on the test set

We evaluate our model on the test set from the evaluation servers of VidSitu [31]. A constant improvement over [31] can be seen in Table 4.5. The trend is similar when compared with Table 4.4 from the main paper that reports performance on the validation set.

Note that we do not report grounding metrics as the ground-truth nouns are not available. We are currently working with the authors of VidSitu [31] to establish this as part of the benchmark.

Table 4.5 Results of SRL with GT verb and role pairs on the test dataset. VidSitu's [31] results are as reported in their paper.

Method	CIDEr	C-Vb	C-Arg	R-L	Lea	IoU@0.3	IoU@0.5
SlowFast+TxE+TxD [31]	47.25	52.92	45.48	43.46	**50.88**	-	-
VideoWhisperer (Ours)	**68.04**	**81.23**	**62.19**	**46.15**	48.77	-	-
Human Level	83.68	87.78	79.29	40.04	71.77	-	-

4.5 Framework 2: GVSR- Joint Prediction of Video Situations

The primary goal of our work is to enable joint prediction of the verb, roles, entities, and grounding.

4.5.1 Verb prediction

is an extremely challenging problem due to the long-tail nature of the dataset. In fact, to alleviate this challenge, verb metrics are computed by comparing predictions against 10 ground-truth human annotations. In Table 4.6, we evaluate verb prediction performance when training the model for verb prediction only (rows 1-3) or training it jointly for GVSR (rows 4, 5). Using a simple video-only

Table 4.6 Verb prediction performance. Rows 1-3 train only for verb prediction. Rows 4, 5 are trained for GVSR.

#	Architecture	Acc@1	Acc@5	Rec@5
1	Baseline [31]	46.79	75.90	23.38
2	V	48.82	78.01	23.32
3	VO	**49.73**	**78.72**	24.72
4	V + RV + C	40.83	70.73	24.37
5	VO + RO + C	45.06	75.59	**25.25**

transformer encoder boosts performance over independent predictions for the five event clips (46.8% to 48.8%, rows 1 vs. 2). Including objects through the video-object encoder (row 3) provides an additional boost resulting in the highest performance at 49.73% on Accuracy@1.

A similar improvement is observed in rows 4 to 5 (V vs. VO stage 1 encoder). Interestingly, the reduced performance of rows 4 and 5 as compared against rows 1-3 is primarily because the best epoch corresponding to the highest verb accuracy does not coincide with highest SRL performance. Hence, while the verb Accuracy@1 of the GVSR model does reach 49% during training it degrades subsequently due to overfitting. Nevertheless, we observe that the macro-averaged Recall@5 is highest for our model, indicating that our model focuses on all verbs rather than just the dominant classes. In section 4.6, we show the challenges of the large imbalance and perform experiments that indicate that classic re-weighting or re-sampling methods are unable to improve performance in a meaningful mannner. Addressing this aspect is left for future work.

4.5.2 Understanding role prediction

The verb-role prediction accuracy is crucial for GVSR, since the SRL task is modelled on role-queries. In Table 4.7 we analyse role prediction in various settings to understand its effect on SRL. Previous work [31] used ground-truth verbs for SRL, while roles and their entities or values are predicted sequentially. This setting is termed “GT, Pred” (row 2) as it uses the ground-truth verb but predicts the roles. We argue that as the verb-role mapping $\mathcal{P}$ is a deterministic lookup table, this setting is less interesting. We enforce a “GT, GT” setting with ground-truth verbs and roles in [31] by teacher-forcing the GT roles while unrolling role-noun predictions (row 1). Another setting is where the verb is predicted and roles are obtained via lookup, “Pred, GT map” (row 3). Note that this enables end-to-end SRL, albeit in two steps. The last setting, “Pred, Pred” predicts both verb and role on-the-fly (row 4).

Comparing within variants of [31], surprisingly, row 1 does not perform much better than row 2 on CIDEr. This may be because the model is trained on GT verbs and is able to predict most of the roles

Table 4.7 Role prediction in various settings. Role F1 is the F1 score averaged over all role classes.

#	Architecture	Verb	Role	V. Acc@1	Role F1	CIDEr
1	VidSitu [31]	GT	GT	-	-	46.01
2		GT	Pred	-	0.88	45.52
3		Pred	GT map	46.79	-	29.93
4		Pred	Pred	46.79	0.47	30.33
5	RO+C	GT	GT	-	-	70.48
6	VO+RO+C	Pred	GT	45.06	-	68.54
7	VO+RO+C	Pred	GT map	45.06	-	51.24
8	VO+RO+C	Pred	Pred	44.05	0.44	52.30

correctly (row 2, Role F1 = 0.88). Subsequently, both rows 3 and 4 show a large performance reduction indicating the over-reliance on ground-truth verb. We see similar trends for our models. Rows 7 and 8 with predicted verb-role pairs lead to reduced SRL performance as compared against rows 5 and 6. Nevertheless, our "Pred, Pred" CIDEr score of 52.3 is still higher than the baseline "GT, GT" at 46.0. Next Role Prediction discusses further challenges of multi-label and imbalance in predicting roles.

4.5.3 Role prediction

Role prediction is critical for end-to-end GVSR. We analyse its performance for each role separately. As can be seen from Table 4.8 roles like *Arg0, Arg1, Ascn, ADir, AMnr* which appears a lot more frequently than other roles in the dataset, have both high precision and recall, suggesting that role prediction can be done with a reasonably high accuracy directly from the video features. Other roles that appears less frequently have a good precision but a very low recall, which is expected due to the long tail nature of roles.

4.5.4 GVSR Evaluation

We evaluate our end-to-end model for grounded video situation recognition. In order to enable end-to-end GVSR in [31], we use it in the "Pred, Pred" setting discussed above, that allows verb, role, and SRL predictions. Table 4.9 shows that our model improves SRL performance over Vidsitu [31] by a margin of 22% on CIDEr score. In addition to that, our model also enables Grounded SRL, not achievable in VidSitu [31].

Table 4.8 Precision, Recall and, F1 score for role-prediction performance on all the role classes. Architecture is VO + RO + C in the "Pred Pred" mode.

Method	Role-name	Precision	Recall	F1
VideoWhisperer	Arg0	0.90	0.97	0.93
	Arg1	0.79	0.93	0.86
	Arg2	0.55	0.26	0.36
	Arg3	0.30	0.05	0.09
	Arg4	0.15	0.04	0.06
	AScn	0.74	0.93	0.83
	ADir	0.66	0.49	0.56
	APrp	0.36	0.03	0.06
	AMnr	0.71	0.66	0.68
	ALoc	0.40	0.12	0.19
	AGol	0.65	0.15	0.24

Table 4.9 GVSR: Results for end-to-end situation recognition. Our model architecture is VO+RO+C.

Model	Prediction			Verb	CIDEr	C-Vb	C-Arg	R-L	Lea	IoU	
	Verb	Role	SRL	Acc@1						0.3	0.5
VidSitu [31]	✓	✓	✓	46.79	30.33	39.56	23.97	29.98	35.92	-	-
VideoWhisperer	✓	✓	✓	44.06	52.30	61.77	38.18	35.84	38.00	0.13	0.05
	✓	GT	✓	45.06	68.54	77.48	61.55	45.70	47.54	0.29	0.12

4.6 Long Tailed Verb Classification

The grounded SRL task depends heavily on the action information. In addition to complex scenes, the VidSitu dataset encompasses a large number of verbs and has a long-tailed distribution. In fact, the number of verbs, 1560, is 2-4x larger than popular large-scale video action recognition datasets (Kinetics400 / Kinetics700). We believe that these are the key challenges that result in lower performance for verb classification which inevitably affects the SRL.

We experiment with three common approaches to handle long-tailed distributions. (i) Loss reweighting applies weights corresponding to the inverse verb frequency to the cross-entropy loss; (ii) Focal loss is applied as described in [20] (with gamma = 2.0); and (iii) Balanced sampling, we apply a weight for each sample such that the `DataLoader` picks samples with a higher weight. The results are presented in Table 4.10.

Table 4.10 Results of three common approaches to handle long-tailed distribution of verbs. V only represents the Video encoder (no object features) trained only for verb prediction.

Method	Verb Acc@1
V only	48.82
V only + Loss Re-weighting	48.91
V only + Focal loss	47.81
V only + Balanced sampling	35.38

Unfortunately, we do not see any significant improvement using these simple approaches. We have observed that the dataset is very challenging and has complex movie events with fast shot changes and many actions can be confusing. For example in Figure 4.1 the woman turns while walking, but the model predicts “Walk” instead of “Turn” which is the dominant, but less significant action (if one considers duration). Balanced sampling in particular leads to a significant drop since our sample consists of 5 event clips, each with a verb. When rare verbs are oversampled, co-occurring event clips with potentially not-so-rare verbs are also oversampled, leading to a skewed training dataset. This is similar to the challenges of applying balanced sampling to multi-label classification.

4.7 Qualitative Evaluation

We visualize the predictions of VideoWhisperer (Pred-GT) in Fig. 4.1 for one video of 10 seconds[1] and see that it performs reasonably well given the complexity of the task. VideoWhisperer correctly predicts verbs for actions like “open” and “walk”. Given the large action space and complex scenes, there can be multiple correct actions, *e.g* in Ev2 we see a reasonable “walk” instead of “turn”.

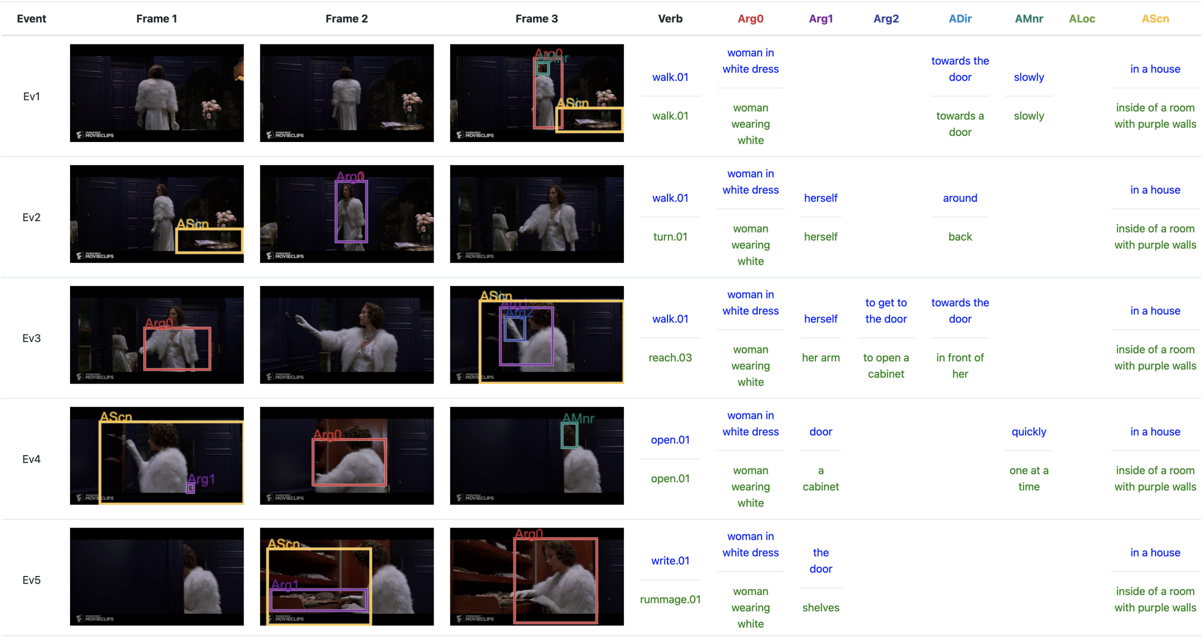
For SRL, the model generates diverse captions with good accuracy, like “woman in white dress”. Even though the ground-truth is syntactically different, “woman wearing white”, they both mean the same. In fact, this is our primary motivation to introduce grounding. In Ev3, the model incorrectly predicts “walk” as the verb instead of ”reach”. While “walk” does not have the role Arg2, we are able to predict a valid caption “to get to the door” while grounding the woman’s arm in Frame3. We see that our model correctly understands the meaning of Arg2 as we use ground-truth role embeddings

[1] More examples on our project page, https://zeeshank95.github.io/grvidsitu/GVSR.html.

combined with event features for SRL. This shows the importance of event embeddings, as they may recall fine-grained details about the original action even when there are errors in verb prediction.

For grounding SRL, we see that the model is able to localize the roles decently, without any bounding box supervision during training. While we evaluate grounding only for Arg0, Arg1, and Arg2 (when available), we show the predictions for other roles as well. In Fig. 4.1, the model is able to ground the visual roles Arg0 and Arg1 correctly. For non-visual roles like *"Manner"*, the model focuses its attention to the face, often relevant for most expressions and mannerisms.

Video ID: v_q6j_0vS_NNM_seg_175_185

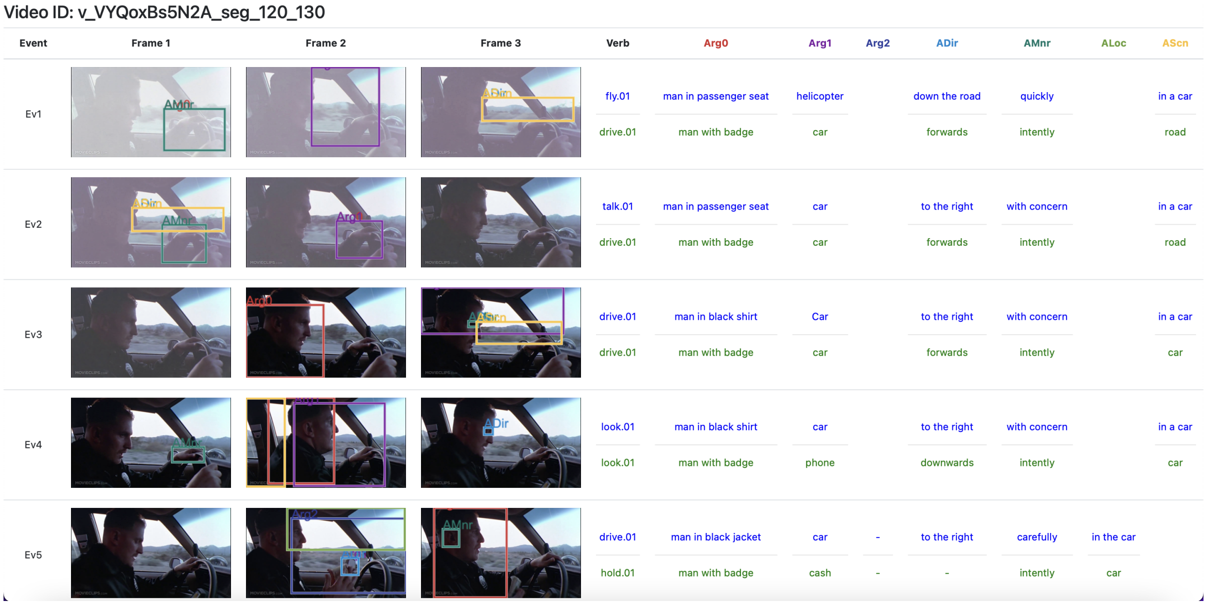

Arg0
AMnr
AScn
AScn
Arg0
Arg0
AScn
Arg1
Arg0
AScn
Arg1
Arg0
AMnr
AScn
Arg1
Arg0

Event	Frame 1	Frame 2	Frame 3	Verb	Arg0	Arg1	Arg2	ADir	AMnr	ALoc	AScn
Ev1				walk.01	woman in white dress			towards the door	slowly		in a house
				walk.01	woman wearing white			towards a door	slowly		inside of a room with purple walls
Ev2				walk.01	woman in white dress	herself		around			in a house
				turn.01	woman wearing white	herself		back			inside of a room with purple walls
Ev3				walk.01	woman in white dress	herself	to get to the door	towards the door			in a house
				reach.03	woman wearing white	her arm	to open a cabinet	in front of her			inside of a room with purple walls
Ev4				open.01	woman in white dress	door			quickly		in a house
				open.01	woman wearing white	a cabinet			one at a time		inside of a room with purple walls
Ev5				write.01	woman in white dress	the door					in a house
				rummage.01	woman wearing white	shelves					inside of a room with purple walls

Figure 4.1 We show the results for a 10s clip that can be viewed here: https://youtu.be/q6j_0vS_NNM?t=175. The video is broken down to 5 events indicated by the row labels Ev1 to Ev5. At a 1fps sampling rate, we obtain boxes from 3 frames for each event (with Frame3 of event $i-1$ being the same as Frame1 of event i). On the right side of the table, we show the predictions for the verb and various roles in the "Pred GT" mode, discussed in Table 4.7 (row 6). Predictions are depicted in blue, while the ground-truth is in green. Each role is assigned a specific color (see table header), and boxes for many of them can be found overlaid on the video frames (with the same edge color).

4.8 More Qualitative Results

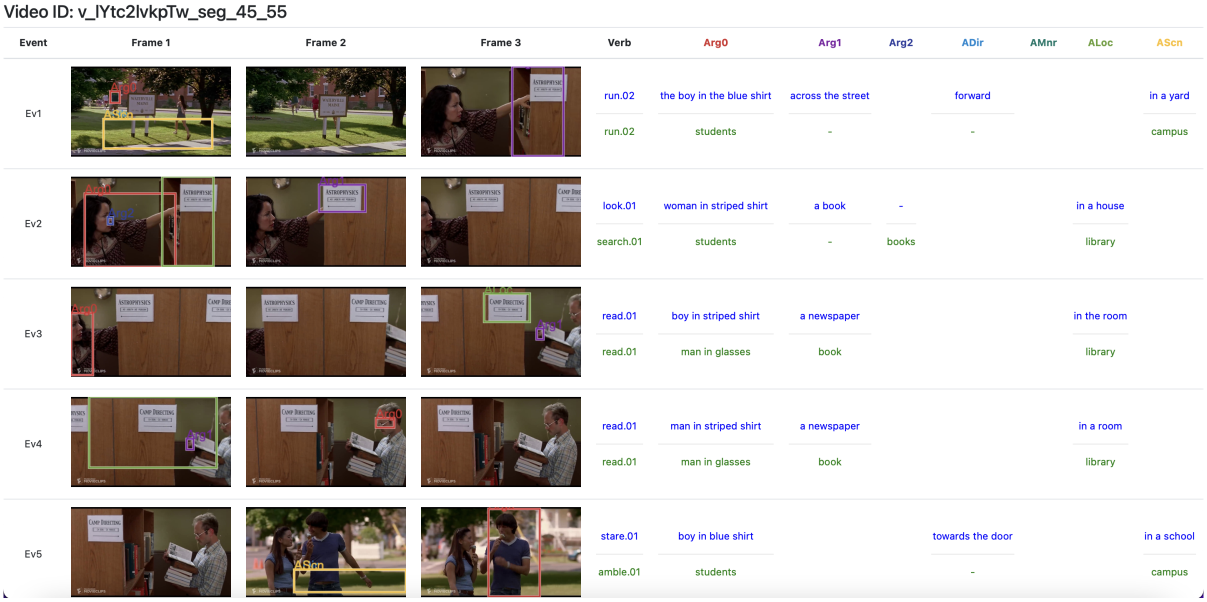
We show more qualitative results on six 10 second video clips taken from movies.



Video ID: v_VYQoxBs5N2A_seg_120_130

Event	Frame 1	Frame 2	Frame 3	Verb	Arg0	Arg1	Arg2	ADir	AMnr	ALoc	AScn
Ev1				fly.01 drive.01	man in passenger seat man with badge	helicopter car		down the road forwards	quickly intently		in a car road
Ev2				talk.01 drive.01	man in passenger seat man with badge	car car		to the right forwards	with concern intently		in a car road
Ev3				drive.01 drive.01	man in black shirt man with badge	Car car		to the right forwards	with concern intently		in a car car
Ev4				look.01 look.01	man in black shirt man with badge	car phone		to the right downwards	with concern intently		in a car car
Ev5				drive.01 hold.01	man in black jacket man with badge	car cash	- -	to the right -	carefully intently	in the car car	

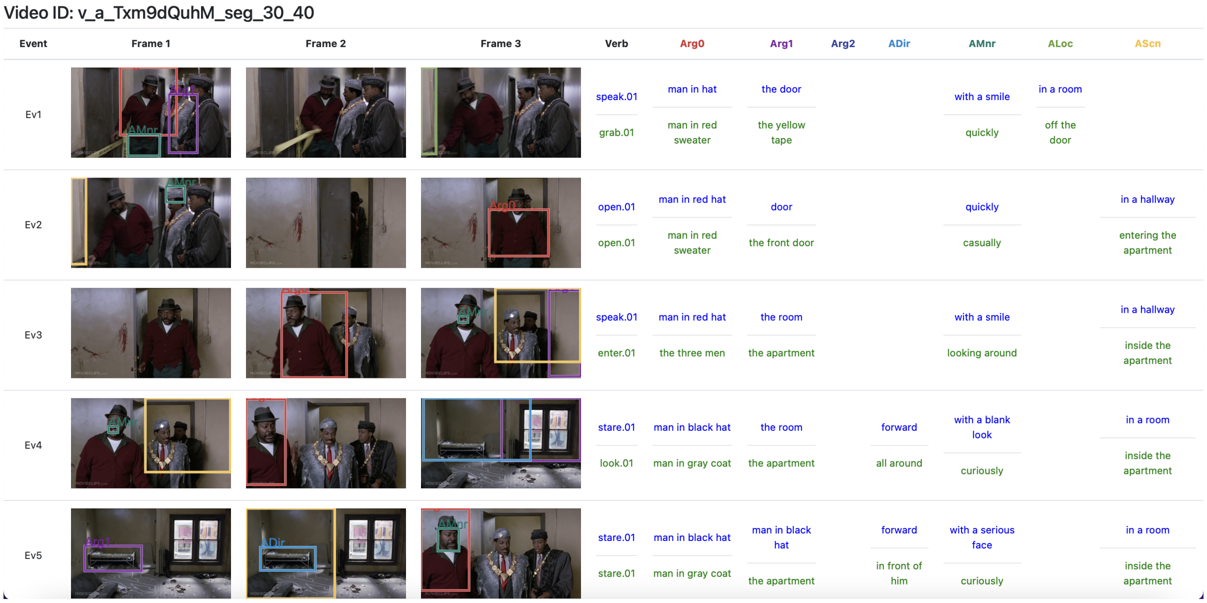
Figure 4.2 We show the results for a 10s movie clip that can be viewed here: https://youtu.be/VYQoxBs5N2A?t=120



Video ID: v_lYtc2lvkpTw_seg_45_55

Event	Frame 1	Frame 2	Frame 3	Verb	Arg0	Arg1	Arg2	ADir	AMnr	ALoc	AScn
Ev1				run.02 run.02	the boy in the blue shirt students	across the street -		forward -			in a yard campus
Ev2				look.01 search.01	woman in striped shirt students	a book -	- books			in a house library	
Ev3				read.01 read.01	boy in striped shirt man in glasses	a newspaper book				in the room library	
Ev4				read.01 read.01	man in striped shirt man in glasses	a newspaper book				in a room library	
Ev5				stare.01 amble.01	boy in blue shirt students			towards the door -			in a school campus

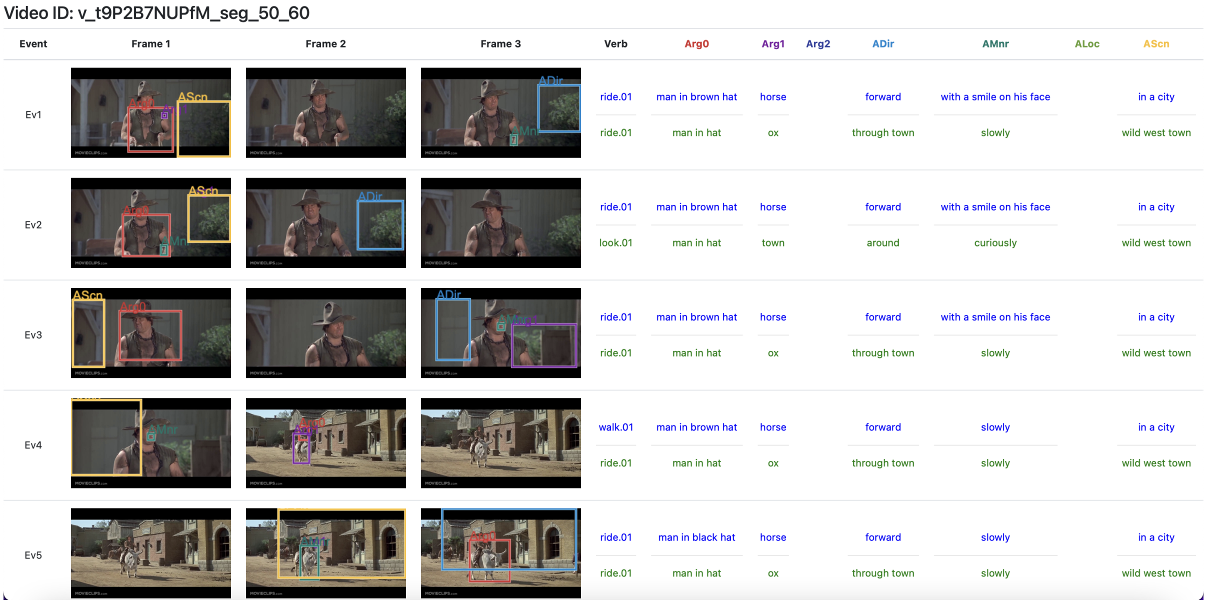
Figure 4.3 We show the results for a 10s movie clip that can be viewed here: https://youtu.be/lYtc2lvkpTw?t=45



Video ID: v_a_Txm9dQuhM_seg_30_40

Event	Frame 1	Frame 2	Frame 3	Verb	Arg0	Arg1	Arg2	ADir	AMnr	ALoc	AScn
Ev1				speak.01 grab.01	man in hat man in red sweater	the door the yellow tape			with a smile quickly	in a room off the door	
Ev2				open.01 open.01	man in red hat man in red sweater	door the front door			quickly casually		in a hallway entering the apartment
Ev3				speak.01 enter.01	man in red hat the three men	the room the apartment			with a smile looking around		in a hallway inside the apartment
Ev4				stare.01 look.01	man in black hat man in gray coat	the room the apartment		forward all around	with a blank look curiously		in a room inside the apartment
Ev5				stare.01 stare.01	man in black hat man in gray coat	man in black hat the apartment		forward in front of him	with a serious face curiously		in a room inside the apartment

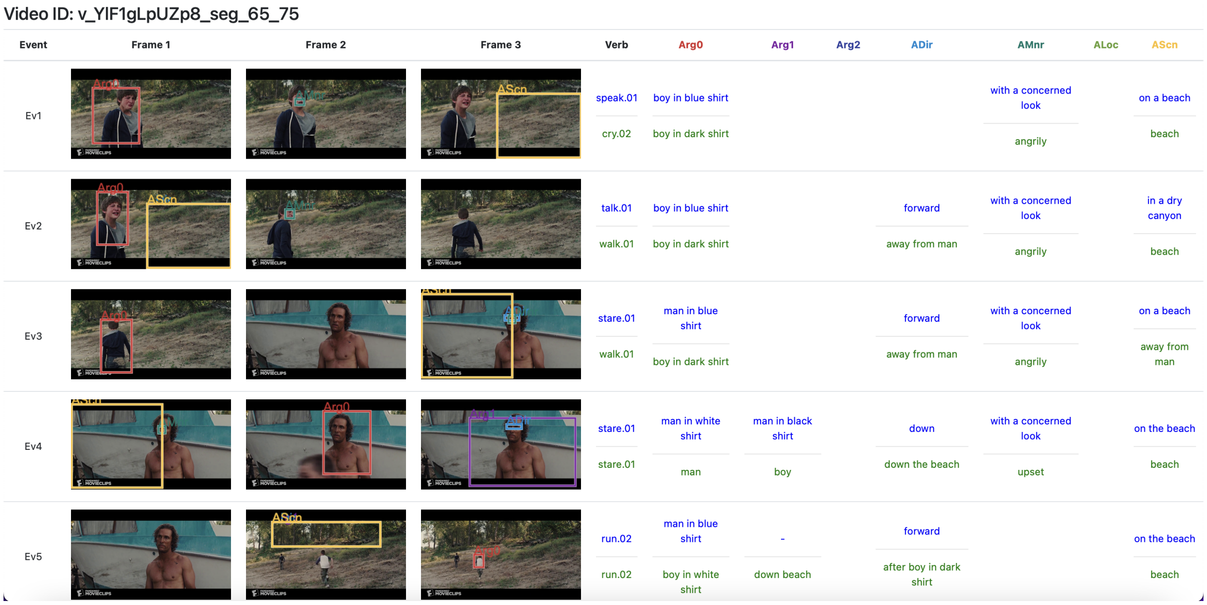
Figure 4.4 We show the results for a 10s movie clip that can be viewed here: https://youtu.be/a_Txm9dQuhM?t=30



Video ID: v_t9P2B7NUPfM_seg_50_60

Event	Frame 1	Frame 2	Frame 3	Verb	Arg0	Arg1	Arg2	ADir	AMnr	ALoc	AScn
Ev1				ride.01 ride.01	man in brown hat man in hat	horse ox		forward through town	with a smile on his face slowly		in a city wild west town
Ev2				ride.01 look.01	man in brown hat man in hat	horse town		forward around	with a smile on his face curiously		in a city wild west town
Ev3				ride.01 ride.01	man in brown hat man in hat	horse ox		forward through town	with a smile on his face slowly		in a city wild west town
Ev4				walk.01 ride.01	man in brown hat man in hat	horse ox		forward through town	slowly slowly		in a city wild west town
Ev5				ride.01 ride.01	man in black hat man in hat	horse ox		forward through town	slowly slowly		in a city wild west town

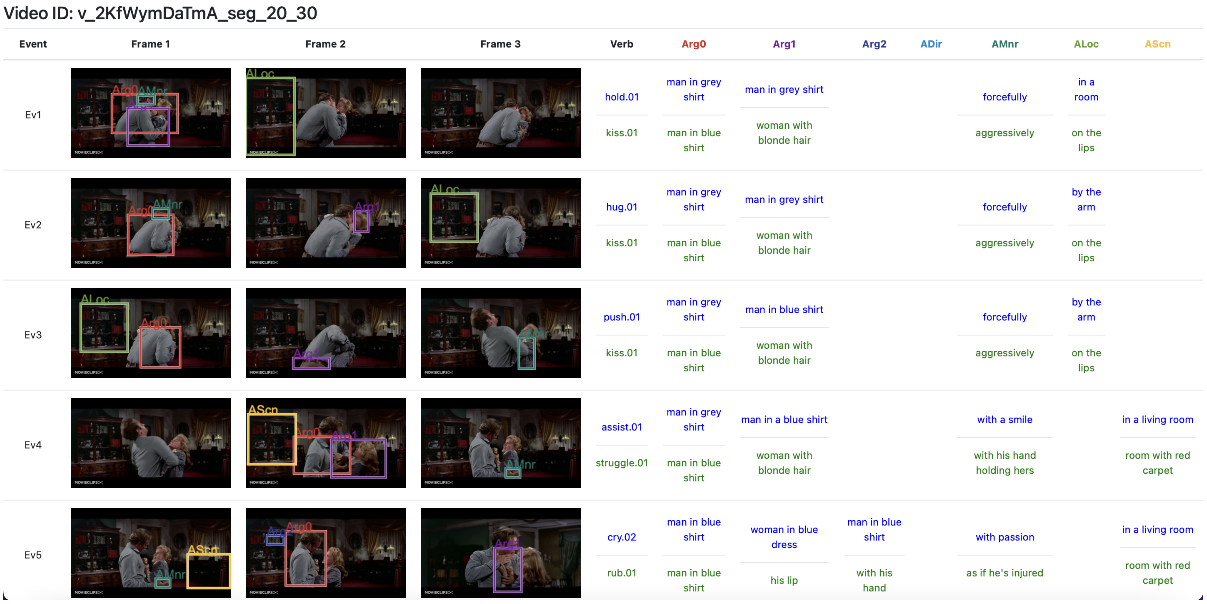
Figure 4.5 We show the results for a 10s movie clip that can be viewed here: https://youtu.be/t9P2B7NUPfM?t=50



Video ID: v_YIF1gLpUZp8_seg_65_75

Event	Frame 1	Frame 2	Frame 3	Verb	Arg0	Arg1	Arg2	ADir	AMnr	ALoc	AScn
Ev1	Arg0	AMnr	AScn	speak.01	boy in blue shirt				with a concerned look		on a beach
				cry.02	boy in dark shirt				angrily		beach
Ev2	Arg0 AScn	AMnr		talk.01	boy in blue shirt			forward	with a concerned look		in a dry canyon
				walk.01	boy in dark shirt			away from man	angrily		beach
Ev3	Arg0		AScn ADir	stare.01	man in blue shirt			forward	with a concerned look		on a beach
				walk.01	boy in dark shirt			away from man	angrily		away from man
Ev4	AScn AMnr	Arg0	Arg1 ADir	stare.01	man in white shirt	man in black shirt		down	with a concerned look		on the beach
				stare.01	man	boy		down the beach	upset		beach
Ev5		AScn	Arg0	run.02	man in blue shirt	-		forward			on the beach
				run.02	boy in white shirt	down beach		after boy in dark shirt			beach

Figure 4.6 We show the results for a 10s movie clip that can be viewed here: https://youtu.be/YIF1gLpUZp8?t=65



Video ID: v_2KfWymDaTmA_seg_20_30

Event	Frame 1	Frame 2	Frame 3	Verb	Arg0	Arg1	Arg2	ADir	AMnr	ALoc	AScn
Ev1	Arg0 AMnr Arg1	ALoc		hold.01	man in grey shirt	man in grey shirt			forcefully	in a room	
				kiss.01	man in blue shirt	woman with blonde hair			aggressively	on the lips	
Ev2	Arg0 AMnr	Arg1	ALoc	hug.01	man in grey shirt	man in grey shirt			forcefully	by the arm	
				kiss.01	man in blue shirt	woman with blonde hair			aggressively	on the lips	
Ev3	ALoc Arg0	Arg1	AMnr	push.01	man in grey shirt	man in blue shirt			forcefully	by the arm	
				kiss.01	man in blue shirt	woman with blonde hair			aggressively	on the lips	
Ev4		AScn Arg0 Arg1	AMnr	assist.01	man in grey shirt	man in a blue shirt			with a smile		in a living room
				struggle.01	man in blue shirt	woman with blonde hair			with his hand holding hers		room with red carpet
Ev5	AScn AMnr	Arg2 Arg0	Arg1	cry.02	man in blue shirt	woman in blue dress	man in blue shirt		with passion		in a living room
				rub.01	man in blue shirt	his lip	with his hand		as if he's injured		room with red carpet

Figure 4.7 We show the results for a 10s movie clip that can be viewed here: https://youtu.be/2KfWymDaTmA?t=20

4.9 Limitations

Limitations for our current model are with verb and role prediction and disambiguation, improving the quality and diversity of captions to go beyond frequent words, and the division of attention towards multiple instances of the same object that appears throughout a video. Nevertheless, we hope that this work inspires the community to couple videos and their descriptions. We describe each challenge in detail: (i) role disambiguation, (ii) descriptive caption generation, and (iii) localisation. We describe each aspect in detail.

Role disambiguation directly depends on the event features, since we use role queries contextualised by event embeddings. Event embeddings help in disambiguation of role even when the predicted action is incorrect. But in many cases when the event embedding captures an action very far from the ground-truth, the role query gets updated based on the incorrect action and this hampers role disambiguation, in turn affecting the quality of SRL captioning and grounding.

Descriptive captioning. We are able achieve descriptive captioning by exploiting object features. Our model is able to predict difficult long-tailed entities like "Monsters" and descriptive captions like "Man in red towel", with high accuracy. However, the presence of "Man in black jacket" or "AMnr: with a smile" is undeniably high.

Localising roles in a weakly supervised manner is a very challenging task, it requires to disambiguate the roles and shift the attention to the right object out of a large pool of objects. Since the supervision comes from captions, which are descriptive and may refer to multiple attributes of an object, the attention is divided among many objects and it is difficult to get the most representative object with high probability. Our model is able to ground the roles reasonably well, but leaves a lot of room for improvement.

4.10 Annotations

Sampling frames and creating an annotation task for a video. In a video of 10 seconds consisting of 5 events, we sample frames at 1fps, $\mathcal{F} = \{f_t\}_{t=1}^{T}$ from the entire video V, resulting in 11 frames. Then, from the SRL annotations, we extract the captions for the typically visual roles: *agent, patient, and instrument* from all the 5 events. We retain all the unique captions from the selected ones and use them as ground-truth labels for the video V. For each video we create a separate annotation task on the CVAT tool [1], with video specific labels as shown in Fig 4.8.

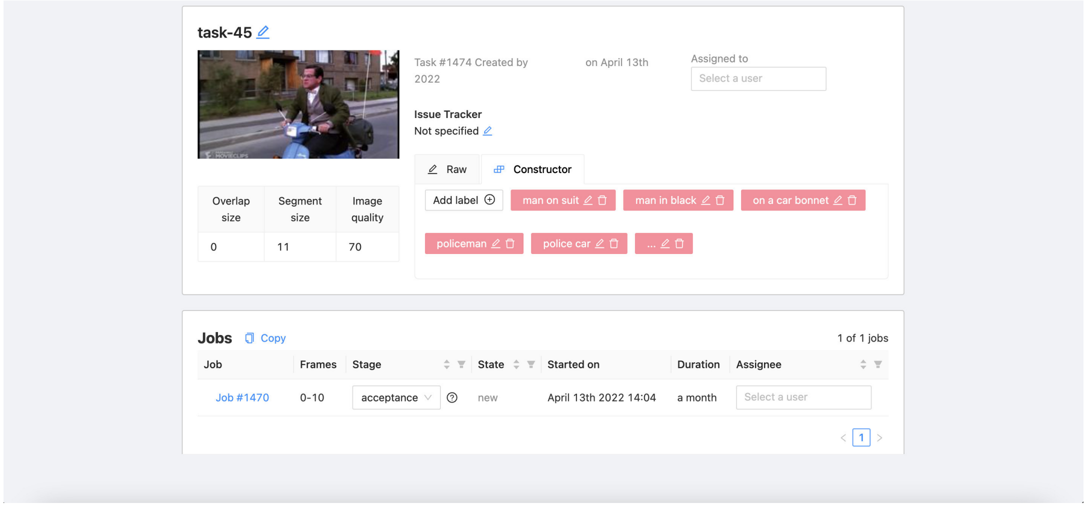
4.10.1 Annotation process

We iterate over every frame in $\mathcal{F}$, and find if any of the label is visually recognised. If it is we select the label, and draw a bounding box around the visual entity as shown in Fig 4.9, 4.10, and 4.11. Some labels might not be visually present in the frames, like *Policeman* is not visible in any of the frames in

Fig. 4.10 or *ground* is not visible in Fig 4.12. Some entities are non-visual like *up* in Fig. 4.13. We do not annotate boxes for such roles.

After the annotations are done, for each event i and role k in a video, we create a dictionary of annotations $\mathcal{G}_{ik}$ with keys as frame number of all the frames that has the role k annotated in it and values as the coordinates of the bounding box corresponding to them. We will share the annotations for further research on our project page.

Compensation. We fairly compensated the annotators for their efforts at almost twice the minimum daily wage.



task-45

Task #1474 Created by on April 13th 2022

Assigned to: Select a user

Issue Tracker: Not specified

Raw | Constructor

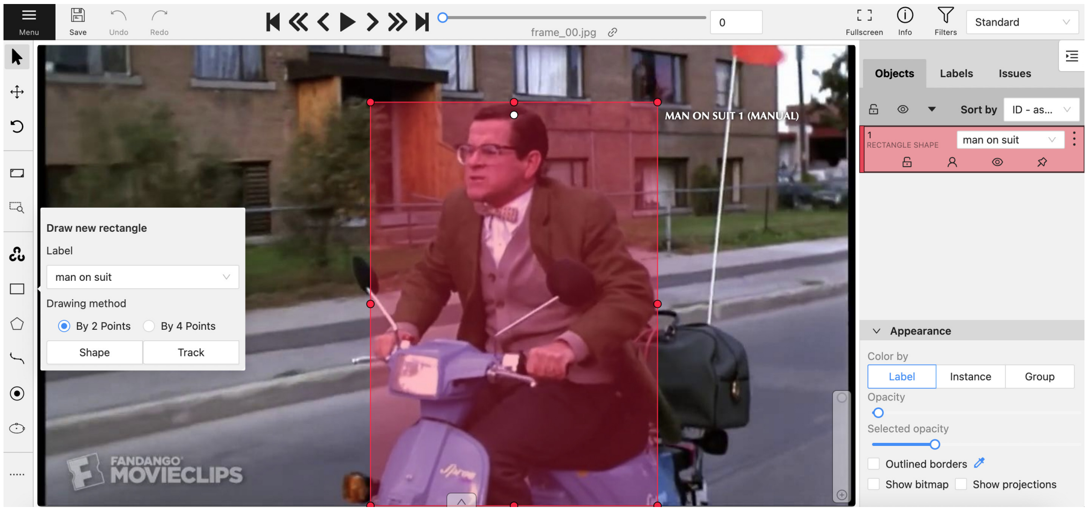
Add label ⊕ | man on suit | man in black | on a car bonnet | policeman | police car | ...

Overlap size	Segment size	Image quality
0	11	70

Jobs Copy 1 of 1 jobs

Job	Frames	Stage	State	Started on	Duration	Assignee
Job #1470	0-10	acceptance	new	April 13th 2022 14:04	a month	Select a user

Figure 4.8 Example annotation task for a video. There are a total of 11 frames subsampled at T=1 second from a 10 second video. Text highlighted in red are the labels.



Draw new rectangle

Label: man on suit

Drawing method: ● By 2 Points ○ By 4 Points

Shape | Track

1 RECTANGLE SHAPE: man on suit

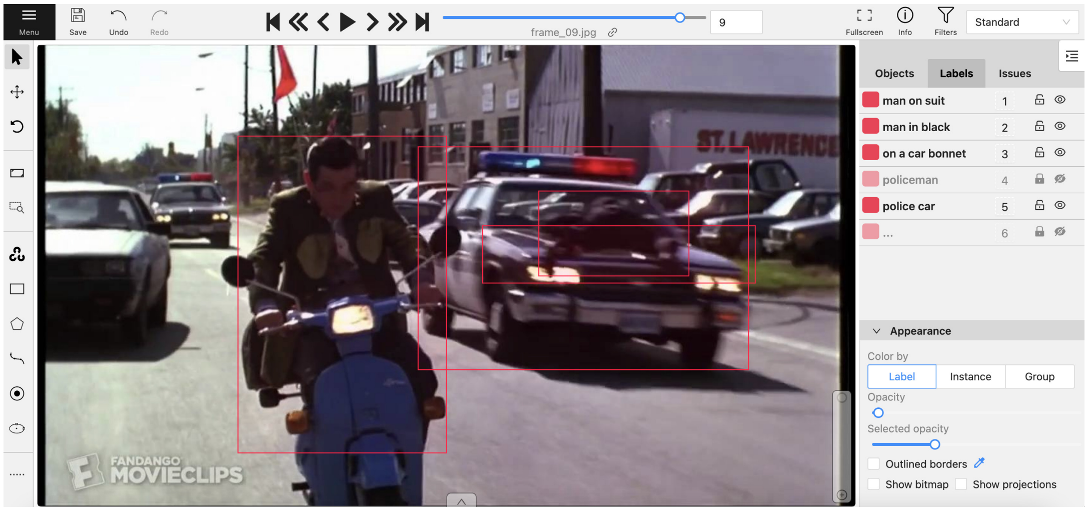
Sort by: ID - as...

Color by: Label | Instance | Group

□ Outlined borders

□ Show bitmap □ Show projections

Figure 4.9 Select a label from the set of labels that can be visually recognised and draw a box around it.

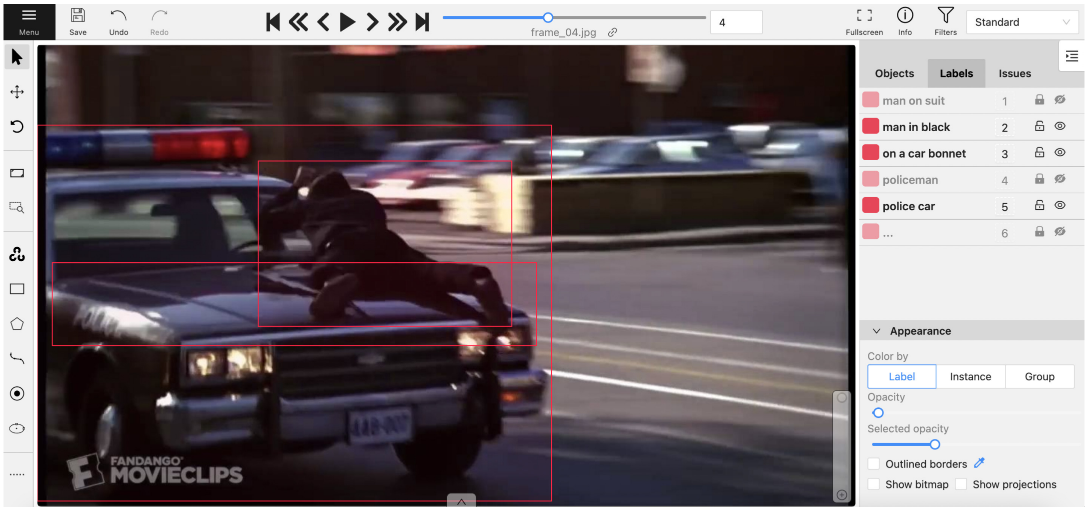


frame_09.jpg 9

man on suit 1
man in black 2
on a car bonnet 3
policeman 4
police car 5
... 6

Color by: Label | Instance | Group
Opacity
Selected opacity
□ Outlined borders
□ Show bitmap □ Show projections

Figure 4.10 Labels *Man on suit, Man in black, on a car bonnet and, Police car* are visible in *frame_09*. Four boxes are drawn around the corresponding four entities

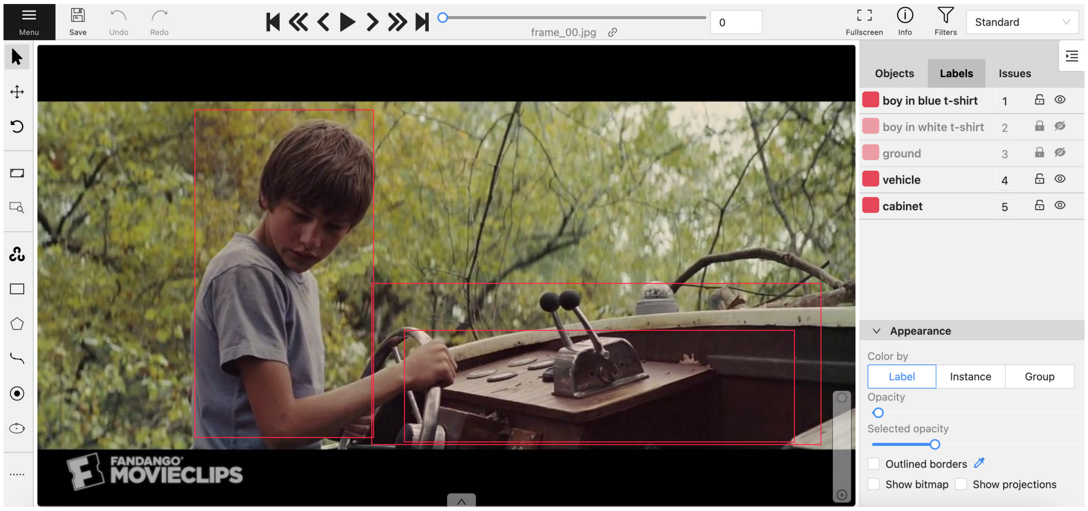


frame_04.jpg 4

man on suit 1
man in black 2
on a car bonnet 3
policeman 4
police car 5
... 6

Color by: Label | Instance | Group
Opacity
Selected opacity
□ Outlined borders
□ Show bitmap □ Show projections

Figure 4.11 Labels *Man in black, on a car bonnet and, Police car* are visible in *frame_04* . Three boxes are drawn around the three corresponding objects.



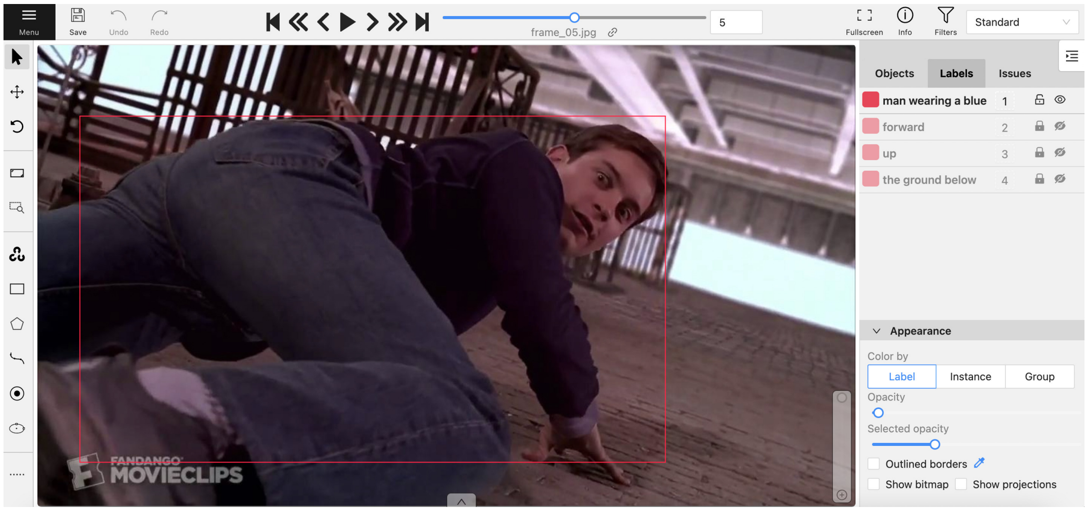
frame_00.jpg

0

Standard

🟥	boy in blue t-shirt	1
🟥	boy in white t-shirt	2
🟥	ground	3
🟥	vehicle	4
🟥	cabinet	5

Figure 4.12 Label *ground* is not visible in *frame_00*, hence it is not annotated



frame_05.jpg

5

Standard

🟥	man wearing a blue	1
🟥	forward	2
🟥	up	3
🟥	the ground below	4

Figure 4.13 Label *up* is a non-visual role, hence it is not annotated.

Chapter 5

Conclusions and Future Works

5.1 Conclusions

In this thesis, we explore the domain of holistic and structured video understanding through a new task: Video Situation Recognition. We introduce Situation recognition by discussing the pioneer works in the domain of image understanding. We show that the unstructured nature of dense understanding tasks like image captioning can be converted into a structured format using the theory of semantic role labeling(SRL). SRL provides a structure to represent any situation using an action verb, all the entities that are taking part in the action (nouns), and the role they play. This gave rise to the first structured and dense image understanding framework: imsitu [45]. Later imsitu was extended to grounded situation recognition [27], which added one more level of complexity, where in addition to predicting the verbs, roles, and nouns, it is now required to ground each noun by drawing a bounding box in the image.

We then discuss Video Situation Recognition [31], framework for holistic video understanding through semantic role labeling. We discussed several challenges of the existing framework mainly (i) The free-form nature and the ambiguity of SRL captions, which makes it difficult to evaluate. And (ii) Separately modeling the three tasks in VidSitu, requiring the Ground Truth of the previous task to model the next. This limits the ability for end-to-end situation recognition. To alleviate the challenges and provide another level of understanding we proposed a new framework- GVSR as a means for holistic video understanding combining situation recognition - recognizing salient actions, and their semantic role-noun pairs with grounding. Grounding helps remove the ambiguity in the captions and allows ananother level of interpretability. Moreover grounding is performed in a weakly supervised setting, without requiring ground truth bounding boxes. We approached this challenging problem by proposing *VideoWhisperer*, which combines a video-object encoder for contextualized embeddings, video contextualized role query for better representing the roles without the need for ground-truth verbs (enables end-to-end situation recognition) and an event-aware cross-attention that helps identify the relevant nouns and ranks them to provide grounding. We achieved state-of-the-art performance on the VidSitu benchmark with large gains and proposed a new benchmark for grounding in a weakly-supervised manner.

5.2 Future Works

One of the most important future direction would be video representation learning, since videos in the VidSitu dataset are extracted primarily from movies. The domain of videos is very different from other general video understanding frameworks. It consists of multiple events with sudden shot changes which is difficult to incorporate in the existing video representation learning models. To account for shot changes, shot detectors can be used to improve the video representations. This will improve the temporal consistency across shots and events, and hence better capture the semantic meaning of the scene.

Another important direction would be towards improving the coreferencing abilities of the model. VidSitu dataset allows for coreferencing multiple entities across multiple events in a video. Current models fail to accurately corefer an entity which results in incoherent results. Identifying and corefer-ncing each entity will allow to accurately track them across all the events and tremendously improve the performance.

My Publications

- **Zeeshan Khan**, C. V. Jawahar and Makarand Tapaswi. 2022. Grounded Video Situation Recognition. In: Advances in Neural Information Processing Systems; 2022.
- **Zeeshan Khan**, Kartheek Akella, Vinay Namboodiri, and C V Jawahar. 2021. More Parameters? No Thanks!. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 96–102, Online. Association for Computational Linguistics. *(Not a part of thesis)*
- Kartheek Akella, Sai Himal Allu, Sridhar Suresh Ragupathi, Aman Singhal, **Zeeshan Khan**, C.V. Jawahar, and Vinay P. Namboodiri. 2020. Exploring Pair-Wise NMT for Indian Languages. In Proceedings of the 17th International Conference on Natural Language Processing (ICON), pages 437–443, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI). *(Not a part of thesis)*

Bibliography

[1] CVAT Annotation Tool. https://github.com/openvinotoolkit/cvat. 19, 31

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. 2018. 1, 18

[3] C. F. Baker, C. J. Fillmore, and J. B. Lowe. The Berkeley FrameNet project. 1998. 5

[4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-End Object Detection with Transformers. 2020. 1

[5] J. Carreira and A. Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. 2017. 10, 11

[6] J. Cho, Y. Yoon, and S. Kwak. Collaborative Transformers for Grounded Situation Recognition. 2022. 5, 7

[7] J. Cho, Y. Yoon, H. Lee, and S. Kwak. Grounded Situation Recognition with Transformers. 2021. 5, 7, 20

[8] C. Feichtenhofer, H. Fan, J. Malik, and K. He. SlowFast Networks for Video Recognition. 2019. 1, 10, 11, 18

[9] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman. Video Action Transformer Network. 2019. 11

[10] S. Gupta and J. Malik. Visual Semantic Role Labeling. *arXiv preprint arXiv:1505.04474*, 2015. 5

[11] Q. Huang, Y. Xiong, A. Rao, J. Wang, and D. Lin. MovieNet: A Holistic Dataset for Movie Understanding. 2020. 1

[12] Z. Khan, C. Jawahar, and M. Tapaswi. Grounded video situation recognition. In *Advances in Neural Information Processing Systems*, 2022. 1

[13] A. Khoreva, A. Rohrbach, and B. Schiele. Video Object Segmentation with Referring Expressions. 2018. 11

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 19

[15] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. 123(1):32–73, 2017. 18

[16] R. Li, M. Tapaswi, R. Liao, J. Jia, R. Urtasun, and S. Fidler. Situation Recognition with Graph Neural Networks. 2017. ix, 5, 6, 7

[17] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004. 18

[18] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen. BMN: Boundary-matching Network for Temporal Action Proposal Generation. 2019. 11

[19] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang. BSN: Boundary Sensitive Network for Temporal Action Proposal Generation. 2018. 11

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal Loss for Dense Object Detection. 2017. 26

[21] J. Lu, D. Batra, D. Parikh, and S. Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. 2019. 1

[22] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[23] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, nov 1995. 5

[24] N. S. Moosavi and M. Strube. Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric. 2016. 18

[25] M. Palmer, D. Gildea, and P. Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, 2005. 1, 12

[26] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. 2019. 18

[27] S. Pratt, M. Yatskar, L. Weihs, A. Farhadi, and A. Kembhavi. Grounded Situation Recognition. 2020. ix, 5, 8, 35

[28] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. 2015. 18

[29] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele. Movie Description. 123:94–120, 2017. 11

[30] A. Sadhu, K. Chen, and R. Nevatia. Video Object Grounding using Semantic Roles in Language Description. 2020. 11

[31] A. Sadhu, T. Gupta, M. Yatskar, R. Nevatia, and A. Kembhavi. Visual Semantic Role Labeling for Video Understanding. 2021. ix, xii, 1, 3, 9, 11, 14, 15, 16, 18, 19, 20, 21, 22, 23, 24, 25, 35

[32] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid. VideoBERT: A Joint Model for Video and Language Representation Learning. 2019. 1

[33] C. Sun, A. Shrivastava, C. Vondrick, K. Murphy, R. Sukthankar, and C. Schmid. Actor-Centric Relation Network. 2018. 11

[34] M. Tapaswi, V. Kumar, and I. Laptev. Long term spatio-temporal modeling for action detection. 210, 2021. 11

[35] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, , and S. Fidler. MovieQA: Understanding Stories in Movies through Question-Answering. 2016. 11

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is All you Need. 2017. 5, 12, 14, 16

[37] R. Vedantam, C. L. Zitnick, and D. Parikh. CIDEr: Consensus-based image description evaluation. 2015. 18

[38] P. Vicol, M. Tapaswi, L. Castrejon, and S. Fidler. Moviegraphs: Towards understanding human-centric situations from videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[39] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool. Temporal segment networks: Towards good practices for deep action recognition. 2016. 11

[40] M. Wei, L. Chen, W. Ji, X. Yue, and T.-S. Chua. Rethinking the Two-Stage Framework for Grounded Situation Recognition. 2022. 5, 7, 20

[41] C.-Y. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krahenbuhl, and R. Girshick. Long-term Feature Banks for Detailed Video Understanding. 2019. 11

[42] F. Xiao, K. Kundu, J. Tighe, and D. Modolo. Hierarchical Self-supervised Representation Learning for Movie Understanding. 2022. x, 3, 10, 11, 14, 15, 16, 19, 21, 22

[43] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid. Learning to answer visual questions from web videos. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2022. 1

[44] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid. TubeDETR: Spatio-Temporal Video Grounding with Transformers. 2022. 11

[45] M. Yatskar, L. Zettlemoyer, and A. Farhadi. Situation Recognition: Visual Semantic Role Labeling for Image Understanding. 2016. ix, 1, 5, 6, 8, 35

[46] S. Young, D. Gandhi, S. Tulsiani, A. K. Gupta, P. Abbeel, and L. Pinto. Visual imitation made easy. In *CoRL*, 2020. 1

[47] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao. ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering. 2019. 11

[48] Z. Zhang, Z. Zhao, Y. Zhao, Q. Wang, H. Liu, and L. Gao. Where Does It Exist: Spatio-Temporal Video Grounding for Multi-Form Sentences. 2020. 11

[49] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong. End-to-end dense video captioning with masked transformer. 2018. 11