Grounded Video Situation Recognition

Zeeshan Khan **C. V. Jawahar** **Makarand Tapaswi**
CVIT, IIIT Hyderabad
https://zeeshank95.github.io/grvidsitu

Abstract

Dense video understanding requires answering several questions such as *who is doing what to whom, with what, how, why, and where*. Recently, Video Situation Recognition (VidSitu) 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 task poses several challenges in identifying, disambiguating, and co-referencing entities across multiple verb-role pairs, but also faces some challenges of evaluation. 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, and 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.

1 Introduction

At the end of *The Dark Knight*, we see a short intense sequencethat 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 [19, 29]. We are interested in assessing how some of these advances can be combined for a holistic understanding of video clips.

A recent and audacious step towards this goal is the work by Sadhu *et al.* [28]. 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 [40]), VidSitu is annotated using Semantic Role Labelling (SRL) [22]. 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-*

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

ArgM(manner), *etc*., and each role is annotated with a free form text caption, *e.g*. *agent: Blonde Woman*, as illustrated in Fig. 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 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). 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.

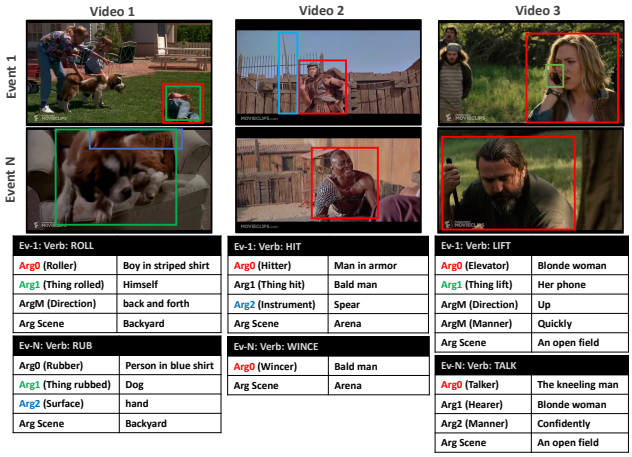

Figure 1: **Overview of GVSR:** Given a video consisting of multiple events, GVSR requires recognising the action verbs, their corresponding roles, and localising them in the spatio-temporal domain. This is a challenging task as it requires to disambiguate between several roles that the same entity may 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).

Joint structured prediction. Previous works [28, 38] 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 [28].

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.

Contributions summary. (i) We present a new framework that combines grounding with SRL for end-to-end Grounded Video Situation Recognition (GVSR). We will release the grounding annotations and also include them in the evaluation benchmark. (ii) We design a new three-stage transformer architecture for joint verb prediction, semantic-role labelling through caption generation, and weakly-supervised grounding of visual entities. (iii) 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. (iv) At the encoder, we 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 [28]. (v) Finally, we present extensive ablation experiments to analyze our model. Our model achieves the state-of-the-art results on the VidSitu benchmark.

2 Related Work

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 [40] proposed situation recognition via semantic role labelling by leveraging linguistic frameworks, FrameNet [3] and WordNet [20] to formalize situations in the form of verb-role-noun triplets. Recently, grounding has been incorporated with image situation recognition [24] 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 [40], LSTMs [24] and Graph neural networks [14] to model the global dependencies among verb and roles. Recently various Transformer [33] based methods have been proposed that claim large performance improvements [6, 7, 36].

Video Situation Recognition. Recently, imSitu was extended to videos as VidSitu [28], 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. Sadhu *et al*. [28] propose 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. A concurrent work to this submission, [38], proposes to improve upon the video features by pretraining the low-level video backbone using contrastive learning objectives, and pretrain the high-level video contextualiser using event mask prediction tasks resulting in large performance improvements on SRL. Our goals are different from the above two works, 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 previous works.

Video Understanding. Video understanding is a broad area of research, dominantly involving tasks like action recognition [5, 8, 9, 30, 35, 37], localisation [16, 17], object grounding [27, 39], question answering [32, 41], video captioning [26], and spatio-temporal detection [9, 31]. These tasks involve visual temporal understanding in a sparse uni-dimensional way. 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 [42, 39] or referring expressions based segmentation [11] 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.

3 VideoWhisperer for Grounded Video Situation Recognition

We now present the details of our three stage Transformer model, *VideoWhisperer*. A visual overview is presented in Fig. 2. For brevity, we request the reader to refer to [33] for now popular details of self- and cross-attention layers used in Transformer encoders and decoders.

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 [22]) 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 sub-sampled keyframes of the video. We evaluate this against ground-truth annotations done at a keyframe level.

3.1 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

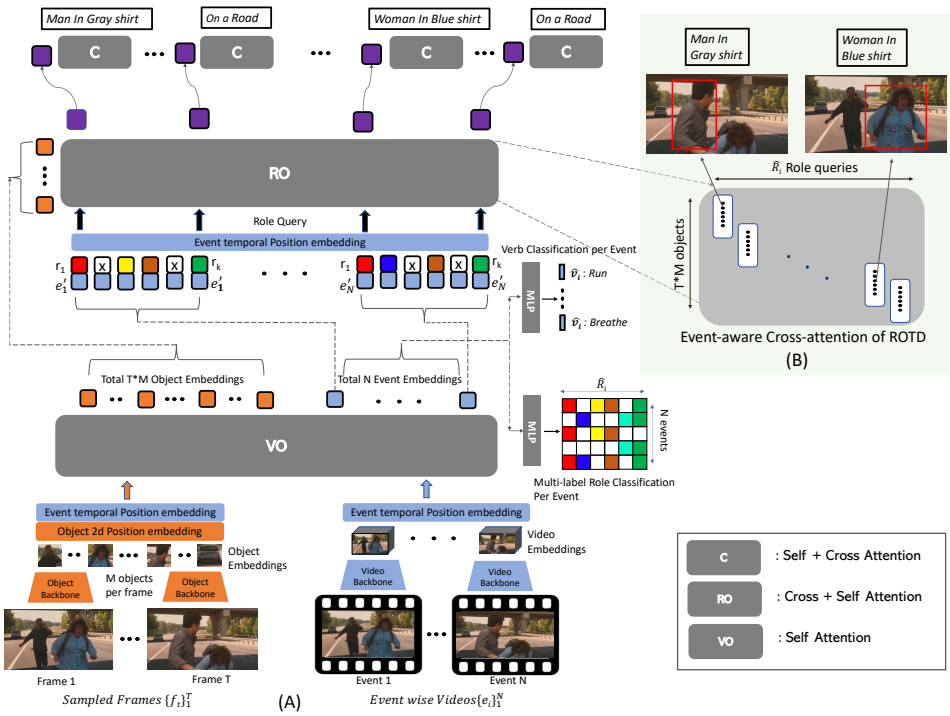

Figure 2: **VideoWhisperer**: We present a new 3-stage Transformer for GVSR. Stage-1 learns the contextualised object and event embeddings through a video-object Transformer encoder (VO), that is used to predict the verb-role pairs for each event. Stage-2 models all the predicted roles by creating role 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.

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 (along with timestamp) and corresponding features are

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

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

$$\mathcal{F}_i = \{f_t | e_i^{\text{start}} \leq t \leq e_i^{\text{end}}\}\,. \tag{2}$$

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 $\mathcal{B}_i$.

Video-Object Transformer Encoder (VO). Since the object and video embeddings come from different spaces, we align and contextualise them with a Transformer encoder [33]. Event-level position embeddings PE_i 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} 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:

$$[\ldots, \mathbf{o}'_{mt}, \ldots, \mathbf{e}'_i, \ldots] = \text{Transformer}_{\text{VO}}\left([\ldots, \mathbf{x}_{mt}^o + \text{PE}_i + \text{PE}_{mt}, \ldots, \mathbf{x}_i^e + \text{PE}_i, \ldots]\right)\,. \tag{3}$$

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,

$$\hat{v}_i = \text{MLP}_e(\mathbf{e}'_i)\,. \tag{4}$$

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 [28, 38] 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

$$\hat{\mathcal{R}}_i = \{r | \sigma(\text{MLP}_r(\mathbf{e}'_i)) > \theta_{\text{role}}\}\,, \tag{5}$$

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 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.

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

$$\mathbf{q}_{ik} = \mathbf{r}_k + \mathbf{e}'_i + \text{PE}_i\,. \tag{6}$$

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

Role-Object Transformer Decoder (RO). It is hard to achieve rich captions while using features learned for action recognition. Different from prior works [28, 38], 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

$$\mathbf{q}'_{ik} = W_Q \mathbf{q}_{ik}, \quad \mathbf{k}'_{mt} = W_K \mathbf{o}'_{mt}, \quad \text{and} \quad \mathbf{v}'_{mt} = W_V \mathbf{o}'_{mt}\,. \tag{7}$$

Here, $W_{[Q|K|V]}$ are learnable linear layers. Next, we apply a mask while computing cross-attention to obtain contextualised role embeddings as

$$\mathbf{r}'_{ik} = \sum_{mt} \alpha_{mt} \mathbf{v}'_{mt}, \quad \text{where} \quad \alpha_{mt} = \text{softmax}_{mt}(\langle \mathbf{q}'_{ik}, \mathbf{k}'_{mt} \rangle \cdot \mathbb{1}(f_t \in \mathcal{F}_i))\,, \tag{8}$$

where $\langle \cdot, \cdot \rangle$ is an inner product and $\mathbb{1}(\cdot)$ 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.

$$[\ldots, \mathbf{z}_{ik}, \ldots] = \text{Transformer}_{\text{RO}}([\ldots, \mathbf{q}_{ik}, \ldots; \ldots, \mathbf{o}'_{mt}, \ldots])\,. \tag{9}$$

Captioning Transformer Decoder (C). The final stage of our model is a caption generation module. Specifically, we use another Transformer decoder [33] whose input context is the output role embedding $\mathbf{z}_{ik}$ from the previous stage and unroll predictions in an autoregressive manner.

$$\hat{C}_{ik} = \text{Transformer}_{\text{C}}(\mathbf{z}_{ik}) \,. \tag{10}$$

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 [28, 38] 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 [28] requires 0.94 seconds.

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 $\text{Transformer}_{\text{RO}}$, and identify the highest scoring box and the corresponding timestep as

$$\hat{b}^*_m, \hat{b}^*_t = \arg\max_{m,t} \alpha_{mt} \,. \tag{11}$$

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

$$L_i^{\text{verb}} = CE(\hat{v}_i, v_i) \quad \text{and} \quad L_i^{\text{role}} = \sum_{r \in \mathcal{R}_i} BCE(r \in \hat{\mathcal{R}}_i, r \in \mathcal{R}_i) \,. \tag{12}$$

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,

$$L_{ik}^{\text{caption}} = \sum_w CE(\hat{C}^w_{ik}, C^w_{ik}) \,, \tag{13}$$

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

$$\mathcal{L} = \sum_i L_i^{\text{verb}} + \sum_i L_i^{\text{role}} + \sum_{ik} L_{ik}^{\text{caption}} \,. \tag{14}$$

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.

4 Experiments

We evaluate our model in two main settings. (i) This setup mimics VidSitu [28], 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) End-to-end GVSR, where all the three tasks are modelled together without using ground truth verb-roles.

Dataset. We evaluate our model on the VidSitu [28] 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 [28]). 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 [34],

CIDEr-Vb: Macro-averaged across verbs, CIDEr-Arg: Macro-averaged across roles, LEA [21], and ROUGE-L [15]. For more details on the metrics pleas refer to [28].

Implementation details. We implement our model in Pytorch [23]. We extract event (video) features from a pretrained SlowFast model [8] for video representation (provided by [28]). For object features, we use a FasterRCNN model [25] provided by BUTD [2] pretrained on the Visual Genome dataset [13]. 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 [28] which uses byte pair encoding. We have 3 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. 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 [12] 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.

4.1 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 E 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. 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,

$$\text{IoU@}\theta = \frac{1}{|\mathcal{R}_i|} \sum_{k=1}^{|\mathcal{R}_i|} \mathbb{1}[\hat{b}_t^* \in \mathcal{G}_{ik}] \cdot \mathbb{1}[\text{IoU}(\hat{b}_m^*, \mathcal{G}_{ik}[t]) > \theta] \,. \tag{15}$$

4.2 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 [38] we observe 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).

Table 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 $\pm$ 0.53	-	-
2	RO + C	Role + GT-verb	**70.48** $\pm$ 1.09	0.14 $\pm$ 0.01	0.06 $\pm$ 0.003
3	VOR + C	Role + Event	67.4 $\pm$ 0.81	0.22 $\pm$ 0.00	0.09 $\pm$ 0.002
4	V + RO + C	Role + Event	69.15 $\pm$ 0.62	0.23 $\pm$ 0.03	0.09 $\pm$ 0.01
5	VO + RO + C	Role + Event	68.54 $\pm$ 0.48	**0.29** $\pm$ 0.013	**0.12** $\pm$ 0.01

Architecture design. We present SRL and grounding results in Table 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 $\mathbf{x}^o_{mt}$.

Role query embeddings design. Prior works in situation recognition [7, 28, 36] 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 2 (architecture: VO + RO + C) row 2 vs. 3 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.

Table 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

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 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 performance.

Table 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

4.3 SRL SoTA comparison

In Table 4, we compare our results against VidSitu [28] and a concurrent work that uses far better features [38]. We reproduce results for VidSitu [28] by teacher-forcing the ground-truth role pairs to make a fair comparison while results for work [38] are as reported in their paper. Nevertheless, we achieve state-of-the-art performance with a 22 points gain in CIDEr score over [28] and a 8 point gain over [38], while using features from [28]. Moreover, our model allows grounding, something not afforded by the previous approaches.

Table 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 [28]	46.01	56.37	43.58	43.04	**50.89**	-	-
Slow-D+TxE+TxD [38]	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 GVSR: Joint Prediction of Video Situations

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

Verb prediction is an extremely challenging problem due to the long-tail nature of the dataset. In Table 5, 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 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 Appendix C, 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.

Table 5: 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 [28]	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**

Understanding role prediction. The verb-role prediction accuracy is crucial for GVSR, since the SRL task is modelled on role-queries. In Table 6 we analyse role prediction in various settings to understand its effect on SRL. Previous work [28] 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 [28] 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 [28], 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 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. Appendix B.3 discusses further challenges of multi-label and imbalance in predicting roles.

Table 6: 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 [28]	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

GVSR. We evaluate our end-to-end model for grounded video situation recognition. In order to enable end-to-end GVSR in [28], we use it in the "Pred, Pred" setting discussed above, that allows verb, role, and SRL predictions. Table 7 shows that our model improves SRL performance over Vidsitu [28] by a margin of 22% on CIDEr score. In addition to that, our model also enables Grounded SRL, not achievable in VidSitu [28].

Table 7: GVSR: Results for end-to-end situation recognition. Our model architecture is VO + RO + C.

Model	Prediction			Verb	CIDEr	IoU	
	Verb	Role	SRL	Acc@1		0.3	0.5
VidSitu [28]	✓	✓	✓	46.79	30.33	-	-
VideoWhisperer	✓	✓	✓	44.06	52.30	0.13	0.05
	✓	GT	✓	45.06	68.54	0.29	0.12

4.5 Qualitative Results and Limitations

We visualize the predictions of VideoWhisperer (Pred-GT) in Fig. 3 for one video of 10 seconds[1] and see that it performs reasonably well given the complexity of the task. VideoWhisperer correctly

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

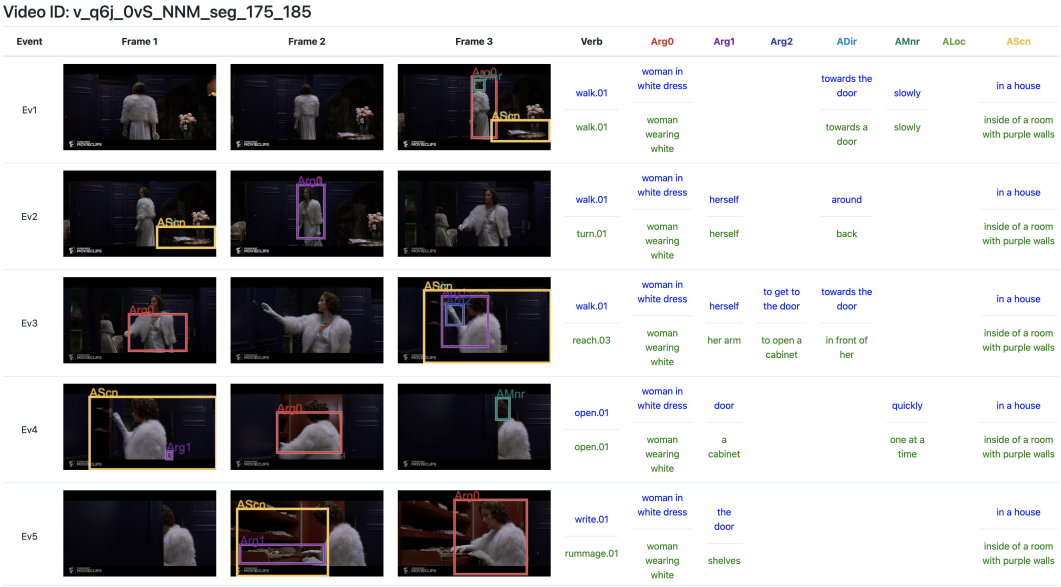

Figure 3: We show the results for a 10s clip that can be viewed here: https://youtu.be/q6j_0vS_NNM?t=175. The video is broken down to 5 events indicated by the row labels Ev1 to Ev5. At a 1fps sampling rate, we obtain boxes from 3 frames for each event (with Frame3 of event $i-1$ being the same as Frame1 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 6 (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).

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 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. 3, 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.

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 (details in Appendix D). Nevertheless, we hope that this work inspires the community to couple videos and their descriptions.

5 Conclusion

We proposed GVSR as a means for holistic video understanding combining situation recognition - recognizing salient actions, and their semantic role-noun pairs with grounding. We approached this challenging problem by proposing VideoWhisperer, that combines a video-object encoder for contextualised embeddings, video contextualised role query for better representing the roles without the need for ground-truth verbs 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 also enabled grounding for roles in a weakly-supervised manner.

References

[1] CVAT Annotation Tool. https://github.com/openvinotoolkit/cvat. 7, 16

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 7

[3] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet project. In *Association of Computational Linguistics (ACL)*, 1998. 3

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[5] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[6] Junhyeong Cho, Youngseok Yoon, and Suha Kwak. Collaborative Transformers for Grounded Situation Recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[7] Junhyeong Cho, Youngseok Yoon, Hyeonjun Lee, and Suha Kwak. Grounded Situation Recognition with Transformers. 2021. 3, 8

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 3, 7

[9] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video Action Transformer Network. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[10] Saurabh Gupta and Jitendra Malik. Visual Semantic Role Labeling. *arXiv preprint arXiv:1505.04474*, 2015. 3

[11] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video Object Segmentation with Referring Expressions. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73, 2017. 7

[14] Ruiyu Li, Makarand Tapaswi, Renjie Liao, Jiaya Jia, Raquel Urtasun, and Sanja Fidler. Situation Recognition with Graph Neural Networks. In *International Conference on Computer Vision (ICCV)*, 2017. 3

[15] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004. 7

[16] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. BMN: Boundary-matching Network for Temporal Action Proposal Generation. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[17] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. BSN: Boundary Sensitive Network for Temporal Action Proposal Generation. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. In *International Conference on Computer Vision (ICCV)*, 2017. 15

[19] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1

[20] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, nov 1995. 3

[21] Nafise Sadat Moosavi and Michael Strube. Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric. In *Association of Computational Linguistics (ACL)*, 2016. 7

[22] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, 2005. 1, 3

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 7

[24] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded Situation Recognition. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 7

[26] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie Description. *International Journal of Computer Vision (IJCV)*, 123:94–120, 2017. 3

[27] Arka Sadhu, Kan Chen, and Ram Nevatia. Video Object Grounding using Semantic Roles in Language Description. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[28] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual Semantic Role Labeling for Video Understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3, 5, 6, 7, 8, 9, 13, 14

[29] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A Joint Model for Video and Language Representation Learning. In *International Conference on Computer Vision (ICCV)*, 2019. 1

[30] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-Centric Relation Network. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[31] Makarand Tapaswi, Vijay Kumar, and Ivan Laptev. Long term spatio-temporal modeling for action detection. *Computer Vision and Image Understanding (CVIU)*, 210, 2021. 3

[32] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, , and Sanja Fidler. MovieQA: Understanding Stories in Movies through Question-Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3, 4, 6

[34] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 6

[35] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[36] Meng Wei, Long Chen, Wei Ji, Xiaoyu Yue, and Tat-Seng Chua. Rethinking the Two-Stage Framework for Grounded Situation Recognition. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 3, 8

[37] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term Feature Banks for Detailed Video Understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[38] Fanyi Xiao, Kaustav Kundu, Joseph Tighe, and Davide Modolo. Hierarchical Self-supervised Representation Learning for Movie Understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 5, 6, 7, 8

[39] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. TubeDETR: Spatio-Temporal Video Grounding with Transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[40] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation Recognition: Visual Semantic Role Labeling for Image Understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 3

[41] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 3

[42] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where Does It Exist: Spatio-Temporal Video Grounding for Multi-Form Sentences. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] **Briefly in the main paper, details in the appendix.**
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] **Please see the project page**

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] **In Section 4, we described all the training details, we will also provide the model code.**

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] **Most experiments in the ablation studies and SoTA comparison are repeated 3 to 10 times, standard deviation is reported.**

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] **Section 4, implementation details.**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] **We use the VidSitu [28] dataset and augment it with grounding annotations for the proposed GVSR task.**

(b) Did you mention the license of the assets? [Yes] **VidSitu [28] is publicly available and licensed under MIT license.**

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] **We will add the grounding annotations over the existing VidSitu dataset on the project page.**

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] **The VidSitu dataset is publicly available for research purposes. Additional data collection did not require consent.**

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] **The VidSitu data does not contain PII information. It may have content that could be deemed to be offensive due to the nature of movies. Our new annotations do not add to this and are only used for evaluation.**

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] **We add it in the Appendix.**

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No] **Annotations were done in-house in the lab. No additional data was collected.**

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] **We add it in the Appendix.**

Grounded Video Situation Recognition - Appendix

We start with a brief mention of additional qualitative results from our model in Appendix A. Following this, we present quantitative results including an extended table that shows SRL performance on all VidSitu [28] metrics, and role prediction performance of our model (Appendix B). We also present challenges of predicting roles (which appear rarely) in Appendix B.3 and long-tail related challenges of predicting verbs in Appendix C. We end this document by talking a bit about the limitations in Appendix D and the annotation process to obtain boxes on the validation set (Appendix E).

A More qualitative results

GVSR is a challenging problem that requires to correctly identify the action, disambiguate the roles taking part in it, localise the roles, and generate descriptive captions. Moreover, the videos are curated from complicated movie scenes with fast motion, shot changes, and diverse scenes. For better visualisations we create an HTML file GVSR.html, with predictions on 10 videos. There are a total of 5 events in each video. As described in the method section, we sample 11 frames from the entire video, divided between 5 events, each with 3 frames with 1 frame sharing the event boundary. Fig. 3 of the main paper illustrates an example of this visualization.

B Additional quantitative results and metrics

B.1 SRL Evaluation on the Test Set

We evaluate our model on the test set from the evaluation servers of VidSitu [28]. A constant improvement over [28] can be seen in Table 8. The trend is similar when compared with Table 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 [28] to establish this as part of the benchmark.

Table 8: Results of SRL with GT verb and role pairs on the test dataset. VidSitu's [28] 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 [28]	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	-	-

B.2 GVSR, all metrics

In Table 9 we show the results of end-to-end GVSR on all the metrics. We can see a clear improvement over the "pred, pred" VidSitu [28] on all the metrics for SRL. Due to lack of space, we showed only the primary metrics in Table 7.

Table 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 [28]	✓	✓	✓	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	68.23	74.15	61.79	45.58	48.22	0.27	0.15

B.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 10 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.

Table 10: 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

C 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 re-weighting applies weights corresponding to the inverse verb frequency to the cross-entropy loss; (ii) Focal loss is applied as described in [18] (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 11.

Table 11: 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 3 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.

D Limitations

Major challenges in GVSR include: (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. As described in Sec. 4.4, 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.

E 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.

E.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 5, 6, and 7. Some labels might not be visually present in the frames, like *Policeman* is not visible in any of the frames in Fig. 6 or *ground* is not visible in Fig 8. Some entities are non-visual like *up* in Fig. 9. 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.

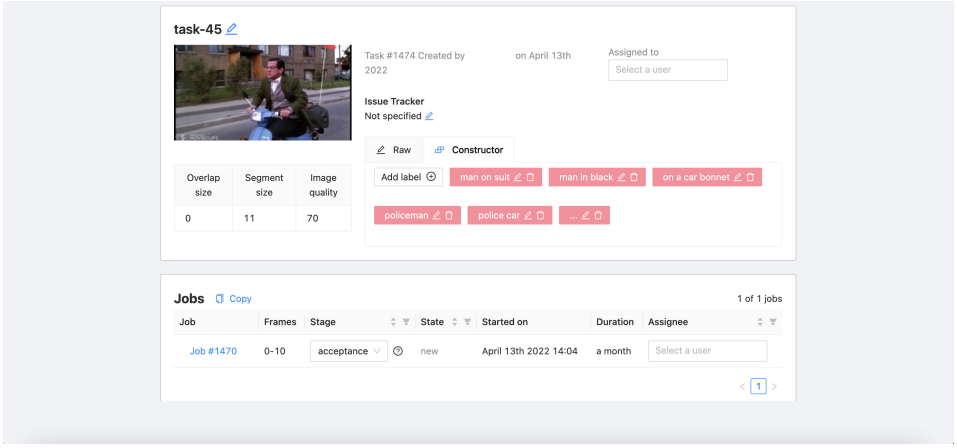

Figure 4: Example annotation task for a video. There are a total of 11 frames subsampled at T=1 second from a 10 second video. Text highlighted in red are the labels.

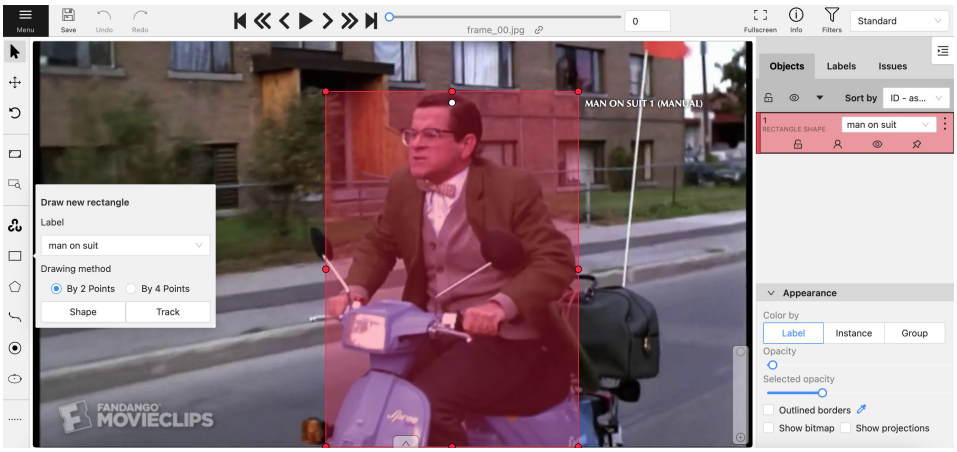

Figure 5: Select a label from the set of labels that can be visually recognised and draw a box around it.

Figure 6: 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

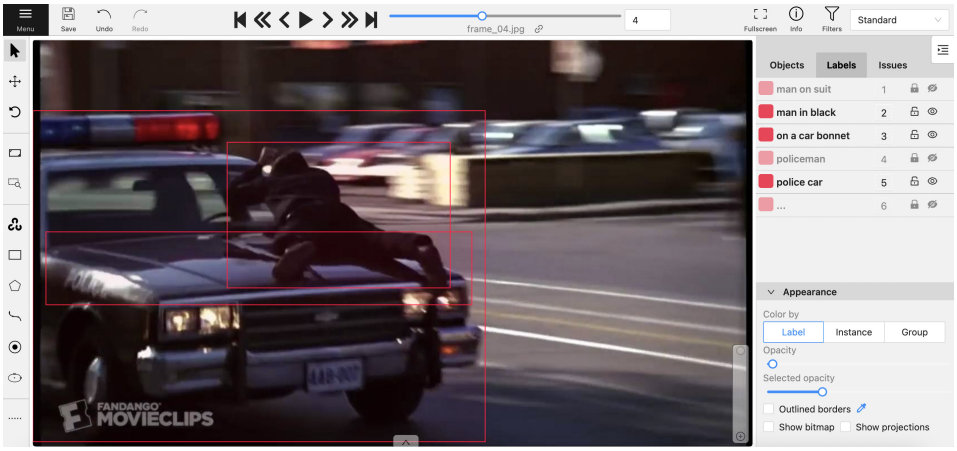

Figure 7: Labels *Man in black, on a car bonnet and, Police car* are visible in *frame_04* . Three boxes are drawn around the three corresponding objects.

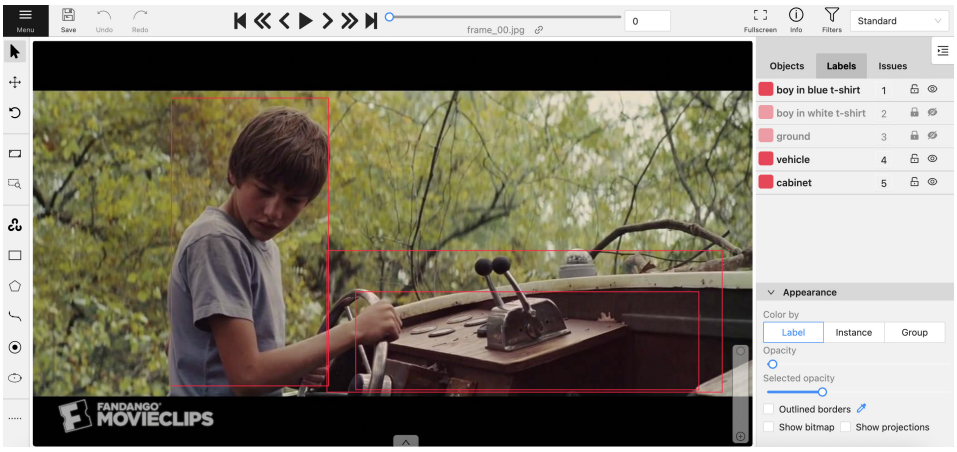

Figure 8: Label *ground* is not visible in *frame_00*, hence it is not annotated

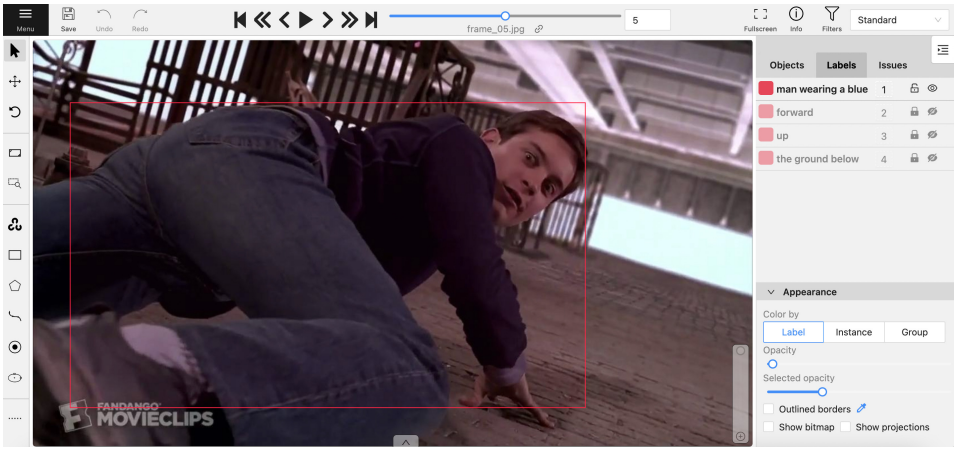

Figure 9: Label *up* is a non-visual role, hence it is not annotated.