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	68.54	77.48	61.55	45.70	47.54	0.29	0.12

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.

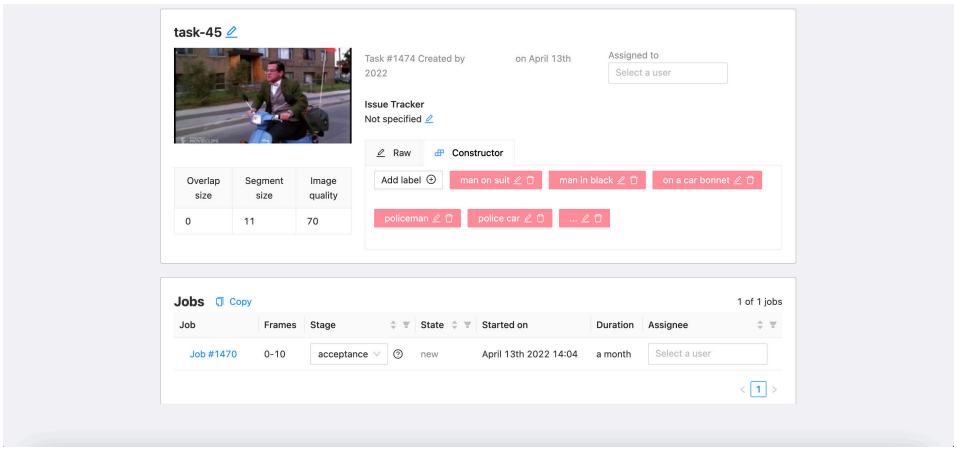

Figure 4: Example annotation task for a video. There are a total of 11 frames subsampled at T=1 second from a 10 second video. Text highlighted in red are the labels.

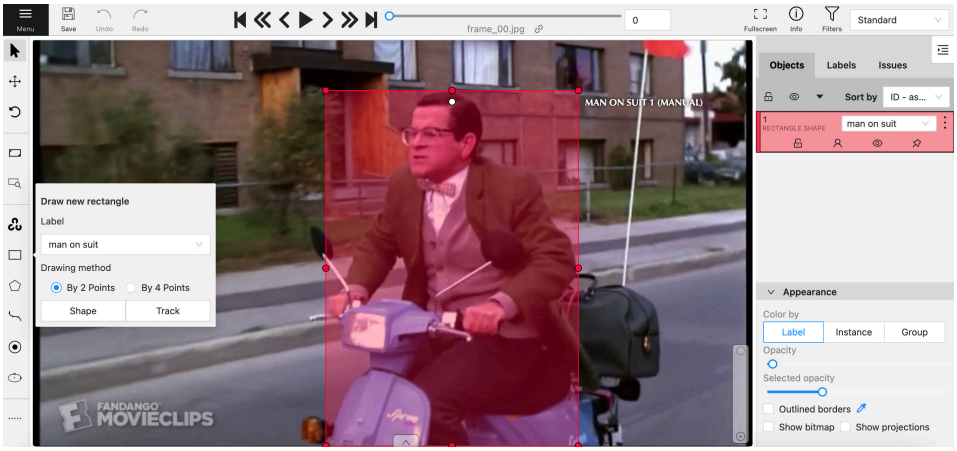

Figure 5: Select a label from the set of labels that can be visually recognised and draw a box around it.

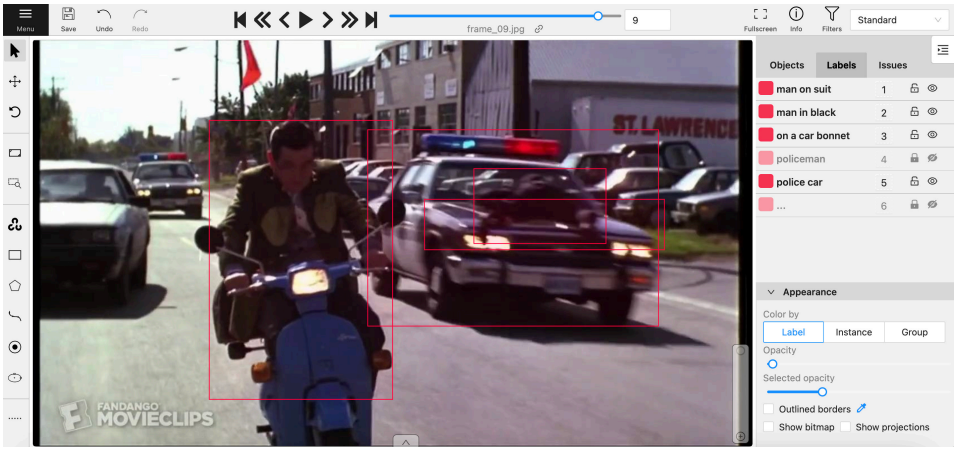

Figure 6: Labels *Man on suit, Man in black, on a car bonnet and, Police car* are visible in *frame_09*. Four boxes are drawn around the corresponding four entities

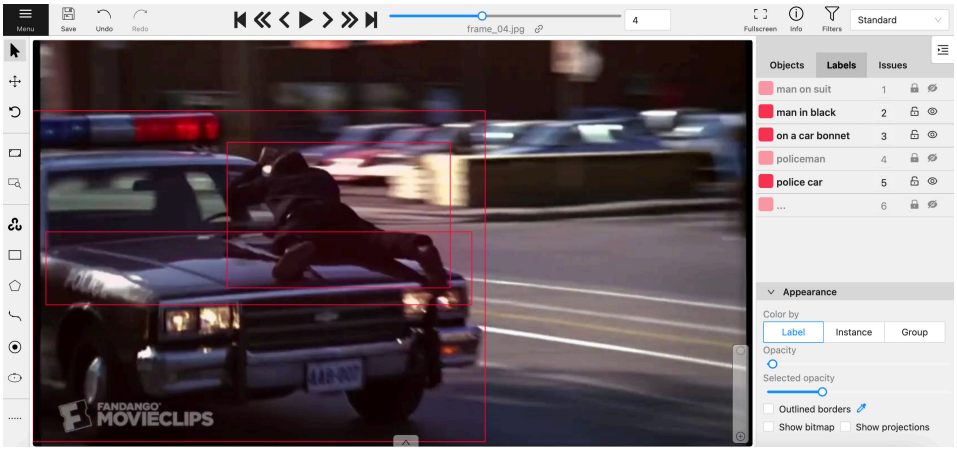

Figure 7: Labels *Man in black, on a car bonnet and, Police car* are visible in *frame_04* . Three boxes are drawn around the three corresponding objects.

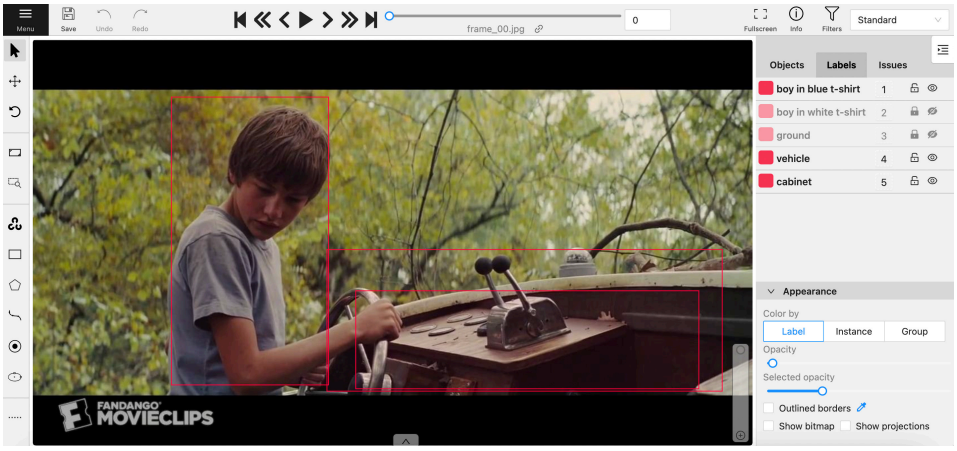

Figure 8: Label *ground* is not visible in *frame_00*, hence it is not annotated

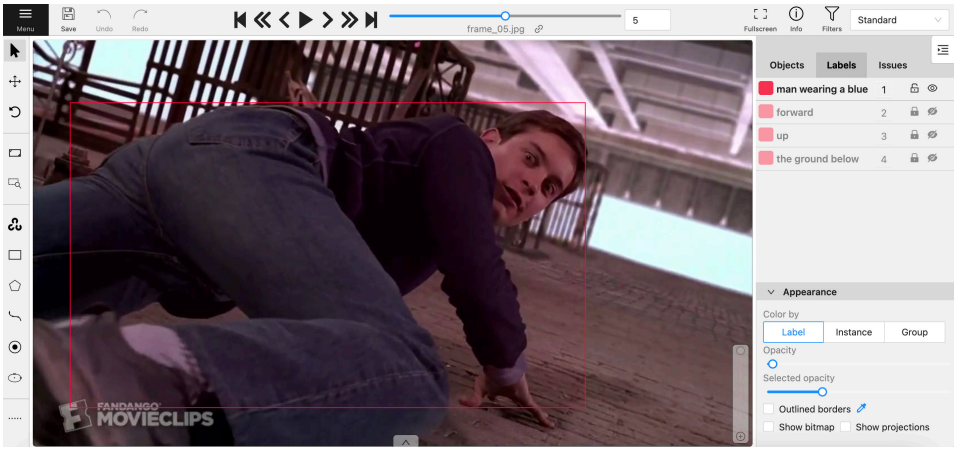

Figure 9: Label *up* is a non-visual role, hence it is not annotated.