Effectively Leveraging CLIP for Generating Situational Summaries of Images and Videos*

Dhruv Verma[1], Debaditya Roy[2], and Basura Fernando[1,2]

[1]Centre for Frontier AI Research Singapore, Agency for Science, Technology and Research Singapore

[2]Institute of High-Performance Computing, Agency for Science, Technology and Research Singapore

Abstract

Situation recognition refers to the ability of an agent to identify and understand various situations or contexts based on available information and sensory inputs. It involves the cognitive process of interpreting data from the environment to determine what is happening, what factors are involved, and what actions caused those situations. This interpretation of situations is formulated as a semantic role labeling problem in computer vision-based situation recognition. Situations depicted in images and videos hold pivotal information, essential for various applications like image and video captioning, multimedia retrieval, autonomous systems and event monitoring. However, existing methods often struggle with ambiguity and lack of context in generating meaningful and accurate predictions. Leveraging multimodal models such as CLIP, we propose ClipSitu, which sidesteps the need for full fine-tuning and achieves state-of-the-art results in situation recognition and localization tasks. ClipSitu harnesses CLIP-based image, verb, and role embeddings to predict nouns fulfilling all the roles associated with a verb, providing a comprehensive understanding of depicted scenarios. Through a cross-attention Transformer, ClipSitu XTF enhances the connection between semantic role queries and visual token representations, leading to superior performance in situation recognition. We also propose a verb-wise role prediction model with near-perfect accuracy to create an end-to-end framework for producing situational summaries for out-of-domain images. We show that situational summaries empower our ClipSitu models to produce structured descriptions with reduced ambiguity compared to generic captions. Finally, we extend ClipSitu to video situation recognition to showcase its versatility and produce comparable performance to state-of-the-art methods. In summary, ClipSitu offers a robust solution to the challenge of semantic role labeling providing a way for structured understanding of visual media. ClipSitu advances the state-of-the-art in situation recognition, paving the way for a more nuanced and contextually relevant understanding of visual content that potentially could derive meaningful insights about the environment that agents observe. Code is available at `https://github.com/LUNAProject22/CLIPSitu-video`.

1 Introduction

Accurately understanding the context and activities shown in visual media is essential for a wide array of applications. Vision based situation recognition defined in [1] stands at the intersection of computer vision and natural language processing and aims to identify and understand the activities, interactions, and scenarios from images and videos (see examples in Fig. 1). Situations as a representation of actions and their effects have been proposed way back by [2, 3] in *situational calculus*. They use situations to model the world at different points in time, actions to change these situations, and a set of axioms to describe how these actions transform one situation into another [4]. In event calculus proposed by [5], a situa-

*This paper is accepted by the International Journal of Computer Vision on Mar 15, 2025.

tion represents a snapshot of the world at a particular point in time. It captures the state of the world, including the events that have occurred up to that point and their effects. By capturing the dependencies and interactions among various entities (such as humans, objects, and scenes), situational description offers a rich depiction of the context in which actions unfold. These structured situational descriptions can contribute to many applications shown in Fig. 1 such as *robot navigation* in a cluttered environment utilizing structured scene representations for precise localization and detection of objects, *autonomous vehicle* interpreting traffic scenes using structural description for safe navigation and decision making, *video surveillance* with structured scene annotations for threat detection and response, and *virtual assistant* interacting with users based on structured understanding of scenes.

Situational descriptions produce *action centric representation* that describe the main action and various interactions between humans and objects in that action which can be evaluated for correctness precisely. Furthermore, situational descriptions are *richer representation* enables more nuanced understanding of scenes, including the relationships between objects, spatial arrangements, and temporal dynamics [11]. Additionally, the structured nature of situational descriptions lends itself well to reasoning and decision-making tasks in assistive systems. By explicitly representing objects, actions, and relationships in a structured format, these descriptions enable more sophisticated reasoning about scenes and support complex decision-making processes [12, 13, 14]. Similarly, structured situational descriptions are easy to integrate with external knowledge bases, allowing AI systems to leverage additional information about objects, events, and concepts [15]. This integration enhances the understanding and interpretation of visual scenes, particularly in domains where external knowledge is crucial, such as in medical imaging or scientific research. Finally, structured situational descriptions enable multi-modal reasoning by providing a unified framework for integrating visual information with other modalities, such as text or knowledge bases. This capability is essential for tasks that require understanding across multiple modalities, such as visual question answering or image-text retrieval [16].

With the need for situational description established, we now describe **situation recognition** as per [1]. Situation recognition comprises of verb prediction that describes the main activity in the image and semantic role labeling, which involves assigning semantic roles to different elements of the scene or action being executed– see Fig. 1. Semantic role labeling presents a several challenges, primarily due to the inherent ambiguity and variability in language and visual context. For instance, a single activity verb like "talking" expresses a spectrum of functional meanings and purposes, depending on the specific context captured in the image or video. Consider scenarios where "talking" might involve a person talking to another person, talking on the phone, or talking on the stage – see Fig. 1. Therefore, effective semantic role labeling necessitates a fine-grained understanding of the event, leveraging contextual cues from both the visual content and linguistic definitions associated with the activity and its specific roles.

Multimodal models for Situation Recognition. Multimodal models like CLIP [17], ALIGN [18], Flamingo [19], VILA [20], and LLaVa [21] have been pivotal in developing a joint semantic representation of vision and language. Trained on millions of image/text pairs, these models excel in capturing cross-modal dependencies between images and text. For instance, CLIP encounters various usages of verbs across visually diverse yet semantically similar images, providing a robust foundation for image semantic role labeling tasks. While CLIP demonstrates prowess in understanding both visual and linguistic information [22], other approaches leverage multimodal models differently. Methods like VL-Adapter [23], AIM [24], EVL [25], and wise-ft [26] apply MLPs on top of image encoders for tasks like image classification or action detection. However, semantic role labeling require a different approach, requiring both verb and role alongside the image. In addressing this challenge, Li et al. [27] propose CLIP-Event, a fine-tuned CLIP model for situation recognition via text-prompt-based prediction. Despite leveraging the vast world knowledge of GPT-3, CLIP-Event falls short in semantic role labeling compared to CoFormer [28], which is directly trained on images.

We present **ClipSitu**, a novel approach for semantic role labeling without the need for fine-tuning the CLIP model. We leverage CLIP-based image, verb and role embeddings and predict all roles for a verb by sharing information across them. We propose a cross-attention Transformer to learn the relation between semantic role queries and CLIP-based visual token representations. We term this model as ClipSitu XTF and it obtains state-of-the-art results for Situation Recognition on imSitu dataset [1]. We also introduce a new MLP model that uses the

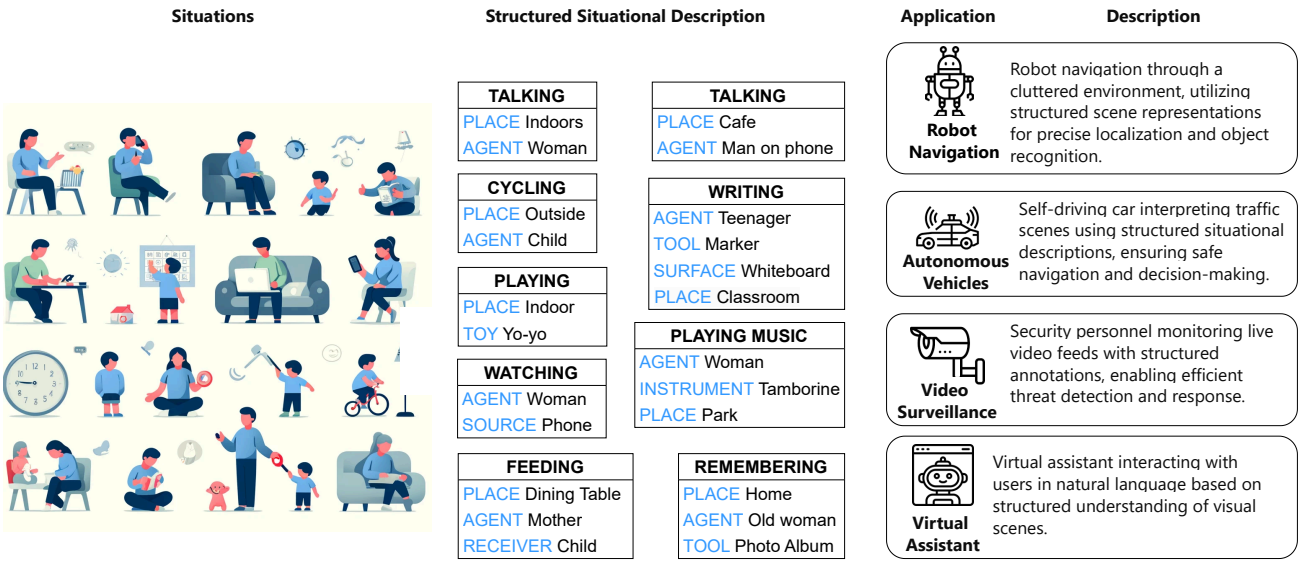


Figure 1: **Visual Understanding with Structured Situational Descriptions.** Each situation structure contains a verb in **bold** followed by the roles in blue and the noun fulfilling that role. Structured situational descriptions provide a detailed and organized view of visual scenes, empowering assistive systems to understand, reason about, and interact with the world more effectively. From robot navigation [6], autonomous vehicles [7], video surveillance [8, 9] to virtual assistants [10], structured representations enable a wide range of applications across diverse domains.

cross-attention scores from ClipSitu XTF to localize the noun for a role in the image and obtain state-of-the-art results for noun localization. We show that *noun localization using cross-attention scores* is more effective than existing approaches even when using predicted verbs which demonstrates the robustness of our XTF cross-attention mechanism.

We presented image situation recognition using ClipSitu at WACV 2024 [29]. We build upon [29] and present the following additional **contributions**. First, for video situation recognition [30], we extend ClipSitu models (MLP, TF and XTF) to accept CLIP image embeddings of multiple frames from a video. Then, we attach a transformer decoder to generate nouns as text phrases from the output of the ClipSitu models. We show that with these extensions, ClipSitu framework produces comparable performance to state-of-the-art on video situation recognition – see Section 4 for the model description and Sections 5.5, 5.6 and 6.2 for the results. Second, we evaluate a state-of-the-art Large Visual Language Model (VLM) on the structured prediction task of semantic role labeling with in-context examples Section 8. On both image and video semantic roles labeling, we observe that semantic role labeling is still a challenging task for VLM with in-context examples. Therefore, we fine-tune multiple VLMs in Section 8 to obtain improvement on both image and video situation recognition. However, we also show that task specific models such as our ClipSitu models perform reasonably well with a fraction of the parameters of the VLMs. Third, existing semantic role labeling frameworks require the semantic role to be provided for noun prediction. This may be a limitation when someone wants to deploy the models in real word as for some situations the exact frame structure may not be known beforehand. Instead, we predict roles associated with a verb with near perfect accuracy using image CLIP features (Section 7). Adding role prediction to ClipSitu allows us to create the first end-to-end framework for describing images with situational frames. We leverage our end-to-end framework to produce situational summaries of out-of-domain images from Conceptual Captions [31] – see Section 7.

i

2 Related Work

2.1 Situation Recognition

Situation Recognition approaches can be broadly classified into two categories – one-stage where the situational verb, and nouns for the semantic roles are predicted simultaneously and two-stage where the verb and nouns are predicted in succession.

2.1.1 One-stage Prediction

Many existing approaches exist for one-stage approaches predict both situational verbs and the associated nouns simultaneously from images [1, 32, 33, 34]. These approaches eliminate the need for sequential prediction of verbs from images and then roles from the verb and image. One-stage prediction reduces computational complexity capturing relationships between roles, nouns, and verbs in a single pass. One-stage models such as Gated Graph Neural Network (GGNN) [33] capture complex and varying relationships between roles without strict sequential constraints. GGNNs can represent non-linear dependencies between entities represented as nodes in a graph. Other one-stage approaches such as Conditional Random Fields (CRF) [1] and tensor decomposition models [32] consider global dependencies across all roles and nouns simultaneously, potentially leading to more coherent predictions. CRFs are probabilistic graphical models used in [1] to model dependencies between random variables i.e. verbs and roles. Tensor decomposition factorize tensors on top of CRF models to handle interactions between roles and nouns efficiently, improving model scalability and performance [32]. Another one-stage approach called mixture kernels [34] provides a prior distribution for predicting nouns based on identified relationships in the image. This approach enhances prediction robustness, especially for rare or underrepresented noun-role combinations.

One-stage approach for video situation recognition [35] leverages spatio-temporal scene graph (HostSG) to model fine-grained spatial semantics and temporal dynamics within and across events. HostSG bridges the gap between the underlying scene structure and high-level event semantics. They jointly predict the event verb, semantic roles, and event relations, avoiding the error propagation across tasks. However, one-stage models require extensive training data to generalize well across diverse scenes and contexts as they need to jointly model the verb and roles directly from the images. Models like CRFs and tensor decomposition struggle with capturing long-range dependencies across multiple roles and verbs. This issue is also prevalent in other approaches [36, 37] that predict nouns in a predefined sequential order. A final limitation of one-stage prediction is that model becomes complex to integrate complex relationships between roles, nouns, and verbs in a single pass increases training time.

2.1.2 Two-stage Prediction

Two-stage prediction approaches in situation recognition introduce an additional stage to refine verb predictions using predicted nouns [38, 39, 28, 40]. Two-stage approaches [40] also allow for refinement of verb predictions based on more accurate noun predictions to improve overall verb prediction accuracy. Another major advantage of these two stage approaches is using transformers that capture dependencies across longer sequences to provide comprehensive understanding of relationships between roles, verbs, and nouns. Two-stage approaches [38, 39, 28] leverage transformers to predict semantic roles using interdependent queries. Other two-stage methods [41] explore pre-trained multimodal transformers i.e. CLIP encoder on the image to predict the situational verb in the first stage and detect the objects in the second stage. Video situation recognition models capture temporal dependencies and contextual information across frames using two-stage approach as well. Most methods [30, 42, 43] use the transformer encoder to learn event wise contextual features from spatio-temporal features such as SlowFast [44]. The decoder is used to learn the role relations and generate the nouns. Grounded two-stage approaches with object features [42] and object state embedding (pixel level changes), object motion-aware embedding (motion of objects across multiple frames) and argument interaction embedding (relative position of multiple objects) [45] improve the precision of noun localization in videos which is crucial for accurately describing events over time.

Two-stage approaches also suffer from limitations such as running separate predictions for verbs and nouns can increase computation compared to one-stage approaches. Inaccurate verb predictions in the first stage can lead to cascading errors in subsequent noun predictions, necessitating robust error-handling mechanisms. In video situation recognition, computational complexity is higher compared to image-based approaches especially with the need to compute spatio-temporal features. We rely on CLIP

image feature derived video representation to train ClipSitu for video situation recognition. We adopt a two-stage mechanism for image situation recognition in ClipSitu for its advantages but train the verb and noun prediction separately to avoid propagation of errors. We also propose efficient cross-attention transformers for noun prediction and cross-attention score based MLP for localization of nouns. Similarly, for video situation recognition, we add a lightweight transformer decoder to ClipSitu for generation of role-wise nouns.

2.2 Structured Representation for Captioning

In the previous subsections, we have summarized approaches that focus on structured summary of images and videos using the situational framework. Structured representations have also been explored extensively for generating balanced image captions [46, 47, 48, 49, 50] and video captions [51, 35, 52, 48, 53, 54]. The structured representation used by these approaches are scene graphs that encode relationships between objects (nouns) and actions (verbs), providing a structured representation of visual scenes [46] and [47]. [53] propose Structured Concept Predictor (SCP) that enhances image captioning by predicting concepts and their relationships before generating captions. Similarly for video captioning, the Set Prediction Guided by Semantic Concepts (SCG-SP) framework [54] generates multiple captions for a single video by encoding semantic concepts detected in the frames. The object and relationships denoted in the scene graph reduces the semantic gap between visual (image/video) and text modalities. Therefore, captions generated using SG are better grounded and more relevant to the visual content [52, 48].

A limitation of constructing and utilizing scene graphs for caption generation is that it requires accurate object detection, relationship tagging, and large-scale graph reasoning. Relationship prediction in images and videos is relatively inaccurate [55, 56, 57]. Another challenge is the integration of scene graphs into caption generation pipelines requires careful alignment of graph representations with model architectures and training objectives. Therefore, we show that a structured summary of an image is in itself a compact and informative description of the image compared to the ground truth captions provided for the image. We leverage the trained ClipSitu on imSitu to generate the structured summary of out-of-domain images without the necessity of fine-tuning, paving the way for adoption of structured prediction approaches for image description.

3 CLIPSitu Models and Training

In this section, first, we present how we extract CLIP [17] embedding (features) for situation recognition. After that we present the verb prediction model. Then, we present three models for Situation Recognition using the CLIP embeddings. Finally, we present a loss function that we use to train our models.

3.1 Extracting CLIP embedding

Every image I has a situational action associated with it, denoted by a verb v. For this verb v, there is a set of semantic roles $R_v = \{r_1, r_2, \cdots, r_m\}$ each of which is played by an entity denoted by its noun $N = \{n_1, n_2, \cdots, n_m\}$. We use CLIP [17] visual encoder $\psi_v()$, and the text encoder $\psi_t()$ to obtain representations for the image, verb, roles, and nouns denoted by X_I, X_V, X_{R_v} and, X_N respectively. Here $X_{R_v} = \{X_{r_1}, X_{r_2}, \cdots, X_{r_m}\}$ for m roles and $X_N = \{X_{n_1}, X_{n_2}, \cdots, X_{n_m}\}$ for corresponding m nouns where $X_V = \psi_t(v)$, $X_{r_i} = \psi_t(r_i)$ and $X_{n_i} = \psi_t(n_i)$ are obtained using text encoder. Similarly, the $X_I = \psi_v(I)$ is obtained using vision encoder. Note that all representations X_I, X_V, X_{r_i} and X_{n_i} have the same dimensions.

3.2 ClipSitu Verb MLP

The first task in situation recognition is to predict the situational verb correctly from the image. We design a simple MLP with CLIP embeddings of the image X_I as input as follows:

$$\hat{v} = \phi_V(X_I). \tag{1}$$

where ϕ_V contains l linear layers of a fixed dimension with ReLU activation to predict the situational verb. Just before the final classifier, there is a Dropout layer with a 0.5 rate. We call this model as ClipSitu Verb MLP and we train it with standard cross-entropy loss.

3.3 ClipSitu MLP

Next, we design a modern multimodal MLP block for semantic role labeling for Situation Recognition that predicts each noun of the semantic role associated with the

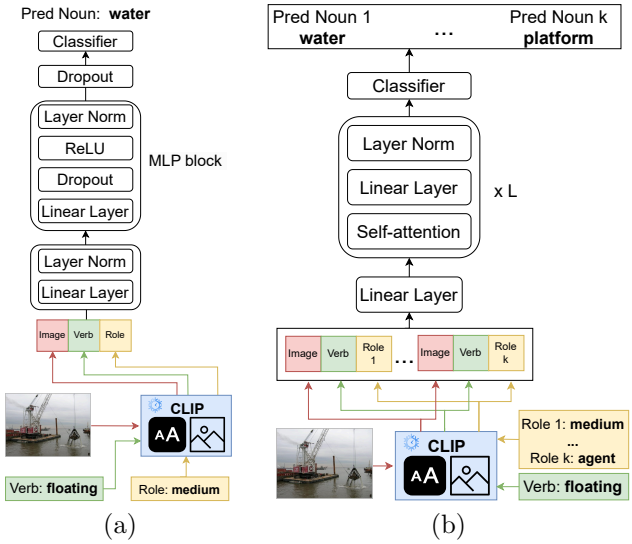


Figure 2: Architecture of (a) ClipSitu MLP and (b) ClipSitu TF. We use pooled image embedding from the CLIP image encoder for ClipSitu MLP and TF.

predicted verb in a given image. We term this method as **ClipSitu MLP**. The detailed architecture of ClipSitu MLP is shown in Fig. 2(a). Specifically, given the image, verb, and role embeddings, the ClipSitu MLP predicts the embedding of the corresponding noun for the role. In contrast to what has been done in the literature, ClipSitu MLP obtains contextual information by conditioning the information from the image, verb, and role embeddings. While the image embedding provides context about the possible nouns for the role, the verb provides the context on how to interpret the image situation.

We concatenate the role embedding for each role r_i to the image and verb embedding to form the multimodal input X_i where $X_i = [X_I, X_v, X_{r_i}]$. Then, we stack l MLP blocks to construct CLIPSitu MLP and use it to transform the multimodal input X_i to predict the noun embedding $\hat{X}_{n_i}$ as follows:

$$\hat{X}_{n_i} = \phi_{MLP}(X_i). \tag{2}$$

In ϕ_{MLP}, the first MLP block projects the input feature X_i to a fixed hidden dimension using a linear projection layer followed by a LayerNorm [58]. Each subsequent MLP block consists of a Linear layer followed by a Dropout layer (with a dropout rate of 0.2), ReLU [59], and a LayerNorm. We predict the noun class from the predicted noun embedding using a dropout layer (rate 0.5) followed by a linear layer which we name as classifier ϕ_c as

$$\hat{y}_{n_i} = \texttt{argmax}\ \phi_c(\hat{X}_{n_i}) \tag{3}$$

where $\hat{y}_{n_i}$ is the predicted noun class. We use cross-entropy loss between predicted $\hat{y}_{n_i}$ and ground truth nouns y_{n_i} as explained later in Section 3.5 to train the model.

3.4 ClipSitu TF: ClipSitu Transformer

The role-noun pairs associated with a verb in an image are related as they contribute to different aspects of the execution of the verb. Hence, we extend our ClipSitu MLP model using a Transformer [60] to exploit the interconnected semantic roles and predict them in parallel. The input to the Transformer is similar to ClipSitu MLP (i.e. $X_i = [X_I, X_v, X_{r_i}]$), however, we build a set of vectors using $\{X_1, X_2, \cdots, X_m\}$ where m denotes the number of roles of the verb. Each vector in the set is further processed by a linear projection to reduce dimensions. Our Transformer model ϕ_{TF} consists of l encoder layers where each encoder layer have h number of multi-head attention heads. Using the Transformer model, we predict the noun embedding of the m roles as output of the transformer

$$\{\hat{X}_{n_1}, \hat{X}_{n_2}, \cdots, \hat{X}_{n_m}\} = \phi_{TF}(\{X_1, X_2, \cdots, X_m\}). \tag{4}$$

Similar to the ClipSitu MLP (Section 3.3), we predict the noun classes using a classifier on the noun embedding as $\hat{y}_i = \texttt{argmax}\ \phi_c(\hat{X}_{n_i})$ where $i = \{1, \cdots, m\}$. The ClipSitu Transformer model for short know as ClipSitu TF is shown in Fig. 2(b).

3.5 ClipSitu XTF: Cross-Attention Transformer

Each semantic role in a situation is played by an object located in a specific region of the image. Therefore, it is important to pay attention to the regions of the image which has a stronger relationship with the role. Such a mechanism would allow us to obtain better noun prediction accuracy. Hence, we propose to use the encoding for each patch of the image obtained from the CLIP model. We design a cross-attention Transformer called **ClipSitu XTF** to model how each patch of the image is related to every role of the verb through attention as shown in Fig. 3.

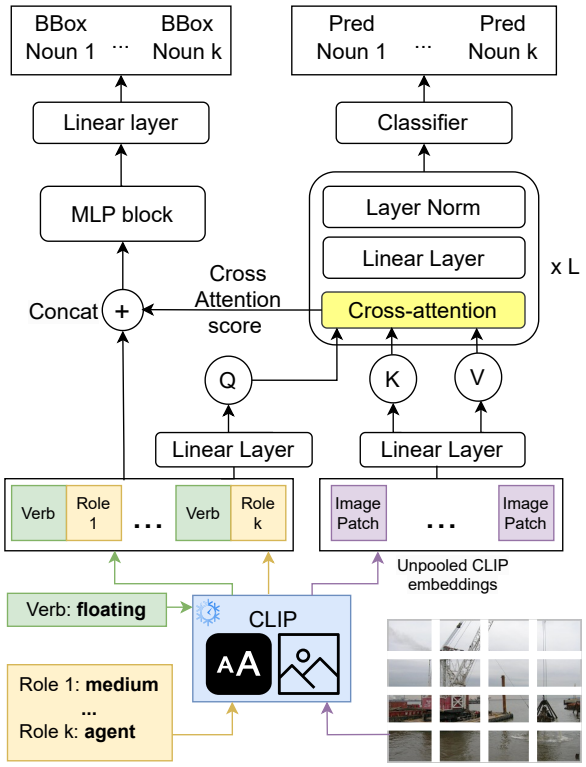


Figure 3: Architecture of ClipSitu XTF. We use embeddings from each patch of the image obtained from CLIP image encoder.

Let the patch embedding of an image be denoted by $X_{I,p} = \{X_I^1, X_I^2, \cdots, X_I^p\}$ where p is the number of image patches. These patch embeddings form the key and values of the cross-attention Transformer while the verb-role embedding is the query in Transformer. The verb embedding is concatenated with each role embedding to form m verb-role embeddings $X_{vr} = \{[X_V; X_{r_1}], [X_V; X_{r_2}], \cdots, [X_V; X_{r_m}]\}$. We project each verb-role embedding to the same dimension as the image patch embedding (i.e the embedding size of X_I^p) using a linear projection layer. Then the cross-attention operator in a Transformer block is denoted as follows:

$$Q = W_Q X_{vr}, \quad K = V = W_I X_{I,p}$$
$$\hat{X} = \text{softmax}\frac{QK^T}{\sqrt{d_K}}V \tag{5}$$

where W_Q and W_I represent projection weights for queries, keys, and values and d_K is the dimension of the key token K. As with ClipSitu TF, we have l cross-attention layers in ClipSitu XTF. The predicted output from the final cross-attention layer contains m noun embeddings $\hat{X} = \{\hat{X}_{n_1}, \hat{X}_{n_2}, \cdots, \hat{X}_{n_m}\}$. Similar to the transformer in Section 3.4, we predict the noun classes using a classifier on the noun embeddings as $\hat{y}_i = \phi_c(\hat{X}_{n_i})$ where $i = \{1, \cdots, m\}$.

Next, we use ClipSitu XTF to perform localization of roles that requires predicting a bounding box $\mathbf{b}_i$ for every role r_i in the image. The cross-attention matrix from the first layer of ClipSitu XTF is denoted by A where $A \in \mathcal{R}^{m \times p}$ is rearranged into m row vectors $\{A_1, \cdots, A_m\}$. Each row vector A_i is p-dimensional and each element in the vector shows how each patch in the image is related to the verb-role pair. To incorporate the verb and role context to the score vector (A_i), we concatenate each score vector to its corresponding verb-role embedding from X_{vr} to obtain input for localization denoted by X_l as follows:

$$\begin{aligned} X_l = &\{[X_V; X_{r_1}; A_1], \\ &[X_V; X_{r_2}; A_2], \cdots, [X_V; X_{r_m}; A_m]\} \end{aligned} \tag{6}$$

We pass X_l through a single MLP block (designed for ClipSitu MLP) followed by a linear layer and a sigmoid function to obtain the predicted bounding box $\hat{b}_i \in [0, 1]^4$ for every role r_i. The four elements in the predicted bounding box indicate the center coordinates, height and width relative to the input image size. Though ClipSitu XTF uses cross-attention as CoFormer [28], the verb role tokens in the query and the image tokens are obtained from CLIP and not learned. Leveraging the power of CLIP embeddings allows us to design a simpler one-stage ClipSitu XTF compared to the two-stage CoFormer [28].

Minimum Annotator Cross Entropy Loss. Popular situation recognition datasets employ multiple annotators to label each noun for a given role of an image instance. In some instances, annotators may not provide the same annotation for the same role as there can be multiple interpretations of the role. Existing approaches [38, 28] make multiple predictions instead of one to tackle this issue. However, this can confuse the network during training as there are multiple annotations for the same example. Furthermore, the loss function should not penalize a prediction that is close to any of the annotators' ground truth but further away from others. We propose minimum cross-entropy loss that considers the ground truth from each annotator. For a prediction $\hat{y}_i$, ground truth

from all the annotators $\mathcal{O} = \{O_1, \cdots O_q\}$ is used to train our network as follows

$$\mathcal{L}_{MAXE} = \min_{\mathcal{O}} -\sum_{c=1}^{C} y_{i,c}^{(O_j)} \log(\hat{y}_{i,c}) \text{ where } \forall O_j \in \mathcal{O}. \quad (7)$$

Here, C denotes the total number of noun classes and $\mathcal{L}_{MAXE}$ stands for Minimum Annotator Cross Entropy Loss.

To train ClipSitu XTF for localization of roles, we employ $L1$ loss to compare the predicted and ground-truth bounding boxes

$$\mathcal{L}_{L1} = \frac{1}{m} \sum_{i=1}^{m} \|b_i - \hat{b}_i\|_1. \quad (8)$$

We train ClipSitu XTF for noun prediction and localization using the combined loss $\mathcal{L} = \mathcal{L}_{MAXE} + \mathcal{L}_{L1}$.

4 Extending ClipSitu for Video Situation Recognition

Video situation recognition [30] (VidSitu) introduces the task of describing the situation in a 10 second video. We need to predict a set of k related events that constitute the situation $\{E_i\}_{i=1}^{k}$. Each event in the situation E_i is described by a verb v_i chosen from a set of verbs V and the nouns (entities, location, or other details of the event) described in text that are assigned to various roles of the verb. While the concept of situation is similar in images and videos, there are key differences between how imSitu and VidSitu are constructed. First, the number of verbs in VidSitu is thrice that that of imSitu (1560 vs. 504) as it distinguishes between homonyms based on verb-senses for e.g. "strike (hit)" is different from "strike (a pose)". Second, for each role in VidSitu, the noun is a text phrase instead of a noun class in imSitu e.g. "man holding shield" vs. "man". The large vocabulary of verbs and noun phrases makes it challenging to predict the exact verb and noun. Therefore, situation recognition in VidSitu is set up as a generation task and evaluated using semantic similarity metrics instead of being a classification task (details in Section 5.4).

4.1 ClipSitu with decoder overview

To extend ClipSitu models for videos and perform text generation, we make the following changes to ClipSitu models. First, we use video embeddings such as X-CLIP[61] for each event of duration 2 seconds in a 10 second video. Second, we add a language decoder based on transformer to our ClipSitu models as shown in Fig. 4 as the noun prediction is a generation task in VidSitu instead of classification in imSitu. The simultaneous noun prediction for each role in image situation recognition is replaced by sequential role-wise noun generation using the transformer decoder. For the ClipSitu MLP and TF models, we use the coarse-grained (video to text) embedding from X-CLIP along with the verb embedding. For the ClipSitu XTF model, we use the fine-grained (image to text) unpooled video-embedding obtained from X-CLIP before temporal pooling similar to our strategy of using unpooled CLIP features for imSitu. We use verb and role embeddings obtained from X-CLIP as roles are unique for every verb in VidSitu. In VidSitu [30], roles are denoted by placeholders – Arg0, Arg1, Arg2, ALoc (Location), AScn (Scene), ADir(Direction), and AMnr (Manner). However, the actual roles depending on the verb which are also available in the VidSitu, for e.g. Arg0 for the verb *driving* is *driver* and for *looking* is *looker*.

For all the ClipSitu models, the output obtained for each event video is fed as input to a transformer decoder to generate the nouns for each role. Apart from the ClipSitu outputs, the verb and ground-truth role and nouns are presented during training. The decoder is trained to generate a verb-noun sequence which is right shifted by one word as shown in Fig. 4. The placeholders Arg0, Arg1, Arg2, ArgScn are useful in evaluation as they are used to separate the generated nouns. During testing, the decoder receives the ClipSitu outputs and generates the rolewise nouns. VidSitu contains 5 events in a video that are semantically connected. Therefore, we also generate the nouns for all 5 events together which is shown to be effective in existing works [30, 42]. For ClipSitu MLP, we collect the output of all 5 events before sending to the transformer decoder. For ClipSitu XTF, verb and role act as query and unpooled X-CLIP embeddings as key and value. During cross-attention of XTF, event-aware masking is applied that restricts the verb and role of event i to attend to only the unpooled X-CLIP features from event i. The transformer decoder (TxD) that we add to our ClipSitu models is inspired by [30, 43]. The encoder outputs from MLP, TF, and XTF are used as encoder attention to start the decoding. During training, cross-

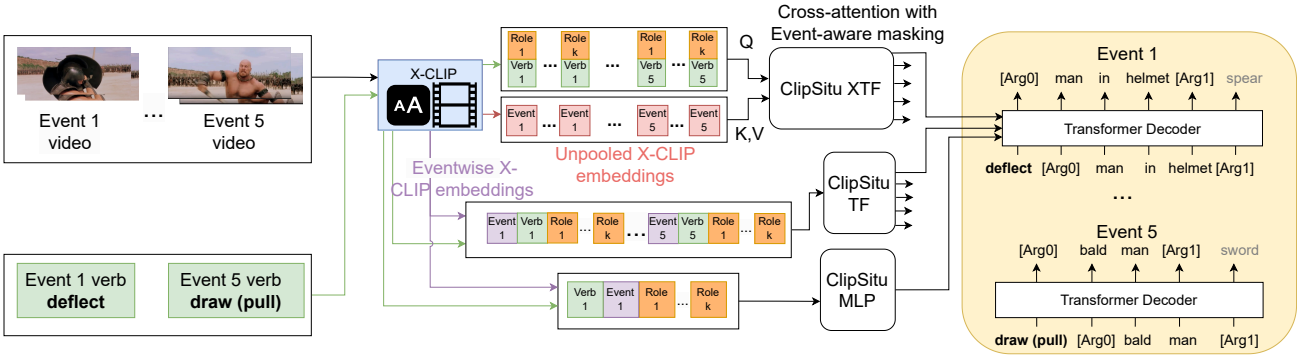


Figure 4: Extending ClipSitu models (XTF, TF, and MLP) for Video Situation Recognition (VidSitu). For every event video (2s duration) from a 10s video, X-CLIP [61] provides the video embeddings. We use a transformer decoder (TxD) for generating the nouns playing different roles for the verb of an event. We generate nouns for all 5 events in a video together. For ClipSitu XTF, verb and role acts as query and unpooled X-CLIP embeddings as key and value. ClipSitu XTF uses *event-aware masking* i.e. verb i (+roles) tokens attend only to event i tokens. For ClipSitu MLP, we collect the output of all 5 events before sending to the transformer decoder.

entropy loss is employed for each token in the predicted sequence.

4.2 ClipSitu with decoder details

For each event E_i with video embedding X_{E_i} (via X-CLIP), CLIP verb embedding X_{v_i} and CLIP role embeddings for each role $X_{r_i}^{Arg0}, X_{r_i}^{Arg1}, X_{r_i}^{Arg2}, X_{r_i}^{ArgScn}$, the input to TxD is:

$$X^{(i)} = \phi_{model}([X_{E_i}; X_{v_i}; X_{r_i}]), \tag{9}$$

where ϕ_{model} is either Clipsitu MLP, TF, or XTF, yielding embeddings for each role's noun.

Event-aware masking. In TF and XTF models, if an event's embeddings attempt to access frames from adjacent events, it could lead to information leakage across events, which would compromise the temporal coherence and accuracy of role-noun assignments within each event. With event-aware masking, we create a binary mask $M_{i,j}$ such that: $M_{i,j} = 1$ if $i = j$ (event E_i can attend only to itself) and $M_{i,j} = 0$ if $i \neq j$ (no cross-event attention is allowed). Let's denote the query matrix for roles in event E_i as $Q^{(i)}$ and the key and value matrices for frames in event E_i as $K^{(i)}$ and $V^{(i)}$, respectively. As the inputs from all events is fed together to the TxD, the attention score between query and key matrices across events is computed as $A_{i,j} = \text{softmax}\frac{Q^{(i)}K^{(j)\top}}{\sqrt{d_K}}$. The masked output is computed using the binary mask and the attention score

$$\hat{X} = A_{i,j} \cdot M_{i,j} \cdot V_i \tag{10}$$

By enforcing that each role-specific query only attends to its corresponding event's frame embeddings, the model is better able to focus on the local visual context relevant to that event. For instance, if event E_1 involves a "driving" action with the "driver" (Arg0) role, and event E_2 involves a "talking" action with a "speaker" (Arg0) role, event-aware masking prevents the "driver" role in E_1 from erroneously attending to frames in E_2. This localized attention structure preserves the temporal structure of the video and ensures that each event's role predictions are more contextually accurate.

Loss for video situation recognition. For image situation recognition, cross-entropy is used for noun prediction but for video situation recognition, cross-entropy loss is computed for sequence generation across multiple events. The transformer decoder TxD takes in ClipSitu output $X^{(i)}$ as input and outputs a right-shifted role-noun sequence $\hat{y}^{(i)}$ for each event of length $T^{(i)}$. The ground truth noun sequence is given by $y^{(i)} = X_{r_i}^{Arg0}, X_{r_i}^{Arg1}, X_{r_i}^{Arg2}, X_{r_i}^{ArgScn}$. The cross-entropy loss is applied to the predicted noun sequence during training,

calculated as:

$$\mathcal{L}^{(i)} = -\sum_{t=1}^{T^{(i)}} \sum_{c=1}^{C} y_{t,c}^{(i)} \log \hat{y}_{t,c}^{(i)} \quad (11)$$

where C is the size of vocabulary. Ground truth and predicted noun sequences ($y^{(i)}$ and $\hat{y}^{(i)}$) are aligned by TxD.

5 ClipSitu Ablations

5.1 imSitu Evaluation Details

We perform our experiments on imSitu dataset [1] and the augmented imSitu dataset with bounding boxes called SWiG [37] for the tasks of situation recognition and noun localization, respectively. Situation recognition with noun localization is called *grounded situation recognition.* The dataset has a total of 125k images with 75k train, 25k validation (dev set), and 25k test images. The metrics used for semantic role labeling are *value* and *value-all* [1] which predict the accuracy of noun prediction for a given role. For a given verb with k roles, *value* measures whether the predicted noun for at least one of k roles is correct. On the other hand, *value-all* measures whether all the predicted nouns for all k roles are correct. A prediction is correct if it matches the annotation of any one of the three annotators. Noun localization metrics *grnd value* and *grnd value-all* compute the accuracy of bounding box prediction [37] similar to value and value-all. A predicted bounding box is correct if the overlap with the ground truth is ≥ 0.5. The metrics value, value-all, grnd value and grnd value-all are evaluated in three settings based on whether we are using ground truth verb, top-1 predicted verb, or top-5 predicted verbs. For our model ablation on semantic role labeling and noun localization, we use the ground truth verb setting for measuring value, value-all, grnd value, and grnd value-all. All experiments are performed on the dev set unless otherwise specified.

5.2 Implementation Details on imSitu

We use the CLIP model with ViT-B32 image encoder to extract image features unless otherwise specified. The input to ClipSitu MLP is a concatenation of the CLIP embeddings of the image, verb, and role, each of 512 dimensions leading to 1536 dimensions. The input to ClipSitu TF is a sequence of tokens where each token is the concatenated image, verb, and role CLIP embeddings similar to ClipSitu MLP. To reduce the size of each token we project it to 512 dimensions using a linear layer. We set the sequence length for ClipSitu TF to be 6 which refers to the maximum number of roles possible for a verb following [38]. Each verb has a varying number of roles and we mask the inputs when a verb has less than 6 roles. The key and value for our patch-based cross-attention Transformer (ClipSitu XTF) is the unpooled image patch embeddings from CLIP image encoder . For example, when using the ViT-B32 model as the CLIP image encoder, we end up with 50 image patch embeddings ($224/32 \times 224/32 + 1$ class) of 512 dimensions that are used as key and value. The query of ClipSitu XTF is a sequence of concatenated verb and role CLIP embeddings (1024 dimensions) tokens where each token is projected to 512 dimensions using a linear layer. Similar to ClipSitu TF, the sequence length of the query in ClipSitu XTF is set to 6 and we mask the inputs when a verb has less than 6 roles. Unless otherwise mentioned, we train all our models with a batch size of 64, a learning rate of 0.001, and an ExponentialLR scheduler with Adamax optimizer, on a 24 GB Nvidia 3090. In imSitu dataset [1], we have 504 unique verbs, 190 unique roles, and 11538 unique nouns. We extract CLIP text embeddings each verb, role, and noun separately.

5.3 Analysis on imSitu with CLIP Image Encoders

Verb Prediction Comparison. In Table 1, we compare the proposed ClipSitu Verb MLP model against zero-shot and linear probe performance of CLIP. We also compare against a CLIP finetuning model called weight-space ensembles (wise-ft) [26] that leverages both zero-shot and fine-tuned CLIP models to make verb predictions. We compare ClipSitu Verb MLP and wise-ft using 4 CLIP image encoders - ViT-B32, ViT-B16, ViT-L14, and ViT-L14@336px. The image clip embeddings for ViT-B32 and ViT-B16 are 512 dimensions and for ViT-L14, and ViT-L14@336px are 768 dimensions. These four encoders represent different image patch sizes, different depths of image transformers, and different input image sizes. We set the hidden layer dimension to 1024 in the ClipSitu Verb MLP. The zero-shot performance of CLIP suggests that CLIP image features are beneficial for situation recognition tasks. The best performance of ClipSitu Verb MLP is obtained with a single hidden layer and increasing the number of hidden layers does not improve performance. Our best performing ClipSitu Verb MLP outperforms lin-

Image Encoder	Verb Model	Hidden Layer	Top-1	Top-5
ViT-B32	zero-shot	-	29.2	65.2
	linear probe	-	44.6	78.4
	wise-ft	-	46.5	74.3
	ClipSitu Verb MLP	1	46.7	76.1
		2	46.5	76.1
		3	44.5	74.2
ViT-B16	zero-shot	-	31.9	67.89
	linear probe	-	49.3	78.8
	wise-ft	-	48.8	83.5
	ClipSitu Verb MLP	1	50.9	89.6
		2	50.8	89.4
		3	48.6	88.5
ViT-L14	zero-shot	-	38.2	79.3
	linear probe	-	52.4	87.7
	wise-ft	-	51.5	84.3
	ClipSitu Verb MLP	1	56.7	84.6
		2	56.6	84.5
		3	53.8	82.4
ViT-L14 @336px	zero-shot	-	39.7	79.2
	linear probe	-	53.4	81.4
	wise-ft	-	52.2	82.9
	ClipSitu Verb MLP	1	**57.9**	**86.1**
		2	56.2	84.5
		3	54.3	82.8

Table 1: Comparing performance of ClipSitu Verb MLP with zero-shot and finetuned CLIP (linear probe and wise-ft [26]).

ear probe by 4.46% on Top-1 and wise-ft by 3.2% on Top-5 when using the same ViT-L14 image encoder. ClipSitu Verb MLP performs better than wise-ft and linear probe for all image encoders. Therefore, verb prediction benefits from a well-designed MLP model such as our ClipSitu Verb MLP than finetuning the CLIP image encoder itself (wise-ft) or using regression (linear probe) for situational verb prediction.

Noun Prediction Comparison. Next, we study the effect of using different CLIP image encoders for noun prediction with ClipSitu MLP, TF and XTF. Again as verb prediction, we use four image encoders for a thorough comparison. Please note that the number of image patch embeddings used as key and value changes based on patch size and image size for ClipSitu XTF. We obtain 197 image patch embeddings ($224/16 \times 224/16 + 1$ class token) for ViT-B16, 257 image patch embeddings for ViT-L14 ($224/14 \times 224/14 + 1$ class token), and 577 image patch embeddings for ViT-L14@336px ($336/14 \times 336/14 + 1$ class token). For ViT-L14, and ViT-L14@336px image encoders, we obtain 768-dimensional embeddings which we project using a linear layer to 512 dimensions. We choose the best hyper-parameters for ClipSitu MLP, TF, and XTF based on the ablations presented in Section 5.7.

Image Encoder	Model	Top-1		Top-5		Ground truth	
		value	v-all	value	v-all	value	v-all
ViT-B32	MLP	45.6	27.1	66.3	37.5	76.9	43.2
	TF	45.7	27.3	66.3	38.0	76.8	43.0
	XTF	44.5	25.9	64.9	35.6	75.2	40.8
ViT-B16	MLP	46.3	28.2	67.3	39.4	77.9	44.8
	TF	46.4	28.6	67.4	39.8	77.2	43.8
	XTF	45.7	27.4	66.1	37.4	75.4	40.6
ViT-L14	MLP	46.5	28.4	67.6	39.7	77.6	43.9
	TF	46.9	29.6	68.2	41.2	78.0	45.2
	XTF	46.9	29.5	68.1	40.6	77.8	44.5
ViT-L14 @336px	MLP	46.7	29.1	67.9	40.5	77.9	44.9
	TF	47.0	29.7	68.3	41.4	78.3	45.8
	XTF	**47.2**	**30.1**	**68.4**	**41.7**	**78.5**	**45.8**

Table 2: Comparison of CLIP Image Encoders on noun prediction task using top-1 and top-5 predicted verb from the best-performing Verb MLP model obtain from Table 1. All models' performance improves by increasing the number of patch tokens either by reducing patch size (32→16→14) or increasing image size (224→336). v-all stands for value all.

In Table 2, we observe that the value and value-all using ground truth verbs steadily improve for all three models as the number of patches increase from 50 (ViT-B32) to 577 (ViT-L14@336px) or the image size increases from 224×224 to 336×336. For smaller backbones such as ViT-B32 and ViT-B16, the best performance is obtained by ClipSitu MLP while ClipSitu XTF shows the most improvement when using a larger backbone (ViT-L14) and the largest image size (ViT-L14@336px). ClipSitu XTF is able to extract more relevant information when attending to more image patch tokens to produce better predictions. To compare noun prediction using top-1 and top-5 predicted verbs, we use the best ClipSitu Verb MLP (ViT-L14@336px) from Table 1. Even when using Top-1 and Top-5 predicted verbs, we observe a similar trend as the ground truth verb. ClipSitu XTF again shows the most improvement in value and value-all to obtain the best performance among the three models across ground truth, Top-1 and Top-5 predicted verbs.

5.4 VidSitu Evaluation Details

Semantic role labeling in VidSitu is evaluated using generation metrics such CIDEr, Rouge-L, and Lea. VidSitu comprises of 23,626 training and 1326 validation videos. Following [30], we generate the nouns 4 roles at most for every verb depicted by placeholders Arg0, Arg1, Arg2, ArgScn (Scene) that vary for every verb. The genera-

tion metrics used for comparison are CIDEr [62] macro-averaged for verb (C-Vb), and macro-averaged CIDEr (C-Arg), Rouge [63] (R-L) and Lea [64] for noun.

5.5 Implementation Details on VidSitu

For video embedding, we use X-CLIP [61] which is a minimal extension to CLIP for videos that captures similarities between video and text at both coarse-grained (video-sentence level) and fine-grained (frame-word level). One of the main distinctions in transformer decoder (TxD) comprises of 3 layers with 8 attention heads inspired by existing works [30, 43]. The encoder outputs from MLP, TF, and XTF are used as encoder attention to start the decoding. We use greedy decoding with temperature 1.0 instead of beam search as it provides better performance.

5.6 Analysis on VidSitu with different X-CLIP Encoders

We compare the performance of ClipSitu MLP, TF and XTF with the decoder(TxD) in Table 3. We use two X-CLIP features – ViT-B32 and ViT-L14 for thorough testing of our models. ClipSituMLP+TxD performs the best among all the models on 4 out of 5 generation metrics using either of the X-CLIP features. We attribute this to ClipSituMLP making the least changes to the input X-CLIP features compared to ClipSituTF and XTF. This property also allows ClipSituMLP+TxD to perform better with a more descriptive feature such as ViT-L14 compared to ViT-B32.

5.7 Ablations on imSitu

ClipSitu MLP ablation In Fig. 5, we explore combinations of MLP blocks and the dimension of the hidden (linear) layer in each block to obtain the best MLP network for semantic role labeling. We train ClipSitu MLP with small to very large hidden layer dimensions i.e. 128→16384 which results in a steady improvement in both value and value-all. Increasing the number of MLP blocks also improves the performance as a model that suggests we need a model with more parameters to predict the large number of unique nouns – 11538. However, we reach saturation at hidden layer dimension of 32768 with 3 MLP blocks as no further improvement in value and value-all is observed. Our best ClipSitu MLP for semantic role labeling obtains 76.91 for value and 43.22

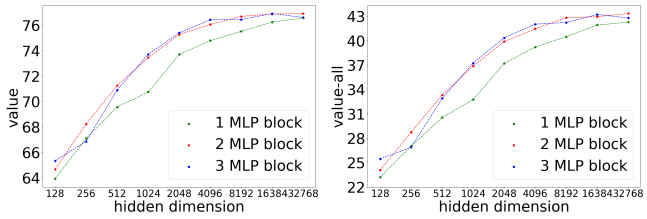


Figure 5: Effect of the number of MLP blocks and hidden dimensions on value and value-all. We train with very large hidden dimensions such as 8192, 16384, and 32768 to obtain state-of-the-art value and value-all results.

for value-all with 3 MLP blocks with each block having 16,384 hidden dimensions which beats the state-of-the-art CoFormer [28]. The main reason our ClipSitu MLP performs well on semantic role labeling is our modern MLP block design that contains large hidden dimensions along with LayerNorm which have not been explored in existing MLP-based CLIP finetuning approaches such as wise-ft [26]. We also compare the performance of ClipSitu MLP with our proposed minimum annotator cross-entropy loss ($\mathcal{L}_{MAXE}$) versus applying cross-entropy using the noun labels of each annotator separately. We find that $\mathcal{L}_{MAXE}$ produces better value and value-all performance (76.91 and 43.22) compared to cross-entropy (76.57 and 42.88).

ClipSitu TF and XTF ablation In Table 4, we explore the number of heads and layers needed in ClipSitu TF and XTF to obtain the best performance in semantic role labeling. We find that a single head with 4 transformer layers performs the best in terms of value for both ClipSitu TF and XTF while for value-all, an 8-head 4-layer ClipSitu TF performs the best and we use this for subsequent evaluation. For both ClipSitu TF and XTF, increasing the number of layers beyond 4 does not yield any improvement in value or value-all when using less number of heads (1,2,4). Similarly, for ClipSitu XTF, increasing the number of heads and layers leads to progressively deteriorating performance. Both of these performance drops can be attributed to the fact that we have insufficient samples for training larger transformer networks [17].

Comparison with other multimodal embeddings We compare the performance of CLIP with another multimodal embedding model – ALIGN [18] which is trained to align images with their captions Table 6. We also compare CLIP text embedding with other text embedding methods such as Glove [65] and BERT [66] to obtain verb

Method	Feature	CIDEr	C-Vb	C-Arg	R-L	Lea
ClipSituMLP+TxD	X-CLIP (ViT-B32)	60.53	**71.21**	52.80	43.29	37.30
ClipSituTF+TxD		58.97	67.75	49.81	42.25	50.19
ClipSituXTF+TxD		57.75	66.17	47.04	41.71	49.11
ClipSituMLP+TxD	X-CLIP (ViT-L14)	**61.93**	70.14	**56.30**	**43.77**	37.77
ClipSituTF+TxD		59.03	68.07	53.51	39.93	45.29
ClipSituXTF+TxD		49.65	58.35	44.91	42.40	**50.91**

Table 3: Comparison of ClipSitu MLP, TF and XTF models on VidSitu with different X-CLIP features. ClipSitu MLP+TxD performs the best on 4 out of 5 generation metrics.

Heads		1				2				4				8			
Layers		1	2	4	6	1	2	4	6	1	2	4	6	1	2	4	6
ClipSitu TF	value	75.73	75.78	**76.87**	24.68	75.80	75.95	75.97	18.28	75.71	76.77	75.87	05.20	75.74	75.93	76.07	75.94
	value-all	41.40	41.52	41.84	00.21	41.64	41.60	41.83	00.21	41.43	42.10	41.75	00.00	41.35	41.58	**42.97**	41.72
ClipSitu XTF	value	72.70	74.33	**75.27**	74.35	53.11	53.17	53.11	53.16	53.18	53.51	53.45	53.49	53.13	53.44	53.38	53.54
	value-all	36.61	39.11	**40.79**	39.06	16.58	16.64	16.38	16.46	16.50	16.77	16.90	16.97	16.42	16.89	16.85	17.02

Table 4: Ablation on Transformer hyperparameters. 1 head with 4 layers is sufficient to obtain best value and value-all performance for ClipSitu XTF. For TF, 1 head and 4 layers produces best value whereas 8 heads and 4 layers produces best value-all performance.

Model	Top-1		Top-5		Ground truth	
	grnd value	grnd v-all	grnd value	grnd v-all	grnd value	grnd v-all
verb-role emb	18.91	05.34	22.45	17.75	27.81	20.07
XAtt. L1	39.30	10.54	46.46	19.70	53.87	30.22
XAtt. L4	33.30	09.34	42.55	17.53	51.32	31.32
XAtt. L1 + L4	36.56	09.88	44.23	11.43	54.56	34.71
verb-role emb + XAtt L1	**41.30**	**13.92**	**49.23**	**23.45**	**55.36**	**32.37**

Table 5: Comparing verb-role embeddings with different ClipSitu XTF inputs for noun localization. XAtt. – cross-attention scores. L1 – first XTF layer and L4 – last XTF layer of best performing model (1 head, 4 layers, ViT-L14@336px). Cross-attention scores vastly improves the performance of noun localization. We use the best Verb MLP from Table 1. v-all stands for value-all.

and role embeddings. ALIGN does not perform as well as CLIP on semantic role prediction perhaps because it does not capture objects well in the images as it trained mainly on alt-text data that may not describe all the objects in the image. Glove and BERT verb and role embeddings with CLIP image embeddings perform well on both verb prediction and semantic role prediction. Therefore, ClipSitu XTF is flexible enough to incorporate any text embedding model and can improve with better text embedding models.

Complexity We compare the number of parameters, computation, and inference time for ClipSitu MLP, TF, and XTF using the ViT-L14-336 image encoder and CoFormer [28] in Table 7. We find that ClipSitu TF is the most efficient in terms of parameters, computation, and inference time closely followed by ClipSitu XTF at half the parameters of CoFormer and 9% of ClipSitu MLP. Even adding the lightweight ClipSitu Verb MLP to ClipSitu XTF for combined verb and noun prediction leads to a very efficient but effective model. Therefore, we conclude that ClipSitu XTF not only performs the best at semantic role labeling but is also efficient in terms of parameters, computation, and inference time compared to ClipSitu MLP and CoFormer [28].

Qualitative Comparison of XTF. In Fig. 6, we compare the qualitative results of ClipSitu XTF with CoFormer. ClipSitu XTF is able to correctly predicts verbs such as cramming (Fig. 6(b)) while CoFormer focuses on the action of eating and hence incorrectly predicts the verb which also makes its noun predictions for the container and theme incorrect. CoFormer predicts the place as table and predicts the verb as dusting (Fig. 6(c)) instead of focusing on the action of nagging. Finally, we see in Fig. 6(d) that CoFormer is confused by the visual context of kitchen as it predicts stirring instead of identifying the action which is drumming. On the other hand, ClipSitu XTF correctly predicts drumming and the tool as drumsticks while still predicting place as kitchen. In Fig. 7, we show some qualitative examples where ClipSitu Verb MLP predicts the verb incorrectly or ClipSitu XTF predicts the noun incorrectly.

Noun Localization ablation. In Table 5, we present noun localization results of ClipSitu XTF. First, we show that cross-attention scores vastly outperform verb-role embeddings. Second, cross-attention scores taken from

Verb & Role embedding	Image emb.	Top-1 predicted verb			Top-5 predicted verb			Ground truth verb	
		verb	value	v-all	verb	value	value-all	value	value-all
ALIGN [18]	ALIGN	52.27	31.17	12.17	80.56	45.26	15.75	52.95	16.86
Glove [65]	CLIP	55.03	42.40	24.27	82.96	62.60	33.94	73.44	37.89
BERT [66]	CLIP	55.03	42.10	23.47	82.96	62.15	33.09	73.03	37.24
CLIP	CLIP	**58.19**	**47.23**	**29.73**	**85.69**	**68.42**	**41.42**	**78.52**	**45.31**

Table 6: Comparing different text and image embedding models on the XTF model. Glove and BERT verb and role embedding perform well with CLIP image embeddings.

Model	# Parameters	GFlops	Inference Time(ms)
CoFormer [28]	93.0M	1496.67	30.62
ClipSitu Verb MLP	1.3M	0.17	0.08
ClipSitu MLP	580.2M	443.18	32.33
ClipSitu TF	**20.2M**	**8.65**	**1.55**
ClipSitu XTF	45.3M	116.01	11.17

Table 7: Comparison of parameters, flop count and inference time for CoFormer [28], Verb MLP, ClipSitu MLP, TF and XTF models.

the first (XAtt. L1) performs better than the fourth layer (XAtt. L4) indicating that cross-attention score is more useful for localization the less transformation it has undergone i.e. closer to the original image. Cross-attention scores are obtained from the patch-based CLIP image features as key and value and verb-role embeddings as query. Therefore, the cross-attention score represents the similarity of each patch to the role. The similarity information helps accurately predict the patches corresponding to each noun. We also concatenate verb and role embeddings to the cross-attention score (XAtt. L1 + verb-role) to provide more context about the role which further improves localization performance. We visualize these quantitative noun localization results using qualitative examples Fig. 8(a)-(e) of noun localization. We visualize a random role per image where the noun is predicted correctly by ClipSitu XTF. We compare the bounding box predicted for that noun using verb-role embeddings and cross-attention scores (XAtt. L1). Cross-attention score produces bounding boxes closer to the ground-truth than verb-role embedding. We also show a challenging in Fig. 8(f) image where the noun occupies a very small region in the image where both cross-attention and verb-role embeddings are not able to localize it accurately.

6 Comparison with state-of-the-art

6.1 Image Situation Recognition

In Table 8, we compare the performance of proposed approaches with state-of-the-art approaches on both grounded and ungrounded situation recognition. Approaches that do not perform grounding and are used for comparison with our approach are as follows: (a) Conditional Random Field (CRF) [1] trained separately for verb prediction and noun prediction, (b) Recurrent Neural Network with fusion (RNN w/ Fus.) [32] that uses CNN for verb prediction and an RNN for sequential prediction of nouns given the verb and image, (c) Graph Attention Network (GraphNet) [33] containing interconnected verb and role nodes used for predicting the verb and nouns, respectively, (d) Context aware Reasoning (CAQ RE-VGG) [38] utilizes a top-down attention mechanism to query the role from region features of an image image and the predicted verb from a CNN, (e) Mixture-Kernel Graph Attention Network (Ker. GraphNet) [34] extend GraphNet [33] with context-aware interactions between role pairs across verbs through a set of shared basis kernel matrices, (f) CLIP-Event [27] uses CLIP for image encoding and GPT3 for decoding the entire situation description as a sentence containing the verb and the nouns. The next set of approaches also perform noun grounding: (g) Joint Situation Localization (JSL) [37] jointly predicts verb, nouns and noun locations using a ResNet-LSTM network, (h) GSRTR [39] uses a vision transformer encoder for verb prediction and a transformer decoder for noun prediction and localization, (i) SituFormer [40] extends GSRTR by re-ranking predicted nouns for every role using other examples of the same verb, (j) Collaborative Transformer (CoFormer) [28] is similar to SituFormer but decomposes noun prediction and localization by predicting nouns first and then their locations based on the verb and predicted nouns based on the image.

We use ViT-L14@336px image encoder for both ClipSitu Verb MLP and ClipSitu XTF. ClipSitu Verb MLP

Set	Method	Top-1 predicted verb					Top-5 predicted verb					Ground truth verb			
		verb	value	v-all	grnd value	grnd v-all	verb	value	value-all	grnd value	grnd value-all	value	value-all	grnd value	grnd value-all
dev	CRF [1]	32.25	24.56	14.28	-	-	58.64	42.68	22.75	-	-	65.90	29.50	-	-
	RNN w/ Fus. [36]	36.11	27.74	16.60	-	-	63.11	47.09	26.48	-	-	70.48	35.56	-	-
	GraphNet [33]	36.93	27.52	19.15	-	-	61.80	45.23	29.98	-	-	68.89	41.07	-	-
	CAQ RE-VGG [38]	37.96	30.15	18.58	-	-	64.99	50.30	29.17	-	-	73.62	38.71	-	-
	Ker. GrphNet [34]	43.21	35.18	19.46	-	-	68.55	56.32	30.56	-	-	73.14	41.68	-	-
	JSL [37]	39.60	31.18	18.85	25.03	10.16	67.71	52.06	29.73	41.25	15.07	73.53	38.32	57.50	19.29
	GSRTR [39]	41.06	32.52	19.63	26.04	10.44	69.46	53.69	30.66	42.61	15.98	74.27	39.24	58.33	20.19
	SituFormer [40]	44.32	35.35	22.10	29.17	13.33	71.01	55.85	33.38	45.78	19.77	76.08	42.15	**61.82**	24.65
	CoFormer [28]	44.41	35.87	22.47	29.37	12.94	72.98	57.58	34.09	46.70	19.06	76.17	42.11	61.15	23.09
	ClipSitu XTF	**58.19**	**47.23**	**29.73**	**41.30**	**13.92**	**85.69**	**68.42**	**41.42**	**49.23**	**23.45**	**78.52**	**45.31**	55.36	**32.37**
test	CRF [1]	32.34	24.64	14.19	-	-	58.88	42.76	22.55	-	-	65.66	28.96	-	-
	RNN w/ Fus [36]	35.90	27.45	16.36	-	-	63.08	46.88	26.06	-	-	70.27	35.25	-	-
	GraphNet [33]	36.72	27.52	19.25	-	-	61.90	45.39	29.96	-	-	69.16	41.36	-	-
	CAQ RE-VGG [38]	38.19	30.23	18.47	-	-	65.05	50.21	28.93	-	-	73.41	38.52	-	-
	Ker. GrphNet [34]	43.27	35.41	19.38	-	-	68.72	55.62	30.29	-	-	72.92	42.35	-	-
	JSL [37]	39.94	31.44	18.87	24.86	09.66	67.60	51.88	29.39	40.6	14.72	73.21	37.82	56.57	18.45
	GSRTR [39]	40.63	32.15	19.28	25.49	10.10	69.81	54.13	31.01	42.5	15.88	74.11	39.00	57.45	19.67
	SituFormer [40]	44.20	35.24	21.86	29.22	13.41	71.21	55.75	33.27	46.00	20.10	75.85	42.13	**61.89**	24.89
	CoFormer [28]	44.66	35.98	22.22	29.05	12.21	73.31	57.76	33.98	46.25	18.37	75.95	41.87	60.11	22.12
	CLIP-Event [27]	45.60	33.10	20.10	21.60	10.60	-	-	-	-	-	-	-	-	-
	ClipSitu XTF	**58.09**	**47.21**	**29.61**	**40.01**	**15.03**	**85.69**	**68.42**	**41.42**	**49.78**	**25.22**	**78.52**	**45.31**	54.36	**33.20**

Table 8: Comparison with state-of-the-art on Grounded Situation Recognition. Robustness of ClipSitu MLP, TF, and XTF is demonstrated by the massive improvement for value and value-all with Top-1 and Top-5 predicted verbs over SOTA.

outperforms SOTA method CoFormer on Top-1 and Top-5 verb prediction by a large margin of 12.6% and 12.4%, respectively, on the test set, which shows the effectiveness of using CLIP image embeddings over directly predicting the verb from the images. We compare ClipSitu XTF to the only other CLIP-based situation recognition approach – CLIP-Event [27]. ClipSitu XTF outperforms CLIP-Event on the tasks of verb prediction by 13%, semantic role labeling (value) by 14%, and noun localization (top-1 predicted verb – grnd value) by 18%. ClipSitu XTF also performs the best for noun prediction based on both the predicted top-1 verb and top-5 verb for value and value-all matrices. ClipSitu XTF betters state-of-the-art CoFormer by a massive margin of 14.1% on top-1 value and by 9.6% on top-1 value-all using the top-1 predicted verb on the test set. For noun localization, ClipSitu XTF uses *cross-attention scores* to obtain better performance than state-of-the-art CoFormer [28] that uses verb and role information only. ClipSitu XTF shows a massive 11% improvement in noun localization over CoFormer When the ground-truth verb is unavailable (grnd value for top-1 predicted verb). We explain this performance improvement with a detailed comparison of cross-attention scores to verb-role embedding in Table 5.

6.2 Video Situation Recognition

We compare the ClipSitu models with transformer decoder on the VidSitu dataset in Table 9. Our results are reported on the validation set following existing approaches. To compare fairly with existing approaches, we test the ClipSitu models using ground-truth roles provided for every verb. Additional information in the form of object bounding boxes is used in [48, 42] due to which [42] performs the best among the existing approaches. When trained only on videos in the absence of object bounding boxes, overall CIDEr drops by about 10% for [42] but they do not provide other metrics for this setting. Our ClipSitu models for VidSitu do not consider object features making them easily extensible from imSitu to VidSitu with the addition of a decoder. We wanted ClipSitu to predict situations directly from the video as object bounding boxes are highly dependent on object detection accuracy and ground truth bounding boxes will not be available in practice. Even when compared to approaches trained with bounding boxes [48, 42], ClipSituMLP+TxD is comparable in performance on Rouge-L. Among all the metrics, ClipSitu results are lower on Lea suggesting that ClipSitu sometimes misidentify key roles with incorrect nouns, resulting in incomplete or incorrect scene interpretations. Lea weights each entity by its "importance" and "resolution score," a low score might suggest that ClipSitu commits ambiguous role assignments such as confusing "driver" with "passenger" weakening the accuracy and coherence of the scene interpretation. In summary, comparing among approaches that do not use object bounding boxes, ClipSituMLP+TxD performs comparable or better than existing approaches on 4 out of 5 metrics.

Method	Object BBoxes	CIDEr	C-Vb	C-Arg	R-L	Lea
OSE [45]	Yes	48.46	56.04	44.60	41.89	-
VideoWhisperer [42]	Yes	**68.54**	**77.48**	**61.55**	**45.70**	47.54
SlowFast+TxE+TxD [30]	No	46.01	56.37	43.58	43.04	<u>50.89</u>
Slow-D+TxE+TxD [43]	No	60.34	69.12	53.87	<u>43.77</u>	46.77
VideoWhisperer [42]	No	47.91	-	-	-	-
ClipSituMLP+TxD	No	<u>61.93</u>	<u>70.14</u>	<u>56.30</u>	<u>43.77</u>	37.77

Table 9: Comparison with state-of-the-art on Video Situation Recognition. TxE – Transformer Encoder, TxD – Transformer Decoder, C-Vb – CIDEr Verb, C-Arg – CIDEr Argument, R-L – Rouge-L. ClipSituMLP+TxD performs the best in 4 out of 5 metrics among all approaches that do not use object bounding boxes. <u>Underline</u> denotes second best performance.

7 Situational Summary of Out-of-Domain Images

In existing semantic role labeling approaches, the roles are provided based on the verb to predict the corresponding noun. For example, as shown in Fig. 6, for the verb *cramming*, we need to provide the roles *theme, agent,* and *container*. This restricts us from building an end-to-end framework that describes any image using situational frames where the ground-truth verb and role annotations are not available. To overcome this limitation, we introduce a novel role prediction model using the imSitu dataset.

7.1 Role prediction on imSitu

Our role prediction model leverages a multi-headed MLP architecture to simultaneously predict multiple roles associated with a given verb. Each head outputs a probability distribution across 191 classes – 190 for the unique roles and an additional class for cases where a role does not exist. The input to our model consists of concatenated image and verb embeddings $[X_I; X_V]$. As roles are context specific, we cannot predict them without knowing the verb first. For example, the image in Fig. 10 can be perceived as *buying* or *selling*. In such a case, the roles are dependent on the verb and we need to use the verb embedding of the predicted verb.

We utilized various CLIP encoder configurations, from ViT-B/16 and ViT-B/32 to ViT-L/14 and ViT-L/14-336, to generate these embeddings. The model output $\hat{r}_{i,j}; i \in \{1, \cdots, m\}, j \in \{1, \cdots, R\}$ is the predicted role out of the 191 classes (R) for each head. As we need to predict a maximum of 6 roles ($m = 6$) per input, we trained the model using *multi-head cross-entropy* loss that computes the cross-entropy loss for each prediction head (role) separately and returns the average loss across all heads.

$$\mathcal{L}_{ROLE} = -\frac{1}{m}\frac{1}{R}\sum_{i=1}^{m}\sum_{j=1}^{R} r_{i,j} log(\hat{r}_{i,j}). \tag{12}$$

$\mathcal{L}_{ROLE}$ ensures balanced learning across all the roles of a verb. Remarkably, our model achieves near-perfect training and validation accuracy, ranging between 99-100%, within one epoch for all tested CLIP encoder configurations. With this exceptional level of accuracy, ClipSitu enables us to predict situational frames directly from images.

7.2 Situational Summary on Conceptual Captions Dataset

Our role prediction model completes our end-to-end pipeline of obtaining situational summaries from an image. Therefore, we present an application of ClipSitu – situational summaries of out-of-domain images. We select images from Conceptual Captions dataset [31] that encompass a wide range of visual concepts and abstractness levels.

We show examples from Conceptual Captions in Fig. 9 comparing situational summaries from ClipSitu and the captions in Conceptual Captions. ClipSitu effectively identifies relevant verbs and roles for action-oriented images, providing a concise representation of the key elements and relationships within an image. ClipSitu results are more grounded in the image than captions which can be sometimes incorrect Fig. 9(g) or contain additional information not observed in the image that cannot be verified Fig. 9(b), (c), (h), (i). ClipSitu faces challenges with more abstract images or cases where it attempts to force an action or activity even when only a faint relation to the action exists Fig. 9(f), (h).

GT verb & Roles	CoFormer	ClipSitu XTF
ENCOURAGING	SCOLDING	ENCOURAGING
PLACE	LAKE	BEACH
AGENT	WOMAN	WOMAN
RECEIVER	MALE CHILD	MALE CHILD

(a)

CRAMMING	EATING	CRAMMING
THEME	SANDWICH	HOTDOG
AGENT	MAN	MAN
CONTAINER	HAND	MOUTH

(b)

NAGGING	DUSTING	NAGGING
AGENT	MAN	WOMAN
NAGGED PERSON	-	MAN
PLACE	TABLE	ROOM

(c)

DRUMMING	STIRRING	DRUMMING
ITEM	-	DRUM
AGENT	MALE CHILD	MALE CHILD
TOOL	SPOON	DRUMSTICK
PLACE	KITCHEN	KITCHEN

(d)

BEGGING	BOWING	BEGGING
ITEM	-	MONEY
PLACE	SIDEWALK	SIDEWALK
GIVER	-	-
AGENT	WOMAN	MAN

(e)

WEIGHING	MEASURING	WEIGHING
PLACE	LAB	INSIDE
TOOL	-	SCALE
MASS	-	-
AGENT	MAN	MAN

(f)

Figure 6: Qualitative comparison of ClipSitu XTF and CoFormer [28] predictions. green refers to correct prediction while red refers to incorrect prediction. '-' refers to predicting blank (a noun class) for this role.

TURNING	DINING
PLACE	-
TURNEDITEM	-
AGENT	-

SLIDING	HITTING
SLIDER	-
DESTINATION	-
SURFACE	-

RAMMING	RAMMING
PLACE	GYMNASIUM
VICTIM	PEOPLE
AGENT	SPOON

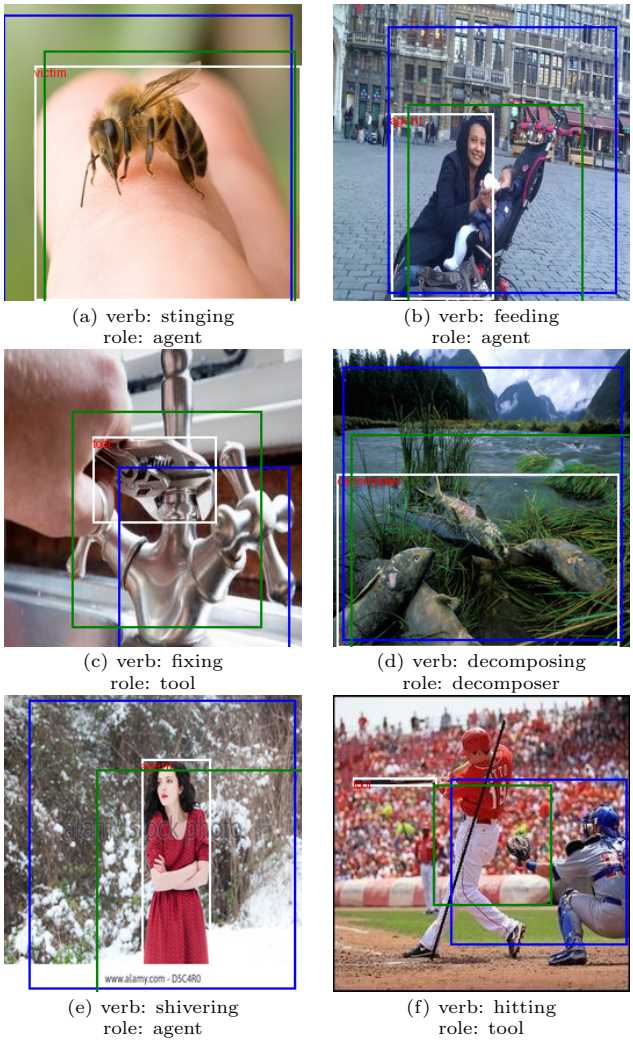
Figure 7: Examples where ClipSitu XTF predicts the verb or noun incorrectly.

8 Situation Recognition with Large Vision Language Model (VLM)

Large vision language models (VLM) are shown to learn new tasks with a few examples provided in context. We study the performance of a contemporary state-of-the-art VLM i.e. VILA [20] on both image and video semantic role labeling when provided in-context examples. We choose VILA as it has been shown to perform better than other vision language models such as LLaVA [67]. Furthermore, it allows in-context examples with interleaved image and text inputs. We also show the effect of completely finetuning an VLM such as VILA on the tasks of image and video situation recognition and the performance improvement obtained after the process.

8.1 Image Situation Recognition with VLM in-context

The prompt sent to VILA with the in-context examples with images as shown in Fig. 11. We fix these 3 in-context examples in the prompt to VILA to generate the nouns across all the imSitu dev and test sets. With this limited context, we provide the image, verb and corresponding roles during testing and only ask VILA to generate the nouns associated with each role following the semantic role labeling experiments with grounded verbs. As shown in Table 10, VILA performs decently on value compared



(a) verb: stinging
role: agent

(b) verb: feeding
role: agent

(c) verb: fixing
role: tool

(d) verb: decomposing
role: decomposer

(e) verb: shivering
role: agent

(f) verb: hitting
role: tool

Figure 8: Examples of noun localization when using verb-role embeddings in blue vs. cross-attention scores in green. In (a) to (e), cross-attention scores localize the noun better than verb-role embeddings with the bounding boxes closer and tighter around the noun (ground-truth bounding box in white). When the noun is very small compared to the image (baseball bat in (f)) it is challenging for both methods to localize the noun. Cross-attention scores from Layer 1 (XAtt. L1 in Table 5)

to the fine-tuned ClipSitu XTF, but in terms of value-all, performance of VILA is poor. This is perhaps because VILA generates semantically similar nouns that do not exactly match the ground-truth nouns as shown in Fig. 12. Therefore, even if a single generated role from VILA matches the ground-truth, all the predictions do not match the ground-truth from any of the 3 annotators leading to poorer value-all performance.

8.2 Instruction-tuning VLM for Image Situation Recognition

Predicting all the roles for an image is quite challenging (value-all) for an VLM such as VILA with in-context and requires additional knowledge of the image-verb-role-noun relations in situational description. To embed this knowledge into VLMs, we fine-tune them using instruction tuning for verb prediction and semantic role labeling. For broader comparison across multiple VLMs, we fine-tune 3 state-of-the-art large visual-language models – LLaVA1.6-mistral-7B [67], MiniCPM-V2.6 [68], and VILA1.5-3B, with the image and provide the verb, roles, and noun as the label. The image and label is provided to all the VLMs in the following instructional form –

`<image>, Task ImSitu: Generate the verb and corresponding roles, then generate the noun based on the verb and roles`
`{"verb": <verb>,`
`"roles": {1: <role1>, 2: <role2>, 3: <role3>, ... }`
`"nouns": { <role1>: <noun1>, <role2>: <noun2>, <role3>: <noun3>, ...} }`

Here, `<image>`, `<verb>`, `<role`i`>`, `<noun`k`>` are placeholders replaced by the image, verb, roles, and nouns. During testing, we only provide the prompt `<image>, Task ImSitu: Generate the verb and corresponding roles, then generate the noun based on the verb and roles` and ask the VLM to generate the text containing verb, roles, and nouns. After collecting the verb and nouns from the generated text, we compare the situation recognition results of fine-tuned VLMs with ClipSitu XTF in Table 11. ClipSitu XTF offers reasonable performance on both verb prediction and semantic role labeling with a fraction of the parameters (45.3M) compared to the VLMs. All fine-tuned VLMs perform only slightly better than ClipSitu XTF at verb prediction which shows that VLMs cannot completely predict the exact situational verb exactly as the human

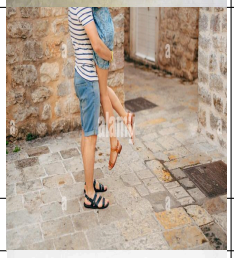
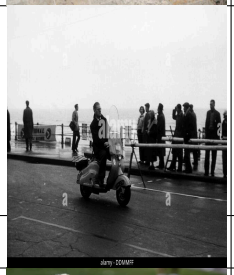
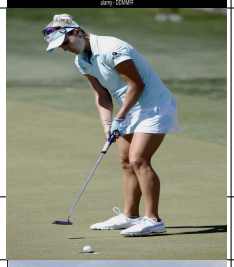
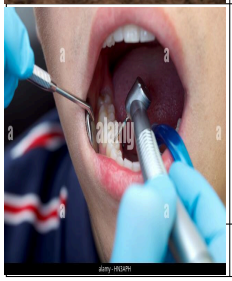
Image	Caption from Conceptual Captions	ClipSitu Situational Summary	Image	Caption from Conceptual Captions	ClipSitu Situational Summary
	beach volleyball player defends against person during the women's main draw.	Verb: Competing Place: Volleyball Court Competition: Agent: Woman		pouring orange juice from a jar into a glass	Verb: overflowing Source: Glass Place: Agent: Water
(a)			(f)		
	racecar driver steers his car during video game subject.	Verb: Racing Place: Racetrack Agent: Car		the groom lifted the bride in his arms	Verb: Embracing Coagent: Woman Place: Outdoors Agent: Man
(b)			(g)		
	a flight was traveling when the animal got free on tuesday night	Verb: Taxiing Place: Airport Agent: Airplane Ground: Runaway		a man on a scooter, taking part in the rally	Verb: Racing Place: Road Agent: Motorcycle
(c)			(h)		
	farm tractor is moving on the field, cultivating land	Verb: Hoeing Place: Farm Agent: Person Ground: Field		putts for birdie on the second hole during the final round of the golf tournament .	Verb: Competiting Place: Golf Competition: Golf Course Agent: Woman
(d)			(i)		
	the dentist drill the tooth with a turbine	Verb: Injecting Source: Syringe Destination: Mouth Substance: Liquid Place: Agent: Person		a land armoured vehicle patrols the highway .	Verb: Guarding Item: Road Weapon: Gun Place: Outdoors Agent: Soldier
(e)			(j)		

Figure 9: Examples of ClipSitu situational summaries on out-of-domain action-oriented images from Conceptual Captions [31]. Compared to the original captions, situational summaries of ClipSitu concisely describes relevant verbs, associated roles and the nouns playing those roles. For certain images, such as (f) "overflowing", (h) "racing", ClipSitu tries to force a verb even when a faint or no relation to the action exists. In some cases, ClipSitu can return null for some roles such as (e), (f).

Figure 10: "buying" or "selling"? GT: buying

Method	dev		test	
	value	value-all	value	value-all
VILA [20] (in-context)	67.44	05.03	67.30	05.17
ClipSitu XTF	**78.52**	**45.31**	**78.52**	**45.31**

Table 10: Comparison of large vision language model (VILA) with in-context examples to ClipSitu XTF on image semantic role labeling in imSitu.

annotators but predicts a semantically similar verb. On single role prediction, all the VLMs perform almost twice as well as XTF demonstrating the prowess of VLMs at identifying at least a single role correctly. However, predicting the nouns for all roles of a verb correctly is still challenging for VLMs as shown by the performance on value-all. VLMs do not drastically outperform ClipSituXTF for verb prediction and value-all as they generate verbs and nouns that are semantically similar to the ground-truth nouns for each role (as shown for VILA in Fig. 12) but not the exact ground-truth nouns which leads to low value-all performance. Therefore, it is difficult to justifiably judge VLM performance on the value-all metric that is not flexible to account for semantic similarity.

Given the image, provide a structured situational summary following the JSON response format template, that captures the essence of the scene. If an action or activity is taking place, describe the action (verb). Then, for each role associated with this verb (action), provide the label (noun) filling these roles.

JSON Response Format Template:

```
{{
  "Verb": "{verb}",
  "Roles_and_Nouns": {{
    {roles_json}
  }}
}}
```

In-Context Examples for Guidance:

<image>

```
{{
  "Verb": "Prowling",
  "Roles_and_Nouns": {{
    "Place": ["Outdoors"],
    "Target": ["Prey"],
    "Agent": ["Wolf"]
  }}
}}
```

<image>

```
{{
  "Verb": "Spraying",
  "Roles_and_Nouns": {{
    "Destination": ["Tree"],
    "Substance": ["Water"],
    "Place": ["Forest"],
    "Agent": ["Person"],
    "Source": ["Hose"]
  }}
}}
```

<image>

```
{{
  "Verb": "Slipping",
  "Roles_and_Nouns": {{
    "Destination": ["Base"],
    "Place": ["Playing Field"],
    "Agent": ["Player"]
  }}
}}
```

Now, based on the image <image>, verb {verb} and roles {roles_list}, fill in the structured summary below and include the label (noun) that fills each role.

Figure 11: Prompt template with constrained JSON output for each imSitu image with 3 fixed in-context examples. Images are appended after the prompt but shown together with the prompt here for clarity. Best viewed on a screen.

Verb	Image	VILA Output	Annotator 1	Annotator 2	Annotator 3
flapping		Bodypart: **Wings** Place: **Air** Agent: **Bird**	bodypart: **wing** place: **outdoors** agent: **bird**	bodypart: **wing** place: **lake** agent: **bird**	bodypart: **wings** place: **sky** agent: **bird**
grieving		Place: **Cemetery** Agent: **People**	place: **funeral** agent: **family**	place: **funeral** agent: **mourner**	place: **funeral** agent: **people**
molding		Place: **Pottery wheel** Substance: **Clay** Goalitem: **Pottery piece** Agent: **Potter**	place: **workshop** substance: **mud** goalitem: **sphere** agent: **potter**	place: **workshop** substance: **mud** goalitem: **sphere** agent: **person**	place: **workroom** substance: **mud** goalitem: **bowl** agent: **person**
skiing		Place: **Snowy Mountains** Agent: **Skier**	place: **ski slope** agent: **skier**	place: **slope** agent: **skier**	place: **ski run** agent: **skier**

Figure 12: Qualitative examples comparing the generated nouns from VILA with the ground-truth nouns for all 3 annotators in imSitu dataset. VILA generates semantically similar nouns but not the same nouns leading to lower value-all performance. Correct nouns in green while incorrect are shown in red.

Method	Top-1 predicted verb dev/test		
	verb	value	value-all
LLaVA [67] (instruction-tuned)	58.77/59.25	94.29/94.43	**30.63/30.61**
MiniCPM-V [68] (instruction-tuned)	57.37/58.04	94.93/95.02	30.02/30.36
VILA [20] (instruction-tuned)	**58.90/59.32**	**95.00/95.11**	26.36/26.44
ClipSituXTF	58.19/58.09	47.23/47.21	29.73/29.61

Table 11: Comparing instruction-tuned VLMs to ClipSitu XTF on image situation recognition in imSitu.

Method	CIDEr-Arg
VILA [20] (in-context)	31.97
ClipSituMLP+TxD	**56.30**

Table 12: Comparing VILA with in-context examples versus ClipSituMLP+TxD on Video Semantic Role Labeling in VidSitu.

8.3 Video SRL with VLM in-context

We also test the performance of large vision language model VILA [20] on VidSitu for semantic role labeling of video events. As VILA accepts only images as in-context examples, we use the strategy of using 2 representative images per event (extracted at 1 frame per second for every 2 second event) as adopted for object extraction in [42]. Similar to imSitu, we provide the image, verb and roles to VILA and ask it to generate role-wise nouns. From every image belonging to an event, we aggregate a list of nouns for every role. We compute the generation metrics in Table 12 by comparing the aggregated list with the ground-truth annotations of 3 annotators per event in VidSitu [30]. The CIDEr metrics of VILA with in-context examples in Table 12 show that it is not able to generalize its understanding of the variety of verbs and their corresponding roles as well as a fine-tuned ClipSitu MLP+TxD model. Another reason is that the events are temporal in nature and providing 2 frames per event might not be sufficient to visually describe the event.

8.4 Instruction-tuning VLMs for Video Situation Recognition

To overcome these limitations of in-context examples, we opt to fine-tune 4 state-of-the-art VLMs (Video LLMs) for the video situation recognition task (verb prediction and semantic role labeling) – VILA [20], MiniCPM-V [68], Qwen2-VL [69], and LLaVA-Video [70]. Specifically, we fine-tune the VILA1.5-3B, MiniCPM-V2.6, Qwen2-VL-7B, LLaVA-Video7B models using instructions by providing the video, verb, roles, and nouns. Our instruction for the video situation recognition task is as follows –

```
<video>, Task VidSitu: Generate the verb and
corresponding roles, then generate the noun
based on the verb and roles
{"verb": <verb>,
"roles": {1: <role1>, 2: <role2>, 3:
```

Method	CIDEr	CIDEr-Vb	CIDEr-Arg	Rouge-L
VILA [20] (instruction-tuned)	40.44	81.68	65.07	41.80
MiniCPM-V [68] (instruction-tuned)	53.04	113.44	77.53	42.33
Qwen2-VL[69] (instruction-tuned)	57.28	127.34	76.92	41.58
LLaVA-Video [70] (instruction-tuned)	60.10	**133.77**	**78.04**	42.81
ClipSituMLP+TxD	**61.93**	70.14	56.30	**43.77**

Table 13: Comparing instruction-tuned VLMs (Video LLMs) to ClipSituMLP+TxD on video situation recognition (VidSitu).

`<role3>, ... }`
`"nouns": { <role1>: <noun1>, <role2>: <noun2>, <role3>: <noun3>, ...} }`
Here,`<video>, <verb>, <role`*i*`>, <noun`*k*`>` are placeholders replaced by the video, verb, roles, and nouns. Similar to imSitu finetuning, we only provide the prompt `<video>, Task VidSitu: Generate the verb and corresponding roles, then generate the noun based on the verb and roles` and ask the VLM to generate the text containing verb, roles, and nouns. After collecting the verb and nouns from the generated text, we compare the situation recognition results of instruction-tuned video LLMs with our best model from Table 9 - ClipSituMLP+TxD in Table 13. ClipSituMLP+TxD being a lightweight model performs quite well on all the metrics compared to existing models but video LLMs with world knowledge and a large number of parameters are able to better describe the verb and nouns on the video. Strangely, on overall CIDEr and Rouge-L, VideoLLMs do not perform as well. Overall CiDEr in VidSitu is evaluated by inserting the predicted verbs and nouns in the template set in [30] i.e. "verb ¡verb name¿ Arg0 ¡noun1 name¿ Arg1 ¡ noun2 name¿ ..." and matching with the ground truth sentence. CIDEr is designed to be evaluated with multiple reference sentences but here we only have a single ground truth sentence. When evaluating long sentences, deviation of a single predicted noun from the ground truth noun is severely penalized which leads to lower overall Rouge-L. Therefore, overall CIDEr and Rouge-L is not a reliable metric for evaluating the video LLMs and CIDEr-Vb and CIDEr-Arg should be considered as the true performance of the video LLMs.

9 Conclusion

In conclusion, our work on ClipSitu represents a significant advancement in situation recognition for both images and videos. By leveraging the powerful capabilities of CLIP-based image embeddings, we have developed an end-to-end framework that excels in predicting semantic roles associated with verbs, thereby generating comprehensive situational summaries even for out of domain images. Our approach enhances the accuracy and contextual relevance of the generated situational summaries of images and videos as shown through extensive experimentation. Furthermore, ClipSitu's integration of a transformer decoder to generate nouns as text phrases from video frames highlights its versatility and robustness. This extension is crucial for applications requiring detailed and structured descriptions of dynamic scenes, such as multimedia retrieval and event monitoring. Our results demonstrate the potential of ClipSitu to serve as a foundational tool in visual media understanding, offering rich, action-centric descriptions that can be evaluated for correctness and integrated with external knowledge bases for enhanced reasoning and decision-making.

In future, there are several promising directions for future research. One area of interest is the enhancement of ClipSitu's ability to handle more complex and nuanced interactions within images and videos, potentially through the incorporation of additional contextual information and multimodal data. Another promising directions is exploring the application of ClipSitu in specialized domains such as medical imaging or autonomous driving could yield valuable insights and improvements in those fields. Finally, we aim to refine the model's scalability and efficiency, ensuring it can be effectively deployed in real-world scenarios with large-scale datasets and diverse visual content.

Acknowledgments This research/project is supported by the National Research Foundation, Singapore, under its NRF Fellowship (Award# NRF-NRFF14-2022-0001). This research is also supported by funding allocation to Basura Fernando by the Agency for Science, Technology and Research (A*STAR) under its SERC Central Research Fund (CRF), as well as its Centre for Frontier AI Research (CFAR).

References

[1] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5534–5542, 2016.

[2] John McCarthy et al. *Situations, actions, and causal laws*. Comtex Scientific, 1963.

[3] John McCarthy and Patrick J Hayes. Some philosophical problems from the standpoint of artificial intelligence. In *Readings in artificial intelligence*, pages 431–450. Elsevier, 1981.

[4] Raymond Reiter. *Knowledge in action: logical foundations for specifying and implementing dynamical systems*. MIT press, 2001.

[5] Robert Kowalski and Marek Sergot. A logic-based calculus of events. *New generation computing*, 4:67–95, 1986.

[6] Xilong Liu, Zhiqiang Cao, Yingying Yu, Guangli Ren, Junzhi Yu, and Min Tan. Robot navigation based on situational awareness. *IEEE Transactions on Cognitive and Developmental Systems*, 14(3):869–881, 2021.

[7] Srikanth Malla, Behzad Dariush, and Chiho Choi. Titan: Future forecast using action priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11186–11196, 2020.

[8] Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, and Zhenzhen Jiao. Towards surveillance video-and-language understanding: New dataset baselines and challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22052–22061, 2024.

[9] Sarvesh Vishwakarma and Anupam Agrawal. A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, 29:983–1009, 2013.

[10] Yuya Chiba and Ryuichiro Higashinaka. Dialogue situation recognition for everyday conversation using multimodal information. In *Interspeech*, pages 241–245, 2021.

[11] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1969–1978, 2019.

[12] Yuke Zhu, Daniel Gordon, Eric Kolve, Dieter Fox, Li Fei-Fei, Abhinav Gupta, Roozbeh Mottaghi, and Ali Farhadi. Visual semantic planning using deep successor representations. In *Proceedings of the IEEE international conference on computer vision*, pages 483–492, 2017.

[13] Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B Tenenbaum, Sanja Fidler, and Antonio Torralba. Watch-and-help: A challenge for social perception and human-ai collaboration. In *International Conference on Learning Representations*, 2020.

[14] Yanming Wan, Jiayuan Mao, and Josh Tenenbaum. Handmethat: Human-robot communication in physical and social environments. *Advances in Neural Information Processing Systems*, 35:12014–12026, 2022.

[15] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 852–869. Springer, 2016.

[16] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018.

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

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan

Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[19] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[20] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pretraining for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26689–26699, June 2024.

[21] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[22] Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2668, 2023.

[23] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, 2022.

[24] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. In *The Eleventh International Conference on Learning Representations*, 2022.

[25] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 388–404. Springer, 2022.

[26] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022.

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

[28] Junhyeong Cho, Youngseok Yoon, and Suha Kwak. Collaborative transformers for grounded situation recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19659–19668, 2022.

[29] Debaditya Roy, Dhruv Verma, and Basura Fernando. Clipsitu: Effectively leveraging clip for conditional predictions in situation recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 444–453, 2024.

[30] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5600, 2021.

[31] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[32] Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, and Ali Farhadi. Commonly uncommon: Semantic sparsity in situation recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7196–7205, 2017.

[33] Ruiyu Li, Makarand Tapaswi, Renjie Liao, Jiaya Jia, Raquel Urtasun, and Sanja Fidler. Situation recognition with graph neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4173–4182, 2017.

[34] Mohammed Suhail and Leonid Sigal. Mixture-kernel graph attention network for situation recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10363–10372, 2019.

[35] Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5281–5291, 2023.

[36] Arun Mallya and Svetlana Lazebnik. Recurrent models for situation recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 455–463, 2017.

[37] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 314–332. Springer, 2020.

[38] Thilini Cooray, Ngai-Man Cheung, and Wei Lu. Attention-based context aware reasoning for situation recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4736–4745, 2020.

[39] Junhyeong Cho, Youngseok Yoon, Hyeonjun Lee, and Suha Kwak. Grounded situation recognition with transformers. In *British Machine Vision Conference (BMVC)*, 2021.

[40] Meng Wei, Long Chen, Wei Ji, Xiaoyu Yue, and Tat-Seng Chua. Rethinking the two-stage framework for grounded situation recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2651–2658, 2022.

[41] Tianyu Jiang and Ellen Riloff. Exploiting commonsense knowledge about objects for visual activity recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7277–7285, 2023.

[42] Zeeshan Khan, CV Jawahar, and Makarand Tapaswi. Grounded video situation recognition. *Advances in Neural Information Processing Systems*, 35:8199–8210, 2022.

[43] Fanyi Xiao, Kaustav Kundu, Joseph Tighe, and Davide Modolo. Hierarchical self-supervised representation learning for movie understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9727–9736, 2022.

[44] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[45] Guang Yang, Manling Li, Jiajie Zhang, Xudong Lin, Heng Ji, and Shih-Fu Chang. Video event extraction via tracking visual states of arguments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3136–3144, 2023.

[46] Junhua Jia, Xiangqian Ding, Shunpeng Pang, Xiaoyan Gao, Xiaowei Xin, Ruotong Hu, and Jie Nie. Image captioning based on scene graphs: a survey. *Expert Systems with Applications*, 231:120698, 2023.

[47] Kien Nguyen, Subarna Tripathi, Bang Du, Tanaya Guha, and Truong Q Nguyen. In defense of scene graphs for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1407–1416, 2021.

[48] Xu Yang, Jiawei Peng, Zihua Wang, Haiyang Xu, Qinghao Ye, Chenliang Li, Songfang Huang, Fei Huang, Zhangzikang Li, and Yu Zhang. Transforming visual scene graphs to image captions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12427–12440, 2023.

[49] Hanhua Ye, Guorong Li, Yuankai Qi, Shuhui Wang, Qingming Huang, and Ming-Hsuan Yang. Hierarchical modular network for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17939–17948, 2022.

[50] Xin Gu, Guang Chen, Yufei Wang, Libo Zhang, Tiejian Luo, and Longyin Wen. Text with knowledge graph augmented transformer for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18941–18951, 2023.

[51] Guorong Li, Hanhua Ye, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. Learning hierarchical modular networks for video captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[52] Dalin Wang, Daniel Beck, and Trevor Cohn. On the role of scene graphs in image captioning. In Aditya Mogadala, Dietrich Klakow, Sandro Pezzelle, and Marie-Francine Moens, editors, *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 29–34, Hong Kong, China, November 2019. Association for Computational Linguistics.

[53] Ting Wang, Weidong Chen, Yuanhe Tian, Yan Song, and Zhendong Mao. Improving image captioning via predicting structured concepts. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 360–370, Singapore, December 2023. Association for Computational Linguistics.

[54] Yifan Lu, Ziqi Zhang, Chunfeng Yuan, Peng Li, Yan Wang, Bing Li, and Weiming Hu. Set prediction guided by semantic concepts for diverse video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3909–3917, 2024.

[55] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16372–16382, 2021.

[56] Jongha Kim, Jihwan Park, Jinyoung Park, Jinyoung Kim, Sehyung Kim, and Hyunwoo J Kim. Group-wise query specialization and quality-aware multi-assignment for transformer-based visual relationship detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28160–28169, 2024.

[57] Hongsheng Li, Guangming Zhu, Liang Zhang, Youliang Jiang, Yixuan Dang, Haoran Hou, Peiyi Shen, Xia Zhao, Syed Afaq Ali Shah, and Mohammed Bennamoun. Scene graph generation: A comprehensive survey. *Neurocomputing*, 566:127052, 2024.

[58] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[59] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

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

[61] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022.

[62] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[63] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[64] Nafise Sadat Moosavi and Michael Strube. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th annual meeting of the association for computational linguistics*, volume 1, pages 632–642. Association for Computational Linguistics, 2016.

[65] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word

representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[66] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[67] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[68] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *CoRR*, 2024.

[69] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[70] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.