# Supplementary Material – ClipSitu: Effectively Leveraging CLIP for Conditional Predictions in Situation Recognition

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

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

## 1. Learned vs. Fixed role tokens for noun prediction

We show in Table 1 that ClipSitu XTF performs worse with learned role tokens for each verb when compared to using fixed role tokens obtained from CLIP.

| ClipSitu XTF | Top-1 | | Top-5 | | Ground truth | |
|---|---|---|---|---|---|---|
| | value | v-all | value | v-all | value | v-all |
| Learnable Role Tokens | 44.82 | 25.44 | 65.04 | 35.37 | 75.62 | 44.31 |
| Fixed Role Tokens | **47.17** | **30.06** | **68.44** | **41.66** | **78.49** | **45.81** |

Table 1. Comparing the performance of noun prediction with fixed and learnable role tokens per verb in ClipSitu XTF.

## 2. More Qualitative Results

In Fig. 1, we show some qualitative examples where ClipSitu Verb MLP predicts the verb incorrectly or ClipSitu XTF predicts the noun incorrectly.

## 3. Complexity analysis

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 [1] in **??**. 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.

| TURNING | DINING |
|---|---|
| PLACE | - |
| TURNEDITEM | - |
| AGENT | - |

| SLIDING | HITTING |
|---|---|
| SLIDER | - |
| DESTINATION | - |
| SURFACE | - |

| RAMMING | RAMMING |
|---|---|
| PLACE | GYMNASIUM |
| VICTIM | PEOPLE |
| AGENT | SPOON |

Figure 1. Examples where ClipSitu predicts the verb/noun incorrectly

## References

[1] 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.