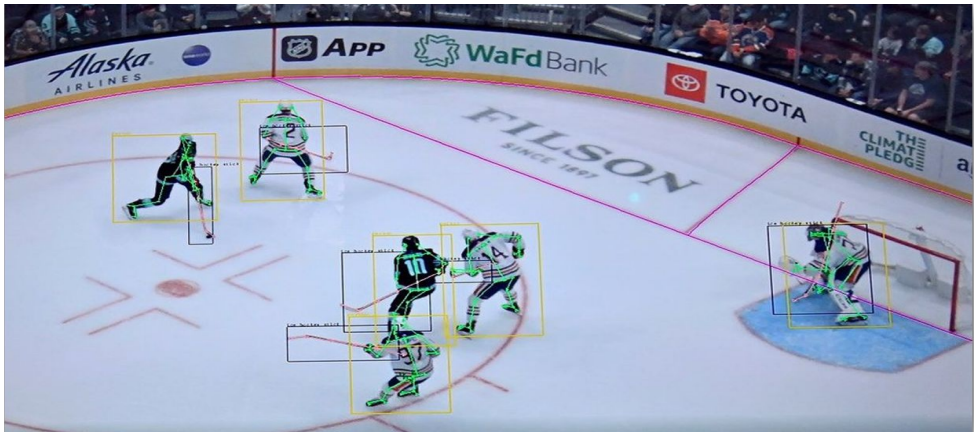
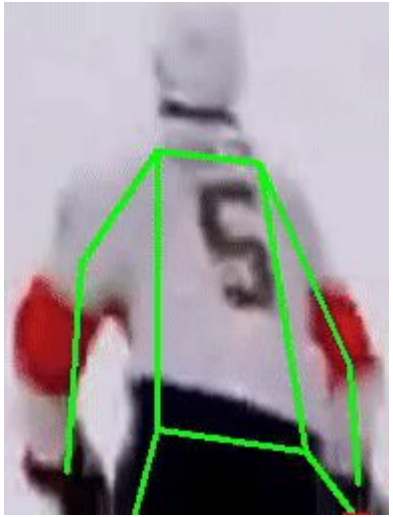
Seeing Beyond the Crop: Using Language Priors for Out-of-Bounding Box Keypoint Prediction

Bavesh Balaji, Jerrin Bright, Yuhao Chen, John Zelek,
Sirisha Rambhatla and David Clausi,
University of Waterloo

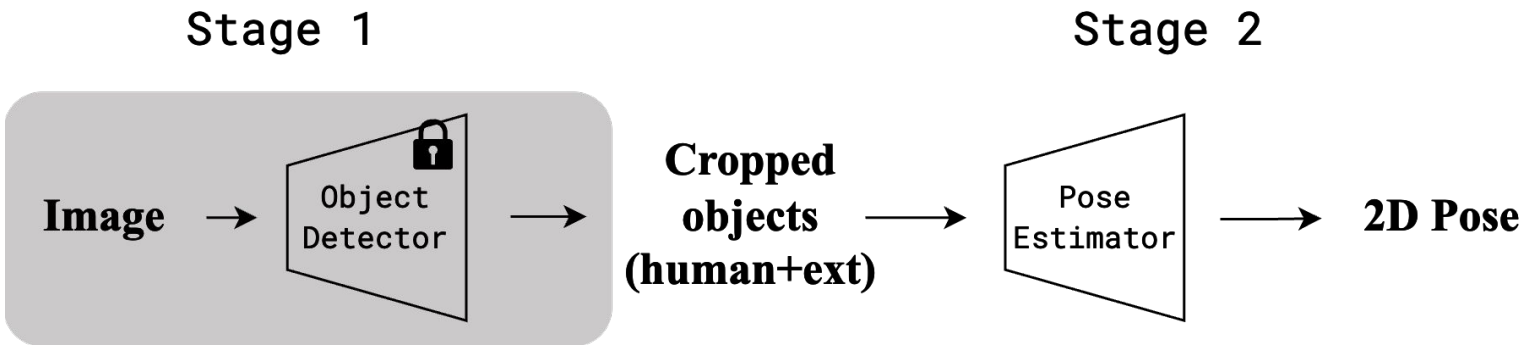
INTRODUCTION

- **Objective:** To jointly predict the pose of a human and the object that they hold and interact with, aka extensions.
- Essential for scene understanding, action recognition and HOI detection.





EXISTING WORKS



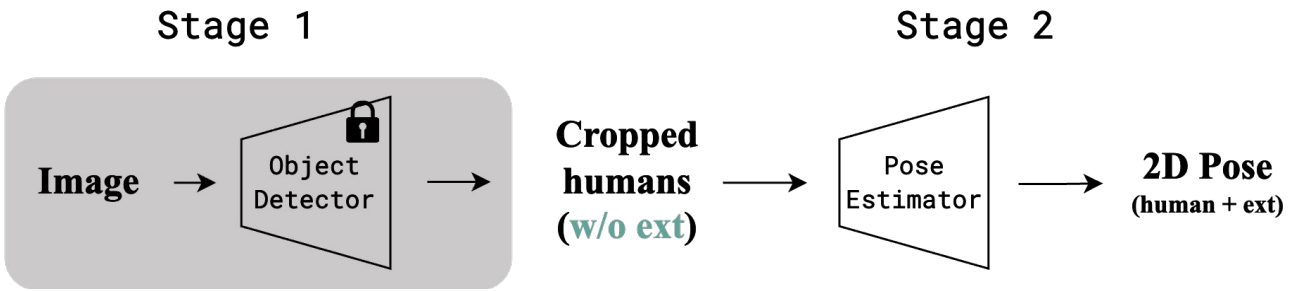
- Predominantly follow top-down paradigm
  - Crop and isolate all the entities (humans along with their extension) in the image.
  - Perform pose estimation on each entity separately.

LIMITATIONS

- Inclusion of extension introduces **noisy features**!
  - Expects all keypoints of interest to be present within the bounding box.
  - Necessitates **larger bounding boxes,** leading to more noise.
- Pose estimators not tuned to handle multiple objects of interest in a single cropped image.

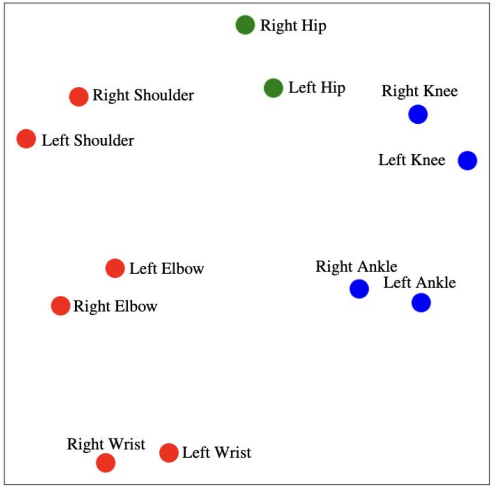
OUR CONTRIBUTIONS

- **Do not capture the extension** in the bounding box!
  - Works (Yao *et al*. 2010 & Neher *et al*. 2018) on human-object interaction identify a strong correlation between human and extension pose.
  - We leverage that to infer extension pose from humans.
  - Reformulate extension pose estimation to unseen keypoint prediction.
- *How can we effectively represent the spatial relations of these unseen keypoints?*



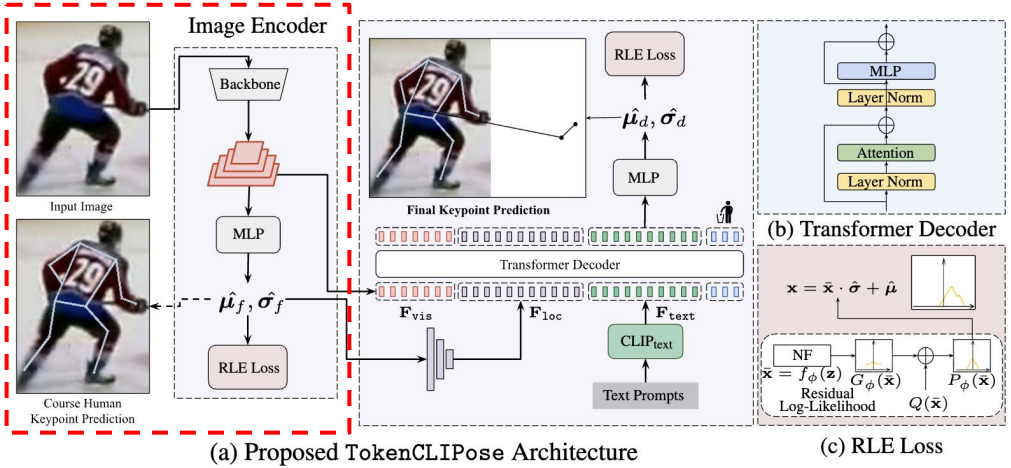
OUR CONTRIBUTIONS

- We leverage **language** to inform our model.
  - Keypoint-specific text prompts are utilized to generate text embeddings .
- Existing multimodal pose estimators align the image features to frozen text embeddings.
  - This is not desirable as text embeddings **do not maintain global positional structure**!
- We create learnable tokens for each keypoint and initialize them with these embeddings.



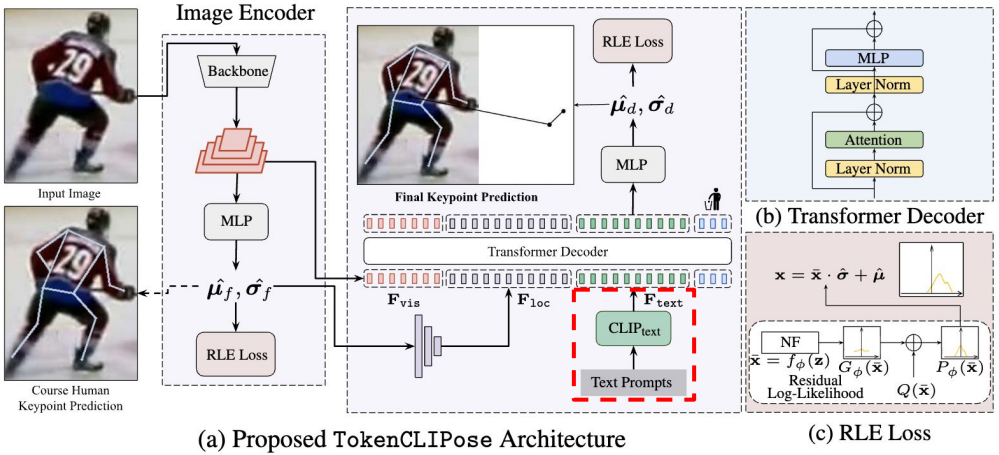
t-SNE plot of CLIP joint embeddings. Upper-body joints (red) and lower-body joints (blue) have no positional correlation among each other.

METHODOLOGY



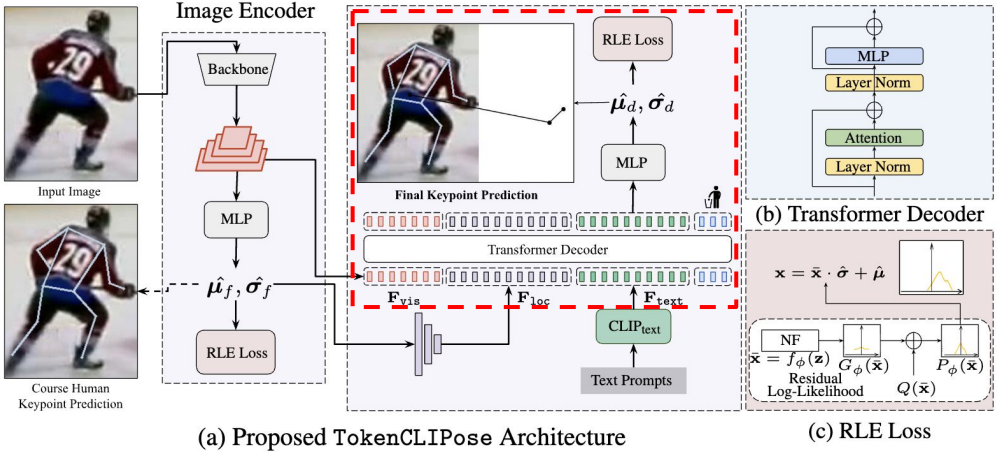
- The cropped image of a human is fed to an image encoder to extract multi-scale feature maps $\mathbf{F}_{\texttt{vis}}$.
- These feature maps are used to predict coarse human keypoint locations $\hat{\boldsymbol{\mu}}_d \in \mathbb{R}^{K_h \times 1}$, which are then reprojected to form location tokens $\mathbf{F}_{\texttt{loc}}$.

METHODOLOGY



- Keypoint-specific text prompts are passed through the CLIP text encoder and projected onto a joint multimodal embedding space to obtain text-initialized keypoint tokens $\mathbf{F}_{\texttt{text}}$.

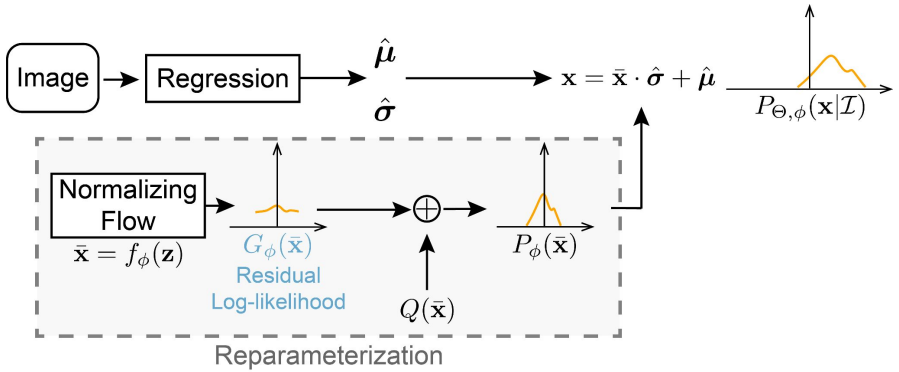
METHODOLOGY



- The visual, location and text tokens are concatenated to produce multimodal tokens $\mathcal{F} = \{\mathbf{F}_{\texttt{vis}}, \mathbf{F}_{\texttt{loc}}, \mathbf{F}_{\texttt{text}}\}$.
- These tokens are then fed to a vision transformer to understand the correlations among the unimodal tokens and in between tokens of different modality.
- The output of the transformer is used to extract final keypoint coordinates $\hat{\mu}_d \in \mathbb{R}^{K_{out} \times 2}$ and scale parameter $\hat{\sigma}_d \in \mathbb{R}^{K_{out} \times 1}$

RLE LOSS

- Standard loss functions like ℓ_2 are more vulnerable to noisy inputs.
- Formulate regression as a distribution learning task.
- Leverage normalizing flows to calculate the deviation between predicted values and ground truth values.
- Adopt MLE to minimize the deviation.
- $$\begin{aligned}\mathcal{L}_{RLE} &= -\log P_{\Theta,\phi}(\mathbf{x}|\mathcal{I})\Big|_{\mathbf{x}=\boldsymbol{\mu}_g} \\ &= -\log P_{\phi}(\bar{\boldsymbol{\mu}}_g) + \log \hat{\boldsymbol{\sigma}} \\ &= -\log Q(\bar{\boldsymbol{\mu}}_g) - \log G_{\phi}(\bar{\boldsymbol{\mu}}_g) - \log s + \log \hat{\boldsymbol{\sigma}}.\end{aligned}$$



DATASETS

- Introduce 2 new datasets for extension pose estimation.
  - Ice hockey dataset collected from 10 NHL games.
  - Lacrosse dataset for zero-shot generalization capabilities.
- We also evaluate on CrowdPose to demonstrate the flexibility of our model.

RESULTS

Table 1: **Comparison with SOTA Methods** on our real-world ice hockey dataset (PCKh@0.5). **BoldFace** represents the best score. Underline represents the top score in existing works.

Method	Backbone	Input Resolution	Body	Butt End	Stick Heel	Stick Toe	Mean
SimpleBaseline [13]	ResNet-50	256x192	93.59	69.57	57.19	52.76	68.83
MSPN [9]	-	256x192	93.61	70.30	59.21	55.69	69.70
HR-Net [12]	HRNet-W48	256x192	94.90	71.48	60.29	55.36	70.44
TokenPose-L/D24 [11]	HRNet-W48	256x192	95.13	70.96	60.93	56.27	70.82
ViTPose [10]	ViT-B	256x192	95.61	71.94	61.33	58.80	71.92
TokenCLIPose	ResNet-50	256x192	95.81	74.86	65.79	65.08	74.92
TokenCLIPose	MSPN	256x192	97.17	75.41	66.70	66.34	75.53
TokenCLIPose	HRNet-W48	256x192	**97.37**	**75.94**	**67.82**	**66.15**	**76.28**
Improvement	-	-	**1.76% ↑**	**4.00% ↑**	**6.49% ↑**	**7.35% ↑**	**4.36% ↑**

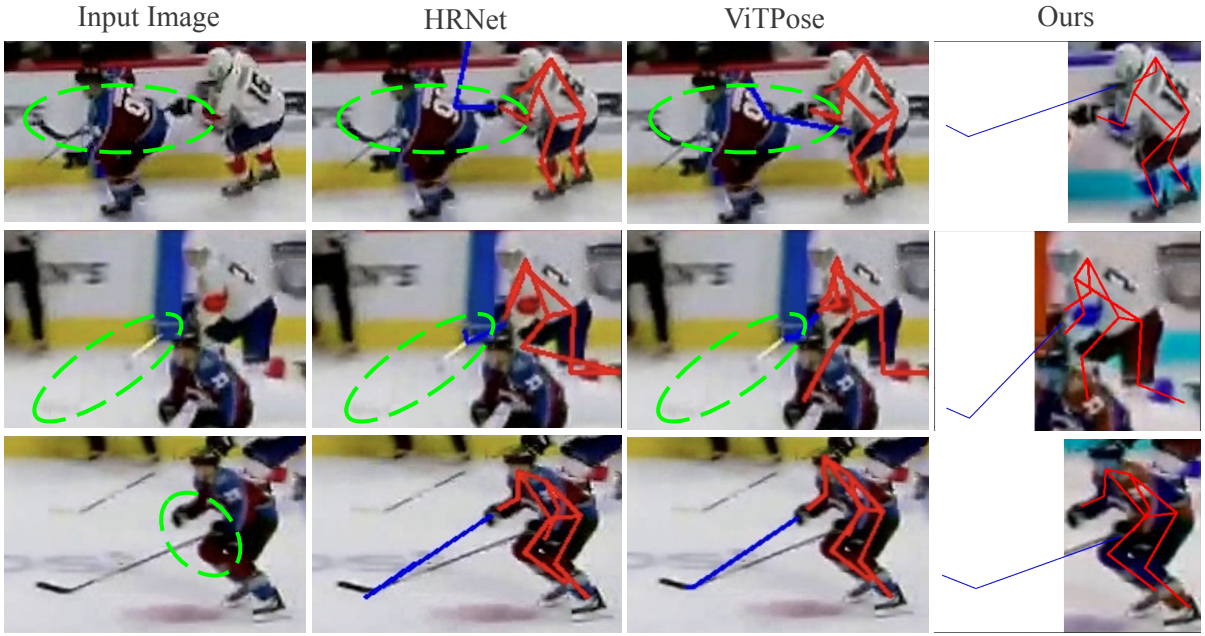
Table 2: **Zero-shot Comparison with SOTA Methods** on our real-world Lacrosse dataset (PCKh@0.5). **BoldFace** represents the best score. Underline represents the second-best score.

Method	Backbone	Body	Butt End	Stick Heel	Mean
SimpleBaseline [13]	ResNet-50	94.73	67.28	53.99	72.00
MSPN [9]	-	95.84	70.68	57.40	74.64
HR-Net [12]	HRNet-W48	95.92	71.35	58.41	75.22
ViTPose [10]	ViT-B	95.77	72.85	60.18	76.26
TokenCLIPose	HRNet-W48	**97.24**	**76.60**	**65.01**	**78.61**
Improvement	-	**1.47% ↑**	**3.75% ↑**	**4.83% ↑**	**2.35% ↑**

Table 3: **Comparison with SOTA Methods** on CrowdPose dataset. **BoldFace** represents the best score. Underline represents the second-best score.

Method	Input Resolution	AP	AP_{50}	AP_{75}	AP_E	AP_M	AP_H
Mask-RCNN [43]	256 × 192	57.2	83.5	60.3	69.4	57.9	45.8
AlphaPose	256 × 192	61.0	81.3	66.0	71.2	61.4	51.1
SimpleBaseline [13]	256 × 192	60.8	81.4	65.7	71.4	61.2	51.2
CrowdPose [6]	256 × 192	66.0	84.2	71.5	75.5	66.3	57.4
Hourglass-104 [44]	384 × 288	65.2	85.9	69.5	-	-	-
KAPAO-L [45]	384 × 288	68.9	89.4	75.6	76.6	69.9	59.5
HRNet-W48 [12]	384 × 288	69.3	89.7	75.6	77.7	70.6	57.8
Transpose-H [22]	384 × 288	71.8	91.5	77.8	79.5	72.9	62.2
HRFormer-B [23]	384 × 288	72.4	91.5	77.9	80.0	73.5	62.4
TokenCLIPose	384 × 288	**76.2**	**93.9**	**82.4**	**83.3**	**77.4**	**66.1**
Improvement	-	**3.8% ↑**	**2.4% ↑**	**4.5% ↑**	**3.3% ↑**	**3.9% ↑**	**3.7% ↑**

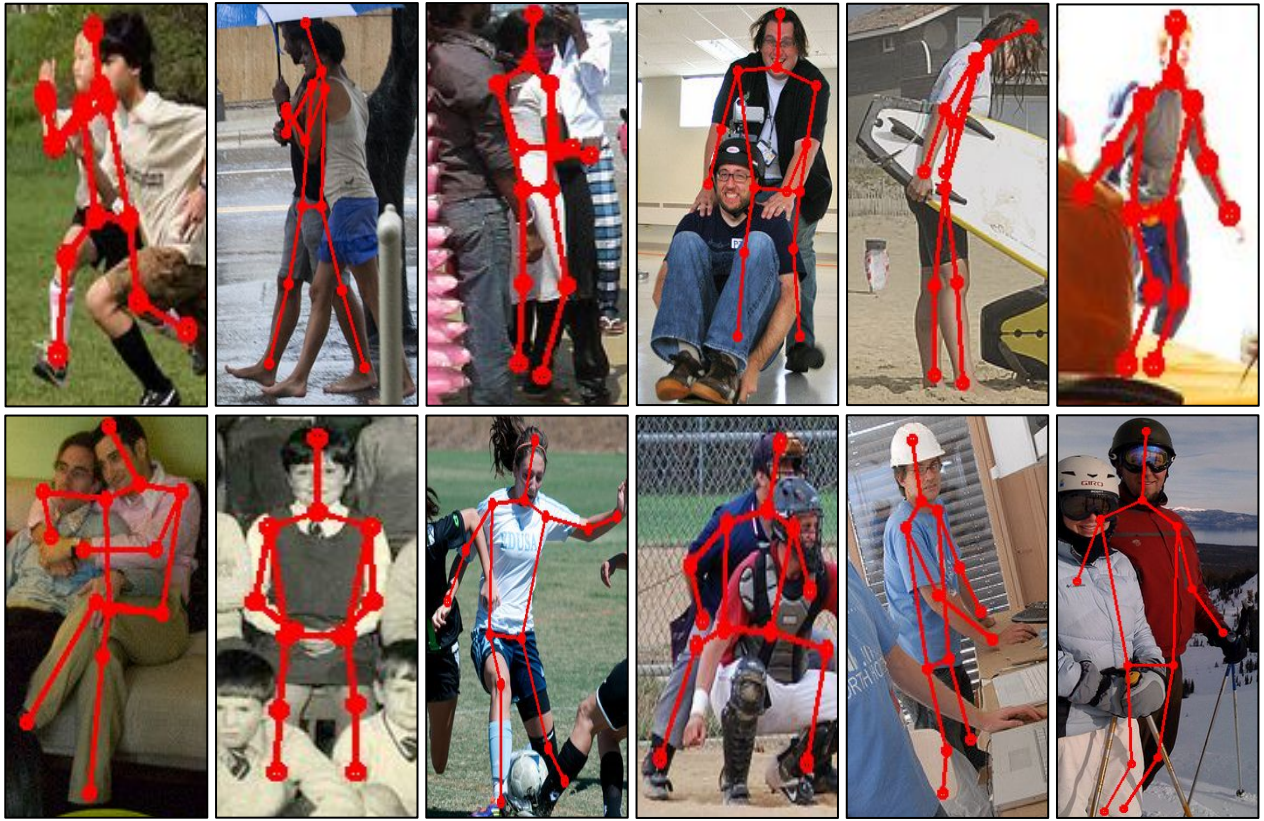
RESULTS



Ice hockey

Lacrosse

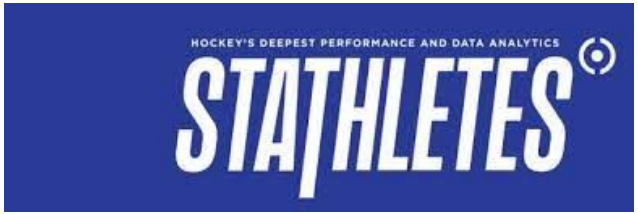
RESULTS



CONCLUSION

- **Out-of-bounding box prediction**: We reformulate the extension pose estimation problem as an unseen keypoint prediction problem and facilitate out-of-bounding box keypoint prediction.
- **Multimodal pose estimation**: We leverage the power of VLMs to augment the spatial relationship of keypoints. Furthermore, we showcase that conventional method of aligning image features to text embeddings is not optimal.
- **Significant improvement on SOTA**: We consistently outperform the existing state-of-the-art by 4.36%, 2.35% and 3.8% on the Ice Hockey, Lacrosse and CrowdPose datasets respectively, underscoring the impact of reducing noise for pose estimation.

ACKNOWLEDGEMENT



Thank You!