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

Bavesh Balaji[1,2,3], **Jerrin Bright**[2,3], **Yuhao Chen**[2,3], **Sirisha Rambhatla**[1,3], **John S. Zelek**[2,3], **David Anthony Clausi**[2,3],
[1]Critical ML Lab, [2]Vision and Image Processing Lab, [3]University of Waterloo,

Abstract

Accurate estimation of human pose and the pose of interacting objects, like a hockey stick, is crucial for action recognition and performance analysis, particularly in sports. Existing methods capture the object along with the human in the bounding boxes, assuming all keypoints are visible within the bounding box. This necessitates larger bounding boxes to capture the object, introducing unnecessary visual features and hindering performance in real-world cluttered environments. We propose a simple image and text-based multimodal solution `TokenCLIPose` that addresses this limitation. Our approach focuses solely on human keypoints within the bounding box, treating objects as *unseen*. `TokenCLIPose` leverages the rich semantic representations endowed by language for inducing keypoint-specific context, even for occluded keypoints. We evaluate the performance of `TokenCLIPose` on a real-world ice hockey dataset, and demonstrate its generalizability through zero-shot transfer to a smaller Lacrosse dataset. Additionally, we showcase its flexibility on CrowdPose, a popular occlusion benchmark with keypoints within the bounding box. Our method significantly improves over state-of-the-art approaches on ice hockey, Lacrosse, and CrowdPose datasets, with gains of 4.36%, 2.35%, and 3.8%, respectively.

1 Introduction

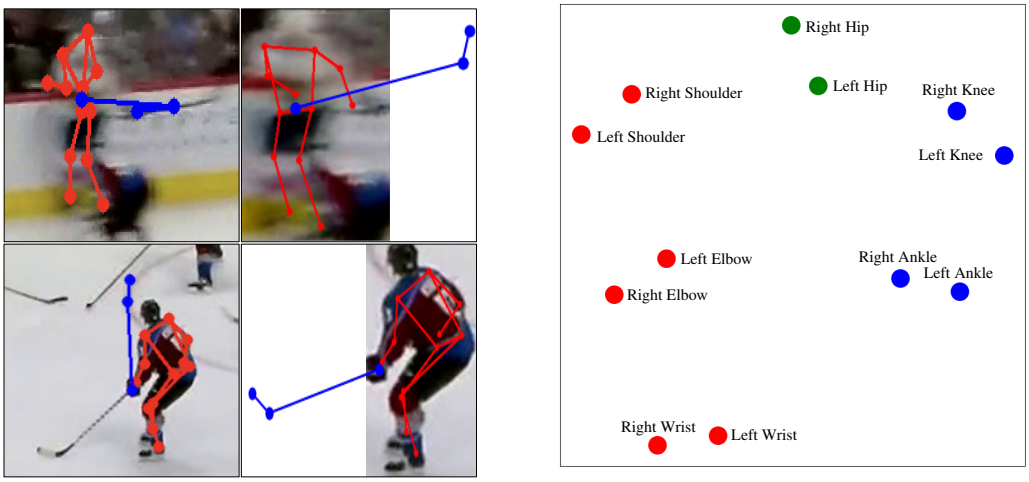
The goal of 2D human pose estimation is to localize the human anatomical keypoints from an image, which is essential for scene understanding, action recognition [1, 2], and human-object interaction detection [3, 4]. This is particularly challenging in cluttered real-world scenarios due to occlusions and other non-idealities [5, 6]. With the emerging applications in Virtual Reality (VR), and Augmented Reality (AR), and real-time sports analysis [7], there is a fundamental need to understand how objects are manipulated via human-object interaction [8]. Often, in such scenarios, the objects that humans hold and interact with, which we define as *extensions*, can provide crucial information that aids in accurately estimating the human pose and the actions being performed [1].

Contemporary SOTA deep learning-based pose estimation methods predominantly follow a top-down approach: cropping each person in an image using bounding boxes before estimating their pose individually[9–16]. While using existing top-down pose estimators seems intuitive for joint prediction of the humans and their extensions, this approach suffers from limitations, yielding suboptimal results as shown in Fig. 1(a).

Our key observation is: capturing the extension in the bounding box expands the field-of-view and introduces unnecessary visual features, which can be confusing to the model. A simple yet powerful fix is to confine the bounding box to capture the human body, treating the extension's keypoints as *unseen*. This approach reduces background interference, as we do not explicitly capture the extension. However, it leads to the loss of important visual information about the extension, making them unseen.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-3268



(a) (b)

Figure 1: Difference between existing networks and our network. **(a)** Qualitative comparisons between HRNet (left) and our pose estimator (right). **(b)** t-SNE Visualization of keypoint-specific prompt embeddings. The different colours represent whether they are upper-body or lower-body joints, with red representing upper-body joints, blue representing lower-body joints and green representing the hip joints. From the figure, it is evident that while these embeddings maintain positional structure within the upper-body joints (the shoulders are placed above the elbows which are placed above the wrists, following the pose of a human standing normally) and the lower-body joints (knees placed above ankles), they fail to maintain positional structure between the upper-body and lower-body joints (elbows and wrists are placed below knees and ankles in the plot which does not follow the pose of a human being).

Here, a crucial question arises: *How can we effectively represent the spatial relationships of these unseen keypoints for accurate pose estimation?*

To answer this, we turn to other ways of informing the model about the unseen keypoints. Recent works have shown that using language to induce semantic context of keypoints can lead to effective feature representations [17–20]. Specifically, existing works [18] on human pose estimation align the image features with keypoint-specific text embeddings generated from Vision Language Models (VLMs) using a contrastive loss. However, these text embeddings primarily capture local details, neglecting the crucial global relationship between lower-body and upper-body joints; In Fig. 1(b) language models encode similar joints together, while losing the global structure. Hence, explicitly imposing the image features to be close to text embeddings could be suboptimal.

Based on the observations, we present a simple yet effective solution to significantly improve reliability under dynamic real-world scenarios by leveraging language to *see beyond the bounding box*. Specifically, We utilize these text embeddings as *priors* and initialize our learnable keypoint tokens (referred to as text tokens) using these text embeddings. By integrating the rich semantic representations of keypoint-specific text embeddings with image features, and employing a transformer to capture global dependencies, we extract superior fine-grained representations which significantly boosts the performance across the board.

We evaluate `TokenCLIPose`'s performance on three real-world datasets containing a lot of occlusions and noise: an ice hockey dataset, a Lacrosse Dataset and the CrowdPose dataset [6]. The ice hockey and Lacrosse datasets are first-of-their-kind datasets that we curated for predicting the pose of human extensions (the sticks). Furthermore, in order to demonstrate the flexibility of `TokenCLIPose` in predicting unseen keypoints that are present within the bounding box, we evaluate it on the CrowdPose dataset. `TokenCLIPose` outperforms existing top-down approaches by 4.36% and 3.8% on the ice hockey and CrowdPose datasets respectively. Furthermore, `TokenCLIPose` demonstrates superior zero-shot capabilities in predicting extension keypoints when tested on the Lacrosse dataset,

outperforming prior works by 2.35%. Our experiments highlight `TokenCLIPose`'s ability to reliably predict *unseen* keypoints.

2 Related Work

2D Human Pose Estimation: Top-down approaches in 2D pose estimation can be broadly classified into two categories: *Heatmap-based pose estimation* has been the *de-facto* standard approach since the introduction of stacked hourglass networks [14]. These methods represent discrete (x, y) coordinates as continuous heatmaps where each pixel indicates the likelihood of a specific joint being located at that position. Most methods [9, 12, 14, 13] rely on powerful convolutional networks to extract high-level multi-scale feature maps. While most research focuses on network architectures, a few works investigate the coordinate representation and the heatmap encoding and decoding process [21, 15]. Recently, researchers have begun exploring transformer-based architectures for pose estimation [10, 22, 11, 23]. Xu *et al.* [10] adopt the original vision transformer [24] and build baselines for pose estimation, showcasing the efficacy of vision transformers. Methods including [22, 11] use CNNs as feature extractors, and utilize transformers to model the relationship between different scale features and the keypoint features respectively.

Regression-based pose estimation, in contrast to the dominance of heatmap-based approaches offers an alternative paradigm. Sun *et al.* [25] leverages a convolutional backbone to extract feature maps and then utilize an integral operation to directly regress keypoint coordinates. Li *et al.* [26] developed a novel pose regression model which aims at minimizing the distance between the predicted and underlying distribution. However, the global pooling operation used in [26] results in loss of spatial information that is crucial for reliable pose estimation. More recent works [27, 16, 28] include transformer-based architectures: [27, 16] use an encoder-decoder strategy and are based on DETR [29] and Deformable-DETR [30] respectively.

While heatmap-based methods have achieved high accuracy, they have some limitations: 1) These methods have a non-differentiable heatmap decoding method; 2) Heatmap representation often leads to quantization error; and 3) They are not designed to predict out-of-bounding-box keypoints. Therefore, we adopt the regression-based approach for directly estimating the coordinates of all joints, both inside and outside the bounding box.

Occlusion-Aware Pose Estimation: Several approaches tackle the problem of occlusions in crowded scenarios. Various methods tackle the problem by studying the relationships between multiple humans present in an image [31–33]. However, we do not compare with their works as we are interested in single-instance approaches. Park *et al.* [34] addresses the problem of *out-of-bounding box* keypoints by refining the bounding box before pose prediction, but this adds computational cost. Our approach, on the other hand, provides a parameter-free technique to induce spatial context for the *out-of-bounding box* keypoints by leveraging the knowledge of VLMs.

Vision-Language Models: Language supervision has been shown to improve feature representations in various vision tasks, such as image classification, semantic segmentation, and pose estimation. Radford *et al.* [35] proposed the contrastive pretraining paradigm CLIP that leverages contrastive learning to optimize a text and image encoder jointly. This work also showcased the significance of large-scale vision-language pretraining by demonstrating accurate zero-shot classification. Following CLIP, various works such as [36–38] focused on improving image classification using CLIP. Zhang *et al.* [39] transfer the 2D pre-trained knowledge to 3D domains, thereby improving zero-shot point-cloud recognition. Rao *et al.* [40] showcased the efficacy of CLIP pretraining on dense prediction tasks by converting the image-text matching problem to a pixel-text matching problem.

Recently, Tevet *et al.* [20] leverage text to inpaint missing poses in a sequence in a spatiotemporally consistent manner. Guo *et al.* [19] use pose-aware prompts to predict 3D hand meshes from images. Recent works on 2D pose estimation [18, 17] utilize joint-specific keypoints to learn richer representations. Particularly, they use a contrastive loss to align the image features extracted from a backbone to their keypoint-specific text embeddings. This, however, is suboptimal as the text embeddings do not capture global structure between human keypoints. Hence, aligning the image features to text embeddings might lead to loss of spatial information and inaccurate localization of joints. We improve this by using a transformer-based network to capture the global spatial dependencies between image and text features, thereby guiding our image features using text supervision rather than biasing them to the text prompts themselves.

 https://doi.org/10.52202/079017-3268

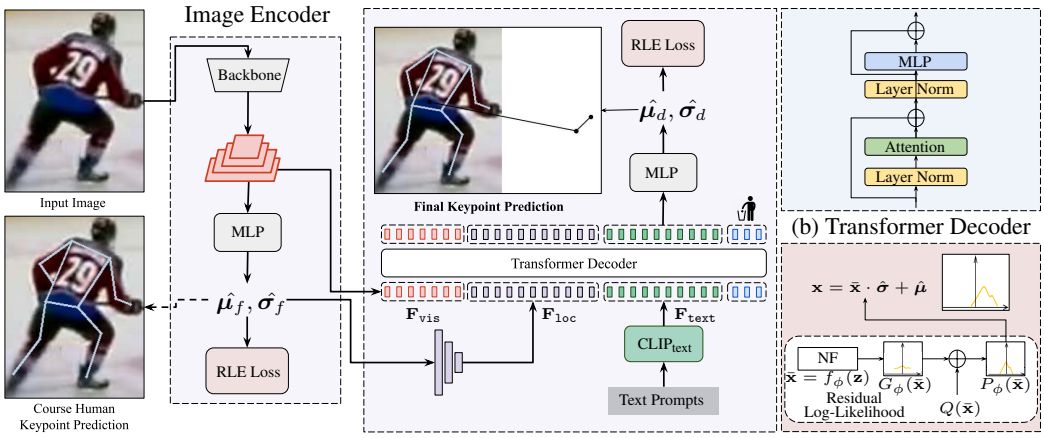


Figure 2: TokenCLIPose Architecture: We first incorporate an image encoder to extract multi-scale image features and coarse human keypoint locations $\hat{\boldsymbol{\mu}}_f$, and project them onto a joint multimodal embedding space obtaining image tokens $\mathbf{F}_{\texttt{vis}}$ and location tokens $\mathbf{F}_{\texttt{loc}}$ respectively. Then, we leverage a text-based keypoint encoder to extract keypoint-specific text tokens $\mathbf{F}_{\texttt{text}}$ from VLMs. These multimodal tokens are fed to a transformer decoder to capture spatial dependencies between them and predict all the 2D keypoints $\hat{\boldsymbol{\mu}}_d$. The coarse human keypoint predictions and the final keypoint predictions are supervised using the RLE loss.

3 Method

3.1 Problem Formulation

Given a cropped image $\mathbf{I} \in \mathbb{R}^{h \times w}$ of a human, generated from bounding boxes obtained through a detection network, we predict K_{out} keypoints $\hat{\boldsymbol{\mu}}_d \in \mathbb{R}^{K_{out} \times 2}$ that represent the 2D poses of extensions and/or humans, along with the scale parameter $\hat{\boldsymbol{\sigma}}_d$ for each keypoint.

3.2 Network Architecture

We propose an encoder-decoder architecture to estimate 2D keypoints that are *not captured* in the bounding box. Firstly, an image encoder extracts multi-scale image features from the cropped image, which are then passed through a Multi-layer Perceptron (MLP) to generate coarse human keypoint proposals. These image features and locations are then projected onto the multimodal embedding space to form image and location tokens. The text-based keypoint encoder leverages text prompts to generate keypoint-specific text tokens. Finally, all these multimodal tokens are concatenated and passed through a transformer decoder to capture global relationships between these tokens and predict final 2D keypoints. The network is exemplified in Fig. 2.

Image Encoder. The cropped input image is initially passed through a pretrained CNN to extract multi-scale dense feature maps. These multi-scale feature maps are fused and projected onto a joint multimodal embedding space to form the image tokens $\mathbf{F}_{vis} \in \mathbb{R}^{N \times C_{emb}}$, where N is the number of tokens and C_{emb} is the joint multimodal space dimension. Furthermore, they are processed through a MLP to generate coarse human keypoint predictions $\hat{\boldsymbol{\mu}}_f \in \mathbb{R}^{K_h \times 2}$, and a scale parameter $\hat{\boldsymbol{\sigma}}_f \in \mathbb{R}^{K_h \times 1}$, which are optimized using the RLE process as detailed in Section 3.3.

Text-based Keypoint Encoder. Following image feature extraction, the keypoint encoder tackles the challenge of *unseen* keypoints in the image. We employ language-guided keypoint representations, where text prompts encode class-related information of each keypoint. By leveraging CLIP's pre-trained text encoder, we generate the text tokens $\mathbf{F}_{\texttt{text}} \in \mathbb{R}^{K_{out} \times C_{emb}}$ which are then concatenated with the extracted image tokens. This process indirectly injects visual context of the missing keypoints into the model, even when it's not directly visible in the cropped input image. Additionally, we incorporate the coarse human keypoint predictions $\hat{\boldsymbol{\mu}}_f$ and convert them to location tokens by projecting

them onto the joint mulitmodal space $\mathbf{F}_{\texttt{loc}} \in \mathbb{R}^{K_h \times C_{emb}}$, and concatenate them with the image and text tokens. The final set of tokens fed to the decoder are denoted as $\mathcal{F} = \{\mathbf{F}_{\texttt{vis}}, \mathbf{F}_{\texttt{text}}, \mathbf{F}_{\texttt{loc}}\}$.

Transformer Decoder. To predict all the keypoints precisely, the relationships between these multimodal tokens ($\mathcal{F}$) must be captured accurately. Therefore, we leverage the transformer decoder to understand the inherent correlations between different text, image and location tokens. Instead of treating each modality separately and utilizing a cross-attention mechanism to understand the associations between them, we treat all tokens together as a homogeneous entity and employ the standard self-attention mechanism. This approach offers a simpler way to gain a holistic view of the relationships between all keypoints, locations, and image patches. The transformer layers are followed by MLPs to estimate the final keypoint predictions $\hat{\boldsymbol{\mu}}_d \in \mathbb{R}^{K_{out} \times 2}$ and the scale parameter $\hat{\boldsymbol{\sigma}}_d \in \mathbb{R}^{K_{out} \times 1}$. To reduce artifacts in feature maps and understand richer relationships, we employ additional register tokens as [41].

3.3 Loss Function

Distribution Learning. Following [26, 16], we formulate the regression task as a distribution learning problem and adopt Maximum Likelihood Estimation (MLE) to effectively predict the output coordinates. We use normalizing flows to estimate the deviation in predicted and ground truth keypoint distributions. In mathematical terms, given an input image $\mathcal{I}$, our network estimates a distribution $P_{\Theta,\Phi}(\mathbf{x}|\mathcal{I})$ representing the probability of the ground truth keypoint appearing at the location $\mathbf{x}$. Here, Θ and Φ denote the parameters of our pose network and the flow model f_ϕ, respectively. The flow model f_ϕ acts as a refinement step, iteratively transforming a preset Gaussian distribution $\bar{\mathbf{z}} \sim \mathcal{N}(0, \mathrm{Id})$ to capture the deviation of the predicted and ground truth distributions using the network's prediction ($\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\sigma}}$).

Equation (1) depicts the mathematical formulation of the RLE loss, where $Q(\bar{\boldsymbol{\mu}}_g)$ is the preset Gaussian distribution, $G_\phi(\bar{\boldsymbol{\mu}}_g)$ is the learned distribution by the flow model, $\bar{\boldsymbol{\mu}}_g = (\boldsymbol{\mu}_g - \hat{\boldsymbol{\mu}})/\hat{\boldsymbol{\sigma}}$ represents the normalized difference between ground truth and predicted keypoints, and s is a constant term.

$$\mathcal{L}_{RLE} = -\log Q(\bar{\boldsymbol{\mu}}_g) - \log G_\phi(\bar{\boldsymbol{\mu}}_g) - \log s + \log \hat{\boldsymbol{\sigma}} \tag{1}$$

Similar to [16], we supervise both the coarse predictions $(\hat{\boldsymbol{\mu}_f}, \hat{\boldsymbol{\sigma}_f})$ and the final predictions from the decoder $(\hat{\boldsymbol{\mu}_d}, \hat{\boldsymbol{\sigma}_d})$ using RLE. Hence, our final loss function is

$$\mathcal{L} = \mathcal{L}^f_{RLE} + \mathcal{L}^d_{RLE} \tag{2}$$

where,

$$\mathcal{L}^f_{RLE} = -\log P_{\Theta_f,\Phi_f}(\mathbf{x}|\mathcal{I})\Big|_{\mathbf{x}=\boldsymbol{\mu}_g} \text{ and } \mathcal{L}^d_{RLE} = -\log P_{\Theta_d,\Phi_d}(\mathbf{x}|\mathcal{I})\Big|_{\mathbf{x}=\boldsymbol{\mu}_g}$$

Here, Θ_f and Φ_f denote the parameters of the backbone network and the flow model of the coarse keypoint predictions, and Θ_d and Φ_d denote the parameters of our decoder regression model and flow model of the final keypoint predictions.

4 Experiments

A critical challenge in evaluating pose networks for human extensions is the lack of publicly available datasets containing both the human and extension pose annotations. To address this limitation, we curate two new sports datasets, ice hockey and Lacrosse, featuring pose annotations for humans and their corresponding extensions. Experiments on these datasets demonstrate the performance of our model in estimating the pose of extension keypoints, which are not captured within the bounding box. We further evaluate the model's ability to predict unseen keypoints (due to occlusion/self-occlusion) within the bounding box using the benchmarked multi-person cluttered dataset, Crowdpose [6], validating the generalizability of our model.

For human pose estimation, the total keypoints (K_{out}) depend on extensions. Without extensions, K_{out} equals the number of human keypoints (K_h). When extensions are present, K_{out} increases to K_h plus the number of extension keypoints (K_e).

 https://doi.org/10.52202/079017-3268

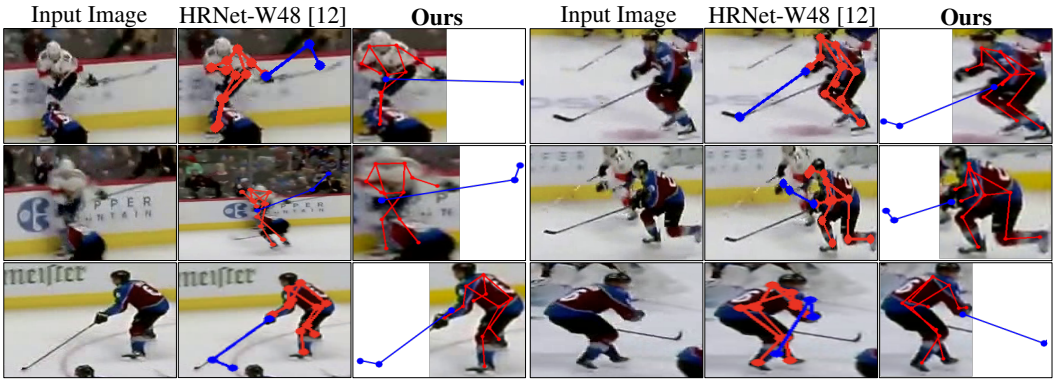


Figure 3: Qualitative Comparison of `TokenCLIPose` with HRNet-W48 on our ice hockey dataset.

4.1 Ice Hockey Dataset

Dataset. The ice hockey dataset, captured from real-world National Hockey League (NHL) videos, presents a challenging environment for pose estimation due to the inherent fast pace of the sport. Players' rapid movements often result in motion blur within frames, making it difficult to distinguish body parts. Furthermore, the nature of the game leads to frequent occlusions, especially when players obstruct each other's bodies. Adding to the complexity is the bulky equipment worn by hockey players, which can obscure keypoints. These combined challenges- motion blur, occlusion, and bulky equipment- make the ice hockey dataset a valuable resource for evaluating the robustness of the pose estimation task.

Our dataset consists of 10 video clips (30-45 seconds each, sampled at 30 fps) from various broadcast NHL videos. We utilize the CVAT tool to annotate the keypoints for multiple players and their hockey sticks in each frame. We followed the standard COCO format to annotate 17 keypoints for the human pose. Additionally, we annotate 3 keypoints for the hockey stick (butt end, heel, and toe). Finally, for each annotated frame, we create two versions of the input image: one cropped to include only the player (*standard bounding box*) and another incorporating both the player and their hockey stick (*extended bounding box*). In total, we generate $11.66K$ pose annotations, with $9.13K$ images from 9 clips used for training and $2.53K$ images from the tenth clip used for testing our model's performance.

Evaluation Metric. We evaluate different State-Of-The-Art (SOTA) pose estimation models on our dataset using the Percentage of Correct Keypoints with head-normalization (PCKh) metric. Due to our method's use of smaller bounding box compared to the other models, achieving the same level of accuracy would result in a higher PCKh threshold for our model. To ensure a fair and direct comparison across all models, we employ the same threshold for evaluation.

Training. All networks are trained for 200 epochs with a batch size of 64 on a single NVIDIA GeForce RTX 4090 GPU. We employ the Adam optimizer with a learning rate of 6×10^{-4} for all CNN-based architectures while the transformer-based architectures are trained using the AdamW optimizer with an initial learning rate of 3×10^{-4}. The weight decay is set to 10^{-5} for all models. A stepLR scheduler was used to linearly reduce the learning rate from the initial value of 10^{-5}. Consistent with [26], we utilize RealNVP [42] as the flow model within our model.

Results. The results, presented in Table 1, emphasize our method's *SOTA performance* over existing methods in robustly predicting the ice hockey player and the hockey stick keypoints, outperforming existing works by 4.36%. This is further validated by the qualitative comparisons in Fig. 5, where our model demonstrates the ability to predict semantic poses even in challenging scenarios with extreme motion blur and occlusion. These results support our hypothesis that *re-considering the extension pose estimation task as an unseen keypoint problem without explicitly capturing it in the bounding box* reduces background noise leading to robust poses.

4.2 Lacrosse Dataset

Dataset. In order to study the efficacy of `TokenCLIPose`'s generalization capabilities in predicting the pose of extensions, we curate another small-scale Lacrosse dataset. Similar to ice hockey, it is

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% ↑**

characterized by motion blur and occlusions, but there are domain differences between the 2 datasets. This dataset consists of 300 pose annotations from one video sampled at 30 fps. We use the same 12 human keypoints used in the ice hockey dataset to denote the human's pose. However, we use only 2 keypoints to represent the Lacrosse stick's pose, as the blade of a Lacrosse stick is circular in nature. Hence, we disregard the 15th keypoint (stick toe) and use the other 14 keypoints to represent the pose of a Lacrosse player along with their stick.

Zero-shot Results. We evaluate TokenCLIPose's generalization capability by performing zero-shot transfer from our pretrained model on ice hockey to the Lacrosse dataset. Due to the difference in the shape of the head of a lacrosse and hockey stick, we predict the two keypoints corresponding to the shaft of a lacrosse stick. Furthermore, the text prompts for the two extension keypoints are also changed while keeping the model frozen. The results presented in Table 2 showcase that our model outperforms the established baselines by 2.35%, thereby demonstrating the efficacy of our proposed model for generalizable pose estimation.

4.3 CrowdPose Dataset

Dataset. The CrowdPose dataset is a large-scale benchmark dataset for human pose estimation, containing $12K$ images and $43.4K$ labeled people in their `trainval` set, and $8K$ images with $29K$ labeled people in the `test` set. Following [32, 18, 31], we use the `trainval` set for training and `test` set for evaluation.

Evaluation Metric. We adopt standard Average Precision (AP) as our evaluation metric on the CrowdPose dataset. AP is calculated based on Object Keypoint Similarity (OKS) denoted by $m_{OKS} \in \mathbb{R}$, which is defined as

$$m_{OKS} = \frac{\sum_i \exp(-\hat{d}_i^2/2s^2k_i^2)\sigma(v_i > 0)}{\sum_i \sigma(v_i > 0)}, \tag{3}$$

where $\hat{d}_i$ is the Euclidean distance between the i-th predicted keypoint coordinate and the corresponding ground truth, v_i is the visibility flag of the keypoint, s is the object scale, and k_i is a keypoint-specific constant.

Training. We employ the AdamW optimizer with an initial learning rate of 6×10^{-4} and weight decay of 0.1. Following ViTPose [10], we apply linear warmup for the first 2000 iterations with a warmup factor of 10^{-3}. Furthermore, we perform gradient clipping to prevent overfitting.

https://doi.org/10.52202/079017-3268

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% ↑**

Table 5: **Effect of Attention Mechanisms** on the overall accuracy.

Attention-Mechanism	Mean
Intention [46]	75.14
Self-attention	76.28
Cross-attention	**76.31** ↑ 0.03

Table 6: **Effect of Text Prompts** on the stick accuracy.

Prompt type	Stick Accuracy	Mean
No text	63.37	72.93
Single Prompt	67.41	75.24
Prompt Ensemble	**69.97** ↑ 2.56%	**76.28** ↑ 1.04%

Results. Table 3 presents a comparison between established pose estimation techniques and our method. As shown in the table, `TokenCLIPose` outperforms all the top-down approaches by 3.8% demonstrating its efficacy in predicting unseen keypoints robustly. We also depict the qualitative results of `TokenCLIPose` to depict its robustness to occlusion. It is notable that even in scenarios where we only see the side-view of humans, TokenCLIPPose estimates the pose reliably.

4.4 Ablation Studies

We conduct comprehensive ablations to study the effect of our design choices and verify the impact of each module on our proposed `TokenCLIPose`. For consistency, all the ablations are conducted on our ice hockey dataset unless specified otherwise.

Gains from Each Modality. The influence of each modality on the proposed model's performance is shown in Table 4. As evidenced by a performance improvement of 3.35%, the inclusion of text tokens plays a significant role in enhancing the pose estimation accuracy. On the other hand, we find that including the location tokens do not improve the performance significantly, only by a small margin of 0.45%.

Table 4: **Effect of Different Modalities** on the overall accuracy.

Text tokens ($\mathbf{F}_{\texttt{text}}$)	Location tokens ($\mathbf{F}_{\texttt{loc}}$)	Image tokens ($\mathbf{F}_{\texttt{vis}}$)	Mean
✗	✗	✓	72.48
✗	✓	✓	72.93
✓	✗	✓	75.72
✓	✓	✓	**76.28**

Do we need to treat each modality heterogeneously? We probe whether multimodal tokens need to be treated heterogeneously by testing out different attention mechanisms for our transformer decoder. As shown in Table 5, utilizing self-attention directly on all the tokens produces similar results to performing self-attention separately on image, location and text tokens, and then applying cross-attention. Therefore, it is not necessary to treat the tokens from each modality separately. We

Figure 4: Qualitative Results of our `TokenCLIPose` on the CrowdPose dataset.

hypothesize that this could be due to all the tokens being projected to the same joint embedding space, thereby eliminating the need to process them differently.

Impact of Text Prompts. We investigate the influence of language on our model's efficacy on the ice hockey dataset by varying the degree and complexity of text prompts that we use. Starting from randomly initializing the text tokens instead of using CLIP embeddings, we study the effects of using single prompts and the prompt ensemble technique proposed in [35] for ImageNet classification. It is evident from Table 6 that incorporating text instead of randomly initializing text tokens results in a significant improvement of 2.31% in the overall accuracy. Furthermore, using prompt ensemble techniques improve the accuracy of the model by 1.04%, showcasing the importance of the quality of text prompts.

Influence of Bounding Boxes. We evaluate the influence of bounding box predictions on our model's performance by using ground truth bounding boxes obtained from ground truth human poses and the bounding boxes from the FasterRCNN object detector. The results are showcased in Table 7. The above table illustrates our model's robustness to the quality of bounding boxes, as using ground truth bounding boxes improves the model's performance by a small margin of 2.47%.

Table 7: **Impact of Bounding Boxes** on the overall accuracy.

Bounding boxes	Mean
Faster-RCNN [47]	76.28
Ground Truth	**78.75** ↑ 2.47%

5 Conclusion

In this work, we proposed `TokenCLIPose`, an innovative solution for robustly predicting human and *extension* poses. Instead of explicitly modeling extensions within the bounding box, we reformulated the *extension* pose estimation as an unseen keypoint prediction problem. We leverage the power of large pre-trained VLMs in augmenting the spatial information of unseen keypoints. We showcased that capturing relationships between multimodal tokens is more effective than aligning image features to text tokens. To evaluate the effectiveness of `TokenCLIPose` in predicting *extension* keypoints, we curated real-world datasets for ice hockey and Lacrosse. We significantly outperform existing top-down methods on these datasets (by 4.36% and 2.35%, respectively). Additionally, we achieve a 3.8%

https://doi.org/10.52202/079017-3268

improvement over prior top-down networks on the CrowdPose dataset. This shows `TokenCLIPose`'s flexibility to predict unseen keypoints within the bounding box as well. Future work will focus on curating and training more extensive and diverse datasets for human and extension pose estimation tasks.

6 Acknowledgement

This work was supported in part by Stathletes, Compute Canada, the Natural Sciences and Engineering Research Council of Canada and MITACS.

References

[1] H. Neher, K. Vats, A. Wong, and D. A. Clausi, "Hyperstacknet: A hyper stacked hourglass deep convolutional neural network architecture for joint player and stick pose estimation in hockey," *2018 15th Conference on Computer and Robot Vision (CRV)*, pp. 313–320, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:57760774

[2] D. C. Luvizon, D. Picard, and H. Tabia, "2d/3d pose estimation and action recognition using multitask deep learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5137–5146.

[3] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, "Pose-aware multi-level feature network for human object interaction detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9469–9478.

[4] Z. Liang, J. Liu, Y. Guan, and J. Rojas, "Pose-based modular network for human-object interaction detection," *arXiv preprint arXiv:2008.02042*, 2020.

[5] S.-H. Zhang, R. Li, X. Dong, P. Rosin, Z. Cai, X. Han, D. Yang, H. Huang, and S.-M. Hu, "Pose2seg: Detection free human instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[6] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 863–10 872.

[7] K. Ludwig, P. Harzig, and R. Lienhart, "Detecting arbitrary intermediate keypoints for human pose estimation with vision transformers," *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 663–671, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:246871051

[8] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 17–24.

[9] W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, G. Yu, H. Lu, Y. Wei, and J. Sun, "Rethinking on multi-stage networks for human pose estimation," *ArXiv*, vol. abs/1901.00148, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:57373771

[10] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose: Simple vision transformer baselines for human pose estimation," *ArXiv*, vol. abs/2204.12484, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248392410

[11] Y. Li, S. Zhang, Z. Wang, S. Yang, W. Yang, S. Xia, and E. Zhou, "Tokenpose: Learning keypoint tokens for human pose estimation," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11 293–11 302, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233181621

[12] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5686–5696, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:67856425

[13] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," *ArXiv*, vol. abs/1804.06208, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:4934594

[14] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:13613792

[15] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7091–7100, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:204509549

[16] W. Mao, Y. Ge, C. Shen, Z. Tian, X. Wang, Z. Wang, and A. van den Hengel, "Poseur: Direct human pose regression with transformers," in *European Conference on Computer Vision*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:246035459

[17] X. Zhang, W. Wang, Z. Chen, Y. Xu, J. Zhang, and D. Tao, "Clamp: Prompt-based contrastive learning for connecting language and animal pose," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2023. [Online]. Available: http://dx.doi.org/10.1109/CVPR52729.2023.02229

[18] S. Hu, C. Zheng, Z. Zhou, C. Chen, and G. R. Sukthankar, "Lamp: Leveraging language prompts for multi-person pose estimation," *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3759–3766, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260126038

[19] S. Guo, Q. Cai, L. Qi, and J. Dong, "Clip-hand3d: Exploiting 3d hand pose estimation via context-aware prompting," ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3581783.3612390

[20] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano, "Human motion diffusion model," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=SJ1kSyO2jwu

[21] J. Huang, Z. Zhu, F. Guo, and G. Huang, "The devil is in the details: Delving into unbiased data processing for human pose estimation," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5699–5708, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:208138553

[22] S. Yang, Z. Quan, M. Nie, and W. Yang, "Transpose: Keypoint localization via transformer," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11 782–11 792, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:236428790

[23] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "Hrformer: High-resolution transformer for dense prediction," *ArXiv*, vol. abs/2110.09408, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:239016306

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[25] X. Sun, B. Xiao, S. Liang, and Y. Wei, "Integral human pose regression," *ArXiv*, vol. abs/1711.08229, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:4055834

[26] J. Li, S. Bian, A. Zeng, C. Wang, B. Pang, W. Liu, and C. Lu, "Human pose regression with residual log-likelihood estimation," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11 005–11 014, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:236318539

[27] K. Li, S. Wang, X. Zhang, Y. Xu, W. Xu, and Z. Tu, "Pose recognition with cascade transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 1944–1953.

[28] P. Panteleris and A. Argyros, "Pe-former: Pose estimation transformer," in *International Conference on Pattern Recognition and Artificial Intelligence*. Springer, 2022, pp. 3–14.

[29] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *ArXiv*, vol. abs/2005.12872, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:218889832

[30] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *ArXiv*, vol. abs/2010.04159, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:222208633

[31] R. Khirodkar, V. Chari, A. Agrawal, and A. Tyagi, "Multi-instance pose networks: Rethinking top-down pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3122–3131.

[32] M. Zhou, L. Stoffl, M. W. Mathis, and A. Mathis, "Rethinking pose estimation in crowds: Overcoming the detection information bottleneck and ambiguity," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 14 689–14 699.

[33] D. Wang and S. Zhang, "Contextual instance decoupling for robust multi-person pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 11 060–11 068.

[34] S. Park and J. Park, "Localizing human keypoints beyond the bounding box," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1602–1611.

[35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[36] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *International Journal of Computer Vision*, vol. 132, no. 2, pp. 581–595, 2024.

[37] Y. Wei, Y. Cao, Z. Zhang, H. Peng, Z. Yao, Z. Xie, H. Hu, and B. Guo, "iclip: Bridging image classification and contrastive language-image pre-training for visual recognition," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2776–2786, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259267527

[38] R. Zhang, Z. Wei, R. Fang, P. Gao, K. Li, J. Dai, Y. J. Qiao, and H. Li, "Tip-adapter: Training-free adaption of clip for few-shot classification," *ArXiv*, vol. abs/2207.09519, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:250698940

[39] R. Zhang, Z. Guo, W. Zhang, K. Li, X. Miao, B. Cui, Y. Qiao, P. Gao, and H. Li, "Pointclip: Point cloud understanding by clip," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8552–8562.

[40] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu, "Denseclip: Language-guided dense prediction with context-aware prompting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 082–18 091.

[41] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, "Vision transformers need registers," *arXiv preprint arXiv:2309.16588*, 2023.

[42] L. Dinh, J. N. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," *ArXiv*, vol. abs/1605.08803, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:8768364

[43] K. He, G. Gkioxari, and P. Dollár, "Girshick ross. mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[44] J. Li, W. Su, and Z. Wang, "Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 11 354–11 361.

[45] W. McNally, K. Vats, A. Wong, and J. McPhee, "Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation," *arXiv preprint arXiv:2111.08557*, 2021.

[46] M. Garnelo and W. M. Czarnecki, "Exploring the space of key-value-query models with intention," *arXiv preprint arXiv:2305.10203*, 2023.

[47] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

 https://doi.org/10.52202/079017-3268

A Appendix

In this appendix, we provide more qualitative results on our ice hockey and Lacrosse datasets. Furthermore, we provide a quantitative comparison with a multimodal network ([18]).

A.1 Comparison with Multimodal Pose Estimators

To investigate the effectiveness of our technique of leveraging text with existing multimodal methods, we compare `TokenCLIPose`'s performance with the multimodal counterpart LAMP [18]. Though LAMP does not follow the top-down approach, it is the only network that uses language for 2D human pose estimation. Thus, we compare and highlight the results in Table 8. As shown in the table, it is evident that `TokenCLIPose` outperforms LAMP on the CrowdPose dataset by 4.8%. This corroborates our claim that *leveraging text as supervisory signals to guide image features provides better performance than aligning image features to the text tokens.*

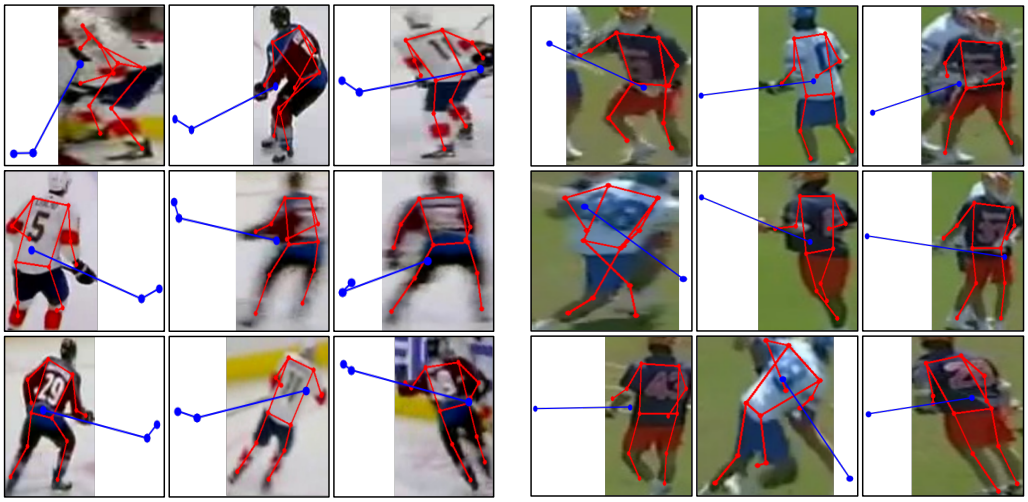
Table 8: **Comparison with LAMP** on CrowdPose dataset

Method	Input Resolution	AP	AP_{50}	AP_{75}	AP_E	AP_M	AP_H
LAMP [18]	512 × 512	71.4	90.3	77.1	77.9	72.1	64.2
`TokenCLIPose`	384 × 288	**76.2**	**93.9**	**82.4**	**83.3**	**77.4**	**66.1**

A.2 Qualitative Results

We showcase additional qualitative results on our ice hockey dataset visualizing the effectiveness of our approach. Furthermore, we also demonstrate the generalization capabilities of our model in transferring to a Lacrosse dataset. The impact of language is visible in our model's robustness to domain changes.

To further illustrate the effectiveness of our approach, we present additional qualitative results on the ice hockey dataset in Figure 5. Furthermore, we demonstrate the model's generalization capabilities by achieving strong performance on a Lacrosse dataset, highlighting the impact of language on the model's robustness to domain changes.



(a) ice hockey (b) Lacrosse

Figure 5: Qualitative Results of `TokenCLIPose` for the extension pose estimation task on our ice hockey and Lacrosse datasets. We plot the pose of extensions outside the cropped image to understand how `TokenCLIPose` works.

B Broader Impact Statement

Our work provides a new way to reconsider the pose estimation task to predict out-of-bounding box keypoints. This improves our understanding of why joint hockey stick pose prediction along with player pose is difficult for existing top-down networks and presents one way to mitigate the issues. This understanding should enable the application of top-down solutions for multi-instance pose estimation which encounters similar issues. Furthermore, accurate stick pose prediction implies that hockey teams can make more data-driven decisions to improve their teams' performance without incurring additional overhead expenses. This also implies that the use of invasive technology such as infrared sensors can be avoided to a large extent to robustly analyze players and teams.

This can be leveraged in various domains where humans closely interact with objects, such as shoveling, where it could be used to study fatigue and biomechanics. Furthermore, as shown in our experiments on the CrowdPose dataset, `TokenCLIPose` also reliably predicts the pose of a human when there is high human-human interaction. Hence, this can be used for pose estimation in highly crowded scenarios, such as surveillance and monitoring people in crowds.

https://doi.org/10.52202/079017-3268

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and precede the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We demonstrate and reason out the claims that we make in the abstract by rigorous quantitative and qualitative experimentation.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

Justification:

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Our paper contains no theoretical proof.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We specify the training and model details explicitly.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

https://doi.org/10.52202/079017-3268

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We do work with an institution which has full access to our work. However, with their permission, we could release the dataset and models soon.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We present all the experimental details necessary for reproduction.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Following prior works in 2D pose estimation, we do not report standard deviations but clearly mention the evaluation metrics and report our performance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the GPU that we run the experiments on and all other details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

https://doi.org/10.52202/079017-3268

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conform to the NeurIPS code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our work is strictly restricted to 2D pose estimation which is mainly used for scene understanding, activity recognition and sports analysis. Hence, we do not see any negative societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The datasets that we utilize, apart from the ones we created are all benchmark datasets who are cited properly.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We clearly mention how we curate our datasets and create our models.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

https://doi.org/10.52202/079017-3268

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.