# Towards Annotation-Free Evaluation: KPAScore for Human Keypoint Detection

Supplementary Material

## Supplementary Overview

The following provides an overview of the content in each Appendix section:

- **Section A**: Supplement to related work.
- **Section B**: The effect of the color of the visual prompt.
- **Section C**: Fine-Tuning Hyperparameters.
- **Section D**: Cases of KPAScors and OKS.
- **Section E**: Human Rating Strategy.
- **Section F**: KPAScore Alignment with Human Perception in More Complex Scenarios.

## A. Supplement to related work

**Evaluation Capabilities of VLMs:** VLMs [19, 20, 23, 31] are trained on large-scale, diverse datasets of image-text pairs, equipping them with rich semantic understanding and strong generalization capabilities. They perform exceptionally well across various downstream tasks, such as image-text retrieval [32], image comprehension [10], visual question answering [25], and object localization [5]. In the VQA score [18], the VLMs were used to estimate the probability that an image contains a specific object as a measure of image-text alignment. VQA score [18] effectively reflects the alignment level between text and image based on the probability of a "Yes" response from the VLM, achieving strong alignment with human perception. However, directly applying VLMs to evaluate human keypoint detection, similar to the approach used in VQA score [18], leads to severe hallucinations, making them unsuitable for evaluation purposes. Our two-stage approach, enhanced with visual prompts, significantly improves the VLM's ability to identify keypoints. This marks the first application of VLM in evaluating human keypoint detection.

## B. The Effect of the Color of the Visual Prompt

To evaluate the impact of visual prompt color on precision localization detection, we conduct experiments by replacing the visual prompt with different colored dots and measuring its effect on the accuracy of GLM-4V-9B [7].

As shown in Table 5, the results indicate that the choice of visual prompt color does not significantly affect the VLM's ability to understand keypoints. This suggests that the model primarily relies on spatial cues rather than color variations when interpreting keypoint locations.

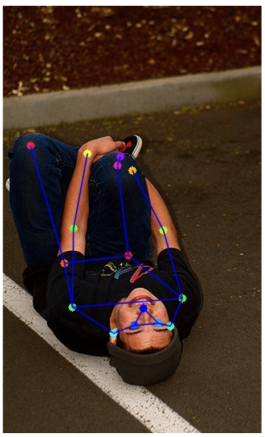
## C. Fine-Tuning Hyperparameters

To further enhance the model's precision, we fine-tuned GLM-4V-9B using LoRA with 10k Precision Localization Detection samples containing visual prompts. We froze the visual encoder and fine-tuned only the linear projection layers and the language model component. The fine-tuning was conducted with a learning rate of $5 \times 10^{-4}$ for 3000 training steps. For LoRA-specific configurations, we set the rank to 8, alpha to 32, and applied a dropout rate of 0.1 to enhance generalization. The experiments were performed on two A100 GPUs, ensuring efficient training and optimization. This setup allows for adapting the model to keypoint localization tasks while maintaining computational efficiency.

| Color | Accuracy (%) | E(Prob) |
|---|---|---|
| Blue | 62.57 | 58.17 |
| Red | 62.34 | 58.11 |
| Green | 62.01 | 58.09 |
| Yellow | 61.78 | 57.91 |
| Black | 61.66 | 57.82 |

Table 5. Impact of visual prompt color on GLM-4V-9B accuracy . The results indicate that color variations have negligible effects on the model's keypoint understanding.

## D. Cases of KPAScors and OKS



| Keypoint | KPAScore | OKS |
|---|---|---|
| Nose | 0.94 | 0.70 |
| Left Eye | 0.88 | 0.99 |
| Right Eye | 0.86 | 0.98 |
| Left Ear | 0.78 | 0.93 |
| Right Ear | None | None |
| Left Shoulder | 0.85 | 0.94 |
| Right Shoulder | 0.84 | 0.95 |
| Left Elbow | 0.77 | 0.99 |
| Right Elbow | 0.72 | 0.99 |
| Left Wrist | 0.75 | 0.98 |
| Right Wrist | None | 0.69 |
| Left Hip | 0.69 | 1.00 |
| Right Hip | 0.68 | 0.99 |
| Left Knee | 0.79 | 1.0 |
| Right Knee | 0.74 | 0.99 |
| Left Ankle | 0.55 | 0.90 |
| Right Ankle | None | 0.02 |
| **Average** | **0.79** | **0.88** |

Figure 6. Keypoint predictions with corresponding KPAScore and OKS values.

To further investigate the differences between KPAScore and OKS, we analyze cases where the two metrics produce divergent scores. Fig. 6 illustrates an example in which a person is lying on the ground, with keypoint predictions


Figure 7. A comparison of the KPA-Score and Human Evaluation Scores for seven algorithms across a variety of scenarios.

overlaid and their corresponding KPAScore and OKS values listed.

KPAScore exhibits greater tolerance for minor spatial deviations compared to OKS, as seen in keypoints like the nose (0.94 vs. 0.70) and left hip (0.69 vs. 1.00), where OKS heavily penalizes small positional shifts despite their perceptual validity. This difference arises because OKS relies purely on Euclidean distance, while KPAScore integrates semantic plausibility via visual-language models (VLMs), allowing for reasonable variations. Additionally, KPAScore accounts for anatomical plausibility, distinguishing between spatial proximity and realistic positioning. For example, it assigns None to the right ankle and right wrist, indicating implausibility, whereas OKS still provides scores (0.02 and 0.69, respectively). This highlights KPAScore's alignment with human perception. However, when keypoints are well-detected—such as the left eye (0.88 vs. 0.99), left shoulder (0.85 vs. 0.94), and left knee (0.79 vs. 1.00)—both metrics yield similar results, confirming KPAScore as a reliable alternative with an added perceptual validation layer.

## E. Human Rating Strategy

To evaluate the performance of various human keypoint detection algorithms in complex environments, we employed KPAScore alongside human ratings as a comparative measure. The human evaluation was conducted by 20 researchers, each independently rating the inference results of different algorithms on the test set.

For each image, raters assigned a score from 0 to 10 based on the perceptual correctness of the predicted keypoints. The scoring criteria considered the following aspects:

- Spatial Accuracy: Whether the keypoints align correctly with the expected anatomical positions.
- Completeness: The presence of all keypoints without missing or redundant detections.
- Plausibility: Whether the overall keypoint arrangement appears natural and realistic.

Each algorithm's final score was computed as the average rating across all raters and test images. This process ensures a robust evaluation aligned with human perception, complementing the quantitative assessment provided by KPAScore.

## F. KPAScore Alignment with Human Perception in More Complex Scenarios

In Fig. 7, we presents a comparison of the KPAScore and Human Evaluation Score for seven algorithms across a variety of multi-person and single-person scenarios. Firstly, the strong correlation between human evaluation scores and KPAScores further validates the effectiveness of KPAScore in reflecting pose estimation quality. Additionally, we used KPAScore as an evaluation metric for individual keypoints and aggregated the results from seven algorithms to calculate the KPAPose for each image.

The results demonstrate that KPAPose consistently outperforms other detection algorithms in both model-based KPA evaluations and human evaluations. In practical applications, we suggest using KPAScore to flexibly ensemble the results of multiple algorithms, thus achieving a robust and generalizable detection method that adapts effectively to various scenarios and movements.

Furthermore, we observed significant differences in algorithm performance in complex multi-person scenarios. ViTPose demonstrated the best performance, with KPAScore of 75.65, 83.26, and 76.25, showcasing its strong ability to handle multi-person keypoint detection tasks. RSN followed closely and also performed well. However, the performance of other algorithms dropped significantly, with KPAScore indicating their limited ability to handle complex multi-person scenarios. Moreover, these algorithms struggle with "Multi 2" and "Multi 3" scenarios in Fig.7, where occlusion and overlapping poses likely affect detection accuracy. This highlights the importance of addressing these issues to improve the robustness of future algorithms.

In contrast, the differences in algorithm performance are less pronounced in single-person scenarios. ViTPose, RSN, and HRNet achieved comparable KPAScore, and human evaluations of their detection results were similarly consistent, indicating relatively balanced performance in single-person keypoint detection tasks. However, it is worth noting that algorithm performance still has substantial potential for improvement, particularly in scenarios involving clothing occlusion or self-occlusion caused by human movements, where current methods have yet to reach saturation levels.

It is evident that KPAScore can not only evaluate keypoint detection algorithms in natural scenes but also serve as an important tool for assessing algorithm performance in complex occlusion scenarios, particularly in situations where precise annotations are costly to obtain.