Pose2Lang3D : Distilling 3D Reasoning from 2D Skeletons via Language Supervision

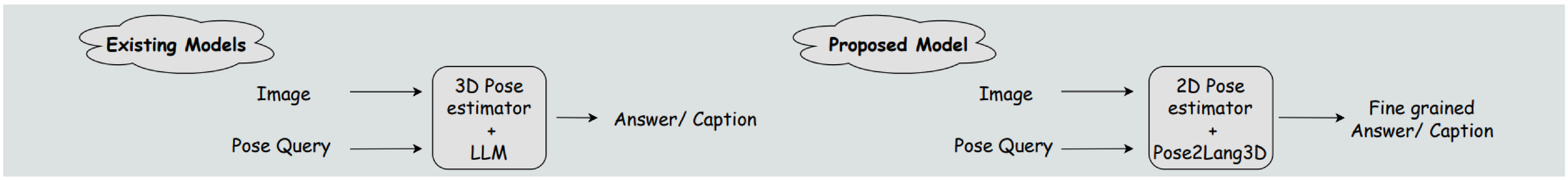
Authors Sai Bhargav Rongali , Kenji Okuma

Affiliations Honda Motor Co., Ltd, Tokyo, Japan

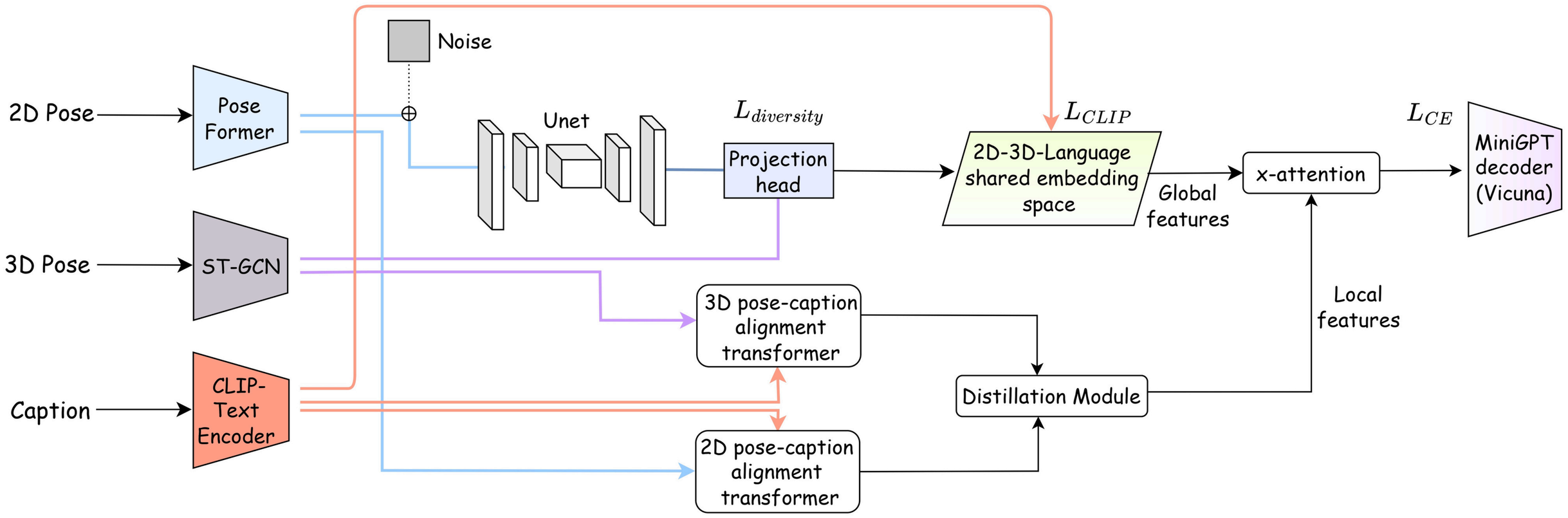
Introduction

Despite the progress in 3D human pose estimation, its reliance on expensive multi-view setups and limited dataset availability hinders scalability in real-world applications. We propose a novel framework that distills 3D spatial understanding into a language-aligned representation space using only 2D and 3D pose skeletons rendered as images. Our method learns a shared embedding space where 2D and 3D pose images are projected close to their corresponding natural language descriptions. During training, the model leverages 3D pose supervision to enrich semantic alignment, while at test time, it operates exclusively on 2D poses inferred from real images. This enables high-quality language-based reasoning such as action descriptions and question answering without an additional computational cost or supervision requirements of 3D pose estimators at inference. Our approach not only reduces reliance on 3D sensors but also demonstrates that 2D pose alone, when trained with 3D-informed language grounding, can achieve rich semantic understanding.

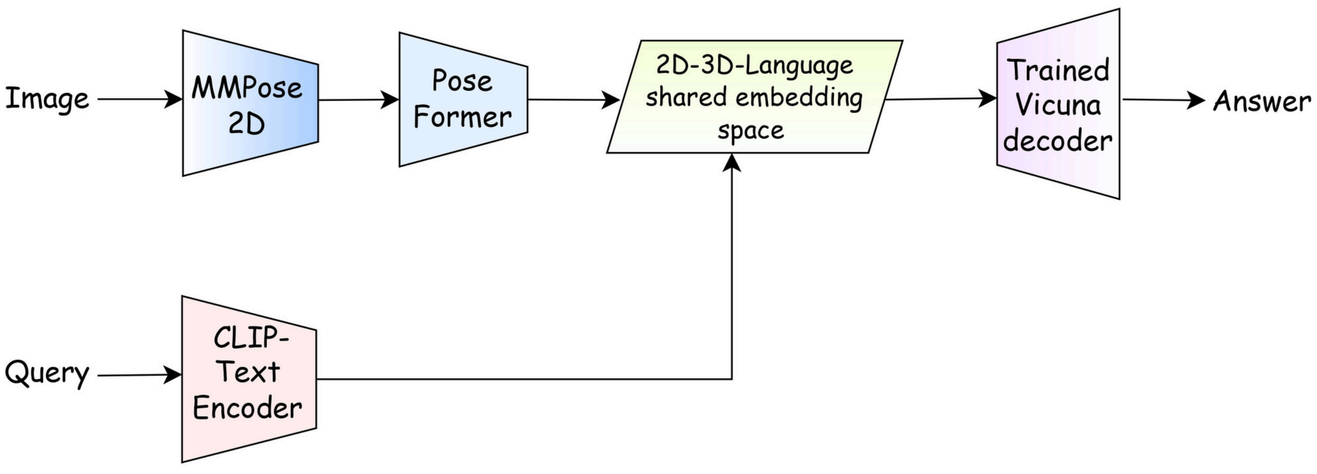
Existing Methods vs Our Method



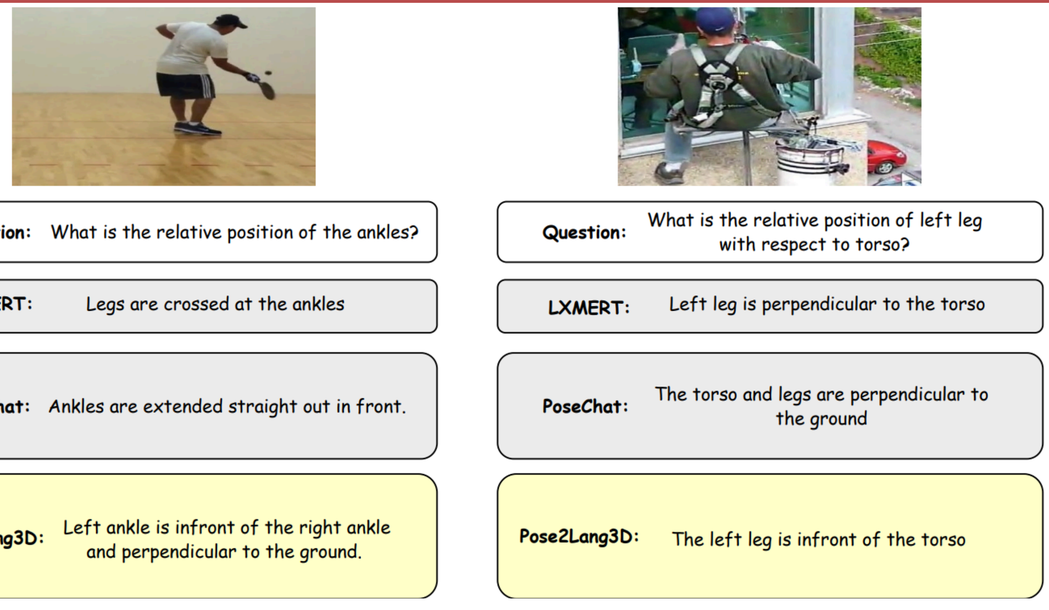
Training Methodology



Inference



Qualitative Results



Conclusion

- Pose2Lang3D distills 3D spatial reasoning into a 2D-only inference model via a shared embedding space and knowledge distillation, achieving fine-grained pose understanding without the computational overhead of 3D pose estimation at test time.
- The model outperforms all baselines on both VQA and captioning tasks — surpassing even 3D-based methods (PoseChat, PoseEmbroider) while running at just 30ms per sample, demonstrating a strong efficiency-accuracy trade-off.
- Strong generalization is shown across datasets (Human3.6M → MPII, MSCOCO), though rare/extreme poses and subtle similar actions remain limitations, with temporal modeling across video sequences as a key future direction.