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

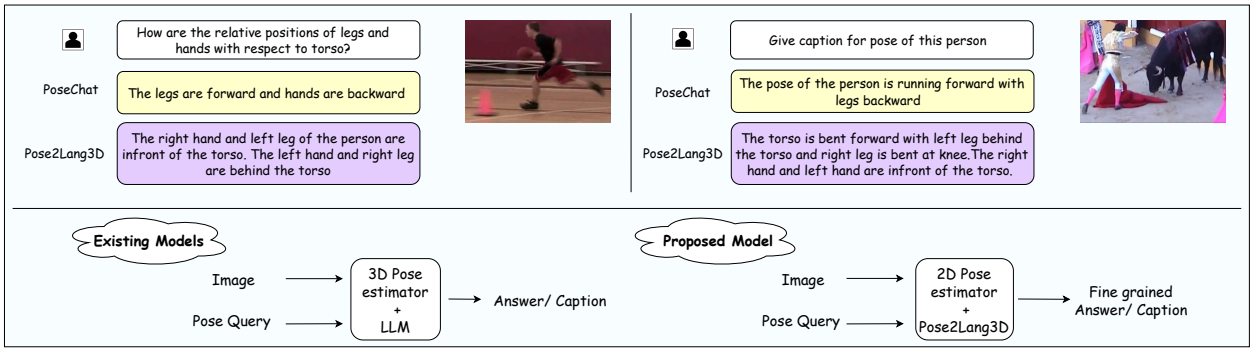

Figure 1: Our model proposes a framework to generate fine grained pose descriptions through 3D pose knowledge distillation using language supervision outperforming existing 3D Pose estimator & VLM approaches that produce generic and spatially unaware responses.

Abstract

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. Experiments on a newly curated dataset of 80K annotated pose images shows the effectiveness of our method, showing 20.8% and 44.1% improvements over 2D-only baselines and 1.8% and 1.3% improvements over 3D methods in VQA accuracy and BLEU-4 scores respectively.

1 Introduction

Understanding human pose is a critical problem in computer vision with applications spanning action recognition, human-computer interaction, and augmented reality. Traditional approaches to this problem have evolved along two primary directions: 2D pose estimation and 3D pose estimation. 2D pose estimation has seen remarkable progress with methods like OpenPose (Cao et al. 2019) and HRNet (Wang et al. 2020) achieving high accuracy in diverse environments. These approaches are computationally efficient and work reliably in unconstrained settings, making them suitable for real-time applications. However, they lack depth information crucial for understanding complex human actions and spatial relationships.

In contrast, 3D pose estimation methods (Liu et al. 2025), (Liu et al. 2021) provide richer spatial understanding but typically require specialized hardware setups like motion capture systems or multi-view cameras. This makes 3D pose estimation prohibitively expensive and impractical for many real-world scenarios, creating a significant barrier to deployment despite its superior representational capacity.

To address this, vision-language models have emerged as a promising alternative, offering better representations along with alignment in language space. However, these models often struggle with fine-grained pose-related queries due to their limited ability to reason about human body configurations. Attempts to address this limitation by incorporating 2D pose keypoints into multimodal frameworks like LXMERT (Tan and Bansal 2019), VilBERT (Lu et al. 2019), and LLaVA (Liu et al. 2023) have shown improvements but still fall short on tasks requiring spatial reasoning. More recently, PoseChat (Feng et al. 2024) demonstrated that incorporating 3D pose information can substantially enhance language-based reasoning about human actions. While effective, this approach introduces significant computational overhead, making it unsuitable for real-time applications due to high latency and resource requirements.

We identify a critical gap in current research: the need for a system that achieves 3D-level reasoning capabilities while maintaining the computational efficiency and accessibility of 2D pose estimation. To address this challenge, we introduce Pose2Lang3D, a novel framework that distills 3D knowledge using language into a model that operates solely on 2D skeletal inputs at inference time. Our key insight is

that while explicit 3D information is valuable, much of this knowledge can be implicitly learned through appropriate supervision during training. By leveraging language as a supervision signal, our approach enables a model to "think in 3D" despite processing only 2D inputs at test time. This distillation-based approach dramatically reduces computational requirements while preserving the rich reasoning capabilities associated with full 3D pose understanding.

The contributions of our work are as follows:

- We propose a knowledge distillation framework that transfers 3D reasoning capabilities to a model operating exclusively on 2D skeletal inputs, eliminating the need for complex 3D pose estimation at inference time.
- We introduce a language-supervised learning approach that grounds pose understanding in natural language, enabling more intuitive and semantically rich pose reasoning without explicit 3D representation.
- We demonstrate that our approach achieves comparable performance to methods using explicit 3D information while requiring significantly fewer parameters and simpler input requirements, making it suitable for real-time applications.

Pose2Lang3D represents a significant step toward making sophisticated 3D-aware pose understanding more accessible and practical for real-world applications. By bridging the gap between computationally efficient 2D pose estimation and semantically rich 3D reasoning, our work opens new possibilities for human-computer interaction systems that can understand and respond to nuanced human activities.

2 Related Work

Our work lies at the intersection of pose estimation, vision-language models, and knowledge distillation. In this section, we review relevant literature across these domains.

2.1 2D Pose Estimation

Human pose estimation in 2D has advanced significantly with deep learning approaches. Early deep learning methods (Toshev and Szegedy 2014) treated keypoint detection as a regression problem. Later, DeepCut (Pishchulin et al. 2016) and DeeperCut (Insafutdinov et al. 2016) introduced part-based models with integer linear programming for joint association. OpenPose (Cao et al. 2019) revolutionized the field with real-time multi-person pose estimation using Part Affinity Fields. Subsequently, top-down approaches like Mask R-CNN (He et al. 2017) and HRNet (Wang et al. 2020) achieved state-of-the-art performance by first detecting people and then estimating their poses. More recently, transformer-based methods (Li et al. 2021; Yang et al. 2021) have shown promising results by modeling long-range dependencies between keypoints.

These approaches are computationally efficient and work well in unconstrained environments, making them suitable for real-time applications. However, they lack depth information crucial for understanding complex spatial relationships and actions.

2.2 3D Pose Estimation

3D pose estimation methods can be broadly categorized into direct regression and 2D-to-3D lifting approaches. Direct regression methods (Pavlakos et al. 2017; Sun, Li, and Lin 2018) predict 3D joint coordinates directly from images. Pavlakos et al. (Pavlakos et al. 2017) used volumetric representations for 3D pose, while Sun et al. (Sun, Li, and Lin 2018) introduced integral regression for improved accuracy.

Lifting-based approaches (Martinez et al. 2017; Zhao et al. 2019; Pavllo et al. 2019) first estimate 2D poses and then "lift" them to 3D. Martinez et al. (Martinez et al. 2017) demonstrated that a simple MLP could effectively lift 2D poses to 3D. Temporal information has been leveraged in works like (Pavllo et al. 2019), which used 1D convolutions over sequential frames to improve 3D predictions. Multi-view approaches (Rhodin et al. 2018; Iskakov et al. 2019) utilize multiple camera perspectives to resolve depth ambiguities, achieving higher accuracy but requiring complex setups.

While these methods provide rich spatial understanding, they typically require specialized hardware, calibrated multi-view systems, or motion capture setups, creating significant barriers to deployment in real-world applications.

2.3 Vision-Language Models for Pose Understanding

General vision-language models like CLIP (Radford et al. 2021), LXMERT (Tan and Bansal 2019), and ViLBERT (Lu et al. 2019) have transformed visual reasoning but often struggle with fine-grained pose understanding. More recent large multimodal models like LLaVA (Liu et al. 2023) have improved capabilities but still face challenges with detailed pose analysis. Several works have attempted to enhance VQA systems specifically for pose reasoning. JRDB-Pose (Vendrow et al. 2023) built a specialized dataset for pose-related questions but was limited to simple spatial relationships. PoseFormer (Zheng et al. 2021) explored generating textual descriptions from pose sequences, while multimodal attention mechanisms (Farinhas, Martins, and Aguiar 2021) investigated the inverse task of generating poses from descriptions. These works highlighted the rich semantic information that can be captured through language-pose alignment but did not address the challenge of distilling 3D reasoning into 2D-only models.

Attention mechanisms for pose estimation (Liu et al. 2020) incorporated 2D pose features as additional inputs to attention mechanisms, demonstrating significant improvements in temporal context exploitation. VisualBERT (Li et al. 2019) integrated multimodal understanding with language processing through unified transformer architectures. Most recently, PoseChat (Feng et al. 2024) demonstrated significant improvements by incorporating 3D pose information into language models, enabling more accurate reasoning about complex spatial relationships. Despite its effectiveness, this approach introduces substantial computational overhead, making it impractical for real-time applications.

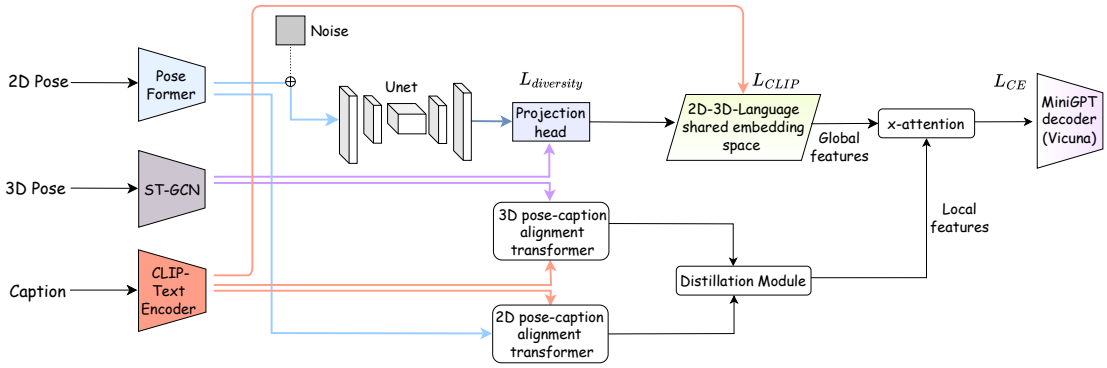

Figure 2: During training, we process triplets of ⟨2Dpose, 3D pose, language⟩to learn a unified embedding space. The encoded 2D, 3D poses are aligned through noise-augmented **U-Net** modules. Specialized alignment transformers establish correspondences between pose modalities and language descriptions. The distillation module transfers rich knowledge representations through cross-attention between global and local features to **miniGPT**.

2.4 Knowledge Distillation for Pose Analysis

Knowledge distillation (Hinton, Vinyals, and Dean 2015) transfers knowledge from a complex teacher model to a simpler student model. In the pose estimation domain, this approach has been applied to improve efficiency (Zhang, Zhu, and Ye 2019; Li et al. 2022). Zhang et al. (Zhang, Zhu, and Ye 2019) distilled knowledge from a high-capacity pose estimator to a lightweight model. Li et al. (Li et al. 2022), proposes an online knowledge distillation framework that trains a single multi-branch network eliminating the need for pretrained teacher models.

Specifically for 3D-to-2D distillation, Zhao et al. (Zhao et al. 2021) proposed a graph transformer that combines graph convolution with multi-head attention to effectively model joint relationships for 2D-to-3D pose estimation. Cheng et al. (Cheng et al. 2021) integrated top-down and bottom-up networks to exploit their complementary strengths. However, these approaches focused primarily on improving pose estimation accuracy rather than enhancing pose-language reasoning capabilities.

Our work builds upon these foundations by introducing a novel approach that distills 3D reasoning capabilities into a model that operates solely on 2D pose inputs, leveraging language supervision as the bridge between these representations. Unlike previous works that focused on either improving pose estimation accuracy or enhancing language understanding, Pose2Lang3D specifically addresses the challenge of achieving 3D-level reasoning without the computational overhead of explicit 3D pose estimation at inference time.

3 Methodology

In this section, we present our novel architecture for unifying 2D pose, 3D pose, and language representations in a shared embedding space. Our approach enables effective multimodal pose understanding and reasoning, which is crucial for human-centric visual understanding tasks. Figure 2 illustrates our proposed methodology.

3.1 Problem Formulation

Given an input image $I \in \mathbb{R}^{H\times W\times 3}$ and a query $Q \in \mathcal{V}^{L_q}$ where $\mathcal{V}$ is the vocabulary and L_q is the query length, our goal is to generate a response $R \in \mathcal{V}^{L_r}$ that accurately answers fine-grained pose-related questions. Our framework addresses the challenge of bridging the semantic gap between 2D visual observations and 3D pose understanding through language, enabling detailed pose reasoning even when only 2D pose estimation is available at inference time.

3.2 Architecture Overview

Our framework consists of four main components that work synergistically to achieve multimodal pose understanding. The **Multimodal Pose Encoders** extract representations from 2D and 3D poses using specialized architectures. The **Noise-Augmented Feature Enhancement** module employs a U-Net architecture to improve feature robustness. The **Unified Embedding Space** incorporates specialized alignment transformers for cross-modal correspondence learning. Finally, the **Knowledge Distillation Module** facilitates cross-modal feature refinement through attention mechanisms. The architecture operates in two phases: a training phase where we learn to align representations from different modalities, and an inference phase where we leverage the learned representations to answer pose-related queries.

3.3 Multimodal Pose Representation

2D Pose Encoding For 2D pose extraction, we employ MMPose (Contributors 2020) to obtain keypoints $P^{2D} = \{p_i^{2D}\}_{i=1}^{K}$ where $p_i^{2D} \in \mathbb{R}^2$ represents the (x, y) coordinates of the i-th keypoint, and K is the total number of keypoints. We encode these keypoints using a Pose Former network. The PoseFormer consists of N_p transformer blocks with multi-head self-attention mechanisms, specifically designed to capture spatial dependencies between keypoints and temporal consistency across frames.

$$F^{2D} = \text{PoseFormer}(P^{2D}) \in \mathbb{R}^{d_{pose}} \quad (1)$$

3D Pose Encoding For 3D poses, we have keypoints $P^{3D} = \{p_i^{3D}\}_{i=1}^{K}$ where $p_i^{3D} \in \mathbb{R}^3$ represents the (x, y, z) coordinates. We employ a Spatial-Temporal Graph Convolutional Network (ST-GCN) (Yan, Xiong, and Lin 2018) to

model the spatial-temporal dependencies.The ST-GCN architecture leverages the inherent skeletal structure of human poses, with adjacency matrices defined according to anatomical connections between body joints.

$$F^{3D} = \text{ST-GCN}(P^{3D}) \in \mathbb{R}^{d_{pose}} \tag{2}$$

Language Encoding To encode textual information, we use a CLIP-Text Encoder for both captions C during training and queries Q during inference.

$$F^C = \text{CLIP-Text}(C) \in \mathbb{R}^{d_{\text{text}}},\ F^Q = \text{CLIP-Text}(Q) \in \mathbb{R}^{d_{\text{text}}} \tag{3}$$

where F^C and F^Q are the caption and query feature representations, respectively.

3.4 Noise-Augmented Feature Enhancement

A key innovation in our approach is the integration of a noise-augmented feature enhancement module using a U-Net architecture. This component serves two purposes: improving the robustness of pose features through controlled noise injection, and facilitating better feature alignment across modalities.The U-Net takes pose features and controlled noise as input.

$$\tilde{F}^{2D} = \text{U-Net}(F^{2D} \oplus \mathcal{N}(0, \sigma^2)) \in \mathbb{R}^{d_{pose}} \tag{4}$$

$$\tilde{F}^{3D} = \text{U-Net}(F^{3D} \oplus \mathcal{N}(0, \sigma^2)) \in \mathbb{R}^{d_{pose}} \tag{5}$$

where $\oplus$ denotes element-wise addition, $\mathcal{N}(0, \sigma^2)$ is Gaussian noise, and σ is a learnable parameter that adapts during training.

3.5 Unified Embedding Space with Alignment

We construct a shared 2D-3D-Language embedding space through a projection head followed by specialized alignment transformers. The projection head maps features from different modalities into a unified space.

$$E^t = \text{Projection}(\tilde{F}^t),\ t \in \{2D,\ 3D,\ C\},\ E^t \in \mathbb{R}^{d_{\text{joint}}} \tag{6}$$

2D Pose-Caption Alignment Transformer The 2D pose-caption alignment transformer establishes fine-grained correspondences between 2D pose features and language representations. Given the projected 2D pose embeddings E^{2D} and caption embeddings E^C, we introduce a **projection-aware attention mechanism** that accounts for depth ambiguity in 2D poses.

The attention weights are computed with an uncertainty-modulated scoring function:

$$\text{Score}(Q, K) = \frac{QK^T}{\sqrt{d_k}} + \lambda \log \sigma^{2D} \tag{7}$$

where $\sigma^{2D} \in \mathbb{R}^{N_j}$ represents learned uncertainty estimates for each 2D joint due to depth ambiguity, and λ is a weighting factor. The multi-head attention becomes:

$$\text{head}_i = \text{softmax}(\text{Score}(QW_i^Q, KW_i^K))VW_i^V \tag{8}$$

$$\hat{E}^{2D} = \text{LayerNorm}(E^{2D} + \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O) \tag{9}$$

This uncertainty-aware mechanism allows the model to dynamically weight attention based on the reliability of 2D pose estimates.

3D Pose-Caption Alignment Transformer The 3D pose-caption alignment transformer learns correspondences between 3D pose features and language representations while preserving spatial geometry. We augment the standard attention with **depth-conditioned position embeddings** that encode both skeletal topology and 3D spatial relationships.

The 3D positional encoding incorporates bone length ratios and joint angles:

$$E^{3D}_{pos} = E^{3D} + \phi_{bone}(P^{3D}) + \phi_{angle}(P^{3D}) \tag{10}$$

where ϕ_{bone} encodes normalized bone length vectors and ϕ_{angle} captures joint angle configurations using sinusoidal encodings:

$$\phi^{(i,j)}_{bone} = \text{MLP}\left(\frac{\|p_j - p_i\|}{\sum_k \|p_k - p_{parent(k)}\|}\right) \tag{11}$$

$$\phi^{(i)}_{angle} = [\sin(\theta_i/\tau^{2k/d}), \cos(\theta_i/\tau^{2k/d})]^{d/2}_{k=0} \tag{12}$$

where θ_i is the joint angle at joint i. The attention mechanism operates on these geometry-enriched embeddings:

$$\hat{E}^{3D} = \text{LayerNorm}(E^{3D}_{pos} + \text{MultiHead}(E^{3D}_{pos}, E^C, E^C)) \tag{13}$$

This depth-conditioned encoding ensures that the 3D alignment captures biomechanically meaningful pose-language correspondences, which are then distilled to guide 2D representation learning.

3.6 Knowledge Distillation Module

The knowledge distillation module is designed to transfer rich 3D structural knowledge to 2D pose representations, enabling 2D features to benefit from 3D understanding during inference. The module operates through a teacher-student paradigm where 3D features act as the teacher and 2D features as the student.The distillation process begins by computing attention weights between aligned 2D and 3D features $A_{2D\to3D}$. The distilled local features F^{local} are obtained from 2D pose caption alignment transformers and global features are extracted from the 3D representations using adaptive average pooling F^{global}. Further, the distillation module employs a gating mechanism to control the flow of information G.

$$A_{2D\to3D} = \text{softmax}\left(\frac{\hat{E}^{2D}(\hat{E}^{3D})^T}{\sqrt{d_{joint}}}\right) \tag{14}$$

$$F^{local} = A_{2D\to3D} \cdot \hat{E}^{3D} \in \mathbb{R}^{d_{joint}} \tag{15}$$

$$F^{global} = \text{AdaptivePool}(\hat{E}^{3D}) \in \mathbb{R}^{d_{joint}} \tag{16}$$

$$G = \sigma(W_g[F^{local}; F^{global}] + b_g) \tag{17}$$

where σ is the sigmoid function, $W_g \in \mathbb{R}^{d_{joint} \times 2d_{joint}}$, and $[;]$ denotes concatenation.

3.7 Cross-Attention for Global and Local Feature Fusion

The cross-attention mechanism fuses global and local features to produce enhanced representations that capture both holistic pose understanding and fine-grained local details. The cross-attention operation is formulated as:

$$\text{CrossAttn}(F^{global}, F^{local}) = \text{Attention}(F^{global}, F^{local}, F^{local}) \tag{18}$$

The enhanced features are computed through a residual connection with layer normalization:

$$F^{enhanced} = \text{LayerNorm}(F^{global} + \text{CrossAttn}(F^{global}, F^{local})) \tag{19}$$

To further refine the features, we apply a feed-forward network with residual connections:

$$F^{refined} = \text{LayerNorm}(F^{enhanced} + \text{FFN}(F^{enhanced})) \tag{20}$$

where $\text{FFN}(x) = W_2 \cdot \text{ReLU}(W_1 \cdot x + b_1) + b_2$ is a two-layer feed-forward network.

3.8 Training Objectives

Our training process involves multiple objectives to ensure effective alignment and knowledge transfer across modalities. To preserve the uniqueness of each modality while encouraging information sharing, we introduce a diversity loss $\mathcal{L}_{diversity}$.To align visual pose representations with language, we adopt $\mathcal{L}_{CLIP}$. We introduce a distillation loss $\mathcal{L}_{distill}$ to transfer 3D knowledge to 2D representations and to improve the language generation we use cross entropy loss $\mathcal{L}_{CE}$. The overall training objective is a weighted combination of the above losses $\mathcal{L}_{total}$.

$$\mathcal{L}_{diversity} = \alpha \cdot \text{MSE}(\hat{E}^{2D}, \hat{E}^{3D}) + \beta \cdot (1 - \cos(\hat{E}^{2D}, \hat{E}^{3D})) \tag{21}$$

where α and β are hyperparameters, MSE is the mean squared error, and cos denotes cosine similarity.

$$\mathcal{L}_{CLIP} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\hat{E}_i^{2D}, E_i^C)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\hat{E}_i^{2D}, E_j^C)/\tau)} \tag{22}$$

where $\text{sim}(\cdot, \cdot)$ computes cosine similarity, τ is a temperature parameter, and N is the batch size.

$$\mathcal{L}_{distill} = \text{KL}(\text{softmax}(F^{local}/T), \text{softmax}(\hat{E}^{3D}/T)) \tag{23}$$

where T is the temperature parameter.

$$\mathcal{L}_{CE} = -\frac{1}{L_r} \sum_{t=1}^{L_r} \log p(r_t | r_{<t}, F^{refined}, E^C) \tag{24}$$

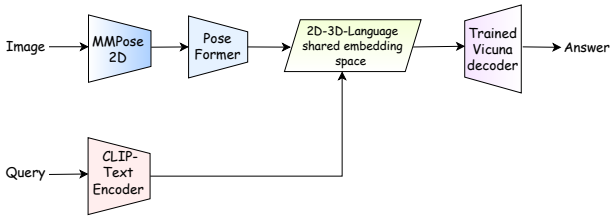

Figure 3: **Inference** pipeline takes image and query as input and using MMPose and pose former extracts 2D pose embeddings and query embeddings from CLIP-text encoder are projected into the learned embedding space which is decoded by the **vicuna** decoder

where $p(r_t | r_{<t}, F^{refined}, E^C)$ is the probability of generating token r_t given previous tokens and multimodal embeddings.

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{diversity} + \lambda_2 \mathcal{L}_{CLIP} + \lambda_3 \mathcal{L}_{distill} + \lambda_4 \mathcal{L}_{CE} \tag{25}$$

where λ_1, λ_2, λ_3, and λ_4 are hyperparameters controlling the contribution of each loss term.

3.9 Inference Pipeline

During inference, given an input image I and a query Q, we follow these steps. First, we extract 2D keypoints using MMPose: $P^{2D} = \text{MMPose}(I)$. Next, we encode the 2D pose: $F^{2D} = \text{PoseFormer}(P^{2D})$, which internally applies noise-augmented enhancement: $\tilde{F}^{2D} = \text{U-Net}(F^{2D} \oplus \mathcal{N}(0, \sigma^2))$. The features are projected to the joint space: $E^{2D} = \text{Projection}(\tilde{F}^{2D})$. We encode the query: $F^Q = \text{CLIP-Text}(Q)$ and project it to the joint space: $E^Q = \text{Projection}(F^Q)$. The alignment transformer is applied: $\hat{E}^{2D} = \text{Alignment-Transformer}_{2D}(E^{2D}, E^Q)$. We generate enhanced features through cross-attention: $F^{refined} = \text{CrossAttn}(F^{global}, F^{local})$. Finally, we generate the answer using the MiniGPT decoder:

$$R = \text{MiniGPT}(F^{refined}, E^Q) \tag{26}$$

4 Experiments

4.1 Experimental Setup

Datasets We evaluate our approach using Human3.6M (Ionescu et al. 2014), which provides 80K poses with paired 2D and 3D keypoint coordinates. We enriched this dataset by generating fine-grained language descriptions for each pose through a carefully designed template that captures detailed information about body parts and their relative positions. Using these descriptions, we created a comprehensive question-answer dataset by prompting LLaVA-1.5-7b to generate three categories of questions: (1) low-level questions focusing on fine-grained pose details, (2) medium-level questions about body configurations, and (3) high-level questions addressing overall pose stability and environmental interactions.

Table 1: **Performance evaluation** of our model against baseline methods on pose understanding tasks. Our approach achieves superior results while operating solely on 2D pose inputs at inference, demonstrating effective knowledge distillation from 3D supervision during training. All models use HRNet as the 2D pose estimator. Best results are in **bold**, second best are underlined.

Method	Test Input	3D Supervision	MPII				MS COCO				Runtime (ms)
			VQA		Captioning		VQA		Captioning		
			Acc (%)	F1	BLEU-4	SPICE	Acc (%)	F1	BLEU-4	SPICE	
HRNet (Wang et al. 2020) + BART (Lewis et al. 2019)	2D Pose	✗	68.9	60.2	31.3	0.22	64.2	55.8	28.6	0.19	28± 3
VilBERT (Lu et al. 2019)	Image+2D	✗	76.9	67.8	40.2	0.30	70.9	63.5	36.5	0.27	58 ± 4
LLaVA (Liu et al. 2023)	Image+2D	✗	77.2	68.9	41.0	0.33	73.1	65.3	37.8	0.29	62 ± 3
LXMERT (Tan and Bansal 2019)	Image+2D	✗	77.5	68.2	40.8	0.31	72.4	64.7	37.2	0.28	55± 2
PoseChat (Feng et al. 2024)	3D Pose	✓	81.7	74.5	44.5	0.36	77.9	71.2	41.3	0.34	35± 8
PoseEmbroider (Delmas et al. 2024)	3D Pose	✓	82.4	75.8	44.8	0.37	78.6	72.1	41.0	0.33	38± 5
Pose2Lang3D (Ours)	2D Pose	✓	**83.2**	**76.3**	**45.1**	**0.38**	**79.5**	**72.8**	**41.3**	**0.34**	30± 4

For training, Pose2Lang3D uses the augmented Human3.6M dataset, while evaluation is performed on the MPII (Andriluka et al. 2014), MSCOCO (Lin et al. 2015) datasets which provides both pose keypoints and corresponding images for real-world testing. This cross-dataset evaluation demonstrates our model's ability to generalize to unseen poses in naturalistic settings.

Metrics We evaluate our approach on two tasks: pose-based Visual Question Answering (VQA) and pose captioning. For VQA, we report accuracy for multiple-choice questions and F1 score for open-ended ones. Captioning performance is measured using BLEU-4 and SPICE. Additionally, we assess computational efficiency via inference time (milliseconds per sample) and model size (millions of parameters).

Baselines We compare Pose2Lang3D against several strong baselines. PoseChatleverages explicit 3D pose information at inference but requires full 3D pose inputs at test time. LXMERT and VilBERT, two vision-language models, are extended to accept 2D pose features from HRNet alongside images. LLaVa, a large-scale vision-language model, is adapted to incorporate pose information. We also consider Pose-guided VQA, which uses only 2D pose features with language supervision, and HRNet+BART, a simple combination of 2D pose extraction and a pretrained language model. For fairness, all methods use HRNet for 2D pose estimation.

4.2 Main Results

Pose2Lang3D consistently outperforms all baselines on VQA and captioning tasks. It achieves a 5.7% higher VQA accuracy and a 4.1% improvement in BLEU-4 compared to the strongest 2D-only baseline. Notably, it surpasses both PoseChat and the recent PoseEmbroider (Delmas et al. 2024) despite relying solely on 2D poses at inference, demonstrating the efficacy of our knowledge distillation strategy. Moreover, Pose2Lang3D outperforms more complex vision language models that use both images and poses. Our method achieves a runtime of 30ms per sample, faster than 3D-based methods (PoseEmbroider: 38ms, PoseChat: 35ms), while delivering superior accuracy, highlighting its optimal efficiency-accuracy trade-off.

Table 2: Ablation study on key components of Pose2Lang3D. The three core components are: 3D pose caption alignment (3D-PC), 2D pose caption alignment (2D-PC), and the Unet-based embedding space (UNet) on MPII dataset. Best results are in **bold**.

Components			VQA	Captioning	
3D-PC Align.	2D-PC Align.	UNet.	Acc (%)	B-4	SPICE
✗	✓	✗	50.8	20.8	0.11
✗	✗	✓	52.1	18.4	0.13
✓	✗	✗	56.9	22.5	0.20
✓	✗	✓	58.3	24.7	0.21
✗	✓	✓	60.6	33.2	0.30
✓	✓	✗	80.2	42.5	0.36
✓	✓	✓	**83.2**	**45.1**	**0.38**

4.3 Ablation Studies

We perform ablation studies to analyze the contributions of key components in our architecture. Removing 3D supervision leads to the largest performance drop, 22.6% in VQA accuracy and 11.9% in BLEU-4, confirming the importance of distilled 3D knowledge for fine-grained pose understanding. Eliminating 2D-language supervision degrades both VQA and captioning performance, while excluding the cross-modal UNet embedding alignment reduces VQA accuracy by 3.0% and BLEU-4 by 2.6%. These results underscore the complementary benefits of each component as shown in Table 2. Table 3 shows the improtance of global and local feature alignment which improves the accuracy by 5% and also shows the performance of our architecture with only 2D components and 3D components.

Further analysis of loss functions shows that diversity loss ($\mathcal{L}_{\text{diversity}}$) preserves modality-specific features and yields modest gains. Removing the CLIP contrastive loss ($\mathcal{L}_{\text{CLIP}}$) leads to substantial drops, highlighting its critical role in aligning pose and language representations. As expected, cross-entropy loss ($\mathcal{L}_{\text{CE}}$) is essential for language generation, and its removal severely impacts performance. Table 4 presents the results with different loss combinations.

Architectural Variations We evaluate various architectural configurations of the pose encoders and projection head (Table 5). Reducing the number of PoseFormer transformer blocks from 6 to 4 results in a minor performance decrease

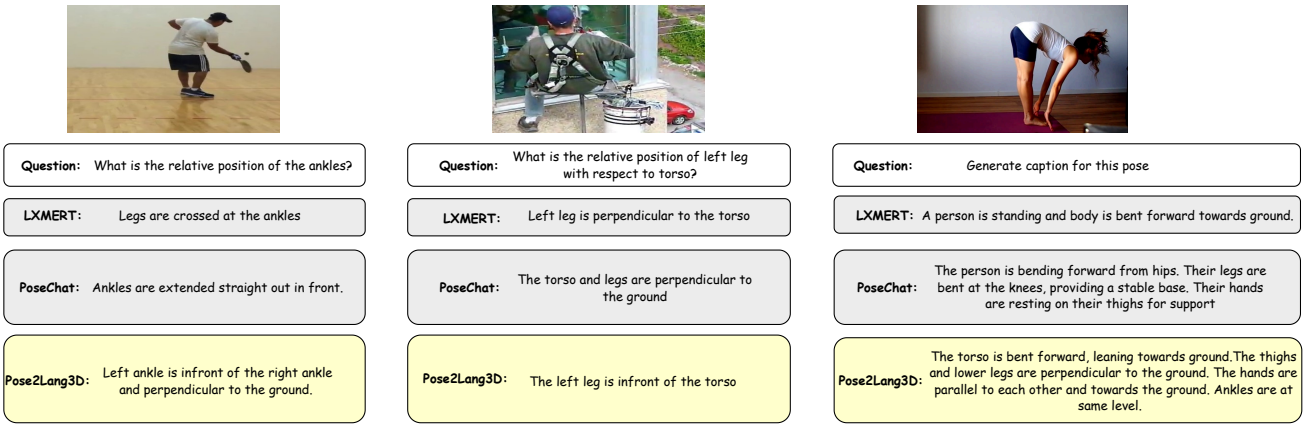

Figure 4: **Qualitative results** on MPII dataset (Andriluka et al. 2014). Pose2Lang3D is fine-grained while captioning the poses and correct while answering the fine-grained questions where are other models struggle to answer finegrained questions and while captioning.

Table 3: Comparison of different model architectures. We compare Pose2Lang3D with 2D-only and 3D-only variants, as well as models without global and local feature alignment. All models use language supervision and results on MPII dataset. Best results are in **bold**.

Model	VQA	Captioning	
	Acc(%)	B-4	SPICE
2D-only model	60.6	33.2	0.24
w/o Global & Local feature alignment	78.3	40.5	0.31
3D-only model	81.7	44.5	0.33
Pose2Lang3D(ours)	**83.2**	**45.1**	**0.38**

Table 4: Ablation study on loss terms in Pose2Lang3D. We evaluate the contribution of diversity loss ($\mathcal{L}_{\text{diversity}}$), CLIP contrastive loss ($\mathcal{L}_{\text{CLIP}}$), and cross-entropy loss ($\mathcal{L}_{\text{CE}}$) on MPII dataset. Best results are in **bold**.

Loss Terms			VQA		Captioning	
$\mathcal{L}_{\text{div}}$	$\mathcal{L}_{\text{CLIP}}$	$\mathcal{L}_{\text{CE}}$	Acc	F1	B-4	CIDEr
✓	✓	✗	67.2	61.5	30.8	0.84
✓	✗	✓	78.3	71.4	40.5	1.10
✗	✓	✓	81.6	74.8	43.9	1.19
✓	✓	✓	**83.2**	**76.3**	**45.1**	**1.23**

which is 0.7% in VQA accuracy and 0.9% in BLEU-4 score, while significantly reducing model size by 14M parameters. Similarly, decreasing ST-GCN graph convolutional layers from 9 to 6 only slightly affects performance, indicating that fewer 3D pose encoding layers suffice to maintain strong results. Regarding the projection head, the default two-layer MLP achieves the best balance, with a simpler single-layer variant yielding slightly lower accuracy and a deeper three-layer head offering negligible improvement at the cost of increased complexity.These results demonstrate that our architectural choices strike a good balance between model complexity and performance. In Figure 4 we show the qualitative results of our model which shows fine grained descriptions

Table 5: Ablation study on architectural choices in Pose2Lang3D. We evaluate different pose encoders and projection head designs on MPII dataset. Best results are in **bold**.

Architecture		VQA		Captioning	
PoseFormer	ST-GCN	Acc (%)	F1	B-4	SPICE
4 blocks	6 layers	81.9	74.8	43.5	0.34
4 blocks	9 layers	82.5	75.6	44.2	0.35
6 blocks	6 layers	82.8	75.9	44.7	0.36
6 blocks	9 layers	**83.2**	**76.3**	**45.1**	**0.38**
1-layer MLP projection		82.6	75.7	44.3	0.35
3-layer MLP projection		83.0	76.1	45.0	0.37

when compared to the other models.

5 Conclusion

We introduced Pose2Lang3D, a framework that distills 3D spatial reasoning capabilities into a model operating solely on 2D skeletal inputs at inference time. We also propose a dataset of 80K 2D, 3D pose images with rich fine grained descriptions. Our knowledge distillation approach and shared embedding space enable high-quality language-based reasoning without the computational overhead of 3D pose estimation. Experiments demonstrate that Pose2Lang3D outperforms existing methods in both VQA and captioning tasks while maintaining efficiency comparable to 2D-only models. Our approach shows strong performance across multiple architectural variants and robust generalization from Human3.6M to MPII and MSCOCO. However, in case of rare or extreme poses, similar and subtle actions are the regions where the current model struggles and would be the future work directions incorporating the temporal information which can also scale across video sequences to better distinguish between nuanced movements.

References

Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. arXiv:1812.08008.

Cheng, Y.; Wang, B.; Yang, B.; and Tan, R. T. 2021. Monocular 3D Multi-Person Pose Estimation by Integrating Top-Down and Bottom-Up Networks. arXiv:2104.01797.

Contributors, M. 2020. OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose.

Delmas, G.; Weinzaepfel, P.; Moreno-Noguer, F.; and Rogez, G. 2024. PoseEmbroider: Towards a 3D, Visual, Semantic-aware Human Pose Representation. arXiv:2409.06535.

Farinhas, A.; Martins, A. F. T.; and Aguiar, P. M. Q. 2021. Multimodal Continuous Visual Attention Mechanisms. arXiv:2104.03046.

Feng, Y.; Lin, J.; Dwivedi, S. K.; Sun, Y.; Patel, P.; and Black, M. J. 2024. ChatPose: Chatting about 3D Human Pose. arXiv:2311.18836.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. arXiv:1503.02531.

Insafutdinov, E.; Pishchulin, L.; Andres, B.; Andriluka, M.; and Schiele, B. 2016. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, 34–50. Springer.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7): 1325–1339.

Iskakov, K.; Burkov, E.; Lempitsky, V.; and Malkov, Y. 2019. Learnable Triangulation of Human Pose. arXiv:1905.05754.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv:1910.13461.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv:1908.03557.

Li, Y.; Zhang, S.; Wang, Z.; Yang, S.; Yang, W.; Xia, S.-T.; and Zhou, E. 2021. TokenPose: Learning Keypoint Tokens for Human Pose Estimation. arXiv:2104.03516.

Li, Z.; Ye, J.; Song, M.; Huang, Y.; and Pan, Z. 2022. Online Knowledge Distillation for Efficient Pose Estimation. arXiv:2108.02092.

Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollár, P. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. arXiv:2304.08485.

Liu, J.; Liu, M.; Liu, H.; and Li, W. 2025. TCPFormer: Learning Temporal Correlation with Implicit Pose Proxy for 3D Human Pose Estimation. arXiv:2501.01770.

Liu, R.; Shen, J.; Wang, H.; Chen, C.; Cheung, S.-c.; and Asari, V. 2020. Attention Mechanism Exploits Temporal Contexts: Real-Time 3D Human Pose Reconstruction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5063–5072.

Liu, W.; Bao, Q.; Sun, Y.; and Mei, T. 2021. Recent Advances in Monocular 2D and 3D Human Pose Estimation: A Deep Learning Perspective. arXiv:2104.11536.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. arXiv:1908.02265.

Martinez, J.; Hossain, R.; Romero, J.; and Little, J. J. 2017. A simple yet effective baseline for 3d human pose estimation. arXiv:1705.03098.

Pavlakos, G.; Zhou, X.; Derpanis, K. G.; and Daniilidis, K. 2017. Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose. arXiv:1611.07828.

Pavllo, D.; Feichtenhofer, C.; Grangier, D.; and Auli, M. 2019. 3D human pose estimation in video with temporal convolutions and semi-supervised training. arXiv:1811.11742.

Pishchulin, L.; Insafutdinov, E.; Tang, S.; Andres, B.; Andriluka, M.; Gehler, P. V.; and Schiele, B. 2016. DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Rhodin, H.; Spörri, J.; Katircioglu, I.; Constantin, V.; Meyer, F.; Müller, E.; Salzmann, M.; and Fua, P. 2018. Learning Monocular 3D Human Pose Estimation from Multi-view Images. arXiv:1803.04775.

Sun, X.; Li, C.; and Lin, S. 2018. An Integral Pose Regression System for the ECCV2018 PoseTrack Challenge. *arXiv preprint arXiv:1809.06079*.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. arXiv:1908.07490.

Toshev, A.; and Szegedy, C. 2014. DeepPose: Human Pose Estimation via Deep Neural Networks. In *2014 IEEE*

Conference on Computer Vision and Pattern Recognition, 1653–1660. IEEE.

Vendrow, E.; Le, D. T.; Cai, J.; and Rezatofighi, H. 2023. JRDB-Pose: A Large-scale Dataset for Multi-Person Pose Estimation and Tracking. arXiv:2210.11940.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; Liu, W.; and Xiao, B. 2020. Deep High-Resolution Representation Learning for Visual Recognition. arXiv:1908.07919.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. arXiv:1801.07455.

Yang, S.; Quan, Z.; Nie, M.; and Yang, W. 2021. TransPose: Keypoint Localization via Transformer. arXiv:2012.14214.

Zhang, F.; Zhu, X.; and Ye, M. 2019. Fast Human Pose Estimation. arXiv:1811.05419.

Zhao, L.; Peng, X.; Tian, Y.; Kapadia, M.; and Metaxas, D. N. 2019. Semantic Graph Convolutional Networks for 3D Human Pose Regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3420–3430. IEEE.

Zhao, W.; Tian, Y.; Ye, Q.; Jiao, J.; and Wang, W. 2021. GraFormer: Graph Convolution Transformer for 3D Pose Estimation. arXiv:2109.08364.

Zheng, C.; Zhu, S.; Mendieta, M.; Yang, T.; Chen, C.; and Ding, Z. 2021. 3D Human Pose Estimation with Spatial and Temporal Transformers. arXiv:2103.10455.

Reproducibility Checklist

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no)

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no)

2.4. Proofs of all novel claims are included (yes/partial/no)

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no)

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no)

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA)

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA)

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) partial

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) yes

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) yes

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) no

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix

(yes/partial/no) no

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) partial

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) no

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) yes

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) partial

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) no

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) no

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) yes