# Unified 2D-3D Discrete Priors for Noise-Robust and Calibration-Free Multiview 3D Human Pose Estimation

**Geng Chen*** **Pengfei Ren*** **Xufeng Jian** **Haifeng Sun†** **Menghao Zhang**
**Qi Qi** **Zirui Zhuang** **Jing Wang** **Jianxin Liao** **Jingyu Wang†**
State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
{chengeng, rpf, jianxf, hfsun, zhangmenghao, qiqi8266, zhuangzirui,
wangjing, liaojx, wangjingyu}@bupt.edu.cn

## Abstract

Multi-view 3D human pose estimation (HPE) leverages complementary information across views to improve accuracy and robustness. Traditional methods rely on camera calibration to establish geometric correspondences, which is sensitive to calibration accuracy and lacks flexibility in dynamic settings. Calibration-free approaches address these limitations by learning adaptive view interactions, typically leveraging expressive and flexible continuous representations. However, as the multiview interaction relationship is learned entirely from data without constraint, they are vulnerable to noisy input, which can propagate, amplify and accumulate errors across all views, severely corrupting the final estimated pose. To mitigate this, we propose a novel framework that integrates a noise-resilient discrete prior into the continuous representation-based model. Specifically, we introduce the *Uni-Codebook*, a unified, compact, robust, and discrete representation complementary to continuous features, allowing the model to benefit from robustness to noise while preserving regression capability. Furthermore, we propose an attribute-preserving and complementarity-enhancing Discrete-Continuous Spatial Attention (DCSA) mechanism to facilitate interaction between discrete priors and continuous pose features. Extensive experiments on three representative datasets demonstrate that our approach outperforms both calibration-required and calibration-free methods, achieving state-of-the-art performance.

## 1 Introduction

3D human pose estimation aims to estimate the 3D locations of keypoints from images or videos. It is important because 3D human skeletons are widely used in scenarios such as action recognition [21, 25], human mesh recovery [8, 12], and robotics manipulation [7, 34]. Although monocular pose estimation offers greater convenience in usage, its performance is limited by depth ambiguity and occlusion issues, which hinder its broader application. Multiview systems [36, 13, 22, 31] offer a promising solution by leveraging complementary information from different views, enhancing accuracy and robustness in pose estimation. Existing multiview methods utilize the geometric constraints provided by camera extrinsics to establish correspondences between views, facilitating the fusion of complementary information to enhance 3D pose prediction. There are several drawbacks to using camera calibration-based methods. First, due to the complexity and computational intensity of camera

*Equal contribution.
†Corresponding authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

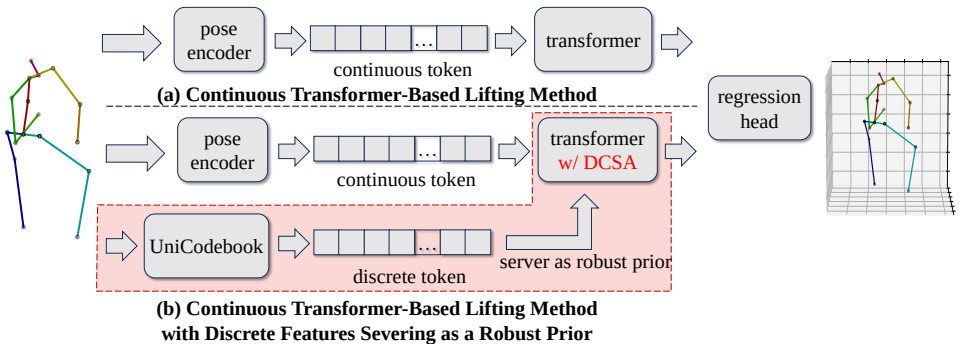


Figure 1: (a) Continuous transformer-based lifting method, which directly processes 2D pose inputs to estimate 3D poses. (b) Proposed method, which integrates discrete features as a robust prior within a continuous transformer-based framework, enhancing robustness to noisy 2D inputs and improving pose estimation accuracy.

calibration, these methods incur higher costs and exhibit limited flexibility. Second, their performance is sensitive to the precision of camera extrinsics, which can be unsatisfactory in complex or dynamic environments. Therefore, uncalibrated multi-view pose estimation has emerged as a significant trend, as it avoids the impact of unreliable camera extrinsics and is more flexible in deployment, leading to better generalization capability and broader applicability in real-world scenarios.

Many uncalibrated methods leverage adaptive multiview interaction patterns, relying on semantic intra-view and inter-view relationships to eliminate reliance on rigid geometric constraints. In recent years, purely transformer-based methods [32, 1, 2] have attracted increasing attention in calibration-free pose estimation, as they provide a unified and scaleable framework for modeling spatial, view, and temporal relationships using attention mechanisms. Notably, FusionFormer [1] and PoseIRM [2] have exhibited performance that surpasses even that of calibrated methods. This observation suggests that uncalibrated methods, rather than being constrained by predefined geometric relationships, can instead exploit data-driven adaptive semantic constraints. Such flexibility enables them to more effectively capture context-aware dependencies across views, thereby learning richer and more robust feature representations. For example, calibrated methods depend heavily on geometric information from external calibration, which occlusions can easily disrupt. In contrast, without calibration, uncalibrated methods are forced to leverage multi-view interactions to dynamically correct errors.

The superiority of transformer-based calibration-free framework can be attributed to their expressive and flexible continuous representations, which enable fine-grained modeling of pose dependencies when equipped with attention mechanism. Their continuity aligns with the continuous nature of both input and output spaces in 2D-to-3D pose lifting, enabling smooth, consistent, and precise mappings that are free from quantization errors and easy to optimize. However, this flexibility comes at the cost of increased susceptibility to overfitting when trained on limited data [5], making them sensitive to noisy 2D poses caused by occlusions and other common real-world challenges. Such errors can propagate, amplify and accumulate across views, affecting the accuracy of predictions from other views. Human pose is fundamentally constrained by inherent anatomical priors. These biological constraints, such as skeletal structure, joint motion limitations, and other biomechanical properties, define a natural boundary for valid poses. By encoding such anatomical knowledge as biomechanical priors in pose estimation models, we can intrinsically restrict the solution space to exclude implausible poses (*e.g.*, hyperextended elbows or inverted knees). Explicitly inject priors for guidance has been proven highly effective in other domains, such as image generation [44] and 3D reconstruction [40, 41]. This prior-based regularization complements data-driven approaches by injecting domain-specific knowledge, significantly improving robustness against noisy inputs, and is particularly valuable in uncalibrated scenarios where semantic constraints are entirely learned from the input data.

Despite the promise of integrating biomechanical prior-based regularization, effectively incorporating them into pose estimation models remains a non-trivial challenge. First, a core difficulty lies in how to represent such priors in a way that is structurally robust in representation. Discrete representations offer a promising solution by clustering large amount of body configurations into a limited set of prototypical poses, which encourages the learning of high-quality dominant pose prototypes. This compression enforces structural regularity by restricting the model to select from a predefined pose vocabulary, thereby reducing overfitting to spurious or implausible patterns. Moreover, such many-to-one mappings, *i.e.*, multiple inputs can be mapped to the same prototype, naturally introduces tolerance to

noisy variations, as structurally similar but perturbed poses are treated equivalently. Although it is data-drive, such quantized space can serve as compact, interpretable, and noise-resilient priors-especially beneficial in ambiguous or weakly constrained estimation scenarios. Second, integrating discrete priors into continuous regression models introduces a representational mismatch. While discrete pose spaces are optimized for compactness and structural regularity, pose regressors demand fine-grained flexibility. Directly constraining outputs to lie in the discrete space may hinder expressiveness.

As illustrated in Figure 1, to address the challenges of constructing and leveraging human pose priors, we propose a unified, compact, and noise-resilient discrete prior-enhanced pipeline for 2D-to-3D human pose estimation. Central to our framework is the *UniCodebook*, which encodes both 2D and 3D poses as compositions of discrete sub-structure tokens within a shared discrete space to server as an informative prior that captures essential pose patterns and promotes resilience to noise. While discrete priors provide strong structural regularization and robustness to noise, their quantized nature inherently mismatches the continuous solution space required for accurate pose regression. Rather than directly regularizing the regressed output, we propose a Discrete-Continuous Spatial Attention (DCSA) module to enable continuous pose features to selectively attend to relevant discrete prototypes, which preserves the continuity of the original regression pipeline while softly integrating structural guidance from the discrete prior. By avoiding hard projection into the discrete space, DCSA mitigates representational mismatch and allows the model to inherit noise robustness from the discrete prior without sacrificing fine-grained precision, thereby effectively reconciling the gap between discrete structure and continuous prediction. We demonstrate the effectiveness by achieving state-of-the-art results on three benchmark datasets covering multiple complex scenarios.

The main contributions of the article are as follows:

- **VQ-VAE as a Robust Prior for Pose Estimation**: We propose the use of VQ-VAE to serve as a robust prior for pose estimation. By unifying 2D and 3D pose representations within a shared codebook, VQ-VAE further enhances its robustness to noisy pose, ensuring more accurate and consistent predictions, especially in challenging scenarios such as occlusion or 2D detection errors.
- **Discrete-Continuous Representation Interaction**: We introduce an attribute-preserving and complementarity-enhancing mechanism for the interaction between discrete (codebook-based) and continuous (transformer-based) representations. This allows the discrete tokens to guide the continuous features, infusing noise-resilient properties into the backbone while preserving the regression capability of the continuous representation in continuous tasks (*i.e.*, pose estimation).
- **State-of-the-Art Performance**: The proposed approach is capable of calibration-free human pose estimation even under significant noise, achieving SOTA performance on three large datasets.

## 2 Related Work

### 2.1 Multiview Calibration-Free 3D HPE

Methods with unknown camera extrinsics typically rely on semantic information rather than geometric information to enable better multiview feature interaction. FLEX [11] exploits the view-invariance characteristics of skeleton representations, such as bone lengths and rotation angles, to reconstruct human poses. MTF-TransFormer [32] and FusionFormer [1] leverage transformers to elegantly and efficiently model multiview, temporal, and spatial information, with the latter notably surpassing the performance of calibrated methods for the first time. PoseIRM [2] treats the uncalibrated task as a domain generalization problem and employs the invariant risk minimization paradigm to enhance performance in unseen camera settings. $A^3$-Net [3] leverages pre-defined alignment proxies, such as meshes and joints, to automatically discover inter-view interaction patterns.

### 2.2 VQ-VAE in Pose-Related Tasks

Recent advancements in pose-related tasks have leveraged VQ-VAE [37] to enhance representation learning and improve task performance. In PCT [10], they introduced a two-step framework for 2D pose estimation, where VQ-VAE is first used to learn a compact, discrete pose representation space, followed by a translator that maps image features to this space. This approach effectively prevents unrealistic poses and provides prior knowledge for occluded pose prediction. In TokenHMR [6], they applied a similar pipeline, focusing on learning pose parameters of the SMPL [23] body model using VQ-VAE. In MEGA [9], they utilized VQ-VAE to discretize 3D human meshes, achieving enhanced

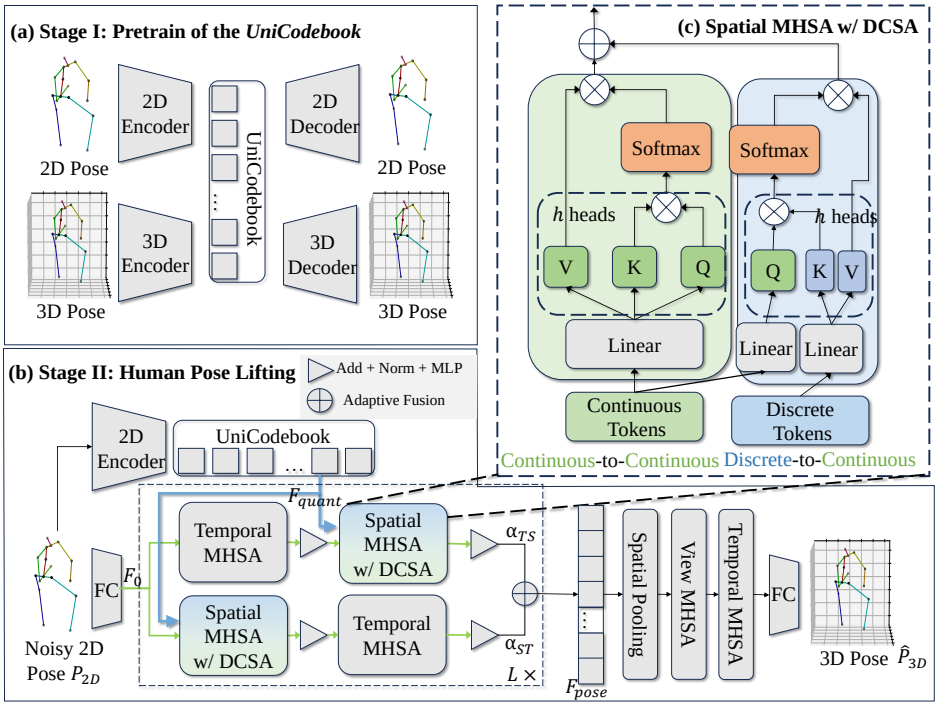


Figure 2: Two stages of the proposed calibration-free multiview 3D human pose lifting pipeline (a, b) and the detailed structure of the Spatial Multi-Head Self-Attention (MHSA) with Discrete-Continuous Spatial Attention (DCSA) (c). In Stage I, we construct the *UniCodebook*, a unified discrete representation space, through a multi-strategy training scheme (2Dto2D, 2Dto3D, 3Dto2D, and 3Dto3D). In this space, both 2D and 3D poses are encoded as sets of discrete tokens in this shared space to bridge the representation gap between 2D and 3D data. In Stage II, a transformer-based continuous model is employed for pose lifting, where codebook tokens generated from the *UniCodebook* are injected into the hybrid spatial attention block. Here, the proposed DCSA mechanism is integrated with conventional MHSA to facilitate effective fusion between the noise-resilient discrete priors and expressive continuous pose features, which enhances the robustness to noisy 2D input.

mesh recovery from single RGB images through both deterministic and stochastic generation modes, significantly improving uncertainty estimation and multi-output predictions. In DVQVAE [48], they proposed a decomposed VQ-VAE approach for generating realistic human grasps, where hand components are separately encoded to enhance interaction with objects. Di$^2$Pose [39] is a discrete diffusion model for occluded 3D pose estimation, which combines pose quantization with a discrete diffusion process to model 3D poses in latent space, leading to superior performance on occlusion handling and physical plausibility. These methods highlight the versatility of VQ-VAE, demonstrating its potential in improving robustness and generalization across a wide range of pose related problems. However, these methods either rely on hard mappings between continuous features and discrete representations [9], or constrain the entire prediction or generation process to discrete latent spaces [48, 39, 10, 6], where quantization can obscure fine-grained details. In contrast, continuous representations excel at capturing subtle spatial variations, which are critical for precise and smooth pose reconstruction. To fully exploit the complementary strengths of discrete structure and continuous precision, our method softly integrates discrete structural priors into the continuous regression stream, enabling noise-robust yet precise pose prediction.

## 3 Methodology

Figure 2 depicts the overall architecture of the proposed methods. In Section 3.1, we describe how to learn an *UniCodebook* which represents 2D and 3D data simultaneously. Section 3.2 explains how the pretrained the *UniCodebook* is used in the human pose lifting task.

### 3.1 Stage I: Pretrain of the *UniCodebook*

In this section, we first introduce the key components used in Stage I. We begin with the Pose Encoder, which processes the original pose into several tokens, followed by the proposed *UniCodebook*, which retrieves features from this unified compact feature space. Finally, we introduce the Pose Decoder and the loss function used to train the entire pipeline.

Existing codebook-based pose representation methods suffer from **modality isolation**, *i.e.*, they encode either 2D poses [10] or 3D poses [6] in separate models. This isolation creates two critical issues: (1) underutilization of heterogeneous pose data that inherently contains complementary 2D-3D correlations, and (2) failure to align the representation gap between 2D observation and 3D geometry. Since the lifting task involves both 2D and 3D features, this gap forces the backbone model to allocate additional capacity to transform unimodal features into a more suitable representation for lifting, rather than focusing entirely on learning robust 2D-to-3D mappings.

To more effectively utilize 2D and 3D data and bridge the gap between them, we propose a unified cross-modal representation learning framework that establishes shared latent embeddings between 2D and 3D human poses. Specifically, while maintaining dedicated modality-specific encoders and decoders for 2D/3D data processing, we use a shared codebook, *i.e.*, the *UniCodebook*, that jointly represents both 2D and 3D pose. We optimize the whole framework through four complementary objectives: 2Dto2D self-reconstruction, 3Dto3D self-reconstruction, cross-modal 2Dto3D projection, and inverse 3Dto2D back-projection. The first two auto-encoding tasks preserve modality-specific reconstruction fidelity and the latter two projection objectives enforce a smaller representation gap through bidirectional translation between modalities.

**Pose Encoder.** Given a human pose $\mathbf{P} \in \mathbb{R}^{J\times D}$ where $D = 2$ for 2D pose $\mathbf{P}_{2d}$ and $D = 3$ for 3D pose $\mathbf{P}_{3d}$, we adopt the **compositional encoding strategy** to capture human priors and structural relationships following [10]. Unlike joint-wise encoding that processes each joint independently, we decompose the pose into $N$ overlapping sub-structures $\mathbf{f}_i$ (e.g., limbs, torso combinations), where each sub-structure corresponds to a group of biomechanically correlated joints. It is noted that the compositional representation contains lots of redundancy due to the fact that different tokens may share overlapping joints. The redundancy, in turn, enhances the robustness of the representation against occlusions of individual parts [10].

We utilize two separate encoders, both identical in structure, to encode 2D and 3D poses, respectively. Following the approach in [10], we adopt the MLP-Mixer architecture [35] as our encoder $f_e$, which is particularly adept at capturing both local joint interactions and global pose semantics. Given an input pose $\mathbf{P}$, we first embed it into $\mathbf{P}_{\text{emb}} \in \mathbb{R}^{J\times D_{\text{feat}}}$ using a linear layer $f_{\text{emb}}$. Subsequently, the embedded features are passed through four MLP-Mixer blocks to effectively model the relationships between different joints. Finally, we employ another linear layer $f_{\text{trans}}$ to map the features to $N$ tokens. The overall process is described by $\mathbf{F}_{\text{en}} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N] = f_{\text{trans}}(f_e(f_{\text{emb}}(\mathbf{P}))) \in \mathbb{R}^{N\times d}$, where each $\mathbf{f}_i \in \mathbb{R}^{D_{feat}}$. The number of tokens is set to $N = 63$, which is larger than the 51 input dimensions (17 joints × 3) of a Human3.6M pose, to introduce representational redundancy.

**Quantization Process With the *UniCodebook*.** We define a shared *UniCodebook* space $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_U]^\top \in \mathbb{R}^{U\times D_{\text{feat}}}$ that jointly represents both 2D and 3D poses. Specifically, each codebook entry $\mathbf{c}_u$ encodes pose patterns applicable to both modalities. This shared representation of discrete tokens helps to minimize the disparity between heterogeneous pose representations.

For each sub-structure token $\mathbf{f}_i$, we perform quantization via nearest neighbor lookup in the codebook:

$$q(\mathbf{f}_i = u \mid \mathbf{P}) = \begin{cases} 1 & \text{if } u = \arg\min_j \|\mathbf{f}_i - \mathbf{c}_j\|_2 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

We abuse $q(\mathbf{f}_i)$ to denote the index of the closest codebook entry for token $\mathbf{f}_i$. The quantized pose representation is given by $\mathbf{F}_{\text{quant}} = [\mathbf{c}_{q(\mathbf{f}_1)}, \mathbf{c}_{q(\mathbf{f}_2)}, \ldots, \mathbf{c}_{q(\mathbf{f}_N)}] \in \mathbb{R}^{N\times D_{\text{feat}}}$.

**Pose Decoder.** The pose decoder is designed to recover pose $\hat{\mathbf{P}}$ from quantized feature $\mathbf{F}_{\text{quant}}$. It mirrors the architecture of the pose encoder but in a reverse manner.

**Loss.** For each strategy, we optimize the entire framework using the following loss function:

$$\mathcal{L}_{\text{pretrain}} = \text{smooth}_{\mathcal{L}_1}(\hat{\mathbf{P}}, \mathbf{P}) + \beta \sum_{i=1}^{M} \|\mathbf{f}_i - \text{sg}[\mathbf{c}_{q(\mathbf{f}_i)}]\|_2^2 \tag{2}$$

where sg denotes stopping gradient, and $\beta$ is a hyperparameter which we set to 10.

### 3.2 Stage II: Human Pose Lifting

Our framework consists of three main components at Stage II: (1) 2D pose discretization and initial continuous feature embedding, (2) dual-stream spatial-temporal feature interaction, and (3) cross-view and temporal feature refinement, and 3D pose regression.

Given a 2D input skeleton sequence $\mathbf{P}_{\text{2D}} \in \mathbb{R}^{V\times T\times J\times D_{\text{2D}}}$, we first feed it into the 2D Pose Encoder of the VQ-VAE trained in Stage I and quantize it into discrete codebook features $\mathbf{F}_{\text{quant}} \in \mathbb{R}^{V\times T\times N\times D_{\text{feat}}}$. Subsequently, within the transformer backbone, we initially project the 2D pose $\mathbf{P}_{\text{2D}}$ into a higher-dimensional feature space $\mathbf{F}^0 \in \mathbb{R}^{V\times T\times J\times D_{\text{feat}}}$ via a fully-connected (FC) layer. We then employ a dual-stream transformer equipped with spatial and temporal Multi-Head Self-Attention (MHSA) mechanisms to derive the feature $\mathbf{F}_{\text{pose}} \in \mathbb{R}^{V\times T\times J\times D_{\text{feat}}}$. Within the spatial MHSA, we implement Discrete-Continuous Spatial Attention (DCSA) to facilitate interaction between continuous features and the discrete features $\mathbf{F}_{\text{quant}}$ along the spatial dimension.

Next, we apply spatial pooling to obtain $\mathbf{F}_{\text{pool}} = \text{AveragePooling}(\mathbf{F}_{\text{pose}}) \in \mathbb{R}^{V\times T\times D_{\text{feat}}}$. This is followed by the application of View MHSA and Temporal MHSA to yield the final continuous pose features $\mathbf{F}'_{\text{pose}} \in \mathbb{R}^{V\times T\times D_{\text{feat}}}$. Finally, we regress the predicted 3D pose $\hat{\mathbf{P}_{\text{3D}}} \in \mathbb{R}^{V\times T\times J\times D_{\text{3D}}}$ through an FC layer. We only apply L1 loss between ground truth(GT) 3D pose $\mathbf{P}_{\text{3D}}$ and $\hat{\mathbf{P}_{\text{3D}}}$. For simplicity, in the following descriptions, we omit layer notations unless cross-layer feature interaction is involved.

#### 3.2.1 Hybrid Spatial Attention Block

The spatial block contains a Spatial Multihead Self-Attention (MHSA) and a Discrete-Continuous Spatial Attention (DCSA) module. For simplicity, we describe the computation for a single attention head. The multi-head version follows standard practice by computing heads in parallel and projecting the concatenated results.

**Spatial MHSA** models relationships between joints within the same view and frame. Given a spatial feature $\mathbf{F}_S \in \mathbb{R}^{J\times D_{\text{feat}}}$ from $\mathbf{F}_{\text{pose}}$, we first compute its query, key, and value vectors:

$$\mathbf{Q}_S = \mathbf{F}_S\mathbf{W}_S^Q, \quad \mathbf{K}_S = \mathbf{F}_S\mathbf{W}_S^K, \quad \mathbf{V}_S = \mathbf{F}_S\mathbf{W}_S^V, \tag{3}$$

where $\mathbf{W}_S^Q, \mathbf{W}_S^K, \mathbf{W}_S^V \in \mathbb{R}^{D_{\text{feat}}\times D_K}$ are learnable projection matrices. The self-attention output is then computed as:

$$\text{S-MHSA}(\mathbf{F}_S) = \text{softmax}\left(\frac{\mathbf{Q}_S\mathbf{K}_S^\top}{\sqrt{D_K}}\right)\mathbf{V}_S, \tag{4}$$

where $D_K$ is the dimension of the key vectors. The temporal MHSA follows the same computation but applies to temporal tokens.

**Discrete-Continuous Spatial Attention (DCSA)** augments the continuous joint interactions with semantic guidance from discrete codebook tokens. We implement this module using a cross-attention mechanism. As detailed in our ablation study in Table 4b and Table 7 in the Supplementary Materials, this approach proved to be the most effective and straightforward compared to other alternatives. The mechanism derives queries from the continuous features $\mathbf{F}_S$ and keys/values from the discrete codebook tokens $\mathbf{F}_{\text{quant}} \in \mathbb{R}^{N\times D_{\text{feat}}}$:

$$\mathbf{Q}_D = \mathbf{F}_S\mathbf{W}_D^Q, \quad \mathbf{K}_D = \mathbf{F}_{\text{quant}}\mathbf{W}_D^K, \quad \mathbf{V}_D = \mathbf{F}_{\text{quant}}\mathbf{W}_D^V. \tag{5}$$

The discrete-to-continuous attention is then computed as:

$$\text{DCSA}(\mathbf{F}_S, \mathbf{F}_{\text{quant}}) = \text{softmax}\left(\frac{\mathbf{Q}_D\mathbf{K}_D^\top}{\sqrt{D_K}}\right)\mathbf{V}_D. \tag{6}$$

The final spatial representation combines features from both modules through addition:

$$\mathbf{F}_S^{\text{out}} = \underbrace{\text{S-MHSA}(\mathbf{F}_S)}_{\text{Continuous-to-Continuous}} + \underbrace{\text{DCSA}(\mathbf{F}_S, \mathbf{F}_{\text{quant}})}_{\text{Discrete-to-Continuous}}. \tag{7}$$

The output features undergo residual connections and layer normalization (LayerNorm), followed by an multilayer perceptron (MLP) block with another residual connection and LayerNorm. We denote the entire spatial block by $\mathcal{S}$.

#### 3.2.2 Dual-Stream Spatial-Temporal Transformer

Following MotionBert [49], we construct a dual-stream spatial-temporal transformer by alternately stacking temporal and hybird spatial blocks in different order, thereby creating two parallel branches.

To enable the model to dynamically adjust the importance of different branches, we compute an adaptive weight for each branch. Adaptive fusion weights $\alpha_{ST}^i, \alpha_{TS}^i \in \mathbb{R}^{L\times V\times T\times J}$ are predicted by linear projection, as $\alpha_{ST}^i, \alpha_{TS}^i = \text{softmax}(\mathcal{W}([\mathcal{T}_1^i(\mathcal{S}_1^i(\mathbf{F}_{\text{pose}}^{i-1})), \mathcal{S}_2^i(\mathcal{T}_2^i(\mathbf{F}_{\text{pose}}^{i-1}))]))$, where $\mathcal{W}$ is a learnable linear transformation, $[\cdot,\cdot]$ denotes concatenation, $i \in \{1,\dots,L\}$ is the current layer number, and $L$ is the number of layers of the dual-stream module.

The features from the two branches are fused additively via predicted weights as $\mathbf{F}_{\text{pose}}^i = \alpha_{ST}^i \circ \mathcal{T}_1^i(\mathcal{S}_1^i(\mathbf{F}_{\text{pose}}^{i-1})) + \alpha_{TS}^i \circ \mathcal{S}_2^i(\mathcal{T}_2^i(\mathbf{F}_{\text{pose}}^{i-1}))$, where $\circ$ denotes element-wise multiplication.

## 4 Experiments

### 4.1 Datasets and Metrics

**Human3.6M** [14] is a widely used 3D human pose dataset with 3.6 million frames captured by 4 synchronized cameras, covering 15 actions performed by 11 subjects. And the poses are represented by 17 human joints. Following common protocol, we use subjects S1, S5, S6, S7, and S8 for training, and S9 and S11 for testing. We evaluate under two settings: using ground-truth (GT) 2D poses and CPN-detected 2D poses. The CPN detector introduces significant 2D pose noise: the average 2D MPJPE between CPN-detected and ground-truth keypoints is 3.9 pixels on the training set and 6.5 pixels on the test set, with maximum errors reaching 55 and 124 pixels, respectively.

**MPI-INF-3DHP** [27] is a challenging multiview dataset for 3D human pose estimation, consisting of 1.3 million frames performed by 8 subjects in diverse indoor and outdoor scenarios, which is widely used to evaluate the generalization performance in complex real-world scenarios. Following previous studies [20, 16], we train on subjects S1-S7 and test on subject S8.

**FreeMan** [38] is a large-scale multiview 3D human pose dataset collected under real-world conditions, containing 11 million frames from 8,000 sequences captured by 8 synchronized smartphones. It covers 40 subjects across 10 diverse indoor and outdoor scenarios with varying lighting. To address the limited distribution gap between the training and testing sets in the original protocol, we re-split the dataset that introduces clearer separation in both scene and camera configurations: scenes 1–8 and cameras (1,3,5,7) are used for training, while scenes 10–12 and cameras (1,3,5,7) (2,4,6,8) are reserved for testing. The resulting training set consists of 566,484 image pairs, and the test set includes 646,608 pairs. This revised split enables a more realistic and challenging evaluation of model generalization.

**AMASS** [26] is a large-scale motion capture dataset with over 40 hours of motion data from 300+ subjects and 11,000+ motions. Since it does not provide direct 2D-3D annotations, we render SMPL meshes and apply official Human3.6M and COCO regressors to extract 3D joints, which are then projected onto 2D planes from 8 virtual views to form 2D-3D pairs. We use the official validation and test splits for training and evaluating the VQ-VAE in Stage I only.

**Evaluation Metrics.** We report Mean Per Joint Position Error (MPJPE), measured in millimeters, Percentage of Correct Keypoints (PCK), and Area Under Curve (AUC) with a 150mm threshold.

### 4.2 Implementation Details

All individual experiments were completed within 1 day using one server equipped with an Nvidia RTX 3090 GPU, 64GB RAM, and a 28-core CPU. We employed the AdamW [24] optimizer with a learning rate of 2e-4 for both the Stage I and Stage II. We used a cosine annealing learning scheduler with 100 warmup steps. The model was trained for 50 and 100 epochs with a batch size of 256 at Stage I and Stage II, respectively. No data augmentation techniques were applied during training and test. For VQ-VAE, we use $N = 63$ tokens to represent a pose, and the codebook is in shape $U \times D_{\text{feat}} = 2048 \times 256$. The original VQ-VAE suffers from codebook collapse, where a significant portion of codebook entries remain underutilized. Following [6], we employ exponential moving average (EMA) and code reset tricks to enhance codebook activation rates.

Table 1: Results on Human3.6M are reported using MPJPE as the evaluation metric. CPN, HRNet and ResNet152 are different 2D pose detectors. GT means using ground truth 2D pose. * means this is an image-to-3d method. † indicates our reimplementation. T represents the number of frames.

| Method | Venue | Input Setting | Dir. | Disc. | Eat. | Greet | Phone | Photo | Pose | Purch. | Sit. | SitD. | Smoke | Wait | WalkD. | Walk | WalkT. | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Monocular Methods | | | | | | | | | | | | | | | | | | |
| MHFormer [19] | CVPR'2022 | (CPN, T = 351) | 39.2 | 43.1 | 40.1 | 40.9 | 44.9 | 51.2 | 40.6 | 41.3 | 53.5 | 60.3 | 43.7 | 41.1 | 43.8 | 29.8 | 30.6 | 43.0 |
| MixSTE [43] | CVPR'2022 | (CPN, T = 243) | 37.6 | 40.9 | 37.3 | 39.7 | 42.3 | 49.9 | 40.1 | 39.8 | 51.7 | 55.0 | 42.1 | 39.8 | 41.0 | 27.9 | 27.9 | 40.9 |
| KTPFormer [29] | CVPR'2024 | (CPN, T = 243) | 30.1 | 32.1 | 29.1 | 30.6 | 35.4 | 39.3 | 32.8 | 30.9 | 43.1 | 45.5 | 34.7 | 33.2 | 32.7 | 22.1 | 23.0 | 33.0 |
| Multi-View Methods With Camera Parameter | | | | | | | | | | | | | | | | | | |
| Epipolar Transformer [13] | CVPR'2020 | (*, T = 1) | 25.7 | 27.7 | 23.7 | 24.8 | 26.9 | 31.4 | 24.9 | 26.5 | 28.8 | 31.7 | 28.2 | 26.4 | 23.6 | 28.3 | 23.5 | 26.9 |
| Crossview Fusion [30] | ICCV'2019 | (*, T = 1) | 24.0 | 26.7 | 23.2 | 24.3 | 24.8 | 22.8 | 24.1 | 28.6 | 32.1 | 26.9 | 31.0 | 25.6 | 25.0 | 28.0 | 24.4 | 26.2 |
| Geometry-Biased Transformer [28] | FG'2024 | (HRNet, T=27) | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 26.0 |
| MTF-Transformer+ [32] | TPAMI'2022 | (CPN, T = 27) | 23.4 | 25.2 | 23.1 | 24.4 | 27.4 | 28.5 | 22.8 | 25.2 | 28.7 | 36.2 | 25.9 | 23.6 | 26.6 | 22.6 | 22.7 | 25.8 |
| LearnTriangulation [15] | ICCV'2019 | (*, T = 1) | 19.9 | 20.0 | 18.9 | 18.5 | 20.5 | 19.4 | 18.4 | 22.1 | 22.5 | 28.7 | 21.2 | 20.8 | 19.7 | 22.1 | 20.2 | 20.8 |
| AdaFuse [47] | IJCV'2021 | (*, T = 1) | 17.8 | 19.5 | 17.6 | 20.7 | 19.3 | 16.8 | 18.9 | 20.2 | 25.7 | 20.1 | 19.2 | 20.5 | 17.2 | 20.5 | 17.3 | 19.5 |
| MvP [42] | NeurIPS'2021 | (*, T = 1) | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 18.6 |
| **Multi-View Methods Without Camera Parameter** | | | | | | | | | | | | | | | | | | |
| ProTriangulation [17] | ICCV'2023 | (*, T = 1) | 24.0 | 25.4 | 26.6 | 30.4 | 32.1 | 20.1 | 20.5 | 36.5 | 40.1 | 29.5 | 27.4 | 27.6 | 20.8 | 24.1 | 22.0 | 27.8 |
| FLEX [11] | ECCV'2022 | (ResNet152, T = 27) | 23.1 | 28.8 | 26.8 | 28.1 | 31.6 | 37.1 | 25.7 | 31.4 | 36.5 | 39.6 | 35.0 | 29.5 | 35.6 | 26.8 | 26.4 | 30.9 |
| FLEX [11] | ECCV'2022 | (CPN, T = 27) | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 31.7 |
| SGraFormer [45] | AAAI'2024 | (CPN, T = 27) | 26.5 | 28.3 | **23.0** | 25.9 | **27.2** | 31.0 | 25.4 | 27.2 | **28.6** | 33.8 | 28.6 | 25.6 | 30.1 | 27.1 | 26.5 | 27.6 |
| FusionFormer [1] † | AAAI'2024 | (CPN, T = 27) | 23.7 | 26.3 | 24.8 | 25.3 | 27.7 | 27.1 | 23.9 | 26.6 | 32.3 | 32.7 | 25.9 | 25.2 | 27.9 | 23.5 | 23.8 | 26.6 |
| Ours | | (CPN, T = 27) | **23.0** | **25.2** | 24.8 | **24.4** | 27.5 | **26.4** | **23.4** | **25.6** | 31.4 | **32.3** | **25.7** | **24.0** | **27.3** | **22.9** | **22.9** | **26.0** |
| MTF-Transfomrer [32] | TPAMI'2022 | (GT, T = 27) | 15.5 | 17.1 | 13.7 | 15.5 | 14.0 | 16.2 | 15.8 | 16.5 | 15.8 | 16.1 | 14.5 | 14.5 | 16.9 | 14.3 | 13.7 | 15.3 |
| SGraFormer [45] | AAAI'2024 | (GT, T = 27) | 11.7 | 13.0 | 10.1 | 12.1 | 10.7 | 13.0 | 12.1 | 10.7 | 10.8 | 11.9 | 11.0 | 11.6 | 12.8 | 11.1 | 12.0 | 11.7 |
| SVTFormer [46] | AAAI'2025 | (GT, T = 27) | 11.6 | 12.3 | 11.4 | 12.2 | 10.6 | 12.1 | 12.5 | 11.7 | 10.3 | 10.7 | 10.7 | 12.1 | 10.8 | 10.9 | 11.4 | 11.4 |
| FusionFormer [1] | AAAI'2024 | (GT, T = 27) | 7.84 | 8.04 | **7.39** | 8.33 | **7.13** | 9.02 | 8.00 | 8.19 | **7.57** | - | **7.37** | **7.83** | - | **7.26** | - | 7.90 |
| Ours | | (GT, T = 27) | **7.77** | **7.82** | 7.70 | **7.71** | 7.80 | **7.76** | **7.87** | **7.60** | 7.81 | **7.66** | 7.76 | 7.92 | **7.63** | 7.66 | **7.74** | **7.74** |

Table 2: Comparison results on the MPI-INF-3DHP.

| Method | Extra Data for Finetune | MPJPE↓ | PCK↑ | AUC↑ |
|---|---|---|---|---|
| ShapeAware [20] | syn data | 62 | 95 | 63 |
| PaFF [16] | × | 48.4 | 98.6 | 67.3 |
| SGraFormer [45] | × | 16.9 | 98.7 | 90.2 |
| MTF-Transformer [32] | h36m | 14.6 | - | - |
| SGraFormer [45] | h36m | 10.6 | 99.9 | 91.7 |
| FusionFormer [1] | h36m | 5.4 | - | - |
| Ours | × | **3.37** | **99.9** | **95.94** |

Table 3: Comparison results on the FreeMan dataset. † indicates our reimplementation.

| | seen camera MPJPE↓ | unseen camera MPJPE↓ | mean MPJPE↓ |
|---|---|---|---|
| SGraFormer [45] † | 22.20 | 28.69 | 25.44 |
| MTF [32] † | 10.85 | 13.51 | 12.18 |
| Fusionformer [1] † | 9.89 | 12.72 | 11.30 |
| Ours | **9.59** | **12.71** | **11.15** |

### 4.3 Comparisons With SOTAs

**Human 3.6M.** Table 1 compares our method with several state-of-the-art (SOTA) monocular and multi-view approaches on the Human 3.6M dataset for 3D human pose estimation (HPE). Our method outperforms all the calibrated methods due to that our method are not limited by predefined geometric relationships between views but instead utilizes data-driven adaptive semantic constraints, allowing for more effective multi-view feature fusion. When compared to calibration-free 3D HPE methods using CPN [4] or GT 2D poses, our approach achieves SOTA performance, with MPJPE values of **26.0mm** and **7.74mm**, respectively, demonstrating its effectiveness. Notably, our method does not rely on any additional tricks but simply adopts L1 Loss.

**MPI-INF-3DHP.** Table 2 presents a detailed comparison with other SOTA calibration-free methods on the MPI-INF-3DHP dataset. We use GT 2D poses as input and keep the same settings as [20]. As shown in the table, our method consistently outperforms all other methods across various metrics. Specifically, we achieve a significant reduction in the MPJPE, outperforming the second-best method by at least **2.03mm (37.5%)**. Additionally, our method shows a notable increase in the AUC, improving it by at least **4.24%**. These indicate that our model not only achieves higher accuracy but also maintains robust performance across different thresholds. It also demonstrates that the *UniCodebook* not only enhances robustness but also improves accuracy without noise.

**FreeMan.** Table 3 reports the evaluation results on the FreeMan dataset under seen and unseen camera settings using HRNet-w48 [33] 2D poses. Our method achieves the best overall performance with a mean MPJPE of 11.15mm, outperforming recent approaches such as Fusionformer [1] (11.30mm) and MTF [32] (12.18mm). In particular, our method attains the lowest error under unseen camera settings (9.59mm), demonstrating strong generalization to novel viewpoints. These results confirm the robustness and adaptability of our approach in real-world scenarios with cross-camera configurations.

### 4.4 Ablation Study

**Impact of Codebook Training Strategy.** Table 4a shows how different codebook training strategies in Stage I affect the lifting performance at Stage II. The best performance is achieved when all strategies are used. Without the codebook, performance drops significantly, highlighting its importance as a robust prior for the continuous model. The MPJPE are around 26.8, 26.7, 26.4 and 26.3 when

Table 4: Ablation studies on key components of the *UniCodebook*, including training strategies, feature interaction methods, and data sources. MPJPE is calculated at Stage II. "CPN" and "GT" denote 2D poses from CPN [4] and ground truth, respectively.

(a) Ablation study on different training strategy combinations for the *UniCodebook* where T=1.

| Number of Strategy | Strategy | | | | MPJPE↓ |
|---|---|---|---|---|---|
| | 2dto2d | 2dto3d | 3dto2d | 3dto3d | |
| 0 | | | | | 27.01 |
| 1 | ✓ | | | | 26.74 |
| | | ✓ | | | 26.86 |
| 2 | ✓ | ✓ | | | 26.70 |
| | ✓ | | ✓ | | 26.78 |
| | ✓ | | | ✓ | 26.74 |
| | | ✓ | ✓ | | 26.73 |
| | | ✓ | | ✓ | 26.75 |
| 3 | ✓ | ✓ | ✓ | | 26.40 |
| | ✓ | ✓ | | ✓ | 26.47 |
| | ✓ | | ✓ | ✓ | 26.53 |
| | | ✓ | ✓ | ✓ | 26.45 |
| 4 | ✓ | ✓ | ✓ | ✓ | **26.34** |

(b) Ablation experiments on different interaction methods between discrete codebook features and continuous features. T=27.

| Method | MPJPE↓ |
|---|---|
| Baseline | 26.7 |
| Conv + Addition | 26.6 |
| Q-Former [18] | 34.1 |
| Concat + MHSA | 26.4 |
| DCSA | **26.0** |

(c) Ablation on training data of VQ-VAE. T=27.

| Data Source | MPJPE↓ |
|---|---|
| H36M CPN | 26.43 |
| H36M GT | 26.35 |
| AMASS | **26.02** |

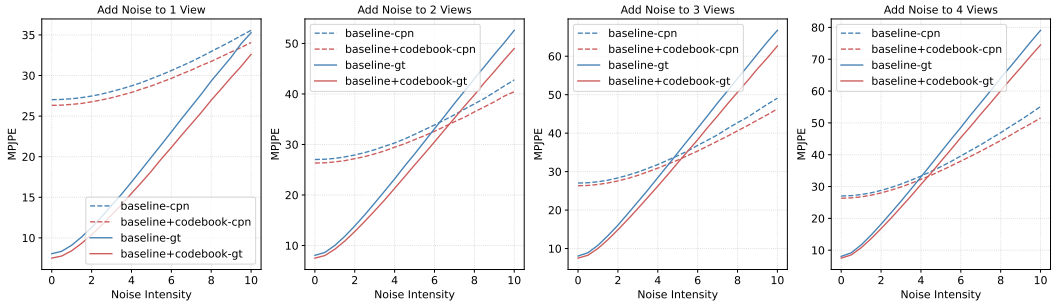


Figure 3: Comparison of MPJPE error across four models (*i.e.*, baseline trained on H36M CPN, baseline with codebook trained on H36M CPN, baseline trained on H36M GT, and baseline with codebook trained on H36M GT) under varying noise intensities without retraining. For each instance (consisting of multi-view 2D poses of the same person at the same timestamp), we randomly select 1 to 4 views and add Gaussian noise with zero mean and a standard deviation of "Noise Intensity" pixels to each 2D joint. For models trained on H36M CPN, we evaluate them using H36M CPN test data with extra noise. Similarly, models trained on H36M GT are evaluated with H36M GT test data with extra noise. **The results show that models with the codebook exhibit robustness across all noise levels, with greater robustness observed at higher noise intensities.**

applying one, two, three and four strategies, respectively. This decreasing trend demonstrates that as more strategies are used, the *UniCodebook* can better capture the relationships between 2D and 3D poses and data utilization rate , making it a robuster prior for lifting tasks.

**Impact of the Interaction Method of Discrete Codebook Features and Continuous Features.** In Table 4b, we aim to investigate the impact of different interaction methods between discrete and continuous features on model performance. "Baseline" means no discrete features are used. "Conv + Addition" involves aligning the token count of the codebook features using a convolutional layer, followed by adding them to the continuous features. "Q-Former [18]" employs learned queries with the same shape as continuous features and maps discrete feature information onto the queries using a transformer layer. The queries and continuous features then interact via cross-attention. "Concat + MHSA" first concatenates discrete and continuous features along the token dimension and then applies self-attention. All methods are applied in multiple spatial blocks. The results demonstrate that our DCSA method achieves the best performance, improving MPJPE by 0.7mm.

**Impact of Noise.** To evaluate the robustness of our model to noisy inputs, we present in Figure 3 the MPJPE performance of four models under varying noise intensities. The results demonstrate that

Table 5: Ablation on the number of pose tokens and codebook size in UniCodebook where the frame number is 1. The active rate represents the mean activation rate in Stage I, while MPJPE is evaluated in Stage II on Human3.6M [14].

| pose tokens | codebook size | Active Rate(%)↑ | MPJPE↓ |
|---|---|---|---|
| 21 | 2048 | 61.8 | 26.82 |
| 63 | | 83.5 | **26.34** |
| 105 | | **84.3** | 26.35 |
| 63 | 1024 | **84.5** | 26.70 |
| | 2048 | 83.5 | **26.34** |
| | 4096 | 75.0 | 26.47 |

our method—whether trained on H36M CPN or H36M GT—consistently outperforms the baseline in the presence of noise. Moreover, the performance gap widens as the noise intensity increases, indicating that the codebook-based model exhibits stronger resilience to noisy 2D observations than the continuous transformer-based baseline. This robustness stems from the discrete nature of the codebook, which provides a form of structural regularization by anchoring predictions to a set of stable pose prototypes. Such discrete priors may help correct noisy or ambiguous joint inputs by guiding the model toward plausible configurations. These findings highlight the effectiveness of our approach in real-world scenarios, where noisy 2D detections are common, and underscore the practical advantages of incorporating discrete structural knowledge into continuous pose estimation models.

**Impact of Codebook Training Data Quality.** To evaluate the impact of data quality on model performance, we conducted experiments in Table 4c using three data sources: H36M CPN, H36M GT, and AMASS. The results show a consistent performance increase among them. As data quality improves, the model performance also increases, indicating that higher-quality data leads to better performance. It is worth noting that even when using only H36M CPN data, *i.e.*, without any external data, our model still achieves near-SOTA performance on the Human3.6M benchmark.

**Impact of Codebook Configurations.** We analyze two key hyperparameters: the **codebook size**, which defines the capacity of the discrete latent space, and the **number of pose tokens**, which reflects the granularity of the pose representation. (1) First, we fixed the codebook size to 2048 and varied the number of pose tokens. Increasing the tokens from 21 to 63 significantly boosted the active rate (from 61.8% to 83.5%) and improved MPJPE (from 26.82 to 26.34). A further increase to 105 tokens yielded no improvement, suggesting that 63 tokens offer a favorable balance between expressiveness and efficiency. (2) Next, we fixed the number of pose tokens at 63 and varied the codebook size. A codebook of 2048 achieved the best MPJPE (26.34). A smaller size (1024) suffered from higher error (MPJPE: 26.70) due to limited capacity, while a larger one (4096) slightly degraded performance (MPJPE: 26.47). Overall, our selected configuration of **63 pose tokens** and a **codebook size of 2048** achieves the best trade-off between token utilization and pose estimation accuracy.

## 5 Conclusion

In this paper, we propose a novel approach for uncalibrated multi-view 3D human pose estimation by integrating a noise-resilient prior, *i.e.*, the *UniCodebook*, into a transformer-based framework. By employing a multi-strategy training scheme, the *UniCodebook* effectively bridges the gap between 2D and 3D pose representations. Furthermore, we introduce Discrete-Continuous Spatial Attention (DCSA) to facilitate interaction between discrete codebook tokens and continuous pose features, ensuring the benefits of noise-resilient priors without disrupting the continuity of the backbone model. Extensive experiments on three benchmark datasets demonstrate that our approach outperforms both calibration-required and calibration-free methods, achieving state-of-the-art performance. These results highlight the effectiveness of integrating discrete priors into multi-view pose estimation, demonstrating their potential to enhance the robustness of continuous models.

## Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grants (62406039, 62321001, 62471055, U23B2001, 62171057, 62201072, 62071067, 62406039, 62101064), the High-Quality Development Project of the MIIT (2440STCZB2584), the Ministry of Education and China Mobile Joint Fund (MCM20200202, MCM20180101), the Project funded by China Postdoctoral Science Foundation (2023TQ0039, 2024M750257, GZC20230320), the Fundamental Research Funds for the Central Universities (2024PTB-004), and the 2025 Education and Teaching Reform Project Funding at Beijing University of Posts and Telecommunications (2025YZ005).

## References

[1] Y. Cai, W. Zhang, Y. Wu, and C. Jin. Fusionformer: A concise unified feature fusion transformer for 3d pose estimation. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 900–908, 2024.

[2] Y. Cai, W. Zhang, Y. Wu, and C. Jin. Poseirm: Enhance 3d human pose estimation on unseen camera settings via invariant risk minimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2124–2133, 2024.

[3] G. Chen, X. Jian, Y. Chen, P. Ren, J. Wang, H. Sun, Q. Qi, J. Wang, and J. Liao. $A^3$-net: Calibration-free multi-view 3d hand reconstruction for enhanced musical instrument learning. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2025.

[4] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. Cascaded pyramid network for multi-person pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7103–7112, 2018.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[6] S. K. Dwivedi, Y. Sun, P. Patel, Y. Feng, and M. J. Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1323–1333, 2024.

[7] K. Ehlers and K. Brama. A human-robot interaction interface for mobile and stationary robots based on real-time 3d human body and hand-finger pose estimation. In *International Conference on Emerging Technologies and Factory Automation (ETFA)*, pages 1–6. IEEE, 2016.

[8] H.-S. Fang, Y. Xu, W. Wang, X. Liu, and S.-C. Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 32, 2018.

[9] G. Fiche, S. Leglaive, X. Alameda-Pineda, and F. Moreno-Noguer. Mega: Masked generative autoencoder for human mesh recovery. *arXiv preprint arXiv:2405.18839*, 2024.

[10] Z. Geng, C. Wang, Y. Wei, Z. Liu, H. Li, and H. Hu. Human pose as compositional tokens. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 660–671, 2023.

[11] B. Gordon, S. Raab, G. Azov, R. Giryes, and D. Cohen-Or. Flex: Extrinsic parameters-free multi-view 3d human motion reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 176–196. Springer, 2022.

[12] R. A. Guler and I. Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10884–10894, 2019.

[13] Y. He, R. Yan, K. Fragkiadaki, and S.-I. Yu. Epipolar transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7779–7788, 2020.

[14] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 36(7):1325–1339, 2013.

[15] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov. Learnable triangulation of human pose. In *IEEE International Conference on Computer Vision (ICCV)*, pages 7718–7727, 2019.

[16] K. Jia, H. Zhang, L. An, and Y. Liu. Delving deep into pixel alignment feature for accurate multi-view human mesh recovery. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, pages 989–997, 2023.

[17] B. Jiang, L. Hu, and S. Xia. Probabilistic triangulation for uncalibrated multi-view 3d human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 14850–14860, 2023.

[18] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ACM International Conference on Machine Learning (ICML)*, pages 19730–19742. PMLR, 2023.

[19] W. Li, H. Liu, H. Tang, P. Wang, and L. Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13147–13156, 2022.

[20] J. Liang and M. C. Lin. Shape-aware human pose and shape reconstruction using multi-view images. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4352–4362, 2019.

[21] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang. Skeleton-based action recognition using spatio-temporal lstm network with trust gates. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 40(12):3007–3021, 2017.

[22] X. Liu, P. Ren, Y. Chen, C. Liu, J. Wang, H. Sun, Q. Qi, and J. Wang. Sa-fusion: multimodal fusion approach for web-based human-computer interaction in the wild. In *ACM International World Wide Web Conference (WWW)*, pages 3883–3891, 2023.

[23] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: a skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015.

[24] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[25] D. C. Luvizon, D. Picard, and H. Tabia. Multi-task deep learning for real-time 3d human pose estimation and action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 43(8):2752–2764, 2020.

[26] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. Amass: Archive of motion capture as surface shapes. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5442–5451, 2019.

[27] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *International Conference on 3D Vision (3DV)*, pages 506–516. IEEE, 2017.

[28] O. Moliner, S. Huang, and K. Åström. Geometry-biased transformer for robust multi-view 3d human pose reconstruction. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2024.

[29] J. Peng, Y. Zhou, and P. Mok. Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3d human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1123–1132, 2024.

[30] H. Qiu, C. Wang, J. Wang, N. Wang, and W. Zeng. Cross view fusion for 3d human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4342–4351, 2019.

[31] P. Ren, H. Sun, J. Hao, J. Wang, Q. Qi, and J. Liao. Mining multi-view information: a strong self-supervised framework for depth-based 3d hand pose and mesh estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20555–20565, 2022.

[32] H. Shuai, L. Wu, and Q. Liu. Adaptive multi-view and temporal fusing transformer for 3d human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 45(4):4122–4135, 2022.

[33] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5693–5703, 2019.

[34] T. Tao, X. Yang, J. Xu, W. Wang, S. Zhang, M. Li, and G. Xu. Trajectory planning of upper limb rehabilitation robot based on human pose estimation. In *International Conference on Ubiquitous Robots (UR)*, pages 333–338. IEEE, 2020.

[35] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:24261–24272, 2021.

[36] H. Tu, C. Wang, and W. Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *European Conference on Computer Vision (ECCV)*, pages 197–212. Springer, 2020.

[37] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

[38] J. Wang, F. Yang, B. Li, W. Gou, D. Yan, A. Zeng, Y. Gao, J. Wang, Y. Jing, and R. Zhang. Freeman: Towards benchmarking 3d human pose estimation under real-world conditions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21978–21988, 2024.

[39] W. Wang, J. Xiao, C. Wang, W. Liu, Z. Wang, and L. Chen. Di$^2$Pose:discrete diffusion model for occluded 3d human pose estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[40] Z. Yuan, C. Liu, F. Shen, Z. Li, J. Luo, T. Mao, and Z. Wang. MSP-MVS: Multi-granularity segmentation prior guided multi-view stereo. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 9753–9762, 2025.

[41] Z. Yuan, J. Luo, F. Shen, Z. Li, C. Liu, T. Mao, and Z. Wang. DVP-MVS: Synergize depth-edge and visibility prior for multi-view stereo. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 9743–9752, 2025.

[42] J. Zhang, Y. Cai, S. Yan, J. Feng, et al. Direct multi-view multi-person 3d pose estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:13153–13164, 2021.

[43] J. Zhang, Z. Tu, J. Yang, Y. Chen, and J. Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13232–13242, 2022.

[44] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023.

[45] L. Zhang, K. Zhou, F. Lu, X.-D. Zhou, and Y. Shi. Deep semantic graph transformer for multi-view 3d human pose estimation. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 7205–7214, 2024.

[46] W. Zhang, M. Liu, H. Liu, and W. Li. Svtformer: Spatial-view-temporal transformer for multi-view 3d human pose estimation. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 10148–10156, 2025.

[47] Z. Zhang, C. Wang, W. Qiu, W. Qin, and W. Zeng. Adafuse: Adaptive multiview fusion for accurate human pose estimation in the wild. *International Journal of Computer Vision (IJCV)*, 129:703–718, 2021.

[48] Z. Zhao, M. Qi, and H. Ma. Decomposed vector-quantized variational autoencoder for human grasp generation. In *European Conference on Computer Vision (ECCV)*, pages 447–463. Springer, 2025.

[49] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang. Motionbert: A unified perspective on learning human motion representations. In *IEEE International Conference on Computer Vision (ICCV)*, pages 15085–15099, 2023.

## 6 NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state three contributions: (1) a VQ-VAE-based UniCodebook unifying 2D/3D pose representations to improve robustness, (2) a DCSA mechanism enabling discrete-continuous interaction, and (3) SOTA results on three representative datasets. These claims are directly validated in the methodology and experiments.

   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: [Yes]

   Justification: We discuss limitations in the supplemental material.

   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: The paper does not include theoretical results.

   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: The methodological workflow is thoroughly detailed in Section 3, while Section 4.1 specifies the dataset's train-test splitting way and preprocessing procedures.

Guidelines:

- The answer NA means that the paper does not include experiments.
- 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: The paper uses publicly available datasets and provides detailed, reproducible processing methods in Section 4.1. The code will be released upon acceptance of the paper, and comprehensive model hyperparameters and training details are provided in Section 4.2.

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: The paper provides comprehensive details on data processing and splits in Section 4.1, along with hyperparameters, optimizer selection, and model configurations in Section 4.2, ensuring full reproducibility and clear understanding of the experimental results.

   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: We do not report error bars or other appropriate information about the statistical significance.

   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 report the compute resources in Section 4.2.

   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.
   - 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: The research fully complies with NeurIPS ethical guidelines.

   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: We discuss potential societal impacts in the supplemental material.

    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: We propose a human pose estimation method and it 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: All datasets and models used in this work are properly licensed and accompanied by accurate citations to their original publications.

    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: [NA]

    Justification: This paper does not release new assets.

    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)?

    Answer: [NA]

    Justification: This 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: This 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.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We only use the LLM to improve writing.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.