DreamPose3D: Hallucinative Diffusion with Prompt Learning for 3D Human Pose Estimation

Jerrin Bright[1,2], **Yuhao Chen**[1,2], **John S. Zelek**[1,2],
[1]Vision and Image Processing Lab, [2]University of Waterloo

Abstract

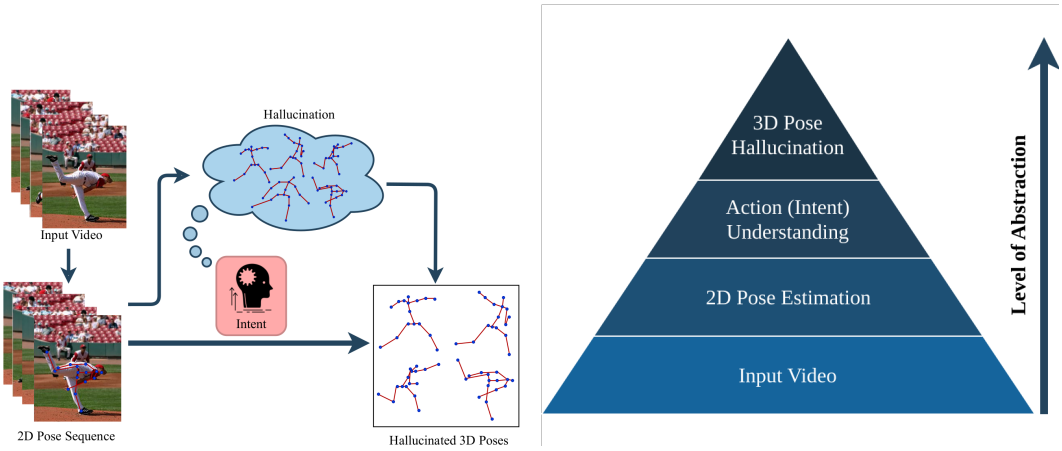
Accurate 3D human pose estimation remains a critical yet unresolved challenge, requiring both temporal coherence across frames and fine-grained modeling of joint relationships. However, most existing methods rely solely on geometric cues and predict each 3D pose independently, which limits their ability to resolve ambiguous motions and generalize to real-world scenarios. Inspired by how humans understand and anticipate motion, we introduce `DreamPose3D`, a diffusion-based framework that combines action-aware reasoning with temporal imagination for 3D pose estimation. `DreamPose3D` dynamically conditions the denoising process using task-relevant action prompts extracted from 2D pose sequences, capturing high-level intent. To model the structural relationships between joints effectively, we introduce a representation encoder that incorporates kinematic joint affinity into the attention mechanism. Finally, a hallucinative pose decoder predicts temporally coherent 3D pose sequences during training, simulating how humans mentally reconstruct motion trajectories to resolve ambiguity in perception. Extensive experiments on benchmarked Human3.6M and MPI-3DHP datasets demonstrate state-of-the-art performance across all metrics. To further validate `DreamPose3D`'s robustness, we tested it on a broadcast baseball dataset, where it demonstrated strong performance despite ambiguous and noisy 2D inputs, effectively handling temporal consistency and intent-driven motion variations.

1 Introduction

Given monocular 2D images or videos, 3D Human Pose Estimation (3D HPE) aims to predict the positions of human body joints in 3D space. It is crucial in applications such as augmented and virtual reality Mehta et al. [2017b], sports analysis Bright et al. [2024a, 2023], Rematas et al. [2018], self-driving Kim et al. [2019], Du et al. [2019], Zheng et al. [2022], and robotics Gui et al. [2018], Erickson et al. [2022]. Traditional monocular 3D HPE Zhang et al. [2022], Pavllo et al. [2019], Zheng et al. [2023], Shan et al. [2022], Zhu et al. [2023] focuses on reconstructing plausible 3D poses from 2D projections Sun et al. [2019], Xu et al. [2022], Chen et al. [2018]. However, despite being conditioned on a 2D pose sequence, these approaches typically predict 3D poses on a per-frame basis, thereby neglecting longer-term motion cues that could improve temporal coherence and robustness.

Recent diffusion-based methods Shan et al. [2023], Gong et al. [2023], Choi et al. [2023], Holmquist and Wandt [2023], Feng et al. [2023], Zhou et al. [2023b] have explored generative modeling for 3D HPE, sampling plausible 3D skeletons conditioned on local 2D pose sequences. Despite their probabilistic formulation, these techniques limit their predictions to a single target frame. Moreover, these models rely solely on geometric information from 2D poses, without considering the higher-level context or intent behind the motion. We refer to this limitation as *intent ambiguity*. This can lead to uncertainty when interpreting motion patterns, especially when different actions exhibit similar

Preprint. Under review.



(a) High-level overview of `DreamPose3D` framework. (b) Illustration of intent-driven motion perception.

Figure 1: **Overview and Effectiveness of `DreamPose3D`.** (a) A conceptual overview of the proposed framework, which mimics human-like reasoning of motion. (b) Qualitative impact of intent and hallucination modules, showing performance with and without each component.

joint movements over short windows, such as waving vs. throwing, or jumping vs. stumbling. This limits their generalizability in scenarios where visual cues may be noisy or corrupted. In such cases, semantic context (i.e., action-level priors) can help guide pose reconstruction more effectively.

Prior work in cognitive science has shown that human perception of motion involves both intent recognition and mental simulation. A recent study on kinematic priming Scaliti et al. [2023] demonstrated that prior knowledge of a person's intent, such as drinking versus pouring, significantly alters how observers interpret subsequent movement kinematics. This suggests that intent strongly shapes motion understanding, especially under ambiguous visual input. The hierarchy illustrated in Figure 1b reflects this cognitive sequence, moving from low-level visual cues to high-level motion simulation. Motivated by these insights, we ask: *Is it possible to build a model that mimics the way humans instinctively reason about motion by identifying intent and imagining plausible motion that aligns with it?*

To mimic this human-like process of reasoning about motion, we reformulate 3D HPE as a **intent-driven motion imagination task**. Instead of predicting a single 3D pose at a time, our method imagines a coherent sequence of 3D poses that span the past, present, and future, given an input window of 2D poses. This allows the model to understand not just the 3D pose at the current frame, but how it evolves explicitly over time. We further condition the generation using action-aware prompts, encoded using a pre-trained Vision-Language Model (VLM). These prompts act as the high-level intent, helping the model generate contextually meaningful poses.

In this work, we propose `DreamPose3D`, a prompt-driven hallucinative diffusion model for 3D human pose estimation (Figure 1a). `DreamPose3D` first employs a transformer network to learn motion representations from 2D pose sequences to predict the action prompt. Then, using a pre-trained VLM (CLIP Radford et al. [2021]), it encodes context embeddings based on the learned action prompt. These embeddings condition a diffusion model that denoises a Gaussian distribution using spatial and temporal multi-head attention. We further modify the attention mechanism by infusing it with the global kinematic joint relationships, ensuring consistent pose structure across frames. The output tokens of the diffusion model are then decoded using a pose hallucinator that predicts a sequence of *hallucinatory* 3D poses, enforcing temporal consistency by regularizing motion dynamics.

Our contributions can be summarized as follows:

- We propose `DreamPose3D`, a prompt-driven hallucinative diffusion model that addresses temporal inconsistencies and intent ambiguity in 3D human pose estimation.
- We introduce a lightweight pose hallucinator that generates temporally coherent 3D pose sequences, improving motion continuity over time.
- Our method achieves state-of-the-art performance on both Human3.6M and MPI-INF-3DHP datasets. Additional experiments on the MLBPitchDB dataset highlight its robustness in scenarios with noisy or corrupted 2D pose inputs.

2 Related Work

2.1 3D Human Pose Estimation

Traditional 3D human pose estimation methods decompose the task by first extracting 2D poses from input images or videos, then reconstructing the 3D pose from these 2D pose sequences Zhang et al. [2022], Pavllo et al. [2019], Zhao et al. [2023], Zhu et al. [2023]. These approaches predict a single, most likely 3D pose for each 2D observation/ sequence. Recently, however, multihypothesis approaches have emerged, generating a set of plausible 3D poses for a given timestamp f from a given 2D pose sequence centered around f Shan et al. [2023], Holmquist and Wandt [2023], Li et al. [2022b]. These approaches aim to better capture pose variability by combining multiple plausible 3D predictions using conditioning or averaging techniques.

By contrast, we predict a set of n sequential 3D poses from the given 2D pose sequence. With our simple yet novel hallucinating approach, we are able to reflect the full dynamic progression of poses.

2.2 Diffusion Models

Diffusion models Ho et al. [2020] have become state-of-the-art in various generative tasks. The diffusion model gradually removes noise from an initialized Gaussian distribution, generating outputs that align with the target distribution. They have demonstrated superior performance for various computer vision tasks including test-to-image synthesis Kwak et al. [2024], Bai et al. [2023], super-resolution Li et al. [2022a], Yue et al. [2024], image inpainting Lugmayr et al. [2022], Corneanu et al. [2024] and segmentation Amit et al. [2021], Wu et al. [2024]. Recent works on 3D HPE Shan et al. [2023], Gong et al. [2023], Holmquist and Wandt [2023], Choi et al. [2023] have focused on using diffusion models, with the aim of naturally handling the indeterminacy and uncertainty in the observation. Diffusion models for 3D HPE offer various advantages: (1) Plausible human poses against noise and occlusions Zhou et al. [2023a]; (2) Does not suffer from phenomena like posterior collapse, vanishing gradients, or training instabilities Holmquist and Wandt [2023]; (3) Captures fine-grained dynamics with fidelity even during inherent ambiguities in representation Zhou et al. [2023a], Feng et al. [2023].

Thus, to address the inherent indeterminacy and uncertainty in 3D HPE, we leverage the strengths of diffusion models, which are well-suited for handling such challenges. Thus, we formulate our 3D HPE task as a reverse diffusion process, enabling `DreamPose3D` to generate accurate 3D pose reconstructions by effectively managing the ambiguous and uncertain nature of pose predictions.

2.3 Prompt Learning

Text prompts have emerged as a valuable tool for guiding various vision tasks, including pose estimation. Several recent works Guo et al. [2023], Hu et al. [2023], Hu and Liu [2024], Balaji et al. [2024] have used text prompts to estimate 2D poses of hands, humans, and animals, respectively. More recently, MDM Tevet et al. [2023] utilized text prompts to generate 3D pose sequences using diffusion models, and FinePOSE Xu et al. [2024] introduced a part-aware learning mechanism that encodes information about action classes and coarse- and fine-grained human parts.

Inspired by these works, we propose the Action Prompt Learning (APL) block, which introduces learnable prompt embeddings to infer high-level motion intent and guide the 3D pose generation process. Unlike previous approaches Xu et al. [2024], Guo et al. [2023] that rely on explicit action class labels, `DreamPose3D` distinguishes itself by *directly learning the action* by encoding the motion through a transformer backbone. This enables the model to operate in scenarios where action labels are unavailable.

3 Method

Given an input 2D skeleton sequence $\mathbf{X} \in \mathbb{R}^{N \times J \times 2}$, defined by N frames with J joints for each skeleton, the goal of our approach is to predict the corresponding 3D pose sequence $\mathbf{Y}_\theta \in \mathbb{R}^{N \times J \times 3}$. To address the challenges of temporal coherence and intent ambiguity in 3D HPE, `DreamPose3D` incorporates three core modules. First, the Action Prompt Learning (APL) block generates action-aware embeddings from the input 2D pose sequence, encoding intent for realistic pose transitions.

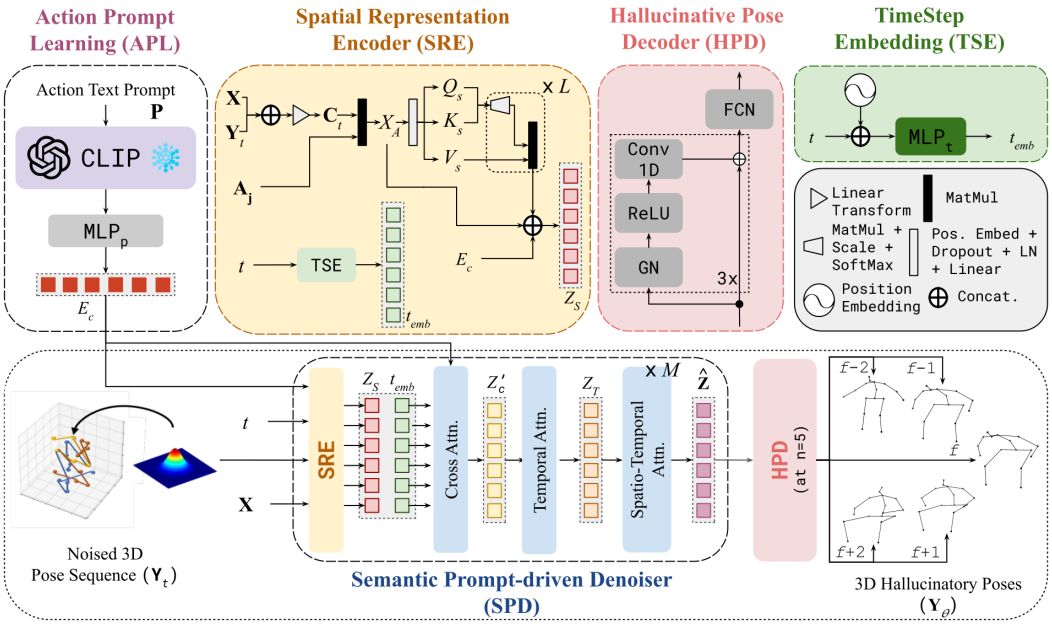


Figure 2: **Overview of** `DreamPose3D` **model architecture.** Given an input 2D pose sequence **X**, `DreamPose3D` predicts temporally coherent 3D poses through the SPD block, which performs diffusion-based denoising. SPD is conditioned on a context embedding, inferred from the 2D sequence via the APL block. The HPD block complements SPD by generating auxiliary 3D hallucinatory poses to reinforce temporal consistency across frames.

Next, a diffusion model utilizes the proposed Semantic Prompt-driven Denoiser (SPD) in the reverse diffusion process, improving denoising performance with spatially and temporally informed embeddings. Finally, a Hallucinative Pose Decoder (HPD) module predicts n 3D hallucinatory poses, ensuring smooth, coherent transitions across frames. Figure 2 shows the network architecture of `DreamPose3D` for 2D-to-3D lifting of the skeleton sequences.

3.1 Action Prompt Learning

Existing works often rely on *manual text prompts* along with input images or pose sequences to guide their models toward a specific task Guo et al. [2023], Xu et al. [2024], Balaji et al. [2024]. To address the limitations of manual prompts, which may not always be available, we propose the APL block that automatically generates task-relevant prompts from the 2D pose sequence (**X**) and encodes context-aware information.

The APL block consists of two main components: an intent classifier and a VLM. Given the input 2D pose sequence, the intent classifier encodes the motion representation using a transformer-based encoder Zhu et al. [2023] denoted as **E**. This motion representation is then passed through a lightweight decoder (ϕ) composed of global average pooling, deconvolution layers, and an MLP with a single hidden layer to generate a context-aware text prompt (**P**), as shown in Equation (1).

$$\mathbf{P} = \phi(\mathbf{E}(\mathbf{X})) \tag{1}$$

The generated prompt, which captures the semantics of the observed 2D motion, is then tokenized using the standard CLIP tokenizer (t_k), and fed as input to a VLM (CLIP Radford et al. [2021]). Specifically, CLIP operates with a fixed token sequence length of 77 tokens, which we structurally partition into two learnable prompt segments: 40 tokens for subject-related context (i.e., template: "*a person*") and 37 for action-related context.

Explicitly regularizing the prompt guides the model towards better convergence and stability while also offering flexibility by allowing the use of custom or predicted prompts. The CLIP output embeddings are then passed through an MLP block (MLP_p) to obtain the context embeddings (E_c), as formulated in Equation (2).

$$E_c = \text{MLP}_p[\text{CLIP}(\text{t}_k(\mathbf{P}))] \tag{2}$$

3.2 Semantic Prompt-driven Denoiser

The SPD block is the key component in the reverse process of `DreamPose3D`, designed to output coherent 3D pose tokens ($\hat{\mathbf{Z}}$) by leveraging spatial and temporal context. As the denoiser network $\mathcal{D}$, the SPD aims to progressively refine noisy 3D pose sequences by conditioning on multiple inputs: the noised 3D pose sequence ($\mathbf{Y}_t$), the input 2D pose sequence $\mathbf{X}$, context embeddings E_c, and time t. The resulting output is a set of denoised pose tokens ($\hat{\mathbf{Z}}$), which effectively incorporate human joint kinematics across spatial and temporal dimensions, formulated as shown in Equation (3).

$$\hat{\mathbf{Z}} = \mathcal{D}(\mathbf{Y}_t, \mathbf{X}, E_c, t) \tag{3}$$

Spatial Representation Encoder (SRE). The SRE encodes spatial dependencies by modeling joint affinities, allowing the SPD to learn local and global relationships between joints in each frame. Initially, the 2D pose $\mathbf{X}$ and the noisy 3D pose $\mathbf{Y}_t$ are concatenated to form the pose tokens ($\mathbf{C}_t$), representing joint configurations across frames.

Inspired by affinity mechanisms Peng et al. [2024], which effectively capture joint dependencies in pose estimation, we construct a joint affinity matrix ($\mathbf{A_j}$) to incorporate local and global joint interactions, which are critical for ensuring joint alignment. The joint affinity matrix $\mathbf{A_j}$ is defined as:

$$\mathbf{A_j} = \frac{(A_L + A_G) + (A_L + A_G)^T}{2} \tag{4}$$

where A_L represents local joint affinities for proximal joints (hand-crafted), and A_G captures global affinities for joints at greater spatial distances (learnable). This symmetric affinity matrix enables the model to understand each joint's structural dependencies with respect to neighboring and distant joints. To integrate the spatial relationships from $\mathbf{A_j}$ with the pose tokens $\mathbf{C}_t$, we apply a series of multi-head attention layers. First, the affinity-modulated pose tokens $X_A = \mathbf{C}_t \cdot \mathbf{A_j}$ are linearly transformed to obtain queries (Q_s), keys (K_s), and values (V_s). These representations are then processed through spatial multi-head attention with L heads. The multi-head attention output is subsequently combined with the affinity-modulated tokens (X_A), along with the context embeddings E_c, which acts as a residual connection to enhance spatial consistency in the final pose tokens (Z_S). The overall output of the SRE, incorporating the spatial multi-head attention, can be expressed as:

$$Z_S = MultiHead(Q_s, K_s, V_s) + X_A + E_c \tag{5}$$

TimeStep Embedding. Following DDPMs, a sinusoidal function encodes the timestep t. This encoded timestep is then processed through an MLP block (MLP_t), which consists of linear layers and GELU activation functions. The output of MLP_t, denoted as t_{emb}, is added to the SRE block and subsequently fed as an additional input to the temporal attention block. This ensures the SPD block can effectively handle 3D noisy poses at various timesteps.

Cross Attention. To incorporate high-level context, we implement a cross-attention mechanism using the output tokens Z_S from the SRE block and the context embeddings E_c. Here, the query, key, and value are defined as $Q_c = W_q Z_S$, $K_c = W_k E_c$, and $V_c = W_v E_c$, respectively, where W_q, W_k, and W_v are the respective parameter matrices. The resulting output tokens from the cross-attention block are denoted as Z_c. Additionally, we add the t_{emb} to the output cross-attention tokens Z_c to optimize based on the timestep context, denoted as $Z_c' = Z_c + t_{emb}$. Next, the concatenated tokens Z_c' are passed through a temporal attention block to model inter-frame relationships between different poses, resulting in the output token representation (Z_T). Finally, a spatiotemporal encoder is applied to extract fine-grained and rich output representation, denoted as $\hat{\mathbf{Z}}$. The spatio-temporal encoder consists of M stacks of alternating spatial and temporal transformers.

3.3 Hallucinative Pose Decoder

Current diffusion-based approaches Xu et al. [2024], Shan et al. [2023] typically produce multiple plausible pose hypotheses for each frame, even when given a sequence of 2D observations. These methods often rely on additional processing to merge or refine the outputs to achieve temporal

consistency. However, this probabilistic formulation lacks an inherent understanding of motion continuity across a sequence, leading to potential inconsistencies in frame-to-frame transitions.

Thus, we introduce the HPD decoder, which leverages the temporal information from the output tokens, $\hat{\mathbf{Z}}$, to generate a coherent set of 3D poses termed as 3D hallucinatory poses. By predicting the hallucinatory poses across a set of consecutive frames, $\{f - \frac{n-1}{2}, \ldots, f, \ldots, f + \frac{n-1}{2}\}$, rather than as isolated frames, the HPD decoder inherently captures temporal dependencies, resulting in smoother and more realistic 3D pose transitions. This approach eliminates the need for further post-processing, providing a streamlined and temporally coherent output. When $n = 5$, the HPD decoder as shown in Figure 2, generates poses over the sequence $\{f - 2, \ldots, f + 2\}$ centered around f.

3.4 Objective Functions

The objective of `DreamPose3D` is to minimize the error between the predicted and groundtruth poses by ensuring joint consistency through the hallucinatory poses, bone length, and action prompt regularization. The primary loss term, $\mathcal{L}_{3D}^{f}$ defined in Equation (6), represents the $L1$ loss for n 3D hallucinatory poses.

$$\mathcal{L}'_{3D} = \sum_{k=-\frac{n-1}{2}}^{\frac{n-1}{2}} \lambda_{3D}^{f+k} \mathcal{L}_{3D}^{f+k} \tag{6}$$

where $\mathcal{L}_{3D}^{f+k}$ is the 3D loss at timestamp $f + k$. To further enhance pose realism, we introduce two additional regularization terms: a cross-entropy classification loss, $\mathcal{L}_{act}$, applied to the action prompt derived from the encoder (**E**), and an $L1$ bone length regularization loss, $\mathcal{L}_{BL}$, which ensures consistent bone proportions by penalizing length discrepancies between paired joints (e.g., left and right limbs). The overall loss function of `DreamPose3D` is given by:

$$\mathcal{L}_{net} = \mathcal{L}'_{3D} + \lambda_{act}\mathcal{L}_{act} + \lambda_{BL}\mathcal{L}_{BL} \tag{7}$$

where λ_{act}, and λ_{BL} are the weights assigned for the action class and bone length regularization terms, respectively.

Sampling Strategy. To enable the model to effectively learn both immediate and long-term dependencies, we adopt a two-stage sampling strategy during training. In the initial stage, we set $n = 1$, focusing on accurate 3D pose prediction for the current frame only. In the second stage, we set $n > 1$. Specifically, the best results were obtained at $n = 3$. The weights for the 3D hallucinatory poses λ_{3D}^{t+k} are set to be $1/(1 + |k|)$ each, allowing the model to incorporate information from both past and future frames with a smaller weight to poses that are further away in time. This shift in weighting ensures the importance of hallucinated poses based on how far they are from the current frame, leading to improved accuracy in long-term pose prediction.

Inference. In the inference phase, `DreamPose3D` begins by initializing the encoder **E** to predict the action class and subsequently generating E_c via the APL block. It is to note that no external input, such as action labels or prompts, is required; all contextual information is inferred directly from **X**. The model then performs the reverse diffusion process, sampling N poses from the 2D input **X** and iteratively passing them through the denoiser $\mathcal{D}$ for K steps to obtain the final 3D pose sequence, represented as $\{Y_T^h\}_{h=1}^{N}$. The HPD decoder was designed mainly to improve the performance of the denoiser by regularizing its temporal understanding during training, so during inference, n was always set to be 1; that is, it captures only the current frame's pose at each timestep. The resulting output for each timestep T is denoted by the 3D pose $\hat{Y}_0^h$, yielding a temporally coherent sequence of N poses.

4 Experiments

4.1 Quantitative Results

Human3.6M. Table 1 and 2 presents a quantitative comparison of `DreamPose3D` with various state-of-the-art (SOTA) 3D HPE techniques on the Human3.6M dataset. We conducted experiments using two different types of inputs following the literature of 3D HPE: 2D poses from a keypoint

Table 1: **Quantitative comparison with the state-of-the-art 3D HPE methods on the Human3.6M dataset.** N: the number of input frames. CPN, HRNet, SH: using CPN Chen et al. [2018], HRNet Sun et al. [2019], and SH Newell et al. [2016] as the 2D keypoint detectors to generate the inputs. GT: using the groundtruth 2D keypoints as inputs. The best and second-best results are highlighted in **bold** and underlined formats. Colored differences indicate the margin between the top two results.

Method	N	Human3.6M (DET)			Human3.6M (GT)			Year
		Detector	mPJPE ↓	P-mPJPE ↓	Detector	mPJPE ↓	P-mPJPE ↓	
VideoPose3D Pavllo et al. [2019]	243	CPN	46.8	36.5	GT	37.8	/	CVPR'19
Anatomy Chen et al. [2020]	243	CPN	44.1	35.0	GT	32.3	/	CSVT'21
P-STMO Shan et al. [2022]	243	CPN	42.8	34.4	GT	29.3	/	ECCV'22
MixSTE Zhang et al. [2022]	243	HRNet	39.8	30.6	GT	21.6	/	CVPR'22
MHFormer Li et al. [2022b]	351	CPN	43.0	34.4	GT	30.5	/	CVPR'22
Diffpose Holmquist and Wandt [2023]	243	CPN	36.9	28.7	GT	18.9	/	CVPR'23
GLA-GCN Yu et al. [2023]	243	CPN	44.4	34.8	GT	21.0	17.6	ICCV'23
ActionPrompt Zheng et al. [2023]	243	HRNet	41.8	29.5	GT	22.7	/	ICME'23
NC-RetNet Zheng et al. [2020]	243	CPN	40.4	32.5	GT	21.5	/	ECCV'24
MotionBERT Zhu et al. [2023]	243	SH	37.5	/	GT	16.9	/	ICCV'23
D3DP Shan et al. [2023]	243	CPN	35.4	28.7	GT	18.4	/	ICCV'23
KTPFormer Peng et al. [2024]	243	CPN	33.0	26.2	GT	18.1	/	CVPR'24
FinePOSE Xu et al. [2024]	243	CPN	31.9	25.0	GT	16.7	12.7	CVPR'24
DreamPose3D (Ours)	243	CPN	**29.5** (-2.4)	**23.4** (-1.6)	GT	**15.9** (-0.8)	**12.2** (-0.5)	

Table 2: **Quantitative comparison with the state-of-the-art 3D HPE methods on the Human3.6M dataset using 2D keypoint detectors to generate the inputs.** *Dir.*, *Disc.*, . . . , and *Walk* correspond to 15 action classes. *Avg* indicates the average MPJPE among 15 action classes. The best and second-best results are highlighted in **bold** and underlined formats.

Method / mPJPE ↓	Human3.6M (DET)															
	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
VideoPose3D Pavllo et al. [2019]	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
SRNet Zeng et al. [2020]	46.6	47.1	43.9	41.6	45.8	49.6	46.5	40.0	53.4	61.1	46.1	42.6	43.1	31.5	32.6	44.8
RIE Shan et al. [2021]	40.8	44.5	41.4	42.7	46.3	55.6	41.8	41.9	53.7	60.8	45.0	41.5	44.8	30.8	31.9	44.3
Anatomy Chen et al. [2020]	41.4	43.5	40.1	42.9	46.6	51.9	41.7	42.3	53.9	60.2	45.4	41.7	46.0	31.5	32.7	44.1
P-STMO Shan et al. [2022]	38.9	42.7	40.4	41.1	45.6	49.7	40.9	39.9	55.5	59.4	44.9	42.2	42.7	29.4	29.4	42.8
MixSTE Zhang et al. [2022]	36.7	39.0	36.5	39.4	40.2	44.9	39.8	36.9	47.9	54.8	39.6	37.8	39.3	29.7	30.6	39.8
MHFormer Li et al. [2022b]	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
Diffpose Holmquist and Wandt [2023]	33.2	36.6	33.0	35.6	37.6	45.1	35.7	35.5	46.4	49.9	37.3	35.6	36.5	24.4	24.1	36.9
GLA-GCN Yu et al. [2023]	41.3	44.3	40.8	41.8	45.9	54.1	42.1	41.5	57.8	62.9	45.0	42.8	45.9	29.4	29.9	44.4
ActionPrompt Zheng et al. [2023]	37.7	40.2	39.8	40.6	43.1	48.0	38.8	38.9	50.8	63.2	42.0	40.0	42.0	30.5	31.6	41.8
NC-RetNet Zheng et al. [2020]	36.9	40.1	38.7	38.3	42.9	48.6	38.2	40.0	52.5	55.4	42.3	38.7	39.7	26.2	27.8	40.4
MotionBERT Zhu et al. [2023]	36.1	37.5	35.8	32.1	40.3	46.3	36.1	35.3	46.9	53.9	39.5	36.3	35.8	25.1	25.3	37.5
D3DP Shan et al. [2023]	33.0	34.8	31.7	33.1	37.5	43.7	34.8	33.6	45.7	47.8	37.0	35.0	35.0	24.3	24.1	35.4
KTPFormer Peng et al. [2024]	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
FinePOSE Xu et al. [2024]	31.4	31.5	28.8	29.7	34.3	36.5	29.2	30.0	42.0	42.5	33.3	31.9	31.4	22.6	22.7	31.9
DreamPose3D (Ours)	**26.5** (-3.6)	**28.9** (-2.6)	**26.3** (-2.5)	**27.0** (-2.3)	**31.6** (-2.7)	**34.9** (-1.6)	**27.1** (-2.1)	**28.0** (-2.0)	**38.8** (-3.2)	**41.1** (-1.4)	**30.3** (-2.0)	**29.4** (-2.5)	**29.5** (-1.9)	**21.6** (-0.5)	**22.0** (-0.7)	**29.5** (-2.4)

detector Chen et al. [2018] and groundtruth 2D poses. The results, as shown in Table 1, demonstrate the superior performance of DreamPose3D on both input settings. DreamPose3D outperforms the previous SOTA method, FinePOSE Xu et al. [2024], by 7.5% and 6.4% in terms of mPJPE and P-mPJPE, respectively, when using a 2D keypoint detector. When using groundtruth 2D keypoints, DreamPose3D still outperforms with improvements of 4.8% and 3.9% for mPJPE and P-mPJPE, respectively. Table 2 further breaks down the results on the 15 action classes used in the Human3.6M dataset. Notably, DreamPose3D achieves significant improvements in action categories such as "Directions", "Sit", and "Phone", with reductions in mPJPE of 11.9%, 7.6%, and 2.7%, respectively.

MPI-INF-3DHP. Table 3 reports a comparison between DreamPose3D and various SOTA techniques on the MPI-INF-3DHP dataset. The results highlight that DreamPose3D achieves SOTA performance across key metrics, including PCK, AUC, and mPJPE. These improvements show the strong generalization capability of DreamPose3D, particularly in challenging outdoor scenes consisting of variations in pose, lighting, and background conditions.

MLBPitchDB. To evaluate the **generalizability** of DreamPose3D, we use the MLBPitchDB dataset Bright et al. [2023], which consists of high inherent motion blur and occlusions. Table 4 demonstrates the SOTA performance of DreamPose3D despite the challenging poses and the noisy 2D pose quality. Our method demonstrates a significant improvement of 8.7% and 6.6% with groundtruth 2D poses and 2D poses from a detector Xu et al. [2022], respectively. This highlights the effectiveness of DreamPose3D in handling complex scenarios where motion blur and occlusion affect the quality of pose estimation.

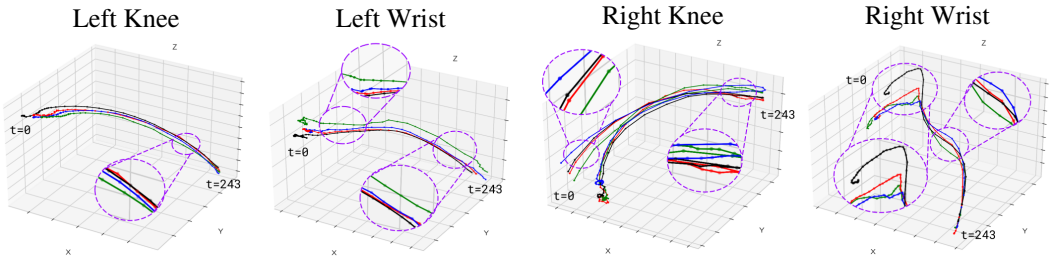


Figure 3: **Comparison of trajectories on the Human3.6M dataset for a sitting action.** The **black** trajectory represents the groundtruth, while the **blue**, **green**, and **red** trajectories correspond to the predictions from FinePOSE Xu et al. [2024], KTPFormer Peng et al. [2024], and `DreamPose3D`, respectively.

Table 3: **Quantitative comparison with the SOTA 3D HPE methods on the MPI-INF-3DHP dataset using groundtruth 2D keypoints as inputs.** N: the number of input frames. The best and second-best results are highlighted in **bold** and underlined formats.

Method	N	MPI-INF-3DHP		
		PCK↑	AUC↑	mPJPE↓
VideoPose3D Pavllo et al. [2019]	81	86.0	51.9	84.0
MixSTE Zhang et al. [2022]	27	94.4	66.5	54.9
PoseFormerV2 Zhao et al. [2023]	81	97.9	78.8	27.8
MHFormer Li et al. [2022b]	9	93.8	63.3	58.0
Diffpose Holmquist and Wandt [2023]	81	98.0	75.9	29.1
D3DP Shan et al. [2023]	243	98.0	79.1	28.1
FinePOSE Xu et al. [2024]	243	98.9	80.0	26.2
KTPFormer Peng et al. [2024]	27	98.9	84.4	19.2
`DreamPose3D` **(Ours)**	81	**99.1** (0.2)	**84.5** (0.1)	**18.9** (-0.3)

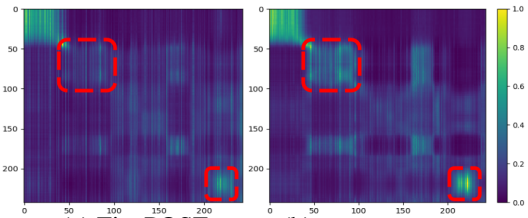
Table 4: **Quantitative comparison with the SOTA 3D HPE methods on the MLBPitchDB dataset using 2D keypoints from groundtruth and a detector (ViTPose) as inputs.** N: the number of input frames. The best and second-best results are highlighted in **bold** and underlined formats.

Method	N	2D Keypoints	
		GT↓	ViTPose↓
VideoPose3D Pavllo et al. [2019]	351	40.2	75.8
MHFormer Li et al. [2022b]	351	34.8	67.5
D3DP Shan et al. [2023]	243	26.6	61.7
FinePOSE Xu et al. [2024]	243	23.9	57.3
`DreamPose3D` **(Ours)**	243	**21.8** (-2.1)	**53.5** (-3.8)

4.2 Qualitative Analysis

The qualitative results of `DreamPose3D` are demonstrated in Figure 5. We compare our method with two SOTA techniques, FinePOSE Xu et al. [2024] and KTPFormer Peng et al. [2024]. All estimated poses are overlaid alongside the groundtruth 3D poses for different actions. It can be observed that, especially for actions like "Eating" and "Waiting", `DreamPose3D` produces poses that are more closely aligned with the groundtruth. This shows the ability of `DreamPose3D` to model robust kinematic joint relationships from the SRE block, resulting in more accurate pose reconstruction, whereas prior methods show greater deviation from the groundtruth.

Figure 3 illustrates the consistency of the trajectories for four different joints of a human performing the "Sitting" action. Compared to prior methods, `DreamPose3D` produces smoother and plausible motion paths, highlighting its ability to generate temporally coherent 3D poses through hallucination. Furthermore, Figure 4 shows the temporal attention map for a sequence, where the x-axis corresponds to the query of 243 frames and the y-axis indicates the attention output. The highlighted regions demonstrate high attention-weight concentration, contributing to the ability of `DreamPose3D` to hallucinate coherent poses and intent-conditioned denoising.



(a) FinePOSE (b) `DreamPose3D`

Figure 4: **Comparison of the attention maps between ours and FinePOSE Xu et al. [2024].** The x-axis corresponds to queries and the y-axis to predicted outputs. Lighter color indicates stronger attention.

4.3 Ablation Study

Effect of different blocks on `DreamPose3D`. To quantify the influence of individual blocks on `DreamPose3D`, we conduct an ablation as presented in Table 5. Starting with a baseline model

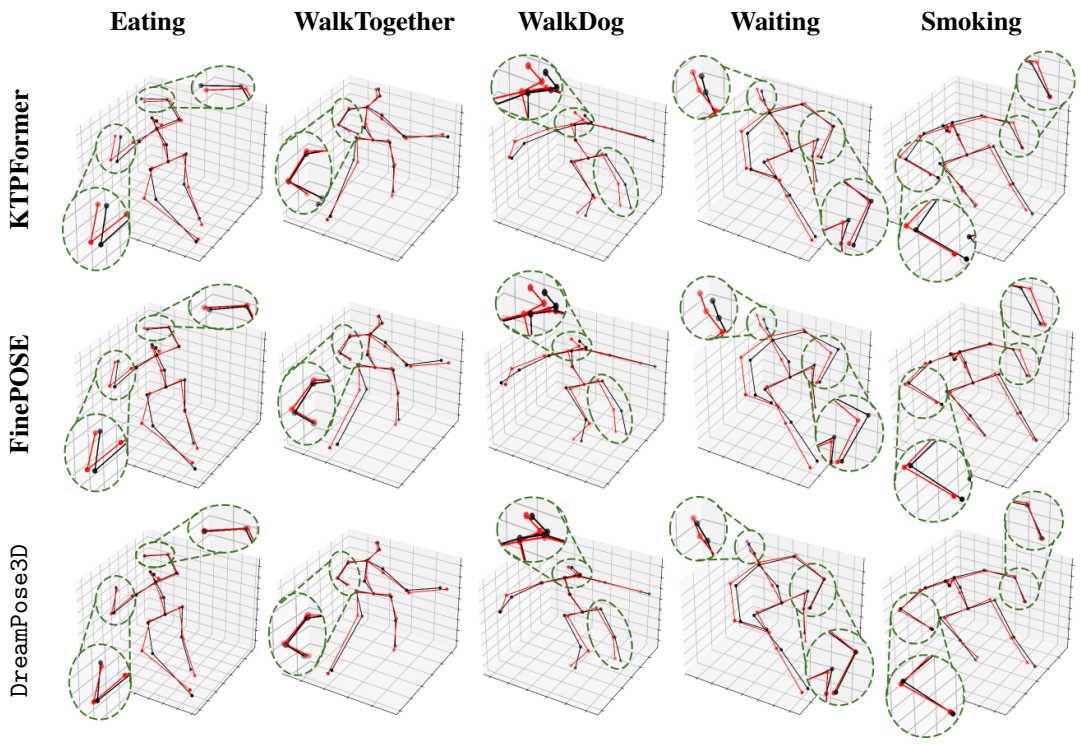


Figure 5: **Qualitative comparison of DreamPose3D with FinePOSE Xu et al. [2024] and KTP-Former Peng et al. [2024] on the Human3.6M dataset**. The **red** skeleton represents the groundtruth pose, while the **black** skeleton shows DreamPose3D's prediction. **Green** dashed lines highlight regions where DreamPose3D demonstrates superior accuracy over the other methods.

Table 5: **Ablation study on different configurations of DreamPose3D on the Human3.6M dataset using 2D keypoint detectors as inputs.** All configurations use the denoiser backbone Zhang et al. [2022].

Method	Module			mPJPE ↓	P-mPJPE ↓
	SRE	APL	HPD		
Baseline				37.4	30.7
w/o APL	✓		✓	31.9	24.9
w/o SRE		✓	✓	32.2	24.8
w/o HPD	✓	✓		30.1	24.2
DreamPose3D **(Ours)**	✓	✓	✓	**29.5** (-0.6)	**23.4** (-0.8)

Table 6: **Impact of objective functions in DreamPose3D on the Human3.6M dataset using 2D keypoint detectors as inputs.** Network fails to converge without $\mathcal{L}'_{3D}$ due to lack of pose regularization.

Method	Module			mPJPE ↓	P-mPJPE ↓
	$\mathcal{L}'_{3D}$	$\mathcal{L}_{act}$	$\mathcal{L}_{BL}$		
w/ $\mathcal{L}'_{3D}$	✓			31.9	25.6
w/ $\mathcal{L}'_{3D} + \mathcal{L}_{act}$	✓	✓		31.1	24.8
w/ $\mathcal{L}'_{3D} + \mathcal{L}_{BL}$	✓		✓	30.4	24.0
DreamPose3D **(Ours)**	✓	✓	✓	**29.5** (-0.9)	**23.4** (-0.6)

consisting solely of a backbone denoiser block Zhang et al. [2022], we incorporated each of the proposed blocks. The final model, incorporating all components, achieved substantial performance gains of 7.9% and 7.3% in terms of mPJPE and P-mPJPE, respectively, compared to the denoiser-only baseline model. These results highlight the significance of intent, hallucinating 3D poses, and kinematic representation modeling for the 3D HPE task.

Effect of different objective functions on DreamPose3D. Table 6 evaluates the influence of the bone length and prompt regularization on DreamPose3D performance. Incorporating these losses into the objective function yields improvements of 2.4% and 2.2% in mPJPE and P-mPJPE, respectively. These results show the importance of explicit regularization of prompts and bone length modeling.

5 Conclusion

In this work, we introduced DreamPose3D, a novel framework for 3D human pose estimation that mimics human-like reasoning by inferring motion intent and hallucinating plausible 3D pose

trajectories. Our method integrates the power of language models to produce action-aware information, which conditions a transformer-based denoiser to predict 3D hallucinatory poses. Furthermore, novel ways to incorporate kinematic joint information into the attention mechanism are explored for the denoiser. Experimentation on three datasets demonstrates the efficacy of our model in capturing robust spatial and temporal information by achieving SOTA performance in challenging scenarios through its intent-conditioned hallucination approach.

Supplementary Material

The supplementary material is organized as follows:

A Additional Formulation of Diffusion Models

Our work is built over DDPMs Ho et al. [2020], which are trained to reverse a diffusion process that gradually adds Gaussian noise to the training data. This can be formulated as a two-stage process:

Forward process. The diffusion process q takes in as input a sequence of groundtruth 3D pose $\mathbf{Y}_0 \in \mathbb{R}^{N \times J \times 3}$ and a sampled time step $t \in [0, T]$, where T is the maximum number of diffusion steps. Then the input is diffused by adding Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ at each step to generate $\mathbf{Y}_t(t \rightarrow T)$. This can be mathematically written as

$$q(\mathbf{Y}_t|\mathbf{Y}_0) = \sqrt{\bar{\alpha}}\mathbf{Y}_0 + \epsilon\sqrt{1-\bar{\alpha}} \tag{8}$$

where $\bar{\alpha}$ is a constant dependent on the variance schedule.

Reverse process. The reverse process aims in reconstructing the uncontaminated 3D poses $\bar{\mathbf{Y}}_0$ using a denoiser $\mathcal{D}$ from the noised 3D poses $\mathbf{Y}_t$ from the forward process:

$$\hat{\mathbf{Y}}_0 = \mathcal{D}(\mathbf{Y}_t, \mathbf{X}, t) \tag{9}$$

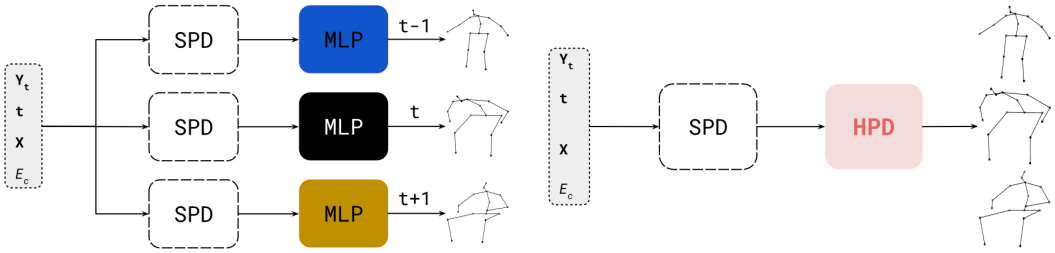
where the denoiser $\mathcal{D}$ takes as input the noised 3D pose sequence $\mathbf{Y}_t$, given the 2D pose sequence $\mathbf{X}$ and the time step to reconstruct the denoised output 3D pose sequence $\hat{\mathbf{Y}}_0$.

B Implementation Details

`DreamPose3D` is implemented in PyTorch Paszke et al. [2019] and optimized using the AdamW optimizer Loshchilov and Hutter [2017] with a learning rate of 1×10^{-5} and a weight decay of 1×10^{-4} per epoch. The model is trained on three A6000 GPUs with a batch size of 4 for 100 epochs, which takes approximately two days. During training, we freeze the pretrained CLIP model to avoid updating its weights. The denoiser backbone (SPD) is based on MixSTE Zhang et al. [2022], which incorporates $M = 16$ stacks of spatial and temporal transformers with a channel size of 512. The SRE block utilizes MHSA with $L = 6$.

C Datasets

Human3.6M. The Human3.6M dataset Ionescu et al. [2014] is the standard benchmark for 3D HPE in controlled indoor environments. It comprises over 3.6 million images capturing 15 daily activities performed by 11 subjects at a frame rate of 50 Hz. Following common practice Xu et al. [2024], Bright et al. [2024b], we train on data from five subjects (S1, S5, S6, S7, S8) and evaluate on two subjects (S9, S11). Some action prompts used to train MLP_p with Human3.6M dataset include "Sitting", "Phoning", "Greeting", "Discussion", etc.



(a) Hallucinator-Free Design (b) Hallucinator Design

Figure 6: **Different module configurations for** `DreamPose3D`**.** (a) illustrates alternative setup for the proposed `DreamPose3D` module (b).

MPI-INF-3DHP. The MPI-INF-3DHP dataset Mehta et al. [2017a] is the largest publicly available 3D human pose dataset capturing diverse motions in both indoor and outdoor environments. Comprising over 1.3 million images, it surpasses the motion variety offered by the Human3.6M dataset. Some action prompts used to train MLP_p with the MPI-INF-3DHP dataset include "Walking", "Running", "Crouching", "Bending", etc.

MLBPitchDB. The MLBPitchDB dataset Bright et al. [2023] comprises 30,000 images spanning 150 pitch sequences from diverse MLB games. Characterized by inherent blur and occlusions due to the rapid motion of baseball players captured at 30fps by broadcast cameras, this dataset rigorously evaluates `DreamPose3D`'s capacity to generate accurate and robust pose predictions under challenging real-world conditions. The MLP_p block was trained with action prompts including "Throwing" and "Hitting" based on the performed action by the baseball players.

D Evaluation Metrics

To evaluate the accuracy of `DreamPose3D`, we compute the mean Per Joint Position Error (mPJPE) Ionescu et al. [2014] and Procrustes-aligned mPJPE (PA-mPJPE) Zhou et al. [2019] measured in millimeters (mm). We report the mPJPE and PA-mPJPE between the predictions and groundtruth 3D poses. In addition, the Percentage of Correct Keypoint (PCK) with a threshold of 150 mm, and Area Under Curve (AUC) for a range of PCK thresholds are used as metrics for the MPI-INF-3DHP dataset. Following typical work on 3D human pose estimation Bright et al. [2024b], Xu et al. [2024], Zhu et al. [2023], we report the metrics in Human3.6M validation data split and the test set of the MPI-INF-3DHP and MLBPitchDB datasets.

E More Ablation Study

Effect of different model configurations on `DreamPose3D`**.** Table 7 presents an ablation study comparing an alternative design based on model simplicity. Hallucinator-free design, depicted in Figure 6(a), where the HPD block in `DreamPose3D` (Figure 6(b)) is replaced with n denoisers having shared weights to predict n poses. Experimental results indicate that `DreamPose3D` achieves significant improvements, with reductions of 3.9% in mPJPE and 2.9% in PA-mPJPE compared to the hallucinator-free design. Additionally, the hallucinator-free design incurred a 24.95-second increase in inference time compared to `DreamPose3D`.

Sampling Strategy. Table 8 presents an ablation study on different sampling strategies evaluated on the Human3.6M dataset. The strategies include: (1) **Fixed Weights Sampling**, where fixed weights, λ_{3D}^{f+k}, are assigned uniformly to all predictions throughout training; (2) **Falloff Sampling**, which assigns the highest weight to the center frame, with weights decreasing as the temporal distance from the center frame increases; and (3) **Controlled Sampling**, which restricts training to predict only the center frame ($n = 1$) for the first 25 epochs. For the remaining epochs, predictions are reduced with a decay, represented as: $\lambda_{3D}^{f+k} = 1/(1 + |k|)$. Results indicate that the controlled sampling strategy achieves the best performance, highlighting the effectiveness of gradually increasing prediction difficulty during training with a decay factor.

Table 7: **Quantitative comparison of different model configurations on `DreamPose3D`.** CT: Computation Time.

	mPJPE ↓	P-mPJPE ↓	CT ↓
Hallucinator-free design	30.7	24.1	85.62
`DreamPose3D`	**29.5** (-1.2)	**23.4** (-0.7)	60.67 (+24.95)

Table 8: **Quantitative comparison of the conventional sampling strategy on `DreamPose3D` during training.**

Method	mPJPE ↓	P-mPJPE ↓
Fixed Weights Sam.	30.1	24.3
Falloff Sam.	30.0	24.1
Controlled Sam.	**29.5** (-0.5)	**23.4** (-0.7)

Table 9: **Impact of the number of hallucinated poses in HPD of `DreamPose3D` on Human3.6M dataset using 2D keypoint detectors as inputs during training.** n denotes the sequence length/ number of output predictions.

Method	mPJPE ↓	P-mPJPE ↓
HPD ($n = 1$)	30.1	24.2
HPD ($n = 5$)	29.9	23.7
HPD ($n = 7$)	30.0	23.9
HPD ($n = 3$)	**29.5** (-0.5)	**23.4** (-0.5)

Table 10: **Impact of the quality of text prompts on `DreamPose3D`**

Method	mPJPE ↓	P-mPJPE ↓
No CLIP module (APL)	31.9	24.9
Random Prompts	<u>31.2</u>	<u>24.0</u>
`DreamPose3D` (Ours)	**29.5** (-0.7)	**23.4** (-0.6)

Effect of different pose decoder configurations on `DreamPose3D`. Table 9 presents an ablation study comparing the performance of the HPD block with varying numbers of hallucinated poses (n) during training. Experimental results demonstrate that the HPD consistently achieves optimal results with three poses ($n = 3$) as output, resulting in 1.7% and 2.1% reductions in mPJPE and PA-mPJPE, respectively, when compared against HPD at $n = 1$ during training. While in inference, n is always set to 1 since the objective of hallucinating poses was only to improve the training coherence.

Impact of Text Prompts. Table 10 shows the impact the quality of text prompts have on `DreamPose3D`. Interestingly, the models performed better with random prompts than when the APL block was entirely omitted. This suggests that the CLIP model positively influences the overall performance, regardless of the specific prompt used. However, employing relevant prompts yields the best results of 2.2% and 2.5% in terms of mPJPE and PA-mPJPE, respectively.

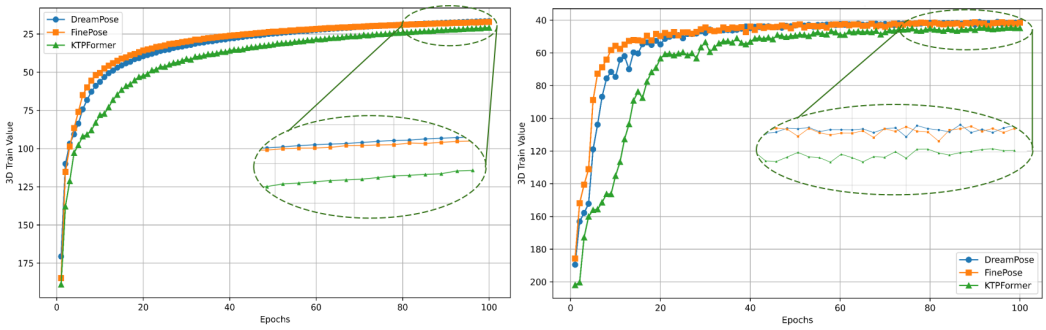


Figure 7: **Training and validation curves** in Human3.6M dataset for FinePOSEXu et al. [2024], KTPFormer Peng et al. [2024] and `DreamPose3D`.

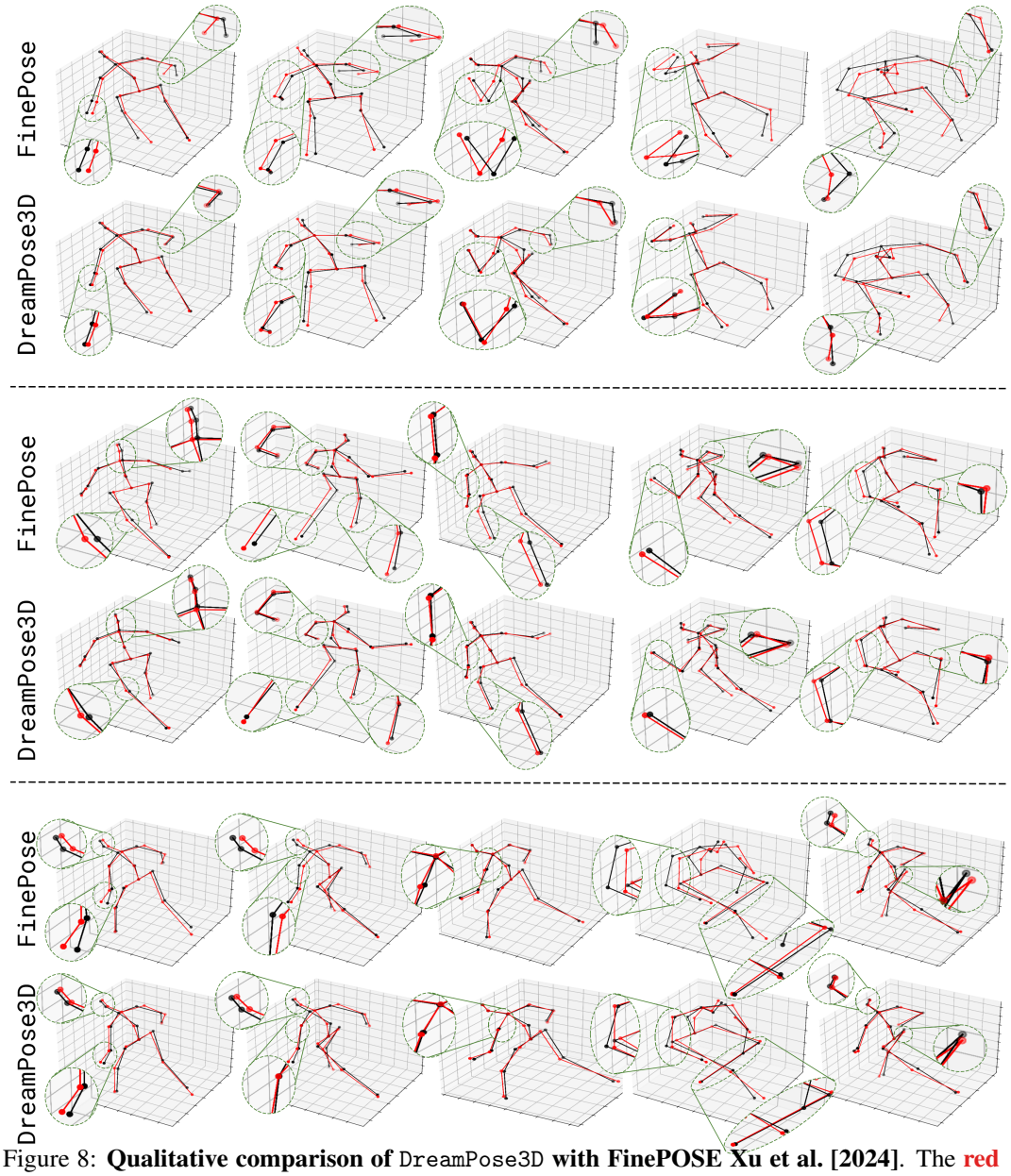


Figure 8: **Qualitative comparison of** `DreamPose3D` **with FinePOSE Xu et al. [2024]**. The **red** skeleton represents the groundtruth pose, while the **black** skeleton shows `DreamPose3D`'s prediction. **Green** dashed lines highlight regions where `DreamPose3D` demonstrates superior accuracy over the other methods.

F More Qualitative Results

In this section, we present more qualitative results of `DreamPose3D`. Figure 8 depicts the visual comparison of `DreamPose3D` with FinePOSE Xu et al. [2024] model which was the previous SOTA in 3DPE task. The highlighted regions depict the regions in which `DreamPose3D` significantly outperforms FinePOSE in pose alignment with the groundtruth. In addition, Figure 7 presents the training and validation curves of KTPFormer Peng et al. [2024], FinePOSE Xu et al. [2024], and the proposed `DreamPose3D` method.

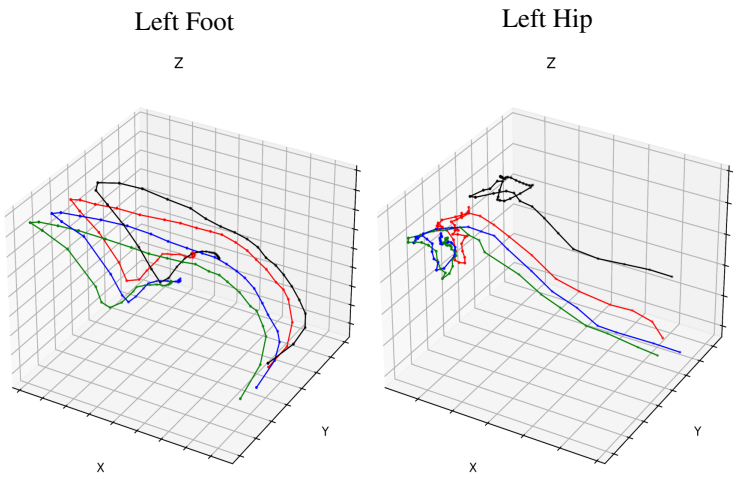


Figure 9: **Gap in Existing 3DPE Methods.** The **black** trajectory represents the groundtruth, while the **blue**, **green**, and **red** trajectories correspond to the predictions from FinePOSE Xu et al. [2024], KTPFormer Peng et al. [2024], and `DreamPose3D`, respectively.

G Limitations and Future Works

Inconsistencies in the predicted poses are observed, particularly in high-occlusion scenarios, including cases of self-occlusion by the target individual. This is illustrated in Figure 9, where the trajectories of the left hip and left foot show noticeable misalignment compared to the groundtruth (black). Despite these misalignments, it is worth noting that `DreamPose3D` demonstrates closer alignment to the groundtruth compared to prior SOTA methods, such as FinePOSE Xu et al. [2024] and KTPFormer Peng et al. [2024].

For future research, we aim to explore the integration of physics-based priors to further enhance pose realism and incorporate image features for richer contextual understanding, potentially improving performance in highly unconstrained, in-the-wild settings.

H Broader Impact Statement

By mimicking aspects of human cognition for human motion understanding, `DreamPose3D` advances 3D pose estimation, benefiting applications in animation, sports analytics, augmented reality, and human-object interaction. The ability for such systems to act under ambiguous or noisy conditions makes it especially valuable for real-world deployment with no specific need for expensive sensors or controlled environmental setup.

More broadly, this work contributes to a growing trend of incorporating human-like reasoning and intent understanding into AI systems. Such domains can be adopted beyond just pose estimation, including assistive robotics, human-computer interaction, and embodied agents, where an implicit understanding of *why* a person moves is necessary to find, rather than just *how* they move.

While our work primarily focuses on technical improvements, we acknowledge that the deployment of pose estimation systems must carefully consider potential ethical concerns, such as misuse of surveillance or privacy violations.

I Summary of Experimental Findings

The experimental results systematically demonstrated the effectiveness of `DreamPose3D`'s two cognitively inspired capabilities: *intent inference* and *hallucinative motion generation*. We evaluated their contributions through targeted ablations and visualizations across multiple datasets and configurations.

Intention. Table 5 (from the main paper) shows that removing the APL block leads to a reduction of **7.5%** in mPJPE, confirming the critical role of intent guidance. Table 10 further demonstrates

that the improvement is not solely due to CLIP embeddings, i.e., prompt structure and relevance significantly affect performance (**2.2**% in mPJPE).

Hallucination. Ablating the HPD block results in a performance drop of **2.0%** in mPJPE as shown in Table 5 (from the main paper), highlighting its importance for temporal consistency. Figure 3 (from the main paper) visualizes smoother joint trajectories compared to prior SOTA, and Figure 4 (from the main paper) shows more concentrated attention weights than FinePose, supporting improved temporal reasoning. Tables 7 and 9 also confirm that both the HPD design and the number of hallucinated outputs contribute positively to performance.

Together, these results confirm that `DreamPose3D` benefits from the proposed modules. The model achieves state-of-the-art results on both Human3.6M and MPI-INF-3DHP, and shows strong generalization on the broadcast baseball dataset, handling occlusions, motion blur, resulting in ambiguous 2D inputs with greater robustness than prior approaches. This reinforces the value of incorporating intent and hallucination into temporally structured 3D pose estimation.

References

Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021.

Yunpeng Bai, Cairong Wang, Shuzhao Xie, Chao Dong, Chun Yuan, and Zhi Wang. Textir: A simple framework for text-based editable image restoration. *arXiv preprint arXiv:2302.14736*, 2023.

Bavesh Balaji, Jerrin Bright, Yuhao Chen, Sirisha Rambhatla, John Zelek, and David Clausi. Seeing beyond the crop: Using language priors for out-of-bounding box keypoint prediction. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 102897–102918. Curran Associates, Inc., 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/ba84da6921f3040b74ee163aa7451f53-Paper-Conference.pdf`.

Jerrin Bright, Yuhao Chen, and John Zelek. Mitigating motion blur for robust 3d baseball player pose modeling for pitch analysis. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, pages 63–71, 2023.

Jerrin Bright, Bavesh Balaji, Yuhao Chen, David A Clausi, and John S Zelek. Pitchernet: Powering the moneyball evolution in baseball video analytics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3420–3429, 2024a.

Jerrin Bright, Bavesh Balaji, Harish Prakash, Yuhao Chen, David A Clausi, and John Zelek. Distribution and depth-aware transformers for 3d human mesh recovery. *arXiv preprint arXiv:2403.09063*, 2024b.

Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, and Jiebo Luo. Anatomy-aware 3d human pose estimation in videos. *CoRR*, abs/2002.10322, 2020. URL `https://arxiv.org/abs/2002.10322`.

Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation, 2018. URL `https://arxiv.org/abs/1711.07319`.

Jeongjun Choi, Dongseok Shim, and H Jin Kim. Diffupose: Monocular 3d human pose estimation via denoising diffusion probabilistic model. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3773–3780. IEEE, 2023.

Ciprian Corneanu, Raghudeep Gadde, and Aleix M Martinez. Latentpaint: Image inpainting in latent space with diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4334–4343, 2024.

Xiaoxiao Du, Ram Vasudevan, and Matthew Johnson-Roberson. Bio-lstm: A biomechanically inspired recurrent neural network for 3-d pedestrian pose and gait prediction. *IEEE Robotics and Automation Letters*, 4(2):1501–1508, 2019. doi: 10.1109/LRA.2019.2895266.

Zackory Erickson, Henry M Clever, Vamsee Gangaram, Eliot Xing, Greg Turk, C Karen Liu, and Charles C Kemp. Characterizing multidimensional capacitive servoing for physical human–robot interaction. *IEEE Transactions on Robotics*, 39(1):357–372, 2022.

Runyang Feng, Yixing Gao, Tze Ho Elden Tse, Xueqing Ma, and Hyung Jin Chang. Diffpose: Spatiotemporal diffusion model for video-based human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14861–14872, 2023.

Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

Liang-Yan Gui, Kevin Zhang, Yu-Xiong Wang, Xiaodan Liang, José M. F. Moura, and Manuela Veloso. Teaching robots to predict human motion. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 562–567, 2018. doi: 10.1109/IROS.2018.8594452.

Shaoxiang Guo, Qing Cai, Lin Qi, and Junyu Dong. Clip-hand3d: Exploiting 3d hand pose estimation via context-aware prompting. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4896–4907, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15977–15987, 2023.

Shengnan Hu, Ce Zheng, Zixiang Zhou, Chen Chen, and Gita Sukthankar. Lamp: Leveraging language prompts for multi-person pose estimation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3759–3766. IEEE, 2023.

Xiaoling Hu and Chang Liu. Animal pose estimation based on contrastive learning with dynamic conditional prompts. *Animals*, 14(12):1712, 2024.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 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, jul 2014.

Wonhui Kim, Manikandasriram Srinivasan Ramanagopal, Charles Barto, Ming-Yuan Yu, Karl Rosaen, Nick Goumas, Ram Vasudevan, and Matthew Johnson-Roberson. Pedx: Benchmark dataset for metric 3-d pose estimation of pedestrians in complex urban intersections. *IEEE Robotics and Automation Letters*, 4(2):1940–1947, 2019. doi: 10.1109/LRA.2019.2896705.

Jeong-gi Kwak, Erqun Dong, Yuhe Jin, Hanseok Ko, Shweta Mahajan, and Kwang Moo Yi. Vivid-1-to-3: Novel view synthesis with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6775–6785, 2024.

Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022a.

Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13147–13156, 2022b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022.

Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017a. doi: 10.1109/3dv.2017.00064. URL `http://gvv.mpi-inf.mpg.de/3dhp_dataset`.

Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *Acm transactions on graphics (tog)*, 36(4):1–14, 2017b.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation, 2016. URL `https://arxiv.org/abs/1603.06937`.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Jihua Peng, Yanghong Zhou, and PY Mok. Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1123–1132, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Konstantinos Rematas, Ira Kemelmacher-Shlizerman, Brian Curless, and Steve Seitz. Soccer on your tabletop. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4738–4747, 2018.

Eugenio Scaliti, Kiri Pullar, Giulia Borghini, Andrea Cavallo, Stefano Panzeri, and Cristina Becchio. Kinematic priming of action predictions. *Current Biology*, 33(13):2717–2727, 2023.

Wenkang Shan, Haopeng Lu, Shanshe Wang, Xinfeng Zhang, and Wen Gao. Improving robustness and accuracy via relative information encoding in 3d human pose estimation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3446–3454, 2021.

Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In *ECCV*, pages 461–478. Springer, 2022.

Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14761–14771, 2023.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=SJ1kSyO2jwu`.

Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Medical Imaging with Deep Learning*, pages 1623–1639. PMLR, 2024.

Jinglin Xu, Yijie Guo, and Yuxin Peng. Finepose: Fine-grained prompt-driven 3d human pose estimation via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 561–570, 2024.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022.

Bruce X.B. Yu, Zhi Zhang, Yongxu Liu, Sheng-hua Zhong, Yan Liu, and Chang Wen Chen. Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8818–8829, October 2023.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36, 2024.

Ailing Zeng, Xiao Sun, Fuyang Huang, Minhao Liu, Qiang Xu, and Stephen Ching-Feng Lin. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In *ECCV*, 2020.

Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13232–13242, June 2022.

Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8877–8886, June 2023.

Hongwei Zheng, Han Li, Bowen Shi, Wenrui Dai, Botao Wan, Yu Sun, Min Guo, and Hongkai Xiong. Actionprompt: Action-guided 3d human pose estimation with text and pose prompting, 2023. URL `https://arxiv.org/abs/2307.09026`.

Jingxiao Zheng, Xinwei Shi, Alexander Gorban, Junhua Mao, Yang Song, Charles R Qi, Ting Liu, Visesh Chari, Andre Cornman, Yin Zhou, et al. Multi-modal 3d human pose estimation with 2d weak supervision in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4478–4487, 2022.

Kaili Zheng, Feixiang Lu, Yihao Lv, Liangjun Zhang, Chenyi Guo, and Ji Wu. 3d human pose estimation via non-causal retentive networks. In *ECCV*, 2020.

Jieming Zhou, Tong Zhang, Zeeshan Hayder, Lars Petersson, and Mehrtash Harandi. Diff3dhpe: A diffusion model for 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2092–2102, October 2023a.

Jieming Zhou, Tong Zhang, Zeeshan Hayder, Lars Petersson, and Mehrtash Harandi. Diff3dhpe: A diffusion model for 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2092–2102, 2023b.

Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G. Derpanis, and Kostas Daniilidis. Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(4):901–914, apr 2019. ISSN 0162-8828. doi: 10.1109/TPAMI.2018.2816031. URL `https://doi.org/10.1109/TPAMI.2018.2816031`.

Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.