This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# Superman: Unifying Skeleton and Vision for Human Motion Perception and Generation

Xinshun Wang[1,*], Peiming Li[1,*], Ziyi Wang[1,*],
Zhongbin Fang[1], Zhichao Deng[1], Songtao Wu[2], Jason Li[3], Mengyuan Liu[1]

[1]State Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School

[2]Sony R&D Center [3]Nanyang Technological University

https://github.com/BradleyWang0416/Superman

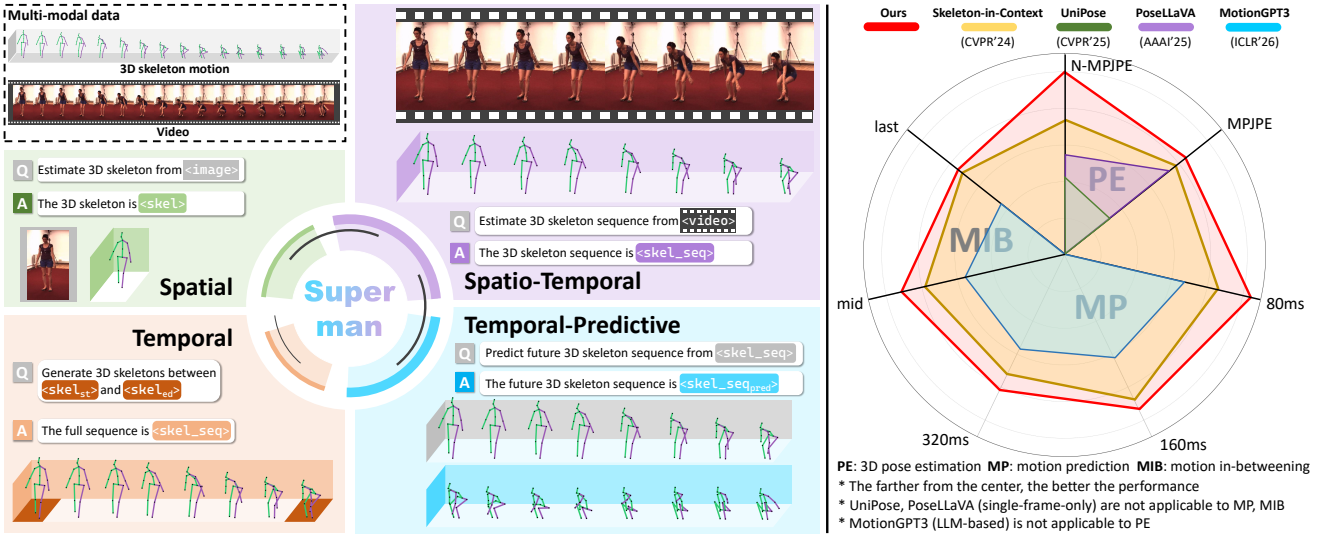



Figure 1. **A Unified Framework for Human Motion Perception and Generation.** Our unified model achieves state-of-the-art performance across three traditionally disparate tasks. Our method consistently outperforms existing methods across all tasks simultaneously: **(1) 3D Pose Estimation**, **(2) Motion Prediction**, and **(3) Motion In-betweening**.

## Abstract

*Human motion analysis tasks, such as temporal 3D pose estimation, motion prediction, and motion in-betweening, play an essential role in computer vision. However, current paradigms suffer from severe fragmentation. First, the field is split between "perception" models that understand motion from video but only output text, and "generation" models that cannot perceive from raw visual input. Second, generative MLLMs are often limited to single-frame, static poses using dense, parametric SMPL models, failing to handle temporal motion. Third, existing motion vocabularies are built from skeleton data alone, severing the link to the visual domain. To address these challenges, we introduce Superman, a unified framework that bridges visual perception with temporal, skeleton-based motion generation. Our solution is twofold. First, to overcome the modality disconnect, we propose a Vision-Guided Motion Tokenizer. Leveraging the natural geometric alignment between 3D skeletons and visual data, this module pioneers robust joint learning from both modalities, creating a unified, cross-modal motion vocabulary. Second, grounded in this motion language, a single, unified MLLM architecture is trained to handle all tasks. This module flexibly processes diverse, temporal inputs, unifying 3D skeleton pose estimation from video (perception) with skeleton-based motion prediction and in-betweening (generation). Extensive experiments on standard benchmarks, including Human3.6M, demonstrate that our unified method achieves state-of-the-art or competitive performance across all motion tasks. This showcases a more efficient and scalable path for generative motion analysis using skeletons.*

## 1. Introduction

Human motion analysis is a cornerstone of computer vision and robotics, with critical applications in human-computer

* Equal contribution. Corresponding author is Mengyuan Liu.

Table 1. Comparison with existing traditional and LLM/MLLM-based models for perceiving and/or generating human poses

| | Perception | | Generation | Tokenizer | Repr. |
|---|---|---|---|---|---|
| | vision | motion | motion | visual guidance | |
| *Traditional Multi-Task Models* | | | | | |
| MotionBERT [39] | × | ✓ | × | N/A | skeleton |
| SiC [30] | × | ✓ | × | N/A | skeleton |
| HiC [19] | × | ✓ | × | N/A | skeleton |
| *LLM / MLLM-based Models* | | | | | |
| MotionLLM [4] | video | ✓ | × | × | SMPL |
| MotionGPT [14] | × | ✓ | ✓ | × | SMPL |
| MotionGPT3 [38] | × | ✓ | ✓ | × | SMPL |
| LocLLM [28] | image | × | × | × | × |
| PoseLLaVA [9] | image | × | × | × | SMPL |
| UniPose [16] | image | × | × | × | SMPL |
| Superman | video/image | ✓ | ✓ | ✓ | skeleton |

*"Motion" refers to a temporal consecutive sequence of poses.*

interaction, augmented reality, and autonomous systems [1, 18, 34, 35]. Several tasks such as 3D pose estimation [18, 34, 35], motion prediction [5, 8, 29, 31], and in-betweening [11, 12, 15, 37] form the core of this field. Traditionally, these tasks have been tackled by highly specialized models, each optimized for a single objective. While this focused approach has yielded significant success, it has led to a fragmented and inflexible ecosystem.

The recent paradigm shift driven by Multi-modal Large Language Models (MLLMs) offers a compelling path toward unification. However, a fundamental disconnect persists, cleaving the field into two distinct camps. On one side, Multi-modal Large Language Models (MLLMs) like MotionLLM [4] and LLaVA-Pose [33] excel at understanding motion. They can reason about poses from raw video and then generate textual descriptions, but lack the capacity to generate new, plausible human poses. On the other hand, specialized generative models like MotionGPT [14] excel at generating motion from text but struggle to process or perceive raw visual inputs. This "read-only" vs. "write-only" dichotomy forces researchers to use disparate architectures for tasks that are inherently interconnected. Moreover, constraint on input modality is also a limitation. Many generative methods operate only on abstract text or single images (*e.g.,* ChatPose [10], PoseLLaVA [9], UniPose [16]), ignoring the rich temporal dynamics of video. Conversely, video-processing models such as SKI Models [23] are typically confined to understanding, not perception, i.e., generating textual results instead of human poses. This fragmentation not only hinders research efficiency but also prevents models from leveraging the synergistic relationship between perception and generation. This leaves a critical gap: the absence of a single, unified model that can fluidly process *multi-modal inputs*, from *full videos* for pose estimation to skeleton sequences for prediction, and generate coherent, structured 3D motion as output.

To unify perception with generation, we conceptualize motion as a universal language. Our core idea is to train a single MLLM to serve as a unified motion processor. The key to achieving this is a cross-modal motion vocabulary. Existing methods, however, build this vocabulary only from skeletal data, severing the link to the visual domain.

To address this, we introduce **Superman**, a unified, multi-task, and multi-modal generative framework that reformulates these disparate challenges as a conditional sequence generation problem. Our approach has two key advantages: 1) it unifies perception and generation in a single model, promoting knowledge sharing between tasks; 2) it is flexible with multi-modal inputs, processing video, sequential skeletons, and text within the same architecture. Our cornerstone is the Vision-Guided Motion Tokenizer, a VQ-VAE [26] architecture building a universal, discrete pose vocabulary. Unlike prior works limited to skeleton data, our tokenizer introduces a **hybrid codebook** for a richer, cross-modal representation. Each token comprises paired visual and geometric prototypes. This structure guides quantization using both visual appearance from frames and 3D skeleton geometry, creating a powerful representation that intrinsically links visual evidence to motion semantics.

Grounded in this motion language, our MLLM flexibly processes diverse inputs within a single architecture. It can: **(1) Estimate 3D Pose** by translating a video into a sequence of pose tokens. **(2) Predict Motion** by auto-regressively completing a pose sequence. **(3) Perform In-betweening** by generating intermediate tokens between keyframes. As shown in Fig. 1, our unified design is validated by extensive experiments on standard benchmarks including Human3.6M [13] and 3PDW [27]. Superman achieves state-of-the-art performances across all tasks, even when compared with traditional multi-task models specializing in perception tasks. For instance, for 3D pose estimation on Human3.6M [13], Superman achieves an 11.97% improvement compared with state-of-the-art multi-task perception method [19], and a 10.91% improvement compared with state-of-the-art LLM/MLLM-based method [9]. Additionally, Superman delivers strong generalization performance. Trained exclusively on Human3.6M, Superman generalizes well to unseen data, namely, 3DPW [27], outperforming existing methods. Our contributions are threefold:

- We propose a unified generative framework that leverages a single MLLM for multi-task, multi-modal human motion analysis, bridging the gap between motion perception and generation.
- We introduce a novel Vision-Guided Motion Tokenizer that creates a robust, cross-modal pose vocabulary. By designing a hybrid codebook where each token is a bimodal entity comprising paired visual and geometric prototypes, we ensure that the quantization process is simultaneously guided by both appearance features from video and the geometric structure of 3D skeletons.
- We demonstrate that our unified method achieves state-

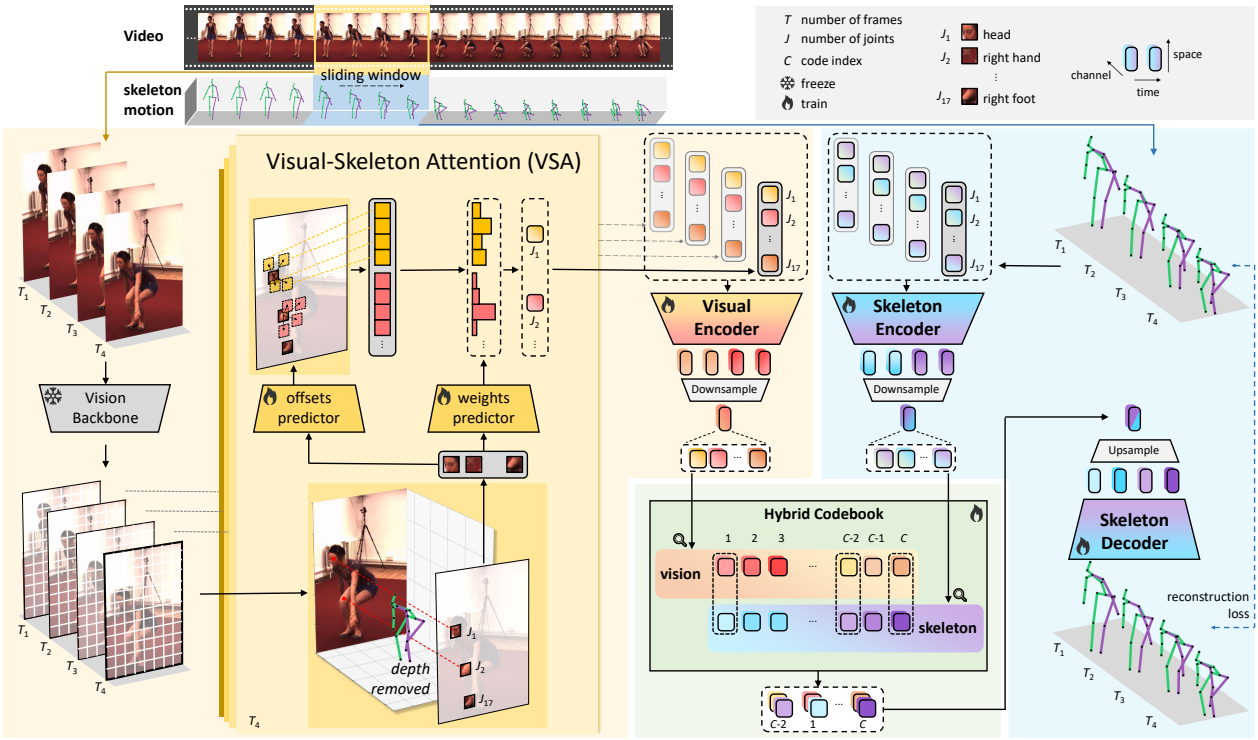



Figure 2. **Architecture of our Vision-Guided Motion Tokenizer (VGMT).** VGMT creates discrete motion vocabulary by jointly fusing information from two modalities. A Skeleton Encoder ($E_s$) captures geometry while a Visual-Skeleton Attention (VSA) module and a subsequent Visual Encoder ($E_v$) ground the pose in visual features. The fused representation is quantized against a learnable hybrid codebook, and a decoder reconstructs the 3D poses.

of-the-art across diverse tasks, validating the efficacy and scalability of our approach.

## 2. Related Work

**Traditional Human Perception Models.** Traditional models have evolved from classical methods to deep learning approaches using CNNs [22, 25], GCNs [21, 29, 31], and Transformers [36]. However, traditional multi-task methods [8, 19, 20, 30, 39] often are limited in the scope of tasks, often focusing on only perception and not generation capabilities. In contrast, our work integrates multiple human perception and generation tasks into a unified generative framework based on MLLMs.

**Human Motion as a Language.** A groundbreaking shift in motion analysis involves representing continuous human motion as a sequence of discrete tokens, a paradigm pioneered by early works in gesture synthesis [14]. A common technique for this is using a Vector Quantized Variational Autoencoder (VQ-VAE) to learn a "codebook" of atomic poses [7]. As summarized in Tab. 1, recent models have framed tasks like motion generation and prediction as language modeling problems [14, 35]. However, their scope is often limited to generating from non-visual modalities, failing to ground the motion language in raw visual input. Our work builds directly upon this paradigm but makes a crucial extension by integrating vision into the tokenization process, creating a visually grounded motion language.

**Multi-modal Large Language Models for Pose Perception and Generation.** The remarkable success of Multimodal Large Language Models (MLLMs), which integrate powerful vision encoders with LLMs for complex visual reasoning [6, 17, 32], has inspired their application to human-centric tasks. In the context of human motion, models such as MotionLLM [4] and LLaVA-Pose [33] have demonstrated the ability to comprehend and answer questions about human actions and poses depicted in images and videos. They effectively connect the visual and semantic aspects of human motion. However, a significant limitation of these models is their primary focus on perception. They can interpret and describe poses, but they are not architecturally designed to generate new, structured motion sequences. Recent works, such as ChatPose [10] and UniPose [16], have begun to bridge this gap by enabling pose generation conditioned on both visual and textual inputs. These models, however, typically operate on static, single-frame images. Our work distinguishes itself by being a fully versatile framework that not only perceives motion from videos but is also a powerful generative model capable of multiple synthesis tasks, thereby unifying the reading and writing of motion within a single, coherent architecture.

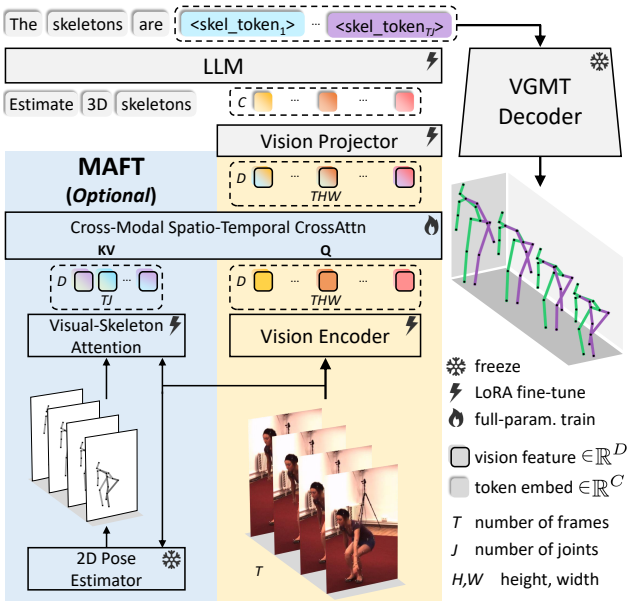



Figure 3. **Network architecture and training paradigm.** Superman fine-tune a single LLM to integrate information from text, video, and 3D skeleton modalities. Optionally, a **Motion-Aware Fine-Tuning (MAFT)** module can be integrated. With <0.2% extra parameters, MAFT enhances motion perception by enabling cross-video-motion fusion, leading to substantial improvement on tasks with visual input, as validated by the experiment results.

## 3. Method

Our method unifies multi-task human motion analysis within a conditional sequence generation framework. It involves two key stages: **(1) a novel Vision-Guided Motion Tokenizer (Section 3.1)** that converts continuous, high-dimensional motion into a discrete sequence of semantic tokens, and **(2) a Multi-modal Large Language Model (MLLM) (Section 3.2)** that autoregressively models these token sequences to perform our diverse set of tasks. The overall network architecture is depicted in Fig. 3.

### 3.1. Vision-Guided Motion Tokenizer

The cornerstone of our approach is a discrete vocabulary for continuous human motion. Unlike prior tokenizers that rely on skeletons alone [14], we design a Vision-Guided Motion Tokenizer, built upon a VQ-VAE, which is explicitly trained to fuse *visual appearance* and *3D skeletal geometry*. This ensures each token is grounded in both modalities. The architecture is illustrated in Fig. 2.

**Dual-Stream Encoding.** Given a video clip $\mathcal{V} \in \mathbb{R}^{F\times H\times W\times 3}$ and its corresponding 3D pose sequence $\mathbf{X} \in \mathbb{R}^{F\times N\times 3}$, our tokenizer employs two parallel streams.

*1) Visual Encoding Stream:* This stream extracts pose-centric visual features. We first use a vision backbone (e.g., HRNet [24]) to obtain feature maps $\mathcal{F}_f$ for each frame. For each joint $j$ at frame $f$, its projected 2D location $\mathbf{p}_{j,f}$ serves as a reference point to sample an initial query feature $\mathbf{q}_{j,f}$ from $\mathcal{F}_f$. To enhance feature representation and robustness to occlusions, we introduce a Visual-Skeleton Attention (VSA) module. The VSA adaptively aggregates features by predicting sampling offsets and aggregation weights from the query. The visual feature for the joint is computed as:

$$\mathbf{v}_{j,f} = \mathrm{VSA}(\mathbf{q}_{j,f}, \mathcal{F}_f, \mathbf{p}_{j,f}), \tag{1}$$

yielding a sequence of contextually-aware visual features $\mathbf{V} \in \mathbb{R}^{F\times N\times D}$.

*2) Skeletal Encoding Stream:* This stream captures the motion's intrinsic geometry. The input pose sequence $\mathbf{X}$ is passed through a series of lightweight 2D convolutions operating on the joint-temporal grid. This effectively models spatio-temporal kinematics and maps the coordinates into a latent space, producing skeletal features $\mathbf{S} \in \mathbb{R}^{F\times N\times D}$.

**Hybrid Codebook Quantization.** To ensure each token represents a short motion pattern rather than a static pose, we first temporally downsample the feature sequences $\mathbf{V}$ and $\mathbf{S}$, yielding window-level representations $\mathbf{z}^v_w$ and $\mathbf{z}^s_w$ for each time window $w$. We then introduce a hybrid codebook that pairs visual and geometric prototypes: $\mathcal{C} = \{(\mathbf{c}^v_k, \mathbf{c}^s_k)\}_{k=1}^K$, where $\mathbf{c}^v_k, \mathbf{c}^s_k \in \mathbb{R}^D$. For each window $w$, $k_w$ is the token index that minimizes the joint distance:

$$k_w = \arg\min_k \left( \|\mathbf{z}^v_w - \mathbf{c}^v_k\|_2^2 + \|\mathbf{z}^s_w - \mathbf{c}^s_k\|_2^2 \right). \tag{2}$$

This hybrid selection, based on minimizing the joint Euclidean distance, grounds the discretization process in both vision and geometry. The selected code vectors for the window are then $\hat{\mathbf{c}}^v_w = \mathbf{c}^v_{k_w}$ and $\hat{\mathbf{c}}^s_w = \mathbf{c}^s_{k_w}$.

**Reconstruction and Training.** The skeletal code $\hat{\mathbf{c}}^s_w$ is upsampled and passed to a decoder to reconstruct the 3D poses $\hat{\mathbf{X}}_w$. The tokenizer is trained end-to-end with a VQ objective that combines reconstruction with modality-wise commitment losses:

$$\mathcal{L}_{\text{VQ}} = \|\mathbf{X}_w - \hat{\mathbf{X}}_w\|_2^2 + \beta_s \|\mathrm{sg}[\mathbf{z}^s_w] - \hat{\mathbf{c}}^s_w\|_2^2 + \beta_v \|\mathrm{sg}[\mathbf{z}^v_w] - \hat{\mathbf{c}}^v_w\|_2^2, \tag{3}$$

where $\mathrm{sg}[\cdot]$ is the stop-gradient operator, and we set the commitment weights $\beta_s = 0.5$ and $\beta_v = 0.5$. This process effectively converts a continuous motion clip of $F$ frames into a discrete sequence of $T$ integer tokens $\mathbf{K}_{1:T}$. After training, the tokenizer provides a robust, visually-grounded motion vocabulary.

### 3.2. Unified Multi-Task Modeling with LLM

Our discrete motion tokenizer enables us to frame diverse motion tasks as a unified conditional sequence generation problem. We employ a decoder-only Multi-modal Large Language Model (MLLM), Qwen2.5-VL-7B [2], as our central sequence processor, which learns to predict motion tokens autoregressively.

**Connecting Tokenizer to Qwen-VL.** Besides the standard tokenizer-MLLM integration paradigm [14, 16], we additionally design an optional light-weight module to inject

Table 2. **Comparison of our model with traditional multi-task models and LLM / MLLM-based models** on three human motion tasks: pose estimation (PE), motion prediction (MP), and motion in-betweening (MIB) on Human3.6M [13]. All tasks are evaluated using Mean Per Joint Position Error (MPJPE) in millimeter, averaged over all test data, where lower is better. "$T$" means how many frames the model inputs and outputs. ("$T$=1" means the video has to be processed image-by-image.) "N/A" means the model is not applicable to the task. For example, single-frame-only models [9, 16, 28] cannot do MP or MIB; LLM-based models [14, 38] cannot do PE.

| Models | Venue | $T$ | PE‡ | | MP† | | | | MIB† | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | N-MPJPE | MPJPE | Avg | 80ms | 160ms | 320ms | Avg | mid | last |
| *Traditional Multi-Task Models* | | | | | | | | | | | |
| MotionBERT [39] | ICCV'23 | 16 | 47.07 | 56.70 | 29.94 | 18.67 | 26.82 | 50.33 | 42.37 | 44.86 | 53.16 |
| Skeleton-in-Context [30] | CVPR'24 | 16 | 45.26 | 55.57 | 33.42 | 21.09 | 30.48 | 54.62 | 31.27 | 36.66 | 36.84 |
| Human-in-Context [19] | Arxiv'25 | 16 | 44.77 | 53.86 | 26.66 | 14.36 | 23.58 | 47.34 | 31.13 | 37.01 | 35.49 |
| **Superman (w/ MAFT)** | Ours'26 | 16 | **39.41** | **51.61** | **26.13**† | **13.70** | **23.30** | **44.90** | **30.61**† | **35.99** | **35.13** |
| *LLM / MLLM-based Models* | | | | | | | | | | | |
| LocLLM [28] | CVPR'24 | 1 | 49.32 | 62.19 | N/A | | | | N/A | | |
| UniPose [16] | CVPR'25 | 1 | 65.24 | 106.96 | N/A | | | | N/A | | |
| PoseLLaVA [9] | AAAI'25 | 1 | 50.40 | 62.43 | N/A | | | | N/A | | |
| MotionGPT [14] | NeurIPS'23 | 16 | N/A | | 48.81 | 42.99 | 49.27 | 55.31 | 52.96 | 56.07 | 66.45 |
| MotionGPT3 [38] | ICLR'26 | 16 | N/A | | 42.30 | 37.25 | 42.70 | 47.94 | 47.66 | 50.47 | 59.81 |
| **Superman** | Ours'26 | 16 | **44.90** | **61.39** | **26.13**† | **13.70** | **23.30** | **44.90** | **30.61**† | **35.99** | **35.13** |

‡ *Traditional models adopt "RGB→2D→3D" paradigm, leveraging a fixed 2D pose estimator; LLM/MLLM-based models adopt "RGB→3D" paradigm.*
† *For MP and MIB, all models take only 3D poses as input. Therefore, our MAFT module (which enhances vision) is deactivated, hence the identical results for Superman and Superman (w/ MAFT) on these two tasks.*

skeletal geometry into the MLLM's visual stream, Motion-Aware Fine-tuning (MAFT). First, we enhance the ViT features using a Visual-Skeleton Attention (VSA) block. The ViT backbone extracts patch features $\mathbf{Z}_{\text{grid}}$. Concurrently, 2D joint projections serve as reference points for a multi-scale deformable sampler [40] to aggregate pose-centric features $\mathbf{Z}_{\text{pose}}$. We then fuse this information using VSA, implemented as a cross-attention layer followed by an FFN, where grid tokens are queries and pose tokens are keys/values. The visual tokens $\hat{\mathbf{Z}}_{\text{grid}}$ are computed as:

$$\hat{\mathbf{Z}}_{\text{grid}} = \text{VSA}(\mathbf{Z}_{\text{grid}}, \mathbf{Z}_{\text{pose}}). \tag{4}$$

These enhanced visual tokens, along with any textual prompts, are fed into the LLM decoder, which generates the final sequence of motion tokens.

Our framework handles multiple tasks by varying the conditioning inputs:

**3D Pose Estimation from Video:** Given a video clip $\mathbf{I}_{1:F}$, the MLLM is conditioned on its enhanced visual features $\hat{\mathbf{Z}}_{\text{grid}}$ to generate the corresponding motion token sequence $\mathbf{K}_{1:T}$. The objective is to maximize the log-likelihood of the ground-truth sequence:

$$\mathcal{L}_{\text{est}} = \sum_{t=1}^{T} \log P(k_t | \mathbf{K}_{<t}, \hat{\mathbf{Z}}_{\text{grid}}). \tag{5}$$

The tokens are then decoded to obtain the final 3D poses.

**Motion Prediction:** Given a historical pose sequence, tokenized as $\mathbf{K}_{1:T'}$, the task is to predict the future. The LLM autoregressively generates subsequent tokens $\mathbf{K}_{T'+1:T}$ conditioned on the past:

$$\mathcal{L}_{\text{pred}} = \sum_{t=T'+1}^{T} \log P(k_t | \mathbf{K}_{<t}). \tag{6}$$

**Motion In-betweening:** Given start and end tokens ($k_1$, $k_T$), we use a prompt like "[START] $k_1$ [MIDDLE] $k_T$ [END]", training the LLM to fill in the missing sequence corresponding to the [MIDDLE] segment.

Our framework's strength is a single LLM trained jointly on a mixture of task-specific formats. This multi-task strategy fosters a rich, unified representation, allowing it to act as a versatile motion processor.

**Training.** Our framework is trained in two stages. First, the Vision-Guided Motion Tokenizer is independently trained, and its weights are frozen. Subsequently, the MLLM is trained on a mixture of all tasks using a single, unified objective: the standard autoregressive cross-entropy loss, summed across all task-specific samples in a batch. This joint strategy facilitates knowledge transfer between perception and generation.

Table 3. **Action-specific N-MPJPEs** for pose estimation on Human3.6M [13].

| Models | Venue | $T$ | Actions | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Sit | SitDown | Smoke | Photo | Wait | Walk | WalkDog | WalkTwo |
| *Traditional Multi-Task Models* | | | | | | | | | | |
| MotionBERT [39] | ICCV'23 | 16 | 62.22 | 69.81 | 49.21 | 49.66 | 39.49 | 40.80 | 44.94 | 40.95 |
| Skeleton-in-Context [30] | CVPR'24 | 16 | 59.34 | 65.89 | 48.70 | 48.11 | 37.28 | 38.52 | 42.12 | 38.97 |
| Human-in-Context [19] | Arxiv'25 | 16 | 59.95 | 65.21 | 47.81 | 47.02 | 36.37 | 38.17 | 41.76 | 37.31 |
| **Superman (w/ MAFT)** | Ours'26 | 16 | **47.12** | **48.97** | **41.31** | **34.37** | 41.21 | **36.12** | **39.23** | 43.23 |
| *LLM / MLLM-based Models* | | | | | | | | | | |
| LocLLM [28] | CVPR'24 | 1 | 64.48 | 68.76 | 50.96 | 55.16 | 42.55 | 39.11 | 49.96 | 43.10 |
| UniPose [16] | CVPR'25 | 1 | 72.30 | 84.24 | 69.20 | 62.65 | 60.02 | 66.97 | 65.71 | 70.69 |
| PoseLLaVA [9] | AAAI'25 | 1 | 64.64 | 71.36 | 52.67 | 54.21 | 44.67 | 46.90 | 48.90 | 44.90 |
| **Superman** | Ours'26 | 16 | **62.68** | 71.93 | **49.09** | **43.54** | **35.21** | **30.89** | **45.91** | **39.60** |

Table 4. **Comparison of results on generalization to unseen data.** Motion prediction (MP) and motion in-betweening (MIB) on 3DPW [27] are reported. All models are trained on Human3.6M only, with 3DPW completely excluded from training.

| Models | MP | | | MIB | |
|---|---|---|---|---|---|
| | Avg | 80ms | 320ms | Avg | mid |
| *Traditional Multi-Task Models* | | | | | |
| MotionBERT [39] | 164.96 | 137.89 | 200.14 | 123.05 | 148.51 |
| Skeleton-in-Context [30] | 140.71 | 110.77 | 183.74 | 103.97 | 127.68 |
| Human-in-Context [19] | 141.90 | 112.53 | 183.62 | 108.54 | 131.99 |
| *LLM / MLLM-based Models* | | | | | |
| MotionGPT [14] | 292.87 | 257.92 | 331.88 | 234.39 | 257.55 |
| MotionGPT3 [38] | 228.17 | 190.83 | 259.49 | 180.10 | 204.71 |
| **Superman (Ours)** | **62.05** | **34.75** | **97.37** | **60.68** | **63.35** |

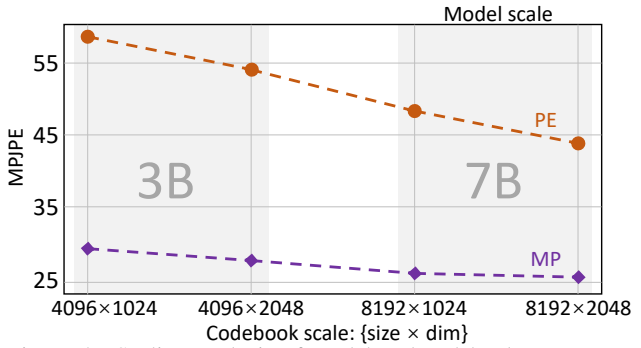



Figure 4. Scaling analysis of model and codebook parameters. Scaling up the model (3B→7B) and the codebook consistently reduces (N-)MPJPE for both pose estimation (PE) and motion prediction (MP) tasks, demonstrating the method's ability to leverage larger capacity for improved accuracy.

## 4. Experiment

**Implementation Details.** Our framework is built upon Qwen2.5-VL-7B [2] as the base MLLM. We employ LoRA and, optionally, MAFT. The trainable parameters accounts for 12.67% (w/o MAFT) and 12.87% (w/ MAFT) of vanilla Qwen2.5-VL-7B. Our Vision-Guided Motion Tokenizer has a vocabulary size of $K = 8192$. The model is trained on 2 NVIDIA H20 GPUs.

**Datasets and Metrics.** For evaluation, we utilize two large-scale 3D human motion datasets: Human3.6M [13] and 3DPW [27]. The models are trained on Human3.6M only, and tested on both datasets. The training set and the test set are in line with their respective splitting standards. Tasks are evaluated using Mean Per Joint Position Error (MPJPE) as the main metric, in line with standard protocols.

**Data Preprocessing.** To unify formats, we first convert 3DPW's SMPL vertices to the Human3.6M skeleton format using a pre-trained matrix [3]. Then, both datasets follow the same processing pipeline. Specifically, for the X and Y spatial dimensions, we transform the 3D camera coordinates (in millimeters) into 2D image coordinates (in pixels), following [39]. For the depth (Z) dimension, we normalize the skeleton by setting the depth of the root joint (pelvis) to 0. The depths of all other joints are then converted to be relative to this root. Crucially, we do not translate the root joint to the coordinate origin (0, 0, 0) during either the training or testing phases. This approach retains the global positional information in the image coordinate space.

### 4.1. Main Results

Our unified model, Superman, was evaluated on three human motion tasks: 3D Pose Estimation (PE), Motion Prediction (MP), and Motion In-Betweening (MIB). As shown in Tabs. 2, 3, and 4, our model demonstrates state-of-the-art (SOTA) performance on both the standard benchmark (Human3.6M [13]) and generalization tests (3DPW [27]).

**Performance on Human3.6M.** In the comprehensive evaluation (Tab. 2), our model surpassed all baselines under both paradigms. Compared to traditional multi-task models (which use a "RGB→2D→3D" paradigm), Superman achieved the lowest errors of all models. Compared to LLM/MLLM-based models (which use a "RGB→3D" paradigm), Superman also secured the best performance with the lowest MPJPE/N-MPJPE. In the action-specific N-MPJPE breakdown (Tab. 3), Superman achieved the lowest error in most of the action categories.

**Generalization on 3DPW.** To evaluate the model's abil-

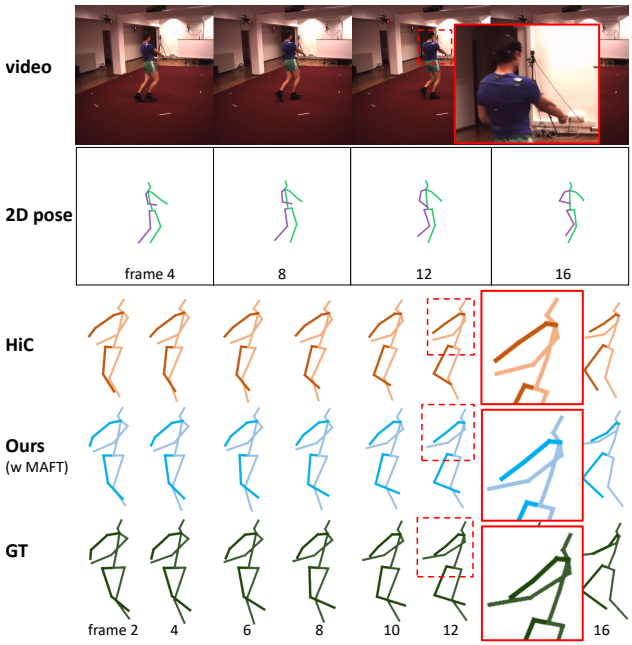



Figure 5. Qualitative results for pose estimation on Human3.6M. Our method with MAFT is compared with HiC [19], the current SoTA. Both methods use video and 2D poses as inputs and leverages a fixed 2D pose estimator.

ity to generalize to unseen data, we conduct a zero-shot test on the 3DPW dataset; all models are trained only on Human3.6M and tested on 3DPW. As shown in Tab. 4, Superman demonstrates exceptional generalization on both motion tasks, outperforming all baseline models by a significant margin. Fig. 6 provides qualitative validation for this, showing our model generating more accurate future poses than the current SoTA baseline [30]. On motion inbetweening (MIB), our model achieved an average error of 60.68, again far surpassing the best traditional model and other LLM/MLLM models. These results confirm that our model not only excels on the benchmark dataset but also possesses superior robustness and generalization capabilities when handling unseen "in-the-wild" data.

**Computational Efficiency.** Tab. 5 breaks down the efficiency metrics for both VGMT and MLLM components. Specifically, it details GFLOPs, parameter counts, and latency for the vision backbone, vision encoder, and VSA. While the large-scale Qwen2.5-VL-7B dominates overall resource consumption, accounting for 99.7% of the total GFLOPs and 98.6% of the total latency, our newly introduced components, such as MAFT and VSA, exhibit remarkably low overhead. Collectively, these proposed modules account for less than 0.03% of the total computation cost, demonstrating the exceptional efficiency and practical scalability of our architecture.

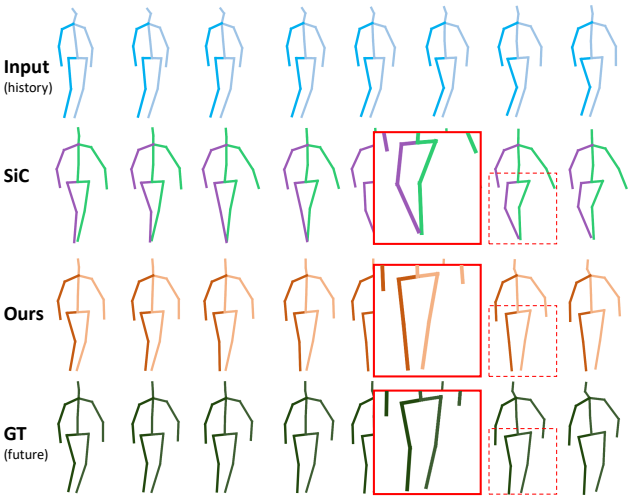



Figure 6. Qualitative results for generalizing to motion prediction on 3DPW (unseen dataset). Our method is compared with SiC [19], the current SoTA on the task.

## 4.2. Visualization

**Analysis of the VQ-VAE Codebook.** To evaluate the efficiency of our codebook utilization, we conducted a quantitative analysis. As shown in Fig. 7 (a), the codes are categorized into four usage frequency levels: Frequent (usage rate $>1\%$), Active (0.01%-1%), Underused (0-0.01%), and Unused (never used). On the Human3.6M dataset, up to 65.4% of the codes remain Active, indicating efficient utilization of the encoding space. This high efficiency is maintained even on the unseen 3DPW dataset, which shows 51.7% Active codes and only 1.0% Unused codes. Furthermore, we analyzed the codebook's latent space by computing pairwise cosine similarities. The distribution in Fig. 7 (b) shows that the similarity of most code pairs is close to 0, confirming the model's ability to learn a discrete latent space with highly decorrelated representations.

**Qualitative Results for 3D Pose Estimation.** Figure 5 presents qualitative visualizations for 3D pose estimation on Human3.6M. The figure provides an intuitive comparison of our model ("Ours" shown in blue) against the baseline method HiC [19] ("HiC" shown in orange) and the Ground Truth ("GT" shown in dark green). Visually, the 3D skeleton sequence generated by our method shows a high degree of fidelity to the Ground Truth in both shape and motion, while the baseline method exhibits noticeable visual discrepancies.

## 4.3. Ablation Study and Analysis

We conduct comprehensive ablation studies to validate the effectiveness of our framework's key design choices.

**Analysis of Model and Codebook Scaling.** To validate the scalability of our framework, we investigate the impact

Table 5. Computation efficiency.

| Module | GFLOPs | params (M) | latency (ms) |
|---|---|---|---|
| *Vision-Guided Motion Tokenizer (VGMT)* | | | |
| vision backbone 🔒 | 31.375 | 28.53 | 43.7 |
| vision encoder | 2.468 | 17.12 | 2.7 |
| VSA | 0.001 | 0.11 | 0.1 |
| *MLLM* | | | |
| MAFT | 5.01 | 25.45 | 0.3 |
| Qwen2.5-VL-7B | ∼129920.00 | ∼9605.00 | ∼3300.0 |

*Vision backbone remains frozen during training.*

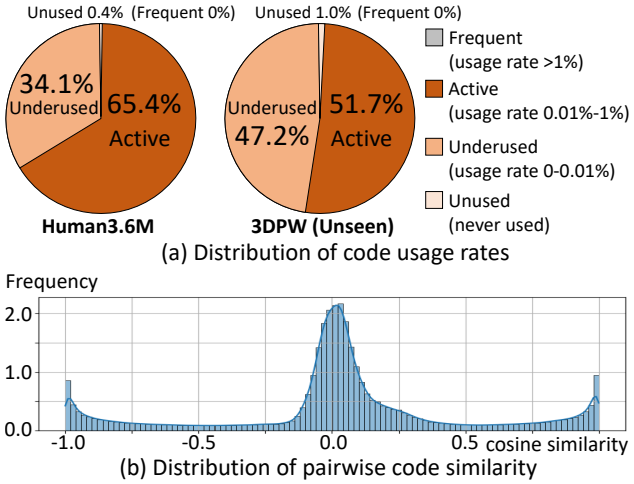



Figure 7. Quantitative analysis of the VQ-VAE codebook. (a) Up to 65.4% of the codes remain active during inference, indicating efficient usage of the encoding space. (b) The cosine similarity of most code pairs is close to 0, confirming the ability to learn a discrete latent space with highly decorrelated representations.

of both the MLLM's parameter size and the codebook's capacity. As illustrated in Fig. 4, we evaluate performance (MPJPE) by scaling the model from 3B to 7B parameters, while simultaneously increasing the codebook scale (in terms of both size and dimension). The results show a clear and consistent trend: (1) Increasing the model size from 3B to 7B significantly reduces the error for both Pose Estimation (PE) and Motion Prediction (MP). (2) Within a given model size, increasing the codebook capacity (e.g., from $4096 \times 1024$ to $4096 \times 2048$ for the 3B model) also consistently lowers the MPJPE. This analysis confirms that our unified framework benefits from larger model and codebook parameters, demonstrating strong scalability and potential for future improvements.

**Effectiveness of the Vision-Guided Tokenizer.** We validate our tokenizer design in Table 6. Results show that jointly learning from vision and geometry is essential. The Visual-Only model fails (22.5 mm error), and while the Skeleton-Only baseline [14] is reasonable (7.7 mm), it is significantly outperformed by all fused configurations. Among these, our chosen balanced weighting ($\beta_s = 0.5, \beta_v = 0.5$) achieves the lowest reconstruction error (4.7 mm), proving superior to other weightings (5.3-6.8 mm) and validating our approach.

Table 6. Ablation on tokenizer components and fusion weights.

| Tokenizer Config. | Recon. Err. | PE (N-MPJPE) |
|---|---|---|
| Visual-Only | 22.5 | 51.3 |
| Skeleton-Only | 7.7 | 47.8 |
| Fused: $\beta_s, \beta_v = 0.3, 0.7$ | 6.8 | 45.8 |
| Fused: $\beta_s, \beta_v = 0.7, 0.3$ | 5.3 | 45.1 |
| Fused: $\beta_s, \beta_v = 0.5, 0.5$ | **4.7** | **44.9** |

Table 7. Ablation on the unified multi-task training strategy. We compare our single unified model against specialized models.

| **Training Strategy** | PE | MP | MIB |
|---|---|---|---|
| Specialized (Est. Only) | 46.5 | N/A | N/A |
| Specialized (Pred. Only) | N/A | 27.3 | N/A |
| Specialized (MiB. Only) | N/A | N/A | 33.1 |
| Unified | **44.9** | **26.1** | **30.6** |

**Effectiveness of Unified Multi-Task Training.** Effectiveness of Unified Multi-Task Training. Our second key design is the use of a single MLLM for joint multi-task training (as described in Sec. 3.2). To validate this, we compare our Unified model against Specialized models, which use the same architecture but are trained exclusively on a single task. As shown in Tab. 7, our Unified model consistently outperforms its Specialized counterparts across all three tasks. The unified model achieves a PE of 44.9 (vs. 46.5), an MP of 26.1 (vs. 27.3), and an MiB of 30.6 (vs. 33.1). This strongly indicates positive knowledge transfer between tasks, as learning them jointly improves performance in all areas. This confirms our unified framework is not only efficient but also more effective by learning a richer, joint representation.

## 5. Conclusion

We propose a unified generative framework that bridges the gap between motion perception and generation. The core of our approach is a novel Vision-Guided Motion Tokenizer, which, unlike prior work, learns a robust cross-modal vocabulary by jointly fusing visual features and geometric data. Grounded in this motion language, our single MLLM architecture achieves state-of-the-art or competitive performance across diverse tasks, including 3D pose estimation, motion prediction, and motion in-betweening. This work demonstrates that a unified, visually grounded generative model is an efficient, scalable paradigm for holistic human motion analysis.

## 6. Acknowledgement

This work was supported by National Natural Science Foundation of China (No. 62473007), Guangdong Outstanding Youth Fund (No. 2026B1515020015), Shenzhen Innovation in Science and Technology Foundation for The Excellent Youth Scholars (No. RCYX20231211090248064).

## References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: a benchmark for human pose estimation and tracking. In *CVPR*, 2018. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 4, 6

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 6

[4] Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. Motionllm: Understanding human behaviors from human motions and videos. *arXiv preprint arXiv:2405.20340*, 2024. 2, 3

[5] Qiongjie Cui and Huaijiang Sun. Towards accurate 3d human motion prediction from incomplete observations. In *CVPR*, 2021. 2

[6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. 2023. 3

[7] Yanbo Ding, Xirui Hu, Zhizhi Guo, Chi Zhang, and Yali Wang. Mtvcrafter: 4d motion tokenization for open-world human image animation. *arXiv preprint arXiv:2505.10238*, 2025. 3

[8] Zhongbin Fang, Xiangtai Li, Xia Li, Joachim M Buhmann, Chen Change Loy, and Mengyuan Liu. Explore in-context learning for 3d point cloud understanding. 2023. 2, 3

[9] Dong Feng, Ping Guo, Encheng Peng, Mingmin Zhu, Wenhao Yu, and Peng Wang. Posellava: Pose centric multimodal llm for fine-grained 3d pose manipulation. In *AAAI*, 2025. 2, 5, 6

[10] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. Chatpose: Chatting about 3d human pose. In *CVPR*, 2024. 2, 3

[11] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM TOG*, 2020. 2

[12] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *ICCV*, 2019. 2

[13] 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 T-PAMI*, 2013. 2, 5, 6

[14] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. 2023. 2, 3, 4, 5, 6, 8

[15] Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and Otmar Hilliges. Convolutional autoencoders for human motion infilling. In *3DV*, 2020. 2

[16] Yiheng Li, Ruibing Hou, Hong Chang, Shiguang Shan, and Xilin Chen. Unipose: A unified multimodal framework for human pose comprehension, generation and editing. In *CVPR*, 2025. 2, 3, 4, 5, 6

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 36:34892–34916, 2023. 3

[18] Mengyuan Liu and Junsong Yuan. Recognizing human actions as the evolution of pose estimation maps. In *CVPR*, 2018. 2

[19] Mengyuan Liu, Xinshun Wang, Zhongbin Fang, Deheng Ye, Xia Li, Tao Tang, Songtao Wu, Xiangtai Li, and Ming-Hsuan Yang. Human-in-context: Unified cross-domain 3d human motion modeling via in-context learning. *arXiv preprint arXiv:2508.10897*, 2025. 2, 3, 5, 6, 7

[20] Mengyuan Liu, Zhongbin Fang, Xia Li, Joachim M Buhmann, Deheng Ye, Xiangtai Li, and Chen Change Loy. Point-in-context: Understanding point cloud via in-context learning. *IJCV*, 2026. 3

[21] Julieta Martinez, Rayat Hossain, Javier Romero, and J. Little. A simple yet effective baseline for 3d human pose estimation. 2017. 3

[22] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. 2018. 3

[23] Arkaprava Sinha, Dominick Reilly, Francois Bremond, Pu Wang, and Srijan Das. Ski models: Skeleton induced vision-language embeddings for understanding activities of daily living. In *AAAI*, 2025. 2

[24] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 4

[25] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. 2013. 3

[26] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. 30, 2017. 2

[27] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 2, 6

[28] Dongkai Wang, Shiyu Xuan, and Shiliang Zhang. Locllm: Exploiting generalizable human keypoint localization via large language model. In *CVPR*, 2024. 2, 5, 6

[29] Xinshun Wang, Qiongjie Cui, Chen Chen, and Mengyuan Liu. Gcnext: towards the unity of graph convolutions for human motion prediction. In *AAAI*, 2024. 2, 3

[30] Xinshun Wang, Zhongbin Fang, Xia Li, Xiangtai Li, Chen Chen, and Mengyuan Liu. Skeleton-in-context: unified skeleton sequence modeling with in-context learning. In *CVPR*, 2024. 2, 3, 5, 6, 7

[31] Xinshun Wang, Wanying Zhang, Can Wang, Yuan Gao, and Mengyuan Liu. Dynamic dense graph convolutional network

for skeleton-based human motion prediction. *IEEE T-IP*, 2024. 2, 3

[32] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 3

[33] Dewen Zhang, Tahir Hussain, Wangpeng An, and Hayaru Shouno. Llava-pose: Enhancing human pose and action understanding via keypoint-integrated instruction tuning. *arXiv preprint arXiv:2506.21317*, 2025. 2, 3

[34] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *CVPR*, 2020. 2

[35] Xinyi Zhang, Qiqi Bao, Qinpeng Cui, Wenming Yang, and Qingmin Liao. Pose magic: efficient and temporally consistent human pose estimation with a hybrid mamba-gcn network. *arXiv preprint arXiv:2408.02922*, 2024. 2, 3

[36] Ce Zheng, Sijie Zhu, Mat'ias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. 2021. 3

[37] Yi Zhou, Jingwan Lu, Connelly Barnes, Jimei Yang, Sitao Xiang, et al. Generative tweening: long-term inbetweening of 3d human motions. *arXiv preprint arXiv:2005.08891*, 2020. 2

[38] Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. Motiongpt3: Human motion as a second modality. *arXiv preprint arXiv:2506.24086*, 2025. 2, 5, 6

[39] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: a unified perspective on learning human motion representations. In *ICCV*, 2023. 2, 3, 5, 6

[40] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 5