Dual Coding Theory in Action: Language-Assisted Human Pose Estimation in Videos

Sifan Wu[1,2], Haipeng Chen[1,2], Yingda Lyu[1,3*], Shaojing Fan[4*], Zhigang Wang[5], Zhenguang Liu[5,6,7*], Yingying Jiao[8*]

[1]College of Computer Science and Technology, Jilin University,
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University,
[3]Public Computer Education and Research Center, Jilin University,
[4]Department of Electrical and Computer Engineering, National University of Singapore,
[5]The State Key Laboratory of Blockchain and Data Security, Zhejiang University,
[6]Shandong Rendui Network Co., Ltd.,
[7]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security,
[8]College of Computer Science and Technology, Zhejiang University of Technology,
wusifan2021@gmail.com, chenhp@jlu.edu.cn, ydlv@jlu.edu.cn, dcsfs@nus.edu.sg,
wangzhigang2024@gmail.com, liuzhenguang2008@gmail.com, yingyingjiao21@gmail.com

Abstract

Video-based human pose estimation aims to localize keypoints across frames, enabling robust analysis of human motion in applications such as sports, surveillance, and healthcare. However, existing methods rely *solely* on visual cues, limiting their robustness in complex scenes involving occlusion, motion blur, or poor lighting. In contrast, dual coding theory from psychology suggests that human cognition is inherently multimodal: we learn by integrating visual perception with linguistic context to form structured, semantic understandings of the world. Visual input provides concrete spatiotemporal grounding, while language offers symbolic abstraction that enhances reasoning and generalization. Motivated by this cognitive principle, we present the first framework that explicitly incorporates language as an auxiliary modality to enhance video-based pose estimation. To address the lack of paired video-text datasets, we first employ a Multimodal Large Language Model (MLLM) to generate textual descriptions of human interactions from videos. We then propose a novel coarse-to-fine multimodal alignment pipeline: a cross-modal semantic interaction module establishes initial grounding between spatiotemporal visual features and textual embeddings, while an optimal transport-based feature matching mechanism enforces fine-grained, geometry-aware alignment. This cognitively inspired design enables more accurate and robust pose estimation, especially in visually challenging scenes like occlusion and motion blur. Extensive experiments on three benchmarks confirm that our method consistently outperforms state-of-the-art approaches.

Introduction

Human pose estimation has been an active research topic in artificial intelligence, focusing on localizing the anatomical positions of human joints (*e.g.,* elbows, knees). As human poses play a central role in mediating our interactions with the external world, pose estimation serves as the cornerstone for numerous human-centric applications, ranging from *interpreting pedestrian intent in autonomous driving* (Tang et al. 2023; Xu et al. 2024) to *enabling realistic virtual try-on systems* (Liu et al. 2022b; Wu et al. 2024b).

Deep learning continues to advance rapidly (Tang et al. 2025a), leading to significant achievements in modern AI applications (Tang et al. 2024, 2025b; Wang et al. 2024). Current pose estimation methods are largely unimodal, relying on either static images or video sequences. Image-based approaches have evolved from hand-crafted features (Zhang et al. 2009) to deep models such as CNNs (Sun et al. 2019), Transformers (Xu et al. 2022), but they inherently lack the ability to capture the temporal dynamics of motion. Video-based methods (Chen et al. 2025; Jiao et al. 2025) model spatiotemporal dependencies, *yet* their unimodal, visual-only design limits their robustness (Wu et al. 2025; Wang et al. 2025). As shown in Figure 1, even state-of-the-art models (Liu et al. 2021; Wu et al. 2024a) can miss key poses (dashed red box) in challenging scenes (*e.g.*, defocus, occlusion), indicating that visual input alone is often insufficient.

In contrast, according to Dual Coding Theory (DCT) (Paivio 1971, 1986), human cognition is intrinsically multimodal: we learn by integrating visual and linguistic information. In dynamic scenes, humans integrate visual motion cues with language and prior knowledge to form coherent, time-aware interpretations, which enables the disambiguation of interactions and action prediction over time (Baltrušaitis, Ahuja, and Morency 2018). Drawing inspiration from this human multimodal interplay, we propose a novel paradigm for video-based human pose estimation that explicitly incorporates language as an auxiliary modality: While visual perception provides concrete spatiotemporal data, language offers symbolic abstraction for higher-level reasoning and generalization (Subramaniam et al. 2024). For example, in visually ambiguous, defocused videos (Fig. 1), textual descriptions (*e.g.*, "**The man in the black shirt is skateboarding ..., while the man in the white shirt is also**

*Corresponding authors.

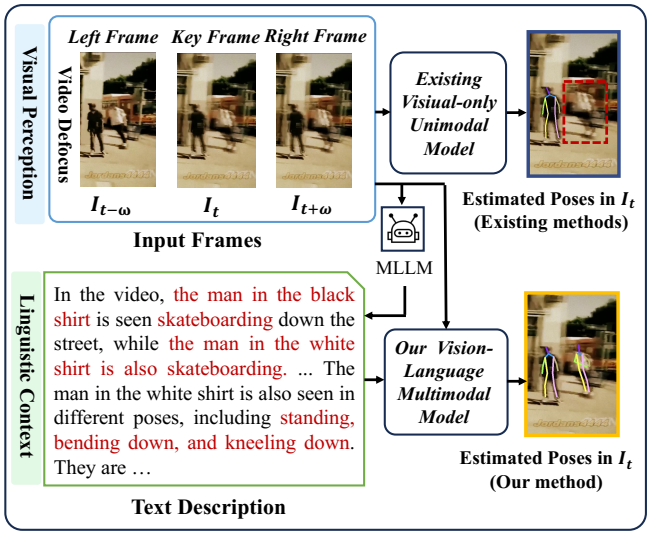

Figure 1: Our work is motivated by Dual Coding Theory (Paivio 1986), which posits that human cognition integrates visual and linguistic inputs for semantic understanding. Inspired by human multimodal reasoning, we propose the first framework incorporating language as an auxiliary modality to enhance video pose estimation. We use MLLM-generated text to overcome the failures of visual-only models (Liu et al. 2021; Wu et al. 2024a) in ambiguous scenes like video defocus (dashed red box).

skateboarding ...") supply crucial semantic anchors. This linguistic context disambiguates identities, actions, and subject count, which are often degraded by visual noise. We argue that this synergy between seeing and linguistic understanding is key to robust video-based pose estimation in challenging scenes, such as pose occlusion and motion blur.

Despite its potential, this language-assisted multimodal paradigm for video-based human pose estimation presents three key challenges: (i) **Lack of paired video-text datasets**: Large-scale, paired video-text datasets for video-based human pose estimation are unavailable, and their manual collection is prohibitively expensive. (ii) **Inexpressiveness of static prompts**: The static prompts used in image-based multimodal models (*e.g.*, predefined joint names or simple templates like "a person on the left") are insufficient to capture the rich, dynamic nature of human actions in video. (iii) **Mismatch between modalities**: Visual features are dense and low-level, whereas language embeddings are sparse and conceptual. This inherent disparity makes establishing meaningful cross-modal correspondences a significant and non-trivial challenge.

To overcome these challenges, we introduce a novel multimodal paradigm for video-based human pose estimation, termed **L**anguage-**A**ssisted Video-based Human **Pose** Estimation (LAPose). To address the limitations of available video-text datasets and static prompts, we employ a pretrained Multimodal Large Language Model (MLLM) (Cheng et al. 2024) as a knowledge engine to generate interaction-aware textual descriptions from videos. To overcome the modality mismatch, we propose a coarse-to-fine alignment strategy. First, a cross-modal semantic interaction module establishes an initial, coarse-grained grounding between spatiotemporal visual features and interaction-aware textual embeddings. Building upon this, an optimal transport-driven feature matching mechanism enforces a fine-grained, geometrically consistent alignment between the two modalities. This robustly anchors the final video pose estimation in a multimodal consensus.

In summary, our contributions are as follows:

- We present the first language-assisted multimodal framework for video-based human pose estimation, inspired by the cognitive principle that humans integrate visual and linguistic information. Our approach leverages multimodal large language models (MLLMs) to generate interaction-aware textual descriptions, enriching visual features with high-level semantic context.
- We propose a coarse-to-fine cross-modal alignment strategy, comprising a cross-modal semantic interaction module for initial feature co-refinement and an optimal transport-driven feature matching mechanism to enforce fine-grained, geometrically consistent semantic correspondences.
- Extensive experiments demonstrate that our framework achieves new state-of-the-art performance on three widely used benchmarks.

Related Work

Unimodal Human Pose Estimation. Early human pose estimation approaches focus on static images, progressing from classic graphical models (Zhang et al. 2009) to deep learning architectures (Xu et al. 2022). To handle temporal information, video-based human pose estimation methods model motion dynamics using techniques such as temporal convolutions (Artacho and Savakis 2020) and feature differencing (Feng et al. 2023). Despite their progress, these human pose estimation methods in images or videos rely solely on the visual modality. The unimodal dependency makes them inherently fragile in visually ambiguous scenarios, such as occlusion or motion blur, where visual cues alone are insufficient for robust inference. Our work tries to address this limitation by integrating language as an auxiliary modality, inspired by the human multimodal cognitive process, which is discussed in the next paragraph.

From Multimodal Cognition to Computation. While most pose estimation models are unimodal, Dual Coding Theory (DCT) (Paivio 1971, 1986) in psychology suggests that human cognition is inherently multimodal, relying on both visual and verbal systems for understanding. Foundational work in cognitive science supports this view, showing that language provides symbolic abstraction to ground concrete visual perception (Sharma and Giannakos 2020; Xu et al. 2025a,b). In dynamic scenes, such as those involving motion, humans integrate visual motion cues with linguistic context and prior knowledge to form coherent, temporally grounded interpretations. This multimodal perception enables disambiguation of complex interactions and supports action prediction over time (Baltrušaitis, Ahuja,

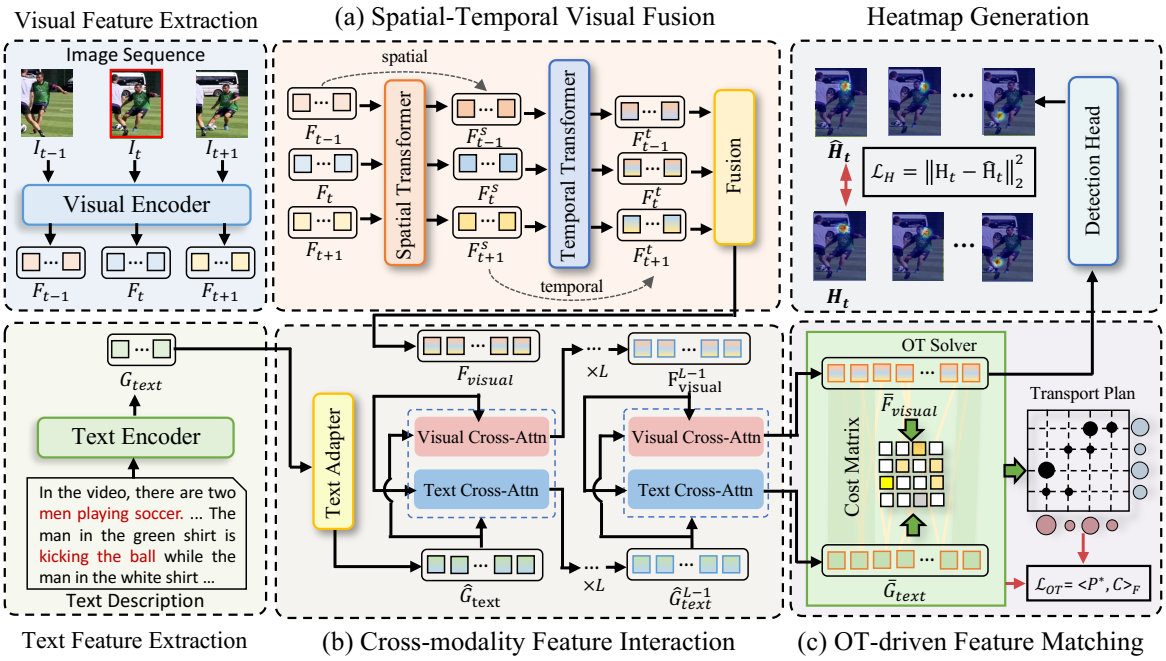

Figure 2: An overview of our LAPose framework. Given a video, we first obtain the frame sequence $\{I^i_{t-\omega}, ..., I^i_t, ..., I^i_{t+\omega}\}$ (for simplicity, we set $\omega = 1$ and omit the person index i in the figure.) and the text description generated from a pretrained multimodal large language model (MLLM). A spatial-temporal visual fusion module integrates multi-frame visual features, while a text encoder produces a textual embedding. We then propose a coarse-to-fine alignment strategy. First, the Cross-modal Semantic Interaction (CSI) module performs coarse-grained feature co-refinement. Subsequently, an Optimal Transport-driven Feature Matching mechanism $\mathcal{L}_{OT}$ enforces fine-grained, geometrically consistent feature correspondences. Finally, the refined visual features are fed into a detection head to generate pose heatmaps $\hat{H}_t$, supervised by a standard heatmap loss $\mathcal{L}_H$.

and Morency 2018). A few recent studies have explored text-guided image-based pose estimation (Wang, Xuan, and Zhang 2024). However, these methods are limited in two key aspects: they are designed for static images, not video, and they rely on fixed, template-based textual prompts that fail to capture the dynamic nature of human motion. In contrast, our work is the first to bring this cognitively-inspired, dynamic language-assisted paradigm to video-based human pose estimation.

Method

Preliminaries

Given an input video $\mathcal{V}$, the goal is to estimate the pose of each individual in every frame. Following previous works (He and Yang 2024; Wu et al. 2024a), we first employ an object detector (Qiao, Chen, and Yuille 2021) to obtain bounding boxes for each person. These boxes are then enlarged by 25% to ensure consistent cropping of the same person across consecutive frames. For each person i, this process yields a keyframe I^i_t and neighboring frames $\{I^i_{t-\omega}, ..., I^i_{t+\omega}\}$, where ω is the temporal span. The core task is to model the spatiotemporal dependencies within this sequence $\{I^i_{t-\omega}, ...I^i_t, ..., I^i_{t+\omega}\}$ to estimate the pose in the keyframe I^i_t. For simplicity, we omit the person index i and set $\omega = 1$ in the following sections.

Method Overview

The overall architecture of our LAPose framework is illustrated in Figure 2. First, we leverage a pretrained Multimodal Large Language Model (MLLM) to generate a detailed textual description representing human interactions and motion dynamics within the video. The text description is encoded into textual features using a text encoder. In parallel, a visual backbone extracts visual features from the video frame sequence, which are then aggregated by a spatial-temporal visual fusion module. To bridge the inherent modality gap, we introduce a coarse-to-fine cross-modal alignment strategy comprising two key components: (1) a Cross-modal Semantic Interaction (CSI) module for initial, coarse-grained feature co-refinement, and (2) an Optimal Transport-driven Feature Matching (OTFM) module for fine-grained, geometrically consistent feature correspondences. Finally, the refined visual representation is fed into a detection head to generate the final pose heatmaps. In the following, we will introduce these key components in detail.

Multimodal Feature Representation

Visual features are critical for joint localization in video-based human pose estimation. However, for challenging scenarios, such as video defocus in Figure 1, visual-only inputs could be insufficient for robust pose estimation. Our

Figure 3: An example of the MLLM-based text description generation process: the input video (top) and a task-specific prompt (middle) are used to generate the final text description (bottom).

core insight is grounded in cognitive science: humans understand physical motion not only through visual perception but also by generating internal linguistic narratives that describe actions and interactions (*e.g.*, (Clark 1998; Francken et al. 2015; Lupyan and Bergen 2020)). This integration of vision and language facilitates structured, semantic understanding. Motivated by this, we treat language as a powerful auxiliary signal—encoding high-level semantics of human-object interactions—to enhance visual representation learning and improve pose estimation robustness. To this end, we propose a multimodal video human pose estimation framework integrating spatiotemporal visual features and interaction-aware textual embeddings. We now detail the extraction of these representations.

Visual Features. Given the frame sequence $\{I_{t-1}, I_t, I_{t+1}\}$, we first utilize ViT (Dosovitskiy et al. 2020) as the visual encoder E_v to generate visual features $\{F_{t-1}, F_t, F_{t+1}\}$. As depicted in Figure 2 (a), these features are subsequently processed by a spatial-temporal fusion module, which aggregates the multi-frame information to yield the spatial-temporal visual representation F_{visual}:

$$\{F_{t-1}, F_t, F_{t+1}\} = E_v(\{I_{t-1}, I_t, I_{t+1}\}), \quad (1)$$

$$F_{visual} = \phi_{ST}(T_T(T_S(\{F_{t-1}, F_t, F_{t+1}\}))), \quad (2)$$

where T_S and T_T denote the spatial and temporal transformer layers, respectively, and ϕ_{ST} represents a temporal fusion layer.

Description Generation. A primary obstacle for multimodal video pose estimation is the absence of large-scale video-text datasets (Wang et al. 2023). To overcome this, we leverage a pretrained Multimodal Large Language Model (MLLM) as a knowledge engine to generate an interaction-aware textual description of human motion. Formally, for a given video $\mathcal{V}$ and a task-specific prompt $\mathcal{P}$, the MLLM produces a detailed description $\mathcal{T} = \text{MLLM}(\mathcal{V}, \mathcal{P})$. An example for description generation is shown in Fig. 3.

Textual Features. Given the text description $\mathcal{T}$, we first perform a standard tokenization process on $\mathcal{T}$. Then, we adopt a pre-trained transformer-based language model (Devlin et al. 2019) as our text encoder E_t to encode the text features G_{text}.

Cross-modal Semantic Interaction

Having extracted visual features F_{visual} and textual features G_{text}, the next challenge is to resolve the misalignment between the visual and textual modalities. To this end, we propose a coarse-to-fine alignment strategy. This section details the coarse-grained alignment designed to establish initial semantic associations, while the subsequent fine-grained stage is elaborated on in the next section. As illustrated in Fig. 2 (b), we propose a Cross-modal Semantic Interaction (CSI) module that facilitates a mutual exchange of information between the two modalities.

Firstly, a text adapter layer is incorporated into the text branch to enhance the semantic representations of the text description. Given initial text features G_{text}, the adapter operation is formulated as:

$$\hat{G}_{text} = G_{text} + \text{Gelu}(\text{G}_{\text{text}}\text{W}_{\text{down}})\text{W}_{\text{up}}, \quad (3)$$

where W_{down} and W_{up} denote the down-projection and up-projection weights, Gelu (Hendrycks and Gimpel 2016) is an activation function.

Next, we employ a dual cross-attention mechanism for the coarse feature interaction, comprising L layers of stacked Visual Cross-Attention (VCA) and Text Cross-Attention (TCA) blocks. This deep, iterative interaction allows each modality to be progressively refined by contextual cues from the other. For the l-th VCA block, the visual features are updated by attending to the textual features:

$$F_{visual}^{(l)} = \Psi\left(\frac{(F_{visual}^{(l-1)}W_Q)(\hat{G}_{text}^{(l-1)}W_K)^T}{\sqrt{d}}\right)\hat{G}_{text}^{(l-1)}W_V, \quad (4)$$

where W_Q, W_K, and W_V denote the query, key, and value projection matrices, respectively. Ψ denotes the softmax operation. The TCA block operates symmetrically. This cross-modal semantic interaction module yields coarse-aligned visual and textual features, denoted as $\bar{F}_{visual} = F_{visual}^{L}$ and $\bar{G}_{text} = G_{text}^{L}$, which serve as the input for our fine-grained matching stage.

Multimodal Feature Alignment based on Optimal Transport

Following the coarse alignment, our objective shifts to establishing a fine-grained correspondence between $\bar{F}_{visual}$ and $\bar{G}_{text}$. Considering the natural differences between textual and visual features, we further propose an Optimal Transport-driven Feature Matching (OTFM) objective, which formulates the cross-modal feature alignment task as an optimal transport problem.

Optimal Transport (OT), originally a mathematical formalism for distribution comparison, provides a geometric framework for comparing probability distributions (Peyré, Cuturi et al. 2019). It seeks the optimal transport plan that minimizes the cost of moving mass from one distribution to another under a specified cost function. Specifically, we have two sets of features $\bar{F}_{visual} = \{f_m\}_{m=1}^{M}$ and $\bar{G}_{text} = \{g_n\}_{n=1}^{N}$, the discrete empirical distribution can be denoted

as follows:

$$\nu = \sum_{m=1}^{M} \alpha_m \delta_{f_m} \text{ and } \tau = \sum_{n=1}^{N} \beta_n \delta_{g_n}, \tag{5}$$

where weights $\alpha = \{\alpha_m\}_{m=1}^{M} \in \Delta_M$ and $\beta = \{\beta_n\}_{n=1}^{N} \in \Delta_N$, Δ represents the M- and N-dimensional probability simplex, *i.e.,* $\sum_{m=1}^{M} \alpha_m = \sum_{n=1}^{N} \beta_n = 1$. M and N denote the number of samples in each empirical distribution, $\delta_{(\cdot)}$ represents the Dirac delta function. Formally, the optimal transport distance can be defined via the Kantorovich formulation as follows:

$$\begin{aligned} D_{OT}(\nu, \tau) &= \inf_{\pi \in \Pi(\nu,\tau)} \mathbb{E}_{(\nu,\tau)\sim\pi}[C(f, g)] \\ &= \min_{P \in \Pi(\nu,\tau)} \sum_{m=1}^{M} \sum_{n=1}^{N} P_{m,n} \cdot c(f_m, g_n), \\ &s.t.\ Pe = \nu, P^T e = \tau, \end{aligned} \tag{6}$$

where $C \in \mathbb{R}^{M \times N}$ denotes the cost matrix whose entry $c(f_m, g_n)$ measures the transportation cost between f_m and g_n, *i.e.,* $c(f_m, g_n) = 1 - \cos(f_m, g_n)$. $P \in \mathbb{R}^{M \times N}$ defines the transport plan to be optimized and $\Pi(\nu, \tau)$ denotes the transportation polytope, which collects all joint probabilities of ν and τ. Considering the prohibitive computational burden and statistical challenges of this original OT problem, we adopt the Sinkhorn-Knopp algorithm, which is more computationally amenable (Cuturi 2013), to solve an entropy-regularized OT problem:

$$\begin{aligned} D_{OT,\lambda}(\nu, \tau) &= \min_{P \in \Pi(\nu,\tau)} \sum_{m=1}^{M} \sum_{n=1}^{N} P_{m,n} \cdot c(f_m, g_n) - \lambda \mathbb{H}(P), \\ &s.t.\ Pe = \nu, P^T e = \tau, \end{aligned} \tag{7}$$

where λ is a regularization weight and $\mathbb{H}(P) = \sum_{m,n} P_{m,n} \mathrm{log} \mathrm{P}_{\mathrm{m,n}}$ denotes the entropy of the transport plan P. The optimized P^* can be computed in a few iterations:

$$P^* = diag(\nu^{(T)}) \exp(-\mathrm{C}/\lambda) \mathrm{diag}(\tau^{(\mathrm{T})}), \tag{8}$$

where T is the iteration step. At each Sinkhorn iteration t, $\nu^{(t+1)}$ and $\tau^{(t+1)}$ are updated as follows:

$$\nu^{(t+1)} = \nu / (\exp(-\mathrm{C}/\lambda) \tau^{(\mathrm{t})}), \tag{9}$$

$$\tau^{(t+1)} = \tau / (\exp(-\mathrm{C}/\lambda) \nu^{(\mathrm{t+1})}), \tag{10}$$

where the initiation $\tau^{(0)} = 1_N$. Therefore, the OT-driven feature matching loss $\mathcal{L}_{OT}$ could be derived from the optimal transport solution:

$$\mathcal{L}_{OT} = \langle P^*, C \rangle_F, \tag{11}$$

where $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product. Minimizing this loss enforces a geometrically consistent alignment, ensuring that semantically related visual and textual features are matched with minimal cost.

Heatmap Generation

The final language-assisted visual features are fed into a detection head to generate the keypoint heatmaps $\hat{\mathbf{H}}_t$. We utilize the standard heatmap loss $\mathcal{L}_H$ to supervise the pose estimation task:

$$\mathcal{L}_H = ||\mathbf{H}_t - \hat{\mathbf{H}}_t||_2^2, \tag{12}$$

where $\hat{\mathbf{H}}_t$ and $\mathbf{H}_t$ represent the generated and ground truth pose heatmaps, respectively.

Training Objectives

Considering the natural differences between textual and visual features, we also incorporate a vision-language contrastive learning objective to match the pairwise visual and text features. Specifically, the proposed visual-language contrastive learning could pull paired visual and textual representations closer in the embedding space while pushing non-paired samples apart. Formally, we formulate the symmetric similarities between $\bar{F}_{visual}$ and $\bar{G}_{text}$ within the batch:

$$p_i^{f2g}(f_i) = \frac{\exp(\mathrm{sim}(\mathrm{f_i}, \mathrm{g_i})/\eta)}{\sum_{j=1}^{B} \exp(\mathrm{sim}(\mathrm{f_i}, \mathrm{g_j})/\eta)}, \tag{13}$$

$$p_i^{g2f}(g_i) = \frac{\exp(\mathrm{sim}(\mathrm{f_i}, \mathrm{g_i})/\eta)}{\sum_{j=1}^{B} \exp(\mathrm{sim}(\mathrm{f_i}, \mathrm{g_j})/\eta)}, \tag{14}$$

where f_i and g_i represent the visual and text features, $\mathrm{sim}(\cdot, \cdot)$ denotes the cosine similarity, η denotes the temperature parameter, and B is the batch size. Following (Xiang et al. 2023), we utilize Kullback-Leibler divergence as a vision-text contrastive loss:

$$\mathcal{L}_{cl} = \frac{1}{2} \mathbb{E}_{f,g\sim D}[KL(P^{f2g}(f), y^{f2g}) + KL(P^{g2f}(g), y^{g2f})], \tag{15}$$

where D represents the full dataset, y^{f2g} and y^{g2f} denote the ground-truth, which is formulated as 0 for negative pairs and 1 for positive pairs.

The proposed vision-language contrastive learning loss $\mathcal{L}_{cl}$ and OT-driven feature matching loss $\mathcal{L}_{OT}$ promote a better feature matching and alignment between the visual features $\bar{F}_{visual}$ and textual features $\bar{G}_{text}$. The total training loss function is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_H + \gamma_1 \mathcal{L}_{cl} + \gamma_2 \mathcal{L}_{OT}, \tag{16}$$

where γ_1 and γ_2 denote two hyper-parameters.

Experiments

Experimental Settings

Datasets. We evaluate our LAPose for video-based human pose estimation on three large-scale benchmarks: PoseTrack2017 (Iqbal, Milan, and Gall 2017), PoseTrack2018 (Andriluka et al. 2018), and PoseTrack21 (Doering et al. 2022). PoseTrack2017 contains 300 video sequences and 80,144 pose annotations, where 250 video clips are the training set and the remaining video sequences are the validation set. Each pose annotation in this dataset

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
PoseWarper (Bertasius et al. 2019)	81.4	88.3	83.9	78.0	82.4	80.5	73.6	81.2
CorrTrack (Rafi et al. 2020)	86.1	87.0	83.4	76.4	77.3	79.2	73.3	80.8
Dynamic-GNN (Yang et al. 2021)	88.4	88.4	82.0	74.5	79.1	78.3	73.1	81.1
DCPose (Liu et al. 2021)	88.0	88.7	84.1	78.4	83.0	81.4	74.2	82.8
FAMI-Pose (Liu et al. 2022a)	89.6	90.1	86.3	80.0	84.6	83.4	77.0	84.8
SLT-Pose (Gai et al. 2023)	88.9	89.7	85.6	79.5	84.2	83.1	75.8	84.2
HANet (Jin et al. 2023)	90.0	90.0	85.0	78.8	83.1	82.1	77.1	84.2
TDMI (Feng et al. 2023)	90.0	91.1	87.1	81.4	85.2	84.5	78.5	85.7
M-HANet (Jin et al. 2024)	90.3	90.7	85.3	79.2	83.4	82.6	77.8	84.8
DSTA (He and Yang 2024)	89.3	90.6	87.3	82.6	84.5	85.1	77.8	85.6
JMPose (Wu et al. 2024a)	90.7	**91.6**	87.8	82.1	85.9	85.3	79.2	86.4
TPSD-MS (Zhang et al. 2025)	91.1	91.5	87.6	82.1	85.9	85.0	79.4	86.4
TPSD-ViT (Zhang et al. 2025)	90.7	91.1	87.9	83.6	85.3	86.3	80.0	86.7
LAPose (Ours)	**91.2**	**91.6**	**88.1**	**84.1**	**89.4**	**88.0**	**81.3**	**87.9**

Table 1: Quantitative results on the PoseTrack2017 dataset.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
PoseWarper (Bertasius et al. 2019)	79.9	86.3	82.4	77.5	79.8	78.8	73.2	79.7
Dynamic-GNN (Yang et al. 2021)	80.6	84.5	80.6	74.4	75.0	76.7	71.8	77.9
DCPose (Liu et al. 2021)	84.0	86.6	82.7	78.0	80.4	79.3	73.8	80.9
FAMI-Pose (Liu et al. 2022a)	85.5	87.7	84.2	79.2	81.4	81.1	74.9	82.2
SLT-Pose (Gai et al. 2023)	84.3	87.5	83.5	78.5	80.9	80.2	74.4	81.5
HANet (Jin et al. 2023)	86.1	88.5	84.1	78.7	79.0	80.3	77.4	82.3
TDMI (Feng et al. 2023)	86.2	88.7	85.4	80.6	82.4	82.1	77.5	83.5
M-HANet (Jin et al. 2024)	86.7	88.9	84.6	79.2	79.7	81.3	78.7	82.7
DSTA (He and Yang 2024)	85.9	88.8	85.0	81.1	81.5	83.0	77.4	83.4
JMPose (Wu et al. 2024a)	86.6	88.7	**86.0**	81.6	83.3	83.2	78.2	84.1
TPSD-MS (Zhang et al. 2025)	87.0	89.0	85.6	81.5	83.4	82.4	78.2	84.1
TPSD-ViT (Zhang et al. 2025)	86.9	89.0	**86.0**	81.6	83.3	83.3	78.0	84.2
LAPose (Ours)	**87.8**	**89.4**	**86.0**	**82.6**	**84.4**	**84.0**	**79.0**	**85.0**

Table 2: Quantitative results on the PoseTrack2018 dataset.

has 15 keypoints. In addition, the training samples deliver dense annotations in the center 30 frames, and validation samples provide additional labels every four frames. PoseTrack2018 comprises more video sequences and pose annotations, including 763 video sequences and 153,615 pose labels. Specifically, there are 593 video sequences for training and 170 video clips for validation. In addition to the locations of the 15 keypoints, the pose annotation also contains a flag indicating whether the joint is visible. Based on the PoseTrack2018, PoseTrack21 further increases the number of pose annotations, including 177,164 pose labels. PoseTrack21 focuses on small persons and persons in crowded scenes, presenting a more challenging evaluation scenario.

Evaluation Metric. Following standard protocol, we report the Average Precision (AP) to evaluate the performance between ground-truth poses and predicted poses. Furthermore, we average the AP of all joints to obtain the final mAP performance.

Implementation Details. All experiments are conducted on $4\times$ NVIDIA RTX 4090 GPUs using the PyTorch framework for 20 epochs. We utilize VideoLLaMA2 (Cheng et al. 2024) as the Multimodal Large Language Model (MLLM). We utilize ViT (Dosovitskiy et al. 2020) as the visual encoder and BERT (Devlin et al. 2019) as the text encoder. The image size is set as 256×192. We employ four types of data augmentation during training, including 1) random rotation $[-45°, 45°]$, 2) random scale $[0.65, 1.35]$, 3) random truncation, and 4) flipping. The time span ω is set to 1. We adopt AdamW as our optimizer with the initial learning rate of 2e-4 (decays to 2e-5, 2e-6, and 2e-7 at the 6-th, 12-th, and 16-th, respectively).

Qualitative Results

Results on the PoseTrack2017 Dataset. As shown in Table 1, LAPose achieves a new state-of-the-art (SOTA) performance of 87.9 mAP, surpassing the previous best method, TPSD-ViT (Zhang et al. 2025), by a significant margin of 1.2 mAP. Notably, our method exhibits consistent gains on challenging joints, achieving 89.4 AP for hip and 88.0 AP for knee, outperforming prior work by 3.5 AP and 1.7 AP, respectively. This is consistent with our expectations, as language-assisted semantic priors are highly effective at resolving visual ambiguities where visual-only models typically struggle.

Results on the PoseTrack2018 Dataset. Table 2 reports comparisons between the proposed LAPose and existing approaches on the PoseTrack2018 dataset. LAPose again sets a new SOTA with 85.0 mAP, outperforming TPSD-ViT (Zhang et al. 2025) by 0.8 mAP. For challenging joints, wrist, hip, and knee joints, we achieve a better performance than the existing method, with an AP of 82.6, 84.4, and 84.0, respectively.

Results on the PoseTrack21 Dataset. PoseTrack21 introduces more challenging scenarios with smaller individuals and heavier occlusion. As detailed in Table 3, LAPose demonstrates its robustness by achieving an SOTA performance of 84.8 mAP. We also find that our method obtains an 84.4 AP for hip and 79.1 AP for ankle joints.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
CorrTrack (Rafi et al. 2020)	-	-	-	-	-	-	-	72.3
DCPose (Liu et al. 2021)	83.2	84.7	82.3	78.1	80.3	79.2	73.5	80.5
FAMI-Pose (Liu et al. 2022a)	83.3	85.4	82.9	78.6	81.3	80.5	75.3	81.2
TDMI (Feng et al. 2023)	85.8	87.5	85.1	81.2	83.5	82.4	77.9	83.5
DSTA (He and Yang 2024)	87.5	87.0	84.2	81.4	82.3	82.5	77.7	83.5
JMPose (Wu et al. 2024a)	85.8	88.1	**85.7**	82.5	84.1	83.1	78.5	84.0
TPSD-MS (Zhang et al. 2025)	87.0	87.6	85.3	81.5	84.0	82.6	77.9	83.9
TPSD-ViT (Zhang et al. 2025)	87.7	88.0	85.0	81.7	83.4	82.8	78.3	84.1
LAPose (Ours)	**88.1**	**88.3**	85.4	**82.6**	**84.4**	**83.7**	**79.1**	**84.8**

Table 3: Quantitative results on the Posetrack21 dataset.

Figure 4: The keyframe (a) and visual comparisons of results obtained from DCPose (b), JMPose (c), and our LAPose (d) on challenging scenes in the PoseTrack dataset. Inaccurate predictions are highlighted with the red solid circles.

Qualitative Results

Qualitative Comparison. We further qualitatively evaluate the performance of our method in challenging scenes. Figure 4 shows the visual comparisons of the proposed LAPose, DCPose (Liu et al. 2021), and JMPose (Wu et al. 2024a). From Figure 4, we can see that compared with existing methods, the proposed LAPose identifies the human poses more accurately and robustly in challenging scenes (such as rapid-motion, self-occlusion, and video defocus).

Ablation Study

Impact of the Coarse-to-Fine Alignment Strategy. We first analyze our two-stage alignment mechanism. When the Cross-modal Semantic Interaction (CSI) module is removed (Table 4, row (a)), performance drops significantly by 1.7 mAP to 86.2 mAP. This highlights the necessity of the initial coarse interaction for establishing a foundational cross-modal context. In the next setting (Table 4 (b)), removing the Optimal Transport-driven Feature Matching (OTFM) mechanism results in the most substantial performance degradation, with mAP falling by 2.4 mAP to 85.5 mAP. This validates the importance of our optimal transport-driven feature matching module in enforcing fine-grained feature correspondence between visual and textual features.

Ablation	CSI	OTFM	$\mathcal{L}_{cl}$	Mean	Declines
LAPose	✓	✓	✓	87.9	-
(a)	✗	✓	✓	86.2	1.7 (↓)
(b)	✓	✗	✓	85.5	2.4 (↓)
(c)	✓	✓	✗	87.4	0.5 (↓)

Table 4: Ablation of different designs in LAPose.

Impact of Contrastive Learning. As shown in Table 4 (c), ablating the auxiliary contrastive loss $\mathcal{L}_{cl}$ leads to a 0.5 mAP decrease. While smaller than the drop from removing the core alignment modules, this demonstrates that explicitly enforcing global instance-level similarity between visual and textual representations provides a valuable complementary supervision to the overall learning objective.

Conclusion

In this work, we introduce LAPose, a novel framework for video-based human pose estimation, inspired by the psychological theory of Dual Coding, which highlights the multimodal nature of human cognition. To address the lack of paired video-text datasets, we leverage Multimodal Large Language Models as knowledge engines to generate rich textual descriptions. For effective video-text multimodal alignment, we propose a coarse-to-fine strategy featuring a cross-modal semantic interaction module and an optimal transport-based feature matching mechanism, enforcing geometrically consistent feature correspondence. Extensive experiments on three large-scale benchmarks show that LAPose sets a new state of the art, especially in visually ambiguous scenarios. We believe this work opens a promising direction for video-based human pose estimation, where models learn not just from pixels, but from the rich semantic context that language provides.

Acknowledgments

This work is supported by the National Key R&D Program of China (No. 2024YFB3311605), National Natural Science Foundation of China (No. 62276112, No. 62372402) and the Key R&D Program of Zhejiang Province (No. 2025C01084).

References

Andriluka, M.; Iqbal, U.; Insafutdinov, E.; Pishchulin, L.; Milan, A.; Gall, J.; and Schiele, B. 2018. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5167–5176.

Artacho, B.; and Savakis, A. 2020. Unipose: Unified human pose estimation in single images and videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7035–7044.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443.

Bertasius, G.; Feichtenhofer, C.; Tran, D.; Shi, J.; and Torresani, L. 2019. Learning temporal pose estimation from sparsely-labeled videos. In *Advances in Neural Information Processing Systems*, 3027–3038.

Chen, H.; Wu, S.; Wang, Z.; Yin, Y.; Jiao, Y.; Lyu, Y.; and Liu, Z. 2025. Causal-Inspired Multitask Learning for Video-Based Human Pose Estimation. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 2052–2060. AAAI Press.

Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; et al. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Clark, H. H. 1998. *Using Language*. Cambridge University Press.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Doering, A.; Chen, D.; Zhang, S.; Schiele, B.; and Gall, J. 2022. Posetrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20963–20972.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Feng, R.; Gao, Y.; Ma, X.; Tse, T. H. E.; and Chang, H. J. 2023. Mutual information-based temporal difference learning for human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17131–17141.

Francken, J. C.; Kok, P.; Hagoort, P.; and De Lange, F. P. 2015. The behavioral and neural effects of language on motion perception. *Journal of cognitive neuroscience*, 27(1): 175–184.

Gai, D.; Feng, R.; Min, W.; Yang, X.; Su, P.; Wang, Q.; and Han, Q. 2023. Spatiotemporal Learning Transformer for Video-Based Human Pose Estimation. *IEEE Transactions on Circuits and Systems for Video Technology*.

He, J.; and Yang, W. 2024. Video-Based Human Pose Regression via Decoupled Space-Time Aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1022–1031.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Iqbal, U.; Milan, A.; and Gall, J. 2017. Posetrack: Joint multi-person pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011–2020.

Jiao, Y.; Wang, Z.; Liu, Z.; Fan, S.; Wu, S.; Wu, Z.; and Xu, Z. 2025. Optimizing Human Pose Estimation Through Focused Human and Joint Regions. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 4102–4110. AAAI Press.

Jin, K.-M.; Lee, G.-H.; Nam, W.-J.; Kang, T.-K.; Kim, H.-W.; and Lee, S.-W. 2024. Masked Kinematic Continuity-aware Hierarchical Attention Network for pose estimation in videos. *Neural Networks*, 169: 282–292.

Jin, K.-M.; Lim, B.-S.; Lee, G.-H.; Kang, T.-K.; and Lee, S.-W. 2023. Kinematic-aware hierarchical attention network for human pose estimation in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5725–5734.

Liu, Z.; Chen, H.; Feng, R.; Wu, S.; Ji, S.; Yang, B.; and Wang, X. 2021. Deep Dual Consecutive Network for Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 525–534.

Liu, Z.; Feng, R.; Chen, H.; Wu, S.; Gao, Y.; Gao, Y.; and Wang, X. 2022a. Temporal feature alignment and mutual information maximization for video-based human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11006–11016.

Liu, Z.; Wu, S.; Xu, C.; Wang, X.; Zhu, L.; Wu, S.; and Feng, F. 2022b. Copy Motion From One to Another: Fake Motion Video Generation. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, 1223–1231. ijcai.org.

Lupyan, G.; and Bergen, B. K. 2020. Effects of language on visual perception. *Trends in Cognitive Sciences*, 24(11): 930–944.

Paivio, A. 1971. *Imagery and Verbal Processes*. Holt, Rinehart Winston.

Paivio, A. 1986. *Mental Representations: A Dual Coding Approach*. Oxford University Press.

Peyré, G.; Cuturi, M.; et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607.

Qiao, S.; Chen, L.-C.; and Yuille, A. 2021. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10213–10224.

Rafi, U.; Doering, A.; Leibe, B.; and Gall, J. 2020. Self-supervised keypoint correspondences for multi-person pose estimation and tracking in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, 36–52. Springer.

Sharma, K.; and Giannakos, M. 2020. Multimodal data capabilities for learning: What can multimodal data tell us about learning? *British Journal of Educational Technology*, 51(5): 1450–1484.

Subramaniam, V.; Conwell, C.; Wang, C.; Kreiman, G.; Katz, B.; Cases, I.; and Barbu, A. 2024. Revealing vision-language integration in the brain with multimodal networks. *ArXiv*, arXiv–2406.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5693–5703.

Tang, H.; Li, Z.; Zhang, D.; He, S.; and Tang, J. 2024. Divide-and-conquer: Confluent triple-flow network for RGB-T salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tang, H.; Liu, J.; Yan, S.; Yan, R.; Li, Z.; and Tang, J. 2023. M3net: multi-view encoding, matching, and fusion for few-shot fine-grained action recognition. In *Proceedings of the 31st ACM international conference on multimedia*, 1719–1728.

Tang, L.; Huang, K.; Chen, C.; Yuan, Y.; Li, C.; Tu, X.; Ding, X.; and Huang, Y. 2025a. Dissecting generalized category discovery: Multiplex consensus under self-deconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 297–307.

Tang, L.; Yuan, Y.; Chen, C.; Zhang, Z.; Huang, Y.; and Zhang, K. 2025b. OCRT: Boosting Foundation Models in the Open World with Object-Concept-Relation Triad. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 25422–25433.

Wang, D.; Xuan, S.; and Zhang, S. 2024. Locllm: Exploiting generalizable human keypoint localization via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 614–623.

Wang, X.; Ma, K.; Zhong, R.; Wang, X.; Fang, Y.; Xiao, Y.; and Xia, T. 2024. Towards dual transparent liquid level estimation in biomedical lab: Dataset, methods and practices. In *European Conference on Computer Vision*, 198–214. Springer.

Wang, Y.; He, Y.; Li, Y.; Li, K.; Yu, J.; Ma, X.; Li, X.; Chen, G.; Chen, X.; Wang, Y.; et al. 2023. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*.

Wang, Z.; Fan, S.; Liu, Z.; Wu, Z.; Wu, S.; and Jiao, Y. 2025. Multi-Grained Feature Pruning for Video-Based Human Pose Estimation. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, 1–5. IEEE.

Wu, S.; Chen, H.; Yin, Y.; Hu, S.; Feng, R.; Jiao, Y.; Yang, Z.; and Liu, Z. 2024a. Joint-Motion Mutual Learning for Pose Estimation in Video. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8962–8971.

Wu, S.; Liu, Z.; Zhang, B.; Zimmermann, R.; Ba, Z.; Zhang, X.; and Ren, K. 2024b. Do as I Do: Pose Guided Human Motion Copy. *IEEE Trans. Dependable Secur. Comput.*, 21(6): 5293–5307.

Wu, S.; Zhang, H.; Liu, Z.; Chen, H.; and Jiao, Y. 2025. Enhancing Human Pose Estimation in Internet of Things via Diffusion Generative Models. *IEEE Internet Things J.*, 12(10): 13556–13567.

Xiang, W.; Li, C.; Zhou, Y.; Wang, B.; and Zhang, L. 2023. Generative action description prompts for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10276–10285.

Xu, H.; Ke, X.; Li, Y.; Xu, R.; Wu, H.; Lin, X.; and Guo, W. 2024. Vision-Language Action Knowledge Learning for Semantic-Aware Action Quality Assessment. In *European Conference on Computer Vision*, 423–440.

Xu, H.; Ke, X.; Wu, H.; Xu, R.; Li, Y.; and Guo, W. 2025a. Language-Guided Audio-Visual Learning for Long-Term Sports Assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23967–23977.

Xu, H.; Wu, H.; Ke, X.; Li, Y.; Xu, R.; and Guo, W. 2025b. Quality-Guided Vision-Language Learning for Long-Term Action Quality Assessment. *IEEE Transactions on Multimedia*, 27: 7326–7339.

Xu, Y.; Zhang, J.; Zhang, Q.; and Tao, D. 2022. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35: 38571–38584.

Yang, Y.; Ren, Z.; Li, H.; Zhou, C.; Wang, X.; and Hua, G. 2021. Learning dynamics via graph neural networks for human pose estimation and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8074–8084.

Zhang, R.; Lin, D.; Wang, X.; Liu, R.; Sheng, B.; Baciu, G.; Chen, C. P.; and Li, P. 2025. Temporal-Interim Pose Synthesis and Distillation for Dynamic Human Pose Estimation. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhang, X.; Li, C.; Tong, X.; Hu, W.; Maybank, S.; and Zhang, Y. 2009. Efficient human pose estimation via parsing a tree structure based human model. In *2009 IEEE 12th International Conference on Computer Vision*, 1349–1356. IEEE.