HuMoCon: Concept Discovery for Human Motion Understanding

Qihang Fang[1,2] Chengcheng Tang[2] Bugra Tekin[2] Shugao Ma[2] Yanchao Yang[1]
[1]The University of Hong Kong [2]Meta
qihfang@gmail.com, {chengcheng.tang,bugratekin,shugao}@meta.com, yanchaoy@hku.hk

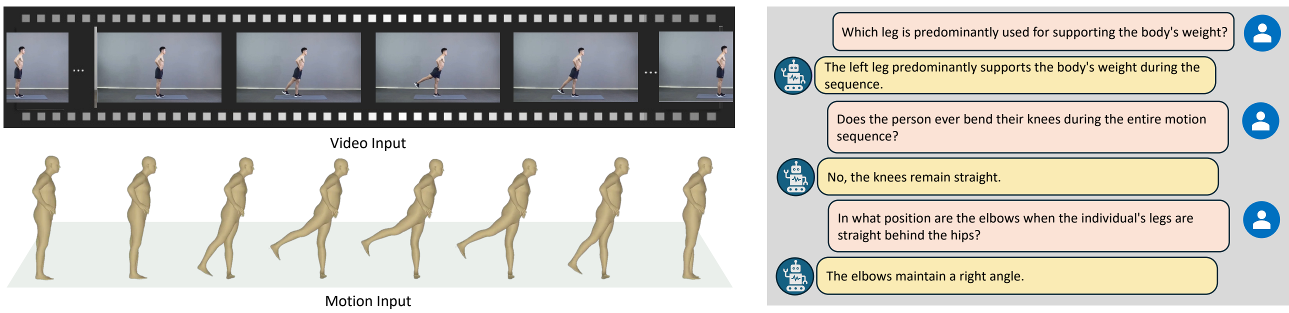

Figure 1. We present HuMoCon for concept discovery and human motion understanding with video or motion input. To address the challenge of feature misalignment and high-frequency information loss, we propose a novel feature alignment strategy and an advanced masked auto-encoder reconstructing velocity. HuMoCon empowers effective motion concept discovery and accurate Question Answering tasks, significantly outperforming state-of-the-art methods qualitatively and quantitatively.

Abstract

We present HuMoCon, a novel motion-video understanding framework designed for advanced human behavior analysis. The core of our method is a human motion concept discovery framework that efficiently trains multi-modal encoders to extract semantically meaningful and generalizable features. HuMoCon addresses key challenges in motion concept discovery for understanding and reasoning, including the lack of explicit multi-modality feature alignment and the loss of high-frequency information in masked autoencoding frameworks. Our approach integrates a feature alignment strategy that leverages video for contextual understanding and motion for fine-grained interaction modeling, further with a velocity reconstruction mechanism to enhance high-frequency feature expression and mitigate temporal over-smoothing. Comprehensive experiments on standard benchmarks demonstrate that HuMoCon enables effective motion concept discovery and significantly outperforms state-of-the-art methods in training large models for human motion understanding. We will open-source the associated code with our paper.

1. Introduction

Understanding human behavior is crucial for gaining deeper insights into human actions and reasoning, which are foundational for developing human-centered AI systems [5, 26, 37, 61] that effectively address real-world tasks [14, 68, 72]. Recent advancements in large language models (LLMs) have led to significant progress in video- and motion-based methods [5, 15, 37, 40, 75] for analyzing human behavior. Despite these advancements, achieving a precise and fine-grained understanding of human actions remains a persistent and complex challenge.

Human behavior is often represented as motion sequences structured based on the kinematics of the human skeleton [32**?**]. Traditional approaches to analyzing motion sequences have predominantly focused on classification [11, 24, 34, 42, 44, 45, 58, 76], relying on predefined action categories. While effective for specific tasks, these methods often lack the flexibility for deeper or more nuanced analysis. Other methods employ statistical models to assess motion quality [3, 6, 18, 43, 51], but they require significant expert knowledge, limiting scalability.

Recent advancements in language models [40, 50, 75] have opened new possibilities in motion analysis, enabling tasks such as motion generation and motion captioning based on free-text inputs [4, 20, 52, 59, 61], advancing the field of motion analysis. However, obtaining high-quality motion sequences and annotations remains costly, and motion sequences generally lack information about the surrounding environment, limiting models' ability to understand human-environment interactions.

In contrast, video data offers rich visual context and is easier to collect. It has been widely used for action classification [16, 62, 65] and video Q&A [5, 37, 40]. However, video data often contains irrelevant elements that can hinder accurate content analysis. For example, a model might learn from training data that most running occurs in stadiums, leading it to mistakenly classify other stadium activities as running during testing.

In recent years, significant progress has been made in developing multi-modality models [5, 87]. Integrating multiple modalities, as opposed to relying on a single modality, enhances the network's ability to learn fine-grained and semantic features. For instance, LART [56] combines video and motion data to ensure that video-based action classification focuses on the human body rather than overemphasizing environmental context. MotionLLM [5] expands large language models (LLMs) to video and motion, building a model that can understand both. MotionLLM demonstrated that learning from video and motion improves the network's ability to analyze human behavior. However, MotionLLM implicitly aligns video and motion using motion-video pairs during LLM training without optimizing the encoder for explicit cross-modal alignment. To address this limitation and further enhance feature representation, we propose explicitly aligning features from different modalities during the encoding process within a human motion concept discovery framework.

Achieving effective cross-modal alignment and feature representation relies heavily on robust encoder training methods. Many existing studies [70, 71, 87] adopt a masked autoencoder framework [27] when training encoders. This approach involves applying a mask to a certain proportion of the input, extracting features from the masked data using the encoder, and reconstructing the original unmasked input with a decoder. While effective in learning high-quality features, this masking strategy inherently causes the loss of high-frequency information, resulting in temporally over-smoothed reconstructions that hinder the LLM's ability to understand human behavior. [47] To address this issue, we propose reconstructing velocity as a means to preserve high-frequency information. Following InfoCon [46], we design a VQ-VAE-based loss function that helps retain the fine-grained features of the original masked autoencoder while further enhancing high-frequency feature expression. This approach significantly boosts the subsequent LLM's capacity to model and understand human motion effectively.

In summary, we present HuMoCon, an effective human motion concept discovery pertaining framework for advanced human behavior understanding from both video and motion data. Our key contributions include: (i) a novel feature alignment strategy that explicitly leverages video data for contextual understanding and motion data for fine-grained interaction modeling, and (ii) an enhanced masked autoencoder framework incorporating velocity reconstruction to significantly improve high-frequency feature expression and mitigate temporal over-smoothing. To evaluate the effectiveness of our proposed method, we conduct experiments on the Activity-QA [78] and BABEL-QA [13] benchmarks [54], comparing our approach with the state-of-the-art methods. Comprehensive quantitative and qualitative evaluations demonstrate that HuMoCon achieves state-of-the-art performance on standard benchmarks for motion-video understanding.

2. Related Work

2.1. Human Motion Understanding

The core of human motion understanding is to learn the semantics of motion sequences or videos. In the early stages, research in this field is often based on predefined action categories. Such work can be divided into two types: action classification [11, 24, 34, 42, 44, 45, 58, 76] and motion generation targeted at specific categories [4, 20, 28, 31, 36, 52, 59–61, 74, 77]. For action recognition, several works [11, 44] utilize Recurrent Neural Networks (RNN) to extract temporal information from human joint sequences, beyond the CNN-based ones [34, 45]. Subsequent works [24, 42, 58, 76] propose representing motion data as a graph structure, enabling the use of spatio-temporal graph convolutional networks (ST-GCNs) to effectively capture the connectivity and interactions between human joints. For motion generation, several studies [28, 31, 36, 60, 74, 77] have concentrated on generating specific, single actions. In contrast, other works [4, 20, 52, 59, 61] aim to generate a broader range of diverse action types, addressing the challenge of variability in motion synthesis.

In recent years, more research has shifted towards captioning [23, 30], generation [1, 22, 23, 30, 35, 53, 64, 81, 82, 86], and analysis [5, 9, 15, 17] based on free-form text. Leveraging alignment of motion latent features with the CLIP [55] embedding space, MotionCLIP [63] enables the generation of out-of-distribution motions. Transformers [1, 22, 23, 53, 81] and diffusion [35, 64, 82] models are also utilized for text-conditioned motion generation. Moreover, MotionGPT [30] and AvatarGPT [86] utilize large language models (LLMs) to handle multiple motion-related tasks simultaneously. Despite these advancements, research focused specifically on motion analysis remains relatively limited. PoseFix [9] and CigTime [15] are proposed to annotate the corrective instructions from source to target for pose and motion sequences. MotionLLM [5] proposes a method to analyze both the video and motion sequences with detailed spatial-temporal awareness and reasoning abilities. However, MotionLLM relies on implicit alignment through paired data and lacks explicit cross-modal feature alignment.

2.2. Video Understanding

In addition to motion understanding, video understanding has emerged as a prominent research area , attracting extensive attention in recent years. Early works on video classification and segmentation relied on handcrafted features [10, 44, 66, 67] or convolutional neural networks [16, 62, 65]. Beyond classification, video analysis [3, 6, 18, 25, 26, 37, 40, 43, 50, 51, 57, 73, 75, 85] has become a highly active field of research. Action quality assessment has been the focus of several works [3, 6, 18, 43, 51], with methods such as Aifit [18] and PoseTrainer [6] proposed to correct user actions. These approaches evaluate actions in conjunction with complex textual instructions [3, 43, 51]. Other studies [25, 26, 57, 73, 85] aim to generate descriptive narratives for videos, broadening the scope of video understanding. Recently, many works [2, 37, 40, 50, 75, 80] have leveraged large language models (LLMs) to understand video content and perform complex reasoning tasks. These methods combine video processing with the advanced reasoning capabilities of LLMs, further advancing the state-of-the-art in video analysis. Video-LLaVA [40] build a model to perform visual reasoning capabilities on images and videos simultaneously. Follow-up works [37, 50, 75] extend Video-LLaVA for faster inference and more accurate reasoning ability. Unlike typical video understanding algorithms, which rely heavily on visual context alone, our approach integrates motion data directly, enhancing the model's ability to understand human-centric actions.

2.3. Multi-modality Pre-training

Multi-modality pre-training has primarily focused on aligning vision and language representations. CLIP [55] pioneered this effort by using contrastive learning to align image and text features on large-scale datasets. Subsequent works [12, 48, 63, 83] extended CLIP's framework to other modalities, including video [48], audio [12] and motion [63].

Recent studies, such as VALOR [7] and VAST [8], demonstrated that cross-modal alignment not only enhances model robustness but also maintains high performance across modalities. Building on this, follow-up works [19, 87] have aimed to expand the number of aligned modalities, further advancing multi-modality research.

In the context of motion and video alignment, MotionLLMs [5] is the first work to address this challenge. However, it relies solely on implicit alignment during the fine-tuning stage of large language models.

3. Method

3.1. Overview

Our objective is to develop an LLM-based system capable of understanding and reasoning both human video and motion inputs. Given a question Q and an input, which can be either a video U or motion sequence M, the model generates the corresponding answer W. *Specifically,* a video U is represented as a sequence of T keyframes, denoted as $\{u_0, u_1, ..., u_{T-1}\}$. Similarly, a motion sequence M is represented as K consecutive pose frames, denoted as $\{m_0, m_1, ..., m_{K-1}\}$.

We present an overview of the pipeline in Fig. 2. To enable effective multi-modal human motion understanding, our framework focuses on: (1) discovering human motion concepts from both video and motion data, (2) preserving high-frequency motion details using velocity reconstruction to reduce temporal smoothing, and (3) aligning video and motion features to enhance multi-modal comprehension. To accomplish this, we train our video and motion encoders with four key learning objectives: *Masked Reconstruction*, *Discriminative Informativeness*, *Actionable Informativeness*, and *Feature Alignment*. Consequently, we leverage the encoded features of videos and motions to facilitate the fine-tuning of LLMs for video-motion understanding and reasoning. The following sections detail each component of our approach, including multi-modality encoder pre-training, modality translation, and instruction tuning.

3.2. Multi-Modality Encoder Pre-training via Human Motion Concept Discovery

Pre-training encoders are critical for multi-modality large language models (LLMs) [70, 71, 87], as they enable the effective extraction of essential features from large-scale input data. To enhance the encoder's ability to capture key information, we design three distinct learning objectives: masked reconstruction, velocity reconstruction, and feature alignment, all contributing to the quality of the discovered human motion concepts as well as the pre-training efficiency. For both videos and motions, we adopt a similar network structure during encoder pre-training. Consequently, in the following section, we focus on detailing the network structure for video inputs.

Encoding Function. In our approach, we design masked autoencoder architectures tailored for both video and motion data. *Specifically,* for videos, we utilize a pre-trained encoder $\mathcal{E}_u$, to encode the input U into a feature representation f^u:

$$F^u = \mathcal{E}_u(U). \tag{1}$$

Then, we employ a codebook C^u, modeled with VQ-VAE to quantize f^u into a discrete feature representation d^u:

$$D^u = \text{Quantize}(F^u, C^u), \tag{2}$$

where 'Quantize' denotes the quantization operation, and C^u is the codebook used for discretization, also serving as the discovered human motion concepts.

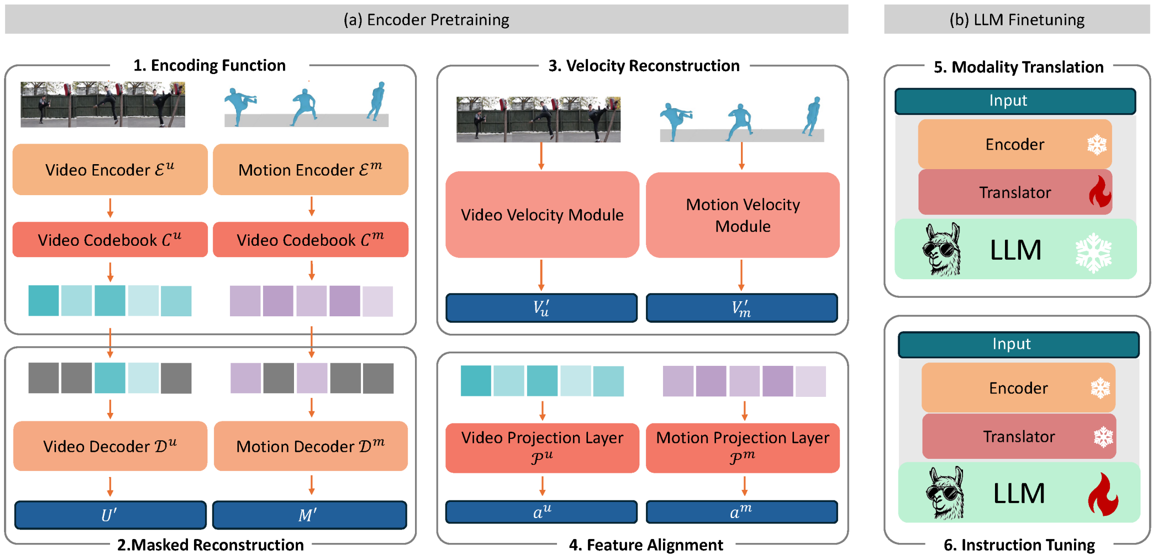

Figure 2. **System overview of our method.** (a) The encoder pre-training process for learning and aligning video and motion features and enhancing high-frequency details through velocity reconstruction. We utilize a VQ-VAE-based structure, and we design effective learning objectives to enhance the encoder to extract semantic meaningful and fine-grained features. (b) The fine-tuning of the large language model (LLM) for video and motion reasoning consists of two stages: Modality Translation and Multi-modality Instruction Tuning. In the Modality Translation stage, we train a translation layer for each modality to map the encoding feature to the LLM space. In the Instruction Tuning stage, we fine-tune the LLM to understand human motion and videos for downstream tasks.

Masked Reconstruction. The masked autoencoder has proven to be highly effective for pre-training [27], as it facilitates learning of meaningful and low-frequency features from input data. Thus, we leverage masked auto-encoding to enhance the discovered concepts and facilitate the multi-modal feature pre-training.

More explicitly, we mask D^u by a probability of σ to obtain the masked feature D^u_{mask}:

$$D^u_{\text{mask}} = \text{Mask}(D^u, \sigma), \tag{3}$$

where 'Mask' is the operation that applies masking to the features, and $\sigma < 1.0$ controls the masking ratio. A video decoder D^u is used to reconstruct the video U' from $\mathcal{D}^u_{\text{mask}}$:

$$U' = \mathcal{D}^u(D^u_{\text{mask}}). \tag{4}$$

We compute the reconstruction loss between the original video U and the reconstructed video U' as follows:

$$\mathcal{L}^{\text{rec}}_{\text{video}} = ||U' - U||^2_2. \tag{5}$$

Similarly, we introduce a reconstruction loss for motion data, $\mathcal{L}^{\text{rec}}_{\text{motion}}$, following the same process. The total reconstruction loss is defined as:

$$\mathcal{L}^{\text{rec}} = \mathcal{L}^{\text{rec}}_{\text{video}} + \mathcal{L}^{\text{rec}}_{\text{motion}}. \tag{6}$$

The above masked reconstruction loss ensures that the encoded features contain sufficient information about the video or motion input, while learning away possible nuisances through masking and making the learning of the concepts (codebook) more focused on critical information. On the other hand, the discretization imposed by the concepts further enhances the meaningfulness and robustness of the pre-trained multi-modal (motion) features.

Velocity Reconstruction. Reconstructing inputs from masked features inevitably results in the loss of high-frequency details, leading to temporally over-smoothed reconstruction outcomes and preventing the discovery of more nuanced yet useful motion concepts.

To address this limitation, we propose an objective that further reconstructs the "velocity" of the input sequence. To be more specific, we define "state" as the encoding or feature of a frame in the video or motion sequence, while the "velocity" means the difference between consecutive "states" that captures high-frequency information by focusing on dynamic changes. For video and motion data, we denote the "states" as U and m, and the "velocity" as the optical flow O (computed from image frames) and delta motion δM, respectively.

However, directly reconstructing velocity from the same set of masked features introduces difficulty in figuring out the dynamic information, given the above-mentioned over-smoothing effect in the encoded features. To mitigate this, we draw inspiration from the robotic manipulation literature [46] on action concept discovery and incorporate two

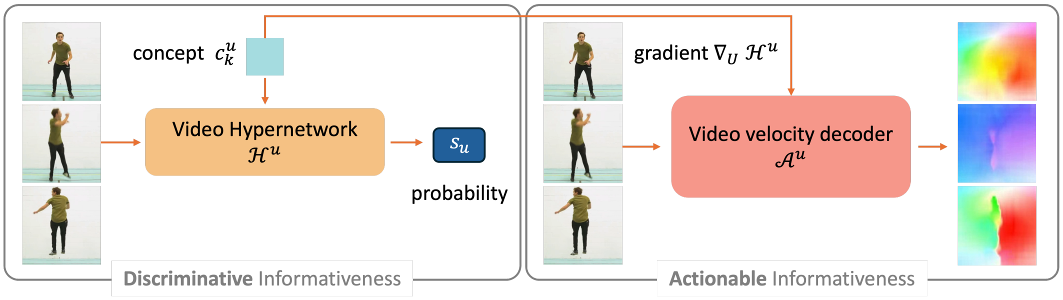

Figure 3. **Overview of the velocity reconstruction components.** We build similar network structures for both video and motion, and we present the video part in this figure. This module is composed of two learning objectives. 1) Discriminative informativeness (left) aims to improve the distinctiveness of encoded features by reducing representational ambiguity. 2) Actionable informativeness (right) focuses on reconstructing the velocity by leveraging gradient information from the discrimination hypernetwork. As for the video data, we employ optical flow as the representation of the velocity.

tailored learning objectives – *discriminative informativeness* and *actionable informativeness* – to enhance the efficiency of representations and improve the quality of temporal dynamics in multi-modality pretaining, as shown in Fig. 3.

Discriminative informativeness. The discriminative informativeness objective aims to improve the distinctiveness of encoded features by increasing the correlation between the discovered concepts and the corresponding state features. *Specifically,* we introduce a video hyper-network $\mathcal{H}^u$, which takes discrete features c_k^u from the codebook C^u as input and generates a classification network. The hyper-network $\mathcal{H}^u$ (together with the generated classifier) is trained to determine whether an input state is compatible with its corresponding code class, which is represented by a confidence value s_u (classification score) from 0 to 1, thereby enhancing the distinguishability and separation of each feature class in the representation space.

The discriminative loss for video is defined as:

$$\mathcal{L}_{\text{video}}^{\text{dis}} = \sum_{k=0}^{|C^u|-1} \sum_{i=0}^{T-1} CE(\mathcal{H}^u(c_k^u, u_i), d_i^u), \tag{7}$$

where $|C^u|$ is the size of the codebook C^u, $CE(\cdot)$ is the cross-entropy loss for classification, d_i^u represents the "ground-truth label" for the input state u_i determined by the VQ-VAE, and T is the number of frames.

Actionable informativeness. The actionable informativeness objective focuses on reconstructing the velocity by leveraging gradient information from the discriminative hyper-network, $\mathcal{H}^u$ introduced above. Inspired by the idea that gradients of a discriminator function can inform the state changes, as observed in [46], this objective ensures that the model captures not only static state information, but also the dynamic changes in the sequence. By leveraging gradient fluctuations around a given state, the approach effectively captures dynamic motion details, such as velocity, resulting in more fine-grained motion concepts, and enriching the motion representation.

Specifically, we initialize a network $\mathcal{A}^u$ for video data, which takes key image frames U and gradients from the discriminator $\nabla_U \mathcal{H}^u(D^u, U)$ as inputs and predicts the velocity V_u'. The actionable informativeness loss for video is then computed as:

$$\mathcal{L}_{\text{video}}^{\text{act}} = ||V_u - V_u'||_2^2. \tag{8}$$

where V_u represents the ground-truth velocity of video frames. This process is repeated for motion data to compute the corresponding loss $\mathcal{L}_{\text{motion}}^{\text{act}}$ using V_m and V_m'.

Combined Objectives. By integrating both video and motion components, we derive the total velocity reconstruction loss for human motion concept discovery and motion feature pre-training as in the following:

$$\mathcal{L}^{\text{dis}} = \mathcal{L}_{\text{video}}^{\text{dis}} + \mathcal{L}_{\text{motion}}^{\text{dis}} \tag{9}$$

$$\mathcal{L}^{\text{act}} = \mathcal{L}_{\text{video}}^{\text{act}} + \mathcal{L}_{\text{motion}}^{\text{act}} \tag{10}$$

With the above objectives, our approach ensures that the model effectively reconstructs high-frequency details, improving the temporal fidelity and overall quality of multimodal representations.

Feature Alignment. Multi-modality models often leverage the shared information across different modalities to enhance performance [5, 56, 87]. In the case of motion and video, video data provides contextual information that motion data alone may be lacking, which is critical for accurate motion prediction and understanding. Conversely, motion data focuses on human-centric features in the video, reducing over-reliance on environmental details. An efficient utilization or balancing of both is beneficial for the human motion LLMs. *However,* prior work on human motion and video understanding [5] only achieves implicit alignment through paired data during LLM fine-tuning, without explicit alignment in encoder training. To address this issue, we propose an alignment loss to improve information sharing and cross-modal coherence.

To achieve alignment between video and motion features, we collect a set of video-motion pairs from Motion-X [41], which are aligned by frames. We instantiate two projection layers: The video projection layer $\mathcal{P}^u$ and the motion projection layer $\mathcal{P}^m$. These layers are designed to map the respective features from the video and motion domains into a shared aligned space.

More explicitly, given paired video and motion data $\{U, M\}$, we use $\mathcal{P}^u$ and $\mathcal{P}^m$ to project the discrete video feature d_i^u and motion feature d_j^m into the aligned feature space, resulting in a_i^u and a_j^m, where $a_i^u, a_j^m \in \mathbb{R}^H$. Here, H denotes the dimensionality of the aligned features in the shared space.

To ensure effective alignment, we define an alignment loss as follows:

$$\mathcal{L}^{\text{align}} = \sum_{0}^{T-1} \frac{\text{norm}(a_i^u) \cdot \text{norm}(a^m_{\text{align}(i)})/\epsilon}{\sum_0^{K-1} \text{norm}(a_i^u) \cdot \text{norm}(a_j^m)/\epsilon}, \tag{11}$$

where $\text{align}(i)$ indicates the pose index in the motion sequence corresponding to the key video frame u_i, norm represents a normalization operation to stabilize feature magnitudes and ϵ is a temperature parameter that controls the sharpness of the alignment distribution.

Note that $\mathcal{L}^{\text{align}}$ is computed only when motion and video are paired. For standalone motion or video inputs, this loss term is omitted for applicability.

The **overall** human motion concept discovery objective is defined as:

$$\begin{aligned}\mathcal{L}^{\text{total}} = &\mathcal{L}^{\text{rec}}_{\text{motion}} + \mathcal{L}^{\text{rec}}_{\text{video}} + \\ &\lambda^{\text{dis}}\mathcal{L}^{\text{dis}} + \lambda^{\text{act}}\mathcal{L}^{\text{act}} + \lambda^{\text{align}}\mathcal{L}^{\text{align}},\end{aligned} \tag{12}$$

where λ^{dis}, λ^{act} and λ^{align} are weighting coefficients for the respective losses $\mathcal{L}^{\text{dis}}$, $\mathcal{L}^{\text{act}}$ and $\mathcal{L}^{\text{align}}$.

3.3. Modality Translation

Although pre-trained encoders capture rich features from the original input, these features are often difficult for LLMs to interpret directly [5, 37, 40]. *Therefore,* it is essential to introduce an additional projection layer to map these features into a space that LLMs can effectively understand.

As illustrated in Fig 2, we add tunable translation layers after the encoders $\mathcal{E}^u$ and $\mathcal{E}^m$. These layers are trained using annotated video and motion sequence data to learn task-specific transformations. During this stage, the encoders and the LLM $\mathcal{F}$ are kept frozen to prevent adverse feature drift, and only the weights of the translation layers are updated. This strategy preserves the pre-trained knowledge of the encoders and language model while optimizing the feature mapping for downstream tasks.

3.4. Multi-Modality Instruction Tuning

Finally, we employ instruction tuning to train the language model $\mathcal{F}$ to respond to user queries using the corresponding video or motion sequences as context. During this process, the encoders and projection layers remain frozen, ensuring their pre-trained features are retained, while the weights of the language model $\mathcal{F}$ are updated to accommodate the task.

In this stage, the model can further enhance cross-modal alignment using paired motion-video data in an implicit manner. To streamline the fine-tuning process, we adopt LoRA [29] with a rank of 8, allowing efficient and lightweight adjustment to the language model.

4. Experiments

4.1. Experiments Settings

Implementation details For videos and motions, we initialize a codebook with 512 codes, a one-layer projector for feature alignment, and a 2-layer MLP to generate the weights for a 2-layer discriminator separately. For video inputs, we utilize pre-trained SigLIP [79] with an image patch size of 16 pixels and hidden dimension of 256 as the video encoder, and we take a 12-layer transformer as the reconstruction decoder. A 4-layer transformer is utilized as the motion encoder and a 2-layer transformer is used for motion reconstruction. For the velocity reconstruction, we initialize a two-layer hypernet and two-layer velocity decoder. Besides, we utilize one-layer MLPs for feature alignment. We pre-train the motion encoder with only the reconstruction loss for 60K iterations with a learning rate of 1e-4. Then, we train both the video and motion encoders with the total loss for 8K iterations with a learning rate of 1e-5.

We utilize Llama3.1-8B as our base LLM model. For the modality translation stage, we initialize a one-layer projector for the video and motion encoder separately. We train the projectors on the annotation dataset for one batch with a batch size of 32 and a learning rate of 1e-4. In the Multi-modality instruction tuning stage, we train the LLM model with a batch size of 16 and a learning rate of 3e-4.

Training datasets For the encoder pre-training stage, we utilize Motion-X [41] datasets which contain 81K motion sequences and 36K motion-video pairs. In order to facilitate the training process of LLMs, we follow MotionLLM [5] to prompt LLMs to generate 20K annotation for motion-video pairs from the Motion-X dataset, and 10K annotation data from the HumanML3D [21] dataset. This data is utilized for modality translation. We additionally utilize the VATEX [69] dataset which contains 20K annotated videos for video translator training. For multi-modality instruction

Table 1. **Comparison of different methods on BABEL-QA test set.** The pred type represent the prediction type, including classification and generaration. For 2s-AGCN and MotionCLIP, 'M' means MLP, 'R' means RNN. Our method outperforms the baselines across all the metrics.

Model	Pred type	Overall	Query type			Relation Type		
			Action	Direction	BodyPart	Before	After	Other
2s-AGCN-M	cls.	0.355	0.384	0.352	0.228	0.331	0.264	0.295
2s-AGCN-R	cls.	0.357	0.396	0.352	0.194	0.337	0.301	0.285
MotionCLIP-M	cls.	0.430	0.485	0.361	0.272	0.372	0.321	0.404
MotionCLIP-R	cls.	0.420	0.489	0.310	0.250	0.398	0.314	0.387
MotionLLM	gen.	0.436	0.517	0.354	0.154	0.427	0.368	0.529
Ours	gen.	**0.711**	**0.809**	**0.697**	**0.623**	**0.707**	**0.635**	**0.797**

tuning, we follow MotionLLM to generate 200K Q&A pairs for both motions and videos. We also utilize the Video-ChatGPT [49] and BABEL-QA [13] datasets to enhance the motion and video understanding ability of our model, which has 100K and 2K Q&A data separately.

Evaluation datasets Although MotionLLM [5] creates a large benchmark for both motion and video understanding, by the time we completed our paper, they had not released the specific details or access methods for the benchmark, nor did they provide a clear timeline for its release. Therefore, we did not perform comparisons on that benchmark but instead selected the test set of ActivityNet-QA [78] and BABEL-QA [13] to separately validate the performance of our algorithm on video and motion data.

Evaluation metrics On the BABEL-QA benchmark, we evaluate the prediction accuracy for evaluation. As for the classification-based baselines, we evaluate the accuracy directly. As for the generative baselines, we adapt the evaluation protocol used in previous works [5, 37] by utilizing LLMs. Specifically, the evaluation LLM is given some examples of questions, answers, and scores (from 0 to 1), and it is required to follow these examples and score the experiment results. On the Activity-QA benchmark, we only compare our methods with our generative baselines. Thus, we directly utilize LLMs to evaluate the methods.

4.2. Quantitative Results

Experiments on ActivityNet-QA. We evaluate the video understanding ability of our method and other baselines on the ActivityNet-QA benchmark as shown in Tab. 2. We follow previous methods [5, 33, 38] to report the accuracy and scores that are evaluated by LLMs. We show in Tab. 2 that our method outperforms all the baselines on both metrics. Compared to the video-only method, our method aligns the video feature with the motion feature, making the encoder concentrate on human-centric information. Compared to MotionLLM, our encoder pre-training helps the network to learn more fine-grained information through velocity reconstruction.

Experiments on BABEL-QA. We evaluate the motion understanding ability of our method as compared to existing studies on the BABEL-QA [13] dataset. Following [5, 13], we compare our method with a diverse set of baselines: 1) **2s-AGCN** [58], a method using 2s-GCN to extract motion features and classify motion by MLP or RNN; 2) **MotionCLIP** [63], a method that aligns the features of motion and text using CLIP [55] and predicts motion class by MLP or RNN; 3) **MotionLLM** [5], an LLM-based method which has motion reasoning ability. We follow [5] to fine-tune our method with the training set of BABEL-QA and evaluate our method with LLMs.

From Tab. 1, our method outperforms other methods in all metrics by a large margin of nearly 60% overall. In the "bodypart" query task, our algorithm achieves 2.29 times the score of the second-place method, MotionCLIP-M (0.623 vs. 0.272). This advantage stems partly from our velocity-based encoder training, which, compared to traditional motion feature extraction, helps us better capture high-frequency information and ensures the learning of fine-grained features. Additionally, unlike MotionLLM, we perform alignment at the encoder pre-training stage, ensuring that the features learned by the encoder are inherently more semantically rich.

Table 2. **Experiments results on ActivityNet-QA benchmark.**

Model	Acc↑	Score ↑
FrozenBiLM [57]	24.7	-
VideoChat [38]	-	2.2
LLaMA-Adapter [84]	34.2	2.7
Video-LLaMA [80]	12.4	1.1
Video-ChatGPT [49]	35.2	2.7
Video-LLaVA [40]	45.3	3.3
VideoChat2 [39]	49.1	3.3
MotionLLM [5]	53.3	3.5
Ours	**54.2**	**3.6**

Table 3. Ablation study on BABEL-QA test set.

Model	Pred type	Overall	Query type			Relation Type		
			Action	Direction	BodyPart	Before	After	Other
MotionLLM	gen.	0.436	0.517	0.354	0.154	0.427	0.368	0.529
Ours-w/o$\mathcal{L}^{rec}$	gen.	0.696	0.741	0.645	0.577	0.600	0.597	0.762
Ours-w/o$\mathcal{L}^{dis}$&$\mathcal{L}^{act}$	gen.	0.637	0.693	0.478	0.606	0.667	0.526	0.709
Ours-w/o$\mathcal{L}^{align}$	gen.	0.675	0.743	0.579	0.523	0.584	0.570	0.743
Ours	gen.	**0.711**	**0.809**	**0.697**	**0.623**	**0.707**	**0.635**	**0.797**

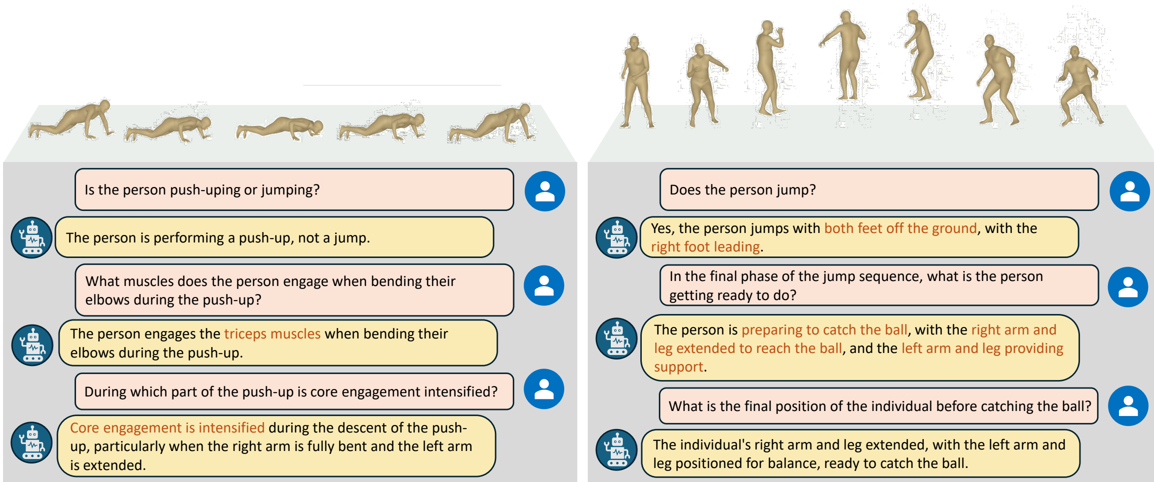

Figure 4. **Example of motion understanding.** We validate the proposed model's ability to understand motion from multiple aspects. Q1 on the left and Q1–3 on the right evaluate its comprehension of motion sequences, while Q2 and Q3 on the left push it to analyze kinematic, kinesthetic, and physical properties. These results demonstrate our algorithm's ability to recognize, describe in detail, and analyze motions.

Ablation study. We present the ablation results in Tab. 3. We conducted ablation studies on masked reconstruction, velocity reconstruction, and feature alignment. MotionLLM serves as an ablation baseline without any of these techniques. As presented, the absence of any single module leads to a decrease in performance. Among these, masked reconstruction has the least impact, as we initially train the motion encoder solely with masked reconstruction loss to ensure stable subsequent training. Removing velocity reconstruction has the greatest impact on pexrformance since velocity provides high-frequency information that other modules lack. Likewise, removing the alignment loss significantly affects results, highlighting the importance of multi-modality alignment in motion understanding.

4.3. Qualitative Results

To validate our algorithm's ability to understand motions and its contextual knowledge, we conducted multiple rounds of Q&A with the motion model. We present two examples in Fig. 4. We discover that, beyond its basic motion recognition capabilities, our algorithm can provide detailed descriptions of motions. In the example on the right, it can precisely describe the actions of the left hand and foot, as well as the right hand and foot. Our algorithm also possesses extensive knowledge about various movements. For instance, when asked, "what muscles does the person engage when bending their elbows during a push-up?" it can accurately respond with "triceps muscles." This demonstrates that our algorithm can understand the input motion sequences and perform corresponding analyses. We present more qualitative results in the supplementary material.

5. Conclusion

In this paper, we presented HuMoCon, a novel motion-video understanding framework for human action understanding and concept discovery. We demonstrated the effectiveness of our approach through comprehensive experiments, showcasing major improvements over state-of-the-art methods. Our work provides a strong foundation for future research in human action understanding and motion concept discovery. There are still some drawbacks of our work, such as the lack of human-environment contact or understanding niche motions. We plan to address those limitations by expanding the framework to incorporate additional modalities, particularly as environments, objects, and contact, and curate a wider range of specialized datasets.

Acknowledgment

This work is supported by the Early Career Scheme of the Research Grants Council (grant # 27207224), the HKU-100 Award, a donation from the Musketeers Foundation, and an Academic Gift from Meta.

References

[1] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *2022 International Conference on 3D Vision (3DV)*, pages 414–423. IEEE, 2022. 2

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. 3

[3] Yang Bai, Desen Zhou, Songyang Zhang, Jian Wang, Errui Ding, Yu Guan, Yang Long, and Jingdong Wang. Action quality assessment with temporal parsing transformer. In *European conference on computer vision*, pages 422–438. Springer, 2022. 1, 3

[4] Pablo Cervantes, Yusuke Sekikawa, Ikuro Sato, and Koichi Shinoda. Implicit neural representations for variable length human motion generation. In *European Conference on Computer Vision*, pages 356–372. Springer, 2022. 1, 2

[5] 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. 1, 2, 3, 5, 6, 7

[6] Steven Chen and Richard R Yang. Pose trainer: correcting exercise posture using pose estimation. *arXiv preprint arXiv:2006.11718*, 2020. 1, 3

[7] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023. 3

[8] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36:72842–72866, 2023. 3

[9] Ginger Delmas, Philippe Weinzaepfel, Francesc Moreno-Noguer, and Grégory Rogez. Posefix: correcting 3d human poses with natural language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15018–15028, 2023. 2

[10] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In *2005 IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance*, pages 65–72. IEEE, 2005. 3

[11] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015. 1, 2

[12] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3

[13] Mark Endo, Joy Hsu, Jiaman Li, and Jiajun Wu. Motion question answering via modular motion programs. In *International Conference on Machine Learning*, pages 9312–9328. PMLR, 2023. 2, 7

[14] Qihang Fang, Yingda Yin, Qingnan Fan, Fei Xia, Siyan Dong, Sheng Wang, Jue Wang, Leonidas J Guibas, and Baoquan Chen. Towards accurate active camera localization. In *European Conference on Computer Vision*, pages 122–139. Springer, 2022. 1

[15] Qihang Fang, Chengcheng Tang, Bugra Tekin, and Yanchao Yang. Cigtime: Corrective instruction generation through inverse motion editing. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 2

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 2, 3

[17] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. Chatpose: Chatting about 3d human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2093–2103, 2024. 2

[18] Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9919–9928, 2021. 1, 3

[19] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 3

[20] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 1, 2

[21] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022. 6

[22] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 2

[23] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022. 2

[24] Tianyu Guo, Hong Liu, Zhan Chen, Mengyuan Liu, Tao Wang, and Runwei Ding. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 762–770, 2022. 1, 2

[25] Tengda Han, Max Bain, Arsha Nagrani, Gul Varol, Weidi Xie, and Andrew Zisserman. Autoad ii: The sequel-who, when, and what in movie audio description. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13645–13655, 2023. 3

[26] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. Autoad iii: The prequel-back to the pixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18164–18174, 2024. 1, 3

[27] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 4

[28] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. 2

[29] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 6

[30] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023. 2

[31] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11107–11116, 2021. 2

[32] Yujiao Jiang, Qingmin Liao, Zhaolong Wang, Xiangru Lin, Zongqing Lu, Yuxi Zhao, Hanqing Wei, Jingrui Ye, Yu Zhang, and Zhijing Shao. Smplx-lite: A realistic and drivable avatar benchmark with rich geometry and texture annotations. *arXiv preprint arXiv:2405.19609*, 2024. 1

[33] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024. 7

[34] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3288–3297, 2017. 1, 2

[35] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8255–8263, 2023. 2

[36] Kosmas Kritsis, Aggelos Gkiokas, Aggelos Pikrakis, and Vassilis Katsouros. Danceconv: Dance motion generation with convolutional networks. *IEEE Access*, 10:44982–45000, 2022. 2

[37] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 2, 3, 6, 7

[38] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 7

[39] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 7

[40] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 1, 2, 3, 6, 7

[41] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[42] Lilang Lin, Jiahang Zhang, and Jiaying Liu. Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2363–2372, 2023. 1, 2

[43] Daochang Liu, Qiyue Li, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. Towards unified surgical skill assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9522–9531, 2021. 1, 3

[44] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 816–833. Springer, 2016. 1, 2, 3

[45] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017. 1, 2

[46] Ruizhe Liu, Qian Luo, and Yanchao Yang. Infocon: Concept discovery with generative and discriminative informativeness. *arXiv preprint arXiv:2404.10606*, 2024. 2, 4, 5

[47] Thomas Lucas, Fabien Baradel, Philippe Weinzaepfel, and Grégory Rogez. Posegpt: Quantization-based 3d human motion generation and forecasting. In *European Conference on Computer Vision*, pages 417–435. Springer, 2022. 2

[48] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 3

[49] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video

understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 7

[50] Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. Pg-video-llava: Pixel grounding large video-language models. *arXiv preprint arXiv:2311.13435*, 2023. 1, 3

[51] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 304–313, 2019. 1, 3

[52] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021. 1, 2

[53] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 2

[54] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 722–731, 2021. 2

[55] 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. 2, 3, 7

[56] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, Christoph Feichtenhofer, and Jitendra Malik. On the benefits of 3d pose and tracking for human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 640–649, 2023. 2, 5

[57] Leonard Salewski, Stefan Fauth, A Koepke, and Zeynep Akata. Zero-shot audio captioning with audio-language model guidance and audio context keywords. *arXiv preprint arXiv:2311.08396*, 2023. 3, 7

[58] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019. 1, 2, 7

[59] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Transactions on Graphics*, 38(6):178, 2019. 1, 2

[60] Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. Local motion phases for learning multi-contact character movements. *ACM Transactions on Graphics (TOG)*, 39 (4):54–1, 2020. 2

[61] Sebastian Starke, Ian Mason, and Taku Komura. Deepphase: Periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 1, 2

[62] Graham W Taylor, Rob Fergus, Yann LeCun, and Christoph Bregler. Convolutional learning of spatio-temporal features. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI 11*, pages 140–153. Springer, 2010. 2, 3

[63] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 2, 3, 7

[64] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2

[65] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2, 3

[66] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 588–595, 2014. 3

[67] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013. 3

[68] Jingbo Wang, Ye Yuan, Zhengyi Luo, Kevin Xie, Dahua Lin, Umar Iqbal, Sanja Fidler, and Sameh Khamis. Learning human dynamics in autonomous driving scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20796–20806, 2023. 1

[69] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4581–4591, 2019. 6

[70] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 2, 3

[71] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024. 2, 3

[72] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. *arXiv preprint arXiv:2309.07918*, 2023. 1

[73] Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. Autoad-zero: A training-free framework for zero-shot audio description. *arXiv preprint arXiv:2407.15850*, 2024. 3

[74] Guo-Hao Xu, Yi-Lin Wei, Dian Zheng, Xiao-Ming Wu, and Wei-Shi Zheng. Dexterous grasp transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17933–17942, 2024. 2

[75] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 1, 3

[76] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 1, 2

[77] Zeshi Yang, Kangkang Yin, and Libin Liu. Learning to use chopsticks in diverse gripping styles. 2022. 2

[78] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134, 2019. 2, 7

[79] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 6

[80] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 3, 7

[81] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. *arXiv preprint arXiv:2301.06052*, 2023. 2

[82] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[83] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8552–8562, 2022. 3

[84] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 7

[85] Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18243–18252, 2024. 3

[86] Zixiang Zhou, Yu Wan, and Baoyuan Wang. Avatargpt: All-in-one framework for motion understanding planning generation and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1357–1366, 2024. 2

[87] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023. 2, 3, 5