









Generative Hierarchical Temporal Transformer for Hand Pose and Action Modeling

Yilin Wen^{1,2} , Hao Pan³ , Takehiko Ohkawa² , Lei Yang^{1,4} , Jia Pan^{1,4} ,
Yoichi Sato² , Taku Komura¹ , and Wenping Wang⁵ 

¹ The University of Hong Kong

² The University of Tokyo

³ Microsoft Research Asia

⁴ Centre for Garment Production Limited, Hong Kong

⁵ Texas A&M University

Abstract. We present a novel unified framework that concurrently tackles recognition and future prediction for human hand pose and action modeling. Previous works generally provide isolated solutions for either recognition or prediction, which not only increases the complexity of integration in practical applications, but more importantly, cannot exploit the synergy of both sides and suffer suboptimal performances in their respective domains. To address this problem, we propose a generative Transformer VAE architecture to model hand pose and action, where the encoder and decoder capture recognition and prediction respectively, and their connection through the VAE bottleneck mandates the learning of consistent hand motion from the past to the future and vice versa. Furthermore, to faithfully model the semantic dependency and different temporal granularity of hand pose and action, we decompose the framework into two cascaded VAE blocks: the first and latter blocks respectively model the short-span poses and long-span action, and are connected by a mid-level feature representing a sub-second series of hand poses. This decomposition into block cascades facilitates capturing both short-term and long-term temporal regularity in pose and action modeling, and enables training two blocks separately to fully utilize datasets with annotations of different temporal granularities. We train and evaluate our framework across multiple datasets; results show that our joint modeling of recognition and prediction improves over isolated solutions, and that our semantic and temporal hierarchy facilitates long-term pose and action modeling.

Keywords: Hand pose action modeling · recognition and future prediction · temporal regularity · semantic and temporal hierarchy · hand pose estimation · hand action recognition · hand motion prediction

1 Introduction

Understanding dynamic hand poses and actions is fundamental in fields such as human-robot interaction and VR/AR applications. In recent years, huge progress

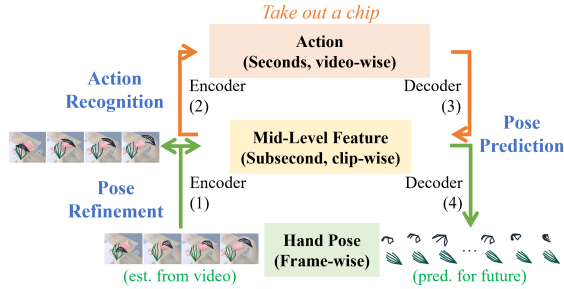


Fig. 1: Jointly modeling recognition and prediction, while following the semantic dependency and temporal granularity for hand pose-action. For recognition, (1)→(2) moves up from short to long spans for input pose refinement and action recognition respectively. For motion prediction, two paths are available: (1)→(4) exploits short-term motion regularity, and (1)→(2)→(3)→(4) enables long-term action-guided prediction.

has been made in recognizing 3D hand poses and actions (*e.g.* take out a chip) from inputs such as RGB videos [12, 20, 43, 52, 54, 59]. Meanwhile, another line of research [2, 26, 27] focuses on predicting future hand motion represented as a sequence of frame-wise poses, where recent research [29] further employs generative models to achieve diverse motion prediction conditioned on a given action.

However, existing literature provides isolated solutions working on either the recognition [12, 20, 43, 52, 54, 59] or prediction [2, 26, 27, 29] side, which brings deployment complexity when integrating both sides in practical applications. Furthermore, we note that recognition and prediction tasks are naturally synergized by the temporal regularities shared through observed and future timestamps: As exemplified in Fig. 1, given the observation of hand poses reaching into the can, which indicates the action of taking out a chip, one can predict that future hand motion will describe grabbing and pulling out a chip, thus completing the action. However, isolated solutions cannot fully exploit this temporal regularity, resulting in a tendency to overfit to specific data distributions and suffer from suboptimal performances in their respective domains (Sec. 4.3).

We present a framework with a generative Transformer VAE architecture to jointly capture both recognition and future prediction for hand pose and action modeling, therefore addressing various tasks including input 3D hand pose refinement, action recognition, and future 3D hand motion prediction. Our transformer encoder and decoder respectively produce outputs for recognition and prediction, while the VAE latent code connecting the two forces the extraction of regular and consistent hand motion and action, by predicting the future from the past and vice versa. In this way, we synergize recognition and prediction tasks and enhance the performance compared to isolated solutions (Sec. 4.3).

Moreover, when it comes to pose-action modeling, extensive literature has shown the benefits of capturing the semantic dependency between the instantaneous pose and action over seconds. For example, [15, 29, 37, 38, 48] generate motion by conditioning on action to enhance realism. On the recognition

side, [9, 47, 52, 54] aggregates the frame-wise hand poses estimated throughout the video to recognize the performed action. Besides modeling the semantic dependency, Wen *et al.* [52] further emphasize capturing the different temporal granularity of hand pose and action. Their framework, namely *Hierarchical Temporal Transformer*, has two cascaded encoders that capture short and long time spans respectively for effective hand pose estimation and action recognition.

In addition to our jointly modeling of both recognition and prediction, we are inspired by [52] to introduce block cascades, thus faithfully respecting the semantic dependency and temporal granularity of hand pose-action. To this end, we decompose our framework into two cascaded blocks that have the same generative transformer VAE structure but focus on different semantic and temporal granularities, thus naming our framework *Generative Hierarchical Temporal Transformer* (G-HTT): The lower pose block (**P** block) models hand poses over short time spans, and the upper action block (**A** block) models action over long time spans. A middle-level representation is further introduced to connect the two blocks: it is simultaneously the pose block VAE latent code and the action block encoder input/decoder output, and semantically encodes clip-wise motion over a subsecond span (Fig. 1).

This decomposition into block cascades offers two key advantages: First, it decouples the complex motion generation into hierarchical subtasks to respectively capture short-term and long-term temporal regularity, which improves over flattened models (Sec. 4.4). Second, it brings training flexibility, as we can train the blocks separately, which not only reduces training computational cost but also allows for using datasets of different annotation granularities (Secs. 3.3 and 4.5).

We train and evaluate the framework across different datasets of two-hand interactions, including H2O [23] for daily activities, Assembly101 [42] and AssemblyHands [36] for (dis-)assembling take-part toys. At test time, given a 3D hand pose sequence (*e.g.* per-frame estimations from the observed RGB video), we first refine it by leveraging the short-term hand motion regularity, (Fig. 1, (1)). Next, we aggregate the clip-wise motions for action recognition (Fig. 1, (1)→(2)). Finally, we decode the observed motions and action into a sequence of future middle-level features for motion prediction (Fig. 1, (1)→(2)→(3)→(4)). Evaluation results across datasets show that our framework can solve recognition problems from various camera views, and generate plausible future hand poses over time. The contribution of this paper can be summarized as follows:

- A generative Transformer VAE architecture to concurrently capture recognition and future prediction for hand pose and action modeling, which exploits the temporal regularity synergized between the past and the future, thus improving over isolated solutions.
- A hierarchical architecture composed of two cascaded generative blocks, which models semantic dependency and temporal granularity of pose-action. This block cascade facilitates capturing both short-term and long-term temporal regularities, and further brings training flexibility.

- A comprehensive evaluation of the system on tasks such as 3D hand pose refinement, action recognition, and 3D hand motion prediction, validating the performance and design of our framework.

2 Related Works

Action Recognition and 3D Hand Pose Estimation Massive literature addresses perceiving hand pose and action from visual observation. For example, a series of works aims to recover the 3D hand skeleton or mesh from the visual input, where the spatial correlation within a single-frame is well exploited [20, 24, 34, 46, 55, 59], and the motion coherence along the temporal dimension is further leveraged to improve robustness under occlusion and truncation [3, 10, 16, 17, 35, 51]. Meanwhile, [5, 11–13, 43] focus on the higher semantic level, where they extract the spatial-temporal feature from the input frames to recognize the semantic hand or body action.

Moreover, many works notice and exploit the benefits of modeling the semantic dependency between hand pose and action, since intuitively action is defined by the pattern of hand motion (*i.e.* verb) and object in manipulation (*i.e.* noun). For example, [9, 25, 32, 45, 47, 52, 54] leverage the hand pose features for action recognition, while Yang *et al.* [54] further refer to the action feature for pose refinement. Wen *et al.* [52] further stress capturing the respective temporal granularity of pose and action when exploiting temporal cues, and propose a framework with two cascaded blocks to respectively work on short- and long-term spans and output per-frame 3D hand pose and video action.

The hierarchical structure of our framework is inspired by [52], but we have extended it to model prediction tasks, which not only covers more tasks but also enhances recognition performance (Sec. 4.3).

3D Human Hand and Body Motion Prediction Previous works predict the 2D or 3D trajectory of hand roots [2, 26, 27, 30] or skeleton [8] from the observed hand motion. On the other hand, [1, 4, 21, 29, 31, 33, 49, 56, 57] capture the distribution of future body motion with powerful generative deep neural networks. Motion prediction can also benefit from semantic dependency modeling, as achieved by taking the past motion together with a specified action as condition, based on cVAEs [4, 33], GPT-like models [21, 29, 57] and diffusion models [49]. For example, PoseGPT [29] first quantizes short motion clips into latent codes by training a VQ-VAE, and then constructs a GPT-like auto-regressive model for motion generation, which learns on sequences of action and latent motion tokens.

Our work builds a hierarchical structure for motion prediction, where consistencies in both short-term motion and long-term action are explicitly ensured through the cascade of generative Transformer VAEs (Sec. 4.4). In addition, we learn prediction and recognition simultaneously, which improves both tasks by exploiting the shared temporal regularity (Sec. 4.3).

Bridging Recognition and Prediction There are previous attempts to bridge recognition and prediction, therefore benefiting recognition or prediction at the

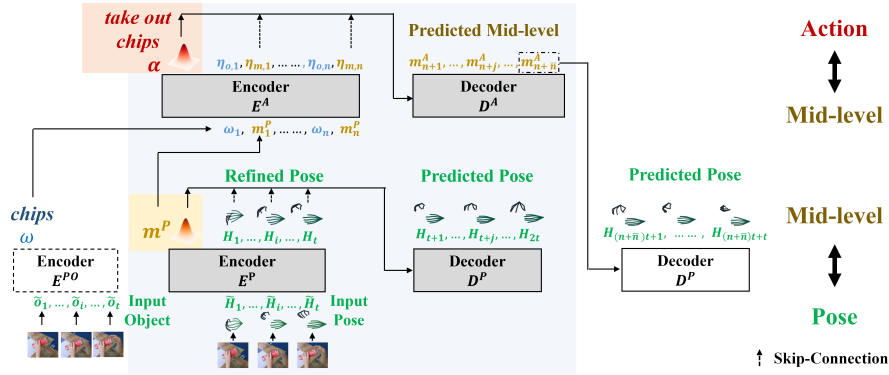


Fig. 2: Overview of our framework. The cascaded **P** and **A** (shaded in blue) of G-HTT jointly model recognition and prediction, and faithfully respect the semantic dependency and temporal granularity among pose, mid-level and action (Sec. 3.1).

pose or action level. For example, [40, 44] learn next-frame pose prediction with cVAEs to model a latent space depicting pose transitions along the temporal dimension. Their learned latent space then serves as a strong regulation in test-time optimization for body pose estimation. [29, 37, 38, 48, 49] learn text-guided motion generation models, with the text guidance in the form of prescribed actions. The generated sequences can then be used as training data for recognition tasks. On the other hand, [7, 14, 50, 53, 58] leverage per-frame prediction for understanding high-level action, benefiting tasks such as action anticipation [14, 50] or early action detection [7, 53, 58].

In comparison, our hierarchical modeling enables capturing both recognition and prediction by modeling the semantic hierarchy between short-term pose and long-term action, which significantly boosts computational efficiency, pose/action estimation accuracy, and long-term generation fidelity. This is not considered by existing works.

3 Methods

The core framework, namely *Generative Hierarchical Temporal Transformer* (G-HTT, Fig. 2), takes as input the object in manipulation and the observed pose sequence of T frames for two interacting hands, where the hand motion and object feature respectively depict the *verb* and *noun* of the action being performed (*e.g.* take out a chip). G-HTT then jointly models both recognition and prediction, while following the semantic-temporal hierarchy of pose-action that captures their dependency and different temporal granularities (Sec. 3.1). In the test stage, we apply G-HTT to recognition tasks of input pose refinement and action recognition, and to the generation task of diverse hand motion prediction (Sec. 3.2). Important implementation details are given in Sec. 3.3, and a table of notations is provided in the supplementary for reference.

3.1 Joint Modeling of Recognition and Prediction with Semantic-Temporal Hierarchy

G-HTT consists of two cascaded blocks, namely the short-term pose block \mathbf{P} and the long-term action block \mathbf{A} , to jointly model recognition and prediction while following the hierarchy of temporal and semantic granularity for pose-action. Both \mathbf{P} and \mathbf{A} have the same VAE structure, with their encoders and decoders respectively outputting for recognition and future motion prediction, but \mathbf{P} and \mathbf{A} model different semantic levels and time spans (Fig. 2).

To bridge the pose and action blocks in the semantic-temporal hierarchy, we explicitly introduce a mid-level feature \mathbf{m} , which represents the hand poses within a sub-second time span. \mathbf{P} and \mathbf{A} then respectively model the mappings between pose *vs.* mid-level, and mid-level *vs.* action. As the two blocks are cascaded, the different semantic levels can refer to each other for globally consistent recognition and prediction (Sec. 3.2). Moreover, our design enables a flexible training scenario, where \mathbf{P} and \mathbf{A} can be decoupled and trained separately based on their respective supervision signals and training data (Sec. 3.3).

P-Block takes a subsecond time span of t ($t < T$) consecutive frames to model the relationship between per-frame hand pose and mid-level feature \mathbf{m}^P , without explicitly leveraging the action information. The mid-level \mathbf{m}^P is learned to be the latent bottleneck of \mathbf{P} , which encodes the input t consecutive frames of hand poses, and is decoded to hand motion of the future t frames. Meanwhile, similar to HTT [52], the input hand poses can be refined via the encoder \mathbf{E}^P by leveraging the short-term temporal regularity.

In detail, \mathbf{E}^P takes as input a sequence of $t + 2$ tokens $(\tilde{\mu}^P, \tilde{\Sigma}^P, \tilde{\mathbf{H}}_1, \dots, \tilde{\mathbf{H}}_t)$. $\tilde{\mathbf{H}}_i$ represents the per-frame hand pose, and $\tilde{\mu}^P, \tilde{\Sigma}^P \in \mathbb{R}^d$ are trainable tokens for parameterizing the distribution of \mathbf{m}^P by aggregating over $\tilde{\mathbf{H}}_{1:t}$, similar to [37, 38]. Denoting the output sequence of $(\mu^P, \Sigma^P, \mathbf{H}_1, \dots, \mathbf{H}_t)$, we obtain \mathbf{H}_i as the refined frame-wise hand poses, and sample \mathbf{m}^P with re-parameterization [22] from the normal distribution $N(\mu^P, \Sigma^P)$.

\mathbf{D}^P predicts the hand pose for the following t frames in a parallel manner. The mid-level \mathbf{m}^P , with $\mathbf{H}_{1:t}$ optionally concatenated as skip-connections, are referred by \mathbf{D}^P through cross-attention. As a parallel transformer decoder, the query input of \mathbf{D}^P are the sinusoidal position encoding of t tokens, and the output t tokens are mapped into the future hand motion $\mathbf{H}_{t+1:2t}$.

Denoting the corresponding GT motion as $\bar{\mathbf{H}}_{1:2t}$, \mathbf{P} is trained by a loss function consisting of three parts:

- The hand component loss to compare the refined and predicted motion with GT:

$$L_{comp} = \frac{1}{2t} \sum_{i=1}^{2t} \|\bar{\mathbf{H}}_i - \mathbf{H}_i\|_1 \quad (1)$$

- The root frame trajectory loss for the predicted part:

$$L_{trj} = \frac{1}{t} \sum_{i=t+1}^{2t} (\|\bar{s}_i^L - s_i^L\|_1 + \|\bar{s}_i^R - s_i^R\|_1) \quad (2)$$

with \bar{s}_i^L, \bar{s}_i^R denote the GT counterpart.

- The KL-loss L_{KL}^P for the regularity of \mathbf{m}^P , as the KL-divergence between $N(\mu^P, \Sigma^P)$ and the standard normal distribution.

The overall loss for \mathbf{P} sums them up: $L_P = \lambda_1 L_{comp} + \lambda_2 L_{trj} + \lambda_3 L_{KL}^P$.

A-Block models the relationship between the mid-level feature and action: it exploits the long-term time span to aggregate the sequence of mid-level features \mathbf{m}^P from the whole observation, and predicts a sequence of mid-level features \mathbf{m}^A for future timestamps, which are further expanded by \mathbf{D}^P into concrete motion. In addition to variational auto-encoding, **A** has its latent bottleneck feature also aligned with text embeddings of the action taxonomy, to enable action recognition of the observation and action-controlled prediction.

The encoder \mathbf{E}^A derives action from hand motion and object feature across the observation. Its input sequence concatenates the trainable tokens $\tilde{\mu}^A, \tilde{\Sigma}^A \in \mathbb{R}^d$ with the clip-wise mid-level $\mathbf{m}_{1:n}^P$ and object feature $\omega_{1:n}$ ($n = \lceil T/t \rceil$). Specifically, \mathbf{m}_i^P is the μ^P from \mathbf{E}^P ; ω_i is comparable to the CLIP [6, 39, 41] feature of object name, which is aggregated by an extra individual \mathbf{E}^{PO} from the per-frame object features (Sec. 3.3). We further add a sinusoidal phase encoding ϕ_i to \mathbf{m}_i^P and ω_i , which denotes the number of clips since the beginning of the performed action. Given μ^A, Σ^A output from \mathbf{E}^A , we follow $N(\mu^A, \Sigma^A)$ to re-parameterize and obtain the bottleneck latent feature α .

We then inject α into the decoder \mathbf{D}^A to enable action-controlled generation. For the cross attention of \mathbf{D}^A , we utilize α and optionally include the clip-wise feature obtained from \mathbf{E}^A for enhanced continuity. The parallel decoder \mathbf{D}^A takes the phase embeddings $\phi_{n+1:n+\bar{n}}$ as input, and outputs $\mathbf{m}_{n+1:n+\bar{n}}^A$ depicting the mid-level features of the future \bar{n} consecutive clips. One can further expand the predicted \mathbf{m}^A into concrete poses through \mathbf{D}^P , completing the cycle of long-term observation for long-term prediction, with consistency in both global action and local poses (Sec. 3.2, P.b).

To train **A**, besides the KL-loss L_{KL}^A , we constrain the bottleneck latent α by matching it with the embedding of action taxonomy $\mathcal{A} = \{a = \mathbf{FC}_1(\bar{a}) \in \mathbb{R}^d\}$, where \bar{a} are CLIP text embeddings [6, 39, 41] for action labels in the taxonomy of size N_A . Therefore, the action recognition loss is:

$$L_{action} = \sum_{i=1}^{N_A} w_i (\|\alpha - a_i\|_1 - \log \Pr(a_i|\alpha)) \quad (3)$$

which penalizes the differences of action features by both l_1 -norm and contrastive similarity. Here, w_i is 1 for the GT action and 0 otherwise, and

$$\Pr(a_i|\alpha) = \frac{\exp(\hat{\alpha} \cdot \hat{a}_i/\tau)}{\sum_{j=1}^{N_A} \exp(\hat{\alpha} \cdot \hat{a}_j/\tau)} \quad (4)$$

measures the probabilistic similarity of predicted and GT labels among candidates from taxonomy. $\hat{z} = z/\|z\|$ denotes the normalized unit vector, and $\tau = 0.07$ is the temperature of contrastive similarity. When testing, we perform action recognition by searching the closest labels to μ^A .

For future motion supervision, instead of expanding down to concrete hand poses, we directly compare mid-level features for efficiency. Specifically, $\mathbf{m}_{n+1:n+\bar{n}}^A$ are compared with pre-computed $\bar{\mathbf{m}}_{n+1:n+\bar{n}}^P$, with $\bar{\mathbf{m}}_j^P = \mu_j^P$ encoding the GT hand motion of the future j -th clip via \mathbf{E}^P . The motion prediction loss is

$$L_{mid} = \sum_{j=n+1}^{n+\bar{n}} \|\mathbf{m}_j^A - \bar{\mathbf{m}}_j^P\|_1 \quad (5)$$

To summarize, the overall loss of \mathbf{A} is $L_A = \lambda_4 L_{mid} + \lambda_5 L_{action} + \lambda_6 L_{KL}^A$.

3.2 Network Flow for Tasks

The framework addresses tasks of recognition (*i.e.* pose refinement and action recognition) and prediction by going through different paths within the network.

Recognition Recognition tasks are performed through the encoders. Specifically, \mathbf{E}^P refines the input per-frame estimated hand pose by referring to the motion regularity over a subsecond clip of t frames, followed by \mathbf{E}^A to output μ^A for action recognition over the entire T input frames.

Prediction Prediction of future hand motion is fulfilled by the decoders. Given observed hand motion $\tilde{\mathbf{H}}_{1:T}$, G-HTT provides two ways for the prediction of diverse and realistic hand motions (Fig. 1):

- (P.a) **(1)→(4)**. It generates locally consistent motions using only \mathbf{P} . It takes the last t observed frames $\tilde{\mathbf{H}}_{T-t+1:T}$ as input, and predicts motion of the following clip $\mathbf{H}_{T+1:T+t}$ by sampling from $N(\mu^P, \Sigma^P)$ for \mathbf{D}^P decoding. The output motion can then be autoregressively fed back to \mathbf{P} as input for longer prediction.
- (P.b) **(1)→(2)→(3)→(4)**. For more realistic long-term prediction with action guidance, we move up the hierarchy to leverage \mathbf{A} and predict $\mathbf{m}_{n+1:n+\bar{n}}^A$, with diversity coming from sampling from $N(\mu^A, \Sigma^A)$. $\mathbf{m}_{n+1:n+\bar{n}}^A$ are further decoded by \mathbf{D}^P into concrete poses $\mathbf{H}_{T+t+1:T+(\bar{n}+1)t}$.

We compare the two paths for motion prediction empirically in Sec. 4.4, and use P.b by default for long-term prediction in other experiments.

3.3 Implementation Details

\mathbf{P} , \mathbf{A} are trained across different datasets separately. To deploy G-HTT for practical RGB video processing, we leverage an external image-based hand object

estimator \mathbf{F} and a sequence-based object aggregator \mathbf{E}^{PO} , to provide the input for G-HTT. The external modules can be trained independently or obtained from off-the-shelf models.

G-HTT Details We set $t = 16$ and allow T to have a maximum value of 256 at 30 fps. Both \mathbf{P} and \mathbf{A} have 9 layers for the encoders and decoders, with a token dimension of $d = 512$. We train a single network across datasets with different pose and action annotation qualities for enhanced capability (Sec. 4.5).

We first train \mathbf{P} on all available pose sequences, regardless of the availability and transition of action labels, thanks to the decoupling of action and \mathbf{P} . We augment the input motion $\tilde{\mathbf{H}}$ with random noise, making \mathbf{P} capable of coping with noisy per-frame estimation in the recognition stage. Then, we fix the pre-trained \mathbf{P} and train \mathbf{A} , whose training data assumes the same action shared between the observation and prediction. We derive the mid-level $\bar{\mathbf{m}}^P$ from the pose annotations with \mathbf{E}^P , randomly divide the training sequence into observed and predicted parts, and correspondingly assign $\bar{\mathbf{m}}^P$ as the input of \mathbf{E}^A or supervision signal of \mathbf{D}^A . For the input of \mathbf{E}^A , we augment the mid-level features with random Gaussian noise, and refer to the noun of GT action for the clip-wise object feature ω .

We use AdamW [28] optimizer with a learning rate of 10^{-4} and weight decay of 0.01 for both \mathbf{P} and \mathbf{A} , where our batch size is 256. We respectively train \mathbf{P} and \mathbf{A} with 80 and 200 epochs, with loss weights as $\lambda_1, \lambda_2 = 1, \lambda_3 = 10^{-5}$, and $\lambda_4 = 1, \lambda_5 = 0.1, \lambda_6 = 10^{-5}$. Other design and training details are illustrated in the supplementary.

Image-based Estimator \mathbf{F} takes an image as input, and outputs for the image its hand pose, along with the object feature $\tilde{\mathbf{o}}$ that is comparable with CLIP [6, 39, 41] feature of object in manipulation. For experiments, we implement \mathbf{F} as a ResNet-18 [18] backbone followed by heads regressing the hand pose and object feature. We provide more details in the supplementary.

Clip-wise Object Detector \mathbf{E}^{PO} extracts ω as the clip-wise object representation aggregated from per-frame object features $\tilde{\mathbf{o}}$ of t consecutive frames. The clip-wise ω is then fed into \mathbf{E}^A to provide a consistent object information for action recognition (Sec. 3.1). We implement \mathbf{E}^{PO} by 2 transformer encoder layers, which is trained based on \mathbf{F} ; details are explained in the supplementary.

4 Experiments

4.1 Datasets

To demonstrate the versatility of our framework, we use three large-scale hand action datasets [23, 36, 42], covering highly diverse motions and actions, for training and testing. Across all datasets, we consider $N = 20$ joints annotated by [23], leaving the carpometacarpal joint of the thumb out as its annotation is unavailable in [42]. As the short-span hand motion is independent of action, we train \mathbf{P} on untrimmed sequences that could contain multiple action annotations, and train \mathbf{A} on trimmed sequences with clean and complete action annotations. For

the evaluation of both pose and action, we use trimmed sequences with clean action labels.

H2O [23] records four subjects performing 36 indoor daily activities, in four fixed (*cam0-3*) and one egocentric camera (*cam4*) viewpoints. We conduct evaluation on validation and test splits, the latter having subjects unseen in training. **Assembly101 [42] and AssemblyHands [36]** Assembly101 [42] contains procedures of people assembling and disassembling toy vehicles, with pose labels computed automatically by UmeTrack [17]. AssemblyHands [36] further improves the pose annotation quality for a subset of Assembly101 sequences. We follow the splits of Assembly101 for training and testing, using actions of its fine-grained taxonomy with 1380 different labels. We conduct evaluations on sequences that have more reliable pose annotations from AssemblyHands, and consider six fixed camera views that have no severe hand occlusions (details in supplementary). We focus evaluation on the validation split, as it contains accessible object labels for action recognition and motion prediction (Secs. 4.3 and 4.4), which the test split lacks.

4.2 Setup and Metrics

Recognition We take a whole video sequence depicting the process of an action as input. For evaluation of pose estimation, we use metrics of **MPJPE-RA** (Mean Per Joint Position Error-Root Aligned) and **MPJPE-PA** (Procrustes Analysis of MPJPE). To deal with the ambiguity of different hand scales, we align estimation and GT by equalizing the average length of palm bones. For action recognition, we report the top-1 classification accuracy (**Action Acc.**).

Prediction We divide each action video into segments of 16 frames. Given each segment, we use its pose and object annotation as the input observation, and predict the rest sequence until reaching the end of action or the maximum duration of 96 frames. We generate 20 random samples from $N(\mu, 5\Sigma)$ as a trade-off between accuracy and diversity (Secs. 4.3 and 4.4). For the evaluation of generative results, we mainly use the widely adopted **FID** (Frechet Inception Distance) [19] to assess quality, which computes the distributional distance of features between generated and GT motion sequences. The features are obtained from the last layer of a pre-trained transformer-based action recognition network, and the GT sequences are from the evaluation split unseen in training. A smaller FID means a more faithful generation. In addition, to explicitly measure generation diversity, we report **APD** (Average Pairwise Diversity) [56] in *mm* that computes the average distance between all pairs of 20 generated samples. A larger APD means a more diverse generation. More details about the setup and metrics are given in the supplementary.

4.3 Joint Modeling of Recognition and Prediction

We first demonstrate the enhanced capability because of our joint modeling of recognition and prediction, by respectively comparing with state-of-the-art

Table 1: Pose estimation and action recognition results. For hand pose estimation we report MPJPE-RA/-PA in *mm* for (left,right) hand respectively, where * denotes training views leveraged for HTT [52]. Please refer to the supplementary for complete results on all camera views, and comparison on the H2O-Val.

H2O-Test				AssemblyHands-Val					
		Resnet-18(F)	HTT [52]	Ours		Resnet-18(F)	HTT [52]	Ours	
cam0*	MPJPE-RA↓	27.0,25.6	26.9, 24.1	26.5 ,25.3	v1	MPJPE-RA↓	35.4,22.7	55.6,39.0	35.1 , 22.4
	MPJPE-PA↓	7.8,10.6	7.3 ,10.4	7.4, 10.3		MPJPE-PA↓	12.0,10.8	17.2,14.2	11.7 , 10.4
	Action Acc.↑	-	85.12	59.92		Action Acc.↑	-	16.55	36.01
cam2	MPJPE-RA↓	19.2,24.8	20.1,25.4	18.9 , 24.5	v3*	MPJPE-RA↓	27.5,27.2	26.7 ,27.3	27.3 , 26.9
	MPJPE-PA↓	6.9,10.6	7.4,11.0	6.6 , 10.3		MPJPE-PA↓	12.2,12.0	12.3,12.1	11.9 , 11.7
	Action Acc.↑	-	73.55	68.18		Action Acc.↑	-	39.42	34.79
cam4	MPJPE-RA↓	18.4,21.4	101.2,137.8	17.9 , 21.0	v8	MPJPE-RA↓	26.1,30.4	91.3,88.5	25.9 , 30.0
	MPJPE-PA↓	6.8,9.4	28.5,33.8	6.4 , 9.1		MPJPE-PA↓	11.8,12.3	24.2,27.3	11.5 , 11.8
	Action Acc.↑	-	2.89	57.85		Action Acc.↑	-	9.98	36.74

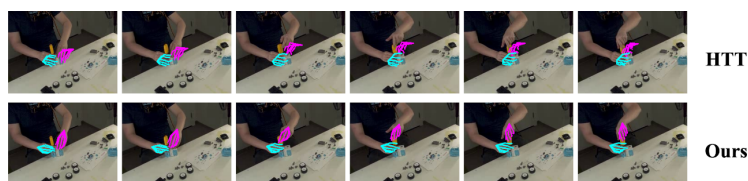


Fig. 3: Qualitative comparison of pose estimation for HTT [52] and our G-HTT, on camera view *v1* of Assembly datasets [36, 42]. More cases on H2O and AssemblyHands datasets are provided in the supplementary.

solutions for either recognition (*i.e.* HTT [52]) or prediction (*i.e.* PoseGPT [29]). More implementation details for the baselines are given in the supplementary.

Recognition The most relevant baseline is HTT [52], which also models the semantic-temporal hierarchy but focuses only on recognition. Based on the pre-trained image-based estimator **F** used by ours for fair comparison (Sec. 3.3), we train HTT on two camera views of H2O (*cam0,1*) and one view (*v3*) of AssemblyHands, where we concatenate image with the estimated hand pose and object from **F** as the per-frame input of HTT. We also take the initial pose estimation of **F** as a reference for comparison. Moreover, to obtain the object input for both methods, on H2O we leverage the network estimation, while on AssemblyHands we use the GT labels, where it is very challenging to recognize objects reliably due to cluttered scenes and frequent occlusions (Fig. 3).

As shown in Tab. 1 and Fig. 3, G-HTT demonstrates robust accuracy on various camera views, for refining the local pose and action recognition, even though G-HTT is never trained on **F**. In comparison, although HTT fits better on views that are trained on or close to trained ones (*e.g.*, *cam2* of H2O), its performance significantly degrades on the other views, even worse than its input obtained from **F**. The results show that our simultaneous modeling of both recognition and prediction enhances generalization by learning regular motion priors across tasks. In contrast, a recognition-only network is more likely to overfit particular data distributions.

Prediction We take PoseGPT [29], a state-of-the-art model for motion prediction with *prescribed* action, as a baseline for performance evaluation. While the

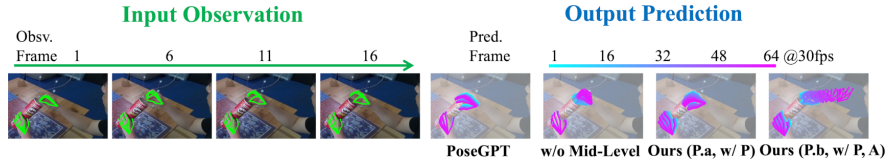


Fig. 4: Qualitative comparison of predicted motions for PoseGPT [29], the ablated settings of w/o mid-level, w/ only \mathbf{P} via path P.a, and the full G-HTT (w/ \mathbf{P} , \mathbf{A} , via path P.b) on H2O. More qualitative cases are provided in the supplementary.

Table 2: Comparison with PoseGPT [29] for motion prediction, on action sequences that are longer than 1 sec. APD in *mm* for (left, right) hand respectively.

	GT action input	H2O-val		H2O-Test		AssemblyHands-Val	
		FID↓	APD↑	FID↓	APD↑	FID↓	APD↑
PoseGPT [29]	✓	5.19	32.3,43.1	11.70	24.1,48.6	16.07	25.3, 33.0
Ours (P.b, w/ \mathbf{P} , \mathbf{A})	×	5.32	22.1,25.7	8.19	20.1,33.9	5.04	28.1,32.8

Table 3: Comparison of long-term prediction decoded from \mathbf{m}^P (P.a) and \mathbf{m}^A (P.b), on action sequences that are longer than 2 sec.

	H2O-val		H2O-Test		AssemblyHands-Val	
	FID↓	APD↑	FID↓	APD↑	FID↓	APD↑
Ours (P.a, w/ \mathbf{P})	8.18	40.7,48.3	12.78	36.7,52.5	8.20	40.8,51.1
Ours (P.b, w/ \mathbf{P} , \mathbf{A})	6.59	29.7,33.3	10.88	26.2,48.3	6.84	34.8,42.3

original PoseGPT trains on body motion, we retrain PoseGPT with its official code by combining the three hand pose-action datasets of Assembly101, AssemblyHands and H2O as we do. We evaluate on action sequences longer than 1 sec to better show the differences in prediction.

As reported in Tab. 2, G-HTT shows significantly better FID on the H2O-test split of unseen subjects and on AssemblyHands; meanwhile, the two methods have comparable accuracy on the H2O-val split of trained subjects. These results show our better generation quality across actions and datasets. Visually from Fig. 4, while PoseGPT suffers from lacking regularity for predicted motion, our prediction shows globally consistent action.

We attribute the differences to two factors: our joint learning of both recognition and prediction, and our hierarchical model for pose and action. In contrast to action-conditioned generation alone, by modeling both recognition and prediction our framework learns strong motion-action regularities across datasets covering highly diverse hand actions, as shown by the consistent high quality across three datasets (Tab. 2). Moreover, different from the vector quantization of PoseGPT for clip-wise motion, our mid-level representation originates from \mathbf{P} , where \mathbf{P} models not only the observed motion but also prediction, enabling global action-guided motion generation that preserves local motion continuity (see also Sec. 4.4).

Table 4: Comparison of motion prediction between ours and the ablated setup without modeling the mid-level, on action sequences that are longer than 2 sec.

	Sampling with	H2O-val		H2O-Test		AssemblyHands-Val	
		FID↓	APD↑	FID↓	APD↑	FID↓	APD↑
w/o Mid-Level	$\mu, 5\sigma$	6.69	18.0,23.7	13.84	17.9,29.4	5.67	16.9,21.9
w/o Mid-Level	$\mu, 10\sigma$	14.05	25.5, 33.4	22.64	24.1,35.2	6.31	23.3,30.2
Ours (P.b, w/ P, A)	$\mu, 5\sigma$	9.30	27.8 ,31.4	13.10	24.6 , 43.3	5.21	33.6 , 40.0

4.4 Modeling Semantic-Temporal Hierarchy

In this section, we examine the effects of modeling the semantic-temporal hierarchy. As HTT [52] has well demonstrated the benefits of leveraging this hierarchy in recognition tasks, here we mainly examine its benefits for motion prediction, especially on sequences longer than 2 secs where global action is more apparent.

Action for Prediction We first compare the long-term prediction decoded from \mathbf{m}^P and \mathbf{m}^A , *i.e.*, the two strategies P.a, P.b described in Sec. 3.2, to examine the effectiveness of involving **A** for long-term prediction. As shown in Tab. 3 and Fig. 4, generations from \mathbf{m}^A are more realistic and plausible, with lower FID and more consistent global motion. In contrast, results of P.a show larger diversity due to its short-term modeling, but lack fidelity or regularity for long-term motion. Overall, the comparison shows the importance of action modeling in generating faithful and action-guided motions.

Mid-level for Prediction In addition to enabling a decoupled training strategy for **P** and **A** (Sec. 3.3), the modeling of mid-level features should enhance the learning of generation. To verify it, we construct a flattened baseline (*i.e.*, w/o mid-level) by removing the mid-level representation and instead using a single transformer VAE to directly model pose and action. For this flattened baseline, its encoder takes hand poses and objects as input and outputs for action recognition; its decoder directly outputs the future hand poses. The flattened baseline has a comparable amount of parameters as **P** and **A**, and is trained on the same dataset as our framework for fair comparison.

From Tab. 4 we find that under a comparable FID, the mid-level representation enables better diversity (3rd, 5th rows); meanwhile, as we increase the generation diversity of the flattened baseline via more noisy sampling, its accuracy significantly degrades (4th, 5th rows). From Fig. 4, we can see that the flattened baseline results lack global regularity, despite its modeling of pose and action through a powerful end-to-end transformer VAE. The comparison shows that the mid-level representation enables easier learning of global motion regularity, as it decouples the complex task of action-guided motion generation into hierarchical subtasks better captured by **P** and **A** respectively.

4.5 Discussion

We make more observations about motion prediction, training strategy, and mid-level representation, to give additional understanding of the framework. Due to space limit, we provide more details in supplementary.

Training with Assembly101 We find that including the large-scale Assembly101 for training G-HTT significantly benefits motion prediction on H2O and action recognition, although the pose annotation of Assembly101 is not sufficiently accurate for training pose estimators (*cf.* [36]). The finding points to the importance of large-scale pretraining of fundamental modules.

Predicting Action Transition We observe that our model can generate smooth transitions between actions, It probably comes from training \mathbf{P} on sequences of mixed action annotations (Sec. 3.3), which allows \mathbf{P} to drive the transition by decoding local motions into new actions. Meanwhile, it is also facilitated by the decoupled training of \mathbf{P} and \mathbf{A} .

The Mid-level Regularity We blend mid-level features of two different input pose sequences, and decode the blended mid-level features into pose sequences. The generated motions naturally interpolate between the tendencies of two given inputs, showing the regularity of the learned mid-level representation.

5 Conclusion

We present a novel unified framework for understanding hand pose and action, which concurrently models both recognition and prediction, and captures the hierarchy of semantic dependency and temporal granularity. The framework addresses tasks of 3D hand pose refinement, action recognition, and 3D hand motion prediction, showing improved performances than isolated solutions. The framework has two cascaded Transformer VAE blocks to model pose and action respectively. Both blocks have their encoder and decoder output respectively for recognition and prediction, while their VAE bottleneck extracts the temporal regularity synergized between the two sides. A mid-level clip-wise motion representation is further introduced to bridge the two blocks. The connected cascade enables regular pose and action modeling over both short and long time spans, and brings flexibility to train the two blocks separately on multiple datasets with different setups and annotation granularities. Extensive experiments validate the performance and design of our framework on both recognition and prediction across different datasets.

Limitations and Future Work We assume a fixed camera viewpoint for input videos; to process cases with drastic camera movement (*e.g.*, egocentric views with large head motions), an explicit decomposition of hand and camera motion would be necessary, which we leave as future work. Another aspect is to leverage hand motion as priors for robust recognition of manipulated objects, therefore further benefiting action understanding. Moreover, extensions to cross-dataset settings and human body pose action modeling are also interesting directions for broader impacts.

Acknowledgement This research is supported by the Innovation and Technology Commission of the HKSAR Government under the InnoHK initiative, Innovation and Technology Commission (Ref: ITS/319/21FP), Research Grant Council of Hong Kong (Ref: 17210222, 17200924), JST ASPIRE (Grant Number: JPMJAP2303), and JST ACT-X (Grant Number: JPMJAX2007).

References

1. Aliakbarian, S., Saleh, F.S., Salzmann, M., Petersson, L., Gould, S.: A stochastic conditioning scheme for diverse human motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5223–5232 (2020)
2. Bao, W., Chen, L., Zeng, L., Li, Z., Xu, Y., Yuan, J., Kong, Y.: Uncertainty-aware state space transformer for egocentric 3d hand trajectory forecasting. arXiv preprint arXiv:2307.08243 (2023)
3. Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.J., Yuan, J., Thalmann, N.M.: Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2272–2281 (2019)
4. Cai, Y., Wang, Y., Zhu, Y., Cham, T.J., Cai, J., Yuan, J., Liu, J., Zheng, C., Yan, S., Ding, H., et al.: A unified 3d human motion synthesis model via conditional variational auto-encoder. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11645–11655 (2021)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
6. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2818–2829 (2023)
7. Chi, S., Chi, H.g., Huang, Q., Ramani, K.: Infogcn++: Learning representation by predicting the future for online human skeleton-based action recognition. arXiv preprint arXiv:2310.10547 (2023)
8. Chik, D., Trumpf, J., Schraudolph, N.N.: Using an adaptive var model for motion prediction in 3d hand tracking. In: 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition. pp. 1–8. IEEE (2008)
9. Cho, H., Kim, C., Kim, J., Lee, S., Ismayilzada, E., Baek, S.: Transformer-based unified recognition of two hands manipulating objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4769–4778 (2023)
10. Fan, Z., Liu, J., Wang, Y.: Adaptive computationally efficient network for monocular 3d hand pose estimation. In: European Conference on Computer Vision. pp. 127–144. Springer (2020)
11. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 203–213 (2020)
12. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)
13. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1933–1941 (2016)
14. Gammulle, H., Denman, S., Sridharan, S., Fookes, C.: Predicting the future: A jointly learnt model for action anticipation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5562–5571 (2019)

15. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2021–2029 (2020)
16. Han, S., Liu, B., Cabezas, R., Twigg, C.D., Zhang, P., Petkau, J., Yu, T.H., Tai, C.J., Akbay, M., Wang, Z., et al.: Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics (ToG)* **39**(4), 87–1 (2020)
17. Han, S., Wu, P.c., Zhang, Y., Liu, B., Zhang, L., Wang, Z., Si, W., Zhang, P., Cai, Y., Hodan, T., et al.: Umetrack: Unified multi-view end-to-end hand tracking for vr. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
20. Iqbal, U., Molchanov, P., Gall, T.B.J., Kautz, J.: Hand pose estimation via latent 2.5 d heatmap regression. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 118–134 (2018)
21. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. arXiv preprint arXiv:2306.14795 (2023)
22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
23. Kwon, T., Tekin, B., Stühmer, J., Bogo, F., Pollefeys, M.: H2o: Two hands manipulating objects for first person interaction recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10138–10148 (2021)
24. Li, M., An, L., Zhang, H., Wu, L., Chen, F., Yu, T., Liu, Y.: Interacting attention graph for single image two-hand reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2761–2770 (2022)
25. Li, Y., Ye, Z., Rehg, J.M.: Delving into egocentric actions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 287–295 (2015)
26. Liu, M., Tang, S., Li, Y., Rehg, J.M.: Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 704–721. Springer (2020)
27. Liu, S., Tripathi, S., Majumdar, S., Wang, X.: Joint hand motion and interaction hotspots prediction from egocentric videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3282–3292 (2022)
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
29. Lucas, T., Baradel, F., Weinzaepfel, P., Rogez, G.: Posegpt: Quantization-based 3d human motion generation and forecasting. In: European Conference on Computer Vision. pp. 417–435. Springer (2022)
30. Luo, R.C., Mai, L.: Human intention inference and on-line human hand motion prediction for human-robot collaboration. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5958–5964. IEEE (2019)
31. Ma, H., Li, J., Hosseini, R., Tomizuka, M., Choi, C.: Multi-objective diverse human motion prediction with knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8161–8171 (2022)

32. Ma, M., Fan, H., Kitani, K.M.: Going deeper into first-person activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1894–1903 (2016)
33. Mao, W., Liu, M., Salzmann, M.: Weakly-supervised action transition learning for stochastic human motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8151–8160 (2022)
34. Moon, G.: Bringing inputs to shared domains for 3d interacting hands recovery in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17028–17037 (2023)
35. Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D., Theobalt, C.: Gnerated hands for real-time 3d hand tracking from monocular rgb. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 49–59 (2018)
36. Ohkawa, T., He, K., Sener, F., Hodan, T., Tran, L., Keskin, C.: AssemblyHands: towards egocentric activity understanding via 3d hand pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12999–13008 (2023)
37. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10985–10995 (2021)
38. Petrovich, M., Black, M.J., Varol, G.: TEMOS: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision (ECCV) (2022)
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
40. Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., Guibas, L.J.: Humor: 3d human motion model for robust pose estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11488–11499 (2021)
41. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5b: An open large-scale dataset for training next generation image-text models. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022), <https://openreview.net/forum?id=M3Y74vmsMcY>
42. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21096–21106 (2022)
43. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12026–12035 (2019)
44. Shi, M., Starke, S., Ye, Y., Komura, T., Won, J.: Phasemp: Robust 3d pose estimation via phase-conditioned human motion prior. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14725–14737 (2023)
45. Singh, S., Arora, C., Jawahar, C.: First person action recognition using deep learned descriptors. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2620–2628 (2016)
46. Spurr, A., Iqbal, U., Molchanov, P., Hilliges, O., Kautz, J.: Weakly supervised 3d hand pose estimation via biomechanical constraints. In: European Conference on Computer Vision. pp. 211–228. Springer (2020)

47. Tekin, B., Bogo, F., Pollefeys, M.: H+o: Unified egocentric recognition of 3d hand-object poses and interactions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4511–4520 (2019)
48. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: European Conference on Computer Vision. pp. 358–374. Springer (2022)
49. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
50. Vondrick, C., Pirsivash, H., Torralba, A.: Anticipating visual representations from unlabeled video. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 98–106 (2016)
51. Wang, J., Mueller, F., Bernard, F., Sorli, S., Sotnychenko, O., Qian, N., Otaduy, M.A., Casas, D., Theobalt, C.: Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. ACM Transactions on Graphics (ToG) **39**(6), 1–16 (2020)
52. Wen, Y., Pan, H., Yang, L., Pan, J., Komura, T., Wang, W.: Hierarchical temporal transformer for 3d hand pose estimation and action recognition from egocentric rgb videos. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
53. Xu, M., Gao, M., Chen, Y.T., Davis, L.S., Crandall, D.J.: Temporal recurrent networks for online action detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5532–5541 (2019)
54. Yang, S., Liu, J., Lu, S., Er, M.H., Kot, A.C.: Collaborative learning of gesture recognition and 3d hand pose estimation with multi-order feature analysis. In: European Conference on Computer Vision. pp. 769–786. Springer (2020)
55. Yu, Z., Huang, S., Fang, C., Breckon, T.P., Wang, J.: Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12955–12964 (2023)
56. Yuan, Y., Kitani, K.: Dlow: Diversifying latent flows for diverse human motion prediction. In: Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 346–364. Springer (2020)
57. Zhang, Y., Huang, D., Liu, B., Tang, S., Lu, Y., Chen, L., Bai, L., Chu, Q., Yu, N., Ouyang, W.: Motiongpt: Finetuned llms are general-purpose motion generators. arXiv preprint arXiv:2306.10900 (2023)
58. Zhao, Y., Krähenbühl, P.: Real-time online video detection with temporal smoothing transformers. In: European Conference on Computer Vision. pp. 485–502. Springer (2022)
59. Zimmermann, C., Brox, T.: Learning to estimate 3d hand pose from single rgb images. In: Proceedings of the IEEE international conference on computer vision. pp. 4903–4911 (2017)