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

# Persistent-Transient Duality: A Multi-mechanism Approach for Modeling Human-Object Interaction

Hung Tran[1], Vuong Le[2], Svetha Venkatesh[1], Truyen Tran[1]
[1]Applied AI Institute, Deakin University, [2]Amazon
{tduy, svetha.venkatesh, truyen.tran}@deakin.edu.au, levuong@amazon.com

## Abstract

*Humans are highly adaptable, swiftly switching between different modes to progressively handle different tasks, situations and contexts. In Human-object interaction (HOI) activities, these modes can be attributed to two mechanisms: (1) the large-scale consistent plan for the whole activity and (2) the small-scale children interactive actions that start and end along the timeline. While neuroscience and cognitive science have confirmed this multi-mechanism nature of human behavior, machine modeling approaches for human motion are trailing behind. While attempting to use gradually morphing structures (e.g., graph attention networks) to model the dynamic HOI patterns, they miss the expeditious and discrete mode-switching nature of the human motion. To bridge that gap, this work proposes to model two concurrent mechanisms that jointly control human motion: the Persistent process that runs continually on the global scale, and the Transient sub-processes that operate intermittently on the local context of the human while interacting with objects. These two mechanisms form an interactive Persistent-Transient Duality that synergistically governs the activity sequences. We model this conceptual duality by a parent-child neural network of Persistent and Transient channels with a dedicated neural module for dynamic mechanism switching. The framework is trialed on HOI motion forecasting. On two rich datasets and a wide variety of settings, the model consistently delivers superior performances, proving its suitability for the challenge.*

## 1. Introduction

In a planned activity that involves interaction with objects, the patterns of human motion vary greatly, contextualized to the situation and stages of the plan. The variety of patterns comes from the fact that the subject needs to switch between multiple modes of operations: navigating the global scene according to the plan, and occasionally concentrate on a small set of objects to act on them. This

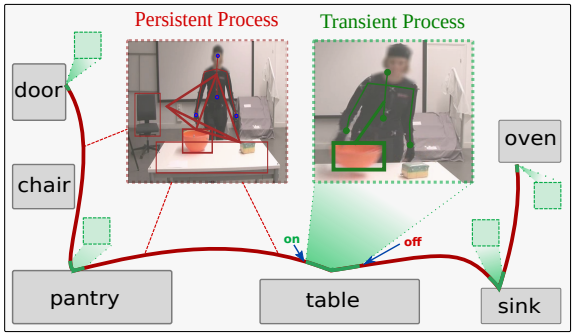



Figure 1: We model the human motions by a duality of two types of processes: A single *Persistent process* (red box) considers the whole activity plan and trajectory (red curve) and use dense relational structure (red graph). Its children *Transient processes* (green boxes) work on the small and local spatiotemporal scope of a human and the interacted objects using egocentric structure (green graph). They get turn on and off on-demand by the Persistent process.

mode-switching nature requires a machine model to adapt quickly in structure, representation, and inference mechanisms to follow the patterns of the behavior. This observation is confirmed by neuroscience findings that human brain activity contains transient networks that deals with particular situation [4]. In cognitive science, human activities are also proved to follow parent-child planning patterns [2].

Recent advancements in graph neural networks [43] allow motion models to dynamically adjust their relational structure to adapt to changing situations [8]. However, with a singular inference mechanism, they can only have *gradual adaptation*: smoothly adjusting parameters of the same model. They cannot account for quick and abrupt changes regarding the discrete switching between distinctive mechanisms and as a result fail to keep up with the movement patterns. This limitation is inherent in human-object interaction motion (HOI-M) prediction. Fig. 1 visualizes an example activity of cooking a dinner meal. Here the subject navigates around the kitchen following a recipe and con-

sider the whole kitchen floor plan (red trajectory in Fig. 1). Occasionally, they deviate to perform a particular action by interacting with an object (green sections in Fig. 1) such as moving a bowl, where only the bowl needs their attention. When this happens, these models will continue to consider the interacted object as an equal member of the scene and miss its importance as the direct recipient of the action.

To address this limitation, we explore a new modeling paradigm that factorizes the human-object interaction into two internal types of processes: a *Persistent process* (red box in Fig. 1) which maintains a continuous large-scale activity progress; and *Transient sub-processes* (green boxes in Fig. 1) which has an adaptive life cycle and a personalized structure to reflects the small-scale local interaction with objects. The Transient act as Persistent's sub-processes, and they can be turned on or off by a switch (on/off arrows in Fig. 1) operating on signal from the main Persistent process. The *parent-child relationship* connects the two types of processes and constitutes the *Persistent-Transient Duality*.

This modeling is related to similar transient concepts in other fields such as control theory [22], electrical [19] and chemical engineering [41, 14]. In computer engineering, the parent-child relationship also resembles the operating system and the children task-specific processes.

We model this concept into a multi-channel neural network called *Persistent-Transient Duality* (PTD) networks for Human-object interaction motion (HOI-M) prediction. The *Persistent channel* contains a recurrent relational network operating on the global scene spatially and throughout the session temporally. The *Transient channels* instead have contextualized structures constructed on the spot whenever the human subjects shift the priority toward interacting with a subset of objects. The life cycles of these spontaneous channels are managed by a neural *Transient Switch*, which anticipates the initialization and termination of many Transient processes along a single Persistent process.

The benefit of PTD is demonstrated via motion forecasting experiments on WBHM and Bimanual Actions datasets where it sets new SOTA performances and generalizability tests. Being a generic framework for human behavior modeling, PTD is readily applicable to other problems such as pedestrian trajectory prediction of which preliminary adaptation and experiment are shown in the supplementary.

The key contributions of this work are:

1. The exploration of *the new Persistent-transient duality concept* to model the multi-mechanism nature of human behavior reflected in the large-small temporal scale and global-local spatial scope of HOI motions.

2. A *parent-children neural framework* with egocentric design that applies the PTD concept in HOI-M prediction.

3. The extensive analysis demonstrating that PTD *sets new SoTA in HOI-M prediction* across multiple datasets and settings, and generalizes better to new scenarios.

## 2. Related Work

**Human-object interaction** in videos is traditionally formulated for predictions of human action and object affordance labels [30, 37, 26, 15]. Approaches include modeling the activity continuity with CNNs [30] and RNNs [37], or modeling the interactions using GNNs [25, 18], which can be improved with dynamic topology [36, 39]. Recent works also leverage the power of transformers [42] to model the spatio-temporal relationship between humans and objects in the videos [26]. In this paper, we consider human-object interaction motion (HOI-M) prediction, the task of forecasting the concrete future locations of humans and objects in an activity which is more well-defined and easier to evaluate than the vague affordance labels traditionally used in HOI. Also, this task is more complex and requires more precise modeling structures, such as the duality proposed in this work.

**Human body motion** predictions are done with MLPs [6, 21], RNNs [16, 17, 38, 3, 45] or GNNs [25, 9, 29, 11, 12, 34, 40, 32, 31] and can be embedded in generative models [20, 23, 10, 27]. More recent works started to consider relations with surroundings in the form of intention toward destinations [7] or interaction with other entities [8, 1]. We advance this line by considering the directional relations between human and objects beyond the simplistic homogenous graphs such as in [8] through a new multi-mechanism adaptive structural of the persistent-transient duality.

## 3. Method

### 3.1. Preliminaries

We consider the problem of modeling the sequential motions of $N$ entities of humans and objects, where each entity $i$ is represented by a class label $c_i$ and sequential visual features $X_i = \{x_i^t\}_{t=1}^{T}$. Given the class labels and the sequential features, we want to predict the future features in the next $L$ steps, $Y_i = \{y_i^t\}_{t=T+1}^{T+L}$. Each time instance of features $x_i^t$ of humans contains skeleton positions; while those of objects include their object types and bounding boxes.

This problem is traditionally approached under the single-mechanism assumption that the inference dynamics stay the same during the whole session [8]. In this section, we present our new multi-mechanism paradigm, modeled with the persistent-transient duality, to capture how humans maintain their global operation mode while swiftly adapt to emerging local interaction situations.

### 3.2. The Persistent-Transient Duality

We consider the motions of the human bodies and the interacted objects to be caused by two mechanisms: 1. the navigation according to the global activity progress; and 2. the individual interactions with objects, which happens in

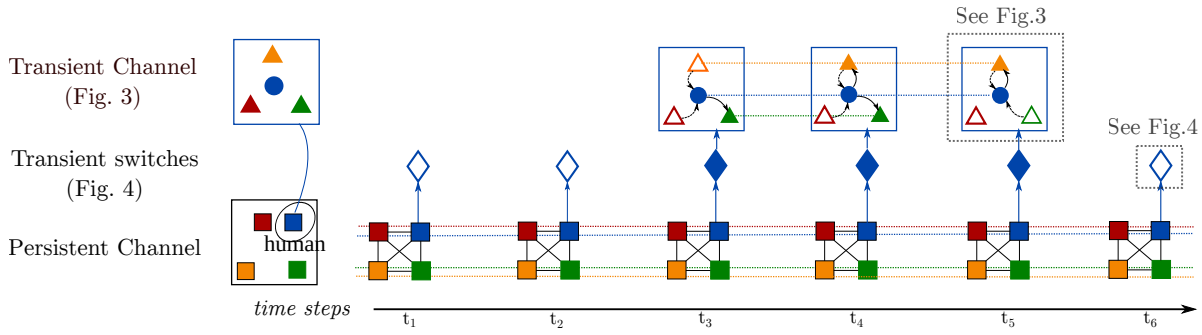



Figure 2: The architecture of *Persistent-transient Duality* Networks (PTD). The *persistent channel* is a heterogeneous graph network connecting all entities at the same time step. In complement, the *Transient channel* zooms into the local context of each human (circles) when they interact with surrounding objects (triangles). This channel works with the egocentric structure, under the viewpoint of the human subject. The transient channels are initialized and terminated on-demand, controlled by the *Transient Switches* (diamonds) which turn on (filled) or off (empty) the corresponding Transient processes.

the local contexts. This multi-mechanism factorization can be viewed as a persistent-transient process duality: The *persistent process* operates in large activity temporal scale and with a global spatial perspective. Its counterpart, the *transient process,* is local in time and egocentric to each human subject's spatial viewpoint.

We model this duality by a hierarchical neural network called *Persistent-Transient Duality Networks (PTD)* whose architecture is drawn in Fig. 2. Within PTD, the *Persistent process* is modeled as a single *Persistent channel*, which operates recurrently along the whole activity sequence. Its children, the *Transient channel* instances, are initiated with personalized structures and representations whenever a human has a new interaction with the surrounding. When the interaction is over, the outdated transient channel is terminated, and the control is returned to the persistent channel waiting in the background. These transient life cycles are managed by the neural modules called *Transient Switches*.

### 3.3. Persistent Channel

The Persistent channel models the common global view of all humans and objects in the scene. It takes the form of a heterogeneous graph attention network [44] with two node types corresponding to human and object entities and *dense spatial edges* connecting entities at the same time step. We extend this model by adding *recurrent temporal edges* along the temporal dimension for each entity:

$$z_i^t = \left[x_i^t, m_i^t, m_{i,\mathcal{T}\to\mathcal{P}}^t\right], \quad h_i^t = \text{RNN}_{c_i}\left(z_i^t, h_i^{t-1}\right), \quad (1)$$

where $\text{RNN}_{c_i}$ is a recurrent unit (e.g., GRU) that corresponds to the class of the $i^{\text{th}}$ entity. The input $z_i^t$ of $\text{RNN}_{c_i}$ is formed using the the entity's intrinsic features $x_i^t$ (skeleton for humans and bounding boxes for objects), the spatial messages $m_i^t$ gathered from spatial edges, and a message from the active transient channel $m_{i,\mathcal{T}\to\mathcal{P}}^t$ (Sec. 3.4).

The spatial message $m_i^t$ in Eq. (1) is aggregated from the *spatial edges* of the graph via attention: $m_i^t = \text{Attn}\left(u_i^t, \left\{u_j^t\right\}_{j\neq i}\right)$ with $u_-^t = \left[x_-^t; h_-^{t-1}\right]$. Here and throughout this work, we use the GAT-based attention function [43], $\text{Attn}\left(q, V\right)$, defined over the query $q$ and the identical key/value pairs $V = \{v_j\}_{j=1}^N$ by:

$$\text{Attn}\left(q, V\right) := \sum_{j=1}^{N} \text{softmax}_j\left(\sigma\left(\text{MLP}\left(\left[\boldsymbol{W_q} q; \boldsymbol{W}_v v_j\right]\right)\right)\right) \boldsymbol{W}_v v_j, \quad (2)$$

where $\boldsymbol{W_q}$ and $\boldsymbol{W}_v$ are the learnable embedding weights for the query and key/value, $[\cdot;\cdot]$ is the concatenation, $\sigma\left(\cdot\right)$ is a non-linear activation.

The Persistent channel generates two outputs from its hidden state: the message to the transient channel of each human $m_{i,\mathcal{P}\to\mathcal{T}}^t$ and the future feature of each entity $\hat{y}_{i,\mathcal{P}}^t$:

$$m_{i,\mathcal{P}\to\mathcal{T}}^t = \text{MLP}\left(h_i^t\right), \qquad \hat{y}_{i,\mathcal{P}}^t = \text{MLP}\left(h_i^t\right). \quad (3)$$

These prediction outputs are later combined with those from the Transient channel as detailed in Sec. 3.6.

### 3.4. Transient Channel

Operating with a consistent graph, the Persistent process alone cannot adapt quickly enough to emerging events that require different perspectives, particularly when the human interacts with an object. When these cases are detected, PTD initiates a *Transient process* that (1) zooms in to the relevant context and (2) take the viewpoint of the subject.

In our framework, the Transient process is implemented by a neural *Transient channel*, available one for each *human* entity. We propose an *egocentric recurrent graph networks* for the Transient channel. The egocentric property of this model is the most important aspect that separates it from the global view of its parent persistent channel. This egocentric

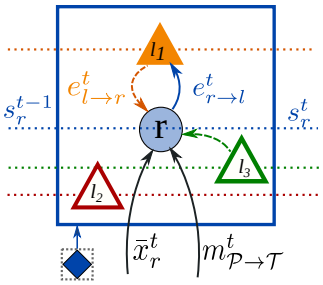



Figure 3: The Transient Channel features an egocentric structure of the center node (circle) and leaves (triangles) with sparse edges (colored arrows). When being active, it uses the egocentric features $\bar{x}^t_r$ and the persistent message $m^t_{\mathcal{P}\to\mathcal{T}}$ to update its state $s^t_r$. The life-cycle of this channel is determined by the Transient Switch (blue diamond) visualized in Fig. 4.

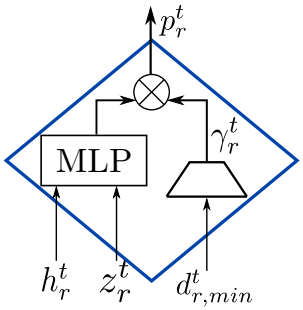



Figure 4: The Transient Switch predicts the transient score $p^t_r$ from the hidden states $h^t_r$ and the input $z^t_r$ of the persistent RNN. This score is modulated by a discount factor $\gamma^t_r$ computed from the minimum distance $d^t_{r,min}$ via the discount gate (trapezoid).

design reflects in three aspects of *computational structure, feature representation*, and *inference logic*.

**Egocentric computational structure.** In switching from the global to personalized view, we start by forming an *egocentric Transient graph* $\mathcal{G}^t_i = (\mathcal{V}^t_i, \mathcal{E}^t_i)$ at time $t$ of the human $i$ and their relations with the objects interacted with. For a particular human, subscript $i$ will be omitted for conciseness. The egocentric characteristic of $\mathcal{G}^t$ reflects in its star-like structure: the nodes $\mathcal{V}^t$ includes a single *center node* $r$ for the considering human, and *leaf nodes* of indices $\{l\}_{l\neq r}$ for objects. The dynamic edges $\mathcal{E}^t$ connect the center with the leaves in two directions: *inward edges* $e^t_{l\to r}$ reflect which objects the human may consider to interact with, and the *outward edges* $e^t_{r\to l}$ represents the objects are being manipulated by the human. In our implementation, the existences of the edges are determined by thresholding the center-leaf distances $d^t_{lr}$:

$$\begin{cases} e^t_{\_} \in \mathcal{E}^t & \text{if } d^t_{lr} \leq \beta_{\_} \\ e^t_{\_} \notin \mathcal{E}^t & \text{otherwise} \end{cases}, \tag{4}$$

where $\_$ indicates either the inward or onward, $\beta_{\_}$ is a pair of threshold hyper-parameters. This thresholding effectively thins out the neighbors, making the graph more efficient and localized around the center node. Inward threshold is commonly smaller than outward one, because humans subject's attention, hence motion, are affected by objects before and after objects get directly manipulated. The edges are estimated dynamically for each time step, allowing the graph's topology to evolve within one single Transient session.

**Egocentric representation.** Paired with the egocentric graph structure, the geometrical features of the entities are also transformed into the egocentric coordinate system corresponding to the viewpoint of the human center node $r$:

$$\bar{x}_{\_}{}^t = f_{\text{ego}}\left(x_{\_}{}^t, x^t_r\right) = x_{\_}{}^t - \text{centroid}(x^t_r), \tag{5}$$

where $\_$ indicate both center or leaf nodes. Conceptually, this change of system puts various patterns of the human's motion into the same aligned space. In the global view, features of two similar actions can be vastly different if they are far away. After this transformation, they are aligned into the common egocentric view of the center human subject. This alignment filters out the irrelevant global information, and facilitates efficient inference of the egocentric model.

**Egocentric inference.** Operating on the transient graph structure, the RNN hidden state of each node $s^t_{\_}$ is updated with the input aggregated through attention-based message passing along the edges in $\mathcal{E}^t$. See Fig. 3 for an illustration.

In detail, for the *center node*, the inward messages from its leaves are aggregated into: $m^t_r = \text{Attn}\left(u^t_r, \{u^t_l\}_{e^t_{l\to r}\in\mathcal{E}^t}\right)$, where $u^t_{\_} = \left[\bar{x}_{\_}{}^t; s_{\_}{}^{t-1}\right]$ and $\text{Attn}\,(q, V)$ is defined in Eq. (2). This message is combined with the egocentric features $\bar{x}^t_r$ and the persistent channel's message $m^t_{\mathcal{P}\to\mathcal{T}}$ (Eq. (3)) to update the recurrent state $s^t_r$:

$$s^t_r = \text{RNN}_r\left(\left[\bar{x}_r, m^t_r, m^t_{\mathcal{P}\to\mathcal{T}}\right], s^{t-1}_r\right). \tag{6}$$

For *leaf nodes*, when they are interacted by the center node (indicated via $e^t_{r\to l}$), they receive a center-broadcasted message $m^t_l = \text{MLP}\left(\left[\bar{x}^t_r; s^{t-1}_r\right]\right)$ and update their state $s^t_l$:

$$s^t_l = \begin{cases} \text{RNN}_l\left(\left[\bar{x}^t_l; m^t_l\right], s^{t-1}_l\right) & \text{if } e^t_{r\to l} \in \mathcal{E}^t \\ s^{t-1}_l & \text{otherwise} \end{cases}, \tag{7}$$

The updated hidden states are used to generate the messages sent to the persistent channel $m^t_{\mathcal{T}\to\mathcal{P}}$ and the transient predictions about the future feature $\hat{y}^{\mathcal{T},t}_{\_}$ of the entities:

$$m^t_{\mathcal{T}\to\mathcal{P}} = \text{MLP}\left(s^t_r\right), \quad \hat{y}^t_{-,\mathcal{T}} = f^{-1}_{\text{ego}}\left(\text{MLP}\left(s^t_-\right)\right), \quad (8)$$

where $f^{-1}_{\text{ego}}$ is the inverse function of Eq. (5) converting the egocentric back to global coordinates. The Transient predictions $\hat{y}^{\mathcal{T},t}_-$ are then combined with the Persistent counterparts as described in Sec. 3.6. As a key modeling feature, the egocentric design plays a crucial goal in the power of the PTD which will be demonstrated later in the Ablation studies (Sec. 4.2).

### 3.5. Switching Transient Processes

The life cycles of the Transient processes are managed based on the situation of human's activity. These decisions are made by a neural *Transient Switch* (See Fig. 4) that makes switch-on and switch-off decisions of the Transient process for each human entity. The switch-on probability $p^t_r$ is computed by considering the current hidden state $h^t_r$ and the input $z^t_r$ of the persistent RNN in Eq. (1):

$$\hat{p}^t_r = \gamma^t_r \cdot \text{sigmoid}\left(\text{MLP}\left(\left[h^t_r; z^t_r\right]\right)\right), \quad (9)$$

where $h^t_r$ and $z^t_r$ is the current hidden state and the input of the persistent RNN. The discount factor $\gamma^t_r \in [0, 1]$ responds to subject's distance to the nearest object $d^t_{r,min}$:

$$\gamma^t_r = \exp\left(-\eta \cdot d^t_{r,min}\right), \quad (10)$$

where $\eta$ is a learnable decay rate. This factor acts as a disruptive shortcut gate that modulates the switching decision based on the spatial evidence of the interaction.

Finally, the binary switch decision is made by thresholding the score with a learnable threshold $\theta$: the switch is *on* when $\hat{p}^t_r >= \theta$, and is *off* otherwise. When the switch changes from *off* to *on*, a new Transient process is created for the subject. It will run until the switch turns *off*, then the persistent process again becomes the single operator.

### 3.6. Future prediction

In PTD, future features are predicted autoregressively: After running through the observed sequence $X^{1:T}$, PTD keeps unrolling to predict the requested $L$ time steps and feeds its prediction back as input to keep unrolling.

At each future time step $t$, the predictions from Persistent and Transient channels $\hat{Y}^t_{\mathcal{P}}$, $\hat{Y}^t_{\mathcal{T}}$ (Eq. (3), Eq. (8)) are combined with the priority on the Transient predictions. For a human entity, if its Transient channel is activated, the Transient prediction will be chosen; otherwise, the Persistent prediction will be used. For an object entity, if it receives an active outward Transient edge, it will take that channel's prediction. If it receives multiple outward edges, it uses the prediction from the channel with the highest transient score $\hat{p}^t_r$. Otherwise, it uses the persistent prediction by default.

### 3.7. Model Training

The model is trained end-to-end with three losses: prediction loss of humans and objects and switch loss:

$$\mathcal{L} = \lambda_h \mathcal{L}_{pred,h} + \lambda_o \mathcal{L}_{pred,o} + \lambda_{switch} \mathcal{L}_{switch}. \quad (11)$$

The **Prediction loss** $\mathcal{L}_{\text{pred}}$ measures the mismatch between predicted values $\hat{Y}$ and ground truth $Y$, implemented as their Euclidean distance:

$$\mathcal{L}_{\text{pred}} = \left\| \hat{Y}^{T+1:T+L} - Y^{T+1:T+L} \right\|_2. \quad (12)$$

Even though the Transient switch can be implicitly trained with the prediction loss, the gradient flowing through this binary gate can be weak. To directly supervise the Transient switch, we further introduce the **Switch loss**:

$$\mathcal{L}_{switch} = \text{BCE}\left(\hat{P}^{1:T+L}, P^{1:T+L}\right), \quad (13)$$

where $\hat{P}^t$ and $P^t$ are the predicted and ground truth switch scores (Eq. (9)) of all human entities at time step $t$.

**Setting switch ground truth label.** $P^t$ from data is an interesting modeling topic. The term represents the true moment where the human $r$ turns on their "Transient mode". A simple way to learn it is through self-supervision using a binary label $q^t$ on whether an interaction happens at that time. In particular, it is determined by the outward edges $e^t_{r\to l}$ in the Transient graph (see Sec. 3.4): $q^t = 1$ if $\exists l : e^t_{r\to l} \in \mathcal{E}^t$, and $q^t = 0$ otherwise.

However, is this the true ground truth to human behavior switch? Humans usually foresee their interaction and change their behavior before the observable interaction occurs; therefore, $q^t$ is actually too late to turn on the Transient channel. We resolve this mismatch by using the future ground truth labels of deviations for the current label of Transient switch:

$$p^t = q^t \vee q^{t+1} \vee ... \vee q^{t+\omega}, \quad (14)$$

where $\vee$ indicates 'bit-wise or' operator, and $\omega$ is the post-dating window meta-parameter, whose values are examined in the Supp, Sec. 5.

**Multistage training.** We train our model in two stages: we first use teacher-forcing [28] to use the ground truth position as the input in the prediction stage, preventing the model from accumulating errors and facilitate faster training. Then, in the second stage, we fine-tune the model with the unrolling mechanism introduced in Sec. 3.6.

This way of training can be thought of as a curriculum learning technique [5] used in previous works [1, 25] where we initially train the model with an easy problem, then increase the difficulty of the problem in the later epochs.

### 3.8. Modeling scope and limitations

The modeling of the PTD in this paper makes an assumption that the Transient processes of different persons are independent of each other. Extensions can be made to break this assumption and support more complex cases, such as collaborative and competitive multi-agent systems.

The model also assumes that there is a hard temporal border between the persistent and transient processes, which may be an approximation as humans sometimes have a mixed-up thinking and acting mechanisms. Future works may involve allowing the two processes to take over each other more softly, hence can cover these cases.

The PTD formulation in this section is done particularly for HOI-M forecasting. Modifications may be necessary for other applications. Example adaptation for *pedestrian trajectory prediction* is given in Supp. Sec 9.

## 4. Experiments

We examine the effectiveness of PTD via quantitative evaluations, generalization trial, visual analysis, and ablation studies on two HOI-M prediction datasets: WBHM [33] and Bimanual Actions [13].

### 4.1. Preliminaries

This section describes the datasets, the baselines, and metrics in used. Further details are in Supp, Sec. 1.

#### 4.1.1 Datasets

**Whole-Body Human Motion (WBHM) Database [33]** is a large-scale dataset featuring 3D motion data of both humans and objects, which is well suited for this paper. From the raw 3D data, the selected visual features include 3D skeleton poses of 18 joints for human entities ($x \in \mathbb{R}^{54}$) and 3D bounding boxes for objects ($x \in \mathbb{R}^{24}$), sampled at 10Hz, consistent with the compared methods [8].

**Bimanual Actions Dataset [13]** contains activities of subjects using both hands to interact with different objects at the same time. Unlike WBHM with 3D geometrical features, all features here are in 2D coordinates making the dataset more challenging for motion prediction. Furthermore, the subjects in this dataset always use two arms to interact concurrently with different objects, requiring a new capability of modeling the collaboration between the arms. PTD naturally support this use case by considering each arm to be one human entity, each with features of 2D locations of the arm key points and hand bounding box.

#### 4.1.2 Compared methods and baselines

We compare PTD with the *HOI-M forecasting SOTAs*: CRNN-OPM, CRNN-OPM-LI [8] and the *pose forecasting SOTAs*: STS-GCN [40], Motion-Mixer [6]. Motion-Mixer and STS-GCN are retrained using the provided codes[1]. The others are re-implemented with the settings provided in the original papers. We also use several common baseline methods of Zero Velocity, Running avg. 2, GRU [35].

| | Human | Obj |
|---|---|---|
| Zero-Velocity | 176.45 | 128.6 |
| Running avg. 2 | 183.95 | 133.3 |
| GRU [35] | 102.86 $\pm$ 1.4 | 119.64 $\pm$ 1.6 |
| STS-GCN [40] | 101.36 $\pm$ 2.4 | - |
| Motion-Mixer [6] | 87.35 $\pm$ 1.2 | - |
| CRNN-OPM [8] | 99.01 $\pm$ 1.1 | 87.52 $\pm$ 1.6 |
| CRNN-OPM-LI [8] | 95.96 $\pm$ 1.7 | 74.27 $\pm$ 1.3 |
| **PTD (Ours)** | **85.53 $\pm$ 0.9** | **70.69 $\pm$ 0.5** |

Table 1: The average errors (mm) on WBHM after 5 independent runs. PTD outperforms other SOTAs in both human and object prediction. The errors at each time step are given in Supp, Sec. 3.

#### 4.1.3 Evaluation metric

The prediction errors at each time step are calculated as the Euclidean distances with the ground-truth for both human poses and object bounding boxes (mm for WBHM, pixel for Bimanual Action Dataset). The error for a sequence is then computed as the average errors across $L$ prediction frames. Prediction performances are reported as the mean and std of the average errors of 5 independent runs for humans and objects entities.

### 4.2. Motion forecasting on WBHM Dataset

For WBHM dataset, we follow the common evaluation protocol [8]: to observe for 1 second ($T = 10$) and predict the next 2 seconds ($L = 20$).

**Quantitative evaluation.** The average errors reported in Tab. 1 clearly indicates that PTD consistently outperforms the SoTAs in both human pose and object box forecasting. Supp, Sec.3 further reports detailed errors at each time step.

**Visual analysis.** We verify the benefit of our model by visualizing the internal output predictions and graph structures of PTD compared to CRNN-OPM-LI [8]. The upper row of Fig. 5 shows that PTD could learn to switch the mechanism from Persistent to Transient when the situation changes from interaction-free (a) to interaction-involved

[1] https://github.com/FraLuca/STSGCN
https://github.com/MotionMLP/MotionMixer

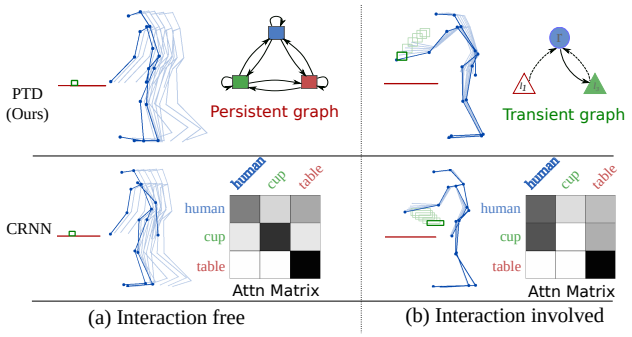



Figure 5: Visual Comparison in WBHM. When the situation changes from interaction-free (a) to interaction-involved (b), PTD *(Upper row)* switches on its Transient channel with egocentric structures and handles the interaction accurately; In contrast, CRNN-OPM-LI [8] (*Lower row*) uses a single mechanism, resulting in the sluggish adaptation of the attention map and inaccurate predictions.

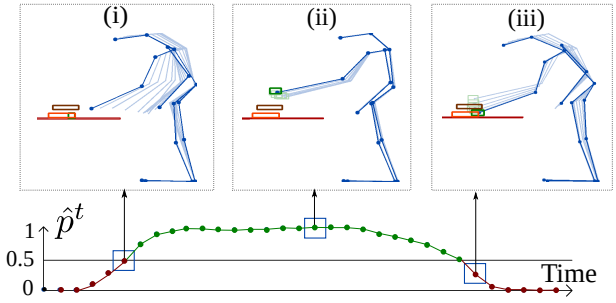


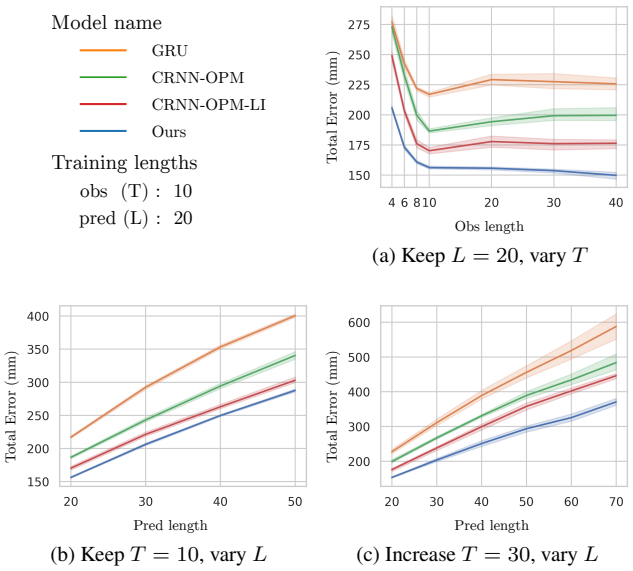

Figure 6: The switching behavior in WBHM. The Transient Switch anticipates the beginning (i), stays stable during (ii), and anticipates the end (iii) of the interaction. This early anticipation is crucial for timely process switching.

(b). The Transient graph in *(b)* reflects the interactions correctly thanks to it being trained on targeted samples.

In contrast, CRNN-OPM-LI [8] (lower row) holds on to a single global mechanism and does not evolve adequately for the swift change in the true relational topology, resulting in inaccurate and unrealistic interactions.

The operation of the Transient Switch is visualized in Fig. 6. When the interaction is about to occur (i), the switch score $\hat{p}_r^t$ (Eq. (9)) increases to reflect the prospective of the interaction. When the score passes the threshold, it switches on the Transient channel a few time-steps before the interaction can be observed, precisely as designed (Eq. (14)). After maintaining high values during interaction (ii), the score falls when it anticipates the end of the interaction, deactivating the transient channel (iii).

**Generalization analysis.** With more accurate modeling, PTD promises a greater generalization in predicting the patterns unseen in training. To evaluate such potential, we use a generalization test [24] to set up a series of experiments where PTD and other models compete on the test sequences of lengths different from the ones used in training: (1) observation length $T$ varies, (2) prediction length $L$ varies, and (3) both observation and prediction lengths varies.




(a) Keep $L = 20$, vary $T$

(b) Keep $T = 10$, vary $L$

(c) Increase $T = 30$, vary $L$

Figure 7: Generalizability test on varying observed lengths ($T$) and predicted lengths ($L$) of the sequences in testing from the ones used in training.

The results are measured as the total average errors (mm) of humans and objects and are visualized in Fig. 7. We exclude STS-GCN and MotionMixer due to their dependence on the sequence length.

**1. Varying observation length.** $T$ (Fig. 7a): Interestingly, when observing longer sequences in testing ($T > 10$), the baseline models failed to generalize and perform worse. In contrast, with mechanism switching, PTD is flexible enough to take advantage of the longer observed data to gain performance without re-training. Also, with shorter observed sequences ($T < 10$), PTD is more resilient than other models, showing its capacity of modeling the generic pattern and avoid overfitting.

**2. Predicting longer sequences.** With same observation ($T = 10$) and increased prediction length $L$ of the testing data (Fig. 7b), PTD keeps the superior performance by extrapolating well to farther future predictions.

**3. Increasing both lengths.** Finally, when both length changes (Fig. 7c), the result is consistent with case 1. and 2. where PTD shows its superior ability to take advantage of more data and generalize well to longer predictions.

| | Ablation | Human | Object |
|---|---|---|---|
| 1 | w/o Transient channel | 89.49 | 74.91 |
| 2 | w/o Persistent channel | 91.41 | 78.67 |
| 3 | w/o egocentric property | 87.90 | 72.30 |
| 4 | w/o heuristic switch | 86.44 | 74.85 |
| 5 | w/o $\gamma$ | 86.17 | 70.71 |
| 6 | w/ only $\gamma$ | 87.11 | 76.95 |
| 7 | w/o switching transient | 86.00 | 71.40 |
| 8 | w/o switch loss | 89.59 | 73.17 |
| 9 | w/o multistage training | 90.21 | 72.22 |
| | **Full PTD model** | **85.53** | **70.69** |

Table 2: Ablation studies on WBHM (avg. errors in mm.)

**Ablation studies.** We examine the roles of PTD's core components by making ablations from the model and report the performances in Tab. 2. They include:

1. **Without Transient channel**: Being alone, Persistent Channel performs significantly worse than when with its Transient partner in the duality.

2. **Without Persistent channel:** Transient Channel also struggles and gives worse performance operating alone.

3. **Without Egocentric property**, including egocentric structure (Eq. (4)) and representation (Eq. (5)), Transient channels operate on a fully-connected graph with global features. This ablated model suffers a significant drop in performance, showing the role of egocentric design.

4. **Heuristic switch:** This experiment probe the need for the *Transient Switch* (Sec. 3.5) by replacing it with a heuristic rule $\hat{p}_r^t = d_{r,min}^t \leq \beta$. This heuristic switch can still provide benefit compared to single channels (row 1,2). However, being too stiff, it cannot represent the switching patterns and fails to reach the duality's full potential.

5. **Switch without spatial discount factor:** Without $\gamma_r^t$ (Eq. (9)), the *Switch* could not respond fast enough to context changes, resulting in slightly weaker performance.

6. **Switch with only discount factor:** This quick-change factor could not do the job by itself as it is susceptible to noisy patterns, performing even worse than row 5.

7. **Without switching transient:** The transient channels is always on and may capture irrelevant patterns outside the HOI-context, leading to a drop in the performance.

8. **Without switch loss:** We study the role of the switch's direct supervision (Sec. 3.7) by setting $\lambda_{switch}$ to 0, taking $\mathcal{L}_{switch}$ out of Eq. (11). This unsupervised switch only relies on weak gradient flowing back from prediction loss and delivers significantly weakened performance.

9. **Without multistage training:** Without such a training procedure (Sec. 3.7), the model suffers from accumulating losses during early epochs and capture less accurate motion pattern, resulting in a decrease in the performance.

| | Arm Keypoints | Hand | BoxObj |
|---|---|---|---|
| Zero-Velocity | 12.11 | 21.52 | 7.02 |
| Running avg. 2 | 12.80 | 22.19 | 7.30 |
| GRU [35] | $12.37 \pm 0.4$ | $20.80 \pm 0.9$ | $7.04 \pm 0.0$ |
| STS-GCN [40] | $11.85 \pm 0.5$ | - | - |
| Motion-Mixer [6] | $11.68 \pm 0.2$ | - | - |
| CRNN-OPM [8] | $12.81 \pm 0.1$ | $21.66 \pm 0.2$ | $7.16 \pm 0.1$ |
| CRNN-OPM-LI [8] | $11.97 \pm 0.3$ | $20.13 \pm 0.2$ | $7.05 \pm 0.1$ |
| **PTD (Ours)** | $\mathbf{10.94 \pm 0.2}$ | $\mathbf{18.80 \pm 0.1}$ | $\mathbf{6.81 \pm 0.0}$ |

Table 3: The avg. errors (pixel) on Bimanual Action dataset from five different runs.

### 4.3. Motion forecasting on Bimanual Action Dataset

While WBHM is rich in patterns and is well-suited for large scale analysis, the Bimanual Action is more realistic, as its 2D RGB data are more available. Due to the increased ambiguity of the 2D data, we use longer observed (2s/20 steps) and shorter predict lengths (1s/10 steps).

**Quantitative evaluation:** The average errors measured in pixel are reported in Tab. 3. Consistent with the results in WBHM dataset, PTD outperforms other SOTAs in both human and object motion forecasting. The errors at each time steps are also detailed in Supp, Sec. 3. **Visual analysis** of PTD's operation in this dataset (similar to that of WBHM in Fig. 5 and Fig. 6) is provided in Supp, Sec 4.

### 4.4. Empirical Complexity Analysis

We did an empirical analysis on the model size and observed that *PTD has a comparable number of parameters to other methods*. Detailed numeric analysis is shown in Supp, Sec 2. This fact confirms that the good performance of PTD is caused by the new multi-mechanism scheme without the negative trade-off in computation cost.

## 5. Discussion

In this work, we have introduced a new concept of Persistent-Transient Duality to model the multi-mechanism nature of human behavior in interaction with objects. The duality is implemented into a parent-children network that demonstrates its effectiveness and generalizability through extensive evaluations in HOI-M forecasting.

Given the ubiquity of persistent-transient relationship in human behavior and the genericity of PTD design, the model can be readily extended to other applications including *pedestrian trajectory prediction* (as demonstrated in Supp., Sec 9) and potentially *social interaction modeling*, and *human-machine collaboration*. Future development may include more fluid dynamics between the two processes and allowing multiple transient instances.

## References


[1] Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13390–13400, 2021. 2, 5

[2] Icek Ajzen. *From intentions to actions: A theory of planned behavior*. Springer, 1985. 1

[3] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7144–7153, 2019. 2

[4] Adam P Baker, Matthew J Brookes, Iead A Rezek, Stephen M Smith, Timothy Behrens, Penny J Probert Smith, and Mark Woolrich. Fast transient networks in spontaneous human brain activity. *Elife*, 3:e01867, 2014. 1

[5] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 5

[6] Arij Bouazizi, Adrian Holzbock, Ulrich Kressel, Klaus Dietmayer, and Vasileios Belagiannis. Motionmixer: Mlp-based 3d human body pose forecasting. *arXiv preprint arXiv:2207.00499*, 2022. 2, 6, 8

[7] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision*, pages 387–404. Springer, 2020. 2

[8] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6992–7001, 2020. 1, 2, 6, 7, 8

[9] Qiongjie Cui and Huaijiang Sun. Towards accurate 3d human motion prediction from incomplete observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4801–4810, 2021. 2

[10] Qiongjie Cui, Huaijiang Sun, Yue Kong, Xiaoqian Zhang, and Yanmeng Li. Efficient human motion prediction using temporal convolutional generative adversarial network. *Information Sciences*, 545:427–447, 2021. 2

[11] Qiongjie Cui, Huaijiang Sun, and Fei Yang. Learning dynamic relationships for 3d human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6519–6527, 2020. 2

[12] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11467–11476, 2021. 2

[13] Christian R. G. Dreher, Mirko WÃ€chter, and Tamim Asfour. Learning object-action relations from bimanual human demonstration using graph networks. *IEEE Robotics and Automation Letters (RA-L)*, 5(1):187–194, 2020. 6

[14] T Michael Duncan and Jeffrey A Reimer. *Chemical Engineering Design and Analysis*. Cambridge University Press, 2019. 2

[15] Victor Escorcia and Juan Niebles. Spatio-temporal human-object interactions for action recognition in videos. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 508–514, 2013. 2

[16] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015. 2

[17] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. In *2017 International Conference on 3D Vision (3DV)*, pages 458–466. IEEE, 2017. 2

[18] Pallabi Ghosh, Yi Yao, Larry Davis, and Ajay Divakaran. Stacked spatio-temporal graph convolutional networks for action segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 576–585, 2020. 2

[19] Allan Greenwood. Electrical transients in power systems. 1991. 2

[20] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. Adversarial geometry-aware human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 786–803, 2018. 2

[21] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4809–4819, 2023. 2

[22] Thomas A Henzinger. The theory of hybrid automata. In *Verification of digital and hybrid systems*, pages 265–292. Springer, 2000. 2

[23] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7134–7143, 2019. 2

[24] Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020. 7

[25] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 5308–5317, 2016. 2, 5

[26] Jingwei Ji, Rishi Desai, and Juan Carlos Niebles. Detecting human-object relationships in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8106–8116, 2021. 2

[27] Jogendra Nath Kundu, Maharshi Gor, and R Venkatesh Babu. Bihmp-gan: Bidirectional 3d human motion prediction gan. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8553–8560, 2019. 2

[28] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 5

[29] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 214–223, 2020. 2

[30] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *European Conference on Computer Vision*, pages 704–721. Springer, 2020. 2

[31] Zhenguang Liu, Pengxiang Su, Shuang Wu, Xuanjing Shen, Haipeng Chen, Yanbin Hao, and Meng Wang. Motion prediction using trajectory cues. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13299–13308, 2021. 2

[32] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6437–6446, 2022. 2

[33] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The kit whole-body human motion database. In *International Conference on Advanced Robotics (ICAR)*, pages 329–336, 2015. 6

[34] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9489–9497, 2019. 2

[35] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017. 6, 8

[36] Romero Morais, Vuong Le, Svetha Venkatesh, and Truyen Tran. Learning asynchronous and sparse human-object interaction in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16041–16050, 2021. 2

[37] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697, 2019. 2

[38] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. *arXiv preprint arXiv:1805.06485*, 2018. 2

[39] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–417, 2018. 2

[40] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11209–11218, 2021. 2, 6, 8

[41] Lou Van der Sluis. *Transients in power systems*. John Wiley & Sons Ltd, 2001. 2

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[43] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 1, 3

[44] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. Heterogeneous graph attention network. In *The World Wide Web Conference*, pages 2022–2032, 2019. 3

[45] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3372–3382, 2021. 2