TransHands: Repurposing Human Pose Encoders as Hand Pose Encoders

Milo Piccioli[1], Gianluca Amprimo[1], Claudia Ferraris[2], and Gabriella Olmo[1]

[1] Politecnico di Torino, Torino, Italy s359009@studenti.polito.it, {gianluca.amprimo, gabriella.olmo}@polito.it
[2] Consiglio Nazionale delle Ricerche, Torino, Italy claudia.ferraris@cnr.it

Abstract. Lifting 3D hand poses from 2D monocular representations remains challenging due to the limited availability of large-scale, diverse 3D-annotated hand datasets, in contrast to the abundance of human body motion data. We address this limitation by transferring motion representations learned from large body pose corpora to the hand domain. We introduce **TransHands**, a backbone-agnostic transfer learning framework that enables pre-trained human motion encoders to be effectively adapted for 3D hand pose estimation from 2D pose inputs. Rather than training hand-specific biomechanical models from scratch, TransHands combines a two-stage training and fine-tuning strategy with a lightweight hand-specific input adaptation module that aligns hand kinematics with the representation space learned for full-body motion. We evaluate TransHands across four state-of-the-art motion modeling architectures, including transformer-based, graph-based, and frequency-domain models. Results demonstrate that motion priors learned from body pose data transfer consistently across architectures, yielding consistent accuracy gains, strong cross-domain generalization, particularly in challenging egocentric settings, and applicability for downstream tasks in real-world contexts.

1 Introduction

Estimating 3D hand pose from 2D joint coordinates (*2D-to-3D lifting*) is a fundamental challenge in computer vision. While significant progress has been made in full-body pose estimation, benefiting from large-scale motion capture datasets such as Human3.6M [11] and AMASS [20], hand pose estimation remains constrained by the limited availability of diverse, large-scale annotated hand datasets. This data scarcity poses a critical bottleneck: hands exhibit complex biomechanical constraints and high-frequency articulations that are difficult to capture and annotate at scale. Existing hand-specific datasets [8,43] are orders of magnitude smaller than their body pose counterparts, limiting the capacity of data-driven models to learn robust kinematic priors. Meanwhile, state-of-the-art (SOTA) human motion encoders [38, 40, 42] have demonstrated remarkable

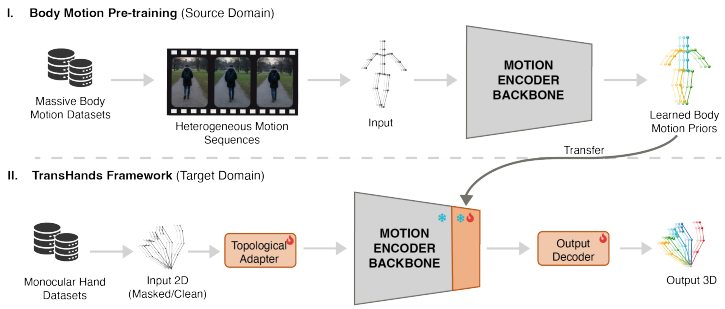

Fig. 1: Framework Overview. TransHands bridges the domain gap between body and hand motion through a backbone-agnostic transfer learning strategy. We leverage **(I) Body Motion Pre-training** to capture spatio-temporal priors from heterogeneous body datasets. These priors are repurposed for hand pose estimation via **(II) TransHands Framework**, using a learnable *Topological Adapter* in input to align monocular hand sequences with the pre-trained body manifold.

proficiency in modeling spatio-temporal dynamics from abundant body motion corpora.

We pose a natural question: *Can motion priors learned from large-scale human body datasets be transferred to model hand biomechanical knowledge?* Despite the topological differences, both domains share fundamental biomechanical principles such as temporal smoothness, joint angle constraints, and hierarchical structure. Recent work has demonstrated the effectiveness of pre-training, transfer learning and cross-domain adaption in pose estimation [13,41,42], suggesting that learned motion representations can generalize across modalities; however, a systematic study on the transfer of body motion priors to the hand domain is still lacking.

To this end, we introduce **TransHands**, a backbone-agnostic framework that enables pre-trained human motion encoders to be effectively repurposed for 3D hand pose estimation from 2D inputs. Our key insight is that the high-level temporal dynamics and biomechanical knowledge encoded in body pose models can be preserved and adapted to hand kinematics through *learnable topological alignment* rather than architectural redesign. TransHands achieves this by encapsulating frozen body encoders within lightweight adaptation modules that bridge the domain gap while preserving the original kinematic reasoning capabilities. As illustrated in Figure 1, our approach comprises three main components: (1) a **Topological Adapter** in input that aligns hand joint hierarchies with the latent space of body-centric encoders, (2) a pre-trained **motion backbone** on large body datasets that extracts spatio-temporal features, and (3) a **task-specific decoder** that reconstructs 3D hand coordinates.

We validate TransHands across four diverse motion encoding architectures spanning transformer-based [11,38], frequency-domain [40], and graph-based [37] models. Our experiments demonstrate that: (1) motion priors transfer consis-

tently across architectural families, yielding consistent accuracy gains, (2) the learned representations exhibit strong cross-domain generalization, particularly in challenging egocentric settings, and (3) the framework requires minimal modification to existing backbones, enabling rapid integration of future motion modeling advances.

In summary, our primary contributions are threefold:

- To the best of our knowledge, we present the first systematic study on transferring body motion priors to hand pose estimation across multiple SOTA architectures.
- We propose a modular framework that decouples topological adaptation from temporal modeling, enabling backbone-agnostic transfer.
- We demonstrate that biomechanical knowledge learned from body datasets provides strong inductive biases for modeling hand kinematics, opening new directions for data-efficient hand pose estimation and downstream applications in gesture recognition.

2 Related Works

Early learning-based approaches formulated 2D-to-3D pose uplifting as a direct regression, showing that even simple feed-forward models can exploit geometric consistency and temporal smoothness when trained with paired 2D–3D data [7,21]. The emergence of large-scale human motion datasets enabled a shift toward data-driven learning of kinematic priors. Datasets such as AMASS [20] facilitated the pretraining of spatio-temporal motion encoders that implicitly capture biomechanical properties, including joint coordination and long-range temporal dependencies. Building on this paradigm, recent work has demonstrated the effectiveness of powerful sequence models for 2D-to-3D lifting, including spatio-temporal transformers [38, 42], frequency-domain formulations [40], and graph-based or hybrid architectures leveraging skeletal structure as an inductive bias [23, 37]. These approaches encode biomechanical knowledge implicitly through learned motion dynamics rather than explicit constraints. In parallel, biomechanical plausibility has been addressed through model-based lifting approaches that fit parametric human models to monocular observations. Statistical body models such as SMPL [16] and GHUM [36] provide strong pose and shape priors and are widely used for monocular 3D reconstruction and tracking, including real-time systems such as BlazePose GHUM Holistic [10]. For hands, analogous constraints are commonly imposed through the MANO statistical model [29], to ensure anatomical validity of estimated coordinates [8, 43]. While effective, these model-based methods rely on carefully engineered optimization pipelines and explicit skeletal modeling, limiting flexibility and cross-domain adaptability. Despite the success of data-driven motion encoders for full-body lifting, hand pose estimation remains comparatively data-limited. Transfer learning offers a natural alternative: pre-training on large body datasets followed by downstream adaptation has shown consistent benefits in tasks related to pose estimation [31, 42]. To date, these approaches have largely assumed a

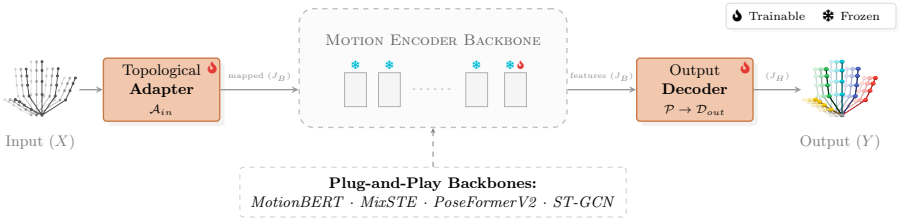

Fig. 2: Overview of the TransHands Framework. To bridge the domain gap between body and hand topology, we employ a learnable **Topological Adapter** ($\mathcal{A}_{in}$) in input that aligns hand joints with the latent space of a pre-trained **Motion Encoder Backbone**. The architecture adopts a *Partial Unfreezing* strategy: initially, the entire backbone is frozen (❄) to leverage pre-trained human motion priors (Stage 1); subsequently, the final block is unfrozen (🔥) to specialize the features for hand-specific kinematics (Stage 2). The Output Decoder then maps these refined features to the final 3D hand coordinates.

shared skeletal topology, leaving open whether motion encoders pre-trained for body pose lifting can be systematically repurposed for hand pose estimation despite substantial topological mismatch. Our work addresses this gap by reframing hand pose estimation as a biomechanical lifting problem and adapting 2D-to-3D body motion encoders to hand kinematics via lightweight, modular alignment modules.

3 Method

3.1 Problem Formulation

We formulate 3D hand pose estimation as a sequence-to-sequence lifting task. Specifically, we focus on the geometric lifting process, assuming the availability of 2D skeletal coordinates either from off-the-shelf detectors or ground-truth projections. This allows us to isolate the motion transfer capabilities from the noise inherent in direct 2D-to-3D estimation from RGB sources. Given a sequence of input coordinates $X \in \mathbb{R}^{T \times J_H \times C_{in}}$, where T denotes the sequence length, $J_H = 21$ represents the hand joints according to the MANO [29] topology, and $C_{in} \in \{2, 3\}$ is the backbone-dependent input dimensionality, our objective is to regress the trajectory of 3D coordinates $Y \in \mathbb{R}^{T \times J_H \times 3}$.

A fundamental challenge in leveraging body motion priors for this task is the *topological mismatch*: SOTA body encoders are typically optimized on skeletal full pose structures, rendering them incompatible with hand geometry. To bridge this domain gap without retraining the kinematic engine from scratch, we propose **TransHands**, the modular framework illustrated in Figure 2. It encapsulates a frozen body encoder $\mathcal{E}_{\text{body}}$ within learnable adaptation layers. The mapping function is formalized as:

$$\hat{Y} = \mathcal{D}_{\text{out}}\left(\mathcal{P}\left(\mathcal{E}_{\text{body}}\left(\mathcal{A}_{\text{in}}(X)\right)\right)\right) \tag{1}$$

Here, $\mathcal{A}_{\text{in}}$ projects the hand topology into the body's latent kinematic space, $\mathcal{P}$ aligns the extracted features via projection, and $\mathcal{D}_{\text{out}}$ decodes the processed representations back to the target $\mathbb{R}^3$ hand manifold. This formulation effectively decouples the learning of temporal dynamics, which is handled by the pre-trained $\mathcal{E}_{\text{body}}$, from the spatial adaptation required for the new topology.

3.2 Topological Adapter

The primary function of the Topological Adapter ($\mathcal{A}_{in}$) is to align hand kinematics with the representation space learned for full-body motion, resolving the dimensional and topological discrepancies between hand joints and body-centric pre-trained encoders. While SOTA body pose estimators are typically optimized on skeletal structures with $J_B = 17$ joints, as in Human3.6M [11] topology or $J_B = 18$, as in ST-GCN [37], standard hand models such as MANO define a distinct topology consisting of $J_H = 21$ joints. Direct application of the frozen body encoder $\mathcal{E}_{\text{body}}$ is therefore infeasible due to this mismatch ($J_H \neq J_B$). Moreover, the two graphs share no common joints except for the wrist, which, however, plays significantly different roles in the two kinematic graphs (i.e., the root node in the hand graph vs. a distal node in the body pose graph).

To address this, we formulate the topological transformation as a continuous dynamic process using a Neural Ordinary Differential Equation (Neural ODE) [5]. Given the flattened input coordinates $x \in \mathbb{R}^{J_H \cdot C_{in}}$, an initial encoder maps them into a latent state $z(0) \in \mathbb{R}^{d_{emb}}$, where d_{emb} is the latent dimension. The continuous transformation of the hidden state is then defined by the ODE:

$$\frac{dz(t)}{dt} = f_\theta(z(t), t) \tag{2}$$

where f_θ is parameterized by a continuous dynamics neural network equipped with time-embeddings. We solve the initial value problem using a fixed-step fourth-order Runge–Kutta (RK4) integrator over the integration interval $t \in [0, 1]$ to obtain the terminal state $z(1)$. Finally, a decoder maps $z(1)$ into the compatibility vector $z_{body} \in \mathbb{R}^{J_B \cdot C_{enc}}$ required by the backbone. This continuous flow formulation allows for a highly non-linear topological alignment without suffering from the rigidity of discrete linear projections.

Critically, $\mathcal{A}_{in}$ acts as a **dynamic interface** that abstracts the underlying backbone requirements from the raw input data:

- **Channel Alignment:** The adapter adjusts the channel dimensionality C_{enc} to match the specific requirement of the selected backbone. For instance, it synthesizes a confidence channel ($C = 3$) for MotionBERT [42], while preserving raw coordinates ($C = 2$) for architectures like MixSTE [38] and PoseFormerV2 [40].
- **Latent Manifold Alignment:** By optimizing the ODE trajectory, the adapter smoothly projects the hand topology onto the body's latent kinematic manifold, effectively reusing pre-trained priors without structural modifications to the frozen $\mathcal{E}_{\text{body}}$.

3.3 Motion Encoder Backbones

The core component of our framework is the kinematic encoder $\mathcal{E}_{\text{body}}$, which processes the adapted sequence $Z_{\text{body}} = \mathcal{A}_{in}(X)$ to extract spatio-temporal features. To enforce the transfer of high-level motion priors (e.g., velocity continuity, biomechanical constraints), we adopt a *Partial Unfreezing* strategy. Instead of fine-tuning, we strategically enable gradients only for the deepest layers of the backbone, together with its embedding and output layers. This design choice ensures that the model reuses learned body dynamics while allowing the final representation stages to adapt to hand-specific micro-articulations. We validate the versatility of our approach by integrating four distinct SOTA architectures serving as Motion Encoder Backbone.

Spatio-Temporal Transformers. We integrate **MotionBERT** [42], utilizing its Dual-Stream Spatio-Temporal Transformer (DSTformer) to capture long-range dependencies. Additionally, we integrate **MixSTE** [38], a Seq2Seq model that decouples spatial and temporal attention, allowing us to evaluate the benefits of fine-grained frame-to-frame modeling for hand micro-articulations.

Frequency-Domain Transformers. To address long-sequence efficiency, we incorporate **PoseFormerV2** [40], which operates in the frequency domain via Discrete Cosine Transform (DCT). This architecture is particularly relevant for analyzing the spectral properties of rapid hand gestures.

Graph-Based Architectures. To benchmark against purely convolutional approaches, we integrate the standalone implementation of **ST-GCN** [37], demonstrating the framework's adaptability to non-transformer backbones. For this backbone the transfer is only partial: the temporal convolutions are initialized from the Kinetics checkpoint, whereas the graph branch (spatial convolutions, edge importance, and adaptive adjacency) is trained from scratch.

3.4 Latent Space Alignment

Since different backbones produce feature maps with varying distributions and sequential dynamics (covariate shift), we introduce a projection module $\mathcal{P}$ to align the encoder output f_{enc} with the decoding task. Instead of a standard linear bottleneck or attention-based transformer, we utilize a multi-scale retention mechanism [32]. Formally:

$$f_{proj} = W_p \cdot \text{LayerNorm}(f_{enc}) \tag{3}$$

Subsequently, the sequence is processed by a stack of retention layers with residual connections:

$$\tilde{f} = f_{proj} + \text{Retention}(\text{LayerNorm}(f_{proj})), \quad f_{ret} = \tilde{f} + \text{FFN}(\tilde{f}) \tag{4}$$

where the retention operator applies learned exponential temporal decay to the position weights. The projection head standardizes heterogeneous encoder outputs to a common 256-dimensional latent before the decoding stage, preventing feature distortion during fine-tuning [12].

3.5 Task-Specific Output Decoder

The reconstruction of the target 3D pose is performed by the module $\mathcal{D}_{out}$, a task-specific regressor. We design this decoder as a deep Multilayer Perceptron (MLP) utilizing Layer Normalization [1] for stable training across sequences. The architecture follows:

$$h_1 = \text{ReLU}(\text{LN}(W_1 f_{ret} + b_1)), \quad h_2 = \text{ReLU}(\text{LN}(W_2 h_1 + b_2)), \tag{5}$$

$$\hat{Y} = \text{Reshape}_{J_H \times 3}(W_3 h_2 + b_3) \tag{6}$$

where $W_1 \in \mathbb{R}^{512\times256}$, $W_2 \in \mathbb{R}^{512\times512}$, and $W_3 \in \mathbb{R}^{63\times512}$ (for $J_H = 21$ joints). The inclusion of Layer Normalization combined with sequential hidden expansion is critical to prevent training instabilities when decoding the narrow bottleneck. The depth of this module (three linear transformations mapping $256 \to 512 \to 512 \to 63$) is essential to model the non-linear inverse kinematics required to recover complex finger articulations from the compressed latent representation.

4 Experiments

4.1 Setup

Datasets and Data Rationale. To validate the effectiveness of TransHands, we adopt **Re:InterHand** [24] as our primary benchmark, a large-scale dataset of relighted 3D interacting hands with high-fidelity MANO-based annotations captured in a controlled multi-view studio setting. To assess robustness across diverse scenarios, we employ **AssemblyHands** [26] (egocentric) and **GigaHands** [9] (bimanual hand activities). These datasets provide high-fidelity 2D ground-truth alongside their 3D annotations, letting us evaluate TransHands as a 2D-to-3D lifting module in isolation from 2D-detection errors. This isolates whether a body-motion encoder's semantic motion priors transfer to accurate 3D hand lifting, independent of detection noise. While this offers a clean controlled setting, TransHands remains inherently *plug-and-play*, allowing the integration of any 2D tracker for real-world deployment (Section 5.1).

Transfer Protocol. To evaluate our transfer strategy, we adopt a three-stage protocol that progressively relaxes the backbone from fully frozen to partially unfrozen, balancing prior preservation with domain adaptation.

- **Stage 1: Geometric Alignment (S1).** We initialize the backbone $\mathcal{E}_{body}$ with pre-trained body pose weights, keeping it *frozen*. Only the topological adapter, the projection head, and the output decoder are optimized on Re:InterHand. Additionally, during this stage a self-supervised Lifted Masked Motion Modelling (L-MMM) objective (detailed in the supplementary materials) is activated to encourage robustness to occlusions by reconstructing full 3D poses from partially masked 2D inputs.
- **Stage 2-A: In-Domain Refinement (S2-A).** Starting from the **S1** checkpoint, we selectively unfreeze the last transformer blocks and optimize on

Re:InterHand using a supervised *Weighted MPJPE* loss ($\mathcal{L}_{sup}$), adapting the model to hand micro-dynamics while preserving the robust representations acquired during S1.
– **Stage 2-B: Multi-Dataset Generalization (S2-B).** To address cross-domain generalization, we specialize the S1 model (aligned on Re:InterHand) on the target domains: **AssemblyHands** and **GigaHands**, using the same supervised objective as in S2-A. We employ a balanced sampling strategy [27, 42] to ensure uniform exposure to all domains during this adaptation phase.

Implementation Details. All models are implemented in PyTorch and optimized using AdamW [18]. We use a learning rate of 10^{-4} for Stage 1. For Stage 2 (both A and B), we apply a differential learning rate: 5×10^{-6} for the unfrozen backbone layers to preserve pre-trained features, and 5×10^{-5} for adapters. A Cosine Warmup scheduler [17,35] stabilizes gradients at the start of each phase. To facilitate reproducibility, we release our source code[3], including the training configurations for all backbones and stages.

Ablation Studies. Section 4.3 presents key ablations on our transfer strategy and its behaviour in low-data regimes, as well as on the key architectural components. Further analyses on the progressive training strategy and the use of the L-MMM objective in Stage S1 are provided in the supplementary materials.

4.2 Main Results: Backbone Comparison

Table 1 reports adaptation efficiency and refinement on Re:InterHand, comparing frozen adaptation (S1) with partial unfreezing (S2-A). To assess the framework's scalability, Table 2 reports the cross-domain generalization performance; here we evaluate the models in two distinct scenarios: the zero-shot capability of the frozen body-centric prior (S1) and the performance of multi-dataset training (S2-B) on the same validation sets. We report the Mean Per-Joint Position Error (**MPJPE**) [21], computed as the average Euclidean distance between predicted and ground-truth joints after root alignment, to measure absolute positional accuracy. As 2D lifting is scale-ambiguous, lifting predictions are normalized to a reference bone length, while the RGB baselines predict metric scale directly. To assess the quality of the reconstructed structure independent of global factors, we also report the Procrustes-Aligned MPJPE (**PA-MPJPE**), which applies a rigid alignment (translation and rotation) to the prediction before error computation. Additionally, qualitative visualizations illustrating the reconstruction quality across these domains are provided in the supplementary materials.

Transformer-based models dominate. Transformer-based architectures significantly outperform Graph Convolutional Networks in the final refinement stage. **MixSTE** stands as the best encoder in our benchmark, achieving the lowest error of **7.16 mm**. This represents a substantial improvement (+21.34%) over its already strong frozen baseline. **MotionBERT** follows closely, confirming the robustness of its dual-stream spatio-temporal attention mechanism with a remarkable gain of +21.45%, reaching 7.47 mm.

[3] Code available at: `https://github.com/picciolimilo/TransHands`

Table 1: Main Results In-Domain. We report parameter efficiency and metric progression on the **Re:InterHand** [24] dataset for the four backbones ([37, 38, 40, 42]). **Total** denotes the backbone size, while **Tr-S1** and **Tr-S2** represent the trainable parameters in Stage 1 and 2 respectively. **S1** refers to frozen adaptation (alignment), and **S2-A** to partial unfreezing (refinement). **Gain** quantifies the relative MPJPE improvement from S1 to S2-A.

Backbone	Params (M)			MPJPE ↓		PA-MPJPE ↓		Gain (%)
	Total	Tr-S1	Tr-S2	S1	S2-A	S1	S2-A	
MotionBERT	65.70	2.24	15.00	9.51	7.47	6.81	5.62	+21.45
MixSTE	36.02	2.24	10.79	**9.09**	**7.16**	**6.69**	**5.41**	+21.34
PoseFormerV2	16.62	2.24	5.89	12.62	8.01	7.88	5.88	+36.53
ST-GCN	6.61	4.10	5.00	12.95	11.56	8.32	7.66	+10.73

Table 2: Multi-Dataset Generalization. We compare the zero-shot performance of the frozen backbone (**S1**) against the multi-dataset refined model (**S2-B**) on **AssemblyHands** [26] and **GigaHands** [9].

	AssemblyHands				GigaHands			
	MPJPE ↓		PA-MPJPE ↓		MPJPE ↓		PA-MPJPE ↓	
Backbone	S1	S2-B	S1	S2-B	S1	S2-B	S1	S2-B
MotionBERT	42.87	11.73	22.06	6.96	139.00	**2.21**	29.80	**1.89**
MixSTE	**38.58**	**10.92**	**21.71**	**6.47**	121.20	2.36	28.27	2.03
PoseFormerV2	39.70	13.05	21.93	7.34	**115.90**	2.82	27.68	2.37
ST-GCN	54.80	20.66	24.15	11.34	137.02	6.12	**26.99**	4.38

Strongest Refinement Margin. A key observation from Table 1 is that the highest adaptation gain does not strictly belong to the models with the largest capacity. **PoseFormerV2**, despite having significantly fewer parameters and a relatively higher initial error in S1 (12.62 mm) compared to the larger transformers in our benchmark, achieves the highest relative improvement in the benchmark (+36.53%). This suggests that frequency-domain transformers possess a latent plasticity that, once unlocked in S2-A, allows for rapid convergence to the target domain geometry, successfully driving the error down to 8.01 mm.

Architecture Limits. ST-GCN, the smallest and only non-transformer model in the benchmark, appears to hit a capacity ceiling. Despite unlocking a parameter budget comparable to lightweight transformers during Stage 2 (5.00 M), it yields the highest S2-A error (11.56 mm) and the lowest overall gain (+10.73%). This suggests that while partial refinement is highly effective for self-attention mechanisms, GCNs may struggle to handle significant domain shifts effectively through partial unfreezing alone.

Cross-Domain Generalization. Table 2 reveals a compelling trade-off between intrinsic robustness and adaptation. In the zero-shot setting (S1) on AssemblyHands, pure transformers exhibit superior robustness: **MixSTE** achieves the best initial alignment (**38.58 mm**), suggesting that its representation effec-

tively encodes relative geometry, making it less sensitive to the global coordinate shifts typical of unseen domains. Conversely, on GigaHands, all architectures yield high initial MPJPE values (> 115 mm). However, their **PA-MPJPE** remains competitive ($\approx 24 - 30$ mm). This indicates that the high error stems from *global rigid misalignment* (i.e., scale/rotation discrepancies between body and hand spaces) rather than structural distortion. The multi-dataset adaptation (S2-B) resolves this misalignment, with high-capacity transformers leveraging their latent plasticity to recover the target distribution. **MixSTE** reaches **10.92 mm** on AssemblyHands, confirming that model capacity is a key driver for resolving geometric ambiguities. On GigaHands, the same adaptation drives all backbones from their initial > 115 mm down to a few millimetres (e.g., **2.21 mm** for **MotionBERT**); this drop reflects the intrinsic nature of the domain rather than capacity alone, as GigaHands features more static, less articulated sequences with temporally smooth annotations, making it an easier target once the global misalignment is removed.

4.3 Ablation Studies

The Value of Transfer Learning We hypothesize that the high performance of our framework relies on *Transfer Learning*. Specifically, the adaptation of robust kinematic priors learned from full-body pose estimation, rather than solely learning from the target hand dataset. To validate this, we test whether our pipeline achieves comparable results learning *from scratch*.

Setup. We compare two initialization strategies using exactly the same two-stage training protocol on the four motion encoder architectures:

1. **Random Initialization (Baseline):** The encoder is initialized with standard random weights. This forces the model to learn 3D structures solely from the Re:InterHand dataset.
2. **Body-Prior Initialization (Ours):** The encoder inherits pre-trained body weights, transferring learned spatial-temporal features and structural priors.

To further evaluate the effectiveness of the transferred structural motion priors, we then assess TransHands in a low-data setting. We train the framework with PoseFormerV2 backbone using incrementally reduced fractions of the target dataset (from 100% down to 10%) and compare it against the scratch baseline.

Results. Table 3 highlights the performance gap across all stages. During S1 (Frozen), the random baseline struggles with a higher error across all backbones, as a frozen random encoder produces noise that the projection head cannot map into valid poses. In contrast, our method starts with an initial semantic understanding, yielding an immediate MPJPE between 12.95 and 9.09 mm for all backbones. When the encoder is allowed to learn during S2 (Partial Unfreeze), the random baseline recovers significantly, bringing the error down. However, it hits a performance ceiling. Our method achieves a superior final result compared to the baseline for all the backbones. PoseFormerV2 shows the strongest benefit from transfer learning, achieving a remarkable 35.7% performance gain

Table 3: Transfer Learning Impact. Comparison between random initialization (*Scratch*) and our **TransHands** framework (*Ours*) on Re:InterHand. The **Gain** column highlights the MPJPE reduction achieved at S2 through motion prior transfer.

Backbone	MPJPE (mm) ↓				PA-MPJPE (mm) ↓				S2 Gain (%)
	S1 (Frz)		S2 (Part)		S1 (Frz)		S2 (Part)		
	Scratch	*Ours*	*Scratch*	*Ours*	*Scratch*	*Ours*	*Scratch*	*Ours*	
PoseFormerV2	19.67	12.62	12.46	8.01	10.97	7.88	8.58	5.88	**35.7%**
MotionBERT	10.76	9.51	8.40	7.47	8.00	6.81	6.47	5.62	11.1%
ST-GCN	14.71	12.95	13.36	11.56	9.31	8.32	8.81	7.66	13.5%
MixSTE	12.71	**9.09**	8.88	**7.16**	9.37	**6.69**	7.00	**5.41**	19.5%

Table 4: Architectural Components Ablation (PoseFormerV2). Impact of replacing standard MLP and Linear layers with our proposed Neural ODE adapter and RetNet projection on Re:InterHand

Protocol	ODE Adap.	RetNet Proj.	MPJPE ↓		PA-MPJPE ↓	
			S1	S2-A	S1	S2-A
Baseline (MLP + Lin.)			18.45	11.69	11.04	8.28
+ RetNet Proj.		✓	16.55	11.65	9.51	8.45
+ ODE Adapter	✓		13.72	9.47	8.81	6.97
Ours (Full)	✓	✓	**12.62**	**8.01**	**7.88**	**5.88**

in stage S2. The data-efficiency analysis using this backbone further shows that TransHands matches the performance of the scratch baseline trained on the full dataset (12.46 mm) while using less than 25% of the available annotations. In other words, TransHands reduces the annotation requirement by more than 75% for PoseFormerV2, indicating that the transferred priors may effectively compensate for the limited availability of large-scale 3D hand datasets.

Architectural components We introduce a Neural ODE [5] adapter and a RetNet [32] projection head to bridge the topological and dimensional gaps between hand and body sequences. Here, we isolate their contributions against standard baselines.

Setup. Using PoseFormerV2 and our two-stage training, we evaluate four configurations:

1. **Baseline (MLP + Lin.):** Standard MLP adapter for input topology mapping and a Linear layer for output projection.
2. **+ RetNet Proj.:** MLP adapter paired with RetNet projection to preserve temporal context.
3. **+ ODE Adap.:** Continuous Neural ODE adapter paired with the baseline Linear projection.
4. **Ours (Full):** The complete TransHands framework combining both the ODE adapter and the RetNet projection.

Results. Table 4 isolates the contribution of each module. The baseline (MLP adapter + linear projection) reaches 11.69 mm at S2-A. Adding the Ret-

Net projection alone yields a marginal change (11.65 mm), indicating that a stronger projection head provides little benefit without a matching input representation. The Neural ODE adapter, by contrast, drives the largest isolated drop (9.47 mm), as its continuous-depth formulation better models the highly non-linear mapping between hand and body joints. The two modules are complementary: combined, they reach the lowest error (**8.01 mm**), a further 1.46 mm over the ODE alone, showing that the RetNet projection becomes effective once the ODE aligns the input to the body manifold.

5 Robustness and Generalization in Real-World Contexts

In real-world deployment, hand pose estimation models must remain reliable under severe domain shift, handle noisy 2D detector inputs from unconstrained environments, and support downstream tasks such as gesture recognition under privacy-preserving constraints, that may preclude the retention of raw RGB data. This section compares TransHands along all these axes with respect to the current state of the art. First, we evaluate cross-site generalization under significant domain shift, testing the framework's resilience to realistic 2D detector noise without any target-domain supervision (Section 5.1). Second, we validate that our compact, skeleton-only representation encodes semantically rich and transferable motion priors through a downstream gesture recognition task across both exo-centric and egocentric vision (Section 5.2).

5.1 RGB-to-3D Hand Pose Estimation

We evaluate robustness under domain shift, from Re:InterHand to the unseen AssemblyHands domain, using an off-the-shelf 2D keypoints detector as input to the TransHands lifting module.

Zero-Shot Generalization. We evaluate our best S2-A models (MixSTE and MotionBERT backbones) directly on the AssemblyHands validation set, without fine-tuning; for each hand we aggregate the four ego camera views, taking the prediction from the first view that observes it. Unlike the in-domain adaptation of Table 2, the S2-A model (trained only on Re:InterHand) here receives noisy keypoints detected using MediaPipe [19] rather than ground-truth 2D annotations. Table 5 shows that our approach reaches an MPJPE of **92.97 mm** and a PA-MPJPE of 23.82 mm with the MotionBERT backbone (93.44 mm / 24.28 mm with MixSTE).

Despite relying only on sparse 2D keypoints from an off-the-shelf detector, our model improves on the reported MPJPE of the RGB baselines ArcticNet-SF [8] (110.76 mm). On a root-aligned basis, TransHands (92.97 mm) is competitive with the dedicated egocentric method V-HPOT [25] (92.09 mm). This compares different regimes rather than claiming superiority: V-HPOT is an end-to-end RGB model that recovers metric depth and adapts at test time, whereas TransHands lifts noisy off-the-shelf 2D keypoints zero-shot, reaching comparable accuracy without any target-domain adaptation. Moreover, this result improves

Table 5: Zero-Shot Robustness Analysis on AssemblyHands. Evaluation on the unseen AssemblyHands domain. *Lift* denotes methods that estimate 2D keypoints with an off-the-shelf detector and then uplift them to 3D (our pipeline and MediaPipe 3D, which share the same 2D input); *RGB* denotes end-to-end methods that estimate 3D directly from the image.

Method	Type	MPJPE ↓	PA-MPJPE ↓
ArcticNet-SF [8]	RGB	110.76†	–
V-HPOT [25]	RGB	**92.09**‡	–
SMPLer-X [3]	RGB	98.20	**19.40**
MediaPipe 3D [19]	Lift	130.77	29.38
Ours (MotionBERT)	Lift	92.97	23.82
Ours (MixSTE)	Lift	93.44	24.28

† As reported in [28]; ‡ As reported in [25].

Table 6: Downstream Gesture Recognition. Evaluation of the frozen S2-B representations on Jester and EgoGesture datasets.

Backbone	Jester		EgoGesture	
	Top-1 (%) ↑	Top-5 (%) ↑	Top-1 (%) ↑	Top-5 (%) ↑
MotionBERT	88.44	**95.80**	**88.29**	97.36
MixSTE	**88.72**	95.76	88.08	**97.53**

the performance, on the same 2D keypoints, of MediaPipe 3D which applies the GHUM model [36] to perform 2D-to-3D uplifting. Mediapipe 3D reaches 130.77 mm MPJPE; our pipeline recovers about 38 mm, denoting a better understanding of the 3D hand geometry compared to the uplifting strategy employed by MediaPipe. We also compare with SMPLer-X [3], a SOTA whole-body mesh model, which motivates a dedicated hand-pose task in egocentric settings where full-body visibility is limited. Despite this limitation, SMPLer-X attains a lower PA-MPJPE (19.40 mm): its parametric SMPL-X prior enforces anatomically valid hand shapes and is, by construction, immune to the depth-sign ambiguity of monocular lifting. Our pipeline, however, achieves a lower MPJPE (92.97 mm vs. 98.20 mm) with approximately one-tenth as many parameters (67.7M, including the MediaPipe 2D hand detector, vs. 703.58M for the end-to-end SMPLer-X pipeline). This parameter comparison is contextual, as SMPLer-X targets the much broader task of full-body mesh recovery.

5.2 Downstream Utility: Gesture Recognition

To evaluate the generalizability and semantic richness of the spatial-temporal representations learned by our framework, we assess TransHands on a down-

stream action recognition task. Specifically, we extract the latent embeddings from the S2-B backbones and use them to classify hand gestures.

Experimental Setup. We freeze the pre-trained encoder and projection of TransHands, training only the input adapter, to strictly evaluate the quality of the pre-computed features. The latent representations are then processed by a Temporal Convolutional Network (TCN) [2] classification head. The network is trained using a Gesture Classification Loss combining Focal Loss [15] and Label Smoothing [33] to handle class imbalance. Optimization is performed via AdamW with a Warmup-Cosine learning rate schedule.

Datasets and Domains. We evaluate the representations on two distinct domains: **Jester** [22], a large-scale dataset of human hand gestures recorded from a fixed, third-person webcam (static domain), and **EgoGesture** [39], a challenging dataset of egocentric gestures recorded from wearable cameras, featuring severe perspective distortions and head motion (dynamic first-person domain).

Results. Table 6 summarizes the Top-1 and Top-5 accuracy of the frozen representations extracted from MotionBERT and MixSTE. Both models demonstrate cross-domain transferability. Reaching 88% on the Jester dataset and ∼88% on the EgoGesture dataset with a *frozen* lifting encoder confirms that our learned motion priors encode high-level semantic dynamics.

Comparison with Literature. To contextualize these results, full-RGB models on **Jester** (e.g., TSM [14]) exceed 95% accuracy by exploiting contextual cues at a high computational cost. In contrast to the pose-only method of Schlüsener et al. [30] (81.2% on a 10-class subset), TransHands achieves over 88% on the full 27-class vocabulary. On **EgoGesture**, our model (∼88%) outperforms RGB baselines like VGG16+LSTM [6] (74.7%) and C3D [34] (86.4%). While complex RGB-Depth ensembles [4] reach 92.2%, TransHands provides a highly efficient and privacy-preserving alternative.

5.3 Limitations and Future Work

Our results establish the benefit of body-motion transfer within our multi-dataset and multi-backbone framework. Further comparisons with hand-specific 2D-to-3D lifting methods are left open for future work. In addition, our evaluation relies mostly on ground-truth 2D keypoints, real-world accuracy remains bounded by the 2D detector (Section 5.1). Further integration between 2D estimation and 3D-uplifting through TransHands is left for future analysis.

6 Conclusion

We presented a framework that repurposes body-motion encoders for 3D hand pose lifting from monocular 2D inputs, decoupling topological adaptation from temporal modeling to enable cross-topology transfer with minimal architectural change. Across multiple encoders, body-derived priors provide strong inductive biases for data-efficient and robust hand pose lifting, highlighting the potential of motion representation reuse across articulated domains and its relevance for downstream applications in real-world contexts.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization (2016), https://arxiv.org/abs/1607.06450
2. Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling (2018), https://arxiv.org/abs/1803.01271
3. Cai, Z., Yin, W., Zeng, A., Wei, C., Sun, Q., Wang, Y., Pang, H.E., Mei, H., Zhang, M., Zhang, L., Loy, C.C., Yang, L., Liu, Z.: Smpler-x: scaling up expressive human pose and shape estimation. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (2023)
4. Cao, C., Zhang, Y., Wu, Y., Lu, H., Cheng, J.: Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 3783–3791 (2017). https://doi.org/10.1109/ICCV.2017.406
5. Chen, R.T.Q., Rubanova, Y., Bettencourt, J., Duvenaud, D.: Neural ordinary differential equations. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 6572–6583 (2018)
6. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Darrell, T., Saenko, K.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2625–2634 (2015). https://doi.org/10.1109/CVPR.2015.7298878
7. Drover, D., M. V, R., Chen, C.H., Agrawal, A., Tyagi, A., Huynh, C.P.: Can 3d pose be learned from 2d projections alone? In: Computer Vision – ECCV 2018 Workshops: Munich, Germany, September 8-14, 2018, Proceedings, Part IV. p. 78–94. Springer-Verlag, Berlin, Heidelberg (2018). https://doi.org/10.1007/978-3-030-11018-5_7, https://doi.org/10.1007/978-3-030-11018-5_7
8. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: Arctic: A dataset for dexterous bimanual hand-object manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12943–12954 (2023). https://doi.org/10.1109/CVPR52729.2023.01244
9. Fu, R., Zhang, D., Jiang, A., Fu, W., Funk, A., Ritchie, D., Sridhar, S.: Gigahands: A massive annotated dataset of bimanual hand activities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
10. Grishchenko, I., Bazarevsky, V., Zanfir, A., Bazavan, E.G., Zanfir, M., Yee, R., Raveendran, K., Zhdanovich, M., Grundmann, M., Sminchisescu, C.: Blazepose ghum holistic: Real-time 3d human landmarks and pose estimation. arXiv preprint arXiv:2206.11678 (2022)
11. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1325–1339 (2014). https://doi.org/10.1109/TPAMI.2013.248
12. Kumar, A., Raghunathan, A., Jones, R., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. In: International Conference on Learning Representations (ICLR) (2022)

13. Li, Z., Zhang, Y., Lin, J., Qin, H., Gu, J., Yuan, X., Kong, L., Yang, X.: Binaryhpe: 3d human pose and shape estimation via binarization (2025), https://arxiv.org/abs/2311.14323
14. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7082–7092 (2019). https://doi.org/10.1109/ICCV.2019.00718
15. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2999–3007 (2017). https://doi.org/10.1109/ICCV.2017.324
16. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM Transactions on Graphics **34**(6), 1–16 (2015). https://doi.org/10.1145/2816795.2818013
17. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (ICLR) (2017)
18. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)
19. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M., Lee, J., Chang, W.T., Hua, W., Georg, M., Grundmann, M.: Mediapipe: A framework for perceiving and processing reality. In: Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019 (2019), https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf
20. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019). https://doi.org/10.1109/ICCV.2019.00554
21. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2659–2668 (2017). https://doi.org/10.1109/ICCV.2017.288
22. Materzynska, J., Berger, G., Bax, I., Memisevic, R.: The jester dataset: A large-scale video dataset of human gestures. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 2874–2882 (2019). https://doi.org/10.1109/ICCVW.2019.00349
23. Mehraban, S., Adeli, V., Taati, B.: Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6905–6915 (2024). https://doi.org/10.1109/WACV57701.2024.00677
24. Moon, G., Saito, S., Xu, W., Joshi, R., Buffalini, J., Bellan, H., Rosen, N., Richardson, J., Mize, M., de Bree, P., Simon, T., Peng, B., Garg, S., McPhail, K., Shiratori, T.: A dataset of relighted 3d interacting hands. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track (2023)
25. Mucha, W., Wray, M., Kampel, M.: Towards egocentric 3d hand pose estimation in unseen domains. In: 2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5776–5786 (2026). https://doi.org/10.1109/WACV61042.2026.00560
26. 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 Recog-

nition (CVPR). pp. 12999–13008 (2023). https://doi.org/10.1109/CVPR52729.2023.01249
27. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3d with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9826–9836 (2024)
28. Prakash, A., Tu, R., Chang, M., Gupta, S.: 3d hand pose estimation in everyday egocentric images. In: European Conference on Computer Vision (ECCV) (2024)
29. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: modeling and capturing hands and bodies together. ACM Transactions on Graphics **36**(6), 1–17 (Nov 2017). https://doi.org/10.1145/3130800.3130883, http://dx.doi.org/10.1145/3130800.3130883
30. Schlüsener, N., Bücker, M.: Fast learning of dynamic hand gesture recognition with few-shot learning models (2022), https://arxiv.org/abs/2212.08363
31. Shan, W., Liu, Z., Zhang, X., Wang, S., Ma, S., Gao, W.: P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In: Computer Vision – ECCV 2022. pp. 461–478. Lecture Notes in Computer Science, Springer (2022). https://doi.org/10.1007/978-3-031-20065-6_27
32. Sun, Y., Dong, L., Huang, S., Ma, S., Xia, Y., Xue, J., Wang, J., Wei, F.: Retentive network: A successor to transformer for large language models (2023), https://arxiv.org/abs/2307.08621
33. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826 (2016). https://doi.org/10.1109/CVPR.2016.308
34. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4489–4497 (2015). https://doi.org/10.1109/ICCV.2015.510
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 5998–6008 (2017)
36. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Ghum & ghuml: Generative 3d human shape and articulated pose models. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6183–6192 (2020). https://doi.org/10.1109/CVPR42600.2020.00622
37. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018). https://doi.org/10.1609/aaai.v32i1.12328
38. Zhang, J., Tu, Z., Yang, J., Chen, Y., Yuan, J.: Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13222–13232 (2022). https://doi.org/10.1109/CVPR52688.2022.01288
39. Zhang, Y., Cao, C., Cheng, J., Lu, H.: Egogesture: A new dataset and benchmark for egocentric hand gesture recognition. IEEE Transactions on Multimedia **20**(5), 1038–1050 (2018). https://doi.org/10.1109/TMM.2018.2808769
40. Zhao, Q., Zheng, C., Liu, M., Wang, P., Chen, C.: Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

(CVPR). pp. 8877–8886 (2023). https://doi.org/10.1109/CVPR52729.2023.00857
41. Zhao, Z., Yang, L., Sun, P., Hui, P., Yao, A.: Analyzing the synthetic-to-real domain gap in 3d hand pose estimation. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12255–12265 (2025). https://doi.org/10.1109/CVPR52734.2025.01144
42. Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: Motionbert: A unified perspective on learning human motion representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15039–15053 (2023). https://doi.org/10.1109/ICCV51070.2023.01385
43. Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 813–822 (2019). https://doi.org/10.1109/ICCV.2019.00090

TransHands: Repurposing Human Pose Encoders as Hand Pose Encoders

Supplementary Material

Milo Piccioli[1], Gianluca Amprimo[1], Claudia Ferraris[2], and Gabriella Olmo[1]

[1] Politecnico di Torino, Torino, Italy s359009@studenti.polito.it, {gianluca.amprimo, gabriella.olmo}@polito.it
[2] Consiglio Nazionale delle Ricerche, Torino, Italy claudia.ferraris@cnr.it

A Qualitative Visualizations

Figure A.1 shows qualitative reconstructions on the same sampled frames across all four backbones, for each of the three datasets. On in-domain Re:InterHand and on GigaHands, predictions (red) closely track the ground truth (blue), while the more challenging egocentric AssemblyHands frames show larger deviations. Per-sample MPJPE and PA-MPJPE (mm) are reported below each pose.

B Lifting Masked Motion Modeling (L-MMM)

To enhance the model's robustness against occlusions (a pervasive challenge in egocentric hand analysis), we employ **Lifting Masked Motion Modeling (L-MMM)** [5, 10]. During training, we apply a binary mask $M \in \{0,1\}^{T \times J}$ to the 2D input sequence, generating a corrupted input $\hat{X}_{2D} = X_{2D} \odot M$. The model is tasked with reconstructing the full 3D structure from this partial observation.

The objective function minimizes the weighted reconstruction error, prioritizing the recovery of masked regions to enforce structural learning:

$$\mathcal{L} = \frac{1}{T \cdot J_H} \sum_{t=1}^{T} \sum_{j=1}^{J_H} \left(w_j \cdot m_{t,j} \cdot \| \hat{y}_{t,j} - y_{t,j}^{GT} \|_2 \right) + \lambda \mathcal{L}_{vel} \tag{B.1}$$

where $w_j \in [1.0, 2.0]$ denotes joint-specific weights emphasizing intermediate joints and fingertips [11], and $m_{t,j}$ is a binary upweighting factor:

$$m_{t,j} = \begin{cases} 1/p & \text{if joint } j \text{ at frame } t \text{ is masked} \\ 1.0 & \text{otherwise} \end{cases} \tag{B.2}$$

with $p = 0.25$ as the spatial masking ratio. This inverse weighting ($\approx 4\times$) incentivizes the model to infer missing geometry from global context. The term

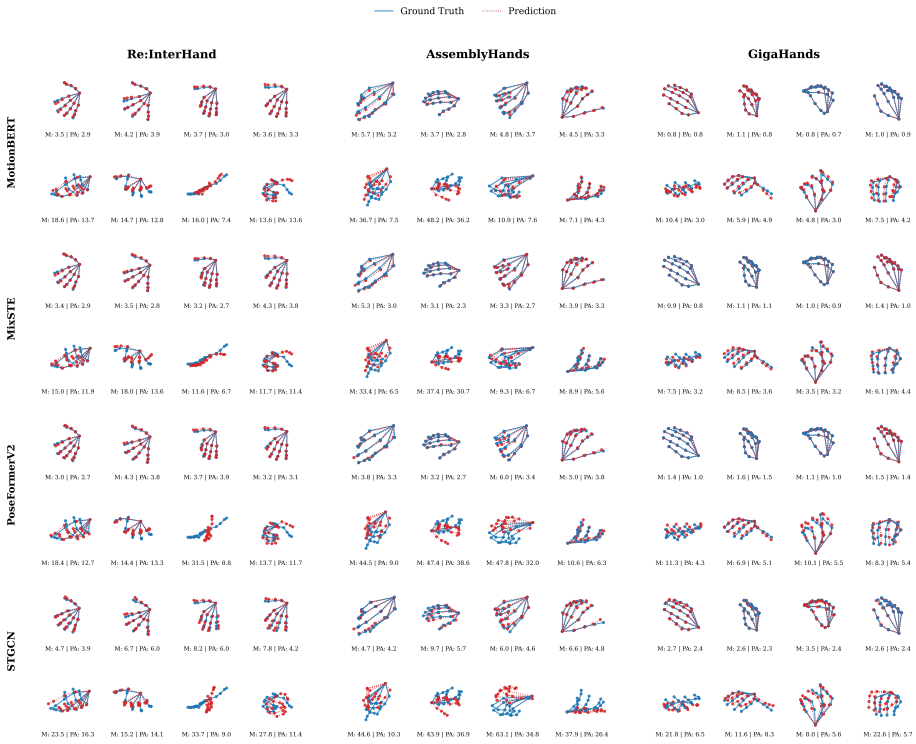

Fig. A.1: Qualitative Results. TransHands predictions (red) vs. Ground Truth (blue). We visualize the reconstruction quality across the four selected backbones ([6–8, 10]) for each scenario: **S2-A** (Refinement) for the in-domain Re:InterHand [3], and **S2-B** (Multi-Dataset) for the cross-domain AssemblyHands [4] and GigaHands [1]. For each backbone, the top row shows representative reconstructions and the bottom row more challenging cases, including failures on occluded frames. The values below each sample denote the MPJPE and PA-MPJPE (mm).

$\mathcal{L}_{vel}$ imposes a temporal consistency regularization on the predicted joint velocities [9]:

$$\mathcal{L}_{vel} = \frac{1}{(T-1)J_H} \sum_{t=2}^{T} \sum_{j=1}^{J_H} \Big\| (\hat{y}_{t,j} - \hat{y}_{t-1,j}) - (y^{GT}_{t,j} - y^{GT}_{t-1,j}) \Big\|_2 \tag{B.3}$$

with $\lambda = 0.5$, ensuring temporally consistent trajectory reconstruction.

Table B.1: Ablation Study on L-MMM Objective (PoseFormerV2 [8]). Impact of L-MMM on source domain (Re:InterHand [3]) and Zero-Shot (AssemblyHands [4]).

Protocol	L-MMM	Best MPJPE ↓		Best PA-MPJPE ↓	
		Re:Inter (In-Domain)	Assembly (Zero-Shot)	Re:Inter (In-Domain)	Assembly (Zero-Shot)
Supervised Only		**11.93**	46.97	**7.42**	23.61
Ours	✓	12.62	**39.70**	8.01	**21.93**

B.1 Ablation Study: Self-Supervised Objectives

We analyze the role of the L-MMM objective. Its inclusion is motivated by previous works in self-supervised pose estimation, which demonstrated that masked reconstruction tasks can improve robustness to occlusions [5, 10].

Results. Comparing the models trained with and without L-MMM (Table B.1), we observe a compelling trade-off. At first, the purely supervised baseline appears superior on the source domain (Re:InterHand), achieving an MPJPE of 11.93 mm compared to our 12.62 mm. However, this marginal in-domain advantage comes at a cost to generalization. When evaluated in a zero-shot setting on the heavily occluded AssemblyHands dataset, the supervised-only model degrades significantly, exhibiting a **15.5%** relative increase in error compared to the L-MMM variant. The strong body motion priors inherited from pre-training already encode robust spatial-temporal patterns, while L-MMM refines the handling of uncertainties [2].

Practical Implication. For practitioners prioritizing simplicity, the supervised-only baseline offers comparable performance with reduced training complexity. We retain L-MMM in our main experiments to ensure comprehensive evaluation, but acknowledge it as an *optional component* rather than a core requirement.

C Ablation Study: Decoupling Analysis

Our training pipeline adopts a progressive strategy: we first align the projection head while keeping the backbone frozen (S1), and only subsequently unfreeze the encoder for refinement (S2-A). We investigate whether this two-step decoupling is necessary compared to a direct fine-tuning approach.

Setup. We compare two protocols using the same pre-trained PoseFormerV2 backbone:

1. **Ours (Progressive):** The standard S1 (Frozen) to S2-A (Partially Unfrozen) pipeline. This corresponds to the *Linear Probing then Fine-Tuning* (LP-FT) paradigm [2].
2. **Baseline (Direct Refinement):** We load the pre-trained weights and **immediately enable gradients** for the last backbone block. This mimics a standard fine-tuning setup without a warm-up alignment phase.

Table C.1: Decoupling Analysis (PoseFormerV2 [8]). Comparison between Direct Fine-Tuning and our Progressive (Frozen → Partial) strategy on Re:InterHand [3].

Protocol	S1 Align.	MPJPE ↓	PA-MPJPE ↓
Direct Refinement		10.02	7.12
Ours (Progressive)	✓	**8.01**	**5.88**

Results. Table C.1 presents the comparison. Without the frozen alignment phase, the direct refinement baseline achieves an error of 10.02 mm, but falls short of our progressive strategy, which improves the result to **8.01 mm**.

We attribute this to *feature distortion*: without a frozen alignment phase, noisy gradients from the randomly initialized head and adapter degrade the pretrained priors before the projection layers converge. Figure C.1 confirms this: S1 absorbs the large-magnitude updates in the Adapter and Head, so the spatio-temporal layers need only minimal adjustment during S2.

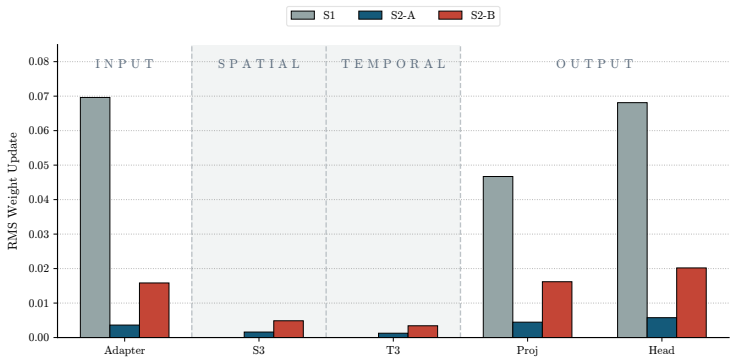

Fig. C.1: Layer-wise Parameter Evolution (PoseFormerV2 [8]). We visualize the RMS weight update across our three-stage training protocol. In **S1**, only the Topological Adapter, Projection, and Head modules are trained. In **S2**, we selectively unfreeze the final spatio-temporal blocks.

References

1. Fu, R., Zhang, D., Jiang, A., Fu, W., Funk, A., Ritchie, D., Sridhar, S.: Gigahands: A massive annotated dataset of bimanual hand activities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
2. Kumar, A., Raghunathan, A., Jones, R., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. In: International Conference on Learning Representations (ICLR) (2022)

3. Moon, G., Saito, S., Xu, W., Joshi, R., Buffalini, J., Bellan, H., Rosen, N., Richardson, J., Mize, M., de Bree, P., Simon, T., Peng, B., Garg, S., McPhail, K., Shiratori, T.: A dataset of relighted 3d interacting hands. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track (2023)
4. 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). https://doi.org/10.1109/CVPR52729.2023.01249
5. Shan, W., Liu, Z., Zhang, X., Wang, S., Ma, S., Gao, W.: P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In: Computer Vision – ECCV 2022. pp. 461–478. Lecture Notes in Computer Science, Springer (2022). https://doi.org/10.1007/978-3-031-20065-6_27
6. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018). https://doi.org/10.1609/aaai.v32i1.12328
7. Zhang, J., Tu, Z., Yang, J., Chen, Y., Yuan, J.: Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13222–13232 (2022). https://doi.org/10.1109/CVPR52688.2022.01288
8. Zhao, Q., Zheng, C., Liu, M., Wang, P., Chen, C.: Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8877–8886 (2023). https://doi.org/10.1109/CVPR52729.2023.00857
9. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11656–11665 (2021). https://doi.org/10.1109/ICCV48922.2021.01145
10. Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: Motionbert: A unified perspective on learning human motion representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15039–15053 (2023). https://doi.org/10.1109/ICCV51070.2023.01385
11. Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 813–822 (2019). https://doi.org/10.1109/ICCV.2019.00090