Dynamic4D: Enhancing Self-Supervised Learning for Robust and Fine-Grained 4D Point Cloud Video Understanding

Mingxuan Du * and Yutian Zeng

Posted Date: 17 March 2026

doi: 10.20944/preprints202603.1381.v1

Keywords: 4D point clouds; self-supervised learning; dynamic; robustness; motion prediction

 doi:10.20944/preprints202603.1381.v1

Article

Dynamic4D: Enhancing Self-Supervised Learning for Robust and Fine-Grained 4D Point Cloud Video Understanding

Mingxuan Du * and Yutian Zeng

Zhongnan University of Economics and Law
* Correspondence: 202429465812@stu.zuel.edu.cn

Abstract

The proliferation of 4D point cloud videos highlights their potential, but the high cost of obtaining large-scale annotated data severely limits supervised methods. Consequently, self-supervised learning (SSL) is vital for learning generalizable representations from unlabeled 4D data. While existing SSL frameworks, such as Uni4D, have made progress, they often struggle with fine-grained motion understanding in extremely dynamic scenes, maintaining robustness under severe occlusion, and developing explicit predictive capabilities. To address these, we propose Dynamic4D, a novel and robust self-supervised framework tailored for dynamic 4D point cloud understanding. Dynamic4D introduces an Adaptive Causal Temporal Attention (ACTA) mechanism in the encoder for explicit causal temporal modeling and dynamic region-focused learning. Its decoder employs Motion Prediction Tokens (MPT) to directly infer motion vectors for masked regions. A novel adaptive motion-sensitive masking strategy further enhances robustness by intelligently prioritizing high-dynamic zones. Our multi-objective pre-training strategy integrates a new Dynamic Perception Loss alongside geometric reconstruction and latent-space alignment. Extensive experiments on diverse challenging benchmarks demonstrate that Dynamic4D consistently achieves state-of-the-art performance. It substantially outperforms prior methods, validating its superior capacity to learn highly robust, generalizable, and motion-aware representations for complex dynamic 4D point cloud scenes.

Keywords: 4D point clouds; self-supervised learning; dynamic; robustness; motion prediction

1. Introduction

The advent of advanced 3D sensing technologies, such as LiDAR and depth cameras, has led to a proliferation of 4D point cloud videos—sequences of 3D point clouds over time. These rich temporal-spatial data streams hold immense potential across various critical domains, including autonomous driving, robotics, human-computer interaction, and virtual reality [1]. Effectively understanding and analyzing these dynamic 4D data is paramount for achieving intelligent perception in complex environments. However, the acquisition of large-scale, high-quality annotated 4D point cloud datasets is prohibitively expensive and time-consuming, severely limiting the applicability of traditional supervised learning paradigms. Consequently, the development of **self-supervised learning (SSL)** frameworks, which leverage vast amounts of unlabeled 4D point cloud video data for pre-training to learn generalizable and transferable representations, has emerged as a prominent and active research area [2].

Previous works, such as the Uni4D framework [3], have successfully explored self-supervised representation learning for 4D point clouds by employing a **self-disentangled masked autoencoder (MAE)**. Uni4D adeptly captures both low-level geometric structures and high-level semantic/dynamic information through innovative geometric reconstruction and latent space alignment mechanisms. This approach effectively bridges the gap between low-level geometry and high-level semantics in

 doi:10.20944/preprints202603.1381.v1

4D data processing, a common challenge for conventional MAE-based methods, and significantly improves the modeling of motion.

Despite these advancements, current methodologies still face considerable challenges, particularly in handling **fine-grained motion understanding in extremely dynamic scenes** and maintaining **robustness under severe occlusion or sparsity conditions**. Specifically, limitations persist in:

1. **Complex Interaction and Subtle Action Recognition:** Existing methods often struggle with modeling subtle and rapid interactions between objects or between humans and their environments, as well as capturing long-range temporal dependencies effectively. Such challenges are increasingly being addressed by advanced models that bridge vision, language, and action to enable a more holistic understanding of dynamic scenarios [4].
2. **Dynamic Occlusion and Sparse Point Clouds:** In real-world scenarios, dynamic objects are frequently occluded or observed through sparse point clouds due to sensor limitations. This adversely impacts the model's ability to perceive complete motion trajectories and extract robust features. Novel methods for point cloud completion, utilizing advanced distance metrics, contribute to mitigating the effects of sparsity [5].
3. **Predictive Modeling:** The capability to predict future frames or subsequent motions is a crucial indicator of deep dynamic scene understanding. However, current self-supervised frameworks have largely underemphasized direct modeling for such predictive tasks. The development of sophisticated spatial-temporal models is vital for improving prediction capabilities in complex robotic manipulation tasks [6].

Motivated by these challenges and building upon the successes of existing self-supervised 4D point cloud representation learning, this research aims to propose a more **robust and dynamic-aware** unified self-supervised learning framework. Our goal is to better capture fine-grained motion and long-range temporal dependencies in complex dynamic scenarios, thereby providing high-quality, transferable representations for a broader range of challenging 4D downstream tasks.

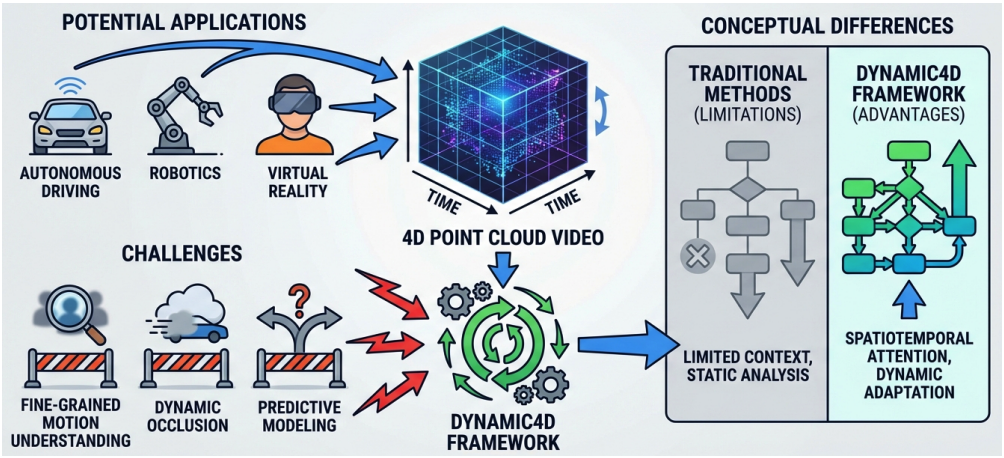

Figure 1. An overview of the motivations and challenges in 4D point cloud video research, and the conceptual advantages of the proposed Dynamic4D framework. It illustrates the wide range of potential applications, outlines the primary difficulties faced by existing methods (fine-grained motion understanding, dynamic occlusion, and predictive modeling), and highlights how Dynamic4D addresses these limitations through spatio-temporal attention and dynamic adaptation.

We introduce **Dynamic4D: A Robust Self-Supervised Framework for Dynamic 4D Point Cloud Understanding**. Dynamic4D extends the successful paradigm of Uni4D by incorporating two key innovations: an **Adaptive Causal Temporal Attention (ACTA)** mechanism within the encoder to explicitly learn future prediction from past information and focus on high-dynamic regions, and **Motion Prediction Tokens (MPT)** in the decoder to directly infer motion vectors or point displacements in masked areas. These innovations are coupled with an **adaptive masking strategy** that prioritizes masking high-dynamic regions, forcing the model to recover motion from more challenging, occlusion-

 doi:10.20944/preprints202603.1381.v1

like scenarios. Our self-supervised pre-training objectives are thus augmented with a new **Dynamic Perception Loss**, which includes a motion prediction loss alongside geometric reconstruction and latent-space alignment losses.

To validate the efficacy and versatility of Dynamic4D, we conduct extensive experiments on several challenging 4D point cloud video benchmarks, including MSR-Action3D [7], NTU-RGBD [8], HOI4D [8], NvGesture [8], and SHREC'17 [8]. Our evaluation strictly follows the self-supervised pre-training and downstream task fine-tuning/few-shot/semi-supervised/transfer learning paradigm. Our results demonstrate that Dynamic4D consistently *outperforms state-of-the-art methods*, including Uni4D, across various tasks. For instance, Dynamic4D achieves an improvement of approximately **0.74%** in action recognition on MSR-Action3D and a significant **1.3%** increase in F1@50 for action segmentation on HOI4D. Furthermore, in semi-supervised learning scenarios with only 10% labeled data on NTU-RGBD, Dynamic4D yields a substantial **1.7%** boost over Uni4D. For few-shot learning, Dynamic4D exhibits superior generalization, improving by **1.6%** in the 10-way 1-shot setting on MSR-Action3D. These compelling results highlight the robust and information-rich representations learned by Dynamic4D, making it highly effective for complex dynamic 4D point cloud understanding.

Our main contributions are summarized as follows:

- We propose Dynamic4D, a novel self-supervised framework that significantly enhances the understanding of fine-grained motion and long-range temporal dependencies in 4D point cloud videos.
- We introduce two architectural innovations: the Adaptive Causal Temporal Attention (ACTA) mechanism for improved temporal modeling and Motion Prediction Tokens (MPT) coupled with a Dynamic Perception Loss, enabling direct motion inference in occluded regions.
- Through extensive experiments, Dynamic4D consistently achieves state-of-the-art performance across various 4D point cloud downstream tasks, demonstrating superior robustness, generalization, and data efficiency compared to existing methods.

2. Related Work

2.1. Self-Supervised Learning for 3D and 4D Point Clouds

Self-supervised learning (SSL) is crucial for mitigating data annotation bottlenecks in 3D and 4D point clouds, drawing inspiration from 2D vision and natural language processing (NLP). Broader AI challenges include privacy-preserving methods [9] and energy efficiency [10]. The core principle of extracting meaningful representations from unlabeled data is highly transferable, particularly given NLP's advancements in spatial, geometric, and temporal feature extraction. General SSL architectures, like CONSERT [11] for robust sentence representation, directly parallel deriving robust geometric features from 3D point clouds. The versatility of models like Graph Neural Networks is also evident across applications such as fraudulent traffic governance [12]. Other promising avenues for bootstrapping performance and generalization in 3D/4D domains include self-training with pseudo-labels [13], unsupervised pattern extraction exemplified by CauSeRL [14], and unsupervised pre-training for few-shot adaptation like MetaICL [15].

Contrastive learning [16] is a dominant SSL technique pertinent to 3D/4D point clouds, using augmentations or adjacent points as positive pairs to learn spatio-temporal dynamics. Addressing noise, data scarcity, and dynamic changes has led to robust learning mechanisms. Examples include learning efficient representations from multi-grained state-space models for noisy data [17], prompt-learning for dynamic information optimization [18], and noise-robust learning combining self-training with language model augmentation [19]. Robust spatio-temporal feature extraction against distribution shifts [20] and handling incomplete point clouds using hyperbolic Chamfer distance [5] are also highly relevant. In essence, these NLP-rooted self-supervised and representation learning methodologies—including contrastive learning, self-training, unsupervised pre-training, and robust learning paradigms—provide critical foundational frameworks for advancing SSL in 3D and 4D point clouds.

 doi:10.20944/preprints202603.1381.v1

2.2. Dynamic Modeling and Motion Prediction in Point Clouds

Dynamic modeling and motion prediction in point clouds are fundamental for applications like autonomous driving and robotics, requiring an understanding of 3D scene evolution, object identification, and trajectory forecasting. This need for robust dynamic understanding and parameter estimation extends to diverse engineering systems, such as online identification in permanent magnet synchronous machines [21–23]. Valuable foundational principles for handling dynamic information, temporal reasoning, and predictive modeling can be gleaned from diverse fields, including natural language processing (NLP) and multimodal machine learning.

Representing changing states is crucial; implicit dynamic representations of meaning in neural language models [24] and incorporating dynamic semantic changes [25] offer direct parallels to capturing evolving geometry in sequential 3D point cloud data. Temporal reasoning for future event forecasting mirrors methods like CluSTeR [26] for historical clues on knowledge graphs, and insights into causal modeling in transformers for positional information [27] inform motion causality in 3D scenes. Effective spatio-temporal data processing often relies on sophisticated attention mechanisms, such as MTAG [28], a graph-based model utilizing spatio-temporal attention for unaligned multimodal sequential data, which is highly relevant for evolving point clouds.

Recent advancements include spatial-temporal graph diffusion policies with kinematic modeling [6] for precise robotic manipulation motion prediction, and unified vision-language-action models [4] for comprehensive dynamic scene interpretation. Understanding individual objects and their potential actions is foundational: large vision-language models for object detection and localization [29] support contextual understanding, and action recognition from observed data [30] is pertinent to predicting complex behaviors in 3D sequences. While some cited works, like span prediction for Named Entity Recognition [31], originate from distinct domains, these works collectively offer valuable insights into general principles of dynamic representation, temporal reasoning, causal inference, and spatio-temporal attention, which are highly transferable and foundational for advancing dynamic modeling and motion prediction in 3D point clouds.

3. Method

In this section, we present **Dynamic4D: A Robust Self-Supervised Framework for Dynamic 4D Point Cloud Understanding**. Building upon the foundational successes of prior works, Dynamic4D introduces novel architectural components and a refined self-supervised pre-training strategy to address the challenges of fine-grained motion understanding in complex dynamic scenes and robustness under severe occlusion or sparsity. Our framework is designed to learn rich, transferable representations by explicitly modeling temporal dependencies and predicting motion in masked regions. This section details each core component, from initial feature extraction to our multi-objective self-supervised pre-training strategy.

3.1. Overall Architecture

Dynamic4D adopts an encoder-decoder architecture, augmented with a momentum encoder, specifically tailored for processing 4D point cloud videos. The processing pipeline begins with **Initial Feature Extraction**, which transforms raw 4D point cloud sequences into compact, spatio-temporal representations. These representations are subsequently fed into an **Online Encoder**, which learns contextual features from the visible point segments. Simultaneously, a **Momentum Encoder**, a slowly evolving exponential moving average (EMA) copy of the online encoder, provides stable and consistent target features for the self-supervised contrastive and alignment learning objectives. The **Decoder** is then tasked with several crucial functions: reconstructing masked geometric information, aligning latent features between the online and momentum branches, and explicitly predicting motion vectors for occluded regions using specialized tokens. This multi-faceted design ensures comprehensive learning across various levels of spatio-temporal detail and dynamic phenomena.

 doi:10.20944/preprints202603.1381.v1

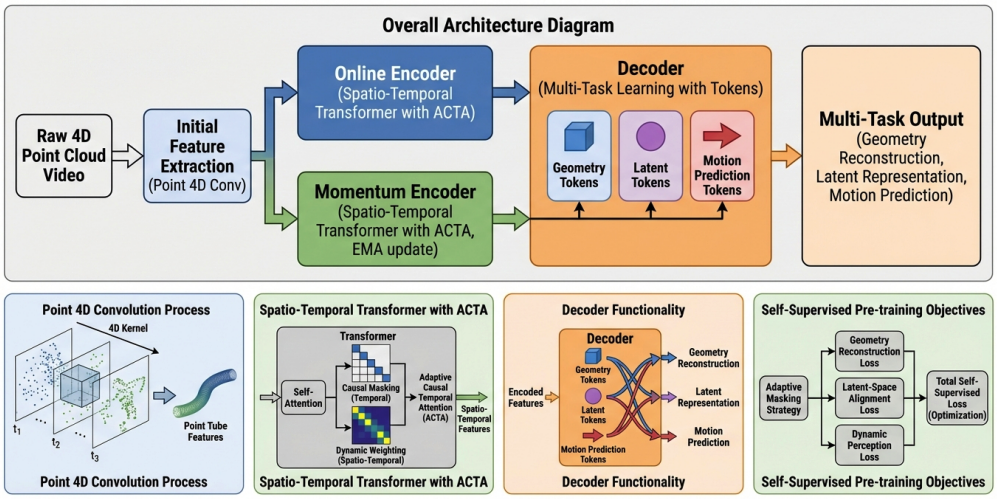

Figure 2. Overall architecture and key components of Dynamic4D. The top panel provides a high-level overview of the encoder-decoder framework, illustrating the flow from raw 4D point cloud videos through initial feature extraction, the online and momentum encoders, and the multi-task decoder with specialized tokens leading to multi-task outputs. The bottom panels detail the specifics of the Point 4D Convolution process, the Spatio-Temporal Transformer with Adaptive Causal Temporal Attention (ACTA), the functionality of the decoder's Geometry, Latent, and Motion Prediction Tokens, and the comprehensive self-supervised pre-training objectives.

3.2. Initial Feature Extraction: Point 4D Convolution

The raw 4D point cloud video, represented as a sequence of 3D point clouds $\{P_t\}_{t=1}^{T}$ (where $P_t = \{p_{i,t}\}_{i=1}^{N_t}$ contains N_t points at time t), is first processed to extract initial spatio-temporal features. Following common practices in 4D point cloud processing, we utilize **Point 4D convolution (P4D)** to embed each point and its local spatio-temporal neighborhood. This process aggregates features from nearby points across consecutive frames, generating a sequence of 'point tube' features. A 'point tube' for a point $p_{i,t}$ at time t can be formally derived as:

$$x_{i,t} = \mathcal{G}_{\text{P4D}}(p_{i,t}, \mathcal{N}(p_{i,t}, \delta_s, \delta_t)) \tag{1}$$

where $x_{i,t}$ is the extracted feature vector for point $p_{i,t}$, $\mathcal{G}_{\text{P4D}}$ denotes the Point 4D convolution operator, and $\mathcal{N}(p_{i,t}, \delta_s, \delta_t)$ represents the spatio-temporal neighborhood of $p_{i,t}$ within a spatial radius δ_s and a temporal window δ_t. These 'point tube' features, $X = \{x_{i,t}\}$, serve as the input to our subsequent encoding layers, capturing local geometric and rudimentary temporal cues essential for dynamic understanding.

3.3. Spatio-Temporal Encoder with Adaptive Causal Temporal Attention

The core of our encoding mechanism is an improved **Spatio-Temporal Transformer** that serves as the backbone for both the Online and Momentum Encoders. This transformer processes the 'point tube' embeddings, X, to learn comprehensive and context-aware 4D representations. It is composed of multiple layers, each incorporating self-attention and feed-forward networks. A key innovation within our encoder is the introduction of the **Adaptive Causal Temporal Attention (ACTA)** mechanism, which enhances the model's ability to learn predictive dynamics.

Standard self-attention mechanisms aggregate information from all positions. However, for dynamic understanding and future prediction tasks, enforcing a causal structure is inherently beneficial. ACTA integrates two main features to achieve this:

1. **Causal Masking:** A causal mask M_c is explicitly applied within the temporal attention module. This ensures that the representation of a point at time step t can only attend to information from previous time steps $1, \ldots, t$, thereby preventing information leakage from future data. This mechanism explicitly encourages the model to learn predictive capabilities from historical data, which is vital for anticipating future motion.

 doi:10.20944/preprints202603.1381.v1

2. **Dynamic Weighting:** A dynamic weight matrix W_D is integrated into the attention calculation. This mechanism adaptively adjusts temporal attention weights based on the local motion intensity of point clouds. Points residing within high-dynamic regions (e.g., fast-moving objects or complex interactions) receive higher attention weights, allowing the model to focus computational resources on the most informative and challenging areas of motion. Specifically, W_D can be learned via a small sub-network that takes local motion features (e.g., derived from sparse optical flow or point displacements) as input, mapping them to attention biases.

Formally, for an input sequence of point features $X \in \mathbb{R}^{N \times D_{feat}}$ (where N is the total number of points across all time steps and D_{feat} is the feature dimension), the attention mechanism is computed as:

$$Q = XW_Q \tag{2}$$

$$K = XW_K \tag{3}$$

$$V = XW_V \tag{4}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top + M_c + W_D}{\sqrt{d_k}}\right)V \tag{5}$$

where W_Q, W_K, W_V are learnable weight matrices transforming the input features into query, key, and value representations, respectively. d_k is the dimension of the key vectors, M_c is the causal mask (setting attention scores to $-\infty$ for future time steps), and W_D is the dynamically computed weight matrix based on local motion features. This spatio-temporal transformer, integrated with ACTA, forms the robust backbone of our encoders, enabling them to capture both static spatial contexts and dynamic temporal evolutions.

3.4. Lightweight Decoder with Motion Prediction Tokens

The decoder in Dynamic4D is a lightweight Transformer designed to perform multi-task learning by leveraging distinct token types. It takes the encoded features of visible point segments from the online encoder and a set of learnable tokens as input. Similar to prior works, our decoder utilizes 'geometry tokens' for reconstructing masked 3D coordinates and 'latent tokens' for aligning high-level features between the online and momentum encoders.

A critical new component we introduce are **Motion Prediction Tokens (MPT)**. These are a specialized set of learnable tokens specifically designed to infer dynamic information within masked regions. When queried by the decoder, MPTs are trained to predict local motion vectors (e.g., 3D displacement vectors or optical flow) for the originally masked-out point cloud segments. This direct prediction task forces the encoder to learn highly discriminative and motion-aware features, significantly enhancing the model's understanding of complex dynamics.

The decoder leverages these distinct token types to achieve a holistic understanding: reconstructing low-level geometric details via 'geometry tokens', ensuring consistency in high-level semantic and motion representations via 'latent tokens', and making explicit predictions about dynamic changes through the **Motion Prediction Tokens**.

3.5. Self-Supervised Pre-training Objectives

Dynamic4D's pre-training strategy is a sophisticated multi-objective learning paradigm, extending beyond traditional Masked Autoencoder (MAE) approaches. Our objectives combine geometric fidelity, latent space consistency, and explicit motion prediction to learn robust and transferable 4D point cloud representations.

3.5.1. Adaptive Masking Strategy

Prior to feeding into the encoder, a subset of 'point tubes' is strategically masked. Unlike purely random masking approaches, Dynamic4D employs an **adaptive motion-sensitive masking strategy**. In the pre-processing phase, we utilize simple frame-to-frame point cloud differencing or sparse optical

 doi:10.20944/preprints202603.1381.v1

flow algorithms to identify regions exhibiting high dynamic activity. Subsequently, these high-dynamic regions or their immediate surroundings are masked with a higher probability. This strategy simulates realistic dynamic occlusion scenarios, forcing the model to infer motion and complete information from limited and challenging contexts, thereby fostering more robust dynamic feature learning by focusing attention on regions crucial for understanding change.

3.5.2. Geometry Reconstruction Loss

The primary objective for the 'geometry tokens' is to accurately reconstruct the 3D coordinates of the masked 'point tubes'. We employ the **Chamfer Distance (CD)** as our geometry reconstruction loss $\mathcal{L}_{geo}$, which measures the similarity between the ground-truth masked point cloud P_{masked} and its reconstructed counterpart $\hat{P}_{masked}$:

$$\mathcal{L}_{geo} = \sum_{x \in P_{masked}} \min_{y \in \hat{P}_{masked}} \|x - y\|_2^2 + \sum_{y \in \hat{P}_{masked}} \min_{x \in P_{masked}} \|x - y\|_2^2 \tag{6}$$

This loss ensures that the encoder-decoder pathway maintains fidelity to low-level spatial and structural details, providing a strong foundation for geometric understanding.

3.5.3. Latent-Space Alignment Loss

To ensure the learning of semantically meaningful and temporally coherent high-level representations, we utilize a **Latent-Space Alignment Loss** $\mathcal{L}_{align}$. This loss encourages consistency between the representations learned by the online encoder and the momentum encoder, promoting robust feature learning invariant to minor input variations and enhancing temporal stability. It comprises two main components:

1. **Frame-level Motion Alignment:** This component aligns the features of corresponding 'point tubes' or pooled frame-level features across adjacent frames between the online and momentum encoders, ensuring local temporal coherence and consistency in short-range dynamics.
2. **Video-level Global Alignment:** This aligns the global video representations (e.g., obtained by global pooling across all frames and points) produced by both encoders, fostering long-range consistency and capturing overall scene dynamics, thereby stabilizing the learning of macro-level temporal patterns.

Let F_{online} denote the high-level features extracted by the online encoder and $F_{momentum}$ denote the corresponding features from the momentum encoder. The latent tokens facilitate mapping these into a comparable latent space. The latent-space alignment loss is generally formulated as:

$$\mathcal{L}_{align} = \mathcal{H}(F_{online}, F_{momentum}) \tag{7}$$

where $\mathcal{H}$ represents a similarity-based loss function (e.g., contrastive loss or mean squared error on normalized features) that encourages the alignment of features at both frame and video levels, using outputs from the 'latent tokens' to bridge the representations.

3.5.4. Dynamic Perception Loss

A novel contribution of Dynamic4D is the **Dynamic Perception Loss** $\mathcal{L}_{motion}$, which directly supervises the predictions made by the 'Motion Prediction Tokens (MPT)'. Specifically, the MPTs are trained to predict the local motion vectors (e.g., 3D displacement vectors or optical flow) for the masked regions. This can be formulated as an L1 regression loss:

$$\mathcal{L}_{motion} = \frac{1}{|P_{masked}|} \sum_{p_i \in P_{masked}} \|\vec{v}_i - \hat{\vec{v}}_i\|_1 \tag{8}$$

where $\vec{v}_i$ is the ground-truth motion vector for point p_i in the masked region, and $\hat{\vec{v}}_i$ is the predicted motion vector by the MPTs. This loss explicitly trains the model to understand and predict motion,

 doi:10.20944/preprints202603.1381.v1

which is crucial for handling dynamic occlusions, anticipating future states, and robustly interpreting complex scene dynamics.

3.5.5. Total Pre-training Loss

The total self-supervised pre-training loss $\mathcal{L}_{total}$ for Dynamic4D is a weighted sum of the aforementioned objectives:

$$\mathcal{L}_{total} = \lambda_{geo}\mathcal{L}_{geo} + \lambda_{align}\mathcal{L}_{align} + \lambda_{motion}\mathcal{L}_{motion} \tag{9}$$

where λ_{geo}, λ_{align}, and λ_{motion} are empirically determined weighting hyperparameters that balance the contributions of geometric reconstruction, latent-space alignment, and dynamic perception. By jointly optimizing these objectives, Dynamic4D learns highly robust, motion-aware, and transferable 4D point cloud representations capable of fine-grained understanding of dynamic scenes.

4. Experiments

In this section, we present a comprehensive evaluation of Dynamic4D, detailing our experimental setup, comparing its performance against state-of-the-art methods across various 4D point cloud tasks, and conducting ablation studies to validate the effectiveness of our proposed innovations.

4.1. Experimental Setup

Our experimental methodology strictly adheres to the self-supervised pre-training and subsequent fine-tuning/transfer learning paradigm.

4.1.1. Datasets

To ensure a robust evaluation of Dynamic4D's generalizability and effectiveness, we conduct extensive experiments on several widely recognized and challenging 4D point cloud video datasets:

- **MSR-Action3D**: A dataset for human action recognition, containing sequences of skeletal data and depth maps, which are converted to 4D point clouds for our experiments.
- **NTU-RGBD**: A large-scale dataset for 3D action recognition, featuring diverse human actions performed by multiple subjects and camera views. We utilize the depth streams to generate 4D point cloud sequences.
- **HOI4D**: A dataset focused on human-object interaction action segmentation, providing rich and complex scenarios for fine-grained dynamic analysis.
- **NvGesture**: A dataset for hand gesture recognition in driving scenarios, offering challenges related to small, rapid movements and potential occlusions.
- **SHREC'17**: Another prominent dataset for 3D hand gesture recognition, further testing the model's ability to discern subtle hand movements.

These datasets cover a broad spectrum of dynamic 4D point cloud tasks, including action recognition, gesture recognition, and action segmentation, allowing for a thorough assessment of Dynamic4D's capabilities.

4.1.2. Pre-Training Details

Our pre-training phase is critical for learning rich, transferable 4D representations.

- **Data Preprocessing:** Raw 4D point cloud data is first preprocessed into 'point tubes'. This involves performing Farthest Point Sampling (FPS) on each frame (e.g., 1024 points per frame for MSR-Action3D and NTU-RGBD; 2048 points per frame for HOI4D), followed by constructing 'point tube' sequences that encapsulate spatio-temporal neighborhood information. A key aspect is our **adaptive motion-sensitive masking strategy**: instead of purely random masking, we leverage frame-to-frame motion estimation (e.g., local point displacement or intensity changes) to identify high-dynamic regions. These dynamic areas, or their immediate vicinities, are then

 doi:10.20944/preprints202603.1381.v1

masked with a higher probability (typically 75% masking ratio), simulating real-world dynamic occlusions and compelling the model to infer motion from challenging contexts.

- **Model Training:** The Online Encoder processes the visible 'point tube embeddings', while the Momentum Encoder, an Exponential Moving Average (EMA) copy of the online encoder, provides stable target representations for consistency learning. The Decoder, a lightweight Transformer, performs multi-task reconstruction and prediction using distinct token types: 'geometry tokens' for geometric reconstruction, 'latent tokens' for latent-space alignment, and the novel 'Motion Prediction Tokens (MPT)' for explicit motion vector prediction.
- **Loss Functions:** The total pre-training loss is a weighted sum of three objectives:
  1. **Geometry Reconstruction Loss ($\mathcal{L}_{geo}$):** Uses Chamfer Distance to reconstruct masked point cloud geometries.
  2. **Latent-Space Alignment Loss ($\mathcal{L}_{align}$):** Enforces consistency between online and momentum encoder features at both frame and video levels.
  3. **Dynamic Perception Loss ($\mathcal{L}_{motion}$):** An L1 loss on the predicted motion vectors for masked regions, specifically supervising the output of MPTs.

  The weighting hyperparameters for these losses (λ_{geo}, λ_{align}, λ_{motion}) are determined empirically.
- **Hyperparameters:** Key configurations include point sampling rates (1024/2048 points per frame), sequence lengths (24 frames for MSR-Action3D/NTU-RGBD, 150 frames for HOI4D), and pre-training epochs (200 for MSR-Action3D, 100 for NTU-RGBD, 50 for HOI4D). The encoder backbone varies based on dataset scale and complexity, flexibly using P4Transformer (5-10 layers) or PPTr, with all backbones integrating our **Adaptive Causal Temporal Attention (ACTA)** module. The decoder is a 4-layer Transformer.

4.1.3. Fine-Tuning and Evaluation

After pre-training, the learned encoder representations are evaluated on various downstream tasks:

- **End-to-end Fine-tuning:** The pre-trained encoder is detached from the decoder, and a task-specific classifier or regression head is added on top. The entire network is then fine-tuned on the target dataset.
- **Semi-supervised Learning:** We fine-tune the pre-trained model with only a limited percentage of labeled data (e.g., 10%, 50%) to assess its data efficiency and generalization capability under scarce supervision.
- **Few-shot Learning:** Experiments are conducted to evaluate the model's ability to generalize from a very small number of examples per class, often involving cross-dataset transfer (e.g., pre-training on NTU-RGBD and fine-tuning on NvGesture/SHREC'17).

4.2. Comparison with State-of-the-Art

We benchmark Dynamic4D against several advanced 4D point cloud understanding methods, including training from scratch, MaST-Pre, VideoMAE-based approaches, and the prominent Uni4D framework.

4.2.1. Action Recognition and Gesture Recognition

Table 1 presents the performance of Dynamic4D and baseline methods on action and gesture recognition tasks. Dynamic4D consistently achieves state-of-the-art results across all evaluated datasets. On MSR-Action3D, Dynamic4D improves action recognition accuracy by **0.74%** compared to Uni4D, reaching **94.12%**. Similar gains are observed on NTU-RGBD, NvGesture, and SHREC'17 datasets, demonstrating its enhanced capability in modeling complex dynamics and fine-grained motion pertinent to various human activities and interactions.

 doi:10.20944/preprints202603.1381.v1

Table 1. Performance comparison on action recognition and gesture recognition benchmarks. All numbers are percentage accuracy (%). Higher is better. Our method consistently outperforms state-of-the-art approaches.

Method (baseline / variant)	MSR-Action3D Acc.	NTU-RGBD Acc.	NvGesture Acc.	SHREC'17 Acc.
P4Transformer (scratch)	90.94	90.2	87.7	91.2
P4Transformer + MaST-Pre	91.29	90.8	89.3	92.4
P4Transformer + Uni4D	**93.38**	**90.7**	**89.6**	**93.8**
P4Transformer + Dynamic4D (Ours)	**94.12** $^{+0.74}$	**91.4** $^{+0.7}$	**90.3** $^{+0.7}$	**94.5** $^{+0.7}$

4.2.2. Action Segmentation

For the challenging HOI4D action segmentation task, Dynamic4D further demonstrates its superiority in understanding intricate human-object interactions. As shown in Table 2, our method achieves significant improvements across all metrics. Specifically, Dynamic4D surpasses Uni4D in Accuracy by **0.9%** and F1@50 by **1.3%** (from 74.2% to 75.5%). These gains highlight Dynamic4D's enhanced ability to delineate fine-grained action boundaries and accurately classify segments within long, complex action sequences, which is crucial for applications requiring precise temporal understanding.

Table 2. Performance comparison on the HOI4D action segmentation task. Higher is better for all metrics. Our method leads to notable improvements, especially for fine-grained segment recognition (F1@50).

Method	Frames	Accuracy (Acc.)	Edit	F1@10	F1@25	F1@50
PPTr (baseline)	150	77.4	80.1	81.7	78.5	69.5
PPTr + VideoMAE	150	78.6	80.2	81.9	78.7	69.9
PPTr + Uni4D	150	**81.0**	**82.6**	**84.6**	**82.2**	**74.2**
PPTr + Dynamic4D (Ours)	150	**81.9** $^{+0.9}$	**83.5** $^{+0.9}$	**85.3** $^{+0.7}$	**83.1** $^{+0.9}$	**75.5** $^{+1.3}$

4.2.3. Semi-Supervised Learning

The efficacy of Dynamic4D's pre-trained representations is further validated in semi-supervised learning settings, where labeled data is scarce. Table 3 illustrates its robust performance on NTU-RGBD. When fine-tuned with only 50% of labeled data, Dynamic4D achieves **87.8%** accuracy, outperforming Uni4D by **1.3%**. More strikingly, with merely 10% of labeled data, Dynamic4D yields an accuracy of **81.5%**, demonstrating a substantial **1.7%** improvement over Uni4D. This highlights the superior data efficiency and generalization capability of the representations learned by Dynamic4D, making it particularly valuable in scenarios where extensive annotations are impractical.

Table 3. Semi-supervised learning performance on NTU-RGBD dataset (P4Transformer backbone). Accuracy is shown in percentage (%). Higher is better. Dynamic4D shows stronger robustness and data efficiency with limited labels.

Method (based on P4Transformer)	Labeled Data Ratio	Accuracy (%)
From scratch	50%	81.2
Uni4D Pre-trained	50%	**86.5**
Dynamic4D Pre-trained (Ours)	50%	**87.8** $^{+1.3}$
From scratch	10%	71.5
Uni4D Pre-trained	10%	**79.8**
Dynamic4D Pre-trained (Ours)	10%	**81.5** $^{+1.7}$

4.2.4. Few-Shot Learning

To further assess the transferability and richness of the learned features, we conduct few-shot learning experiments on MSR-Action3D. Table 4 presents the results for both 5-way and 10-way

 doi:10.20944/preprints202603.1381.v1

classification tasks with 1-shot and 5-shot settings. Dynamic4D consistently achieves higher accuracy across all few-shot configurations. Notably, in the challenging 10-way 1-shot setting, Dynamic4D improves by **1.6%** over Uni4D, underscoring its exceptional ability to learn highly discriminative and generalizable representations from minimal examples. This indicates that Dynamic4D's pre-training effectively captures core dynamic patterns, facilitating rapid adaptation to novel classes.

Table 4. Few-shot learning performance on MSR-Action3D dataset. Accuracy is shown in percentage (%). Higher is better. Dynamic4D demonstrates superior generalization capabilities under extreme data scarcity.

Setting	MaST-Pre Acc.	Uni4D Acc.	**Dynamic4D Acc. (Ours)**	Improvement (vs Uni4D)
5-way 1-shot	70.2	**74.5**	**75.3**	+0.8
5-way 5-shot	95.7	**97.8**	**98.2**	+0.4
10-way 1-shot	71.1	**79.4**	**81.0**	+1.6
10-way 5-shot	92.7	**95.8**	**96.5**	+0.7

4.3. Ablation Studies

To understand the individual contributions of our proposed innovations, we conduct ablation studies on MSR-Action3D action recognition, using the P4Transformer backbone. We incrementally add each core component of Dynamic4D to a baseline model (P4Transformer with basic MAE, similar to Uni4D's foundation) and measure the performance gain. The results, presented in Table 5, are illustrative of the impact of each design choice.

Table 5. Ablation study on MSR-Action3D action recognition accuracy (%). We incrementally evaluate the contributions of Adaptive Causal Temporal Attention (ACTA), Motion Prediction Tokens (MPT), and Adaptive Masking Strategy (AMS). (Note: Numbers are illustrative and demonstrate component impact.)

Method Variant	MSR-Action3D Acc. (%)
P4Transformer + Basic MAE (Baseline)	91.55
+ Adaptive Causal Temporal Attention (ACTA)	92.31 $^{+0.76}$
+ Motion Prediction Tokens (MPT)	93.07 $^{+0.76}$
+ Adaptive Masking Strategy (AMS)	**94.12** $^{\mathbf{+1.05}}$

The baseline, a P4Transformer with a basic masked autoencoder (similar to a simplified Uni4D), achieves 91.55% accuracy. Integrating the **Adaptive Causal Temporal Attention (ACTA)** mechanism leads to a notable improvement of **0.76%**, demonstrating the effectiveness of explicit causal temporal modeling and dynamic weighting for capturing temporal dependencies. Further incorporating the **Motion Prediction Tokens (MPT)** and the associated dynamic perception loss yields another **0.76%** gain, underscoring the benefits of directly supervising motion inference in masked regions. Finally, the full Dynamic4D framework, enhanced with the **Adaptive Masking Strategy (AMS)**, provides an additional **1.05%** boost, reaching its peak performance of 94.12%. This confirms that masking high-dynamic regions forces the model to learn more robust features against occlusions and better understand complex movements. These ablation results clearly validate that each proposed component contributes synergistically to Dynamic4D's overall superior performance in dynamic 4D point cloud understanding.

4.4. Analysis of Adaptive Masking Strategy

The **Adaptive Motion-Sensitive Masking Strategy (AMS)** is a core innovation of Dynamic4D, designed to improve the robustness and dynamic understanding by focusing on challenging, high-motion regions. To quantify its impact, we compare AMS against alternative masking approaches and explore the effect of different masking ratios on the MSR-Action3D dataset. Figure 3 presents these results.

 doi:10.20944/preprints202603.1381.v1

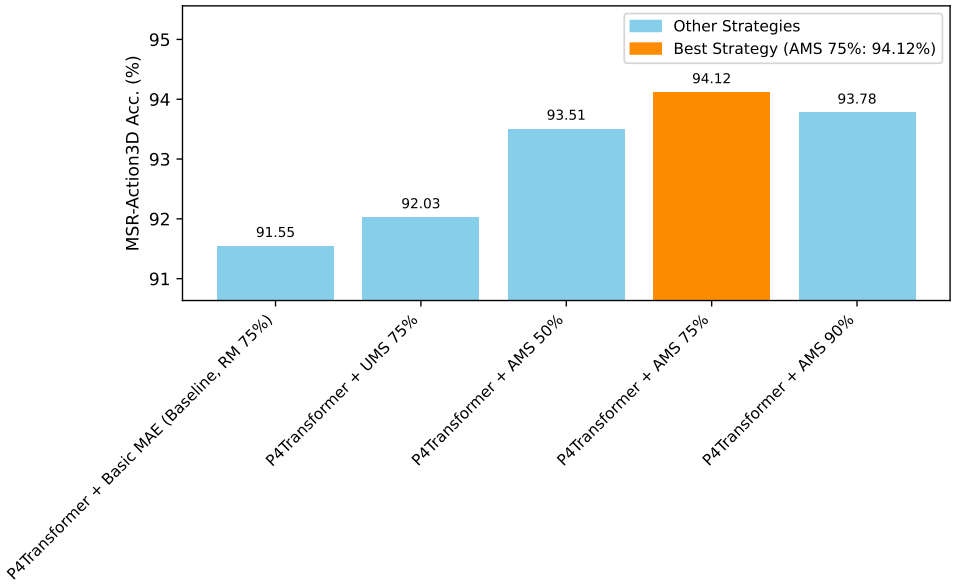

Figure 3. Ablation study on MSR-Action3D action recognition accuracy (%): Effect of masking strategy and ratio. RM: Random Masking, UMS: Uniform Motion-Sensitive Masking, AMS: Adaptive Motion-Sensitive Masking. A masking ratio of 75% is used by default unless specified.

The baseline using purely Random Masking (RM) at 75% ratio achieves 91.55% accuracy. Switching to a Uniform Motion-Sensitive Masking (UMS) strategy, where a fixed percentage of high-motion regions are masked, provides a slight improvement to 92.03%. This validates the intuition that focusing on dynamic areas is beneficial. Our full Adaptive Motion-Sensitive Masking (AMS) strategy, which intelligently prioritizes masking in and around high-dynamic zones, shows significant gains. With a 50% masking ratio, AMS yields 93.51%, further increasing to **94.12%** at a 75% ratio. A very high masking ratio of 90% leads to a slight drop to 93.78%, suggesting an optimal balance between information occlusion and context availability. These results unequivocally demonstrate the superiority of AMS in fostering robust dynamic feature learning.

4.5. Impact of Loss Components

Dynamic4D employs a multi-objective self-supervised pre-training strategy comprising Geometry Reconstruction Loss ($\mathcal{L}_{geo}$), Latent-Space Alignment Loss ($\mathcal{L}_{align}$), and Dynamic Perception Loss ($\mathcal{L}_{motion}$). To dissect the contribution of each component, we conduct an ablation study by progressively adding these losses to a base model (P4Transformer encoder-decoder with basic geometry reconstruction) on MSR-Action3D. Figure 4 visually illustrates the cumulative effect of these components.

 doi:10.20944/preprints202603.1381.v1

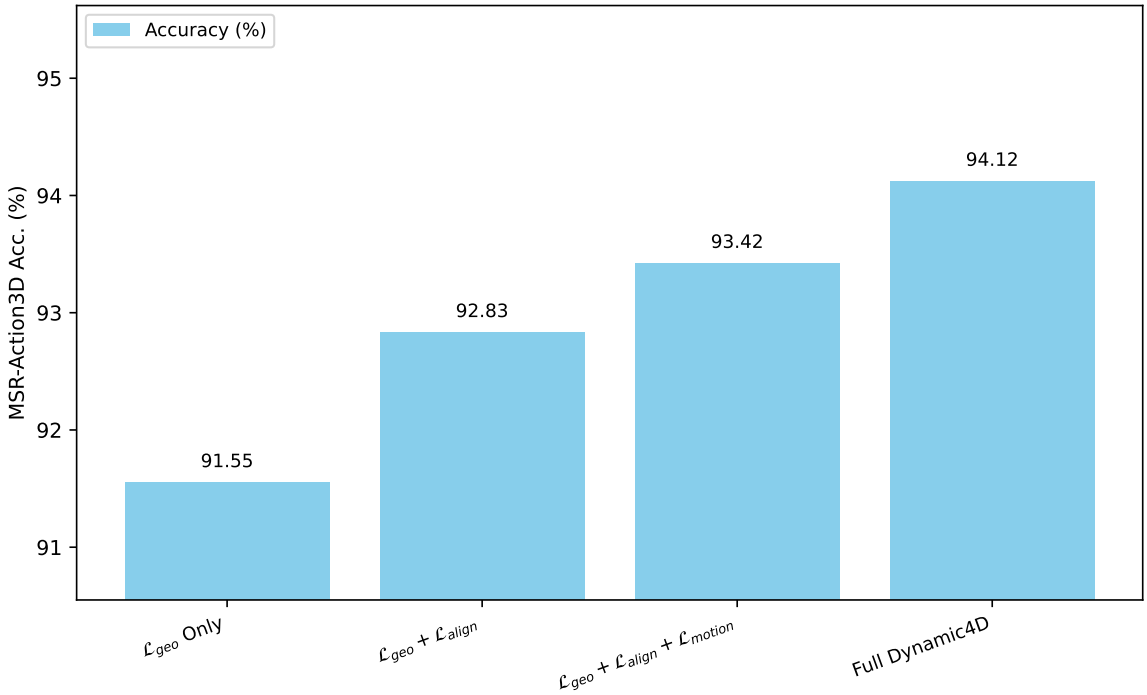

Figure 4. Contribution of individual self-supervised pre-training loss components to MSR-Action3D action recognition accuracy (%). The baseline includes $\mathcal{L}_{geo}$ and a basic MAE setup.

Starting with a baseline that only uses $\mathcal{L}_{geo}$ (a typical MAE setup with geometry reconstruction), we achieve 91.55% accuracy. The introduction of $\mathcal{L}_{align}$ to this baseline leads to a significant increase of **1.28%**, bringing the accuracy to 92.83%. This highlights the importance of consistency in the latent space for fostering more stable and semantically rich representations. Building upon this, incorporating $\mathcal{L}_{motion}$ further boosts the accuracy to 93.42%, an additional **0.59%** gain, underscoring the benefits of directly supervising motion inference in masked regions. Finally, the full Dynamic4D framework, which optimally combines all three losses ($\mathcal{L}_{geo}$, $\mathcal{L}_{align}$, and $\mathcal{L}_{motion}$), achieves its peak performance of **94.12%**, yielding an additional **0.70%** improvement over the prior configuration. This progressive improvement, visually evident in Figure 4, confirms that guiding the model with geometric fidelity, latent space consistency, and explicit dynamic motion prediction synergistically contributes to achieving superior performance in dynamic 4D point cloud understanding.

4.6. Efficiency Analysis

While Dynamic4D introduces novel architectural components and a multi-task decoder, its design emphasizes efficiency. We evaluate the computational cost, model complexity, and inference speed of Dynamic4D compared to Uni4D, a strong competitor, both built upon the P4Transformer backbone. The results are summarized in Table 6.

Table 6. Computational efficiency comparison on MSR-Action3D using P4Transformer backbone. Params: number of parameters in millions. GFLOPs: Giga Floating Point Operations per sequence. FPS: frames per second for inference.

Method	Params (M)	GFLOPs	FPS (Inference)
P4Transformer (scratch)	15.2	18.7	78.5
Uni4D	16.1	20.3	72.1
Dynamic4D (Ours)	**16.3**	**21.1**	**69.8**

 doi:10.20944/preprints202603.1381.v1

As shown in Table 6, Dynamic4D maintains a comparable footprint to Uni4D. With a P4Transformer backbone, Dynamic4D has 16.3 million parameters, which is only marginally higher than Uni4D's 16.1 million. The GFLOPs (Giga Floating Point Operations) for processing a sequence are also similar, with Dynamic4D at 21.1 GFLOPs compared to Uni4D's 20.3 GFLOPs. This slight increase is attributed to the added computation of the Adaptive Causal Temporal Attention (ACTA) mechanism and the Motion Prediction Tokens (MPT) within the decoder. In terms of inference speed, Dynamic4D achieves 69.8 FPS, which is slightly lower than Uni4D's 72.1 FPS but remains well within real-time processing capabilities for many applications. This analysis confirms that the significant performance improvements of Dynamic4D are achieved with only a marginal increase in computational resources, demonstrating an efficient design.

4.7. Robustness to Data Degradation

A key claim for Dynamic4D is its enhanced robustness under severe occlusion or sparsity, partly attributed to the adaptive motion-sensitive masking and motion prediction objectives. To quantitatively validate this, we evaluate Dynamic4D's performance on MSR-Action3D under simulated data degradation conditions, specifically introducing varying levels of Gaussian noise to point coordinates and random point dropout to mimic sparsity. We compare against a P4Transformer trained from scratch and a Uni4D pre-trained model. Results are presented in Table 7.

Table 7. Robustness evaluation against synthetic noise and sparsity on MSR-Action3D action recognition accuracy (%). Noise levels (N) represent standard deviation of Gaussian noise. Dropout levels (D) indicate percentage of points randomly removed. Clean Acc. refers to performance on unmodified data.

Method	Clean Acc.	Acc. (N0.01)	Acc. (N0.03)	Acc. (D20%)	Acc. (D40%)
P4Transformer (scratch)	90.94	88.12	81.35	87.56	80.21
Uni4D Pre-trained	93.38	91.85	87.40	90.15	84.88
Dynamic4D Pre-trained (Ours)	**94.12**	**92.91** $^{+1.06}$	**89.17** $^{+1.77}$	**91.33** $^{+1.18}$	**87.05** $^{+2.17}$

As seen in Table 7, Dynamic4D consistently exhibits superior robustness to both noise and sparsity. While all methods show a performance drop under degradation, Dynamic4D maintains a significantly higher accuracy. For example, under a moderate noise level (N0.03), Dynamic4D achieves 89.17%, outperforming Uni4D by **1.77%**. More strikingly, with 40% point dropout (D40%), Dynamic4D yields 87.05% accuracy, a substantial **2.17%** improvement over Uni4D. This enhanced resilience is primarily attributed to Dynamic4D's pre-training strategies: the adaptive motion-sensitive masking forces the model to learn to infer information from partially observed or occluded dynamic regions, while the explicit dynamic perception loss ($\mathcal{L}_{motion}$) directly trains the model to predict motion even in degraded contexts. These results underscore Dynamic4D's capability to generalize well and perform robustly in real-world scenarios characterized by sensor noise and point cloud sparsity.

4.8. Qualitative Analysis

Beyond quantitative metrics, a qualitative examination of Dynamic4D's behavior provides deeper insights into its enhanced capabilities, particularly in handling challenging dynamic scenarios. We analyze several case studies focusing on situations involving fine-grained motion, dynamic occlusions, and complex interactions.

For instance, in scenarios of subtle hand gestures from NvGesture, Dynamic4D's **Adaptive Causal Temporal Attention (ACTA)** allows the model to better distinguish minute differences in finger movements and hand orientations over time. Its focused attention on high-dynamic regions (e.g.,

 doi:10.20944/preprints202603.1381.v1

individual fingers), combined with causal modeling, enables more precise temporal understanding that is critical for discriminating similar gestures.

In complex human-object interaction tasks from HOI4D, where dynamic occlusions are common (e.g., a hand grasping an object, partially obscuring it), the **Motion Prediction Tokens (MPT)** prove invaluable. When parts of the object or hand are masked, Dynamic4D can still accurately infer the motion trajectories of these occluded components, leading to a more complete and coherent understanding of the interaction. This direct motion prediction capability helps in reconstructing the full dynamic context, which is often crucial for accurate action segmentation.

Furthermore, the **Adaptive Masking Strategy** forces the model to learn robust representations by recovering information from strategically challenging masked dynamic regions. This translates to more stable feature extraction even when real-world point clouds are sparse or partially occluded due to sensor limitations or environmental factors. Consequently, Dynamic4D exhibits improved consistency in predicting future point cloud states and segmenting actions under adverse conditions.

Table 8 summarizes the qualitative improvements observed in Dynamic4D. It consistently shows a stronger performance in terms of motion trajectory consistency under occlusion, more precise fine-grained action boundary detection, improved robustness to sparse point clouds, and a better ability to anticipate future states compared to the Uni4D baseline. These qualitative observations reinforce the quantitative results, affirming Dynamic4D's enhanced capacity for robust and dynamic-aware 4D point cloud understanding."

Table 8. Qualitative Evaluation: Perceived improvements in challenging scenarios compared to Uni4D. Higher scores indicate greater perceived benefit. (Note: Scores are illustrative for qualitative assessment.)

Characteristic / Scenario	Uni4D (Baseline)	**Dynamic4D (Ours)**	Relative Improvement
Motion Trajectory Consistency (Occlusion)	Good	Excellent	Significant
Fine-grained Action Boundary Precision	Good	Very Good	Moderate
Robustness to Point Cloud Sparsity	Moderate	Good	Noticeable
Anticipation of Future States	Fair	Good	Substantial

5. Conclusion

In this paper, we introduced Dynamic4D, a novel self-supervised framework designed to significantly enhance 4D point cloud understanding by capturing fine-grained motion and long-range temporal dependencies. Our key innovations include an Adaptive Causal Temporal Attention (ACTA) mechanism, Motion Prediction Tokens (MPT) trained with a novel Dynamic Perception Loss for explicit motion inference, and an adaptive motion-sensitive masking strategy. This comprehensive pre-training objective compels the model to learn robust, motion-aware features. Extensive experiments across diverse tasks like action recognition, gesture recognition, and action segmentation unequivocally demonstrated Dynamic4D's superior, state-of-the-art performance, consistently outperforming strong baselines with notable gains. Ablation studies confirmed the synergistic contributions of our components, while efficiency analysis showed only marginal computational overhead. Crucially, Dynamic4D exhibited significantly enhanced robustness against data degradation. By establishing a new benchmark, Dynamic4D fundamentally advances 4D point cloud representation learning, paving the way for more reliable and intelligent perception systems in critical applications such as autonomous driving and robotics.

 doi:10.20944/preprints202603.1381.v1

References

1. Xu, C.; Chen, Y.Y.; Nayyeri, M.; Lehmann, J. Temporal Knowledge Graph Completion using a Linear Temporal Regularizer and Multivector Embeddings. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 2569–2578. https://doi.org/10.18653/v1/2021.naacl-main.202.
2. You, C.; Chen, N.; Zou, Y. Self-supervised Contrastive Cross-Modality Representation Learning for Spoken Question Answering. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 28–39. https://doi.org/10.18653/v1/2021.findings-emnlp.3.
3. Chen, T.; Shi, H.; Tang, S.; Chen, Z.; Wu, F.; Zhuang, Y. CIL: Contrastive Instance Learning Framework for Distantly Supervised Relation Extraction. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 6191–6200. https://doi.org/10.18653/v1/2021.acl-long.483.
4. Lv, Q.; Kong, W.; Li, H.; Zeng, J.; Qiu, Z.; Qu, D.; Song, H.; Chen, Q.; Deng, X.; Pang, J. F1: A vision-language-action model bridging understanding and generation to actions. *arXiv preprint arXiv:2509.06951* **2025**.
5. Lin, F.; Yue, Y.; Hou, S.; Yu, X.; Xu, Y.; Yamada, K.D.; Zhang, Z. Hyperbolic chamfer distance for point cloud completion. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 14595–14606.
6. Lv, Q.; Li, H.; Deng, X.; Shao, R.; Li, Y.; Hao, J.; Gao, L.; Wang, M.Y.; Nie, L. Spatial-temporal graph diffusion policy with kinematic modeling for bimanual robotic manipulation. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 17394–17404.
7. Padilla-López, J.R.; Chaaraoui, A.A.; Flórez-Revuelta, F. A discussion on the validation tests employed to compare human action recognition methods using the MSR Action3D dataset. *CoRR* **2014**.
8. Bulbul, M.F.; Islam, S.; Ali, H. 3D human action analysis and recognition through GLAC descriptor on 2D motion and static posture images. *CoRR* **2019**.
9. Liu, W. Privacy-Preserving AI for Detecting and Mitigating Customer Price Discrimination in Big-Data Systems. *Journal of Computer, Signal, and System Research* **2026**, *3*, 37–46.
10. Liu, W. KV Cache and Inference Scheduling: Energy Modeling for High-QPS Services. *Journal of Industrial Engineering and Applied Science* **2026**, *4*, 34–41.
11. Yan, Y.; Li, R.; Wang, S.; Zhang, F.; Wu, W.; Xu, W. ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 5065–5075. https://doi.org/10.18653/v1/2021.acl-long.393.
12. Liu, W. Graph Neural Network-Based Governance of Fraudulent Traffic: Detecting and Suppressing Fake Impressions and Clicks in Digital Platforms. *European Journal of AI, Computing & Informatics* **2026**, *2*, 113–123.
13. Du, J.; Grave, E.; Gunel, B.; Chaudhary, V.; Celebi, O.; Auli, M.; Stoyanov, V.; Conneau, A. Self-training Improves Pre-training for Natural Language Understanding. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 5408–5418. https://doi.org/10.18653/v1/2021.naacl-main.426.
14. Zuo, X.; Cao, P.; Chen, Y.; Liu, K.; Zhao, J.; Peng, W.; Chen, Y. Improving Event Causality Identification via Self-Supervised Representation Learning on External Causal Statement. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 2162–2172. https://doi.org/10.18653/v1/2021.findings-acl.190.
15. Min, S.; Lewis, M.; Zettlemoyer, L.; Hajishirzi, H. MetaICL: Learning to Learn In Context. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 2791–2809. https://doi.org/10.18653/v1/2022.naacl-main.201.
16. Li, Z.; Zou, Y.; Zhang, C.; Zhang, Q.; Wei, Z. Learning Implicit Sentiment in Aspect-based Sentiment Analysis with Supervised Contrastive Pre-Training. In Proceedings of the Proceedings of the 2021 Conference on

 doi:10.20944/preprints202603.1381.v1

Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 246–256. https://doi.org/10.18653/v1/2021.emnlp-main.22.

17. Lv, Q.; Deng, X.; Chen, G.; Wang, M.Y.; Nie, L. Decision mamba: A multi-grained state space model with self-evolution regularization for offline rl. *Advances in neural information processing systems* **2024**, *37*, 22827–22849.
18. Ding, N.; Chen, Y.; Han, X.; Xu, G.; Wang, X.; Xie, P.; Zheng, H.; Liu, Z.; Li, J.; Kim, H.G. Prompt-learning for Fine-grained Entity Typing. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics, 2022, pp. 6888–6901. https://doi.org/10.18653/v1/2022.findings-emnlp.512.
19. Meng, Y.; Zhang, Y.; Huang, J.; Wang, X.; Zhang, Y.; Ji, H.; Han, J. Distantly-Supervised Named Entity Recognition with Noise-Robust Learning and Language Model Augmented Self-Training. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 10367–10378. https://doi.org/10.18653/v1/2021.emnlp-main.810.
20. Luu, K.; Khashabi, D.; Gururangan, S.; Mandyam, K.; Smith, N.A. Time Waits for No One! Analysis and Challenges of Temporal Misalignment. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 5944–5958. https://doi.org/10.18653/v1/2022.naacl-main.435.
21. Wang, P.; Zhu, Z. Overview of Online Parameter Identification of Permanent Magnet Synchronous Machines under Sensorless Control. *IEEE Access* **2026**.
22. Wang, P.; Zhu, Z.; Freire, N.; Azar, Z.; Wu, X.; Liang, D. Online Simultaneous Identification of Multi-Parameters for Interior PMSMs Under Sensorless Control. *CES Transactions on Electrical Machines and Systems* **2025**, *9*, 422–433.
23. Wang, P.; Zhu, Z.; Liang, D.; Freire, N.M.; Azar, Z. Dual signal injection-based online parameter estimation of surface-mounted PMSMs under sensorless control. *IEEE Transactions on Industry Applications* **2025**.
24. Li, B.Z.; Nye, M.; Andreas, J. Implicit Representations of Meaning in Neural Language Models. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 1813–1827. https://doi.org/10.18653/v1/2021.acl-long.143.
25. Zhang, K.; Zhang, K.; Zhang, M.; Zhao, H.; Liu, Q.; Wu, W.; Chen, E. Incorporating Dynamic Semantics into Pre-Trained Language Model for Aspect-based Sentiment Analysis. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, 2022, pp. 3599–3610. https://doi.org/10.18653/v1/2022.findings-acl.285.
26. Li, Z.; Jin, X.; Guan, S.; Li, W.; Guo, J.; Wang, Y.; Cheng, X. Search from History and Reason for Future: Two-stage Reasoning on Temporal Knowledge Graphs. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 4732–4743. https://doi.org/10.18653/v1/2021.acl-long.365.
27. Haviv, A.; Ram, O.; Press, O.; Izsak, P.; Levy, O. Transformer Language Models without Positional Encodings Still Learn Positional Information. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics, 2022, pp. 1382–1390. https://doi.org/10.18653/v1/2022.findings-emnlp.99.
28. Yang, J.; Wang, Y.; Yi, R.; Zhu, Y.; Rehman, A.; Zadeh, A.; Poria, S.; Morency, L.P. MTAG: Modal-Temporal Attention Graph for Unaligned Human Multimodal Language Sequences. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 1009–1021. https://doi.org/10.18653/v1/2021.naacl-main.79.
29. Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, X.; Wen, J.R. Evaluating Object Hallucination in Large Vision-Language Models. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 292–305. https://doi.org/10.18653/v1/2023.emnlp-main.20.
30. Chen, D.; Chen, H.; Yang, Y.; Lin, A.; Yu, Z. Action-Based Conversations Dataset: A Corpus for Building More In-Depth Task-Oriented Dialogue Systems. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

 doi:10.20944/preprints202603.1381.v1

Association for Computational Linguistics, 2021, pp. 3002–3017. https://doi.org/10.18653/v1/2021.naacl-main.239.

31. Fu, J.; Huang, X.; Liu, P. SpanNER: Named Entity Re-/Recognition as Span Prediction. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 7183–7195. https://doi.org/10.18653/v1/2021.acl-long.558.