Harmonizing Stochasticity and Determinism: Scene-responsive Diverse Human Motion Prediction

Zhenyu Lou[1] **Qiongjie Cui**[2*]
Tuo Wang[4] Zhenbo Song[2] Luoming Zhang[1] Cheng Cheng[5] Haofan Wang[3]
Xu Tang[3] Huaxia Li[3] Hong Zhou[1]
[1]Zhejiang University, [2]Nanjing University of Science and Technology, [3]Xiaohongshu Inc,
[4]University of Texas at Austin, [5]Concordia University,
11915044@zju.edu.cn cuiqiongjie@126.com

Figure 1: Comparison of our DiMoP3D with the SoTA baseline [5]. Purple meshes represent observations, and yellow meshes denote predictions. DiMoP3D produces high-fidelity, diverse sequences tailored to real-world 3D scenes, while BeLFusion's inadequate scene context integration leads to issues such as object penetration, motion incoherence, and scene inconsistency.

Abstract

Diverse human motion prediction (HMP) is a fundamental application in computer vision that has recently attracted considerable interest. Prior methods primarily focus on the stochastic nature of human motion, while neglecting the specific impact of the external environment, leading to the pronounced artifacts in prediction when applied to real-world scenarios. To fill this gap, this work introduces a novel task: predicting diverse human motion within real-world 3D scenes. In contrast to prior works, it requires harmonizing the deterministic constraints imposed by the surrounding 3D scenes with the stochastic aspect of human motion. For this purpose, we propose DiMoP3D, a diverse motion prediction framework with 3D scene awareness, which leverages the 3D point cloud and observed sequence to generate diverse and high-fidelity predictions. DiMoP3D can comprehend the 3D scene and determine the probable target objects and their desired interactive pose based on the historical motion. Then, it plans the obstacle-free trajectories toward these interested objects and generates diverse and physically consistent future motions. On top of that, DiMoP3D identifies deterministic factors in the scene and integrates them into stochastic modeling, making the diverse HMP in realistic scenes become a controllable stochastic generation process. On two real-captured benchmarks, DiMoP3D has demonstrated significant improvements over state-of-the-art methods, showcasing its effectiveness in generating diverse and physically consistent motion predictions within real-world 3D environments. More details and the video demo are available at the webpage https://sites.google.com/view/dimop3d.

*Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-1256

1 Introduction

Human motion prediction (HMP), *i.e.,* forecasting future human poses based on observation, is crucial for applications including autonomous vehicles and human-robot collaboration [15, 20, 39, 44, 54, 60, 66]. Many existing works [1, 22, 40, 49, 76, 89] formulate HMP as a deterministic problem, aiming to generate a single future sequence. However, it fails to capture the inherent stochasticity of human motion, where multiple plausible outcomes can arise from a single observation. Recent research has shifted towards diverse or stochastic HMP, to achieve multiple plausible predictions [3, 6, 30, 59, 72], which holds the potential in real-world applications and is the focus of our work.

Recent advances in diverse HMP primarily focus on stochastic predictions, where a random factor from the latent space conditions the diversity of predictions alongside observed motions [3, 12, 30, 59, 71, 72, 82]. While these methods predict multiple plausible futures from a single past motion, they typically disregard the 3D environment, operating within an idealized, context-free framework. This limitation becomes apparent in real-world applications, where motion must conform to physical and semantic scene constraints [29, 67, 68, 78, 90], leading to issues like obstacle penetration and unrealistic interactions, as in Figure 1. This gap underscores the need for a new task that merges diverse HMP within real-world 3D scenes, enhancing both practicality and applicability.

Recognizing existing limitations, this work introduces a novel task, making diverse HMP within real-world 3D scenes. Its objective is to break the previous idealized context-free setup towarding a realistic and practical setting, which involves several key challenges: **(1) Harmonizing Stochasticity and Determinism:** This task necessitates a delicate harmonization between the stochastic nature of human motion and the deterministic constraints from 3D scenes, thereby broadening the scope of traditional HMP; **(2) Scene-Motion Intermodal Coordination:** It requires analyzing coordination between human motion and scene dynamics to align predictions with contextual elements, which involves identifying human intentions and potential interactive objects; **(3) Behaviorally Coherent Physical Consistency:** The predictions must adhere to deterministic constraints, including physical consistency (*e.g.,* avoid collision) and behavior coherence (*e.g.,* sitting on a chair, not lying).

To tackle these challenges, we introduce `DiMoP3D` (**Di**verse **M**otion **P**rediction in 3D Scenes), a framework for generating diverse, physically consistent, and plausible human motion predictions in real-world 3D scenes, which comprises three main components: (1) Context-aware Intermodal Interpreter analyzes potential areas of human interest and goals, essentially intentions within a scene. Our method enhances traditional scene understanding by integrating 3D point clouds with observed motions for context-aware intermodal analysis. It first encodes the point cloud, segments object instances, and then pinpoints potential interaction targets, emphasizing intended factors while filtering out less probable ones. This strategy further transforms the task of diverse HMP into a controllable stochastic generation process; (2) Behaviorally-consistent Stochastic Planner then constructs behaviorally consistent action plans, representing stochastic conditional factors. Recognizing that human interaction with specific objects often follows deterministic patterns, we prioritize predicting the final human pose upon reaching each target, and generate obstacle-free trajectories toward it; (3) Self-prompted Motion Generator harmonizes the stochastic nature of human motion with deterministic constraints in a self-prompted manner to produce varied predictions based on the conditional factor. To ensure coherence with this factor, it employs a denoising diffusion model, guiding the motion denoising process toward a deterministic, obstacle-free final state.

Our contributions are threefold: (1) We introduce a novel and challenging task of predicting diverse human motions within real-world 3D scenes, advancing beyond the traditional scope of diverse HMP from an idealized context-free setting to a more realistic and practical one. (2) We propose `DiMoP3D` to tackle this task. It harmonizes the deterministic constraints of 3D scenes with the stochastic nature of human motion, enabling diverse and plausible motion predictions in real-world scenarios. (3) Evaluations on two real-captured benchmarks, GIMO and CIRCLE, show that `DiMoP3D` significantly outperforms existing SoTA methods, particularly in terms of physical consistency.

2 Related Work

2.1 Stochastic Human Motion Prediction

Human motion prediction diverges into deterministic and stochastic methods. Deterministic models aim to predict a singular sequence that closely aligns with future movements [18, 19, 40, 49, 76],

yet often encounter quality degradation over longer time spans (> 1-sec) due to the stochastic nature of human movement. In contrast, stochastic HMP [2, 3, 5, 44, 82] embraces this variability by generating a range of plausible future motions and modeling the distribution of human behaviors. Notably, it also encompasses the most probable future sequence targeted by deterministic models, as a likely scenario within its broader distribution. Such methodologies enhance applications across autonomous driving [7, 33, 43], patient care [10, 42], and human flow prediction [32, 36], by embracing a wider range of potential scenarios and have become a focal point of contemporary research.

Dominant approaches include VAEs, GANs, flow networks, and diffusion models [3, 6, 12, 30, 38, 59, 71, 72, 82]. Despite promising progress in modeling stochastic human motions, a critical challenge persists: human motion is not only stochastic but also heavily influenced by the external environment. In real-world settings, human motion is intricately intertwined with surrounding scenes, necessitating that future trajectories comply with the scene's physical constraints (*e.g.,* avoid object penetration). Additionally, predicted motions should be semantically consistent with the expected human-object interactions (*e.g.,* sit for a chair, lie for a bed). Addressing these requirements calls for a sophisticated approach to diverse HMP that balances the stochastic aspect of human motion with the deterministic factors imposed by the surrounding 3D scenes, which is the focus of our work.

2.2 3D Scene Encoding and Scene-aware Motion Prediction

3D scene understanding is essential in various applications, prompting extensive research on representations like RGB-D maps [9, 52, 61], scene graphs [77, 79, 84], 3D voxels [23, 48, 74], and notably, 3D point clouds [55, 56, 65, 87]. Recognizing the importance of 3D scene information in human motion prediction [14, 21, 26, 35, 80], our work integrates 3D point clouds as the scene representation due to their direct derivation from sensing technologies.

To enhance fidelity of HMP in real-world scenarios, the connection between human actions and scene context has made scene-aware motion generation a major research focus [8, 17, 28, 88]. Early methods [8, 17, 67] relied on object bounding boxes, 2D images, and depth maps, which are insufficient for capturing real 3D environments. Recent advances use 3D point clouds for scene representation [29, 69, 88]: GIMO [88] employs a bidirectional transformer to fuse human motion and scene features, [50] predicts future contact maps, and [46] extracts global and local salient points.

However, these methods predict a single sequence [4, 50, 88], whereas our approach models a distribution of potential outcomes. Some stochastic methods add diversity, but [8] relies on 2D inputs, limiting 3D interactions, and [28] uses predefined objects without inferring targets. In contrast, our model parses the 3D scene through cross-modal, object-specific human interest analysis and predicts scene-aware motions in real 3D scenarios, greatly enhancing adaptability and authenticity.

2.3 Motion Synthesis in 3D Scenes

Recent advancements in paired scene-motion data [4, 27, 88] have sparked a new research direction in synthesizing motions in 3D scenes [29, 41, 47, 53, 69, 78]. Specifically, SAMP [28] employs a conditional variational autoencoder (cVAE) to generate one frame per forward pass within a three-stage stochastic pipeline, COUCH [86] introduces a human-chair dataset with an autoregressive, contact-satisfying method, and DN-Synt [68] proposes a hierarchical framework for effective scene-aware human motion synthesis. With the rise of language models, methods utilizing natural language prompts have emerged [16, 69, 75, 85]. HUMANISE [69] proposes an attention-based language-prompted synthesis method. Additionally, diffusion models have shown significant promise [16, 29, 37, 63, 64, 70]. Notably, AffordMotion [70] employs scene affordance as an intermediate representation, achieving state-of-the-art performance in object interaction synthesis.

Our novel task of scene-aware diverse HMP is partly inspired by advances in motion synthesis but is tailored for distinctly different applications and challenges. This task is further distinguished by key innovations: (1) Temporal-Dependent Prompting. Contrary to motion synthesis, which often lacks temporal context and follows static user instructions, our approach conditions predictions on the unique prompt of historical human motion, enabling autonomous inference of human intentions. (2) Context-Aware Scene Parsing. Moving beyond traditional scene understanding, our method infers potential movement targets through the integration of context-aware intermodal insights. Utilizing scene determinism to enhance the fidelity of diverse HMP predictions.

 https://doi.org/10.52202/079017-1256

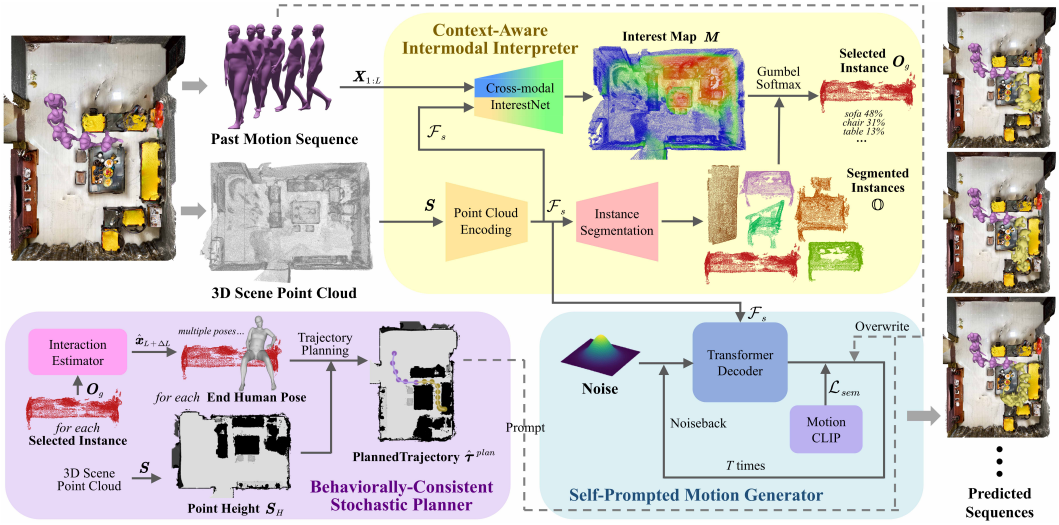

Figure 2: The architecture of DiMoP3D. DiMoP3D incorporates two modalities of input, the past motion and the 3D scene point cloud. Initially, the Context-aware Intermodal Interpreter encodes the point cloud to features $\mathcal{F}_s$, identifies interactive objects $\mathbb{O}$, and uses a cross-modal InterestNet to pinpoint potential interest areas, sampling a target instance $\boldsymbol{O}_g$ according to interest map $\boldsymbol{M}$. Following this, the Behaviorally-consistent Stochastic Planner forecasts the interactive human end-pose $\hat{\boldsymbol{x}}_{L+\Delta L}$, and devises an obstacle-free trajectory $\hat{\boldsymbol{\tau}}^{plan}$ towards this pose. The sampled end-pose and trajectory are incorporated as a stochastic conditional factor to prompt the Self-prompted Motion Generator to generate physically consistent future motions.

3 Method

3.1 Problem Setup

Given L historical human poses $\boldsymbol{X}_{1:L} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_L\}$ within 3D scenes represented by point clouds $\boldsymbol{S} \in \mathbb{R}^{n_p \times 6}$, our goal is to predict K different scene-consistent future motions $\{\hat{\boldsymbol{X}}^{(i)}_{L:L+\Delta L}\}_{i=1}^{K}$. Here, n_p denotes the number of points, each described by 3D spatial coordinates and RGB color information. L and ΔL denote the lengths of the observed and predicted sequences, respectively. Each pose is described as the SMPL-X representation $\boldsymbol{x}_l = (\boldsymbol{t}_l, \boldsymbol{o}_l, \boldsymbol{p}_l)$ [88], where $\boldsymbol{t}_l \in \mathbb{R}^3$ denotes the global translation, $\boldsymbol{o}_l \in SO(3)$ denotes the orientation, and $\boldsymbol{p}_l \in \mathbb{R}^{32}$ refers to the body pose embedding. We set $L = 3\text{-}sec$ and $\Delta L = 5\text{-}sec$ to achieve a long-term prediction [40, 49]. Then, the task can be formulated as:

$$P(\hat{\boldsymbol{X}}_{L:L+\Delta L}|\boldsymbol{X}_{1:L}, \boldsymbol{S}) = \int \max_{\theta} P(\hat{\boldsymbol{X}}_{L:L+\Delta L}|\boldsymbol{X}_{1:L}, \boldsymbol{S}, \theta) P(\theta|\boldsymbol{X}_{1:L}, \boldsymbol{S}) d\theta. \tag{1}$$

The stochastic process θ is sampled jointly from the past motion $\boldsymbol{X}_{1:L}$ and the scene $\boldsymbol{S}$, and then utilized to condition the prediction motion $\hat{\boldsymbol{X}}_{L:L+\Delta L}$. We propose DiMoP3D to solve this novel task, which involves the following novelties: **Context awareness**: unlike traditional diverse HMP methods [12, 59, 71] that focus solely on human motion, our task is more challenging as it requires harmonizing the stochastic nature of human motion with the deterministic constraints of the surrounding 3D scenes. **Autonomous intention estimation**: different from motion synthesis, our task requires independent intention estimation based on past motion, to prompt future motion prediction.

3.2 Context-Aware Intermodal Interpreter

Scene information plays a crucial role in predicting future motion [8, 17, 67]. Despite the stochastic manner of human motion, it is still feasible to deduce human interests and likely goals within a scene. For instance, in Figure 2, the door behind the person is unlikely to be the target based on their trajectory away from it, while the sofa, coffee table, and distant chair may emerge as potential points of interest. Diverging from traditional scene parsing methods, our approach integrates a scene-motion intermodal coordination to better suppose human intentions in real-world settings.

https://doi.org/10.52202/079017-1256

In machine vision, 3D point clouds sourced directly from sensing devices have become fundamental for scene representation [55, 65, 87] and serve as the input for our scene interpreter. Acknowledging the significant influence of past motions and scene context on future human movements, we emphasize the need for a context-aware intermodal analysis that integrates historical motion with scene point clouds to infer potential human intentions. Furthermore, since human movements typically involve interactions with target objects [11, 24, 81], our approach identifies objects within the scene and computes a human interest score for each, rather than analyzing isolated points. This helps determine specific interactive targets, making the motion prediction process more controllable and enabling diverse prediction by sampling different interactive targets.

To achieve this, our scene interpreter employs a UNet-like encoder-decoder architecture [58], comprising an instance segmenter for object recognition and an interest net for human interest inference. These two modules share the same point cloud encoder for efficiency but use different decoders, as illustrated in the yellow box in Figure 2. The shared encoder processes and downsamples the point cloud $\boldsymbol{S}$ into a compact feature representation $\mathcal{F}_s \in \mathbb{R}^{n'_p \times c}$, with $n'_p \ll n_p$ ($8 \leq n'_p \leq 50$ in our cases), and c represents the feature dimension. Subsequently, two decoders are employed to segment objects $\mathbb{O}$, and predict human interests $\boldsymbol{M}$ in the scene, respectively:

$$\mathcal{F}_s = \text{SceneEncoder}(\boldsymbol{S}), \;\; \mathbb{O} = \text{InstanceSegmenter}(\mathcal{F}_s), \;\; \boldsymbol{M} = \text{InterestNet}(\mathcal{F}_s, \boldsymbol{X}_{1:L}). \quad (2)$$

Here, $\mathbb{O} = \{\boldsymbol{O}_1, \boldsymbol{O}_2, ..., \boldsymbol{O}_{n_o}\}$ denotes the set of n_o segmented objects, each being a subset of the scene pointcloud ($\boldsymbol{O}_i \subseteq \boldsymbol{S}$). $\boldsymbol{M} \in \mathbb{R}^{n_p \times 1}$ denotes the per-point interest map, with higher values indicating a greater likelihood of targeting specific scene elements. Once $\boldsymbol{M}$ is obtained, we compute the probability P_i of each object $\boldsymbol{O}_i \in \mathbb{O}$ being selected as an interaction target:

$$M_i = \sum \boldsymbol{M}[p] \,/\, len(\boldsymbol{O}_i), \;\; p \in \boldsymbol{O}_i, \quad (3)$$

$$\{P_1, P_2, ..., P_{n_o}\} = \text{Softmax}(\{M_1, M_2, ..., M_{n_o}\}; \phi), \quad (4)$$

where $len(\boldsymbol{O}_i)$ denotes the number of points in $\boldsymbol{O}_i$, $\boldsymbol{M}[p]$ is the interest value of point p, and $\phi = 0.5$ represents the temperature factor that controls randomness in sampling. During inference, we sample the target object $\boldsymbol{O}_g (g \in \{1, 2, ..., n_o\})$ based on the probability distribution $\{P_1, P_2, ..., P_{n_o}\}$.

In cases where there is no human-object interaction or objects are beyond reach within ΔL, no explicit target object exists. To handle this, we include the ground as a potential target, voxelized into smaller patches to improve granularity. Each ground patch, treated as an individual object, has a side length of $s = 0.5$ meters to balance accuracy and efficiency. The interest scores and interaction probabilities for ground patches are then calculated similarly to other objects, as in Eq.3 and Eq.4.

Our scene interpreter aligns observed motions with scene context, filtering out less likely engagement areas and identifying potential targets. We note that this approach makes the diverse HMP controlled by those deterministic elements of the scene, thereby enhancing the physical consistency of predictions. Representing human intention through scene-motion intermodal analysis, the selected target object $\boldsymbol{O}_g$ directs the subsequent planning process, as outlined below.

3.3 Behaviorally-Consistent Stochastic Planner

Ensuring collision-free and scene-consistent human behavior is crucial but often overlooked, while learning these patterns directly requires impractically large datasets. To tackle these challenges, we employ an action planner to plan obstacle-free trajectories toward the target $\boldsymbol{O}_g$ and deterministically predict the interactive human end-pose $\hat{\boldsymbol{x}}_{end}$ associated with $\boldsymbol{O}_g$. These intermediate predictions serve as stochastic conditional factors, guiding to craft future motions that respect physical constraints and typical human-environment interactions while incorporating motion diversity.

To enable effective navigation and collision avoidance, we utilize a scene height map $\boldsymbol{S}_H \in \mathbb{R}^{n_s \times 1}$ to delineate accessible areas and detect obstacles, inspired by [68, 73]. We first compute obstacle-free trajectories toward the target object $\boldsymbol{O}_g$ using a modified A* algorithm (details on generating diverse trajectories are in Appendix B), and then employ a single-layer transformer ψ to predict per-frame human velocity, sampling discretize points from the planned trajectory:

$$\boldsymbol{\tau}^{plan} = \text{A*}(\boldsymbol{S}_H; \boldsymbol{X}_{1:L}, \hat{\boldsymbol{x}}_{end}), \quad (\hat{\boldsymbol{\tau}}^{plan}, t_{end}) = Sample(\boldsymbol{\tau}^{plan}, \psi(\boldsymbol{X}_{1:L})). \quad (5)$$

Here, $\boldsymbol{\tau}^{plan}$ denotes the continuous trajectory, $\hat{\boldsymbol{\tau}}^{plan} = \{\hat{\boldsymbol{t}}^{plan}_{L+1}, \hat{\boldsymbol{t}}^{plan}_{L+2}, ..., \hat{\boldsymbol{t}}^{plan}_{L+\Delta L}\}$ represents the sampled discretize trajectory points, and t_{end} is the estimated timestamp for the end of the interactive

https://doi.org/10.52202/079017-1256

motion. Since the length of this trajectory varies across sequences, the human may not reach the target object exactly at the prediction horizon $\Delta L = 5\text{-}sec$. In these cases, if the target object is too distant to reach within ΔL ($t_{end} > \Delta L$), we truncate the trajectory $\boldsymbol{\tau}^{plan}$ to fit within ΔL and adjust the target to the nearest ground patch. Conversely, if the target is reached too early ($t_{end} < \Delta L$), we keep the subject relatively static after t_{end}. This adjustment ensures that the planned trajectory aligns with the prediction horizon. We then discretize the continuous trajectory to obtain the per-frame global translation $\hat{\boldsymbol{\tau}}^{plan}$ to guide subsequent motion generation:

$$\hat{\boldsymbol{\tau}}^{plan} = \{\hat{\boldsymbol{t}}_{L+1}^{plan}, \hat{\boldsymbol{t}}_{L+2}^{plan}, ..., \hat{\boldsymbol{t}}_{L+\Delta L}^{plan}\}. \tag{6}$$

In addition to physical constraints, human interactions with specific objects often follow deterministic patterns despite potential action diversity. For instance, "set," and "wipe," are reasonable actions for a table, whereas "sit" and "lie" are not. Traditional diverse HMP methods typically overlook these Human-Object Interaction (HOI) patterns, leading to motion and scene inconsistencies. Differing from these methods, our approach predicts the target object $\boldsymbol{O}_g$ in advance, enabling us to predict the interactive HOI end-pose $\hat{\boldsymbol{x}}_{end}$ before the full-sequence prediction:

$$\hat{\boldsymbol{x}}_{end} = \text{HOI-Estimator}(\boldsymbol{O}_g). \tag{7}$$

This end-pose represents the final state of the prediction, secures appropriate human interaction with the scene. To this end, our planner constructs an obstacle-free trajectory, and the predicted end-pose (along with t_{end}) as a stochastic conditional factor θ from a scene-motion intermodal perspective:

$$\theta = (\hat{\boldsymbol{\tau}}^{plan}, \hat{\boldsymbol{x}}_{end}, t_{end}). \tag{8}$$

This factor further prompts the motion generator to predict behaviorally consistent motions.

3.4 Self-Prompted Motion Generator

By constructing a stochastic conditional factor θ in advance, DiMoP3D operates as a self-prompted motion generator, which harmonizes the stochastic factor with deterministic motion generation rooted in θ. To generate diverse predictions that closely align with the predicted conditional factor θ, we utilize a motion diffuser, taking advantage of the diffusion model's ability to effectively guide intermediate results [12, 59, 71]. Additionally, to further maintain semantic coherence and physical consistency, we propose a semantic alignment inspector to supervise the denoising process.

For simplicity, we denote the sequence at noising step t as $\boldsymbol{X}^t$. Diffusion is modeled as a Markov noising process $\{\boldsymbol{X}^t\}_{t=0}^{T}$, with $\boldsymbol{X}^0$ drawn from the data distribution, and:

$$q(\boldsymbol{X}^t|\boldsymbol{X}^{t-1}) = \mathcal{N}(\sqrt{\alpha_t}\boldsymbol{X}^{t-1}, (1-\alpha_t)\boldsymbol{I}). \tag{9}$$

Here $\alpha_t \in (0, 1)$. When α_T approaches 0, we approximate $\boldsymbol{X}_{1:L+\Delta L}^T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, where $\boldsymbol{0}$ and $\boldsymbol{I}$ represent the zero matrix and the identity matrix, respectively. To effectively integrate scene and motion features in our predictions, our diffusion model $\mathcal{M}_D$ employs a transformer decoder to model distribution akin to the reversed diffusion process, leveraging its capacity for cross-modal attention. Instead of predicting noise, we predict the clean sample directly, following [63, 64, 83]:

$$\bar{\boldsymbol{X}}^0 = \mathcal{M}_D(\hat{\boldsymbol{X}}^t, \mathcal{F}_s, t), \tag{10}$$

where $\bar{\boldsymbol{X}}^0$ represents the intermediate denoising result at each denoising step.

To align the predicted sequence with the observed $\boldsymbol{X}_{1:L}$, the planned trajectory $\hat{\boldsymbol{\tau}}^{plan}$, and the forecasted end-pose $\hat{\boldsymbol{x}}_{end}$ at time t_{end}, we adjust the corresponding segments after each denoising step. To be sepecific, for each frame $\bar{\boldsymbol{x}}_i^0 = (\bar{\boldsymbol{t}}_i^0, \bar{\boldsymbol{o}}_i^0, \bar{\boldsymbol{p}}_i^0)$:

$$\bar{\boldsymbol{x}}_i^0 = \begin{cases} \boldsymbol{x}_i^0 & \text{if } i \leq L, \\ (\hat{\boldsymbol{t}}_i^{plan}, \bar{\boldsymbol{o}}_i^0, \bar{\boldsymbol{p}}_i^0) & \text{if } L < i < L + t_{end}, \\ \hat{\boldsymbol{x}}_{L+\Delta L} & \text{if } i + t_{end}, \\ \bar{\boldsymbol{x}}_i^0 & \text{if } i > L + t_{end}, \end{cases} \quad \text{for each frame } i. \tag{11}$$

This modified prediction is then noised back before the next denoising step:

$$\hat{\boldsymbol{X}}^{t-1} = \mathcal{N}(\sqrt{\alpha_t}\boldsymbol{X}^{t-1}, \gamma_t\boldsymbol{I}), \tag{12}$$

where γ_t represents the posterior variance based on α at step t. Upon completing T denoising steps, the motion generator yields a cohesive sequence $\hat{\boldsymbol{X}}^0$, which integrates smoothly with the observed sequence and aligns with the planned trajectory and goals.

To enhance the consistency between the predicted motion and the target object, we introduce a semantic alignment inspector leveraging MotionCLIP [62]. It computes a HOI semantic loss via natural language descriptions as follows:

$$\mathcal{L}_{sem} = \frac{1}{L+\Delta L} \sum_{l=1}^{L+\Delta L} \Big(1 - \cos\Big(\text{MC}_{Motion}(\hat{\boldsymbol{X}}^t_{l:L+\Delta L}), \text{MC}_{Text}(\mathbb{D}_g)\Big)\Big). \tag{13}$$

Here, $\text{MC}_{Motion}, \text{MC}_{Text}$ denote MotionCLIP's motion and text encoders, respectively, with $\mathbb{D}_g$ signifying the interaction description template related to the class of the sampled target object $\boldsymbol{O}_g$. Acknowledging that HOI predominantly occurs later in motion sequences, our semantic loss formulation weights later motion frames more heavily to accurately capture these interactions.

4 Experiment

4.1 Experimental Setup

Dataset-1: GIMO [88], which records motion sequences represented by full-body SMPL-X poses with $\approx 129K$ frames. It consists of 14 scenes with 3D point clouds, each scene is captured by a 3D LiDAR sensor, containing 10-20 objects with $\approx 500K$ vertices. For a fair comparison, we follow the official split to divide the dataset into training and testing sets according to the scenes.

Dataset-2: CIRCLE [4] comprises 10 hours of high-fidelity full-body motion sequences from 5 subjects across nine apartment scenes. Utilizing a Vicon system with 12 cameras at 120 FPS and the AI Habitat VR environment for virtual world simulation, CIRCLE achieves precise motion and scene capture. It offers an integrated apartment mesh, designating each room as an individual scene. The dataset encompasses motion sequences for 128 tasks, totaling over 7,000 sequences.

We also notice other related datasets [27, 28, 50], yet find limitations precluding their use (*e.g.,* jittering, sequence length, absence of human meshes).

Baselines. Our DiMoP3D is compared with four contemporary methods: DLow [82], SmoothDMP [72], BeLFusion [5], and BiFU [88]. DLow [82] applies a flow network, SmoothDMP [72] is VAE-based, and BeLFusion [5] is diffusion-based, which achieves SoTA performance in diverse HMP. These three, however, do not focus on scene-aware diverse HMP, setting BiFU [88], a deterministic scene-aware method, apart as an essential control for our analysis.

Metrics. To align with existing literature that evaluates human skeleton metrics, we utilize the SMPL model [45] to convert body parameters $\boldsymbol{X}_{1:L+\Delta L}$ into skeletons $\boldsymbol{J}_{1:L+\Delta L} = \{\boldsymbol{j}_1, \boldsymbol{j}_2, ..., \boldsymbol{j}_{L+\Delta L}\}$, where each $\boldsymbol{j}_i \in \mathbb{R}^{n_j \times 3}$ represents a skeleton with $n_j = 22$ joints, following [25].

DiMoP3D is evaluated for diversity, accuracy, and physical consistency in scene-aware predictions. We begin with the well-established pipeline in [82]: Prediction diversity is quantified using the Average Pairwise Distance **(APD)** by computing the L2 distance across predicted sequences. The Average Displacement Error **(ADE)** measures the reconstruction accuracy among the whole predicted sequence, while the Final Displacement Error **(FDE)** measures accuracy of the furthest frame, alongside their multimodal counterparts, **MMADE** and **MMFDE**, for diverse HMP scenarios.

To measure the physical consistency of the predicted motion within 3D scenes, an additional metric, the Average Cumulated Penetration Depth **(ACPD)** is introduced, following [78, 83]:

$$\text{ACPD}(\boldsymbol{J}) = \frac{1}{\Delta L} \sum_{l=L+1}^{L+\Delta L} \sum_{n=1}^{n_j} max\Big(-\text{SDF}(\boldsymbol{j}_l[n], \boldsymbol{S}),\ 0\Big), \tag{14}$$

where $\boldsymbol{j}[n]$ denotes the position of the n-th joint in the skeleton, and $\text{SDF}(\cdot, \boldsymbol{S})$ refers to the signed distance function [51] of the scene point cloud $\boldsymbol{S}$. For training details, please refer to Appendix A.

https://doi.org/10.52202/079017-1256

Table 1: Comparison of DiMoP3D with baselines on GIMO [88] and CIRCLE [4] datasets. The best outcomes are highlighted in bold. Given that BiFU [88] employs a deterministic prediction approach, diversity metrics such as APD, MMADE, and MMFDE are not applicable.

	Method	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	ACPD ↓
GIMO [88]	Dlow [82]	55.12	13.70	16.88	15.96	17.31	14.55
	SmoothDMP [72]	**68.80**	11.17	14.51	13.67	15.42	15.08
	BeLFusion [5]	38.04	9.69	11.19	11.28	12.02	13.73
	BiFU [88]	–	7.11	8.39	–	–	3.73
	DiMoP3D	48.30	**5.66**	**6.81**	**6.57**	**7.44**	**0.98**
CIRCLE [4]	Dlow [82]	49.37	11.70	14.46	13.49	14.95	12.52
	SmoothDMP [72]	**57.38**	9.75	12.04	11.31	13.57	13.92
	BeLFusion [5]	39.46	7.91	9.30	9.56	10.04	14.06
	BiFU [88]	–	5.80	6.99	–	–	2.11
	DiMoP3D	42.24	**5.09**	**6.12**	**5.95**	**6.48**	**0.87**

Table 2: Ablation of four main components in DiMoP3D over the sequences of the GIMO [88].

Ablation	APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓	ACPD↓
w/o InterestNet	52.63	6.17	7.48	7.20	8.09	1.00
w/o HOI-Estimator	46.97	5.95	7.27	6.72	7.84	1.53
w/o TrajectoryPlanner	**57.29**	6.39	**6.81**	7.28	7.45	3.29
w/o SemanticInspector	47.79	5.82	6.82	6.75	7.46	1.06
DiMoP3D	48.30	**5.66**	**6.81**	**6.57**	**7.44**	**0.98**

4.2 Main Results

Table 1 demonstrates DiMoP3D's superiority over the baseline methods across nearly all evaluation metrics on both datasets. The non-scene-aware methods (Dlow, SmoothDMP, BeLFusion) exhibit limited motion accuracy (ADE, FDE, MMADE, MMFDE) and physical scene consistency (ACPD), which we hypothesize is due to (1) lack of scene awareness, resulting in notable inconsistency in real-world applications, and (2) the absence of explicit motion goals, which hinders precise long-term (5-*sec*) motion forecasting. Despite their higher scores in diversity (APD), this is attributed to their erratic and unpredictable predictions, disregarding the scene context (detailed in Sec 4.4).

DiMoP3D's enhanced performance stems from three key factors: (1) Diversity. The stochastic conditional factor introduces diversity through multiple mechanisms: the intermodal interpreter sets broad motion objectives, the stochastic planner generates a variety of end-poses and trajectories, and the motion generator achieves diverse motion poses. This multi-faceted approach ensures a breadth of plausible actions are considered, enabling DiMoP3D to achieve a considerable APD score. (2) Accuracy. DiMoP3D outperforms every baseline in ADE, FDE, MMADE, and MMFDE for a large margin, even the deterministic BiFU. By estimating future human action based on a scene-motion intermodal analysis, DiMoP3D implicitly infers the subject's intent. This boosts the probability of accurately identifying the subject's genuine intent as the basis for prediction, thereby improving the prediction precision. The combination of accurate intermodal scene interpreting and stochastic planning ensures precise motion prediction for each sequence. (3) Physical consistency. Our motion generator employs a diffusion model, prompted by the predicted stochastic factors. It also ensures motion coherence through priors overwrite at each denoising step. This dual focus on deterministic constraints enables DiMoP3D to achieve superior motion-scene consistency.

This superior performance demonstrates the efficacy of our DiMoP3D in predicting diverse human motion in 3D scenes, as further evidenced by subsequent ablation studies and visualizations.

4.3 Ablation Studies

In Table 2, we dissect the impact of excluding four pivotal components from DiMoP3D. First, eliminating InterestNet markedly decreases performance across ADE, FDE, MMADE, and MMFDE (first row). This decline stems from the process of selecting the target object $\boldsymbol{O}_g$ from $\mathbb{O}$, which reverts to random sampling without scene-motion crossmodal analysis, impairing DiMoP3D's ability to deduce human intentions. Consequently, the accuracy of predicting real actions diminishes, highlighting the significance of integrating multimodal scene-motion analysis for scene-aware HMP.

https://doi.org/10.52202/079017-1256

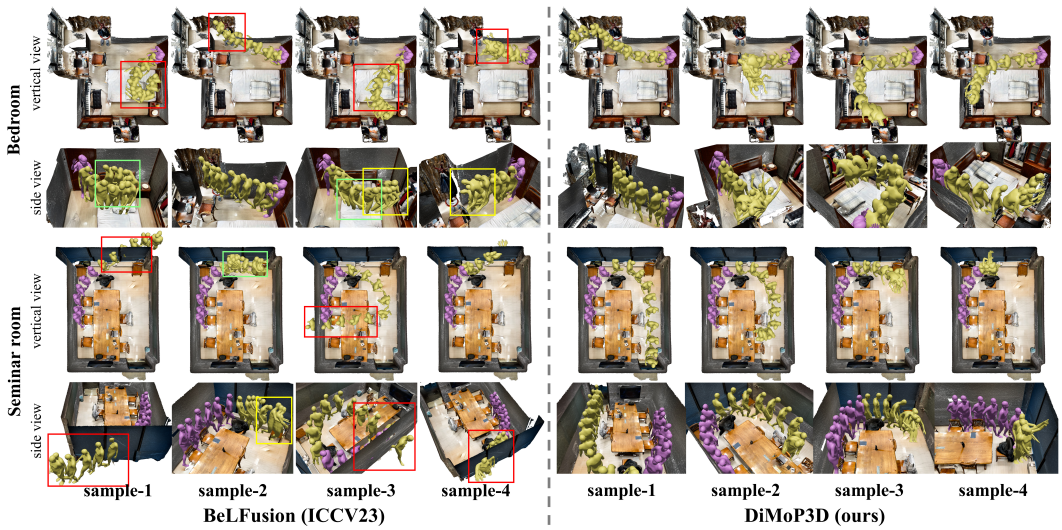

Figure 3: Visual comparisons between DiMoP3D and SoTA BeLFusion in bedroom and seminar room scenarios. BeLFusion's predictions, which rely solely on past human motion without considering 3D scene context, are shown on the left. In contrast, DiMoP3D, displayed on the right, incorporates interactive goals and designs obstacle-free trajectories for each sequence. Purple meshes depict observed motions, while yellow ones signify predicted future motions. For clarity, distortions in BeLFusion's predictions are marked: red boxes for object penetration, green boxes for motion incoherence, and yellow boxes for scene inconsistency.

Addressing the role of stochastic planner, its absence undermines the planning of actions, including the prediction of end-poses by the HOI-Estimator and trajectory planning via the A* TrajectoryPlanner. Without these two components, the motion generator struggles to predict end-poses or future trajectories with scene consistency. Notably, omitting the HOI-Estimator results in imprecise interactive end-poses, often causing the subject to intersect with the target object in later frames, as evidenced by increased ACPD and reduced FDE and MMADE (second row). Similarly, excluding the TrajectoryPlanner significantly elevates ACPD (third row), indicating frequent subject penetrations into the scene context while approaching the end-pose. These findings underscore the vital role of coordinated end-pose and trajectory prediction in predicting motion within 3D scenes effectively.

Finally, the SemanticInspector enhances the scene-motion alignment through natural language, with its omission resulting in higher ADE and MMADE. Please refer to Appendix C for further ablations.

4.4 Visualization

To delve deeper into DiMoP3D, in Figure 3, we showcase DiMoP3D's predictions across two scenarios, contrasting them with the SoTA baseline BeLFusion [5].

In bedroom scenario, the subject stands still behind the door. BeLFusion's predictions show notable issues with object and wall penetrations. Furthermore, sample-1 and 3 are marked by abrupt, illogical movements, and sample-3 and 4 display glaring scene inconsistencies: sample-3 has the subject sitting on the bare floor, and sample-4 involves the subject reaching for non-existent items. Conversely, DiMoP3D ensures physical consistency, directing each prediction towards a specific movement goal: opening a window, lying on the bed, accessing a cabinet, and sitting on a chair.

In the seminar room scene, the subject is moving forward. BeLFusion struggles again, with sample-1, 3, and 4 depicting the subject unrealistically exiting the room, and sample-2 showing an inconsistent motion of picking up a curtain. DiMoP3D, however, delivers high-fidelity predictions, depicting the subject walking through a door, grasping items, sitting, and pulling down curtains, respectively.

To emphasize DiMoP3D's predictive diversity, we further visualize various end-poses generated by the HOI-Estimator in Figure 4. Overall, DiMoP3D consistently delivers diverse, realistic, and physically-consistent motion predictions with clear objectives, benefiting from our conditional factor prediction schema which models human-object interactions and navigates obstacle-free trajectories.

https://doi.org/10.52202/079017-1256

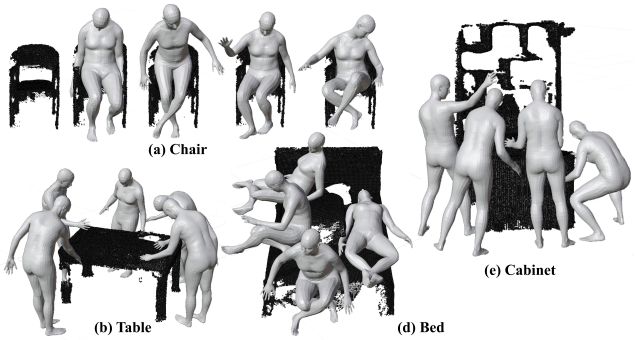

Figure 4: Visualizations of diverse predicted end-poses across five object point clouds. The HOI-Estimator can generate a variety of human-object interactive poses tailored to specific scenarios.

Table 3: Comparison of DiMoP3D with scene-aware motion synthesis methods.

Method	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	FID ↓	ACPD ↓
SAMP [28]	31.73	9.83	9.28	11.16	10.13	1.493	1.69
DN-Synt [68]	44.60	9.71	7.16	11.29	7.83	1.026	1.21
AffordMotion [70]	**52.54**	8.96	8.14	10.38	8.95	**0.687**	1.26
DiMoP3D	48.30	**5.66**	**6.81**	**6.57**	**7.44**	0.769	**0.98**

4.5 Compared with Motion Synthesis Methods

In this section, we compare our DiMoP3D with three scene-aware motion synthesis approaches on GIMO [88]: SAMP [28], DN-Synt [68], and AffordMotion [70]. SAMP and DN-Synt utilize VAE architectures, while AffordMotion employs a diffusion-based model.

To adapt these methods for our diverse HMP task, we initialize motion synthesis from the last observed frame $\boldsymbol{x}_L$ and encode the complete observed sequence $\boldsymbol{X}_{1:L}$ into a unified embedding for historical motion conditions using a 2-layer transformer encoder, similar to the embedding technique in [57]. Additionally, we introduce the FID metric, commonly used in motion synthesis, to evaluate the discrepancy between the distributions of generated and original dataset motions.

The results in Table 3 reveal an intriguing pattern: synthesis methods exhibit higher ADE than FDE. This occurs because, although these methods include explicit end-pose estimators yielding accurate final pose predictions, they struggle to condition on past motion. Consequently, they produce motion incoherence and significant prediction errors along the trajectory from the observation to the final pose. Notably, AffordMotion achieves the best APD and FID scores, which we attribute to its design that prioritizes fidelity over accuracy and allows a higher degree of freedom. Meanwhile, DiMoP3D also demonstrates competitive performance in these metrics. These findings underscore DiMoP3D's capability to harmonize the stochastic nature of human motion and the deterministic constraints from the scene and the past motion, achieving superior performance in diverse scene-aware HMP, and maintaining competitive diversity and fidelity even when compared to SoTA synthesis method.

5 Conclusion and Limitation

This work introduces a novel task of predicting diverse human motion in 3D scenes, along with a novel framework, DiMoP3D, to address it. By incorporating multimodal motion-scene analysis, DiMoP3D identifies areas or objects the subject is likely to interact with, enabling diverse, accurate, and physically consistent human motion prediction. Evaluated on the GIMO and CIRCLE datasets, DiMoP3D reduces ADE and FDE by nearly half compared to the state-of-the-art baseline BeLFusion, while maintaining high physical consistency. These results underscore the importance of scene awareness in diverse human motion prediction for real-world applications.

Despite its strong performance, DiMoP3D predicts motion in a fixed sequence length. When the actual sequence length differs, it either keeps the subject relatively static or truncates the sequence. Future work could explore variable-length motion prediction or end-to-end prediction, where the motion generator predicts sequence length and generates motion simultaneously.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China (2022YFC3602601), in part by the Natural Science Foundation of Jiangsu Province (BK20220939), and in part by the National Natural Science Foundation of China (62306141).

References

[1] E. Aksan, M. Kaufmann, P. Cao, and O. Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *3DV*, pages 565–574. IEEE, 2021.

[2] S. Aliakbarian, F. S. Saleh, M. Salzmann, L. Petersson, and S. Gould. A Stochastic Conditioning Scheme for Diverse Human Motion Prediction. In *CVPR*, pages 5223–5232, 2020.

[3] S. Aliakbarian, F. Saleh, L. Petersson, S. Gould, and M. Salzmann. Contextually Plausible and Diverse 3D Human Motion Prediction. In *ICCV*, pages 11333–11342, 2021.

[4] J. P. Araújo, J. Li, K. Vetrivel, R. Agarwal, J. Wu, D. Gopinath, A. W. Clegg, and K. Liu. Circle: Capture in rich contextual environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21211–21221, 2023.

[5] G. Barquero, S. Escalera, and C. Palmero. Belfusion: Latent diffusion for behavior-driven human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2317–2327, 2023.

[6] E. Barsoum, J. Kender, and Z. Liu. HP-GAN: Probabilistic 3D Human Motion Prediction via GAN. In *CVPR*, pages 1418–1427, 2018.

[7] D. E. Benrachou, S. Glaser, M. Elhenawy, and A. Rakotonirainy. Use of social interaction and intention to improve motion prediction within automated vehicle framework: A review. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):22807–22837, 2022.

[8] Z. Cao, H. Gao, K. Mangalam, Q.-Z. Cai, M. Vo, and J. Malik. Long-term human motion prediction with scene context. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 387–404. Springer, 2020.

[9] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

[10] P. Chang, J. Dang, J. Dai, and W. Sun. Real-time respiratory tumor motion prediction based on a temporal convolutional neural network: Prediction model development study. *Journal of Medical Internet Research*, 23(8):e27235, 2021.

[11] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to Detect Human-object Interactions. In *WACV*, pages 381–389. IEEE, 2018.

[12] L.-H. Chen, J. Zhang, Y. Li, Y. Pang, X. Xia, and T. Liu. Humanmac: Masked motion completion for human motion prediction. *arXiv preprint arXiv:2302.03665*, 2023.

[13] S. Chen, J. Fang, Q. Zhang, W. Liu, and X. Wang. Hierarchical aggregation for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15467–15476, 2021.

[14] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

[15] H.-K. Chiu, E. Adeli, B. Wang, D.-A. Huang, and J. C. Niebles. Action-Agnostic Human Pose Forecasting. *WACV*, pages 1423–1432, 2019.

[16] P. Cong, Z. W. Dou, Y. Ren, W. Yin, K. Cheng, Y. Sun, X. Long, X. Zhu, and Y. Ma. Laserhuman: Language-guided scene-aware human motion generation in free environment. *arXiv preprint arXiv:2403.13307*, 2024.

[17] E. Corona, A. Pumarola, G. Alenya, and F. Moreno-Noguer. Context-aware Human Motion Prediction. In *CVPR*, pages 6992–7001, 2020.

 https://doi.org/10.52202/079017-1256

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

[19] Q. Cui, H. Sun, and F. Yang. Learning Dynamic Relationships for 3D Human Motion Prediction. In *CVPR*, pages 6519–6527, 2020.

[20] Q. Cui, H. Sun, Y. Kong, X. Zhang, and Y. Li. Efficient human motion prediction using temporal convolutional generative adversarial network. *Information Sciences*, 545:427–447, 2021.

[21] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[22] L. Dang, Y. Nie, C. Long, Q. Zhang, and G. Li. MSR-GCN: Multi-Scale Residual Graph Convolution Networks for Human Motion Prediction. In *ICCV*, pages 11467–11476, 2021.

[23] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1201–1209, 2021.

[24] G. Gkioxari, R. Girshick, P. Dollár, and K. He. Detecting and Recognizing Human-object Interactions. In *CVPR*, pages 8359–8367, 2018.

[25] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022.

[26] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7:187–199, 2021.

[27] M. Hassan, V. Choutas, D. Tzionas, and M. J. Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019.

[28] M. Hassan, D. Ceylan, R. Villegas, J. Saito, J. Yang, Y. Zhou, and M. J. Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11374–11384, 2021.

[29] S. Huang, Z. Wang, P. Li, B. Jia, T. Liu, Y. Zhu, W. Liang, and S.-C. Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16750–16761, 2023.

[30] D. K. Jain, M. Zareapoor, R. Jain, A. Kathuria, and S. Bachhety. Gan-poser: an improvised bidirectional gan model for human motion prediction. *Neural Computing and Applications*, 32(18):14579–14591, 2020.

[31] L. Jiang, H. Zhao, S. Shi, S. Liu, C.-W. Fu, and J. Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, pages 4867–4876, 2020.

[32] R. Jiang, Z. Cai, Z. Wang, C. Yang, Z. Fan, Q. Chen, K. Tsubouchi, X. Song, and R. Shibasaki. Deepcrowd: A deep model for large-scale citywide crowd density and flow prediction. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):276–290, 2021.

[33] P. Karle, M. Geisslinger, J. Betz, and M. Lienkamp. Scenario understanding and motion prediction for autonomous vehiclesreview and comparison. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):16962–16982, 2022.

[34] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[35] M. Kolodiazhnyi, A. Vorontsova, A. Konushin, and D. Rukhovich. Oneformer3d: One transformer for unified point cloud segmentation. *arXiv preprint arXiv:2311.14405*, 2023.

[36] X. Kong, K. Wang, M. Hou, F. Xia, G. Karmakar, and J. Li. Exploring human mobility for multi-pattern passenger prediction: A graph learning framework. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):16148–16160, 2022.

[37] N. Kulkarni, D. Rempe, K. Genova, A. Kundu, J. Johnson, D. Fouhey, and L. Guibas. Nifty: Neural object interaction fields for guided human motion synthesis. *arXiv preprint arXiv:2307.07511*, 2023.

[38] J. N. Kundu, M. Gor, and R. V. Babu. BiHMP-GAN: Bidirectional 3D Human Motion Prediction GAN. In *AAAI*, volume 33, pages 8553–8560, 2019.

[39] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, and Q. Tian. Dynamic Multiscale Graph Neural Networks for 3D Skeleton Based Human Motion Prediction. In *CVPR*, pages 214–223, 2020.

[40] M. Li, S. Chen, Z. Zhang, L. Xie, Q. Tian, and Y. Zhang. Skeleton-Parted Graph Scattering Networks for 3D Human Motion Prediction. In *ECCV*, pages 18–36. Springer, 2022.

[41] S. Li, K. Wu, C. Zhang, and Y. Zhu. On the learning mechanisms in physical reasoning. *Advances in Neural Information Processing Systems*, 35:28252–28265, 2022.

[42] H. Lin, W. Zou, T. Li, S. J. Feigenberg, B.-K. K. Teo, and L. Dong. A super-learner model for tumor motion prediction and management in radiation therapy: development and feasibility evaluation. *Scientific reports*, 9(1):14868, 2019.

[43] Y. Liu, J. Zhang, L. Fang, Q. Jiang, and B. Zhou. Multimodal motion prediction with stacked transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7577–7586, 2021.

[44] Z. Liu, K. Lyu, S. Wu, H. Chen, Y. Hao, and S. Ji. Aggregated multi-gans for controlled 3d human motion prediction. In *AAAI*, volume 35, pages 2225–2232, 2021.

[45] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

[46] Z. Lou, Q. Cui, H. Wang, X. Tang, and H. Zhou. Multimodal sense-informed forecasting of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2144–2154, 2024.

[47] S. Lu, L.-H. Chen, A. Zeng, J. Lin, R. Zhang, L. Zhang, and H.-Y. Shum. Humantomato: Text-aligned whole-body motion generation. *arXiv preprint arXiv:2310.12978*, 2023.

[48] J. Mao, Y. Xue, M. Niu, H. Bai, J. Feng, X. Liang, H. Xu, and C. Xu. Voxel transformer for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3164–3173, 2021.

[49] W. Mao, M. Liu, M. Salzmann, and H. Li. Learning Trajectory Dependencies for Human Motion Prediction. In *ICCV*, pages 9489–9497, 2019.

[50] W. Mao, R. I. Hartley, M. Salzmann, et al. Contact-aware human motion forecasting. *Advances in Neural Information Processing Systems*, 35:7356–7367, 2022.

[51] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.

[52] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji. Rgbd salient object detection: A benchmark and algorithms. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pages 92–109. Springer, 2014.

[53] X. Peng, Y. Xie, Z. Wu, V. Jampani, D. Sun, and H. Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553*, 2023.

[54] A. Piergiovanni, A. Angelova, A. Toshev, and M. S. Ryoo. Adversarial Generative Grammars for Human Activity Prediction. In *ECCV*, pages 507–523. Springer, 2020.

[55] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[56] G. Qian, Y. Li, H. Peng, J. Mai, H. Hammoud, M. Elhoseiny, and B. Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Advances in Neural Information Processing Systems*, 35:23192–23204, 2022.

 https://doi.org/10.52202/079017-1256

[57] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[58] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[59] S. Saadatnejad, A. Rasekh, M. Mofayezi, Y. Medghalchi, S. Rajabzadeh, T. Mordan, and A. Alahi. A generic diffusion-based approach for 3d human pose prediction in the wild. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8246–8253. IEEE, 2023.

[60] T. Sofianos, A. Sampieri, L. Franco, and F. Galasso. Space-Time-Separable Graph Convolutional Network for Pose Forecasting. In *ICCV*, pages 11209–11218, 2021.

[61] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from rgbd images. *plan, activity, and intent recognition*, 64, 2011.

[62] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022.

[63] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[64] J. Tseng, R. Castellon, and K. Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023.

[65] T. Vu, K. Kim, T. M. Luu, T. Nguyen, and C. D. Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2708–2717, 2022.

[66] H. Wang, J. Dong, B. Cheng, and J. Feng. PVRED: A Position-Velocity Recurrent Encoder-Decoder for Human Motion Prediction. *IEEE Transactions on Image Processing*, 30:6096–6106, 2021.

[67] J. Wang, S. Yan, B. Dai, and D. Lin. Scene-aware generative network for human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12206–12215, 2021.

[68] J. Wang, Y. Rong, J. Liu, S. Yan, D. Lin, and B. Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20460–20469, 2022.

[69] Z. Wang, Y. Chen, T. Liu, Y. Zhu, W. Liang, and S. Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems*, 35:14959–14971, 2022.

[70] Z. Wang, Y. Chen, B. Jia, P. Li, J. Zhang, J. Zhang, T. Liu, Y. Zhu, W. Liang, and S. Huang. Move as you say, interact as you can: Language-guided human motion generation with scene affordance. *arXiv preprint arXiv:2403.18036*, 2024.

[71] D. Wei, H. Sun, B. Li, J. Lu, W. Li, X. Sun, and S. Hu. Human joint kinematics diffusion-refinement for stochastic motion prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6110–6118, 2023.

[72] M. S. Wei Mao, Miaomiao Liu. Generating Smooth Pose Sequences for Diverse Human Motion Prediction. In *ICCV*, pages 474–489, 2021.

[73] J. Won, D. Gopinath, and J. Hodgins. Physics-based character controllers using conditional vaes. *ACM Transactions on Graphics (TOG)*, 41(4):1–12, 2022.

[74] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[75] Z. Xiao, T. Wang, J. Wang, J. Cao, W. Zhang, B. Dai, D. Lin, and J. Pang. Unified human-scene interaction via prompted chain-of-contacts. *arXiv preprint arXiv:2309.07918*, 2023.

[76] C. Xu, R. T. Tan, Y. Tan, S. Chen, X. Wang, and Y. Wang. Auxiliary tasks benefit 3d skeleton-based human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9509–9520, 2023.

[77] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017.

[78] S. Xu, Z. Li, Y.-X. Wang, and L.-Y. Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14928–14940, 2023.

[79] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018.

[80] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19313–19322, 2022.

[81] Z. Yu, S. Kim, R. Mallipeddi, and M. Lee. Human intention understanding based on object affordance and action classification. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2015.

[82] Y. Yuan and K. Kitani. Dlow: Diversifying Latent Flows for Diverse Human Motion Prediction. In *ECCV*, pages 346–364, 2020.

[83] Y. Yuan, J. Song, U. Iqbal, A. Vahdat, and J. Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16010–16021, 2023.

[84] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5831–5840, 2018.

[85] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[86] X. Zhang, B. L. Bhatnagar, S. Starke, V. Guzov, and G. Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision*, pages 518–535. Springer, 2022.

[87] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021.

[88] Y. Zheng, Y. Yang, K. Mo, J. Li, T. Yu, Y. Liu, C. K. Liu, and L. J. Guibas. Gimo: Gaze-informed human motion prediction in context. In *European Conference on Computer Vision*, pages 676–694. Springer, 2022.

[89] C. Zhong, L. Hu, Z. Zhang, Y. Ye, and S. Xia. Spatio-temporal gating-adjacency gcn for human motion prediction. In *CVPR*, pages 6447–6456, 2022.

[90] W. Zhu, X. Ma, D. Ro, H. Ci, J. Zhang, J. Shi, F. Gao, Q. Tian, and Y. Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

 https://doi.org/10.52202/079017-1256

A Traning Details

Training DiMoP3D encompasses separate phases for InterestNet, HOI-Estimator, and the self-prompted motion generator. This section delineates the specific training methodologies for each component. All training is conducted on a single NVIDIA RTX3090 GPU, with the complete pipeline converging in ~ 8 hours.

Training InterestNet. We adopt the ScanNet [21] pretrained SoftGroup model as our encoder and segmenter to enhance performance, and a transformer decoder for the interest net to achieve cross-modal analysis. The original datasets [4, 88] lack annotations for the interest map $\boldsymbol{M}$; hence, we enrich them with such annotations for each motion clip. InterestNet is employed to elucidate the relationship between the observed motion and the scene, facilitating the prediction of human interests, including likely destinations and objects of interaction. To ensure the interest map accurately represents these aspects, we annotate the human interest map $\boldsymbol{M}$ based on three critical factors: the contact area between humans and objects ($\boldsymbol{M}_{cont}$), the proximity of humans to scene elements ($\boldsymbol{M}_{dist}$), and the spatial relationship between each object $\boldsymbol{O}_i$ and the target object $\boldsymbol{O}_g$ ($\boldsymbol{M}_{obj}$):

$$\boldsymbol{M}_{cont} = f_{\mathcal{N}(0,\sigma_1)}(\text{Dist3D}(\boldsymbol{S}, p_{cont})), \tag{15}$$

$$\boldsymbol{M}_{dist} = \frac{1}{\Delta L} \cdot \sum_{l=L+1}^{L+\Delta L} \frac{(l-L)^2}{(\Delta L)^2} \cdot f_{\mathcal{N}(0,\sigma_2)}(\text{Dist2D}(\boldsymbol{S}, \boldsymbol{t}_l)), \tag{16}$$

$$\boldsymbol{M}_{obj} = \sum_{\boldsymbol{O}_i \in \mathbb{O}} f_{\mathcal{N}(0,\sigma_3)}(\text{Dist2D}(\boldsymbol{O}_i, \boldsymbol{O}_g)) \cdot \boldsymbol{O}_i, \tag{17}$$

$$\boldsymbol{M} = \lambda_{cont}\boldsymbol{M}_{cont} + \lambda_{dist}\boldsymbol{M}_{dist} + \lambda_{obj}\boldsymbol{M}_{obj}. \tag{18}$$

Here, p_{cont} denotes the human-object contact position, $f_{\mathcal{N}(\mu,\sigma)}(\cdot)$ denotes the Gaussian function with mean μ and standard deviation σ. Smaller σ concentrates the interest map while larger σ distributes attention more broadly. We set $\sigma_1 = 0.3$ and $\sigma_2 = \sigma_3 = 1.0$ for balance. Dist3D$(\cdot)$ and Dist2D$(\cdot)$ are the 3D and X-Z 2D distance functions (Y-axis denotes height). Hyperparameters $\lambda_{cont} = 3, \lambda_{dist} = 10, \lambda_{obj} = 1$ are adjusted to maintain a balance among the factors.

Upon annotating the interest map $\boldsymbol{M}$, InterestNet is trained using a KL divergence loss to minimize the discrepancy between the distributions of the predicted interest map and the annotated $\boldsymbol{M}$:

$$\mathcal{L}_{\text{Interest}} = KL(\boldsymbol{M} \,||\, \text{InterestNet}(\mathcal{F}_s, \boldsymbol{X}_{1:L})). \tag{19}$$

Training HOI-Estimator. The HOI-Estimator employs an autoregressive conditional variational autoencoder (cVAE) [34] architecture, designed to predict a range of feasible end-poses. It predicts interactive end-poses $\hat{\boldsymbol{x}}_{end}$ based on the target object's point cloud $\boldsymbol{O}_g$. Initially, we centralize the object on the X-Z plane (assume the Y-axis denotes height) by normalizing its position to the origin:

$$\boldsymbol{O}_g^{Norm} = \boldsymbol{O}_g - \frac{1}{len(\boldsymbol{O}_g)} \sum_{p \in \boldsymbol{O}_g} p_{xz}, \tag{20}$$

where $len(\boldsymbol{O}_g)$ indicates the point count in $\boldsymbol{O}_g$ and p_{xz} is the X-Z coordinates of point p. The HOI-Estimator utilizes a conditional Variational Autoencoder (cVAE) [34] for encoding-decoding:

$$\mu, \sigma = \text{Encoder}(\boldsymbol{O}_g^{Norm}, \boldsymbol{X}_{1:L}), \tag{21}$$

$$\hat{\boldsymbol{x}}_{end} = \text{Decoder}(\boldsymbol{O}_g^{Norm};\ \mu, \sigma). \tag{22}$$

The HOI-Estimator is trained to minimize the L2 loss between the reconstructed interactive pose $\hat{\boldsymbol{x}}_{end}$ and the ground-truth $\boldsymbol{x}_{end}$:

$$\mathcal{L}_{\text{HOI}} = ||\hat{\boldsymbol{x}}_{end} - \boldsymbol{x}_{end}||_2^2. \tag{23}$$

Training motion generator. The motion generator is pre-trained on the HumanML3D dataset [25]. We predict the clean sample $\bar{\boldsymbol{X}}^0$ directly instead of predicting noise following [63, 64, 83], consisting of the basic diffusion loss:

$$\mathcal{L}_{\text{base}} = ||\bar{\boldsymbol{X}}^0 - \boldsymbol{X}||_2^2, \tag{24}$$

https://doi.org/10.52202/079017-1256

Table 4: Description template of semantic inspector among 18 objects in the ScanNet dataset [21].

[OBJ]	[ACT]			
ground	walks on	stands on		
cabinet	opens	searches	reaches hands to	picks things from
bed	sits on	lies on		
chair	sits on			
sofa	sits on			
table	sets	wipes	reaches hands to	picks things from
door	opens	closes	reaches hands to	passes
window	opens	closes	reaches hands to	
bookshelf	arranges	leans on	reaches hands to	picks things from
picture	hangs	takes down	reaches hands to	
counter	wipes	leans on	reaches hands to	
desk	wipes	organizes	reaches hands to	picks things from
curtain	opens	closes	reaches hands to	
refrigerator	opens	closes	reaches hands to	picks things from
shower curtain	opens	closes	reaches hands to	
toilet	sits on	flushes	reaches hands to	
sink	washes		reaches hands to	picks things from
bathtub	flushes		reaches hands to	takes a shower in
other furniture	uses	checks	reaches hands to	interacts with

supplemented by the semantic alignment loss (detailed in Section 3.4 of the main paper), forming the total training loss for motion generator:

$$\mathcal{L}_{\text{diff}} = \mathcal{L}_{\text{base}} + \mathcal{L}_{\text{sem}}. \tag{25}$$

Description Template of Semantic Inspector. The semantic inspector incorporates language description to supervise the predicted motion to be consistent with target objects. To enable this supervision, we design a language description template for each class of objects in the format of:

$$\text{The person walks forward then } [ACT] \text{ the } [OBJ], \tag{26}$$

where $[ACT]$ and $[OBJ]$ are placeholders for the action and object, respectively. There are a total of 18 classes of objects (excluding wall and floor) in the pre-trained ScanNet dataset [21], and we design corresponding $[ACT]$ for each class, detailed in Table 4.

B Modified A* Trajectory Planner

To facilitate diverse trajectory prediction, we enhance the conventional A* trajectory planner, allowing for the iterative generation of valid paths while penalizing previously traversed positions.

Initially, we generate the scene height map $\boldsymbol{S}_H$ as described in Alg. 1. To streamline height analysis and A* pathfinding, we voxelized the scene point cloud $\boldsymbol{S}$ along the X and Z axes (with the Y-axis representing height) into a grid of 0.02-meter resolution. We then identify the maximum height within each grid cell, omitting ceilings and tall cabinets, which, despite their height, do not impede movement and would otherwise be inaccurately marked as obstacles.

Leveraging the generated height map $\boldsymbol{S}_H$, we outline our modified dynamic A* trajectory planner in Alg. 2. Our method transcends traditional one-hot encoding for marking obstacles by enabling navigation over lower barriers through a continuous cost function derived from $\boldsymbol{S}_H$. We adopt a power function to model the cost, with cell height acting as the exponent. To ensure smoother trajectories and reduce sudden changes, an L2 penalty on angular velocity is integrated into the cost framework. Additionally, to prevent the selection of repetitive paths, we increase the cost of cells once traversed. The modified A* algorithm then iteratively generates paths until the cumulative cost exceeds a predefined threshold or the maximum path count is attained, as in Figure 5.

C Additional Ablation Studies

Scene feature for baselines. To assess the influence of scene features on baseline methods, we conduct supplemental experiments integrating scene features into these methods. Despite the baseline models mentioned in the main paper [5, 72, 82] not inherently accommodating scene features, we

https://doi.org/10.52202/079017-1256

Algorithm 1 get_heightmap(S):

```
# S: Scene point cloud with shape (Np, 3), where Np is the point count

invalid = -10000, grid = 0.02, h_th = 1.2
Sg = VoxelGridXZ(S, grid)        # X,Z coordinates are gridded, while Y is not
H = ones((Sg.maxx-Sg.minx)/grid, (Sg.maxz - Sg.minz)/grid) * invalid
For p in Sg:
  If p.y < h_th:        # Excluding the ceiling and high cabinets
    x, z = Sg.grid(p.x, p.z)
    H[x, z] = max(Hmap[x, z], p.y)
return H
```

Algorithm 2 dynamic_astar(X, d, H):

```
# X: Observed motion trajectory with shape (L, 2), where L is the input length
# d: Destination position with shape (2)
# H: Scene height map with shape (lenX, lenZ)

def cost_function(lastcost, p, path):
  return lastcost + C[p] + angular_vel(path)

invalid = -10000, grid = 0.02, h_th = 1.2, noise = 1.0
costbase = 1000, costth = 1000, costinc = 10
H[H==invalid] = h_th
C = costbase ** H
paths = []
For npath in range(MAX_PATHS):
  path, cost = astar(C + randlike(C)*noise, X[-1], d, cost_function)
  If cost > mincost + costth: break
  C[path] += costinc
  paths.append(bessel_smoothing(path))
return paths
```

Table 5: Results of appending scene features for the baseline methods on GIMO [88].

Method	scene	APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓	ACPD↓
Dlow [82]		55.12	13.70	16.88	15.96	17.31	14.55
SmoothDMP [72]		**68.80**	11.17	14.51	13.67	15.42	15.08
BeLFusion [5]		38.04	9.69	11.19	11.28	12.02	13.73
Dlow [82]	✓	–	–	–	–	–	–
SmoothDMP [72]	✓	66.36	11.02	14.49	13.43	15.37	14.71
BeLFusion [5]	✓	36.89	9.55	11.04	11.12	11.76	13.52
DiMoP3D	✓	48.30	**5.66**	**6.81**	**6.57**	**7.44**	**0.98**

augmented their input by appending the scene feature $\mathcal{F}_s$ along the temporal dimension for [5, 82], and along the joint dimension for [72], to explore potential performance.

The results in Table 5 reveal that Dlow failed to convergence, likely attributed to its recursive architecture, wherein the concatenated scene features diverge significantly from the distribution of motion features, making the network difficult to learn. Both SmoothDMP and BeLFusion exhibit marginal improvements, suggesting that the direct integration of scene features into the motion generator yields limited effectiveness. Conversely, these results underscore the efficacy of DiMoP3D's strategy of predicting the conditional factor, highlighting its superiority in harmonizing the scene and the observed motion.

Ablation on the scene segmenter. To evaluate DiMoP3D's robustness across various point cloud instance segmentation methods, we performed additional experiments with different segmentators. Higher-quality segmenters yield more precise object delineations, enabling the selection of more accurate targets and reducing object boundary violations. Table 6 illustrates that DiMoP3D main-

Table 6: Results of DiMoP3D with various scene segmenters. "mAP50" denotes the mean average precision at 50 IoU threshold for each segmenter on the ScanNetv2 dataset [21]. Higher "mAP50" represents better segmenter performance.

Segmentator	mAP50	APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓	ACPD↓
PointGroup [31]	63.6	46.22	5.70	6.93	6.63	7.56	1.05
HAIS [13]	69.9	46.97	5.67	6.87	6.61	7.50	1.02
SoftGroup [65]	76.1	**48.30**	**5.66**	**6.81**	**6.57**	**7.44**	**0.98**

Table 7: Results of DiMoP3D with various trajectory planners. Trajectory planner has no effect on FDE and MMFDE as they are end-pose errors, so we omit them in this table.

Planner	APD↑	ADE↓	MMADE↓	ACPD↓
Naive A*	39.26	5.93	6.82	0.98
ours w/o continuous cost	45.59	5.77	6.72	0.98
ours w/o angular penalty	**50.17**	5.91	6.88	**0.97**
ours	48.30	**5.66**	**6.57**	0.98

tains stable performance even when the quality of the segmentation method declines, underscoring its robustness to variations in segmentation.

Ablation on the trajectory planner. To validate the significance of our enhanced A* trajectory planner, we conducted an ablation study. Results presented in Table 7 show that the naive A* method underperforms due to its deterministic nature. Excluding the continuous cost model results in a binary scene representation, disregarding traversable lower obstacles and thus reducing path diversity. Additionally, omitting the angular velocity penalty leads to paths with abrupt turns, detracting from prediction accuracy. These outcomes significant the critical role of our modified A* trajectory planner in achieving nuanced and reliable path predictions.

D More Visualizations

Instance segmentation and interest map. Exploring our multimodal scene parser, we present visualizations of 3D scene instance segmentation alongside the corresponding interest maps $\boldsymbol{M}$ for three distinct samples. Figure 6 (upper) showcases the segmented instances within the scene, while Figure 6 (lower) evidences our multimodal InterestNet's capacity to deduce potential human intentions.

In the bedroom setting, with the person initially stationary and then beginning to move forward, the future action remains ambiguous. Consequently, almost all objects are highlighted as potential targets in $\boldsymbol{M}$, except for the door situated behind the individual. In the living room scene, as the person navigates the narrow gap between a chair and a table, the sofa, an additional chair, and the table emerge as points of interest. Conversely, the proximate chair and a distant door garner lesser attention, aligning with the observed motion pattern. A similar observation is noted in a laboratory environment.

These results underscore our multimodal scene parser's adeptness at inferring potential human intentions, and the generated interest map $\boldsymbol{M}$ can accurately reflect anticipated human interests and interactions.

Multiple samples with single target. To further explore DiMoP3D's capacity for predicting diverse motions toward a deterministic target object, additional visualizations are presented.

The left panel of Figure 7, showcases predicted sequences where the person navigates different paths to reach the destination, performing varied actions such as looking down and reaching toward the target. The right panel illustrates sequences of the person adopting different positions for lying and sitting on the bed. These results affirm DiMoP3D's adeptness at generating varied and coherent human motions aimed at a specific target object, further ensuring diversity.

Comparison with synthesis methods. To further explore the performance of synthesis methods in diverse scene-aware HMP tasks, we present visual comparisons between DiMoP3D and the SoTA synthesis method AffordMotion in Figure 8. AffordMotion's predicted sequences tend to remain relatively static at the onset of prediction, which we attribute to its method of smoothly synthesiz-

https://doi.org/10.52202/079017-1256

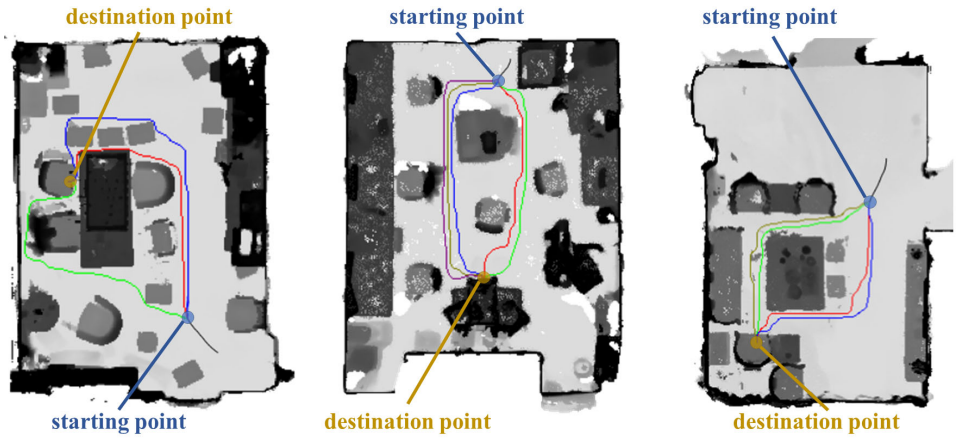

Figure 5: Visualization samples of the modified A* trajectory planner. Black lines denote the observed trajectory, while colored lines represent the generated paths.

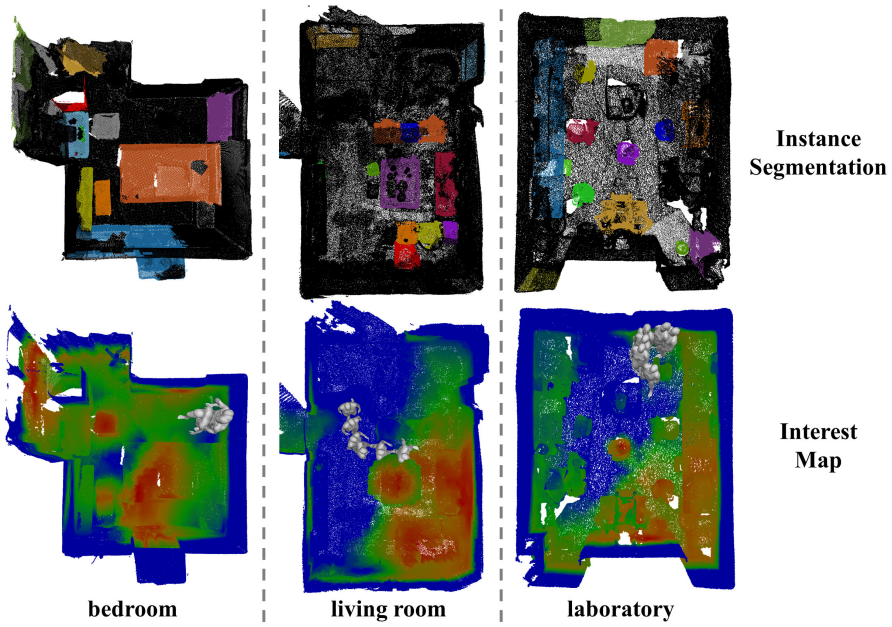

Figure 6: Visualization of 3D scene instance segmentation (upper) and the corresponding interest map (lower). Red points in the interest map denote higher interest, while blue points denote lower interest. Leveraging the insight provided by the predicted interest map enables the exclusion of improbable or illogical targets, thereby enhancing the reliability and scene congruency of predictions.

ing human motion irrespective of prior motion states, resulting in significant errors along the paths. Particularly in samples 3 and 4, AffordMotion shows notable motion incoherence, characterized by abrupt changes at the transition between prediction and observation. In contrast, DiMoP3D considers past motion, predicting coherent and plausible sequences, thereby demonstrating its superior performance in diverse scene-aware HMP tasks. While motion synthesis methods show promise in handling diverse scene-aware HMP, they face the critical challenge of integrating deterministic cues from past motions effectively.

E Potential Broader Impacts

The proposed DiMoP3D introduces a novel diverse scene-aware motion prediction framework, which may involve the following broader impacts:

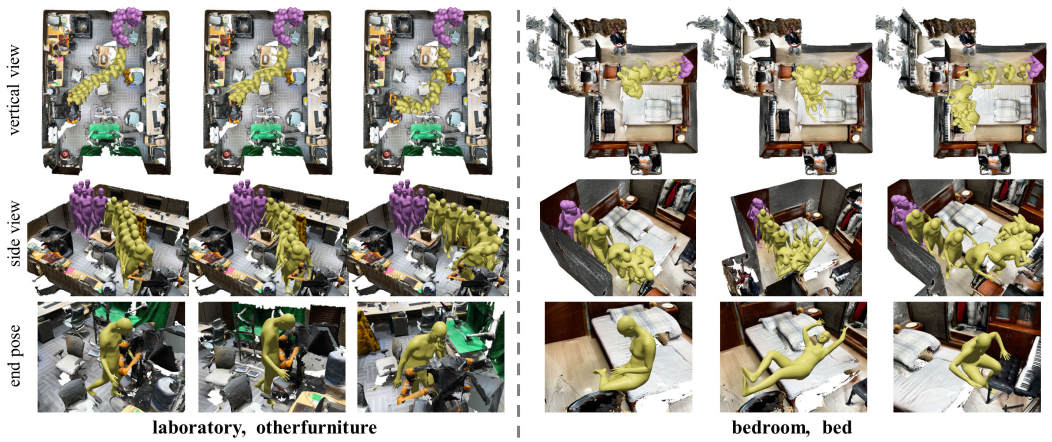

Figure 7: Visualization of multiple samples with fixed target object. DiMoP3D is able to predict motions with diverse trajectories and actions (or end poses) toward a deterministic target object, while maintaining each motion sequence to be consistent with the observation and the scene.

Figure 8: Visualization comparisons between DiMoP3D and SoTA synthesis method AffordMotion. The results from AffordMotion demonstrate significant motion incoherence.

- **Enhanced Safety in Robotics and Automation.** DiMoP3D can improve the interaction between humans and robots. By predicting human motions accurately, robots can avoid collisions and unsafe interactions, making environments safer for both humans and machines.
- **Improved VR and Gaming Experiences.** In VR and video games, this framework can lead to more realistic and responsive interactions with virtual characters and environments. By understanding and predicting how a human might move within a scene, VR systems can offer more immersive and natural experiences, enhancing user engagement and satisfaction.
- **Advancements in Assistive Technologies.** For people with disabilities or the elderly, assistive technologies equipped with human motion prediction can anticipate needs or actions (like falling or reaching for an object) and provide timely assistance or interventions, thereby enhancing independence and quality of life.
- **Applications in Autonomous Vehicles.** Integrating this framework into autonomous vehicle systems can improve pedestrian safety and traffic management. By predicting human movements, autonomous vehicles can better navigate complex urban environments where interactions with pedestrians are frequent and unpredictable.

https://doi.org/10.52202/079017-1256

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 3 tell the method, Section 4 demonstrates results, proving the effectiveness of the method.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Secon 5 analyzes the efficiency limitation of the proposed DiMoP3D.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Section 3 explains the assumptions, theorems, and formulas in detail.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix A tells the training details and the supplemental material provides our source code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/079017-1256

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Appendix A tells the training details and the supplemental material provides our source code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Appendix A tells the training details and the supplemental material provides our source code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Appendix A tells the training details and the supplemental material provides our source code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix A, we conduct all the experiments on single NVIDIA RTX3090 GPU within 8 hours.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and ensured that our research conforms to it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The proposed DiMoP3D may have broader impacts on robotics, autonomous vehicles, video games, assistive technologies, *etc.*, as detailed in Appendix E.

    Guidelines:

https://doi.org/10.52202/079017-1256

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There is no high risk for misuse in the data or models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The licenses and terms of use are explicitly mentioned and properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: There are no new assets introduced in the paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: There is no crowdsourcing experiments and research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: There is no research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

 https://doi.org/10.52202/079017-1256

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.