DREAMoR: Diffusion-based REconstruction And Motion prioR

Sihan Ren Jiashen Du Jingfeng Yang Yidi Zhang
UC Berkeley
{sihan_ren, jason_du, yangjingfeng0705, yidi.lily}@berkeley.edu

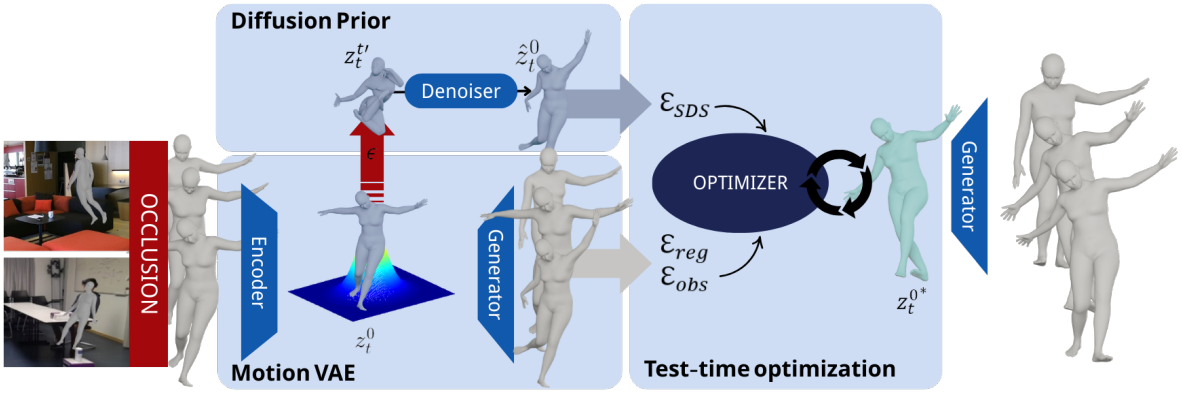

Figure 1. Overview of DREAMoR. Given a partially occluded or noisy motion sequence, we first encode motion using a MotionVAE to obtain a latent transition sequence z_t^0. Each latent transition is then used in two parallel branches: (1) applies noise to latent motion and a latent diffusion denoising to produce a distribution-aligned latent $\hat{z}_t^0$; (2) the current latent is decoded into a predicted motion frame using a transition generator. Both outputs contribute to our energy function, which includes SDS loss, observation loss, and regularization. By minimizing this energy through optimization, we refine z_t and decode the optimized latents into the final high-quality motion sequence.

Abstract

*Monocular human motion capture systems often suffer from noisy or incomplete predictions due to occlusions, poor visibility, or ambiguous poses. To address these limitations, we propose **DREAMoR**: a diffusion-based motion prior framework for reconstructing physically plausible human motion from corrupted sequences. Our method first learns a latent space of motion transitions using a MotionVAE trained on clean data. We then train a diffusion model in this latent space to capture the distribution of realistic motion transitions, conditioned on previous frames. At inference time, we use multistep DDIM-style denoising and score distillation to optimize the latent sequence, ensuring that the resulting motion both aligns with the noisy input and adheres to the learned motion prior. Experiments on AMASS show that DREAMoR outperforms prior methods in recovering occluded joints and produces smoother, more realistic motion. Our ablation studies further highlight the effectiveness of latent diffusion priors for motion refinement.*

1. Introduction

Markerless human motion capture has made significant progress in recent years. Modern methods[2, 6, 13, 14, 17, 25] can even estimate full 3D human mesh and joint trajectories from monocular RGB videos or images. Yet, these monocular reconstruction pipelines remain fragile in practice. When parts of the body are occluded or poorly estimated, the resulting 3D motion can exhibit severe noise, jitter, or even physically implausible behaviors.

To overcome these limitations, many works [31, 32] have explored the use of *human motion priors* to regularize or refine motion predictions. These priors aim to capture how humans typically move, using models ranging from Gaussian processes [38] to VAEs [18, 47] and transformers. They are often used to repair corrupted sequences or fill in missing joints.

At the same time, *diffusion models* have demonstrated remarkable success in modeling complex, high-dimensional distributions. Recent motion generation works [7, 32, 33, 43, 43, 46] have shown that diffusion models can produce highly realistic and diverse human motions from text, audio, or sparse inputs. Inspired by these advances, we investigate

whether diffusion can also serve as a *motion prior* for refining noisy motion observations—particularly those affected by occlusion or sensor noise.

In this work, we propose **DREAMoR**, a framework that reconstructs clean and physically plausible human motion from occluded sequences using a diffusion-based prior. Our key insight is to learn a latent space of *motion transitions*, capturing how motion evolves over time, and train a conditional latent diffusion model directly in that latent space. At test time, we encode the corrupted motion into this latent space, add noise and apply denoising, and then optimize the latent motions via score distillation sampling (SDS), observation matching losses, and motion regularization. We further demonstrate that applying *multi-step* DDIM [35] denoising during SDS yields better results than single-step estimation, improving realism and convergence.

We evaluate DREAMoR on a subset of the AMASS dataset [23] with simulated occlusion. Our method outperforms strong baselines in reconstructing missing joint positions, and our ablation studies confirm that the learned diffusion prior significantly improves both realism and consistency.

Our main contributions are:

- We introduce **DREAMoR**, a novel motion reconstruction framework that combines latent-space diffusion with motion prior optimization.
- We incorporate **multi-step DDIM denoising** into the SDS optimization, which leads to more stable and accurate reconstructions compared to single-step prediction.
- Our experiments demonstrate that DREAMoR produces more accurate and realistic motion reconstructions than prior methods, especially in the presence of occlusion.

2. Related Work

2.1. Motion Priors

Human motion models aim to learn the distribution of plausible human movements and have long been applied to tasks such as tracking, prediction, and synthesis. Prior work including mixtures-of-Gaussians [11], Gaussian processes [38], pose embeddings [26, 27, 36, 39], VAEs [18, 47], 2D convolutional models [15], and normalizing flows [8]. While these approaches capture motion statistics in seen datasets, they often fail to generalize to out-of-distribution behaviors. Physics-based methods [4, 5, 12, 21, 22, 30, 34, 41, 44] improve realism by enforcing physical laws via simulation.

HuMoR [31] proposed a powerful alternative: instead of generating an entire motion sequence from scratch, it rolls out motion transitions step-by-step, optimizing intermediate representations using motion priors. This rollout-style paradigm helps maintain consistency across time and allows fine-grained control guided by partial or noisy observations.

2.2. Diffusion Models for Motion

Diffusion models have recently shown strong performance in human motion generation, where the goal is to produce natural sequences from high-level conditions such as text, audio, or sparse pose inputs. These models are trained to reverse a gradual noising process, and are known to capture expressive, high-dimensional data distributions. Notable examples include MDM [32], MotionCLIP [37], and UniMuMo [42], which generate compelling long-term motion from text or music. Given the powerful ability of diffusion models to capture the distributions of human motion, numerous studies have extended their applications to enhance generation efficiency and improve performance in specific tasks, such as generation of human poses guided by text and motion [33, 45, 48] and human motion reconstruction [7, 43, 46].

RoHM [46] uses a diffusion model to reconstruct human motion from noisy and occluded input. Instead of test-time optimization, it learns directly from synthetic noisy data, enabling faster inference. In addition, recent work[1, 3] demonstrated that processing motion data in a compact latent space allows diffusion models to generate more realistic and coherent motion with improved efficiency.

2.3. Score Distillation Sampling

A recent trend in generative modeling involves using diffusion models not just for generation, but as powerful priors to guide reconstruction in under-constrained or corrupted settings. This includes tasks like 3D shape completion [28, 40], scene reconstruction from sparse views [19], and consistency across generated images [16]. The central idea behind these methods is Score Distillation Sampling (SDS), where a pretrained diffusion model provides gradient-based feedback to steer inputs toward more realistic outputs.

In motion reconstruction, this is especially relevant: real-world capture setups, especially monocular and markerless systems, often suffer from occlusion, missing joints, or jitter. By training diffusion models on clean, high-quality datasets, we gain access to a rich prior that reflects how humans truly move. This prior can then be used to guide the refinement of corrupted sequences, even when the observation is limited or noisy.

3. Method

We propose **DREAMoR**, a framework that learns a generative prior over human motion transitions and leverages it to refine occluded or incomplete motion sequences. The pipeline consists of three key stages: 1) A Motion VAE that learns a latent space z_t capturing the transition between consecutive motion frames (x_{t-1}, x_t). 2) A latent diffusion model trained on the $\{z_t\}$ space to model the distribution

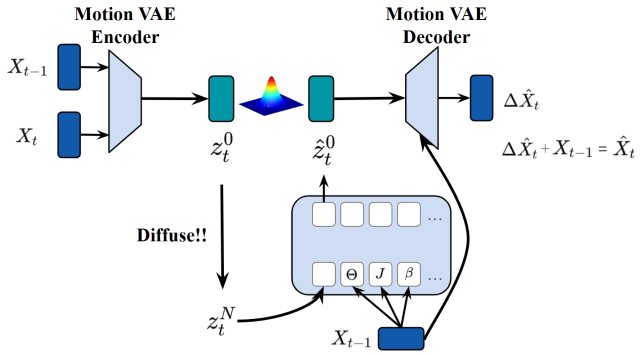

Figure 2. The pipeline of DREAMoR. Each motion pair (x_{t-1}, x_t) is encoded into a latent transition z_t via the MotionVAE. The diffusion model operating in latent space uses a transformer-based denoiser, which takes as input the noisy latent $z_t^{t'}$, a set of tokenized conditioning vectors from x_{t-1}, and a time embedding for step t'. The input sequence is processed with self-attention, and the first token (representing z_t) is used to predict the noise $\hat{\epsilon}_t$. At test time, we apply *multi-step DDIM denoising* from a random timestep t' to $t = 0$, obtaining prior-aligned latents $\hat{z}_t^0$. These are combined with generator outputs and observation constraints in an energy-based optimization over the entire latent sequence.

$p(z_t \mid x_{t-1})$, providing strong priors for motion refinement at test time. 3) Optimization over the sequence of latent motions to make the motion decoded from those latent motions more accurate and realistic.

Given an initial corrupted sequence with L frames, we encode it into the latent space, then optimize the full set of latent motions $\{z_{0:L}\}$, finally decode them to roll out a natural and better motion sequence guided by this prior.

3.1. Problem Setup

We represent each frame of human motion using SMPL parameters and derived quantities. Specifically, the motion at t frame x_t is composed of:

$$x_t = [r_t, \dot{r}_t, \Phi_t, \dot{\Phi}_t, \Theta_t, J_t, \dot{J}_t]$$

Where $r_t \in \mathbb{R}^3$ is root translation, $\Phi_t \in \mathbb{R}^9$ is root orientation, $\Theta_t \in \mathbb{R}^{21\times 9}$ is body joint rotation, $J_t \in \mathbb{R}^{66}$ are 3D joints positions, and $\dot{r}_t, \dot{\Phi}_t, \dot{J}_t$ are velocities of the corresponding parameters.

We use rotation matrices to represent joint rotations and root orientation. Compared to axis-angle or quaternion formats, rotation matrices allow composition via matrix multiplication, which is especially useful when modeling transition dynamics across frames.

In addition to the kinematic variables, we also include a contact vector $c_t \in [0, 1]^9$ that denotes the probability of whether specific body parts (hips, legs, feet, hands, toes) are in contact with the ground. This vector is used during optimization to enforce physical plausibility. The contact vector will be part of the output, and it will not be presented in the input data.

3.2. Motion VAE

We begin by pre-training a variational autoencoder to model the dynamics of human motion transitions. Given a motion pair (x_{t-1}, x_t), the encoder network outputs the parameters of a Gaussian distribution:

$$z_t \sim q_\phi(z_t \mid x_t, x_{t-1}) = \mathcal{N}(\mu_\phi(x_t, x_{t-1}), \sigma_\phi(x_t, x_{t-1}))$$

To simplify, we denote this procedure as:

$$z_t = E_\phi(x_t, x_{t-1})$$

A latent sample z_t is then combined with the previous frame x_{t-1} in the decoder to reconstruct the current motion frame:

$$\hat{x}_t = x_{t-1} + G_\theta(z_t, x_{t-1})$$

Here, G_θ is a learned residual function that predicts the delta motion in state space, which improves stability and prediction accuracy over direct generation.

The training objective optimizes a variational lower bound augmented with physically motivated regularizers:

$$\mathcal{L} = \lambda_{\text{rec}} \cdot \mathcal{L}_{\text{rec}} + \lambda_{\text{KL}} \cdot \mathcal{L}_{\text{KL}} + \lambda_{\text{reg}} \cdot \mathcal{L}_{\text{reg}}$$

Reconstruction loss: We measure the deviation between the predicted and ground-truth motion states using an ℓ_2 norm:

$$\mathcal{L}_{\text{rec}} = \|x_t - \hat{x}_t\|_2^2$$

KL divergence: The KL loss regularizes the approximate posterior towards the standard normal distribution:

$$\mathcal{L}_{\text{KL}} = D_{KL}(q_\phi(z_t|x_t, x_{t-1})\|\mathcal{N}(0, I))$$

Physical regularizers: To ensure plausibility and consistency with the SMPL model, we use predicted parameters to run SMPL inference $[\hat{J}_t^{\text{SMPL}}, \hat{V}_t] = M_{\text{SMPL}}(\hat{r}_t, \hat{\Phi}_t, \hat{\Theta}_t, \beta)$, where M_{SMPL} is SMPL model [20]. Then, we introduce the following additional losses: A joint position loss $\mathcal{L}_{\text{joint}} = \|J_t^{\text{SMPL}} - \hat{J}_t^{\text{SMPL}}\|_2^2$ enforces consistency between predicted and reconstructed SMPL joints, while a mesh vertex loss $\mathcal{L}_{\text{vtx}} = \|V_t - \hat{V}_t\|_2^2$ supervises surface-level precision. To align different prediction branches, we introduce a joint consistency loss $\mathcal{L}_{\text{consist}} = \|\hat{J}_t - \hat{J}_t^{\text{SMPL}}\|_2^2$ between directly regressed joints and SMPL-inferred ones. Finally, a contact-aware velocity penalty $\mathcal{L}_{\text{vel}} = \sum_j \hat{c}_t^j \cdot \|\hat{v}_t^j\|_2^2$ suppresses foot sliding by discouraging high velocities at predicted contact joints.

These terms are collectively grouped into:

$$\mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{joint}} + \mathcal{L}_{\text{vtx}} + \mathcal{L}_{\text{consist}} + \mathcal{L}_{\text{vel}}$$

3.3. Latent Diffusion Prior

Once we have trained the Motion VAE and obtained a compact latent space for motion transitions, we train a diffusion model directly in this space. The goal is to model the distribution of plausible latent transitions z_t, conditioned on the previous motion frame x_{t-1}. This provides a powerful generative prior that can later guide reconstruction via optimization.

Given a latent transition $z_t^0 = z_t = E_\phi(x_t, x_{t-1})$, we simulate a forward noising process by sampling a timestep $t' \sim \mathcal{U}[1, T]$, and applying:

$$z_t^{t'} = \sqrt{\bar{\alpha}_{t'}} \cdot z_t^0 + \sqrt{1 - \bar{\alpha}_{t'}} \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

where $\bar{\alpha}_{t'}$ denotes the cumulative product of the noise schedule [10]. The diffusion model D_θ is trained to predict the noise $\hat{\epsilon}$ given the noisy latent and conditioning frame:

$$\hat{\epsilon} = D_\theta(z_t^{t'} \mid x_{t-1}, t')$$

Then, we train this diffusion model with a simple loss:

$$\mathcal{L}_{\text{diff}} = \|\epsilon - \hat{\epsilon}\|_2^2$$

Transformer-Based Denoising Architecture Following BUDDI [24], we implement D_θ as a transformer encoder that operates on a sequence of tokens. The input consists of: A **latent token** representing the noisy latent $z_{t'}^t$; A set of **conditioning tokens** derived from x_{t-1}.

We tokenize $x_{t-1} = [r, \dot{r}, \Phi, \dot{\Phi}, \Theta, J, \dot{J}]$ using separate MLP-based tokenizers. Each tokenizer projects its input to a unified token dimension d_{token}: $\tau_{\text{r}} = T_r([r, \dot{r}]), \tau_\Phi = T_\Phi([\Phi, \dot{\Phi}]) \in \mathbb{R}^{d_{\text{token}}}$, etc. The conditioning sequence is constructed as:

$$\tau(x_{t-1}) = [\tau_r, \tau_\Phi, \tau_\Theta, \tau_J, \tau_j, \dots]$$

The latent vector is also projected into the same token dimension: $\tau_z = T_{\text{latent}}(z_{t'}^t)$. We prepend τ_z to the conditioning tokens to form the transformer input: $[\tau_z, \tau(x_{t-1})]$

To encode the diffusion timestep t', we add a learned time embedding $\gamma(t')$ to each token prior to attention.

Classifier-Free Guidance for Conditional Denoising To improve the model's controllability and avoid over-reliance on conditioning inputs, we adopt **Classifier-Free Guidance (CFG)** during both training and inference time [9].

During training, we randomly drop the conditioning tokens with a fixed probability p_{drop}. Specifically, for each training sample, with probability p_{drop}, we replace the conditioning sequence $\tau(x_{t-1})$ with a learned *null token* τ_{null}

At inference time, we evaluate the model twice:

$$\hat{\epsilon}_{\text{cond}} = D_\theta(z_{t'}^t \mid x_{t-1}, t') \tag{1}$$

$$\hat{\epsilon}_{\text{null}} = D_\theta(z_{t'}^t \mid \tau_{\text{null}}, t') \tag{2}$$

The final guided prediction is computed by linear interpolation:

$$\hat{\epsilon} = \hat{\epsilon}_{\text{CFG}} = \hat{\epsilon}_{\text{null}} + \omega \cdot (\hat{\epsilon}_{\text{cond}} - \hat{\epsilon}_{\text{null}}) \tag{3}$$

where ω is the *guidance scale*, controlling the strength of conditional adherence.

Prediction and DDIM Denoising After processing through the transformer, we extract the output corresponding to the latent token and map it back to the latent space, as the predicted noise $\hat{\epsilon}$. During inference or optimization, we perform DDIM-style [10] inversion to obtain the denoised latent:

$$\hat{z}_t^0 = \frac{1}{\sqrt{\bar{\alpha}_{t'}}} \left(z_t^{t'} - \sqrt{1 - \bar{\alpha}_{t'}} \cdot \hat{\epsilon} \right)$$

This yields a clean latent $\hat{z}_t^0$ that reflects the model's best estimate of a plausible transition given context.

3.4. Run-Time Optimization

At test time, our goal is to reconstruct a clean and physically plausible motion sequence from a noisy or incomplete observation $x_{0:L}$, such as a motion capture sequence with occlusions or jitter.

Latent Rollout and Prior Estimation Given the corrupted sequence $x_{0:L}$, we first use the pretrained encoder E to obtain latent motions $z_{0:L}$: $z_t = E(x_{t-1}, x_t)$.

To align each z_t with the diffusion prior, we perform K-step DDIM-style denoising. For each latent z_t in the sequence, we randomly sample a diffusion timestep $t' \in [1, T]$ as the starting noise level. We then construct a denoising trajectory of K decreasing timesteps: $t_K = t', t_0 = 0$, where $\{t_k\}_{k=0}^K$ is a predefined sampling schedule. We first perturb the clean latent z_t to the noisy version at step t', denoted $z_t^{t'}$, as: $z_t^{t'} = \sqrt{\bar{\alpha}_{t'}} \cdot z_t + \sqrt{1 - \bar{\alpha}_{t'}} \cdot \epsilon$. Then, for each step $k = K, \dots, 1$, we apply DDIM denoising using the predicted noise $\hat{\epsilon}_{t_k}$ via classifier-free guidance (CFG). Letting $\bar{\alpha}_{t_k}$ denote the cumulative noise schedule, we estimate:

$$\hat{z}_t^0 = \frac{1}{\sqrt{\bar{\alpha}_{t_k}}} \left(z_t^{t_k} - \sqrt{1 - \bar{\alpha}_{t_k}} \cdot \hat{\epsilon}_{t_k} \right),$$

$$z_t^{t_{k-1}} = \sqrt{\bar{\alpha}_{t_{k-1}}} \cdot \hat{z}_t^0 + \sqrt{1 - \bar{\alpha}_{t_{k-1}}} \cdot \hat{\epsilon}_{t_k}.$$

The final denoised latent after K steps is denoted as $\hat{z}_t^0$, which serves as the aligned version of z_t under the learned diffusion prior. We have $\hat{z}_{0:L}$ now.

We recursively reconstruct motion frames using the generator G: $\hat{x}_t = \hat{x}_{t-1} + G(z_t, \hat{x}_{t-1})$. This generates the sequence $\hat{x}_{0:L}$ based entirely on the latent motions.

Optimization Objective Now review what we have: $x_{0:L}$ the observation motions, $z_{0:L}$ latent motions, $\hat{z}_{0:L}$ latent motions that follow prior, and $\hat{x}_{0:L}$ the real motions rolled out from latent motions. We now use the latent motions $z_{0:L}$ as **optimization parameters**, try to minimize the following energy:

$$\min_{z_{0:L},\beta,g} \quad \mathcal{E}_{\text{SDS}} + \mathcal{E}_{\text{obs}} + \mathcal{E}_{\text{reg}}$$

- Score Distillation Loss: Encourages latent motions to stay close to diffusion-prior estimates.

$$\mathcal{E}_{\text{SDS}} = \sum_{t=0}^{L} \|z_t - \hat{z}_t^0\|_2^2$$

- Observation Loss: Encourages the decoded motions to look like observations.

$$\mathcal{E}_{\text{obs}} = \sum_{t=0}^{L} \sum_{j=1}^{J} \lambda_{\text{data}} \|\hat{p}_t^j - y_t^j\|^2$$

  Where $\hat{p}_t^j$ is the j-th joint position inferred from $\hat{x}_t$.
- Regularization Loss: We adopt four regularizers inspired by prior work (e.g., HuMoR [31]): First, the skeleton consistency term ensures that joint positions decoded from latent codes match those inferred by SMPL, while preserving stable bone lengths over time:

$$\mathcal{E}_{\text{skel}} = \sum_{t=1}^{L} \left(\lambda_c \sum_j \|p_t^j - \hat{p}_t^j\|^2 + \lambda_b \sum_i (l_t^i - l_{t-1}^i)^2 \right).$$

  To prevent foot sliding and ensure plausible contact with the ground, we include a ground contact consistency term that penalizes both velocity at contact points and floating artifacts:

$$\mathcal{E}_{\text{env}} = \sum_{t=1}^{L} \sum_j \lambda_{\text{cv}} c_t^j \|p_t^j - p_{t-1}^j\|^2 + \lambda_{\text{ch}} c_t^j \cdot \max(|p_{z,t}^j| - \delta, 0).$$

  We also constrain body shape and global translation through a shape regularizer and a ground constraint:

$$\mathcal{E}_{\text{shape}} = \lambda_{\text{shape}} \|\beta\|^2, \quad \mathcal{E}_{\text{gnd}} = \lambda_{\text{gnd}} \|g - g_{\text{init}}\|^2,$$

  Which respectively prevent unnatural shape deformation and enforce consistency in global body position. Lastly, a smoothness term enforces temporal continuity by penalizing abrupt changes in joint positions:

$$\mathcal{E}_{\text{smooth}} = \sum_{t=1}^{L} \sum_j \|p_t^j - p_{t-1}^j\|^2.$$

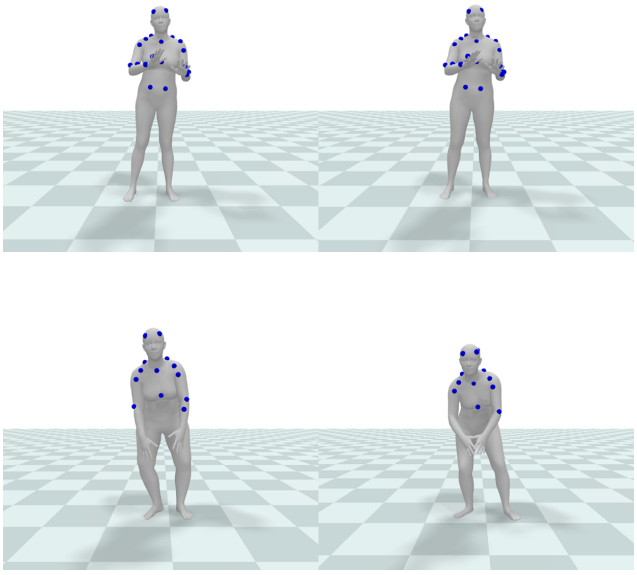

Figure 3. Qualitative result of DREAMoR(right) compared with groundtruth(left). DREAMoR captures active body movements while maintaining steady posture on easy motion under occluded conditions.

The total regularization loss is the weighted sum of the above terms:

$$\mathcal{E}_{\text{reg}} = \mathcal{E}_{\text{skel}} + \mathcal{E}_{\text{env}} + \mathcal{E}_{\text{shape}} + \mathcal{E}_{\text{gnd}} + \mathcal{E}_{\text{smooth}}.$$

Finally, we use optimized $z_{0:L}^*$ to roll out a real motion sequence, all the occluded motions are recovered and match the observed data while ensuring physical plausibility.

4. Experiments

We evaluate DREAMoR on its capability as a prior to estimate and refine motion from partial 3D observations. We recommend watching the qualitative evaluation videos on our website to compare the performance of our method to others.

4.1. Dataset

Our experiments utilize the AMASS dataset[23], a large-scale collection of motion capture (MoCap) data standardized on the SMPL body model. The dataset comprises diverse motions, ranging from daily activities to dynamic and expressive movements such as dancing, sports, and complex interactions. We ran our experiments on part of the AMASS dataset (ACCAD, CMU, DanceDB, EyesJapanDataset, MPI_mosh, and SOMA) with a total of 3282 motion sequences and a total length of 1205.45 minutes. We subsample the dataset at 30 Hz for our experiments, con-

Method	Input	Positional Error			Joints	Mesh			Ground Pen	
		Vis	Occ	All	Legs	Vtx	Contact	Accel	Freq	Dist
VPoser-t	Occ Keypoints	**0.72**	28.67	13.36	26.10	11.86	/	2.89	22.54%	14.29
HuMoR	Occ Keypoints	1.56	24.03	11.94	19.09	11.42	0.88	2.72	11.26%	1.63
RoHM	Occ Keypoints	1.23	9.70	5.91	12.92	7.07	**0.96**	**2.00**	**5.73%**	0.62
Ours(DREAMoR)	Occ Keypoints	1.68	**8.19**	**4.11**	**9.98**	**4.39**	0.95	2.59	6.25%	**0.58**

Table 1. Motion and shape estimation from 3D observations: partially occluded keypoints. *Positional Error (cm)* is reported w.r.t. the input modality. Acceleration is in m/s^2 and penetration distance in *cm*.

sistent with standard practices in previous literature such as HuMoR[31].

4.2. Evaluation Metrics

The metrics adopted follow standard conventions established in prior work:

Error Metrics. 3DPositional errors are measured on joints, keypoints, or mesh vertices (Vtx) and compute global mean per-point position error unless otherwise specified. We report positional errors for all (All), occluded (Occ), and visible (Vis) observations separately. Finally, we report the binary classification accuracy of the 9 person-ground contacts (Contact) predicted by methods.

Plausibility Metrics. We use additional metrics to measure qualitative motion characteristics that joint errors cannot capture. Smoothness is evaluated by mean per-joint accelerations (Accel) [13]. Another important indicator of plausibility is ground penetration [29]. We use the true ground plane to compute the frequency (Freq) of foot-floor penetrations: the fraction of frames for both the left and right toe joints that penetrate more than a threshold. We measure frequency at 0, 3, 6, 9, 12, and 15 cm thresholds and report the mean. We also report mean penetration distance (Dist), where non-penetrating frames contribute a distance of 0 to make values comparable across differing frequencies.

4.3. Estimation and Refinement from 3D Observations

We conduct experiments by masking the lower-body joints of AMASS data[23], thereby removing all positional information for the lower body. Under this challenging setting, DREAMoR must regenerate the lower-body motion solely conditioned on the upper-body motion. Experimental results indicate DREAMoR demonstrates significant capabilities:

- When the upper body performs complex actions such as dancing, the regenerated lower-body movements exhibit realistic and expressive patterns rather than remaining static. This demonstrates that DREAMoR successfully learns latent correlations between upper and lower-body motions from the diffusion-based generative prior.
- Conversely, when upper-body actions are relatively simple (e.g., waving), the lower-body motions generated by DREAMoR appropriately remain stable and minimally active, closely mirroring realistic human behavior.

These findings highlight DREAMoR's capacity to robustly estimate plausible lower-body motions through its generative diffusion process, maintaining both dynamism and consistency with observed upper-body motion.

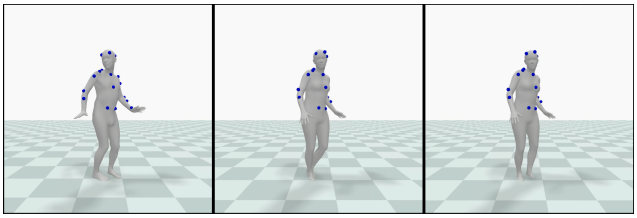

Figure 4. Ablation study of denoising steps. Groundtruth(Left), denoise 1 step(Middle), and denoise 10 steps(Right) are demonstrated at the same time stamp.

4.4. Ablation study

We conduct two ablation experiments to validate key components of DREAMoR's framework.

4.4.1. Denoising steps

From Method part 3.4, we apply denoising using the predicted noise $\hat{\epsilon}_{t_k}$, where we did for each step $k = K, \ldots, 1$. We conducted ablation experiments on the denoising steps and tested the final results for denoising 1 step compared with denoising 10 steps. The results are shown in Figure 4. While both use the same underlying model, Denoising 10 steps consistently produces more accurate and smoother motions. In particular, lower-body generations benefit from iterative denoising, yielding more plausible dynamics under partial observations. From the figure 4, on the same time stamp, the result from denoising 1 step shows that jitteriness has occurred because the left leg jittered to overlap with the right leg, whereas the denoising 10 steps result fol-

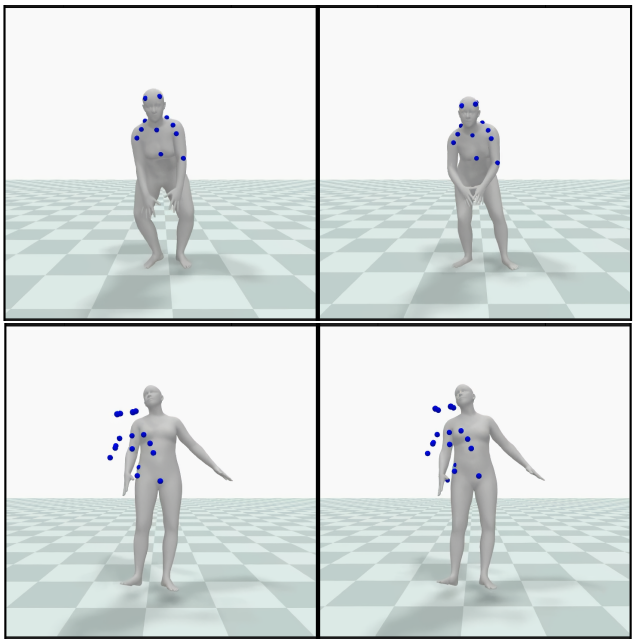

Figure 5. Ablation study of motion optimization. Ground truth (Top-left), denoise 10 steps w/ opt.(Top-right), denoise 10 steps w/o opt.(Bottom-left), and MotionVAE w/o opt.(Bottom-right) are demonstrated at the same time stamp.

lows the ground truth and no overlapping occurred.

Denoising step	**Occ** Positional Error	**Legs** Joints
1	9.38	11.44
10	8.19↓	9.98↓

Table 2. Ablation on denoising steps. Denoising 10 steps shows better performance on related metrics.

4.4.2. Motion optimization from motion prior

We also evaluate the impact of our third-stage optimization by removing it entirely. Instead of performing the optimization, we directly encode the observed motion sequence $x_{0:L}$ to obtain the latent motion sequence $z_{0:L}$. We considered two variants: The first one is to directly roll out the motion using the latent sequence as predicted by MotionVAE, another one is to apply noise to the $z_{0:L}$, followed by denoising steps to obtain a prior-conforming latent sequence $\hat{z}_{0:L}$ and then rolled out. The results of ground truth, w/ optimization, w/o optimization variants 1 and 2 are shown in Figure 5. Without this refinement, motion quality degrades significantly, often failing to match observed upper-body motions or resulting in implausible lower-body artifacts. From Figure 5, there's a significant pose error without optimization, and the result is pretty much unusable.

Method	**Vis** Positional Error	**Vtx** Mesh
w/ opt.	1.68↓	4.11↓
w/o opt. diff.	16.96	18.73
w/o opt. vae.	17.56	19.19

Table 3. Ablation on optimization step. Utilizing our optimization step yields a significant performance increase on related metrics.

5. Conclusion and Future Work

In this work, we introduced **DREAMoR**, a diffusion-based framework for reconstructing realistic and physically plausible human motion from occluded input sequences. Our approach combines a MotionVAE encoder–decoder with a latent-space diffusion prior, trained to model plausible motion transitions. At inference time, we apply multi-step DDIM-based denoising and score distillation to iteratively refine corrupted motion into clean reconstructions. Through experiments on AMASS with simulated occlusion, we show that DREAMoR achieves improvements over existing motion priors.

While DREAMoR demonstrates promising performance, several avenues remain open for further research. First, our current model conditions on the previous frame during both decoding and denoising. A promising direction is to explore *unconditional* decoding from latent variables alone, potentially improving generalization. Second, we currently condition only on a single previous frame; incorporating *longer temporal context* (e.g., $x_{t-2}, x_{t-3}, \dots$) could enable the model to capture more complex motion dynamics. Third, our occlusion and noise settings are synthetic. Integrating *real-world observations*, such as RGB video, depth, or sparse 2D keypoints, would bring DREAMoR closer to practical deployment. We also plan to evaluate the system on custom-captured test sequences with real occlusions and motion ambiguity to validate robustness in the wild.

References

[1] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 2

[2] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *CVPR*, 2021. 1

[3] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. In *ECCV*, pages 390–408, 2025. 2

[4] Erik Gärtner, Mykhaylo Andriluka, Erwin Coumans, and Cristian Sminchisescu. Differentiable dynamics for articulated 3d human motion reconstruction. In *CVPR*, 2022. 2

[5] Erik Gärtner, Mykhaylo Andriluka, Hongyi Xu, and Cristian Sminchisescu. Trajectory optimization for physics-based reconstruction of 3d human pose from monocular video. In *CVPR*, 2022. 2

[6] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 1

[7] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

[8] Gustav Henter, Simon Alexanderson, and Jonas Beskow. MoGlow: Probabilistic and controllable motion synthesis using normalising flows. *ACM TOG*, 39(4):1–14, 2020. 2

[9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 4

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 4

[11] Nicholas Howe, Michael Leventon, and William Freeman. Bayesian reconstruction of 3D human motion from single-camera video. In *NeurIPS*, 2000. 2

[12] Buzhen Huang, Liang Pan, Yuan Yang, Jingyi Ju, and Yangang Wang. Neural mocon: Neural motion control for physically plausible human motion capture. In *CVPR*, 2022. 2

[13] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 6

[14] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, 2019. 1

[15] Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and Otmar Hilliges. Convolutional autoencoders for human motion infilling. In *2020 International Conference on 3D Vision (3DV)*, 2020. 2

[16] Subin Kim, Kyungmin Lee, June Suk Choi, Jongheon Jeong, Kihyuk Sohn, and Jinwoo Shin. Collaborative score distillation for consistent visual synthesis, 2023. 2

[17] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 1

[18] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel van de Panne. Character controllers using motion vaes. *ACM TOG*, 39(4):1–12, 2020. 1, 2

[19] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 2

[20] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 3

[21] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3d human motion estimation via motion compression and refinement. In *ACCV*, 2020. 2

[22] Zhengyi Luo, Shun Iwase, Ye Yuan, and Kris Kitani. Embodied scene-aware human pose estimation. In *NeurIPS*, 2022. 2

[23] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019. 2, 5, 6

[24] Lea Muller, Vickie Ye, Georgios Pavlakos, Michael J. Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3D social interaction from images. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 4

[25] Hyeongjin Nam, Daniel Sungho Jung, Yeonguk Oh, and Kyoung Mu Lee. Cyclic test-time adaptation on monocular video for 3d human mesh reconstruction. In *ICCV*, 2023. 1

[26] Dirk Ormoneit, Hedvig Sidenbladh, Michael Black, and Trevor Hastie. Learning and tracking cyclic human motion. In *NeurIPS*, 2000. 2

[27] Vladimir Pavlovic, James Rehg, and John MacCormick. Learning switching linear models of human motion. In *NeurIPS*, 2001. 2

[28] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 2

[29] Davis Rempe, Leonidas J. Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 6

[30] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *ECCV*, 2020. 2

[31] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 5, 6

[32] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 1, 2

[33] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2

[34] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM TOG*, 39(6):1–16, 2020. 2

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 2

[36] Graham Taylor, Geoffrey Hinton, and Sam Roweis. Modeling human motion using binary latent variables. In *NeurIPS*, 2007. 2

[37] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 358–374. Springer, 2022. 2

[38] Raquel Urtasun, David Fleet, and Pascal Fua. 3D people tracking with gaussian process dynamical models. In *CVPR*, 2006. 1, 2

[39] Raquel Urtasun, David Fleet, and Pascal Fua. Temporal motion models for monocular and multiview 3d human body tracking. *CVIU*, 104(2), 2006. 2

[40] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *arXiv preprint arXiv:2212.00774*, 2022. 2

[41] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *ICCV*, 2021. 2

[42] Han Yang, Kun Su, Yutong Zhang, Jiaben Chen, Kaizhi Qian, Gaowen Liu, and Chuang Gan. Unimumo: Unified text, music and motion generation, 2024. 2

[43] Brent Yi, Vickie Ye, Maya Zheng, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. *arXiv preprint arXiv:2410.03665*, 2024. 1, 2

[44] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. Simpoe: Simulated character control for 3d human pose estimation. In *CVPR*, 2021. 2

[45] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatio-temporal motion generation and editing. *NeurIPS*, 2023. 2

[46] Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo. Rohm: Robust human motion reconstruction via diffusion. In *CVPR*, 2024. 1, 2

[47] Yan Zhang and Siyu Tang. The wanderings of odysseus in 3d scenes. In *CVPR*, 2022. 1, 2

[48] Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast, high-quality motion generation. *arXiv preprint arXiv:2312.02256*, 2023. 2