A Unified Diffusion Framework for Scene-aware Human Motion Estimation from Sparse Signals

Supplementary Material

1. Model details

We first supplement our pipeline's omitted implementation details, including the architecture of VAE-based motion prior and periodic autoencoder, the pre-training of our motion prior and periodic autoencoder, and the hyperparameter settings.

1.1. VAE-based motion prior

Figure 1 shows the detailed structure of the VAE-based motion prior. Our VAE-based motion prior consists of a conditional encoder $\mathcal{E}$ and decoder $\mathcal{D}$.

Given a sequence of sparse tracking signals $\mathbf{p}^{1:N}$ and paired full-body motions $\mathbf{x}^{1:N}$, the conditional encoder outputs a latent code,

$$z = \mathcal{E}(\mathbf{x}^{1:N}|\mathbf{p}^{1:N}). \tag{1}$$

The conditional decoder $\mathcal{D}$ learns to reconstruct the full-body motions by given latent code z and sparse tracking signals $\mathbf{p}^{1:N}$,

$$\hat{\mathbf{x}}^{1:N} = \mathcal{D}(z|\mathbf{p}^{1:N}). \tag{2}$$

The conditional encoder $\mathcal{E}$ is discarded during inference, and we only use the conditional decoder $\mathcal{D}$.

Training objectives. The training objectives of the VAE-based motion prior is

$$\mathcal{L}_{\text{VAE}} = \lambda_{\text{KL}} \cdot \mathcal{L}_{\text{KL}} + \lambda_{\text{recon}} \cdot \mathcal{L}_{\text{recon}} + \lambda_{\text{geometric}} \cdot \mathcal{L}_{\text{geometric}}. \tag{3}$$

The KL divergence $\mathcal{L}_{\text{KL}}$ minimizes the distribution distance between the learned conditional distribution $p_{\mathcal{E}}(z|\mathbf{x}^{1:N}, \mathbf{p}^{1:N})$ and the standard Gaussian distribution $q(z) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

The reconstruction loss forces the model to learn an informative latent z and be able to recover full-body motions from such latents,

$$\mathcal{L}_{\text{recon}}(\hat{\mathbf{x}}^{1:N}, \mathbf{x}^{1:N}) = \left\|\hat{\mathbf{x}}^{1:N} - \mathbf{x}^{1:N}\right\|^2. \tag{4}$$

The geometric loss $\mathcal{L}_{\text{geometric}}$ is the same as we mentioned in the text; it regularizes the generated motion to lie on the motion manifold.

We empirically set $\lambda_{\text{KL}} = 0.002$, $\lambda_{\text{recon}} = 1.0$, $\lambda_{\text{geometric}} = 0.5$ during training.

1.2. Periodic autoencoder

The full pipeline of the periodic autoencoder, shown in Figure 2, consists of an encoder and decoder, whereas in the text we only discuss the encoder part.

Starting with encoded feature maps $\mathbf{f}^{1:N}$ in the temporal domain, we reconstruct the original tracking signals $\tilde{\mathbf{p}}^{1:N}$ by a 1D deconvolution,

$$\tilde{\mathbf{p}}^{1:N} = \text{DeConv}(\mathbf{f}^{1:N}). \tag{5}$$

The entire PAE is pre-trained using reconstruction loss,

$$\mathcal{L}_{\text{PAE}} = \left\|\mathbf{p}^{1:N} - \tilde{\mathbf{p}}^{1:N}\right\|. \tag{6}$$

During inference, we kept only the encoder part to extract the temporal periodic feature maps and the related phase features.

2. Implementation details

The overall inputs to our model consist of encoded scene feature $\mathbf{E}_{\mathcal{S}} \in \mathbb{R}^n$, sparse tracking signals $\mathbf{p}^{1:N} \in \mathbb{R}^{N\times c}$, the extracted periodic motion features $\mathbf{f}^{1:N} \in \mathbb{R}^{N\times h}$. We choose $n = 256$, $c = (6+3)\times 3\times 2 = 54$ following previous works, and set the number of latent periodic channels h to 6. We set the input sequence length N to 120.

We train our VAE-based motion prior and conditional denoiser with a batch size of 64 and use AdamW for tuning parameters. The learning rate is fixed to 0.001 for both models. The feature dimension d_{model} of our models are set to 256, and the stacked transformer layers of the VAE-based motion prior and conditional denoiser are 9 and 8, respectively.

To keep the scale of the guidance score at the same level, we set the scaling factor $\lambda_{\text{penetration}}$ for scene-penetration loss $\ell_{\text{penetration}}$ to 0.1, and the scaling factor λ_{phase} for phase-matching loss ℓ_{phase} to 0.01. We observe larger order of magnitude of scaling factors will result in performance degradation and severe jittering of generated motions.

3. Extra qualitative results

We show the generated motions of our method against others on the GIMO dataset in Figure 6. We highlight the implausible motions in rectangle marks, it is clear that our method learns the correct human-scene interactions and avoids scene penetration as much as possible.

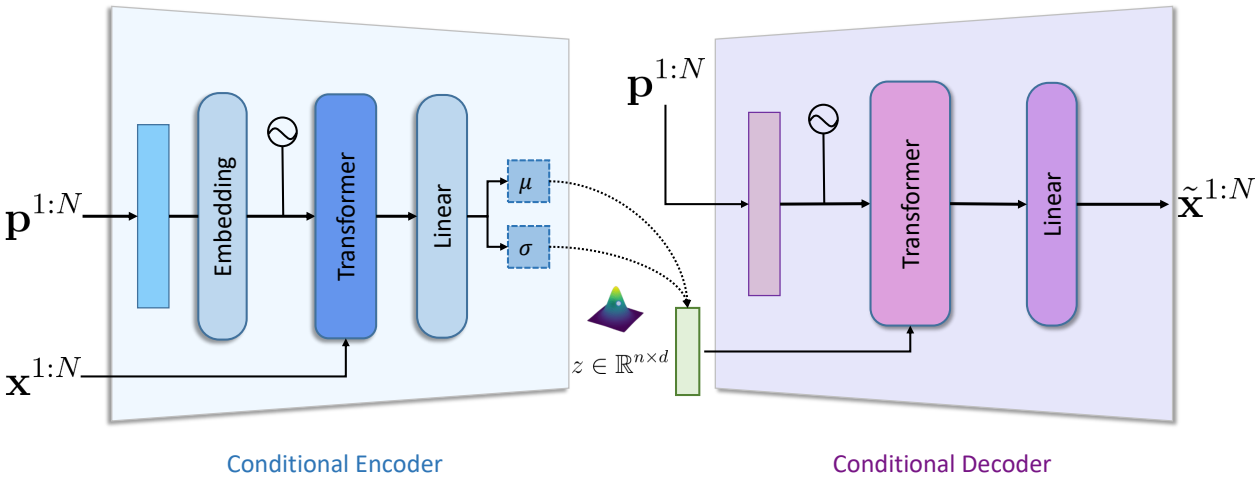

Figure 1. The structure of our VAE-based motion prior, consists of encoder $\mathcal{E}$ and decoder $\mathcal{D}$.

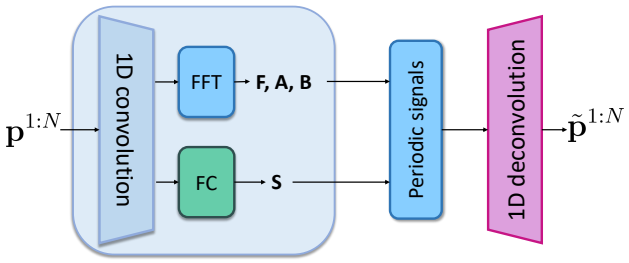

Figure 2. The structure of the periodic autoencoder.

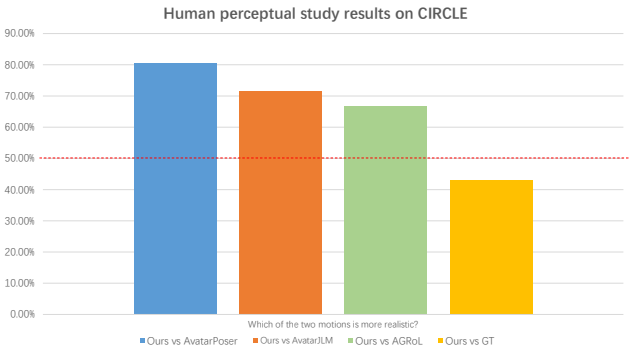

Figure 3. Human perceptual study results on the CIRCLE dataset.

Figure 4. Human perceptual study results on the GIMO dataset.

Figure 5. The failure cases of our motion generation pipeline.

Failure cases and analysis. We also show the failure cases of our motion generation pipeline in Figure 5. While focusing on generating realistic lower body motions, our method failed to faithfully capture fine-grained hand-object interactions, such as picking up clothes or wiping the blackboard. Incorporating more sophisticated full-body physical constraints may resolve the failure cases and be considered in our future work.

4. Extra evaluation of scene modality

In this section, we evaluate the effect of the scene modality on the task of reconstructing full-body motion from *head motion* only. The sparser inputs make the reconstruction task even more difficult. We compare our scene-conditioned diffusion backbone with a recent method EgoEgo which

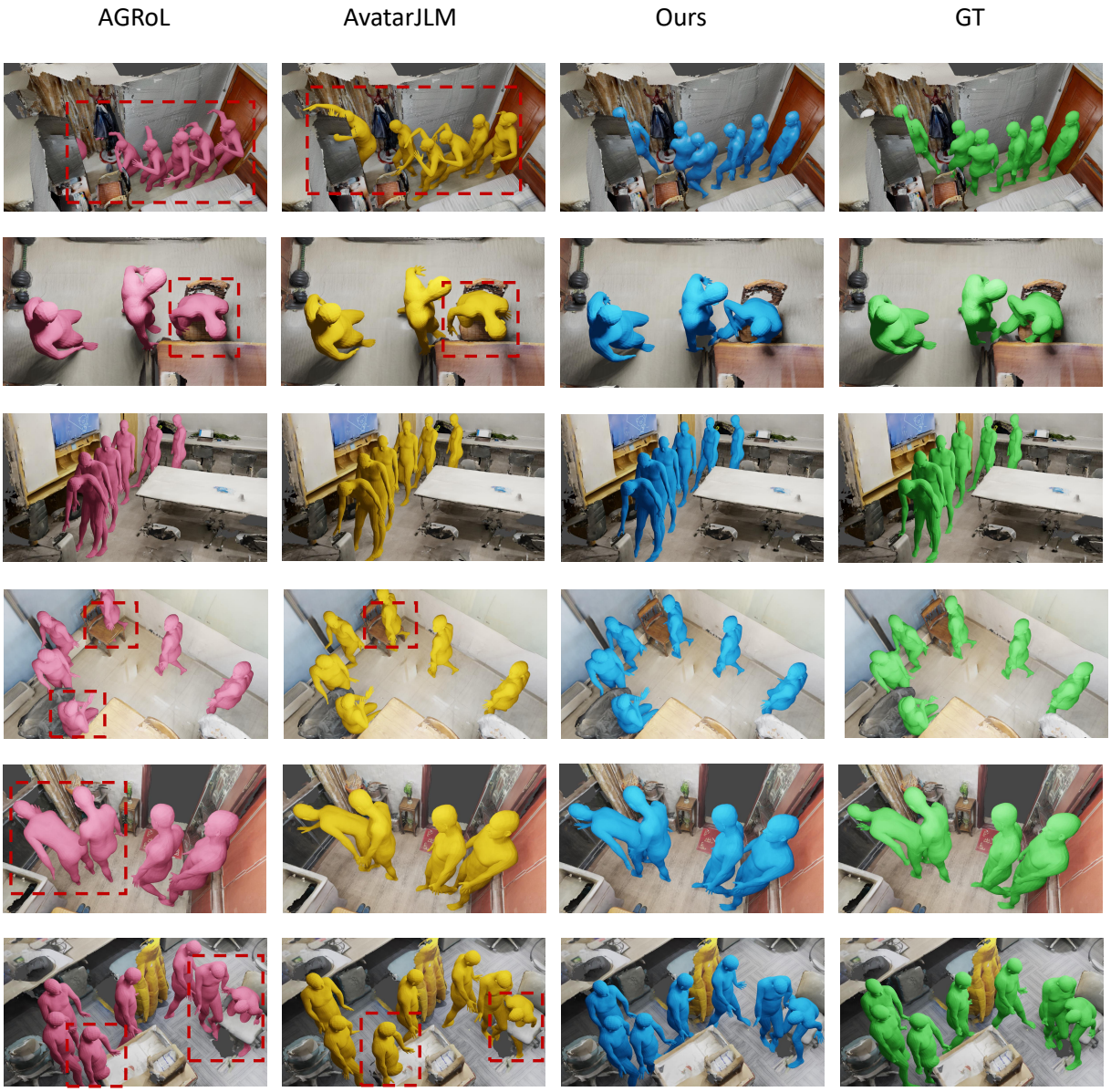

Figure 6. The extra qualitative experiment evaluated on the GIMO dataset.

estimates full-body motion from egocentric videos. For a fair comparison, we provide ground truth head motions to EgoEgo and use its conditional diffusion network to generate full-body motions.

The quantitative results are shown in Table 1, where we report the same metrics as EgoEgo. Although using similar diffusion backbones, by using extra scene modality, our method has higher estimation accuracy, showcasing the benefit of incorporating scene information.

	GIMO [77]			CIRCLE [3]		
Method	**MPJPE**	**Accel**	**FS**	**MPJPE**	**Accel**	**FS**
EgoEgo	125.7	10.2	1.7	96.9	8.3	2.0
Ours	**108.1**	**10.1**	1.7	**73.5**	**7.5**	**1.8**

Table 1. Full-body motion estimation results evaluated on GIMO [77] and CIRCLE [3], given head motion only.

5. Human perceptual study

We conducted a human perceptual study to investigate the quality of the motions generated by our model. We invite 25 users to provide three comparisons. For each comparison, we ask the users *"Which of the two motions is more realistic?"*, and each user is provided 10 sequences to evaluate.

The results are shown in Figure 3 and Figure 4. Our results were preferred over the other state-of-the-art and are even competitive with ground truth motions on the CIRCLE dataset.