Supplementary Material: Towards Open Domain Text-Driven Synthesis of Multi-Person Motions

Mengyi Shan[1], Lu Dong[2], Yutao Han[3], Yuan Yao[4], Tao Liu[3], Ifeoma Nwogu[2], Guo-Jun Qi[5], and Mitch Hill[3]

[1] University of Washington, Seattle WA 98105, USA
[2] University at Buffalo, Buffalo NY 14261, USA
[3] Innopeak Technology, Seattle WA 98004, USA
[4] University of Rochester, Rochester NY 14627, USA
[5] Westlake University, China

1   Video Results

We present motion samples from WebVid-Motion and motion samples generated by our model in video format. Please refer to the videos folder and/or the HTML file to view our results. The HTML page includes:

(i) Video/motion pairs from the WebVid-Motion dataset. Top is original video; bottom is the estimated motion. If the motion is longer than 60 frames, we present the first 60 frames.
(ii) Multi-person samples from our model. This showcases our model's ability to generate multi-person motion sequences with an arbitrary number of subjects for diverse text prompts.
(iii) Comparison between 2-person motion samples from our model and 2-person motion baselines. This showcases our model's ability to generate 2-person motion sequences with open domain prompts.
(iv) Visualization of our model results with different pose/motion guidance terms. This illustrates how a higher motion guidance scale is able to improve the motion quality of each person, while a higher pose guidance scale is able to coordinate the interaction among multiple subjects.

2   Limitations

Pose Dataset. Several limitations of our pose dataset relate to the performance of BEV. BEV often does not provide precise 3D location estimates. This is due to some extent to the innate ambiguity of predicting 3D information from a 2D image. Poses predicted by BEV sometimes do not accurately reflect poses in the source image. The poses predicted by BEV might not capture nuanced details that are obtained from motion capture data. The issues discussed above could potentially be alleviated by employing generative post-processing such as BUDDI [2] to refine the BEV estimates. We performed initial experiments applying BUDDI to our pose data and saw promising results. However, we leave

further investigation for future work. Another limitation of our pose dataset is hallucinations from the Instruct BLIP captioner. Nonetheless, Instruct BLIP essentially always provided better pose captions than the LAION text. An additional limitation is that our pose samples over-represent certain pose configurations, especially people standing for photos. Balancing the dataset across different types of poses and motions is a direction for future work. Our approach will benefit as more robust methods for estimating 3D poses from in-the-wild images become available.

Motion Dataset. We use two motion datasets with more than two people in our training. However, neither can be considered a multi-person motion dataset that is grounded in real-world motions with a high level of accuracy like HumanML3D and Interhuman. *HumanML3D-Comp* does not involve any interaction among subjects. *WebVid-Motion* does not match the quality of motion capture datasets since TRACE often suffers from inaccurate estimation especially for translation. Even after curating the samples where TRACE performed at its best, we still find the predicted global translations show noticeable jitter over nearby frames. The motion estimation quality is severely downgraded especially if the human movement is entangled with complicated camera movements. We believe the quality of multi-person motion data is still the major bottleneck of our model. Our joint training would very likely benefit newly emerging data either from multi-person motion capture or from more accurate methods for estimation 3D group motions from video.

Evaluation. It is hard to evaluate text-driven multi-person motion generation results without a good shared embedding space for text and multi-person motion. Prior works train a shared encoder based on HumanML3D [1] for single person motion. As our WebVid-Motion multi-person motion dataset is relatively small and not a real motion capture dataset, we believe training a motion encoder using our multi-person motion data would not yield a reliable model. Thus we decide to evaluate the model in a decomposed way by measuring the quality of poses for snapshot frames and motions of single subjects. We believe the LAION-Pose data quality is high enough to make this is a reliable solution to measure the quality of the poses in snapshot frames. Yet there remain limitations in our ability to evaluate the quality of individual motions. We use FID between HumandML3D motions and single motions from our multi-person motion samples to evaluate model motion quality. However, there is a distribution mismatch between our generated samples and the reference samples because the motions described by the LAION-Pose prompts along with center-frame pose conditions will likely not result in a motion distribution close to that of HumanML3D motions even if our model is perfect. Nonetheless, we do find some evidence motion FID can provide a rough measure of model motion quality. This is likely because the motions in HumanML3D like contain representative sequences that are echoed in a large proportion of in-the-wild motions. Establishing more robust methods for measuring the quality of multi-person motions is an important direction for future work.

Model. Our model is not designed for complex sequences along the temporal axis (i.e. do A and then do B). It also cannot generate motion for specific texts describing individual motions in a group (i.e. three people do A and two people do B). This is because our training prompts are mostly describing short motions on a group level. Better datasets with specific textual annotations could extend the capabilities of our model.

3   Data Processing Details

We elaborate on some of our data filtering and processing steps below. Fig. 1 shows LAION images removed by different stages of the filtering pipeline, and Fig. 2 presents images that make it through all stages. Fig. 3 presents LAION-Pose samples before and after processing.

3.1   LAION-Pose

LAION Metadata Filter. Before processing data, we used the LAION metadata to remove unsuitable samples. Images with height or width fewer than 256 pixels are removed because BEV performs poorly on these samples. We only process images where the likelihood of harmful content is annotated as `UNLIKELY` and remove all others.

Hand-Crafted Prompt Filter. A list of all hand-crafted prompts is given below.

- `a picture of a piece of clothing`
- `a picture of a shirt`
- `a cd cover`
- `a dvd cover`
- `a book cover`
- `a video game cover`
- `a photo of shoes`
- `a webpage screen with text`
- `a poster with text`

Vertical Height Adjustment. Vertical height adjustment ensures that the height coordinate of the lowest vertex on each mesh is set to the height of the ground. This is accomplished by calculating the SMPL meshes from the SMPL parameters, finding the vertex on the mesh with the lowest height, and adjusting the height coordinate of the global translation pose parameter so that the this vertex is on the ground. In the majority of cases, vertical height adjustment does not change the essence of the group pose while making the group pose appear more natural in the empty rendering environment and compensating for errors in the 3D locations estimated by BEV. We retain the non-adjusted BEV estimates for each sample and use a small bank of 10K non-adjusted samples during

training. Accurate estimation of relative heights of group poses, and better 3D localization overall, is an important direction for future work.

Mesh Separation. We address mesh overlap by optimizing each pair of SMPL parameters to minimize a mesh collision loss. Overlap is measured using the SDF (signed distance function) from all vertices in a source pose to a reference mesh:

$$L(\theta) = \sum_{k=1,\ldots,V} - \min \left\{ \mathbf{sdf} \left(\boldsymbol{v}_k(\theta) \right), 0 \right\}$$

where $\boldsymbol{v}_k(\theta) \in \mathbb{R}^3$ is the k-th vertex on the source pose with parameters θ and **sdf** is computing the signed distance function from that vertex to the reference mesh. Given a set of pose parameters $(\theta_1, \ldots, \theta_n)$, the total collision loss is the sum of collision losses been each subject i and the SDF of all other meshes with parameters θ_j:

$$\mathcal{L}(\theta_1, \ldots, \theta_n) = \sum_{i=1}^{n} \sum_{j \neq i} L_j(\theta_i)$$

where L_j is the SDF loss defined by treating subject j as a fixed reference mesh. We run gradient descent on this loss for a fixed number of 25 steps. The resulting meshes have no or very little overlap. Vertical height adjustment is applied after each gradient step to ensure that the meshes remain on the ground throughout the separation steps.

3.2   WebVid-Motion

Motion Grouping. After applying TRACE to the selected videos, the TRACE outputs are processed by detecting all clips where two or more subjects appear simultaneously in the video for at least 30 frames. In practice, we start by scanning through all pairs of subjects and keep those with more than 30 frames overlap in time. Then we merge any clips of subjects with at least 2 people in common, and more than 30 frames of temporal intersection. Frames that do not contain all merged subjects are discarded. In this way, we may sacrifice the motion lengths a little bit to get the maximum possible number of subjects appearing in one motion sequence. Fig. 4 shows a visual explanation of the grouping effect. Here each row represents a subject and each column represent a frame. White means certain identity is detected at that frame and black means they are not detected. By grouping and merging, we find those gray regions within red circles with at least two subjects and at least a certain length. This video sequence, for example, produces five multi-person motion sequences at different frames with different subjects.

4   Data Format and Attention Masking

Fig. 5 visualizes our unified format for multiple datasets by drawing out the the masked/padded and non-masked poses of different sample types. All samples
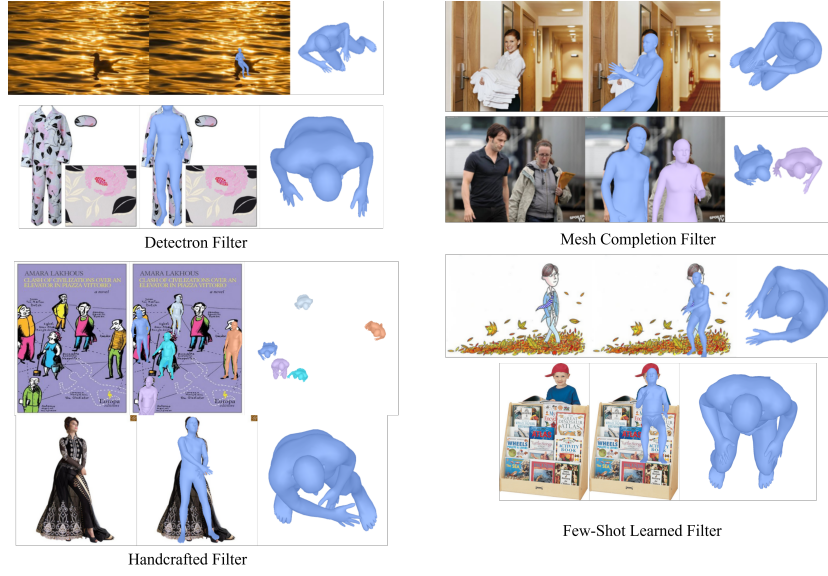
Fig. 1: Examples BEV applied to images discarded in different data filtering steps. Filtering removes images where BEV is not suitable for application or where it produces unsatisfactory results.
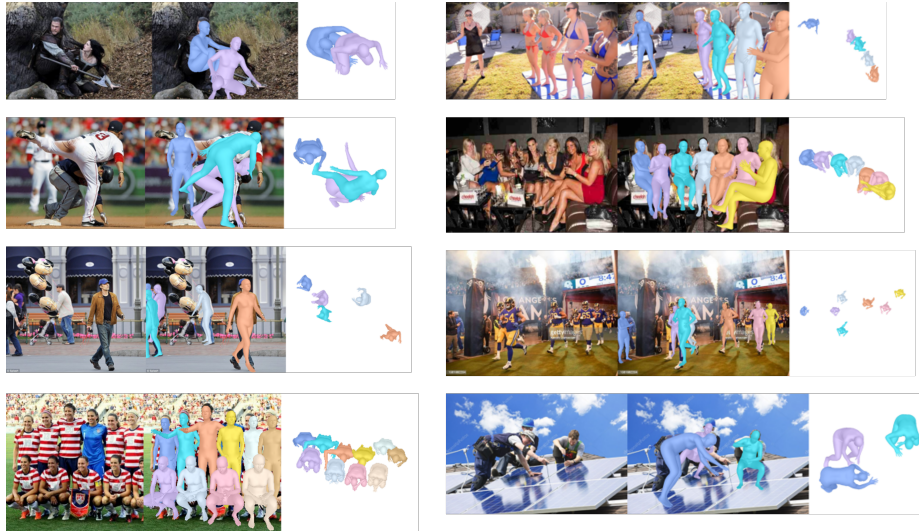

Fig. 2: Examples of BEV applied to images that get through all the filters. BEV produces accurate results for the majority of images that make it through all filtering stages.
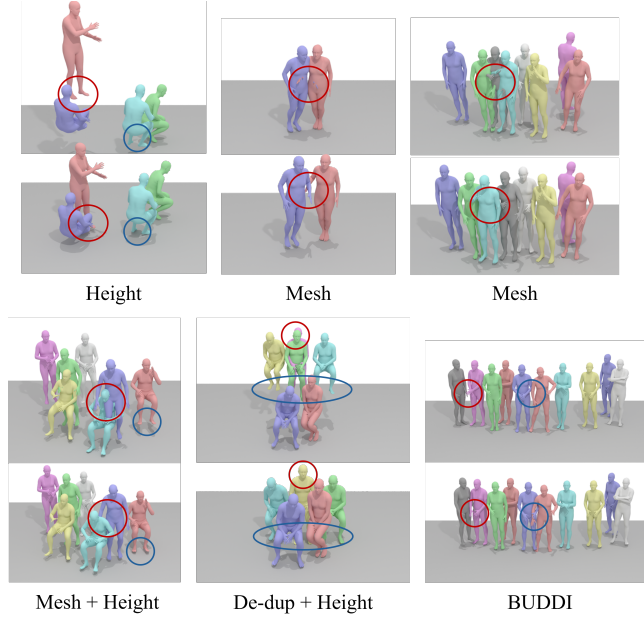
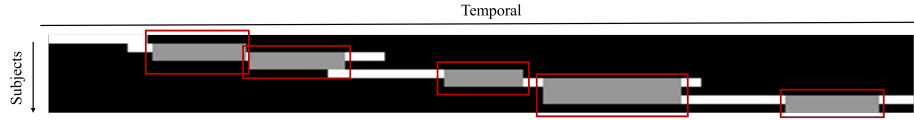**Fig. 3:** Examples before (top) and after (bottom) each pose correction steps.
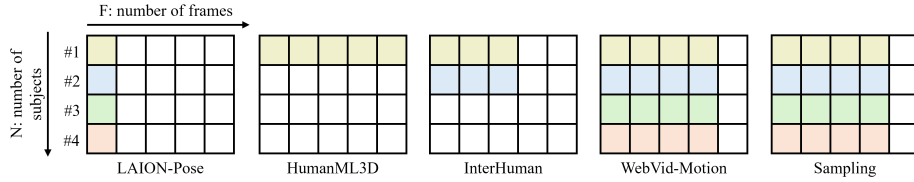

Fig. 4: Motion grouping visual explanation.

Fig. 5: Exemplar padding/masking of samples from our multiple datasets. Our model can also accommodate samples where different subjects are unmasked for varying time windows which do not fully overlap, although this is not done in the current work.

are treated as motion sequences with a F maximum frames and N maximum subjects. Sequences will generally have fewer than F frames (possible only a single frame) and/or fewer than N subjects (possibly only a single subject). For the purpose of parallel batch processing, padding is added until each sequence has the maximum number of frames and subjects so that all samples in a batch have the same shape. Padded states are masked out when passing through attention layers. Pose layers will mask out padded subjects, and motion layers will mask out padded frames. In principle F and N can be set to arbitrary numbers, although our model will not work on sequences/groups that are longer/larger than those seen during training.

5 Evaluation Details and Metrics

We use similar set of metrics as in [1]. The details of each metric are explained below. Metric computation involve three types of features: ground-truth motion features f_{gt}, generated motion features f_{pred}, and text features f_{text}. These features are obtained from a contrastive feature extractor that we trained following CLIP [3]. We train a pose feature extractor and text feature extractor using text and pose pairs from LAION-Pose and a single-person motion feature extractor using HumanML3D text and SMPL motion pairs. We do not use the text feature extractor trained with motion data and primarily use the motion feature extractor to measure the realism of single-person motions using FID. 10K reference samples of the LAION-Pose validation data and HumanML3D training data are used to compute pose and motion f_{gt} respectively. Reference samples are used only for FID. All metrics use 1024 generated samples except Multimodality which uses 100 generated samples. f_{pred} is obtained from the pose and motion encodings of the generated samples. Pose features are extracted every 15 frames and motion features are extracted for each subject in the generated motions. The maximum number of people is set to 10 and the number of generated frames is 61. All samples are generated with prompts from LAION-Pose. f_{text} is calculated from these prompts using pose text feature extractor.

FID (Frechet Inception Distance) measures the distance between generated and testing motion distribution, and thus assesses the overall quality of generated motions. FID is computed as

$$\text{FID} = \|\mu_{\text{gt}} - \mu_{\text{pred}}\|^2 - \text{Tr}\left(\Sigma_{gt} + \Sigma_{\text{pred}} - 2\left(\Sigma_{\text{gt}}\Sigma_{\text{pred}}\right)^{\frac{1}{2}}\right)$$

where μ_{gt} and μ_{pred} are mean of f_{gt} and f_{pred}. Σ is the covariance matrix and Tr is the trace of a matrix.

R-precision measures retrieval accuracy by comparing the text to the generated motions. 1024 generated samples are split into 32 groups, each with 32 pairs. We compute the Top 1, 2, 3 retrieval accuracy within each group for each of the 32 pairs, and take average over the 32 groups.

Similarity (Sim) measures the cosine similarity between pairs of generated motions and text prompts in the feature space:

$$\text{Sim} = \sum_{i=1}^{N} f_{\text{pred},i} \cdot f_{\text{text},i}$$

where $f_{\text{pred},i}$ and $f_{\text{text},i}$ are the features of the i-th text-motion pair. The encodings $f_{\text{pred},i}$ and $f_{\text{text},i}$ are normalized and similarity scores have an upper bound of 1. A similarity near 0 indicates a lack of alignment.

Diversity (Div) evaluates the differences between independently sampled motions. tion sequences in the dataset. We calculate this by randomly selecting S_{dis} pairs of motion features $f_{\text{pred},i}$ and $f'_{\text{pred},i}$ and then computing:

$$\text{Diversity} = \frac{1}{S_{dis}} \sum_{i=1}^{S_{dis}} \left\| f_{\text{pred},i} - f'_{\text{pred},i} \right\|.$$

S_{dis} is set to 300 in our experiments.

Multimodality (MM) assesses the variance within motions generated from a single text. For each of 100 text descriptions, we generate 20 motions which are split into two subsets containing 10 motions each. The features of the j-th pair for the i-th text description are denoted as $f_{\text{pred},i,j}$ and $f'_{\text{pred},i,j}$. Multimodality is defined as:

$$\text{MModality} = \frac{1}{10N} \sum_{i=1}^{N} \sum_{j=1}^{10} \left\| f_{\text{pred },i,j} - f'_{\text{pred },i,j} \right\|.$$

6   Ablation on Guidance Terms

Classifier-Free Guidance. In Tab. 1 we provide additional ablation results on classifier-free guidance terms at sampling time. Note how the pose scores increase along with the classifier-free guidance strength in both first and second stages. As found in other diffusion and motion generation works, classifier-free guidance improves the sample quality and text-sample alignment.

Pose Guidance. In Tab. 2 we provide additional ablation results on multiple pose guidance terms at sampling time. Note how the pose metrics improve along with the pose guidance strength.

Motion Guidance. The effect of motion guidance strength can be more directly seen through visualized samples. Refer to the bottom of the video page for a grid comparison of different pose and motion guidance strength. Note how the motion quality gets better with a higher motion guidance score (smoother, more natural). In particular, we found motion guidance very helpful for reducing translational jitters that are likely learned from imperfections in WebVid-Motion samples. On the other hand, the text-motion alignment gets worse as guidance

increases due to the unconditional nature of our motion model. The guidance strength is set to a value where motion quality is visually improved without too much degradation in the pose quality metrics. For example, the motion often disregards the textual description itself and starts to show HumanML3D type motions like walking instead of standing still. As discussed in the limitation section, motion FID provides only a rough estimate of motion realism and motion FID scores slightly increase even as visual motion quality clearly improves when guidance is applied.

Table 1: Ablation results for different classifier-free guidance term scales. The "1/2" here means different scales for the first and second stage of our model.

1/2 CFG	P-R-Precision ↑			P-FID ↓	P-Sim ↑	P-Div →	M-FID ↓	M-Div →
	Top-1	Top-2	Top-3					
Data	0.621	0.737	0.819	0.000	0.378	1.366	0.002	1.342
1.0/1.0	0.380	0.537	0.626	0.224	0.245	1.318	0.682	1.220
1.75/1.0	0.438	0.589	0.672	0.220	0.260	1.329	0.674	1.210
1.0/1.5	0.420	0.561	0.647	0.221	0.253	1.328	0.677	1.205
1.75/1.5	0.539	0.704	0.776	0.229	0.304	1.329	0.684	1.220

Table 2: Ablation results for different pose and motion guidance term scales.

P/M Guidance	P-R-Precision ↑			P-FID ↓	P-Sim ↑	P-Div →	M-FID ↓	M-Div →
	Top-1	Top-2	Top-3					
Data	0.621	0.737	0.819	0.000	0.378	1.366	0.002	1.342
0.0/0.0	0.525	0.685	0.766	0.253	0.297	1.328	0.687	1.204
0.2/0.0	0.546	0.696	0.780	0.218	0.303	1.341	0.673	1.231
0.4/0.0	0.560	0.723	0.795	0.200	0.310	1.342	0.683	1.234
0.0/0.2	0.515	0.677	0.769	0.207	0.300	1.335	0.700	1.207
0.2/0.2	0.549	0.710	0.791	0.192	0.307	1.340	0.713	1.220
0.4/0.2	0.547	0.713	0.778	0.173	0.309	1.332	0.706	1.211
0.0/0.4	0.535	0.694	0.776	0.193	0.354	1.377	0.727	1.229
0.2/0.4	0.529	0.702	0.792	0.181	0.354	1.382	0.715	1.210
0.4/0.4	0.556	0.718	0.790	0.175	0.315	1.382	0.736	1.118

References

1. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5152–5161 (June 2022)
2. Müller, L., Ye, V., Pavlakos, G., Black, M.J., Kanazawa, A.: Generative proxemics: A prior for 3D social interaction from images. arXiv preprint 2306.09337v2 (2023)

3. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)