Length Aware Motion Synthesis via Latent Diffusion

Alessio Sampieri*, Alessio Palma*, Indro Spinelli, and Fabio Galasso
Sapienza University of Rome, Italy
surname@di.uniroma1.it

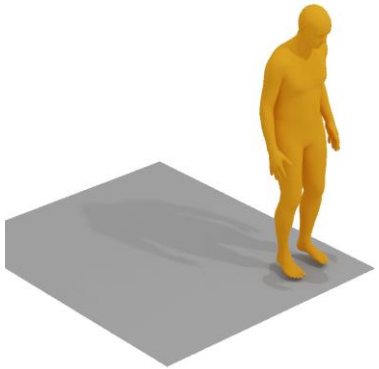
EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO 2024

* Authors contributed equally

The task

Text-to-motion synthesis is the task of generating human motions that conform to an input textual description w. Applications in virtual reality, animation, videogames, robots-human interaction.

E.g. *"A person moves backwards, sits down, then stands back up."*



Related Works

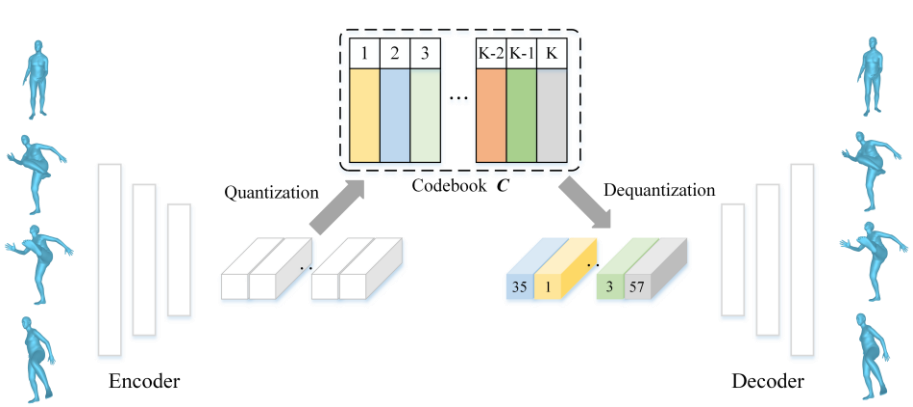
- Earlier works (TEMOS, MDM) **directly** leveraged generative models.
- As of today**,** learning a latent representation and conditional generation are the two building blocks of this task.
- **Learning a latent representation:** can occur in either a discrete (VQ-VAE) or continuous (VAE) space.
- **Generation:** autoregressive (GPT) or one-shot (latent DDPM).
- **RAG** methodologies have begun to be included in pipelines more recently.

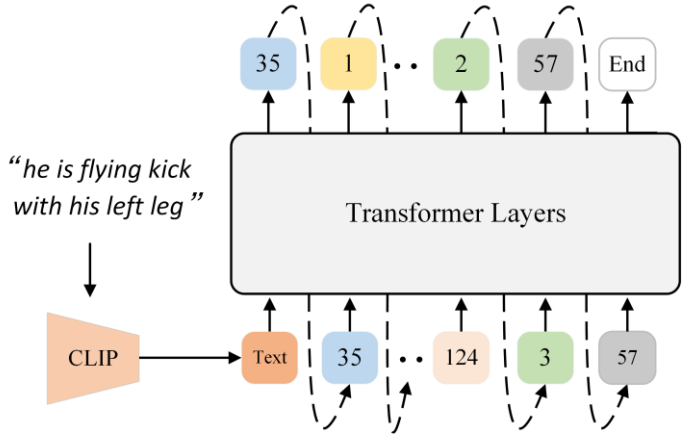
VQ-VAE + GPT
Discrete

VAE + DDPM
Continuous

Related Works: T2M-GPT

VQ-VAE + GPT



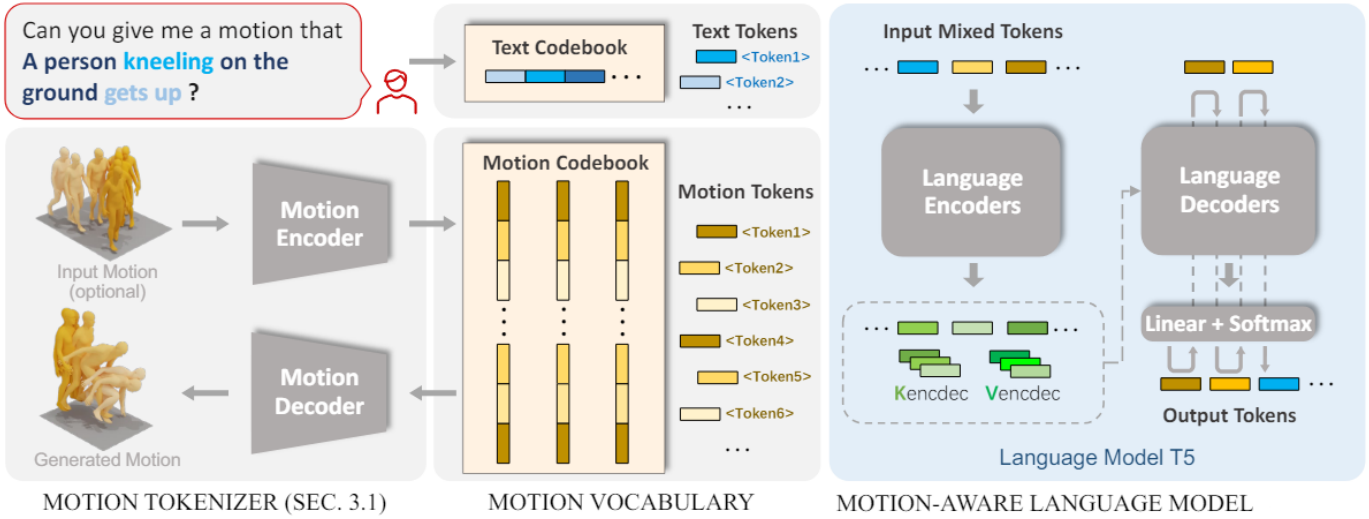


Discrete tokens yield limited expressive capability.

No control over the length of the produced motion: it just depends on when the *End* token is predicted.

Related Works: MotionGPT

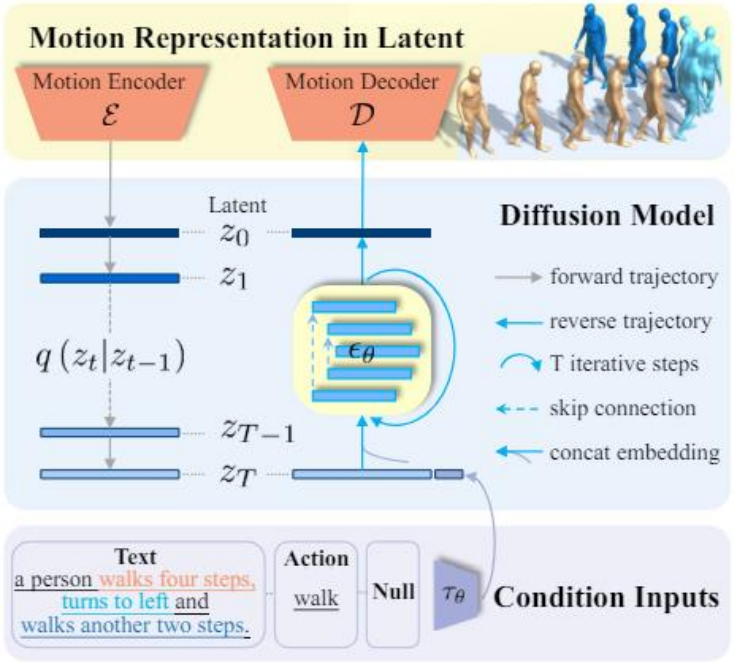
VQ-VAE + GPT



Same limitations as previous model.

Related Works: MLD

VAE + DDPM



VAE representation is agnostic of the desired output sequence size.

Diffusion-based synthesis does not account for mechanisms which affect the output sequence style depending on the desired length.

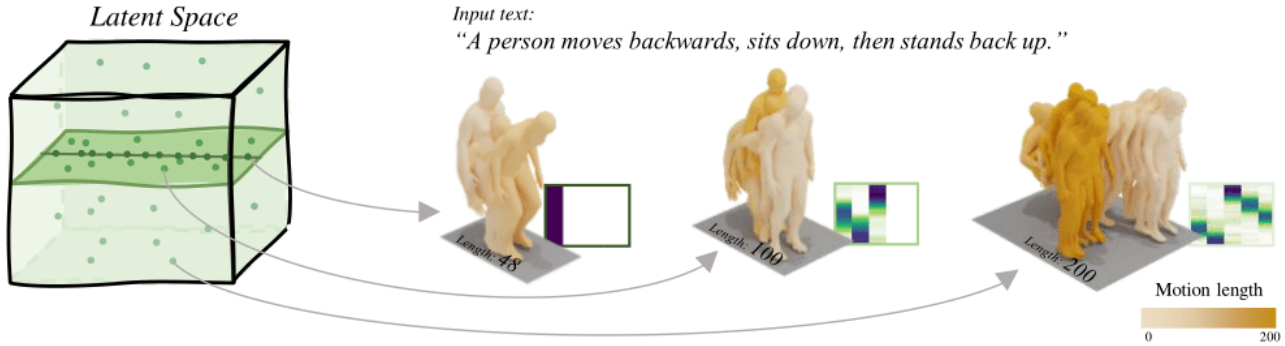
Our idea

- SOTA approaches cannot control the synthesized motion's length or consider the target sequence's variable length as a stylizing attribute.
- Suppose we want to generate a short kick motion: it is **not enough** to subsample a lengthy one, as it will not capture the intricate variations that occur when humans perform actions at different speeds.
- The embedding space should encode longer sequences with **larger** capacity, because they need more details to be generated, and shorter sequences with **less** capacity.

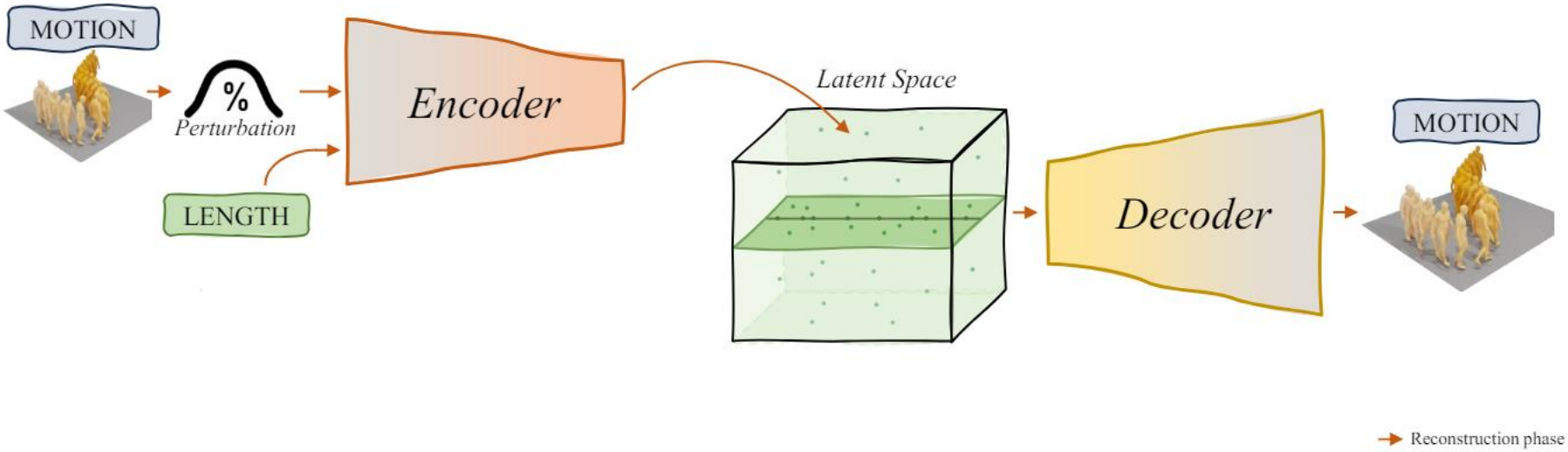
For all these reasons, we introduce **"Length-Aware Latent Diffusion"** (**LADiff**)

LADiff: contributions

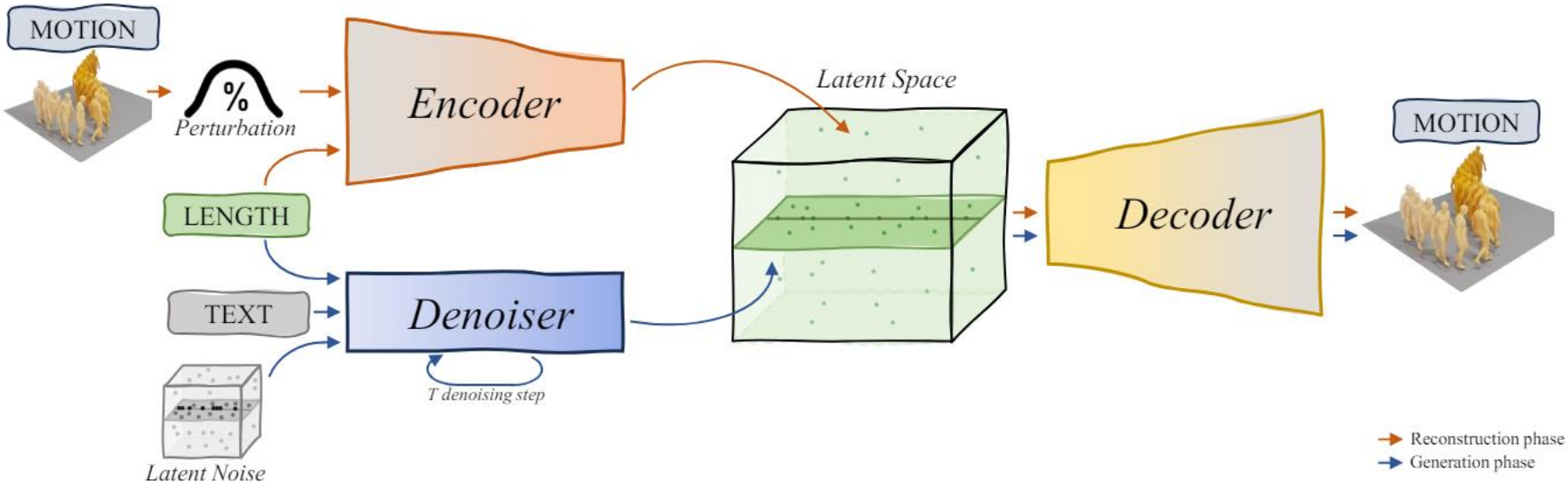
- **Novel length-aware VAE**, designed to learn motion representations with length-dependent latent codes.
- **Novel length-aware latent DDPM**, which generates motions with a richness of details that increases with the target sequence duration.
- Organize the latent representation space into **subspaces**, which activate progressively with the increasing target sequence length.



LADiff: Overview



LADiff: Overview



LADiff: Length-Aware VAE

- **Decompose** the entire latent space into K subspaces: the complete space has dimension $R^{K\times D}$, and the smallest subspace is $1\times D$-dimensional.
- As the motion length grows, we **stepwise unlock** bigger subspaces following the activation rate $k = [\frac{f}{r}]$, where f is the length of the motion, and r is the number of frames assigned to each subspace.
- Each motion $\boldsymbol{x}$ will have its own embedding $\boldsymbol{z} \in R^{k\times D}$.

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mu_1(\mathbf{x}), \sigma_1^2(\mathbf{x})\mathbf{I}, \ldots, \mu_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\mathbf{I})$$

$$\mathbf{z}^{(i)} = [\boldsymbol{\mu}_1^{(i)} + \boldsymbol{\sigma}_1^{(i)2} \odot \boldsymbol{\rho}_1, \ldots, \boldsymbol{\mu}_k^{(i)} + \boldsymbol{\sigma}_k^{(i)2} \odot \boldsymbol{\rho}_k]$$

LADiff: Length-Aware VAE

- **Decompose** the entire latent space into K subspaces: the complete space has dimension $R^{K\times D}$, and the smallest subspace is $1\times D$-dimensional.
- As the motion length grows, we **stepwise unlock** bigger subspaces following the activation rate $k = \lceil \frac{f}{r} \rceil$, where f is the length of the motion, and r is the number of frames assigned to each subspace.
- Each motion $\boldsymbol{x}$ will have its own embedding $\boldsymbol{z} \in R^{k\times D}$.

- The transformer-based decoder handles the varying dimensional space by using the **masked attention** mechanism.

LADiff: Length-Aware DDPM

- Forward diffusion process gradually converts latent representations $\boldsymbol{z}_0 = \boldsymbol{z}$ into random noise $\boldsymbol{z}_T$ in T timesteps, following:

$$g(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{\bar{\alpha}_t}\mathbf{z}_{t-1}, (1-\bar{\alpha}_t)\mathbf{I})$$

- The reverse denoising process gradually refines the noised vector to a suitable representation $\boldsymbol{z}_0$ through $\boldsymbol{z}_{t-1} = \epsilon_\psi\big(\boldsymbol{z}_t, t, \gamma(w)\big)$, with ϵ_ψ being a denoising transformer u-net.
- At inference time we **initialize** $\boldsymbol{z}_T$ using our activation rate k derived from the desired motion length f^*:

$$\mathbf{z}_T = \mathcal{N}(\mathbf{0}, \mathbf{I}) \in \mathbb{R}^{\lceil \frac{f^*}{r} \rceil \times D}$$

LADiff: Denoising VAE

- We train VAE and latent DDPM in cascaded stages.
- VAEs aim to reconstruct clean inputs as faithfully as possible. Conversely, latent DDPMs generate latent vectors from pure Gaussian noise.
- DDPM-generated latents may contain some **residual noise**, which poses challenges for the VAE decoder.
- **DVAE** to align the latent variable distributions: perturb a percentage of input frames with Gaussian noise $\epsilon \sim \mathcal{N}(0,1)$.

Quantitative results: HumanML3D

We evaluate our model on **HumanML3D** (14.6k motion sequences, 44.9k textual descriptions) and KIT-ML.

Methods	R Precision ↑			FID ↓	MM-Dist ↓	Diversity →	MModality ↑
	Top 1	Top 2	Top 3				
Real	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
MotionDiffuse [11]	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$0.630^{\pm.001}$	$3.113^{\pm.001}$	$9.410^{\pm.049}$	$1.553^{\pm.042}$
MDM [21]	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$	$2.799^{\pm.072}$
MLD [9]	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$	$2.413^{\pm.079}$
T2M-GPT [5]	$0.491^{\pm.003}$	$0.680^{\pm.003}$	$0.775^{\pm.002}$	$0.116^{\pm.004}$	$3.118^{\pm.011}$	$9.761^{\pm.081}$	$1.856^{\pm.011}$
MotionGPT [6]	$0.492^{\pm.003}$	$0.681^{\pm.003}$	$0.778^{\pm.002}$	$0.232^{\pm.008}$	$3.096^{\pm.008}$	$\mathbf{9.528}^{\pm.071}$	$2.008^{\pm.084}$
Fg-T2M [33]	$0.492^{\pm.002}$	$0.683^{\pm.003}$	$0.783^{\pm.002}$	$0.243^{\pm.019}$	$3.109^{\pm.007}$	$9.278^{\pm.072}$	$1.614^{\pm.049}$
M2DM [16]	$0.497^{\pm.003}$	$0.682^{\pm.002}$	$0.763^{\pm.003}$	$0.352^{\pm.005}$	$3.134^{\pm.010}$	$9.926^{\pm.073}$	$\mathbf{3.587}^{\pm.072}$
AttT2M [15]	$0.499^{\pm.006}$	$0.690^{\pm.002}$	$0.786^{\pm.002}$	$0.112^{\pm.006}$	$\mathbf{3.038}^{\pm.007}$	$9.700^{\pm.090}$	$2.452^{\pm.051}$
Ours$_{(r=32)}$	$0.493^{\pm.002}$	$0.686^{\pm.002}$	$0.784^{\pm.001}$	$\mathbf{0.110}^{\pm.004}$	$3.077^{\pm.010}$	$9.622^{\pm.071}$	$2.095^{\pm.076}$
Ours$_{(r=48)}$	$\mathbf{0.503}^{\pm.002}$	$\mathbf{0.696}^{\pm.003}$	$\mathbf{0.792}^{\pm.002}$	$0.182^{\pm.004}$	$3.054^{\pm.008}$	$9.795^{\pm.076}$	$2.115^{\pm.063}$
ReMoDiffuse [23]	$\mathbf{0.510}^{\pm.005}$	$\mathbf{0.698}^{\pm.006}$	$\mathbf{0.795}^{\pm.004}$	$0.103^{\pm.004}$	$\mathbf{2.974}^{\pm.016}$	$9.081^{\pm.075}$	$1.795^{\pm.028}$
Ours$_{(r=48)}$ RAG	$0.494^{\pm.002}$	$0.691^{\pm.003}$	$0.786^{\pm.001}$	$\mathbf{0.054}^{\pm.002}$	$3.112^{\pm.008}$	$\mathbf{9.517}^{\pm.077}$	$\mathbf{2.453}^{\pm.074}$

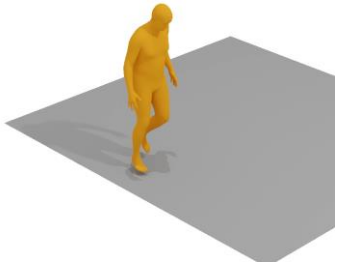
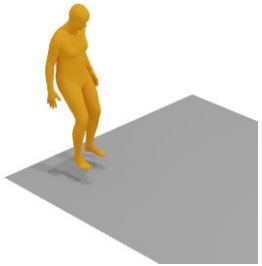
Quantitative results: KIT-ML

We evaluate our model on HumanML3D and **KIT-ML** (3.9k motions, 6.2k textual descriptions).

Methods	R Precision ↑			FID ↓	MM-Dist ↓	Diversity →	MModality ↑
	Top 1	Top 2	Top 3				
Real	$0.424^{\pm.005}$	$0.649^{\pm.006}$	$0.779^{\pm.006}$	$0.031^{\pm.004}$	$2.788^{\pm.012}$	$11.08^{\pm.097}$	-
MotionDiffuse [11]	$0.417^{\pm.004}$	$0.621^{\pm.004}$	$0.739^{\pm.004}$	$1.954^{\pm.062}$	$2.958^{\pm.005}$	$\mathbf{11.10}^{\pm.143}$	$0.730^{\pm.013}$
MDM [21]	$0.164^{\pm.004}$	$0.291^{\pm.004}$	$0.396^{\pm.004}$	$0.497^{\pm.021}$	$9.191^{\pm.022}$	$10.85^{\pm.109}$	$1.907^{\pm.214}$
MLD [9]	$0.390^{\pm.008}$	$0.609^{\pm.008}$	$0.734^{\pm.007}$	$\mathbf{0.404}^{\pm.027}$	$3.204^{\pm.027}$	$10.80^{\pm.117}$	$2.192^{\pm.071}$
T2M-GPT [5]	$0.416^{\pm.006}$	$0.627^{\pm.006}$	$0.745^{\pm.006}$	$0.514^{\pm.029}$	$3.007^{\pm.023}$	$10.92^{\pm.108}$	$1.570^{\pm.039}$
MotionGPT [6]	$0.366^{\pm.005}$	$0.558^{\pm.004}$	$0.680^{\pm.005}$	$0.510^{\pm.016}$	$3.527^{\pm.021}$	$10.35^{\pm.084}$	$2.328^{\pm.117}$
Fg-T2M [33]	$0.418^{\pm.005}$	$0.626^{\pm.006}$	$0.745^{\pm.004}$	$0.571^{\pm.047}$	$3.114^{\pm.015}$	$10.93^{\pm.083}$	$1.019^{\pm.029}$
M2DM [16]	$0.416^{\pm.004}$	$0.628^{\pm.004}$	$0.743^{\pm.004}$	$0.515^{\pm.029}$	$3.015^{\pm.017}$	$11.41^{\pm.097}$	$\mathbf{3.325}^{\pm.037}$
AttT2M [15]	$0.413^{\pm.006}$	$0.632^{\pm.006}$	$0.751^{\pm.006}$	$0.870^{\pm.039}$	$3.039^{\pm.021}$	$10.96^{\pm.123}$	$2.281^{\pm.047}$
Ours$_{(r=48)}$	$\mathbf{0.429}^{\pm.007}$	$\mathbf{0.647}^{\pm.004}$	$\mathbf{0.773}^{\pm.004}$	$0.470^{\pm.016}$	$\mathbf{2.831}^{\pm.020}$	$11.30^{\pm.108}$	$1.243^{\pm.057}$
ReMoDiffuse [23]	$\mathbf{0.427}^{\pm.014}$	$\mathbf{0.641}^{\pm.004}$	$\mathbf{0.765}^{\pm.055}$	$\mathbf{0.155}^{\pm.006}$	$\mathbf{2.814}^{\pm.0.12}$	$10.80^{\pm.105}$	$1.239^{\pm.028}$
Ours$_{(r=48)}$ RAG	$0.415^{\pm.006}$	$0.632^{\pm.007}$	$0.758^{\pm.005}$	$0.386^{\pm.003}$	$2.978^{\pm.020}$	$\mathbf{11.20}^{\pm.008}$	$\mathbf{1.732}^{\pm.066}$

Qualitative results: SOTA comparison

"A person walks in a complete circle and then sits down."

	LADiff	MLD	T2M-GPT †	MotionGPT †
84 frames				
170 frames				

† means no control over motion length

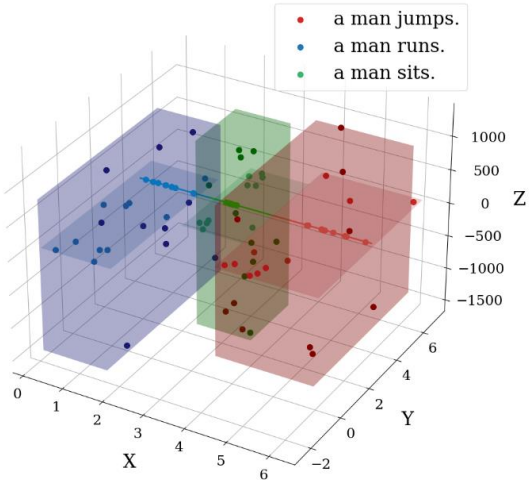
Qualitative results: length comparison

	48 frames	84 frames	130 frames	170 frames	200 frames
"A person has to crawl under an obstacle to continue."					
"A person stepping over something."					

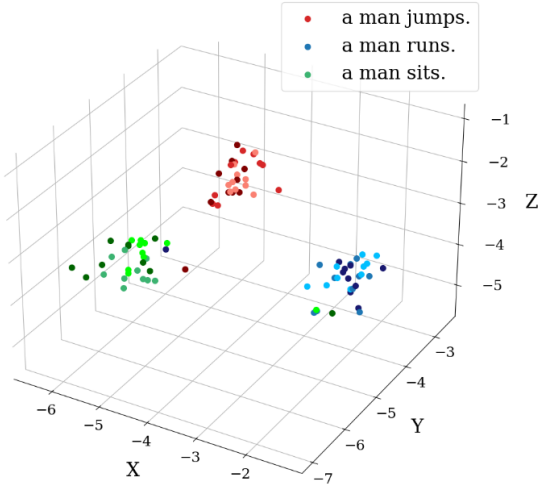
Latent Space Representations

Comparison of the learned latent space between our LA-VAE and a standard VAE, using t-SNE dimensionality reduction. We generate ten times sequences of 30, 96, and 144 frames per 3 actions.

LA-VAE



VAE



Motion length:

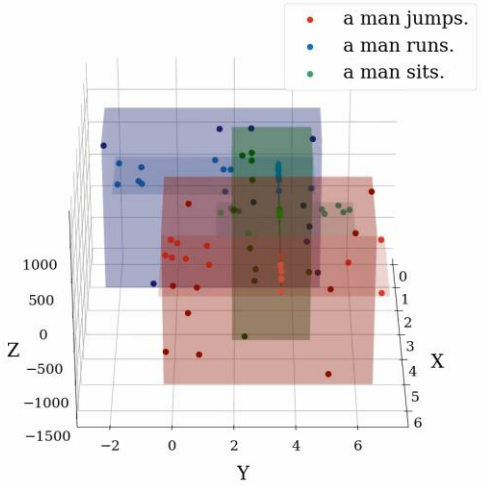
- Light 30
- Mid 96
- Dark 144

Latent Space Representations

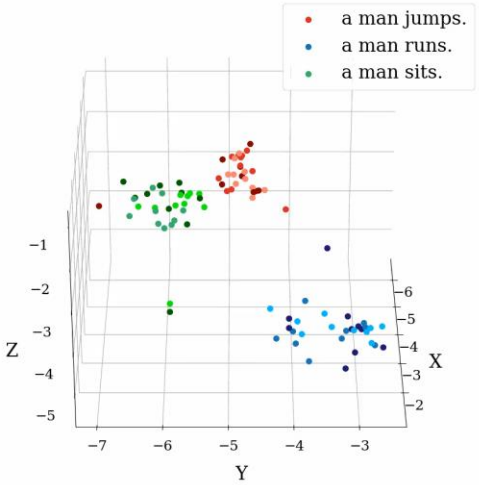
Comparison of the learned latent space between our LA-VAE and a standard VAE, using t-SNE dimensionality reduction. We generate ten times sequences of 30, 96, and 144 frames per 3 actions.

Here embeddings are organized in clustered spaces based on action **and** motion length.

LA-VAE



VAE



Motion length:

Light 30

Mid 96

Dark 144

Conclusions

- LADiff introduces **novel length-aware models** to exploit the length of the target sequence as a given input.
- Sets a **new SOTA** in R-precision and produces motions that adapt in terms of style and dynamics to the desired length.
- **Organizes** motions in the latent space by separating the action types along the first subspace and then by clustering the lengths on the other available subspaces.
- **Code** is available at https://github.com/AlessioSam/LADiff.

Thank you for your attention!

Extra: Motion dynamics

- For MLD, the averages of acceleration and velocity remain **unchanged** as the generated length varies.
- By contrast, for LADiff the statistics increase by 44.4% and 17.3% respectively when the motion is shorter. We confirm therefore that LADiff produces **different motion styles for different lengths**.
- We also repeat the same analysis on a set of atomic actions.

	Avg. Vel. (m/s)		Avg. Acc. (m/s^2)		Max Acc. (m/s^2)	
	84	170	84	170	84	170
MLD	0.39	0.39	0.06	0.05	0.31	0.33
LADiff	0.61	0.52	0.13	0.09	0.61	0.55

	Avg. Vel. (m/s)			Avg. Acc. (m/s^2)			Max Acc. (m/s^2)		
Actions	48	84	170	48	84	170	48	84	170
Sit	0.27	0.17	0.10	0.05	0.04	0.02	0.16	0.16	0.14
Walk	1.31	1.01	0.72	0.11	0.09	0.06	0.19	0.18	0.14
Throw	0.16	0.15	0.13	0.04	0.04	0.03	0.14	0.15	0.13
Mean	0.58	0.44	0.31	0.06	0.05	0.03	0.16	0.16	0.14

Extra: Pose vector

The goal of motion synthesis is to generate a motion $\boldsymbol{x}$ based on textual input w. We define motion as a sequence of poses $\boldsymbol{x} = [x_1, .., x_F] \in R^{F \times V}$ and textual description as a vector $w \in R^{1 \times D}$.

A pose vector $x_i \in R^{1 \times V}$ in our work is defined by a tuple:

$$\left(\dot{r^a}, \dot{r^x}, \dot{r^z}, r^y, j^p, j^v, j^r, c^f\right)$$

containing root angular velocity, root linear velocities, root height, joints positions, velocities, rotations and heel-toe joint velocities.

Motions in the HumanML3D dataset follow the skeleton structure of SMPL with 22 joints. Poses have 21 joints in KIT-ML.

Extra: Reconstruction results

Here we present the results of our proposed length-aware VAE for the **reconstruction** phase on HumanML3D. **MPJPE** is defined as the mean Euclidean distance between the predicted 3D joint locations and the corresponding ground truth joint locations.

	MPJPE ↓	R Prec. Top 3 ↑	FID ↓	MM-Dist ↓	Diversity →
Real	0.0	0.797	0.002	2.974	9.503
MLD [9]	54.4	0.772	0.247	3.101	9.630
T2M-GPT [5]	—	0.785	0.070	3.072	9.593
MotionGPT [6]	55.8	—	0.067	—	9.675
M2DM [16]	—	0.791	0.063	3.015	**9.577**
Ours$_{(r=48)}$ w/o DVAE	**53.2**	**0.792**	**0.048**	**3.002**	9.643
Ours$_{(r=48)}$	**52.6**	**0.797**	**0.054**	**2.993**	9.613

Extra: Metrics (1)

- **R-precision**: for each generated motion, its ground-truth text description and 31 randomly selected mismatched descriptions from the test set form a description pool. Following, calculate and rank the Euclidean distances between the motion feature and each text feature in the pool. We then count the average accuracy at top-1, top-2 and top-3 places.
- **MM-Dist**: the average Euclidean distance between the motion feature of each generated motion and the text feature of its corresponding description.
- **FID**: measures the similarity between the distributions of generated and real motions. It is computed by measuring the L2 loss of the latent representations obtained through the feature extractor.

Extra: Metrics (2)

- **Diversity**: randomly divide the motion feature vectors into two equal-sized subsets $\{x_1, \dots, x_{X_d}\}$ and $\{x'_1, \dots, x'_{X_d}\}$, then compute Diversity as follows:

$$\frac{1}{X_d} \sum_{i=1}^{X_d} ||x_i - x'_i||$$

- **MultiModality**: randomly select T_d text descriptions from the entire collection, then for each t-th description generate two subsets of motions $\{x_{t,1}, \dots, x_{t,X_d}\}$ and $\{x'_{t,1}, \dots, x'_{t,X_d}\}$ of size X_d. MultiModality is then computed as:

$$\frac{1}{T_d \cdot X_d} \sum_{t=1}^{T_d} \sum_{i=1}^{X_d} ||x_{t,i} - x'_{t,i}||$$