HumanTOMATO: Text-aligned Whole-body Motion Generation

Shunlin Lu*, Ling-Hao Chen*

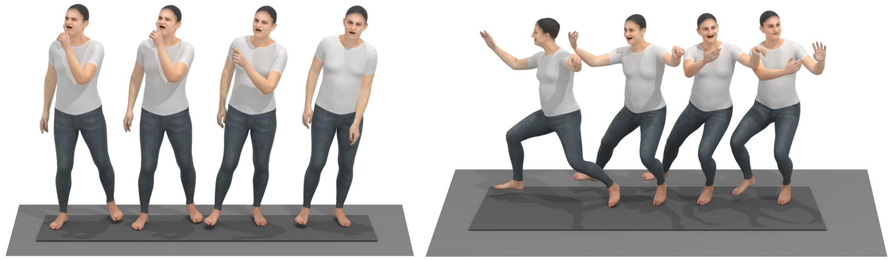
Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, Heung-Yeung Shum

**co-first author, random listing order* *Correspondence: R. Zhang, and H.-Y. Shum*

Project page: https://lhchen.top/HumanTOAMTO

Project Page | Codes | TMA Retrieval | Video demo

Highlight

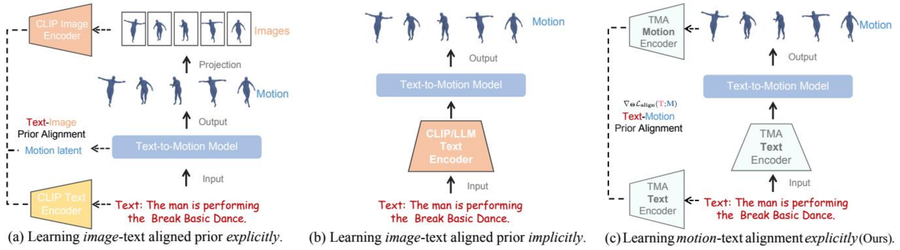


stand and shush, angrily.

Yang-style 40 formTai Chi Competitionroutinestep34, happily.

The proposed HumanTOMATO can generate text-aligned whole-body motions with vivid and harmonious face, hand, and body motion.

Key insight 1: Text-motion alignment



(a) Learning *image*-text aligned prior *explicitly*. (b) Learning *image*-text aligned prior *implicitly*. (c) Learning *motion*-text alignment *explicitly* (Ours).

We clarify the importance of how to use the text and motion data to generate motions for the first time. We highlight the two method.

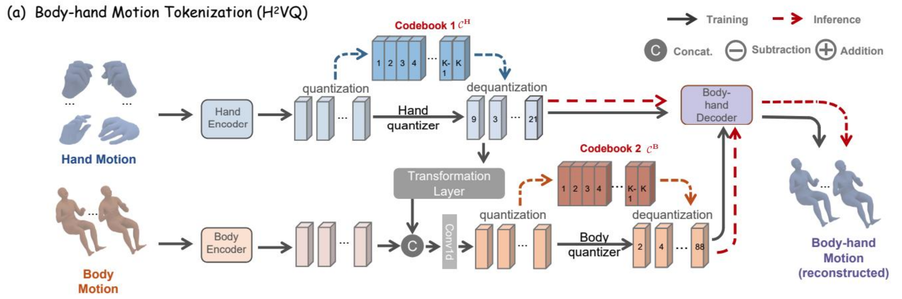
- ✓ Replace your **CLIP** text encoder with **TMA** text encoder.
- ✓ Introduce the text-motion alignment **supervision** to your motion generation model during training.

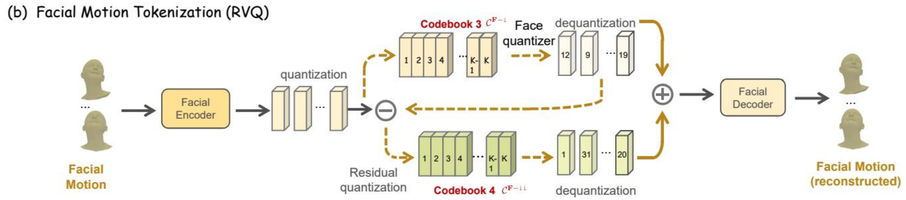
Try TMA!

How to use?

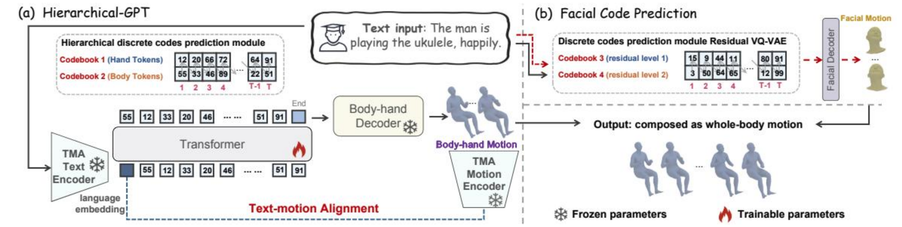
```
from OpenTMA import textencoder, motionencoder
import torch

motion = torch.randn(1, 64, 126)    # B = 1, T = , D = , need normalization
lengths = [64]
t_emb = textencoder(["a man is running"]).loc
m_emb = motionencoder(motion, lengths).loc
```

Key insight 2: Motion quantization





The framework overview of tokenization method for body-hand, and facial motions. (a) Holistic Hierarchical Vector Quantization (H²VQ) to compress fine-grained body-hand motion into two discrete codebooks with hierarchical structure relations. (b) Residual Vector Quantization (RVQ) to compress facial motion into two discrete codebooks with hierarchical structure relations.



Auto-regressive motion generation. Both parts take textual description as input and predict tokens in an auto-regressive manner. The final whole-body motion is composed of both part motions decoded by the corresponding decoders.

Results

SoTA!

	FID↓	R-Precision(32) Top1↑	Top2↑	Top3↑	TMA-R-Precision(256) Top1↑	Top2↑	Top3↑	Matching Score↓	TMA-Matching Score↓	MModality↑	Diversity↑
GT	-	0.500	0.708	0.814	0.407	0.578	0.673	2.888	0.768	-	11.087
TEMOS	9.147	0.279	0.442	0.555	0.258	0.389	0.444	5.482	0.928	1.195	9.764
T2M-GPT	1.366	0.368	0.553	0.655	0.310	0.446	0.527	4.316	0.881	2.356	10.753
MDM	3.800	0.352	0.547	0.634	0.310	0.430	0.530	4.050	0.840	**2.530**	**11.400**
MLD	3.407	0.385	0.571	0.683	0.333	0.477	0.561	3.901	0.883	2.448	10.420
MotionDiffuse	**1.129**	0.391	0.587	0.695	0.368	0.493	0.584	3.950	0.829	1.654	10.580
HumanTOMATO	1.174	**0.416**	**0.603**	**0.703**	**0.399**	**0.555**	**0.638**	**3.894**	**0.809**	1.732	10.812

Main results of motion generation on the Motion-X dataset.

Quantization works!

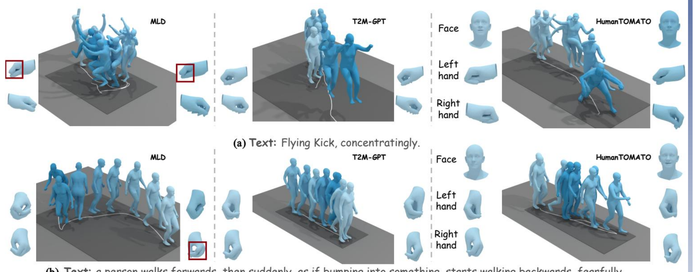
	Motion-X All↓	Motion-X Body↓	Motion-X Hand↓	GRAB All↓	GRAB Body↓	GRAB Hand↓	HumanML3D Body↓
Vanilla VQ (512)	140.66	92.20	46.45	78.23	38.29	31.48	77.21
Vanilla VQ (1024)	139.33	91.77	46.40	76.01	37.34	29.89	71.34
RVQ (512×2)	110.94	73.97	40.01	62.94	31.12	27.28	**63.05**
H²VQ (512×2)	**92.97**	**62.34**	**37.20**	**46.74**	**24.33**	**24.59**	-

Comparison of the motion reconstruction errors (MPJPE in mm) of different quantization methods.

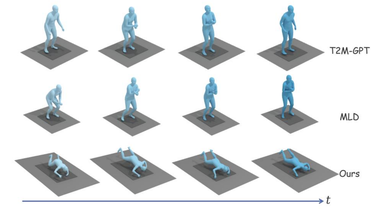
TMA works!

embedding	supervision	FID↓	R-Precision(32) Top1↑	Top2↑	Top3↑	TMA-R-Precision(256) Top1↑	Top2↑	Top3↑	Matching-score↓	TMA-Matching-score↓
GT		0.002	0.500	0.708	0.814	0.407	0.578	0.673	2.888	0.768
CLIP	✗	**1.086**	0.405	0.588	0.695	0.345	0.490	0.573	3.917	0.844
TMA	✗	1.290	0.416	0.596	0.699	0.395	0.550	0.637	3.950	0.815
TMA	✓	1.174	**0.416**	**0.603**	**0.703**	**0.399**	**0.555**	**0.638**	**3.894**	**0.809**

Ablation on a pre-trained text-motion-aligned model for motion generation.



Qualitative comparisons with SOTA models trained on Motion-X



Test-in-the-wild result (unseen text)

Reaching out first authors

Shunlin Lu | Ling-Hao Chen