MotionGPT: Human Motion as a Foreign Language

NeurIPS 2023

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, Tao Chen
紹介者： 渡邉研斗 (AIST)

なぜこの論文を紹介するか？

LLMの登場より自然言語タスクの性能が想像以上に向上

自然言語処理技術： 生活に先端の言語処理技術があるのが当たり前
LLMの解明・効率化研究 / LLM活用研究

マルチ/クロスモーダル： 言語と画像/動画/音声などと組み合わせた手法の研究

コンピュータの中に閉じている

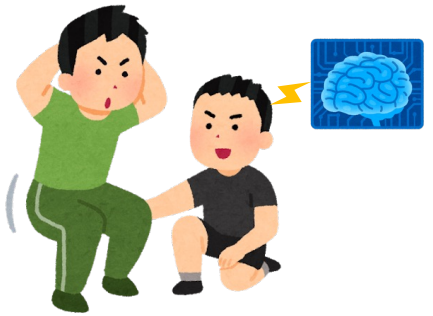
なぜこの論文を紹介するか？

コンピュータの外との干渉/やり取り

言語と地理空間(物理空間)、言語と生理指標(医療情報)、 言語とロボット 等…

「モーション」は現実世界と言語の橋渡し的役割

- モーションは物理空間の制約を受ける
- モーションは生物由来のもの、医療情報とも言える
- モーションは人型ロボットの動きを表現



他にも面白そうな方向性が…

「言語による指導」と「実技(スポーツ/リハビリ/演技/演奏等)モーション」

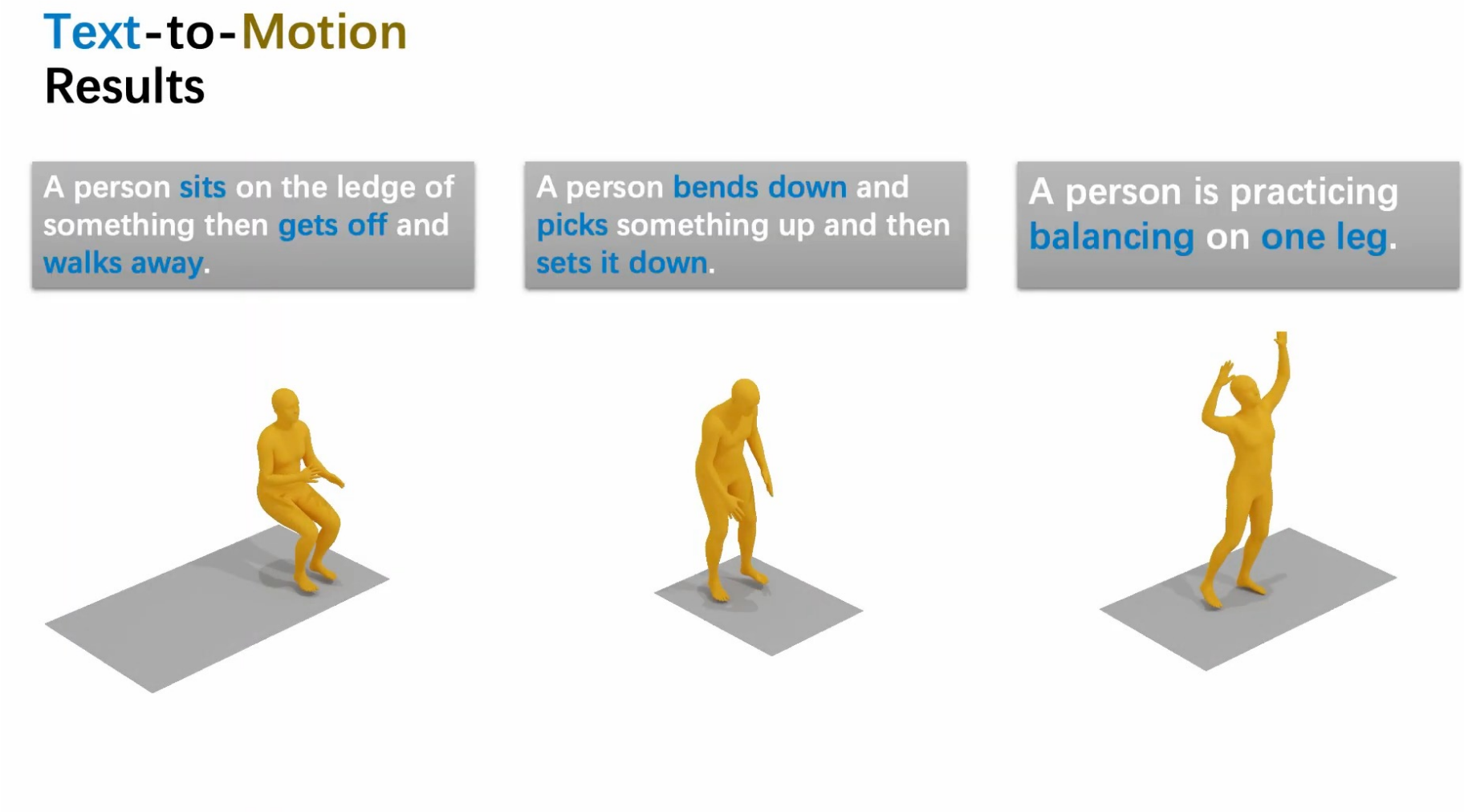
言語とモーションの関係性を扱う技術は、
言語処理研究をコンピュータの外へ持ち出すための基礎技術の一つ

本紹介の目的

最先端の言語-モーション研究の紹介を通して、
「自分でも言語とモーションの研究できそうじゃん」と感じてほしい

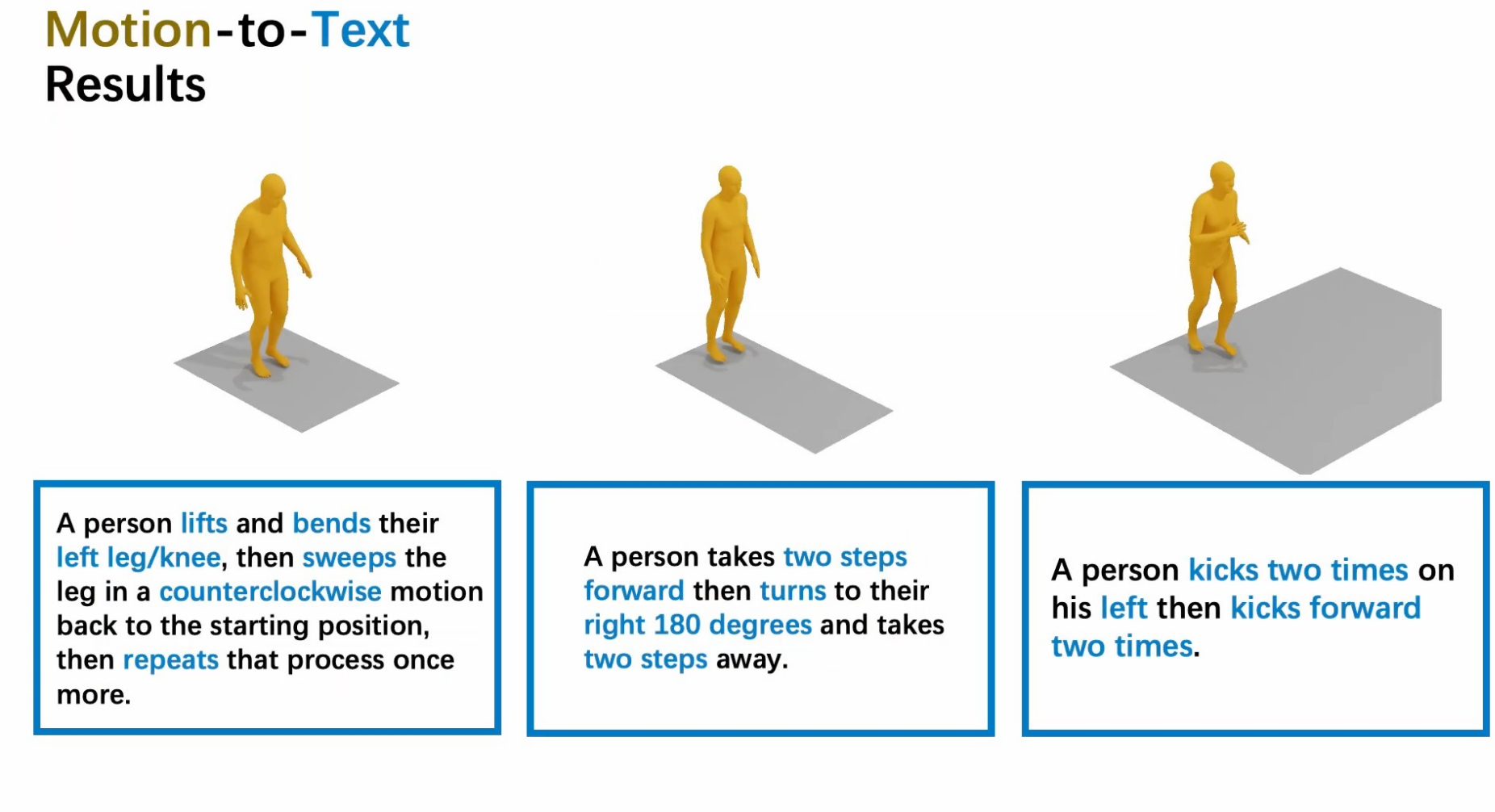
画像・動画は論文/デモページより引用

何が出来るか？



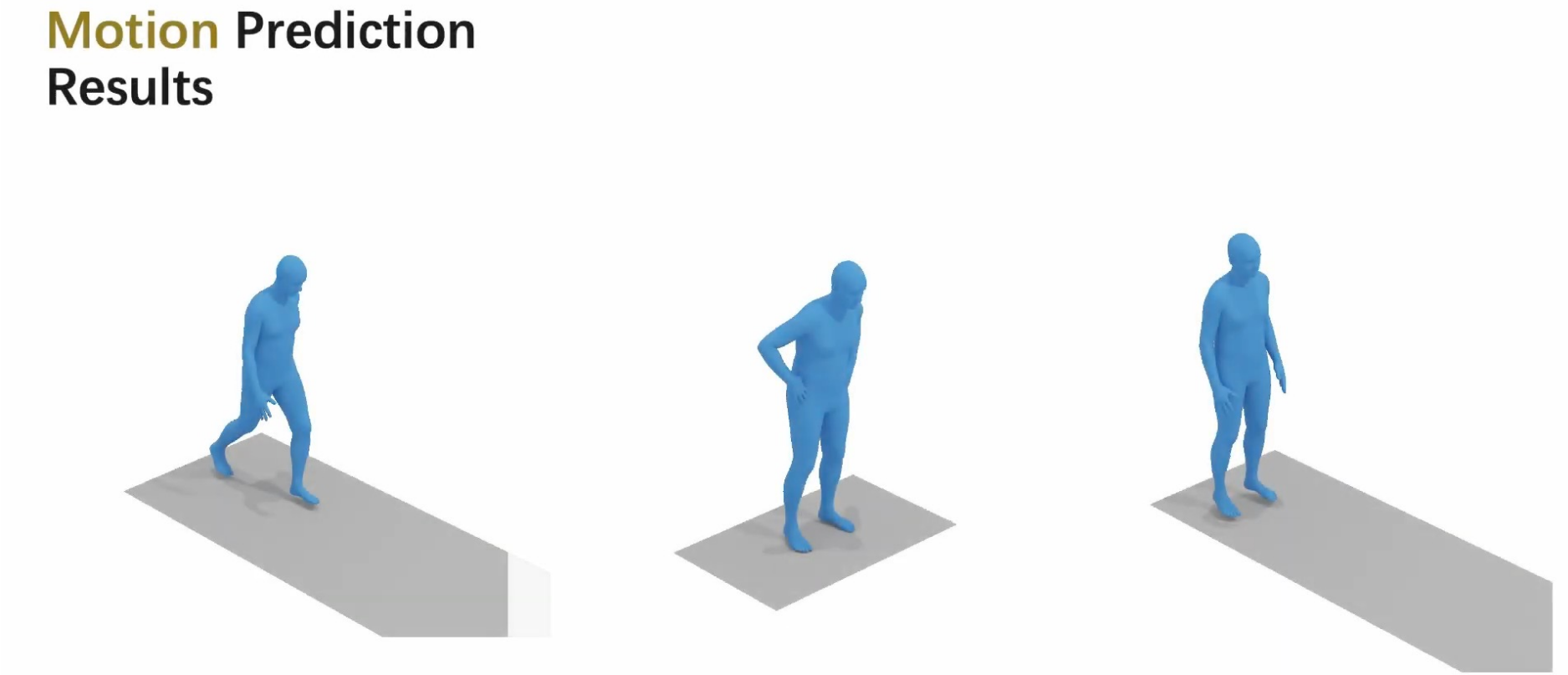
画像・動画は論文/デモページより引用

何が出来るか？



画像・動画は論文/デモページより引用

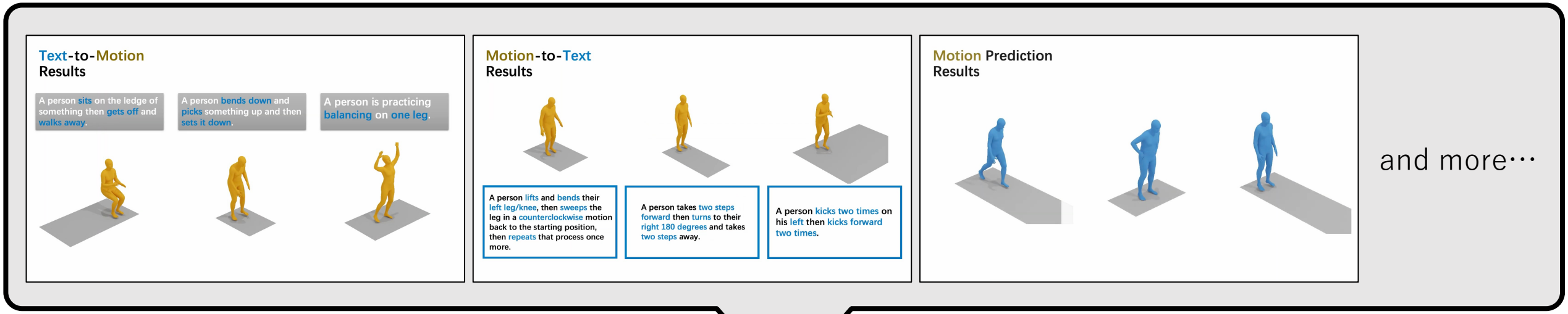
何が出来るか？



画像・動画は論文/デモページより引用

何が出来るか？

15種類のtext-languageタスク



Motion GPT

全てのタスクを一つのモデルで実現

性能

定量評価

Methods	Text-to-Motion			Motion-to-Text			Motion Prediction		Motion In-between	
	R TOP1↑	FID↓	DIV→	R TOP3↑	Bleu@4↑	Cider↑	FID↓	DIV→	FID↓	DIV→
Real	$0.511^{\pm.003}$	$0.002^{\pm.000}$	$9.503^{\pm.065}$	0.828	-	-	0.002	9.503	0.002	9.503
MLD [54]	$0.481^{\pm.003}$	$0.473^{\pm.013}$	$9.724^{\pm.082}$	-	-	-	-	-	-	-
T2M-GPT [48]	$0.491^{\pm.003}$	$\mathbf{0.116}^{\pm.004}$	$9.761^{\pm.081}$	-	-	-	-	-	-	-
TM2T [12]	$0.424^{\pm.017}$	$1.501^{\pm.003}$	$8.589^{\pm.076}$	0.823	7.00	16.8	-	-	-	-
MDM [48]	$0.320^{\pm005}$	$0.544^{\pm.044}$	$9.559^{\pm.086}$	-	-	-	6.031	7.813	2.698	8.420
MotionGPT (Ours)	$\mathbf{0.492}^{\pm.003}$	$0.232^{\pm.008}$	$\mathbf{9.528}^{\pm.071}$	**0.827**	**12.47**	**29.2**	**0.905**	**8.972**	**0.214**	**9.560**

Table 2: Comparison of four motion-related tasks on HumanML3D [11] dataset. The evaluation metrics are computed using the encoder introduced in [11]. The empty columns of previous methods indicate that they can not handle the task. The arrows (→) indicate that closer to *Real* is desirable. **Bold** and underline indicate the best and the second best result on text-to-motion task.

全タスクにおいて性能が良い
（めでたし　めでたし）

これ以上の情報は無いので、
以降、実験の説明はしない

Human Evaluation

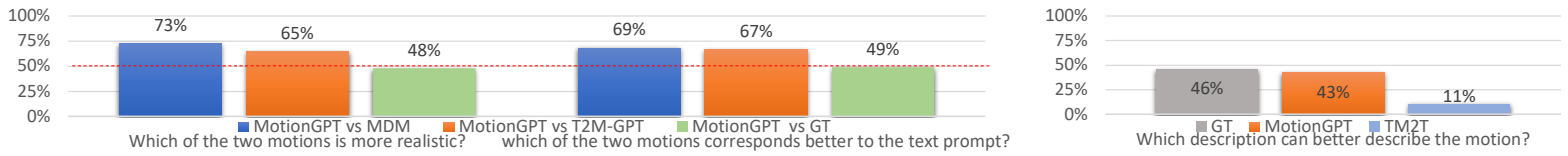


Figure 8: User Study. We investigate our motion quality and the alignment with test descriptions. The left part is the user study for text-to-motion. The right part is for motion captioning.

一つのモデルで複数のタスクが
解けることを明らかにした

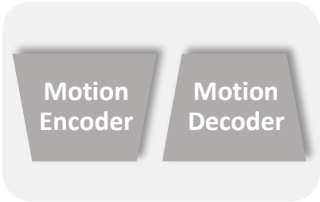
どう実装するか？

Step 1

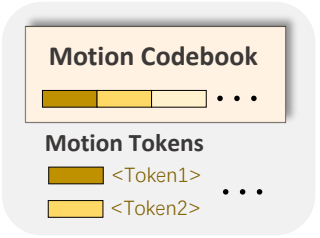
Training of Motion Tokenizer.

A motion sequence is sampled from 3D motion dataset.

Motion tokenizer learns motion representation.



Motion codebook is used to represent human motion as discrete tokens.

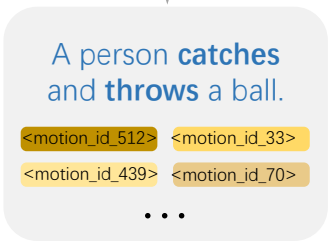


Step 2

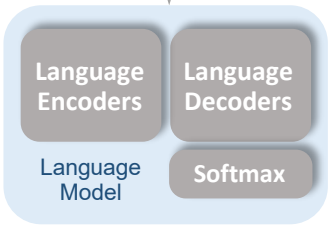
Motion-language Pre-training.

A motion and a language description are sampled.

This motion is mapped to discrete motion indices and mixed with words.

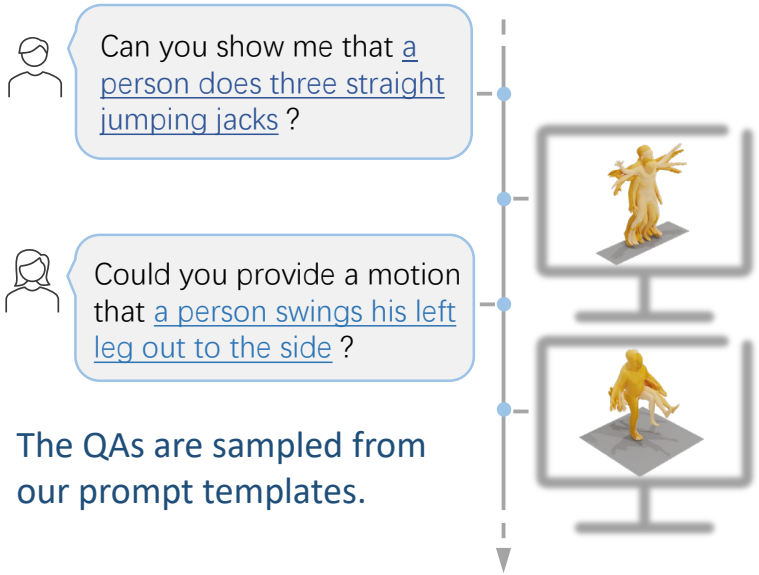


This data is used to pre-train our motion-language model.



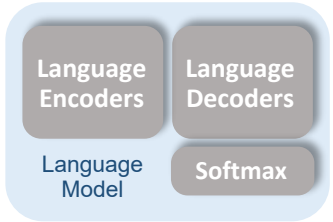
Step 3

Instruction Tuning.



The QAs are sampled from our prompt templates.

The prompts are used to fine-tune our model on diverse motion tasks



どう実装するか？

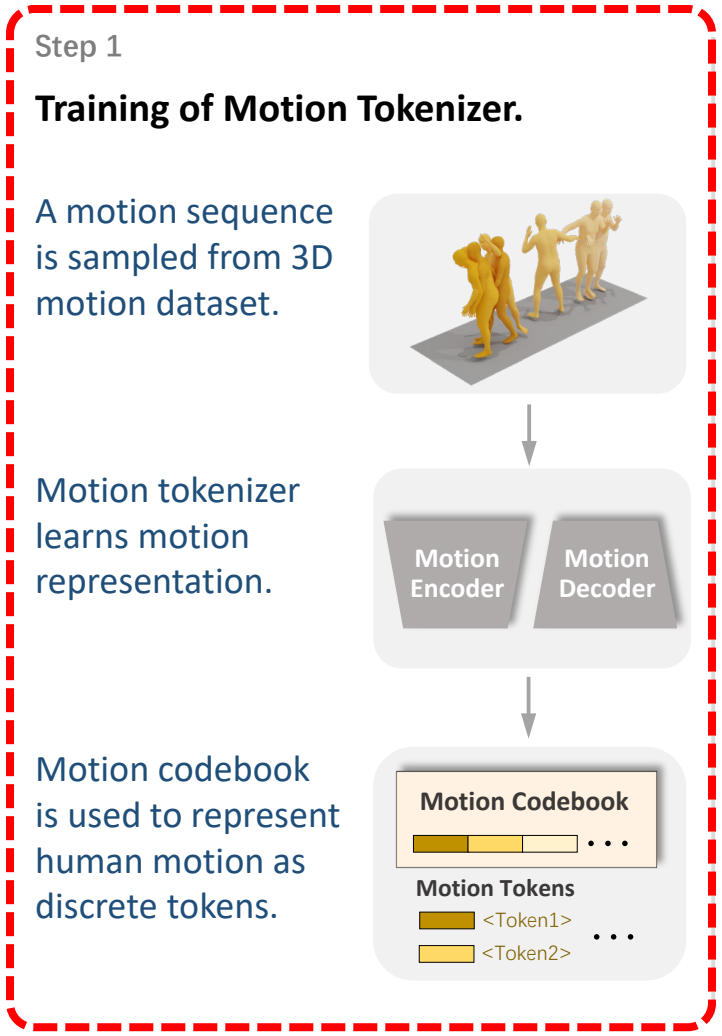


Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks (CVPR2023) から着想

キーアイディア

モーションデータを疑似単語へと変換して
自然言語と一緒に言語モデルを学習すれば
モーションと言語の取り扱いが一緒にでき
一つのモデルで様々なタスクに対応可能に

具体的な実装法

VQ-VAEでモーションを量子化（疑似単語変換）
デコーダで疑似単語からモーション合成も可能

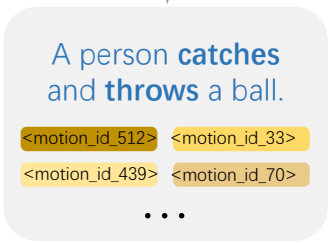
どう実装するか？

Step 2

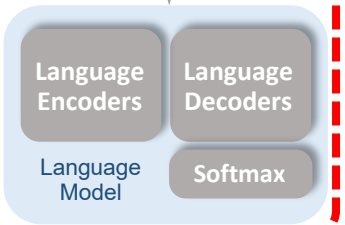
Motion-language Pre-training.

A motion and a language description are sampled.

This motion is mapped to discrete motion indices and mixed with words.



This data is used to pre-train our motion-language model.



モーション-言語の関係学習

1. 大量テキストで事前学習されたT5モデルに対し、
2. 量子化されたモーションとテキストペアを学習

HumanML3D データセット:
14,616のモーションと対応する44,970文のText Description。

ユーザに優しくないテキストフォーマット

どう実装するか？

Instruction Tuning

1. 15種類のモーション-テキストタスクを定義
2. 各タスクのInstruction テンプレートを作成

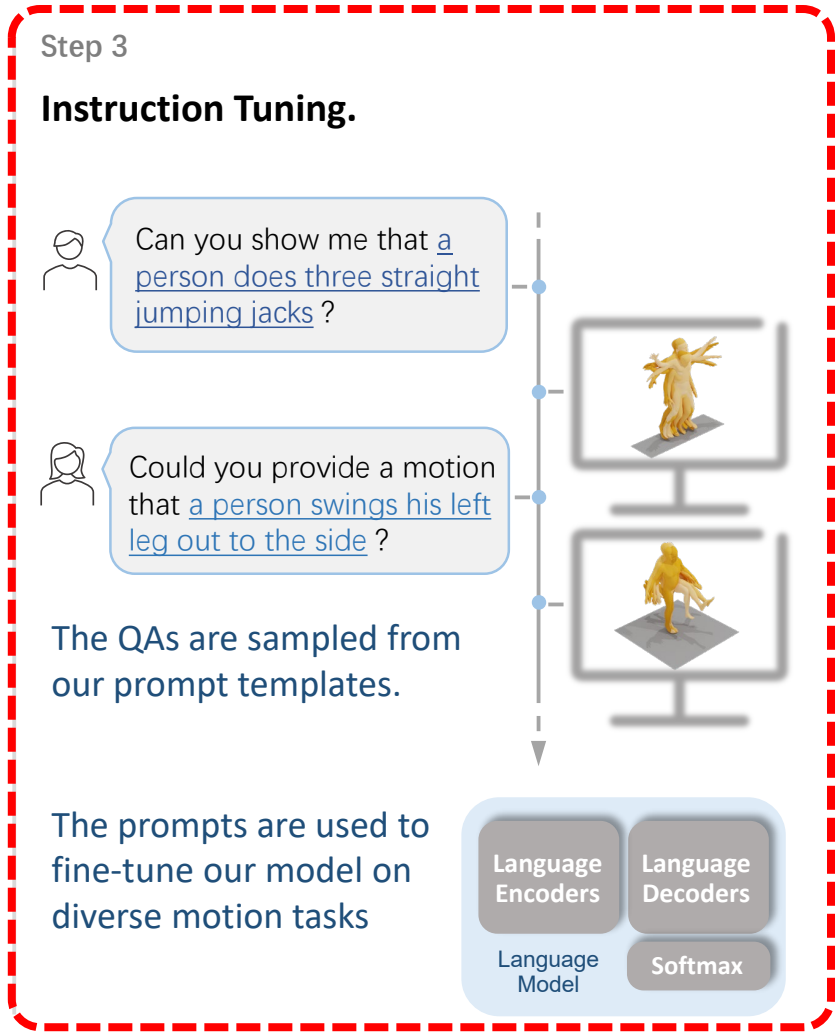
例: Text-to-motion: 青字がtext description
Can you generate a motion sequence that depicts ‘a person emulates the motions of a waltz dance’?

例: Motion-to-text: 青字が量子化されたモーショントークン
Provide an accurate caption describing the motion of <motion_tokens>

テンプレートは公開済み: https://github.com/OpenMotionLab/MotionGPT/blob/main/prepare/instructions/template_instructions.json

3. Instruction テンプレートを使い学習データを作成し、T5モデルをFine-tuneする

Motion GPT 完成



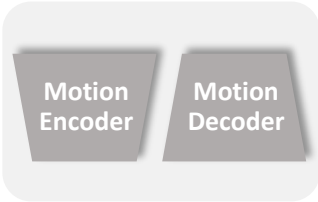
どう実装するか？

Step 1

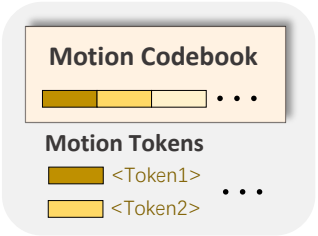
Training of Motion Tokenizer.

A motion sequence is sampled from 3D motion dataset.

Motion tokenizer learns motion representation.



Motion codebook is used to represent human motion as discrete tokens.

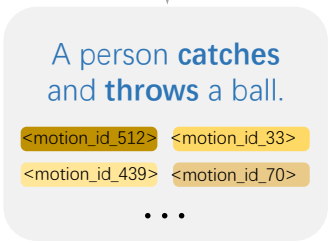


Step 2

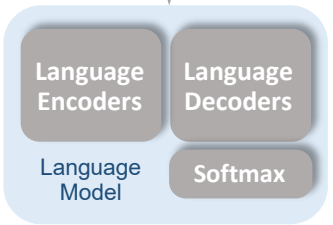
Motion-language Pre-training.

A motion and a language description are sampled.

This motion is mapped to discrete motion indices and mixed with words.

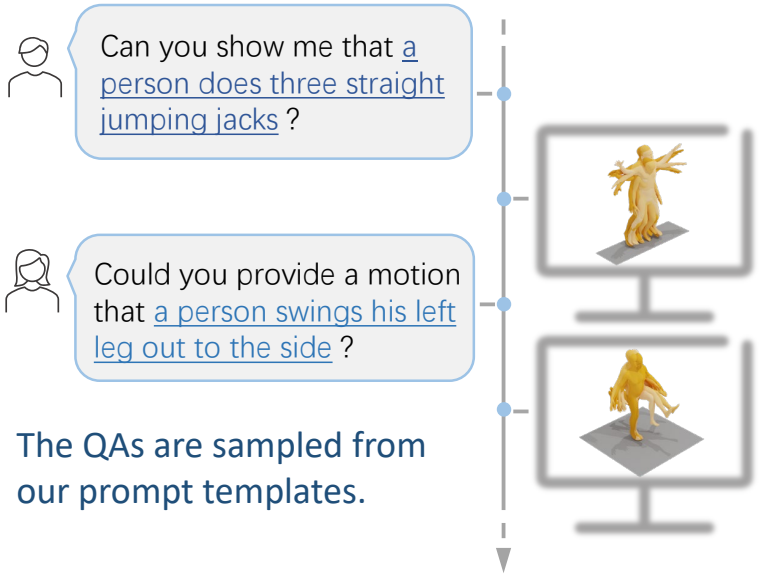


This data is used to pre-train our motion-language model.



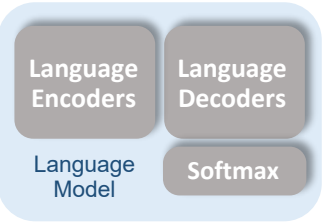
Step 3

Instruction Tuning.



The QAs are sampled from our prompt templates.

The prompts are used to fine-tune our model on diverse motion tasks



NLPerにとって未知の領域

NLPerにとって特に目新しい技術はない

モーションデータってどんなデータ？

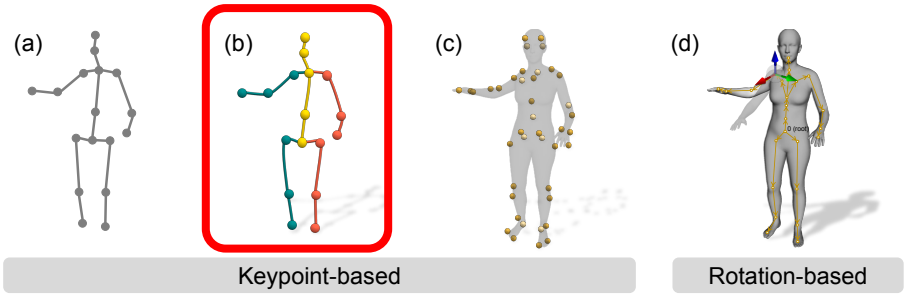


Fig. 3: Typical human pose and shape representations with the same pose in (a) 2D keypoints, (b) 3D keypoints, (c) 3D marker keypoints, and (d) rotation-based model.

3Dキーポイント表現

- 24~53くらいの関節で構成された骨格構造
- 各関節ごとに3D空間上のxyz座標、関節角度
- 30FPS/60FPSのデータ

モーションデータってどんなデータ？

Name	Venue	Collection	Representation	Subjects	Sequences	Frames	Length	Condition	Remarks
Human3.6M [39]	TPAMI 2014	Marker-based	Kpts. (3D)	11	-	3.6M	5.0h	-	15 actions
CMU Mocap [65]	Online 2015	Marker-based	Rot.	109	2605	-	9h	-	6 categories, 23 subcategories
AMASS [40]	ICCV 2019	Marker-based	Rot.	344	11265	-	40.0h	-	Unifies 15 marker-based MoCap datasets
HuMMan [42]	ECCV 2022	Markerless	Rot.	1000	400K	60M	-	-	500 actions
KIT Motion Language [100]	Big data 2016	Marker-based	Kpts. (3D)	111	3911	-	10.3h	Text	6.3k Text descriptions
UESTC [91]	MM 2018	Markerless	Kpts. (3D)	118	25.6K	-	83h	Text	40 Action classes
NTU-RGB+D [87]	TPAMI 2019	Markerless	Kpts. (3D)	106	114.4K	-	74h	Text	120 Action classes
HumanAct12 [7]	MM 2020	Markerless	Kpts. (3D)	12	1191	90K	6h	Text	12 Action classes
BABEL [96]	CVPR 2021	Marker-based	Rot.	344	-	-	43.5h	Text	260 Action classes
HumanML3D [3]	CVPR 2022	Marker-based & Markerless	Kpts. (3D)	344	14.6K	-	28.5h	Text	44.9K Text descriptions
Tang *et al.* [117]	MM 2018	Marker-based	Kpts. (3D)	-	61	907K	1.6h	Music	4 genres
Lee *et al.* [118]	NeurIPS 2019	Pseudo-labeling	Kpts. (2D)	-	361K	-	71h	Music	3 genres
Huang *et al.* [119]	ICLR 2021	Pseudo-labeling	Kpts. (2D)	-	790	-	12h	Music	3 genres
AIST++ [115]	ICCV 2021	Markerless	Rot.	30	1,408	10.1M	5.2h	Music	10 genres
PMSD [120]	TOG 2021	Marker-based	Kpts. (3D)	8	-	-	3.8h	Music	4 genres
ShaderMotion [120]	TOG 2021	Marker-based	Kpts. (3D)	8	-	-	10.2h	Music	2 genres
Chen *et al.* [86]	TOG 2021	Manual annotation	Rot.	-	-	160K	9.9h	Music	9 genres
PhantomDance [123]	AAAI 2022	Manual annotation	Rot.	-	260	795K	3.7h	Music	13 genres
MMD-ARC [8]	MM 2022	Manual annotation	Rot.	-	213	-	11.3h	Music	-
MDC [126]	MM 2022	Manual annotation	Rot.	-	798	-	3.5h	Music	2 genres
Aristidou *et al.* [128]	TVCG 2022	Marker-based	Rot.	32	-	-	2.4h	Music	3 genres
AIOZ-GDANCE [129]	CVPR 2023	Pseudo-labeling	Rot.	>4000	-	-	16.7h	Music	7 dance styles, 16 music genres
Trinity [134]	IVA 2018	Pseudo-labeling	Kpts. (2D)	1	23	-	4.1h	Speech	Casual talks
TED-Gesture [136]	ICRA 2019	Pseudo-labeling	Kpts. (3D)	-	1,295	-	52.7h	Text	TED talks
Speech2Gesture [130]	CVPR 2019	Pseudo-labeling	Kpts. (2D)	10	-	-	144h	Speech	TV shows, Lectures
TED-Gesture++ [135]	TOG 2020	Pseudo-labeling	Kpts. (3D)	-	1,766	-	97.0h	Speech, Text	Extension of [136]
PATS [132]	ECCV 2020	Pseudo-labeling	Kpts. (2D)	25	-	-	251h	Speech, Text	Extension of [130]
Speech2Gesture-3D [137]	IVA 2021	Pseudo-labeling	Kpts. (3D)	6	-	-	33h	Speech	Videos from [130]
BEAT [144]	ECCV 2022	Marker-based	Rot.	30	2508	30M	76h	Speech, Text, Emotion	8 emotions, 4 languages
Chinese Gesture [146]	TOG 2022	Marker-based	Rot.	5	-	-	4h	Speech, Text	Chinese
ZEGGS [150]	CGF 2023	Marker-based	Rot.	1	67	-	2.3h	Speech, Style	19 Styles
SHOW [148]	CVPR 2023	Pseudo-labeling	Rot.	-	-	-	27h	Speech	Videos from [130]
WBHM [153]	ICAR 2015	Marker-based	Rot.	43	3704	691K	7.68h	Object	41 objects
PiGraph [182]	TOG 2016	Markerless	Kpts. (3D)	5	63	0.1M	2h	Scene, Object	30 scenes, 19 objects
PROX [155]	ICCV 2019	Markerless	Rot.	20	60	0.1M	1h	Scene, Object	12 indoor scenes
i3DB [159]	SIGGRAPH 2019	Pseudo-labeling	Kpts. (3D)	1	-	-	-	Scene, Object	15 scenes
GTA-IM [154]	ECCV 2020	Marker-based	Kpts. (3D)	50	119	1M	-	Scene	Synthetic, 10 indoor scenes
GRAB [97]	ECCV 2020	Marker-based	Rot.	10	1334	1.6M	-	Object	51 objects
HPS [187]	CVPR 2021	Marker-based	Rot.	7	-	300K	-	Scene	8 large scenes, some > 1000 m^2
SAMP [160]	ICCV 2021	Marker-based	Rot.	1	-	185K	0.83h	Scene, Object	7 objects
COUCH [66]	ECCV 2022	Marker-based	Rot.	6	>500	-	3h	Scene, Chairs	3 chairs, hand interaction on chairs
HUMANISE [16]	NeurIPS 2022	Marker-based	Rot.	-	19.6K	1.2M	-	Scene, Object, Text	643 scenes
CIRCLE [168]	CVPR 2023	Marker-based	Rot.	5	>7K	4.3M	10h	Scene	9 scenes

モーションデータだけでも
これだけある

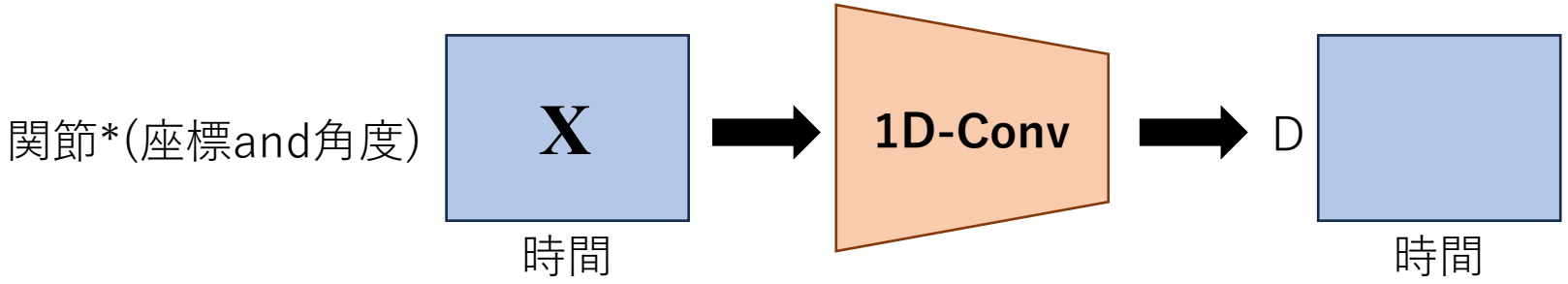
近年になって利用可能な
データが増えてきた

最近、**Motion-X**という8万以上のテキストとモーションのペアデータセットが出てきた。
(Motion-X++も出てくる予定)

Motion-X: A Large-scale 3D Expressive Whole-body Human Motion Dataset, Lin+ , NeurIPS2024

モーションデータをどうエンコードする？

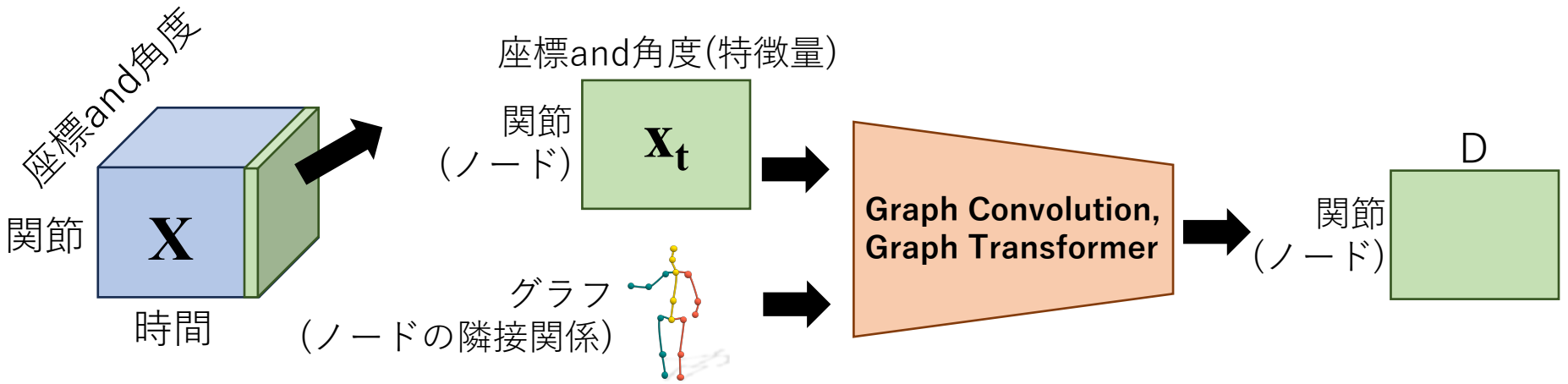
大抵は1D-Convolutionで、「関節*(座標and角度)」次元をチャンネルとして畳み込む



- 実装が楽
- 骨格構造は教えていない
  勝手に学習してるとも言える

Motion-GPTは1D-Convでエンコードしている

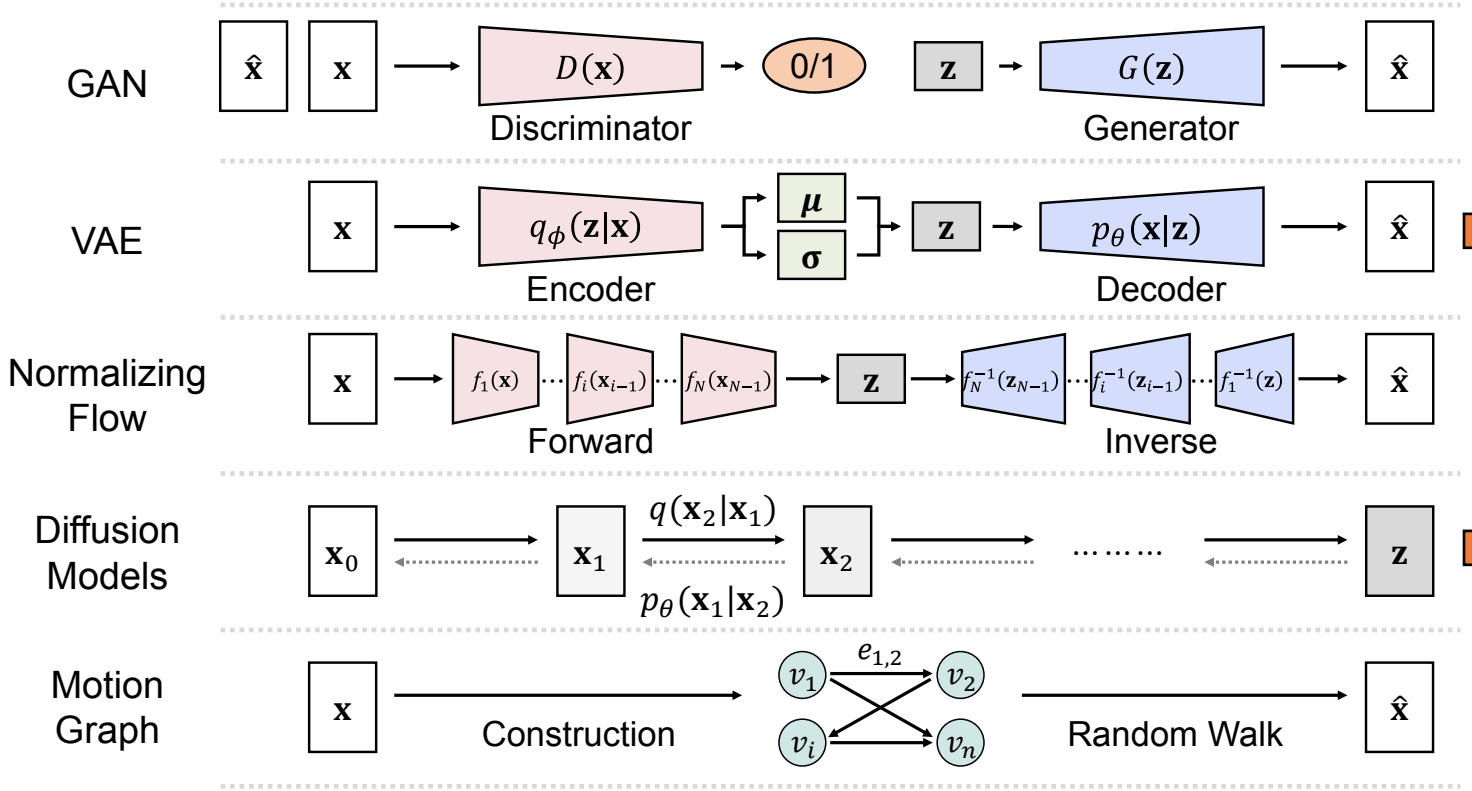
骨格構造をグラフ構造(関節をノード)とみなし、Graph ConvolutionやGraph Transformerを使う



- 実装は若干複雑

モーションデータをどう生成(デコード)する？

基本的には画像生成を踏襲している



この研究ではVQ-VAEで量子化と
エンコード/デコードを兼ねている

→言語処理手法で扱うには妥当な選択

今のトレンド、画像生成で成功し
たように、モーション生成も成功

Fig. 5: An overview of different generative models.

まとめ

- 言語と現実世界を経由するための「モーション」という方向性
- 「モーションデータ」の扱い方の簡単な導入
- Motion-GPT： 先端であるにも関わらず簡単な実装
  利用可能なデータも増えている
  ＝ 参入しやすい
- 鍵は量子化：シンボルにしてしまえば、NLPの領域
  - 調査されていないこと/気になること
    - モーションにも文法や、言語に共通する性質があるのか？
    - モーションと一緒に学習した言語モデルは物理法則を学習しているのか？
      - 「ゆっくり-速い」「右-左」などの速度や方向、大きさなどがどう特徴量化されている？
    - 現状は「モーション」と「モーションの説明文」ペアのみ
      - 別種のモーションとテキストペアの可能性
- モーションがNLPを「コンピュータの外」へ連れ出すきっかけになる？