

# MonSTeR : a Unified Model for Motion, Scene, Text Retrieval

Luca Collorone<sup>1,\*</sup>, Matteo Gioia<sup>1,\*</sup>, Massimiliano Pappa<sup>1</sup>,  
 Paolo Leoni<sup>1</sup>, Giovanni Ficarra<sup>1,3</sup>, Or Litany<sup>2</sup>, Indro Spinelli<sup>1</sup>, Fabio Galasso<sup>1</sup>

<sup>1</sup> Sapienza University of Rome

<sup>2</sup> Technion, NVIDIA

<sup>3</sup> WSense

{name.surname}@uniroma1.it or.litany@gmail.com

## Abstract

*Intention drives human movement in complex environments, but such movement can only happen if the surrounding context supports it. Despite the intuitive nature of this mechanism, existing research has not yet provided tools to evaluate the alignment between skeletal movement (**motion**), intention (**text**), and the surrounding context (**scene**).*

*In this work, we introduce MonSTeR, the first MOTion-Scene-TExt Retrieval model. Inspired by the modeling of higher-order relations, MonSTeR constructs a unified latent space by leveraging unimodal and cross-modal representations. This allows MonSTeR to capture the intricate dependencies between modalities, enabling flexible but robust retrieval across various tasks.*

*Our results show that MonSTeR outperforms trimodal models that rely solely on unimodal representations. Furthermore, we validate the alignment of our retrieval scores with human preferences through a dedicated user study. We demonstrate the versatility of MonSTeR’s latent space on zero-shot in-Scene Object Placement and Motion Captioning. Code and pre-trained models are available at [github.com/collaroneluca/MonSTeR](https://github.com/collaroneluca/MonSTeR).*

## 1. Introduction

*“She asked me to stay and she told me to sit anywhere. But I looked around and I noticed there wasn’t a chair.”*

— Norwegian Wood, The Beatles

Humans can navigate and interact with environments by balancing their intentions with the opportunities presented by their surroundings. For instance, when the intended action is to “sit on a chair” and the corresponding movement is performed, it would be perplexing if the environment lacked a chair (Fig 1). This highlights the necessity for strong **coherence** and plausibility among intentions (text),

\* Authors contributed equally.



Figure 1. MonSTeR can estimate coherence among text, motion, and scene by embedding them into a unified latent space. In the left image, all three modalities are coherent. However, in the right image, this coherence decreases, as there is no chair available in the scene.

movements (motion), and environments (scene) to create a scenario where human-scene interaction can naturally take place. Also, multiple captions could potentially describe the motion in Fig. 1 in isolation, but given the scene context, it is evident that the provided one is the most accurate. This confirms that exists a strong **interdependence** between these modalities.

However, existing human motion generation and retrieval approaches are far from effectively leveraging this trimodal relation and fully capturing the interdependence between texts, motions and the environment in which they occur. In fact, recent retrieval models [10, 13, 27, 39] are unable to incorporate environmental context, causing motions to be depicted in an empty space and overlooking their intrinsic connection to the scene. Similarly, the methods for evaluating Human Scene Interaction models fail to provide a global measure of coherence/realism [4, 34, 35]. In fact, they try to obtain a score based on a number of different metrics, including person-scene interpenetration or distance to the target location. By contrast, we argue that these eval-

uations should take into account other critical aspects often neglected, such as the path taken, the plausibility of the motion, and the coherence among the three modalities.

Thus motivated, we introduce the first MOtioN-Scene-Text Retrieval model, which we call MonSTeR. This model is designed with a unified latent space that enables the assessment of coherence across three modalities: text ( $t$ ), motion ( $m$ ), and scene ( $s$ ). Motivated by recent progress in topological deep learning [25], we model inter-modal dependencies using higher-order interactions that go beyond pairwise relations. Unimodal terms, encoded by variational encoders, can be represented as nodes in a graph describing their interaction. In the same graph, we use cross-modal encoders to embed pairs of modalities,  $ts$ ,  $ms$ ,  $mt$ , representing its edges. To capture the higher-order interactions, we align unimodal and unimodal-bimodal representations:  $(st, m)$  aligns the scene-text cross-modal representation with motion,  $(mt, s)$  aligns the motion-text one with the scene, and  $(ms, t)$  aligns the scene-motion one with text. The proposed modeling produces cohesive representations that deliver superior retrieval performance compared with approaches that align the three modalities using only unimodal terms. Furthermore, it provides the flexibility to perform various related tasks: retrieving a sample from one modality given another, a sample from one modality given a representation of two modalities, or vice versa.

Additionally, we show that MonSTeR can serve as an evaluator for text-conditioned Human Scene Interaction models. Specifically, we conduct (1) a dedicated user study confirming that its assessments align with human judgment, and (2) an evaluation revealing that MonSTeR exhibits scene-motion grounding capabilities by assigning lower scores to motions that follow incorrect paths or interpenetrate the scene, confirming that our method can be used to assess path-scene plausibility. Finally, we leverage the descriptiveness of the latent representations by performing Motion Captioning and zero-shot in-Scene Object Placement.

In summary, our contributions include:

- A new retrieval model that explicitly models higher-order relations to unify text, motion, and scene, for the first time, within a unique latent space.
- A new method to ground the generations of text-conditioned Human Scene Interaction models within the scene and consequently evaluate their quality.
- Zero-shot evaluation of our model on the tasks of in-Scene Object Placement and evaluation on the downstream task of Motion Captioning.

## 2. Related Works

This section introduces existing text-to-motion retrieval models, multimodal alignment, and aggregation techniques.

### 2.1. Text-Motion Retrieval

Text-to-motion retrieval involves discriminating which motion sequence corresponds to a given text ( $t2m$  task) and, conversely, identifying the correct text description for a query motion sequence ( $m2t$  task). These tasks are often modeled using Contrastive Learning: a shared latent space is constructed between text and motion, where cosine similarity serves as a measure of coherence between the two modalities [1, 10, 13, 27, 32, 39]. In particular, [32] initiated this field by aligning texts, motions, and image renderings of the motions. Additionally, [1, 27] proposed an effective and lightweight model to address this task and introduced several evaluation protocols for benchmarking. Existing approaches [27] represented motions through velocities and rotations in root space [13] or tackled this by encoding motions as image patches [39] processed via a pre-trained Vision Transformer (ViT). Unlike previous work, we offer a latent space that can estimate coherence not only between text and motion but also with the scene itself.

### 2.2. Beyond Two-Modal Alignment

Recent works extending beyond text and motion have proposed various strategies for aligning multiple modalities within a shared latent space. Specifically, [11, 20, 21, 36, 37] align all modalities to one modality (or a subset) chosen as a reference. Alternatively, [22, 30, 38] align each modality with all the others, creating intricate interactions among modalities. Particularly relevant to our task, [38] presents a trimodal learning framework that incorporates human-centric videos, motion, and textual instructions.

In this work, we explore both approaches. Since MonSTeR is designed to harness the interdependence between modalities, it is trained by aligning individual modality representations and paired modalities, achieving an all-to-all alignment. We validate our approach through an ablation study, aligning scene and motion solely with text, showing degraded performance compared to MonSTeR.

### 2.3. Modality Aggregation

Early approaches used separate encoders for each modality, combining them via latent averaging [11, 30], or adopted conditional generation—e.g., generating scenes from motion [3, 9, 23] or motion from scenes [18, 19]. However, these approaches do not explicitly model modality interdependence in the latent space, limiting the ability to learn joint representations. Other strategies, such as those in [5, 12], process all modalities through a shared encoder to exploit cross-modal interactions, yet still produce separate embeddings per modality. This prevents the model from representing multiple modalities jointly or retrieving a single modality from a composite latent representation. Additionally, [28] aligns global (multi-modal) and per-modality encoders, but lacks the flexibility to encode partial modal-

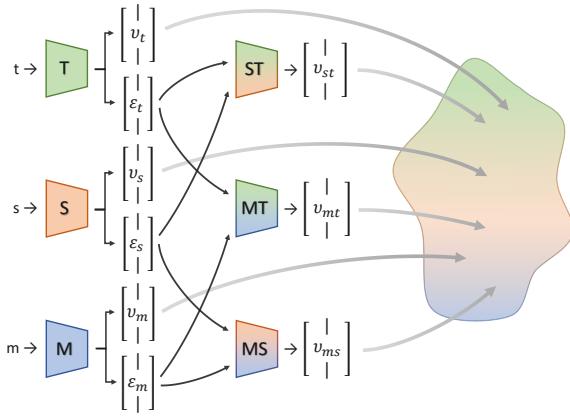


Figure 2. MonSTeR’s Architecture Overview. Each input modality  $t, s, m$  is processed by its single-modality encoder. From the first output tokens of  $T, M, S$  we sample vectors  $v_t, v_m$ , and  $v_s$ . The remaining tokens of each encoder’s output, namely  $\varepsilon_t, \varepsilon_m, \varepsilon_s$ , are pairwise concatenated and passed through cross-modal encoders to generate joint latent vectors ( $v_{st}, v_{mt}, v_{ms}$ ).

ity combinations into a unified representation. Finally, [7] encodes pose, image, and instruction together into a unified embedding and can reconstruct the original modalities. In contrast, MonSTeR promotes cross-modal learning by explicitly combining unimodal and pairwise cross-modal encoders, which provides richer representations that span all the available modalities.

### 3. MonSTeR

In this section, we present the design of MonSTeR. We start with data representation, then detail the high-order decomposition that informed our latent space, and conclude with the chosen optimization objective.

#### 3.1. Data Representation

Following [27], we use DistilBERT [31] to obtain an initial representation for text  $t$ , in the form of a 768-dimensional feature vector, i.e.  $t \in \mathbb{R}^{768}$ . As in [15, 40], the scene  $s$  is treated as a colored point cloud of  $N$  points containing objects with semantic information, i.e.  $s \in \mathbb{R}^{N \times 6}$ , where the six features are the concatenation of the  $(x, y, z)$  coordinates and the RGB colors. We represent the 3D human motion  $m$  as vectors  $m \in \mathbb{R}^{T \times 3 \times 22}$ , where  $T$  is the number of frames, 3 the  $x, y, z$  coordinates of joints, and 22 the number of joints.

#### 3.2. Designing MonSTeR’s latent space

Aligning scene, motion, and text into a unified multimodal latent manifold is a challenging goal. The complexity arises from the need to capture the many-to-many relationship among these modalities: a single text may correspond to

various motions, and the appropriate motion may differ depending on the scene context. The triplet (scene, motion, text) defines an event with minimal ambiguity<sup>1</sup>. Unlike pairwise alignments, this triplet captures higher-order interactions between the modalities, providing a richer and more coherent representation of the underlying event [27].

**Higher-Order Relations in Topological Spaces.** Topology provides a robust framework to analyze and extract insights from data [2, 25]. Therefore, we set to exploit this theory to describe the process of embedding text, scene, and motion in a unified latent space via contrastive learning. This requires modeling *part-whole* relations and it is therefore different from other contrastive objectives targeting pairwise relation between unimodal terms. Turning to topological geometry, this three-way relation represented by a polygonal face  $\mathcal{P} = \{tsm\}$  can be decomposed into the composition of the three edges that are a *part* of the triangle  $\mathcal{E} = \{ts, sm, mt\}$ . Similarly, each edge can be decomposed into the vertices that define it, resulting in a set of vertices  $\mathcal{V} = \{t, s, m\}$ . By aligning unimodal terms, we encode their pairwise relationship in the latent space. On the other hand, by aligning an edge with the opposite vertex, we encode their higher-order relation. All in all, topology [26] demonstrates that representing the three-way relation of  $(s, m, t)$  requires representing the single modes  $\{s, m, t\}$  and the pairwise cross-modal terms  $\{sm, mt, ts\}$ . We verify this empirically, cf. Sec. 4.3.

**Contrastive Learning to Learn High Order Relations.** We align unimodal to unimodal and unimodal to cross-modal terms via contrastive learning. The goal is to embed this trimodal interaction in the latent space. Following the schematics of Fig. 2, we encode the unimodal terms  $\{t, m, s\}$  using transformer-based variational autoencoders. The output of each encoder is split into two: (1) the first two tokens are interpreted as the mean and log-variance of the latent distribution, and they are used to sample latent vectors  $v_t, v_m$ , and  $v_s$ . (2) The rest of these tokens, namely  $\varepsilon_t, \varepsilon_m, \varepsilon_s$ , are pairwise concatenated and processed by cross-modal encoders (e.g.  $\varepsilon_s$  and  $\varepsilon_t$  are concatenated to obtain the input of the  $ST$  encoder). Each one of the three cross-modal encoders outputs mean and log-variance used to sample the cross-modal latent vectors  $v_{st}, v_{mt}$ , and  $v_{ms}$ . This process and the architecture of its encoders are further detailed in the Supplementary Material.

#### 3.3. Training MonSTeR

Contrastive objectives are about maximizing the similarity between positive pairs, or different modalities representing the same object. In our objective, we include all the terms needed to recover unimodal and cross-modal relations of the three modalities. We exclude terms that could encourage

<sup>1</sup>Some ambiguity still exists, as underlying intentions and personal preferences influence human movement

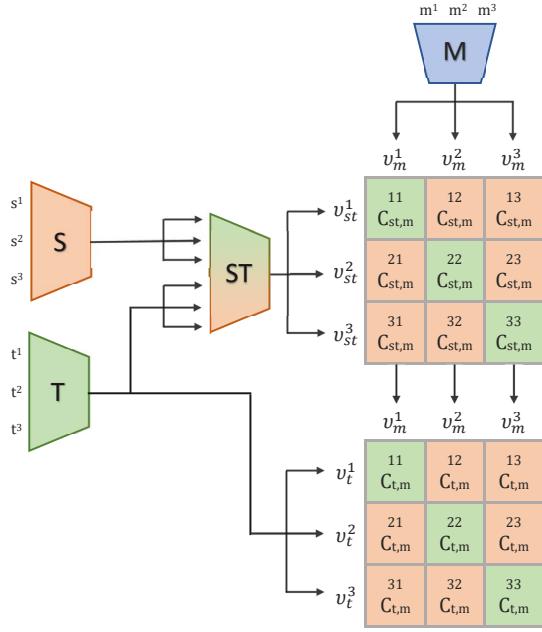


Figure 3. Composition of  $C_{st,m}$  and  $C_{t,m}$  similarity matrices. For  $C_{st,m}$ , motion latents  $v_m^n$  are compared with the cross-modal scene-text latents  $v_{st}^n$ , while for  $C_{t,m}$ , they are compared only with text latents  $v_t^n$ . Green cells are the locations of similarity scores between positive samples. Our optimization objective promotes assigning higher scores to these cells in all matrices  $C_{i,j}$ .

the model to ignore one of the modalities during alignment, leading to degenerate solutions, e.g.,  $(st, t)$  or  $(st, s)$ . This is to prevent a collapse in the cross-modal encoders, that could simply learn an identity function. This leaves us with the resulting terms for the contrastive objective:

$$K = \{(t, s), (m, t), (m, s), (st, m), (mt, s), (ms, t)\}$$

Once we collect all the latent vectors for all the modalities in a batch, we proceed with the alignment procedure. Formally, for a batch of  $N$  tuples of latent vectors  $(v_t^1, v_m^1, v_s^1, v_{st}^1, v_{mt}^1, v_{ms}^1), \dots, (v_t^N, v_m^N, v_s^N, v_{st}^N, v_{mt}^N, v_{ms}^N)$ , we compute the  $N \times N$  cosine similarity matrices  $C_{i,j}$  for each pair of unimodal or cross-modal vectors  $(i, j) \in K$ . Positive pairs appear on the main diagonal, while the off-diagonal terms represent negative pairs. For clarity, Fig. 3 illustrates the composition process of two example similarity matrices:  $C_{st,m}$  which aligns unimodal to cross-modal embeddings and  $C_{t,m}$  which relies solely on unimodal ones. Each similarity matrix is then used to compute the InfoNCE loss [24], which we aggregate in the final loss:

$$\mathcal{L}_{\text{tot}} = \frac{1}{|K|} \sum_{(i,j) \in K} \frac{\mathcal{L}_{\text{NCE}}(C_{i,j})}{N}$$

This process pushes the model to bring the positive vectors (those belonging to the same tuple) closer while distancing them from other vectors in the batch.

**Cross-Modal Retrieval.** The unified latent space enables retrieval of scenes, motions, and texts, as well as their combinations, using either single modalities or by specifying two of them. For instance, to retrieve the correct motion given a scene and text ( $st2m$  task), we compute the scene-text cross-modal embedding using the  $ST$  encoder in Fig 2, and then calculate its cosine similarity with the outputs of the  $M$  encoder.

## 4. Experiments

In this section, we describe the evaluation datasets, retrieval protocols, and MonSTeR ablations, as well as its application to Human-Scene Interaction motion generation. We also perform In-Scene Object Placement and Motion Captioning tasks to demonstrate the effectiveness of the learned representations.

### 4.1. Datasets

To benchmark MonSTeR, we use two recent datasets, [17, 34]. We select these datasets as they are large-scale options that feature a substantial variety of motions across numerous scenes, all accompanied by textual annotations.

**HUMANISE+.** HUMANISE [34] is a synthetic dataset containing 19.6K text-annotated human motion carefully aligned in 643 3D scenes.

The text descriptions in this dataset primarily reference an object of interaction, omitting details about its spatial location, attributes, and other objects in the scene. When multiple instances of the object exist, the descriptions fail to specify which one is relevant to the associated motions. To address this limitation, we use LLAMA3 [8] to recaption the original texts to include references to nearby objects and express spatial context around the motion path. This approach substantially enhances the grounding of text descriptions within the scene. A detailed description of this process is provided in the Supplementary Material. We refer to this dataset as HUMANISE+.

**TRUMANS+.** TRUMANS [17] comprises 15 hours of motion-captured data representing human interacting with 100 varied indoor scenes and 20 distinct object types. While TRUMANS texts' describe the motion accurately, their connection to the scene is not strong, primarily focusing on the object of interaction. Furthermore, the 3D scans of the environment are significantly larger than those in HUMANISE, often containing irrelevant elements such as rooms unrelated to the motion. Although a recaptioning process identical to that of HUMANISE+ was not feasible due to the lack of detailed scene segmentation labels, we applied a similar strategy to strengthen the alignment between text, motion, and scene content. We refer to this dataset as TRUMANS+.

Table 1. Tasks and protocols’ results on HUMANISE+ [34] test set. Reported metrics are mRecall computed across ranks {1,2,3,5,10}. Greyed-out results are not directly comparable with those in the same column, as they do not leverage scene information.

Protocol	Method	<i>st2m</i>	<i>m2st</i>	<i>ms2t</i>	<i>t2sm</i>	<i>tm2s</i>	<i>s2mt</i>	<i>t2m</i>	<i>m2t</i>	<i>s2m</i>	<i>m2s</i>	<i>t2s</i>	<i>s2t</i>	avg.
All	TMR [27]	2.73	—	2.05	—	—	—	2.73	2.05	—	—	—	—	—
	MoPa [39]	4.49	—	3.53	—	—	—	4.49	3.53	—	—	—	—	—
	TMR + S	4.10	3.30	5.81	4.79	1.08	1.98	3.40	3.22	1.41	1.02	1.24	1.38	2.72
	MoPa + S	2.10	2.45	1.62	1.94	3.28	3.06	1.15	0.70	1.86	2.21	1.07	1.19	1.88
	MonSTeR	13.91	13.14	8.46	10.39	4.09	4.45	3.62	3.11	1.12	1.16	1.68	2.51	4.80
Small Batches	TMR [27]	56.0	—	56.01	—	—	—	56.0	56.01	—	—	—	—	—
	MoPa [39]	61.12	—	61.39	—	—	—	61.12	61.39	—	—	—	—	—
	TMR + S	57.08	55.49	70.77	67.05	33.86	37.26	59.42	58.17	27.78	25.73	38.31	38.45	47.44
	MoPa + S	48.83	40.73	40.46	39.23	48.37	40.73	28.47	27.00	36.80	36.66	35.17	37.45	38.32
	MonSTeR	79.15	79.01	76.30	77.42	57.55	57.06	63.20	61.89	34.11	34.47	50.18	49.76	60.00

Table 2. Tasks and protocols’ results on TRUMANS+ [17] test set. Reported metrics are mRecall computed across ranks {1, 2, 3, 5, 10}. Greyed-out results are not directly comparable with those in the same column, as they do not leverage scene information.

Protocol	Method	<i>st2m</i>	<i>m2st</i>	<i>ms2t</i>	<i>t2sm</i>	<i>tm2s</i>	<i>s2mt</i>	<i>t2m</i>	<i>m2t</i>	<i>s2m</i>	<i>m2s</i>	<i>t2s</i>	<i>s2t</i>	avg.
All	TMR [27]	17.55	—	17.71	—	—	—	17.55	17.71	—	—	—	—	—
	MoPa [39]	4.22	—	6.58	—	—	—	4.22	6.58	—	—	—	—	—
	TMR + S	4.04	4.44	9.01	4.67	9.14	1.61	5.65	12.03	1.11	5.89	6.74	1.65	5.49
	MoPa + S	4.83	5.20	4.11	4.39	13.20	2.14	3.52	3.85	2.06	10.79	8.30	1.04	5.28
	MonSTeR	10.54	10.29	8.59	8.51	12.71	3.94	5.67	5.90	1.94	7.88	10.35	2.80	7.42
Small Batches	TMR [27]	82.2	—	82.13	—	—	—	82.2	82.13	—	—	—	—	—
	MoPa [39]	48.55	—	56.44	—	—	—	48.55	56.44	—	—	—	—	—
	TMR + S	48.46	47.63	54.05	51.85	30.38	26.7	53.38	55.12	23.24	25.27	30.06	25.37	39.29
	MoPa + S	42.40	44.84	52.76	52.99	46.50	43.66	42.20	42.35	39.41	43.62	31.59	29.12	42.60
	MonSTeR	60.8	61.4	58.2	57.86	46.99	42.62	49.49	50.13	30.41	33.88	37.65	35.08	47.05

## 4.2. Retrieval Tasks

**Tasks and Metrics.** Given samples from three modalities, text (*t*), motion (*m*) and scene (*s*), we evaluate our model across several retrieval tasks: retrieving one modality given two others (*st2m*, *ms2t*, *mt2s*), retrieving two modalities given one (*m2st*, *t2ms*, *s2mt*), and retrieving one modality given another single modality (*t2m*, *m2t*, *s2m*, *m2s*, *t2s*, *s2t*). For instance, the *m2st* task requires finding the samples *s* and *t* which best suit the given *m* across a pool of possible pairs of scenes and texts. For brevity, we refer to these tasks as double-to-single, single-to-double, and single-to-single.

Retrieval tasks are typically evaluated using Recall@*K*, where *K* is a rank in {1, 2, 3, 5, 10}. This metric measures the percentage of times the correct sample is among the top *K* results, ordered by the similarity scores provided by the model. However, due to space constraints, we omit detailed Recall results at all ranks, instead presenting the more concise Mean Recall (mRecall) from [7], which represents the average Recall across ranks {1, 2, 3, 5, 10}. Comprehensive Recall results are provided in the Supplementary Material.

**Protocols.** The evaluation adheres to the retrieval protocols outlined in [13, 27]. Specifically, the “All” protocol requires distinguishing correct modality samples across all samples within the test set, while the “Small Batches” protocol requires discrimination among batches of 32 samples.

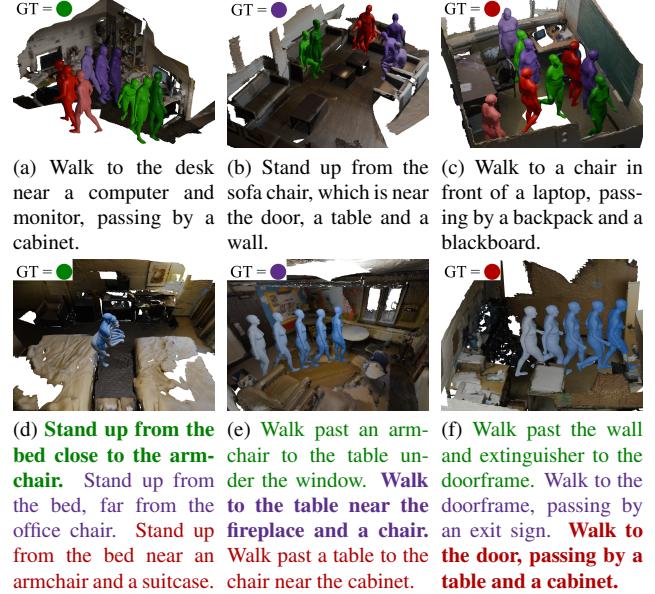


Figure 4. Qualitative examples for *st2m* (4c, 4b, 4a) and *ms2t* (4f, 4e, 4d). **First**, **second**, and **third** retrieved samples are shown. In each pictorial, GT=color (*top-left corner*) indicates the correct corresponding motion (*bottom row*).

Table 3. Ablation Studies for MonSTeR performed on HUMANISE+ [34]. We report the average mRecall computed across ranks.

Protocol	Method	<i>st2m</i>	<i>m2st</i>	<i>ms2t</i>	<i>t2sm</i>	<i>tm2s</i>	<i>s2mt</i>	<i>t2m</i>	<i>m2t</i>	<i>s2m</i>	<i>m2s</i>	<i>t2s</i>	<i>s2t</i>	avg.
All	MonSTeR	<b>13.91</b>	<b>13.14</b>	8.46	<b>10.39</b>	4.09	<b>4.45</b>	3.62	3.11	<u>1.12</u>	1.16	1.68	<u>2.51</u>	<b>5.63</b>
	- w/o cross-modal	5.20	3.77	<b>8.91</b>	7.95	2.49	2.59	<b>4.35</b>	<u>3.21</u>	1.01	0.91	<b>2.52</b>	<b>2.64</b>	3.79
	- w/o single	<u>11.91</u>	<u>12.93</u>	8.35	<u>8.96</u>	<b>4.33</b>	<u>4.31</u>	0.22	0.29	0.09	0.25	0.54	0.24	<u>4.36</u>
	- w tri-modal	6.14	6.00	7.56	7.93	3.02	3.12	<u>4.16</u>	<b>4.37</b>	<b>2.00</b>	<b>1.66</b>	<u>2.10</u>	1.63	4.14
Small Batches	MonSTeR	<b>79.15</b>	<b>79.01</b>	76.30	<b>77.42</b>	<b>57.55</b>	<b>57.06</b>	<u>63.20</u>	<u>61.89</u>	<u>34.11</u>	<u>34.47</u>	<b>50.18</b>	<b>49.76</b>	<b>60.00</b>
	- w/o cross-modal	65.47	63.26	<b>78.11</b>	<u>76.76</u>	49.89	49.31	<b>63.98</b>	<b>62.76</b>	32.23	32.52	<u>49.66</u>	<u>48.95</u>	<u>56.07</u>
	- w/o single	<u>75.31</u>	<u>76.21</u>	<u>76.42</u>	75.50	<u>53.76</u>	<u>53.86</u>	15.71	16.15	12.03	14.56	17.02	14.82	41.77
	- w tri-modal	62.48	62.46	69.02	68.29	48.33	48.20	61.42	60.00	<b>36.57</b>	<b>37.15</b>	45.80	44.70	53.70

Table 4. Ablation Studies for MonSTeR performed on TRUMANS+ [17]. We report the average mRecall computed across ranks.

Protocol	Method	<i>st2m</i>	<i>m2st</i>	<i>ms2t</i>	<i>t2sm</i>	<i>tm2s</i>	<i>s2mt</i>	<i>t2m</i>	<i>m2t</i>	<i>s2m</i>	<i>m2s</i>	<i>t2s</i>	<i>s2t</i>	avg.
All	MonSTeR	<u>10.54</u>	<u>10.29</u>	8.59	<b>8.51</b>	<u>12.71</u>	3.94	5.67	5.9	1.94	7.88	<b>10.35</b>	<b>2.83</b>	<b>7.42</b>
	- w/o cross-modal	5.89	6.26	6.58	6.3	12.01	3.1	<b>6.95</b>	<b>6.84</b>	<u>1.98</u>	<b>9.05</b>	<u>8.26</u>	2.06	<u>6.27</u>
	- w/o single	<b>12.95</b>	<b>13.01</b>	<b>7.01</b>	6.28	<b>12.99</b>	<b>4.76</b>	0.19	0.06	0.49	3.67	2.71	0.07	5.34
	- w tri-modal	5.62	4.91	6.24	6.34	10.29	3.05	<u>6.31</u>	<u>6.3</u>	<b>2.04</b>	<u>8.36</u>	8.19	<u>2.58</u>	6.02
Small Batches	MonSTeR	60.8	61.4	<b>58.2</b>	<b>57.86</b>	46.99	42.62	<u>49.49</u>	<u>50.13</u>	30.41	33.88	<b>37.65</b>	<b>35.08</b>	<b>47.04</b>
	- w/o cross-modal	50.96	51.12	50.12	50.75	38.12	33.98	<b>52.76</b>	<b>52.91</b>	29.91	33.56	33.73	31.08	42.41
	- w/o single	<b>66.96</b>	<b>68.43</b>	<u>56.38</u>	<u>56.46</u>	<b>51.59</b>	<b>48.38</b>	12.5	12.43	16.84	20.86	15.9	12.18	36.57
	- w tri-modal	47.35	48.72	50.77	49.61	41.4	36.48	48.89	49.24	<b>31.33</b>	<b>35.23</b>	37.08	<u>33.52</u>	<u>46.72</u>

**Retrieval Results.** In Table 1 and Table 2, we present MonSTeR retrieval results. Since no other model has previously addressed retrieval across scene, text, and motion, we compare it with two state-of-the-art text-to-motion retrieval models, TMR [27] and MoPa [39]. We report *t2m* and *m2t* as proxies for the *st2m* and *m2st* tasks and grey out these results, as they provide the extra input *s* which neither TMR nor MoPa can use. To ensure a fair comparison (TMR+S and MoPa+S), we start from their pretrained architecture and extend it with the same scene encoder as MonSTeR, finetuning it on the target dataset. As they do not leverage cross-modal encoders we average their unimodal representations to produce cross-modal ones [11].

In the *t2m* and *m2t* tasks, MonSTeR consistently outperforms all baselines on HUMANISE+, while TMR leads on TRUMANS+. We attribute this discrepancy to the limited value of scene cues in TRUMANS+: individual scenes often support several distinct motions. The dataset’s high motion diversity also makes text and motion signals more discriminative than scene context, even after recaptioning. TMR capitalizes on this with Guo features [13], which ignore spatial ties between scene and motion and thus make tasks other than *t2m/m2t* impossible. TMR accuracy drops sharply without them (see Supplementary Material). By contrast, HUMANISE+ contains similar text–motion pairs, so scene information is essential for disambiguation. Note that this setting is the closest match to our target application of evaluating Human-Scene Interaction generations.

Overall, MonSTeR emerges as the top performer for the majority of double-to-single and single-to-double tasks. It improves “All” protocol’s *st2m* relative to the best among the evaluated models by 209% on HUMANISE+. On av-

erage, MonSTeR outperforms the best scene-aware models by 76.47% and 35.15% on the “All” protocol on HUMANISE+/TRUMANS+, and by 26.47% and 10.44% on “Small Batches”, showing superior multimodal alignment.

**Additional Results.** In Fig. 4 we provide qualitative results of MonSTeR’s rankings for the *st2m* and *ms2t* tasks. This evaluation highlights both the complexity and challenging nature of these tasks, as many samples exhibit high semantic relevance to the conditioning modalities. Moreover, it demonstrates the effectiveness of MonSTeR in retrieving motions that maintain a strong correspondence with the ground truth in its top-ranked results. More qualitative results are available in the Supplementary Material, including examples for other tasks, such as *m2t* and *mt2s*.

Finally, we evaluate our model on HumanML3D [13], a common text-to-motion dataset that does not include scene information. In Sec. 9 of the Supplementary Material we show that MonSTeR achieves results comparable to state-of-the-art text-to-motion retrieval models [27, 39].

### 4.3. Ablation Studies on the Model Design

In Table 3 and Table 4 we test ablative variants of MonSTeR: the tested variants still have the same architecture but they differ by taking into consideration only unimodal or cross-modal relations in the loss. Recalling Sec. 3.3, the ablation studies vary the set *K* used for the contrastive terms. In “MonSTeR w/o cross-modal”, we remove all cross-modal encoders, which also eliminates all cross-modal terms from the set *K*, resulting in the optimization of the set  $\{(t, s), (m, t), (m, s)\}$ . In this case we use text as the bridging modality to align all three modalities and average latents during inference to com-

pute the cross-modal representations [11]. Predictably, this model has significant performance drops in most double-to-single and single-to-double retrieval tasks. Next, in “MonSTeR w/o single” we remove all terms involving only single-modality from  $K$ , resulting in optimizing the reduced set  $\{(st, m), (mt, s), (ms, t)\}$ . This model aligns single modalities’ embedding to cross-modal embeddings only. In some cases, the double-to-single retrieval task of the ablated model outperforms the corresponding performance of MonSTeR on TRUMANS+. This improvement likely stems from a closer alignment between training and test objectives in these instances. However, on average, this model’s performance appears diminished compared to that of MonSTeR (see *avg.* in Table 3 and Table 4). Indeed single-to-single retrieval performance shows a significant drop, highlighting the limitations of the ablated approach. In addition to these ablation we also evaluate “MonSTeR w tri-modal”, a single tri-modal architecture inspired by [7]. Here, all modalities are embedded by the same tri-modal encoder and decoded jointly to compute cross-modal embeddings. While this model performs well on certain single-to-single retrieval tasks, it underperforms on double-to-single and single to double tasks. These experiments confirm that superior performance can be achieved by modeling higher-order relationships through the alignment of both unimodal-to-unimodal and unimodal-to-cross-modal interactions while retaining the rich feature semantics of the single encoders.

#### 4.4. MonSTeR for Evaluation

In this section, we introduce MonSTeR as a tool to assess the quality of text-conditioned Human Scene Interaction generative models (HSI). After testing its motion-scene grounding properties, we validate the alignment of MonSTeR’s scores with human judgment.

**MonSTeR Metrics.** FID and Recall@{1,2,3} are commonly employed in the evaluation of text-to-motion models [6, 13, 14, 33]. However, these metrics do not account for the scene and its interactions with text and motion, rendering them less suited for evaluating HSI models. By leveraging MonSTeR’s latent space, we can derive more informative versions of these metrics.

We extend FID to assess both the motions’ plausibility and their coherence with the scene leveraging the embeddings from the cross-modal encoder  $MS$  of MonSTeR (cf. Fig. 2). We extend the Recall [13], by using the  $m2st$  task embeddings to measure the adherence of a motion to a text and a scene.

**Evaluating Path Plausibility.** To evaluate the scene-motion grounding capabilities of MonSTeR, we assess if its latent space can differentiate between motions that adhere to paths coherent with the provided text and scene affordances, and those that do not. To this end, we rotate motions from

the test sets of [17, 34], starting at an angle of 0 radians and increasing incrementally up to  $\pi$ , with the rotation pivot at the end position of the original motion. We ensure that each rotated motion stays within the scene boundaries and avoids interpenetrating any scene objects.

Solid lines in Fig. 5 show that both FID and Recall values computed using MonSTeR embeddings deteriorate as the rotation increases. This is both expected and desirable, as increased rotation results in a larger divergence between the original and modified paths, leading to a loss of coherence between the motion, the scene, and the conditioning text. This is more evident in [17] due to the precise descriptions of contact interactions with small objects (e.g. *pick up the keyboard on the desk*); when motions are rotated, the coherence is disrupted. The gap in Fig. 5 reflects datasets’ properties. TRUMANS+ mixes dynamic motions with small targets in wide spaces, so small rotations quickly misalign modalities and lower scores. HUMANISE+, dominated by static actions around large, nearby objects, is far less sensitive to the same rotations.

**Evaluating Compliance with the Scene.** We repeat the experiment above but enable motion rotations that would lead to interpenetration. In Fig. 5, comparing the dotted lines to the solid ones, we see that including motions penetrating scene objects causes an even further decline in these metrics. This suggests that the latent space of MonSTeR has also internalized that natural motions should not involve interpenetration with the scene.

**Alignment to Human Preferences.** We perform a user study to test the alignment of MonSTeR’s ranking with human preferences. We generate motions using two state-of-the-art models for text-conditioned human-scene interaction: the conditional variational auto-encoder from [34] and the two-stage diffusion model from [35]. For each pair of generated motions, we rank them using  $st2m$  score and compare this ranking to human preferences. We collect 1122 annotations from 224 evaluators. The results show that MonSTeR’s rankings align with human evaluations 66.5% of the time, indicating agreement with human judgment.

#### 4.5. Downstream Tasks

**In Scene Object Placement.** We question if MonSTeR latent space can discriminate if an object is placed on the correct  $x, y, z$  coordinates. To this aim, we extrapolate objects of interest, and set up a  $5 \times 5 \times 5$  grid around their ground-truth positions. We shift the object into each of the 125 cells in the grid, creating an equal number of modified scenes.

Then, we leverage the  $mt2s$  score of MonSTeR to evaluate the similarity between the text-motion embedding and each scene variant. This process provides a score for each pairing  $((text, motion), scene_i)$ , with  $i \in [1, 125]$ .

We locate the object of interest by selecting the highest score provided by our model and compute the  $L2$  dis-

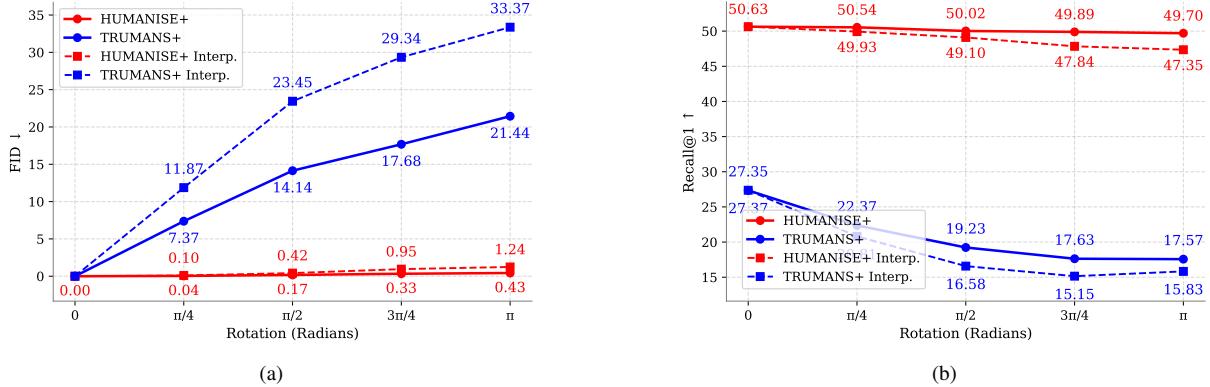


Figure 5. MonSTeR’s FID ↓ (a) and Recall@1 ↑ (b) trends when motions are increasingly rotated, from 0 to  $\pi$  radians.

tance to its real position. As we use  $25 \times 25 \times 25$  cm cells the error ranges in [0 cm, 86.6 cm], with the maximum being twice the length of a cell’s diagonal. While the average error is 58.98 cm, our zero-shot prediction scores only 18cm. MonSTeR can accurately localize objects in a zero-shot setting, leveraging spatial knowledge grounded in the text-motion latent representations. Further details and visualizations are available in Sec. 5 of the Supplementary Material.

**Motion Captioning.** To test the descriptiveness of MonSTeR’s latent representations, we use its embeddings for the downstream task of Motion Captioning. In particular, we freeze MonSTeR’s latent space and feed its motion embeddings to GPT2 [29] which we then train to caption motions. We then compare our results with MotionGPT [16], retrained from scratch on both HUMANISE+ and TRUMANS+, adopting the evaluation protocol from [16].

As shown by Table 5 and Table 6, MonSTeR is able to surpass MotionGPT on most metrics by a significant margin. Intuitively, the motion embedding from our model carries more semantic information as it has previously been aligned with the scene. Moreover, the higher ROUGE<sub>L</sub> and Bert<sub>F1</sub> suggest that the model not only predicts captions that preserve the word ordering of the ground truth but also generates descriptions that are closer to natural human language.

Method	BLEU 1	BLEU 4	ROUGE L	CIDER	BERT F1
mGPT	42.16	17.47	40.23	11.13	22.16
MonSTeR + GPT2	<b>42.93</b>	<b>23.59</b>	<b>50.85</b>	<b>13.70</b>	<b>35.57</b>

Table 5. Captioning performances on HUMANISE+ [34].

Method	BLEU 1	BLEU 4	ROUGE L	CIDER	BERT F1
mGPT	39.62	18.13	40.62	<b>15.08</b>	17.59
MonSTeR + GPT2	<b>42.82</b>	<b>21.59</b>	<b>45.98</b>	12.24	<b>26.66</b>

Table 6. Captioning performances on TRUMANS+ [17].

## 5. Limitations

We train cross-modal encoders only on aligned modality pairs, due to the unfeasible computational expense of training on unpaired data. Also, our scenes are static, meaning human actions do not alter the scene layout. While only a few works explore dynamic scene encoders, this remains a promising direction for future extensions of MonSTeR.

## 6. Conclusion

We introduced a novel retrieval model that unifies text, motion, and scene within a shared latent space. Inspired by topological deep learning, MonSTeR captures higher-order interactions across all three modalities. We validated it through retrieval tasks, demonstrating the effectiveness of our unified representation. Through comprehensive ablation studies, we highlighted the contribution of each component to the model’s overall performance. Also, we proposed MonSTeR as a tool for grounding Human-Scene Interaction models within the scene context and evaluated its alignment with human preferences. We applied our model on in-Scene Object Placement and Motion Captioning, showing its ability to generalize beyond traditional retrieval scenarios.

## 7. Acknowledgement

We acknowledge partial financial support from Wsense S.r.l., from the PNRR MUR project PE0000013-FAIR (CUP: B53C22003980006), from the Sapienza grants RG123188B3EF6A80 (CENTS), and RM1241910E01F571 (V3LI). This work has been carried out while M.P. was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome, funded by the European Union – Next Generation EU, Mission 4 Component 1 CUP B53C22003870006. O.L is a Taub fellow and is supported by the Azrieli Foundation Early Career Faculty Fellowship. He is also supported by the Israel Science Foundation through a personal grant (ISF 624/25) and an equipment grant (ISF 2903/25).

## References

- [1] Léore Bensabath, Mathis Petrovich, and Gü̈l Varol. Tmr++: A cross-dataset study for text-based 3d human motion retrieval. In *CVPRW*, 2024. 2
- [2] Cristian Bodnar, Fabrizio Frasca, Yuguang Wang, Nina Otter, Guido F Montufar, Pietro Lió, and Michael Bronstein. Weisfeiler and lehman go topological: Message passing simplicial networks. In *Proceedings of the 38th International Conference on Machine Learning*, pages 1026–1037. PMLR, 2021. 3
- [3] Tim Brooks and Alexei A. Efros. Hallucinating pose-compatible scenes, 2022. 2
- [4] Zhi Cen, Huajin Pi, Sida Peng, Zehong Shen, Minghui Yang, Shuai Zhu, Hujun Bao, and Xiaowei Zhou. Generating human motion in 3d scenes from text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1855–1866, 2024. 1
- [5] Jiaqi Chen, Daniel Barath, Iro Armeni, Marc Pollefeys, and Hermann Blum. Where am i? scene retrieval with language. *European Conference on Computer Vision 2024*, 2024. 2
- [6] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space, 2023. 7
- [7] Ginger Delmas, Philippe Weinzaepfel, Francesc Moreno-Noguer, and Grégory Rogez. Poseembroider: Towards a 3d, visual, semantic-aware human pose representation. *European Conference on Computer Vision 2024*, 2024. 3, 5, 7
- [8] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 4
- [9] David F. Fouhey, Vincent Delaitre, Abhinav Gupta, Alexei A. Efros, Ivan Laptev, and Josef Sivic. People watching: Human actions as a cue for single view geometry. In *Computer Vision – ECCV 2012*, pages 732–745. Springer, Berlin, Heidelberg, 2012. 2
- [10] Kent Fujiwara, Mikihiro Tanaka, and Qing Yu. Chronologically accurate retrieval for temporal grounding of motion-language models. In *European Conference on Computer Vision*, pages 323–339. Springer, 2024. 1, 2
- [11] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 2, 6, 7
- [12] Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder. *arXiv preprint arXiv:2210.07839*, 2022. 2
- [13] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 1, 2, 5, 6, 7
- [14] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions, 2023. 7
- [15] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *European Conference on Computer Vision*, pages 289–310. Springer, 2025. 3
- [16] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024. 8
- [17] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1737–1747, 2024. 4, 5, 6, 7, 8
- [18] Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei Efros, and Krishna Kumar Singh. Putting people in their place: Affordance-aware human insertion into scenes. In *Computer Vision – CVPR 2023*, pages 17089–17099, 2023. 2
- [19] Jiye Lee and Hanbyul Joo. Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3d environments, 2023. 2
- [20] Weixian Lei, Yixiao Ge, Kun Yi, Jianfeng Zhang, Difei Gao, Dylan Sun, Yuying Ge, Ying Shan, and Mike Zheng Shou. Vit-lens: Towards omni-modal representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26647–26657, 2024. 2
- [21] Ruiping Liu, Jiaming Zhang, Kunyu Peng, Junwei Zheng, Ke Cao, Yufan Chen, Kailun Yang, and Rainer Stiefelhagen. Open scene understanding: Grounded situation recognition meets segment anything for helping people with visual impairments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1857–1867, 2023. 2
- [22] Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3):2276–2289, 2022. 2
- [23] Yinyu Nie, Angela Dai, Xiaoguang Han, and Matthias Nießner. Pose2room: Understanding 3d scenes from human activities, 2022. 2
- [24] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4
- [25] Theodore Papamarkou, Tolga Birdal, Michael M. Bronstein, Gunnar E. Carlsson, Justin Curry, Yue Gao, Mustafa Hajij, Roland Kwitt, Pietro Lio, Paolo Di Lorenzo, Vasileios Maroulas, Nina Miolane, Farzana Nasrin, Karthikeyan Natesan Ramamurthy, Bastian Rieck, Simone Scardapane, Michael T Schaub, Petar Veličković, Bei Wang, Yusu Wang, Guowei Wei, and Ghada Zamzmi. Position: Topological

- deep learning is the new frontier for relational learning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 39529–39555. PMLR, 2024. 2, 3
- [26] Mathilde Papillon, Sophia Sanborn, Mustafa Hajij, and Nina Miolane. Architectures of topological deep learning: A survey of message-passing topological neural networks, 2024. 3
- [27] Mathis Petrovich, Michael J. Black, and Gü̈l Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *ICCV*, 2023. 1, 2, 3, 5, 6
- [28] Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S Melo, Ana Paiva, and Danica Kragic. Geometric multimodal contrastive representation learning. In *International Conference on Machine Learning*, pages 17782–17800. PMLR, 2022. 2
- [29] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 8
- [30] Yue Ruan, Han-Hung Lee, Yiming Zhang, Ke Zhang, and Angel X. Chang. Tricolo: Trimodal contrastive loss for text to shape retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5815–5825, 2024. 2
- [31] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. 3
- [32] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *ECCV*, pages 358–374. Springer, 2022. 2
- [33] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model, 2022. 7
- [34] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems*, 35:14959–14971, 2022. 1, 4, 5, 6, 7, 8
- [35] Zan Wang, Yixin Chen, Baoxiong Jia, Puahao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say interact as you can: Language-guided human motion generation with scene affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 433–444, 2024. 1, 7
- [36] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1179–1189, 2023. 2
- [37] Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27091–27101, 2024. 2
- [38] Kangning Yin, Shihao Zou, Yuxuan Ge, and Zheng Tian. Tri-modal motion retrieval by learning a joint embedding space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1596–1605, 2024. 2