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

Supplementary Material

We complement the main paper with additional details, insights, and retrieval results at all ranks. In particular, Sec. 1 provides more information about the recaptioning process and Sec. 2 about the Motion Captioning pipeline. Then, Sec. 3 includes additional details on training and evaluation, while Sec. 4 discusses using `MonSTeR` as an evaluator for text-conditioned human-scene-interaction models, compared to standard metrics. Also, Sec. 5 presents additional considerations about In-Scene Object Placement, while Sec. 6 elaborates further on the user study. Then, Sec. 7 and Sec. 8 provides implementations details and additional qualitative results. Finally, Sec. 9 includes the results for the main retrieval experiments at all ranks described in Sec. 4.2 of the Main Paper and Sec. 10 includes results for the *t2m* and *m2t* tasks on HumanML3D [4].

1. Recaptioning Process

This section describes the process used to reception HUMANISE [11] and obtain the HUMANISE+ and TRUMANS+ datasets. Note that the recaptioning on TRUMANS is less fine-grained as it lacks detailed scene segmentation labels.

1.1. Overview of the Recaptioning Process

The recaptioning process leverages LLAMA3 [2], specifically the "8-Billion-Instruct" version. This model can generate an answer by following a set of instructions and examples. After being given an initial context, the LLM is tasked to ground captions using environmental cues from the scene. To provide the LLM with this information, we first generate a list of scene elements located within a small distance from the path traversed by the human in motion. Next, we supply the LLM with the original caption, appending the generated list of objects. The object classes or descriptions provided to the model are derived from the segmentation labels included in the original HUMANISE dataset. For what regards TRUMANS, only the objects mentioned in the caption and explicitely loaded in the mesh file are used to perform the recaptioning since the dataset lacks detailed segmentation labels. The rest of the process is the same as HUMANISE's. The final process results in captions that better ground the motion in the original scene, as seen in Fig. 1.

1.2. LLM Prompts

Listing 1 includes the full recaptioning prompt used to condition the LLM for the recaptioning process. The initial context is fed to the model through the *"system"* role, that provides global information on how to execute the instructions. The request is to augment each sentence with information on nearby objects, preserving the original phrasing as much as possible. Afterwards, the LLM is conditioned on 5 different examples, each representing a potential output based on the input sentence and a list of objects. Each caption also includes a reference to the human subject, an element not present in the original dataset. If an object appears in the original caption, the text describing the interaction with it is kept unchanged. The remaining objects are simply included mentioning they are *"nearby"* the motion, in order to not alter the original ground truth caption.

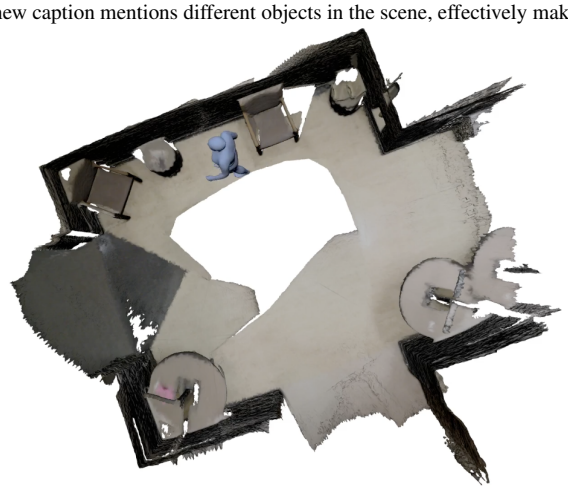
2. Motion Captioning

This section briefly discusses more details of the motion captioning process, specifically the training details for `MonSTeR` + GPT2 [10] and the justification behind the model's choice.

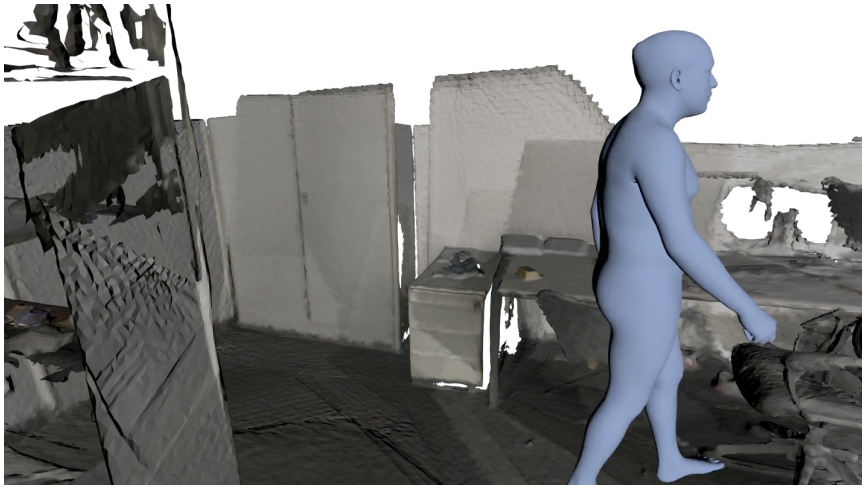
2.1. Further Details on Training Motion Captioning

To train the motion captioning model, we first process all the motions in the entire dataset using `MonSTeR`'s frozen motion encoder and feed them as token-embeddings to GPT-2. We train the encoder-decoder for 50 epochs on TRUMANS+ [6] and 100 on HUMANISE+ [11] respectively. We use a very low learning rate of 1e-6, similarly to [5], in order to ensure stability during the tuning process. During training, we perform ground truth forcing to ensure model convergence, teaching the model to match the original input caption and associate `MonSTeR`'s motion embeddings to a new learned $\langle motion \rangle$ token, prepended to the start of the input tokens. During inference, we provide the model with the input embedding followed by the start-of-sequence token, and generate text autoregressively. At the end of training, we validate the performance of GPT2 on the test set following the same protocol of [5]. In particular, we replicate each evaluation 20 times to compute their average scores. Each metric measures a different aspect of text generation, specifically BLEU 1 evaluates unigram overlap, BLEU 4 assesses n-gram precision up to four words, CIDER measures consensus using weighted n-grams, ROUGE L identifies the longest common subsequence for sequence similarity and BERT F1 quantifies semantic similarity using contextual embeddings. Our model achieves competitive scores on each metric when compared with [5].

(a) HUMANISE caption, *"walk to the bed"*, compared to HUMANISE+ recaptioning, *"a person walks to the bed, passing by a nightstand, a dresser, some books, and finally reaches the bed"*. The new caption mentions different objects in the scene, effectively making it more coherent with it.



(b) HUMANISE caption, *"walk to the end table that is close to the door"*, compared to HUMANISE+ recaptioning, *"a person walks towards the end table that is close to the door, and passes by a wall and a chair"*.



(c) HUMANISE caption, *"walk to the chair that is next to the telephone"*, compared to HUMANISE+ recaptioning, *"a person walks to a chair that is next to a telephone, passing by a bag, a desk, and a whiteboard with some paper on it"*.

Figure 1. Comparison between HUMANISE captions and HUMANISE+ recaptioning.

Listing 1. LLAMA3 prompts used in the recaptioning process. The “system” role is used to give the model the initial context.

```
    "role": "system",
    "content": "You are a bot made to rephrase sentences. You will receive in input a pair
        of sentences that describe the motion of a human inside a scene and an ordered list
        of objects. The input will be of the form 'sentence describing action', (list of
        objects encountered separated by a comma), where the object list is enclosed in
        parenthesis and each object in the list is separated by a comma. Your goal is to
        rephrase the original sentence to include the objects mentioned in order. If an
        object is mentioned multiple times, include only one entry unless the exact same
        object was included in the original sentence. Do not consider synonims as the same
        object. You only have to reply with the sentences used to rephrase the action. Do
        not alterate the order in which objects appear in the original sentence. **Refrain
        from outputting text not related to the task**."


    "role": "user",
    "content": "walk from chair to table, (chair, wardrobe, table)"


    "role": "assistant",
    "content": "a human is walking from a chair, passing nearby a wardrobe, finally reaching
         a table"


    "role": "user",
    "content": "sit on the toilet,(bathroom cabinet) "


    "role": "assistant",
    "content": "a person is sitting on the toilet near a bathroom cabinet"


    "role": "user",
    "content": " lie on the sofa chair,(bed)"


    "role": "assistant",
    "content": " someone lies on the sofa chair nearby the bed"


    "role": "user",
    "content": "sit on the toilet,(toilet paper)"


    "role": "assistant",
    "content": "an individual is sitting on the toilet and toilet paper is nearby"


    "role": "user",
    "content": "Sounds great! Whenever an object that could be interacted with is mentioned,
         such as toilet paper or clothing, be sure to include it in the sentence. Do not
        mention any interaction with the objects, just their presence. Let's try another one
        : sit on the chair that is farthest from the bathtub,(desk)"


    "role": "assistant",
    "content": "a human is sitting on the chair, nearby a desk, that is farthest from the
        bathtub"
```

Encoder-Decoder's Choice To create a motion captioning model, we choose to combine MonSTeR with a large language model. Our goal is to make the LLM learn the new $\langle motion \rangle$ token associated with the motions in our dataset. We thus use the pre-trained version of GPT2 [10] from HuggingFace and subsequently finetune it on the motion captioning task. We specifically choose GPT2 due to its robust pre-training and its attention mechanism that can effectively leverage the new $\langle motion \rangle$ token prepended to the start of the input to caption.

3. Further Details on Training and Evaluation

To ensure reproducibility and comparability of the results in this work, we provide additional information about the model selection process and the dataset splits employed.

3.1. Further Implementation Details

All models are trained for 24 hours on a single NVIDIA A100 64 GB. We find it beneficial to employ small training batches, 16-sized, and a constant learning rate of $1e-5$. Regarding model selection, all presented model checkpoints were chosen based on their best validation *st2m* scores. For models such as TMR [8] and MoPa [13], which do not utilize scene information, checkpoints were selected based on the best validation *t2m* scores.

3.2. Datasets

As no prior work has addressed the task of text-motion-scene retrieval, we selected large-scale datasets featuring skeletal motions described with text and accompanied by scenes. Specifically, as detailed in the Main Paper (cf. Sec 4.1), we utilized TRUMANS+ [6] and HUMANISE+ [11]. However, these datasets have not been previously used for retrieval tasks.

For dataset splits, the approach varies by dataset. With TRUMANS+, we preprocess the dataset following the authors' guidelines closely to minimize leakage while making it suitable for the retrieval task. Motions contained in the dataset are in fact very long, with some spanning over thousands of frames (for reference, HUMANISE's motions are often in the low hundreds of frames). Processing very long motions presents challenges because state-of-the-art transformer-based motion encoders, such as the one employed in MonSTeR, struggle to efficiently compute attention and maintain awareness of long-term dependencies, compared with models specifically designed for that case. We therefore manually divide each motion into smaller ones, making sure that we preserve the original starting and ending frames for each action in order not to pick actions *in fieri*. For HUMANISE+, we use the original splits provided by the authors and select the first motion for each scene when testing, which ensures that there is no scene leakage between these sets.

To make retrieval results comparable, we align the dimension of the test sets. Specifically, we cap TRUMANS+'s test set at 1600 samples, roughly the same amount of test samples in HUMANISE+.

3.3. Evaluation Procedure

We adhere to evaluation procedures and protocols defined in [13]. Going into more detail, such procedures consider a retrieved sample correct for a given conditioning modality not only if the retrieved sample and the conditioning modality belong to the same datapoint (tuple), but also if the retrieved sample is identical to the one in that datapoint. For instance, in the computation of Recall at all ranks for the *ms2t* task, a retrieved text sample t_j is considered correct for a conditioning motion-scene tuple (m_i, s_i) not only if $i = j$ but also if $t_j = t_i$. We employ the above evaluation for both single-to-single and double-to-single retrieval tasks. For what concerns single to double, a retrieved couple of samples is considered to be correct if it belongs to the same datapoint of the conditioning modality or both the elements in the retrieved tuple are equal to those in the datapoint. For instance, in the computation of Recall at all ranks for the *t2ms* task, a retrieved motion-scene tuple (m_j,s_j) is considered correct for a conditioning text sample t_i if $i = j$ or ($m_j = m_i \wedge s_j = s_i$).

4. Human Scene Interaction Evaluation

In this section, we present evaluation metrics computed using MonSTeR's latent space for the task of generating text-conditioned human-scene interactions and evaluate their contribution in comparison to traditional metrics.

Specifically, we report Recall@1,2,3 and FID (cf. Sec. 4.4 of the Main Paper), calculated using MonSTeR, alongside commonly used metrics for this task [11, 12], including goal distance, contact, and non-collision.

The evaluation is conducted on two state-of-the-art models: Move-as-You-Say (MaYS) [12] and HUMANISE (cVAE) [11].

Traditional metrics, such as goal distance, contact, and non-collision, do not account for the coherence between the text prompt and the generated motion. In contrast, Recall@1,2,3 evaluates the alignment of all three modalities—text, motion, and scene—by assessing the coherence between the conditioning text prompt and the combined motion-scene signal. This makes Recall a more comprehensive metric for evaluating cross-modal consistency. Additionally, traditional metrics provide only limited insights into scene alignment. Conversely, FID, computed using MonSTeR's motion-scene combined representation, effectively captures discrepancies in scene-motion coherence.

Table 1 summarizes the results. Both MaYS and cVAE demonstrate strong performance in maintaining appropriate motion-scene contact and avoiding interpenetration with

Table 1. Generation Results on HUMANISE+ [11].

Model	goal distance ↓	contact ↑	non-collision ↑	Recall@1 ↑	Recall@2 ↑	Recall@3 ↑	FID ↓
MaYS [12]	**0.168**	**0.94**	0.996	**0.43**$^{\pm 0.002}$	**0.62**$^{\pm 0.003}$	**0.74**$^{\pm 0.002}$	**2.157**
cVAE [11]	0.596	0.855	**0.997**	0.33$^{\pm 0.004}$	0.49$^{\pm 0.003}$	0.59$^{\pm 0.002}$	32.857

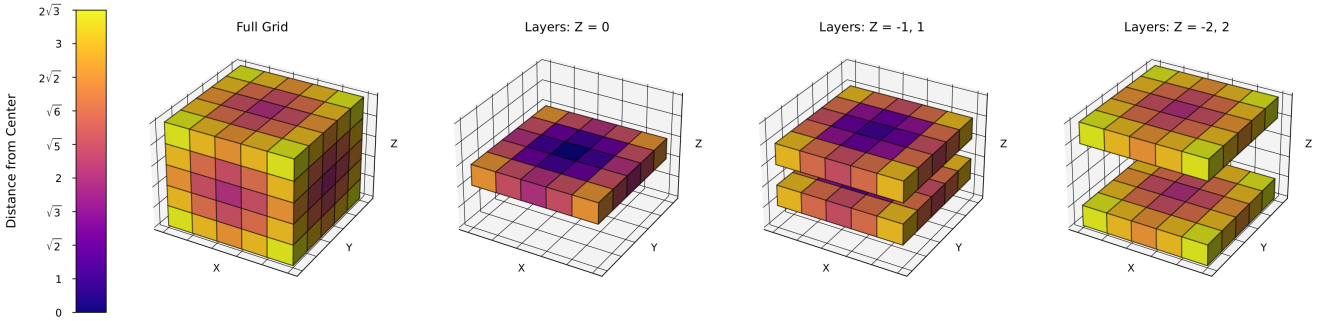


Figure 2. Visualization of the $5 \times 5 \times 5$ grid used for the in-scene object placement evaluation. The full grid (leftmost) shows all cells. Subsequent panels depict sliced layers along the z-axis: $z = 0$ (second from left), $z = (-1, 1)$ (third), and $z = (-2, 2)$ (rightmost). Distances from the center represented with colors, with values going from 0 (darker colors) to $2\sqrt{3}$ (brighter colors).

the environment, with neither model achieving a significant edge in these metrics.

However, Recall@1,2,3 scores reveal that MaYS consistently outperforms cVAE by a significant margin, indicating that MaYS generates motions better aligned with the conditioning text descriptions within the scene. Additionally, MaYS exhibits superior performance in goal distance, reflecting its ability to generate motion trajectories that more closely align with the target destination.

These trends are further corroborated by the FID scores in Table 1, where cVAE reports substantially higher values. This can be attributed to its greater distance from the goal, compromising motion-scene coherence, as well as the presence of generation artifacts (e.g., foot-skating) and motion implausibilities. Notably, MonSTeR-based FID is sensitive to implausible or incorrect paths followed by motion trajectories (cf. Sec. 4.4 of the Main Paper), an aspect not captured by goal distance, contact, or non-collision metrics.

5. In-Scene Object Placement Details

In this section, we further motivate and provide details on the experimental setup used to evaluate MonSTeR's performance on the in-Scene Object Placement task (cf. Sec. 4.5 of Main Paper).

Why In-Scene Object Placement?: Given a text description and a motion input, MonSTeR can retrieve not only the most relevant scene but also the most appropriate scene layout, distinguishing between scenes that differ only in the positions of one or a few objects. We evaluate this capability in this downstream task. Notably, it can serve as a valuable prior when synthesizing scenes from textual and motion inputs. A qualitative illustration is shown in Fig. 3.

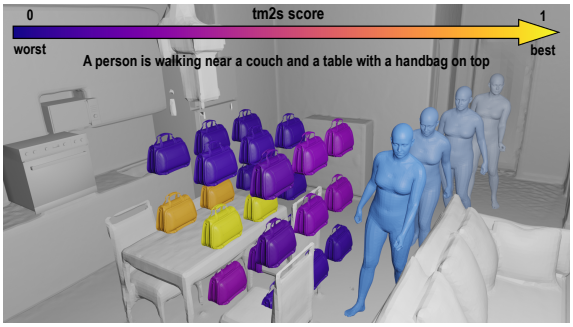


Figure 3. In-Scene Object Placement scores visualized

Further Details: HUMANISE+ is not suited for this experiment, as it integrates objects into the scene mesh, restricting the flexibility required for this task. Although objects can still be detected and their points selected in the scene's point cloud using bounding boxes for a different placement, this process also inadvertently moves points from nearby objects and parts of the scene, which may result in perforated versions of the scene and shattered objects. These missing sections make identifying incorrect scene arrangements easier for our network. Therefore, we conduct this experiment exclusively on TRUMANS+, which provides separated object and scene meshes, enabling straightforward placement of interactive objects.

To determine the optimal placement for an object, we set up a $5 \times 5 \times 5$ grid around the object's initial position. Fig. 2 illustrates this grid, with z-axis slices used to display all the available distances. Different colors correspond to different distances from the center.

The 125 positions in the grid correspond to specific cells with known distances from the grid center. Assuming a unit

step length along each edge, the distances of cells from the center are distributed as follows:

- 1 cell at a distance of 0, representing the grid's center;
- 6 cells at a distance 1;
- 6 cells at a distance 2;
- 8 cells at distance $\sqrt{3}$, corresponding to the one-step cube diagonal;
- 8 cells at distance $2\sqrt{3}$, corresponding to the two-step cube diagonal;
- 12 cells at distance $\sqrt{2}$, corresponding to the one-step diagonal along a cube face;
- 12 cells at distance $2\sqrt{2}$, corresponding to the two-step diagonal along a cube face;
- 24 cells at distance $\sqrt{5}$, corresponding to the diagonal length of 2×1 rectangles on the cube's faces;
- 24 cells at distance $\sqrt{6}$, corresponding to the diagonal length of $2 \times 1 \times 1$ rectangular parallelepipeds in the cube;
- 24 cells at distance 3, corresponding to the diagonal length of $2 \times 2 \times 1$ rectangular parallelepipeds in the cube.

In this scenario, the *maximum possible error* occurs when two steps are taken along a diagonal path, while the *average error* is given by

$$\frac{L}{125}(6(3) + 8(3\sqrt{3}) + 12(3\sqrt{2}) + 24(\sqrt{5} + \sqrt{6} + 3))$$

where L represents the translation step. With a step size of 25 cm, their values correspond to 86.60 cm and 58.98 cm, respectively.

6. User Study

In order to evaluate the agreement between MonSTeR's and human preferences, we conduct a User Study on Amazon Mechanical Turk. We recruit 224 evaluators, filtered based on their acceptance rate on previous tasks, i.e., a score they received from past requesters. They provide 1122 preferences over a sample of 200 motions generated by cVAE [11] and MaYS [12].

Fig. 4 shows the annotation interface that includes:

- A description of the task;
- A textual prompt describing a motion;
- A pair of GIFs depicting the corresponding motion as generated by cVAE and MaYS, in a random order;
- Two buttons for selecting the user's preferred GIF.

The results reveal that MonSTeR 's rankings match human evaluations 66.5% of the time, reflecting a strong agreement with human judgments.

7. Qualitative Results

Fig. 5 presents qualitative results of MonSTeR' rankings for the *t2m* task and Fig. 6 for the *mt2s* task.

For each example in the first task, we show the first three motions retrieved from the given text prompt. Note that here

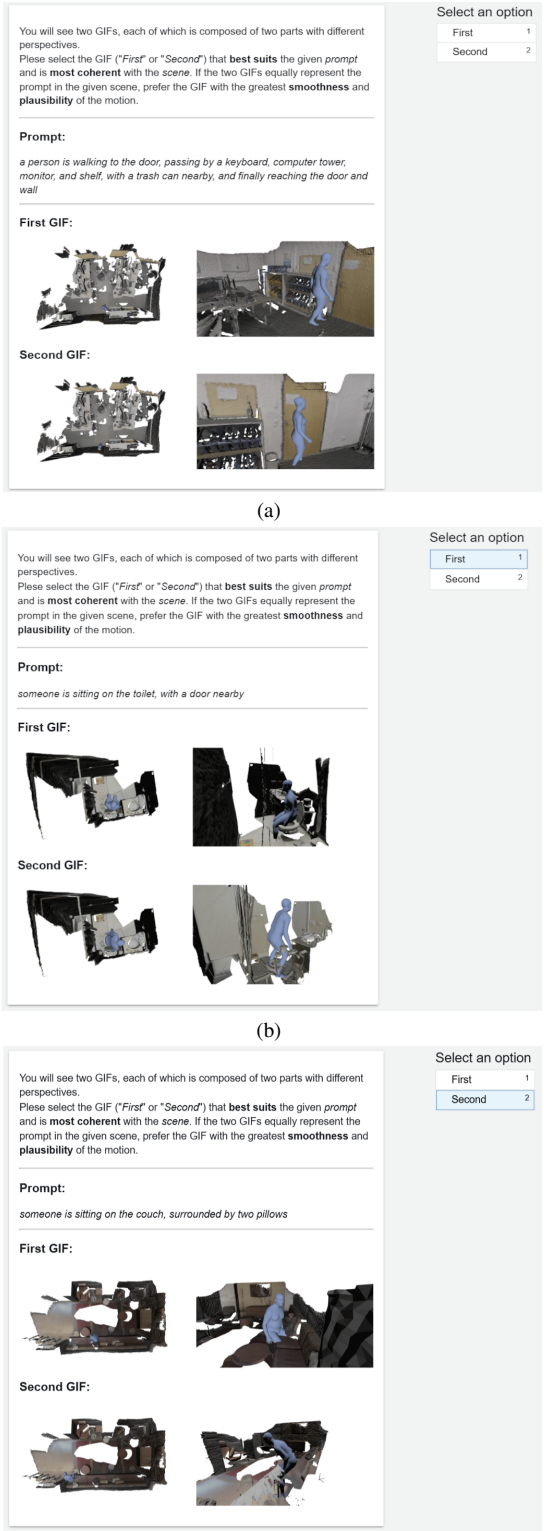


Figure 4. Examples of the interface used for the User Study, described in Sec. 6.

they are shown together with the respective scene, even if it is not part of the task, just to provide some context.

In the second task, a scene is retrieved based on a pair composed of a motion and a text prompt so, for this task, each result in every example shares the same text and motion, while the scene varies.

8. Implementation details

Algorithm 1 details the complete encoding pipeline used in MonSTeR. T and M encoders are based on [7], while we use [9] as S, augmented with Spatial Attention layers. For the cross-modal encoders, we start from the unified encoder from [14] and modify it to produce a single joint modalities' representation via cross-attention between modalities and projection layers. Our model incorporates a motion decoder [7] that stabilizes training by introducing a reconstruction loss between the input motion and the decoded motion. Additionally, the model's encoders are VAE-based. Therefore, we add to the $\mathcal{L}_{tot}$ the terms for KL divergence, Reconstruction, and Latent Loss, as in [8].

Algorithm 1 Encoders' Input–Output Processing[1]

Require: Token sequence $\mathbf{x}_i$ (modality $i \in \{t, m, s\}$)
Ensure: • Unimodal latent vectors $\mathbf{v}_i, \mathbf{v}_j \in \mathbb{R}^{256}$
• Crossmodal latent vector $\mathbf{v}_{ij} \in \mathbb{R}^{256}$

$\tilde{\mathbf{x}}_i \leftarrow [\mu_i, \sigma_i, \mathbf{x}_i]$ ▷ prepend learnable tokens
$(\hat{\mu}_i, \hat{\sigma}_i, \epsilon_i) \leftarrow \mathrm{Enc}_i(\tilde{\mathbf{x}}_i)$
$\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$\mathbf{v}_i \leftarrow \hat{\mu}_i + \hat{\sigma}_i \odot \boldsymbol{\eta}$
Repeat Lines 1–4 for a second modality j to obtain $(\epsilon_j, \mathbf{v}_j)$
$\epsilon_{ij} \leftarrow [\epsilon_i, \epsilon_j]$ ▷ concatenate residual tokens
$(\hat{\mu}_{ij}, \hat{\sigma}_{ij}, -) \leftarrow \mathrm{Enc}_{ij}(\epsilon_{ij})$
$\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$\mathbf{v}_{ij} \leftarrow \hat{\mu}_{ij} + \hat{\sigma}_{ij} \odot \boldsymbol{\eta}$

9. Retrieval Results

In this section, we report the Retrieval Results at all ranks. In particular, in Table 4-9, and Table 10-15 we complement Table 2 and Table 1 of the Main Paper, respectively. In Table 16-21, and Table 22-27 we complement Table 4 and Table 3 of the Main Paper, respectively.

For what regards retrieval tasks, following the same trend described in the Main Paper, MonSTeR outperforms the other models on most tasks or is comparable to them.

Similarly, results reported in the extended ablation studies' tables confirm the trend already discussed in the Main Paper, with MonSTeR being the best performing model on average and beating the *"w/o cross-modal"*, *"w trimodal"* and *"w/o single"* variants.

[1]For readability we write the encoder outputs as $(\hat{\mu}, \hat{\sigma})$. In practice the encoder returns $(\hat{\mu}, \widehat{\log \sigma^2})$, and we recover $\hat{\sigma} = \exp(\frac{1}{2}\widehat{\log \sigma^2})$ before applying the re-parameterisation trick.

Absolute vs. Relative Features on TMR As discussed in Sec. 4 of the main paper, TMR uses Guo features from [4] to generate motion descriptors that are independent of absolute spatial position. The transformation process reorients each motion toward a common direction in an empty space, effectively removing its dependence on original scene coordinates. This leads to a notable performance boost—particularly on TRUMANS+, where the scene-motion correlation is already weak. To validate this, we evaluated TMR using Guo features without the alignment step. This resulted in a significant average performance drop on the *t2m/m2t* tasks of 4.71% on the "All" protocol and of 12.10% on the "Small Batches" protocol on TRUMANS+. Moreover, the use of Guo descriptors inherently prevents scene-related tasks, as the transformation eliminates information about the motion's original location within the scene.

10. Text-to-Motion results

In this section, we evaluate MonSTeR on HumanML3D dataset [4] for the task of Text-to-Motion retrieval. Neither the task or the dataset involve using scene information, therefore we remove the high-order modeling components, e.g. cross-modal encoders, from our network, resulting in a dual tower encoder structure. Additionally, we resort to representing motions as velocities and rotations in root space as commonly used for this task [1, 3, 8]. Table 2 and Table 3 present the experimental results, where bolded entries indicate the best performance among the compared models and underlined entries represent the second-best.

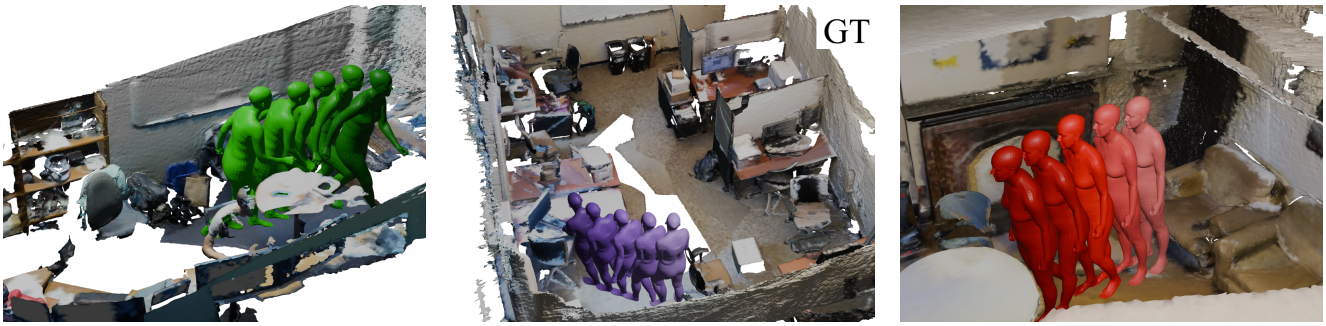
Table 2. *t2m* metrics on HumanML3D [4]

Protocol	Model	R@k (%)				
		R@1	R@2	R@3	R@5	R@10
All	TMR	8.92	12.04	16.33	22.06	33.37
	MoPa	**10.80**	**14.98**	**20.00**	**26.72**	**38.02**
	MonSTeR	7.07	11.20	15.25	20.43	30.55
Small Batches	TMR	67.45	80.98	86.22	91.56	95.46
	MoPa	**71.61**	**85.81**	**90.02**	**94.35**	**97.69**
	MonSTeR	67.38	82.96	88.89	93.75	97.45

Table 3. *m2t* metrics on HumanML3D [4]

Protocol	Model	R@k (%)				
		R@1	R@2	R@3	R@5	R@10
All	TMR	9.44	11.84	16.90	22.92	32.21
	MoPa	**11.25**	**13.86**	**19.98**	**26.86**	**37.40**
	MonSTeR	7.68	9.50	14.28	19.19	27.94
Small Batches	TMR	68.59	81.73	86.75	91.10	95.39
	MoPa	**72.11**	**85.26**	**90.21**	**94.44**	**97.76**
	MonSTeR	67.91	82.64	88.14	93.06	97.43

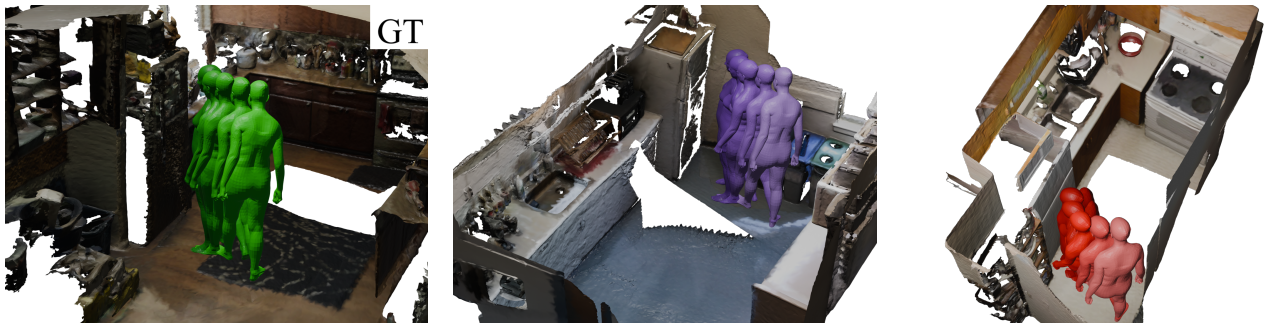
(a) A person is standing up from the bed, which is near a suitcase, and clothes are nearby.



(b) A person is walking to a chair near a desk.

(c) A person is walking to the toilet, passing by a sink.

Figure 5. Qualitative examples for *t2m*. **First**, **second**, and **third** retrieved samples are shown.



(a) A person is walking to the refrigerator, passing by a recycling bin, a cutting board, a trash can, and a sink, and then noticing a window.

(b) A person is walking to the chair, passing by a backpack, surrounded by books on a bookshelf.

(c) A person stands up from the toilet, with a mirror in sight.

Figure 6. Qualitative examples for *mt2s*. **First**, **second**, and **third** retrieved samples are shown.

Table 4. Retrieval Results on TRUMANS+ [6]: st2m and m2st

		st2m					*m2st*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	TMR	7.19	12.44	15.50	20.69	31.94	-	-	-	-	-
	TMR + S	1.06	2.00	2.81	4.75	9.56	1.25	2.31	3.56	5.75	9.31
	MoPa	1.50	2.75	3.62	5.19	8.06	-	-	-	-	-
	MoPa + S	1.38	2.62	3.62	6.25	10.25	1.38	2.69	3.81	6.06	12.06
	MonSTeR	3.81	6.75	9.56	13.06	19.50	3.94	6.81	8.75	12.50	19.44
Small Batches	TMR	57.31	78.12	85.06	92.93	97.56	-	-	-	-	-
	TMR + S	25.00	37.50	46.18	58.56	75.06	22.81	36.62	44.93	57.31	76.50
	MoPa	23.25	36.56	45.38	58.25	79.31	-	-	-	-	-
	MoPa + S	17.62	29.88	38.25	52.44	73.81	21.56	34.50	42.50	52.75	72.88
	MonSTeR	36.25	50.93	59.81	71.31	85.68	36.75	52.06	60.75	71.87	85.56

Table 5. Retrieval Results on TRUMANS+ [6]: ms2t and t2sm

		ms2t					*t2sm*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	TMR	6.44	11.38	15.38	21.75	33.62	-	-	-	-	-
	TMR + S	4.88	6.88	8.31	10.06	14.94	1.69	3.00	3.94	5.31	9.44
	MoPa	1.81	4.56	6.12	8.44	11.94	-	-	-	-	-
	MoPa + S	1.56	2.75	3.38	5.06	7.81	1.38	2.06	3.38	5.56	9.56
	MonSTeR	3.31	5.31	7.38	10.75	16.19	3.25	5.56	7.75	10.75	15.25
Small Batches	TMR	57.87	77.43	85.93	91.87	97.56	-	-	-	-	-
	TMR + S	31.25	43.00	51.37	64.12	80.50	26.50	40.93	49.00	63.37	79.43
	MoPa	29.56	45.75	55.31	67.50	84.06	-	-	-	-	-
	MoPa + S	26.00	40.69	50.25	64.62	82.25	27.12	41.81	50.69	63.12	82.19
	MonSTeR	32.75	47.31	57.31	69.06	84.56	31.56	47.87	57.68	68.81	83.37

Table 6. Retrieval Results on TRUMANS+ [6]: tm2s and s2mt

		tm2s					*s2mt*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	TMR	-	-	-	-	-	-	-	-	-	-
	TMR + S	4.56	6.56	8.12	10.88	15.56	0.44	0.75	1.38	1.94	3.56
	MoPa	-	-	-	-	-	-	-	-	-	-
	MoPa + S	5.75	9.19	11.56	15.94	23.56	0.56	0.94	1.62	2.50	5.06
	MonSTeR	7.44	9.69	11.44	14.44	20.56	1.19	2.31	3.31	4.75	8.12
Small Batches	TMR	-	-	-	-	-	-	-	-	-	-
	TMR + S	12.87	20.75	26.00	36.18	56.12	9.81	17.37	23.50	32.25	50.56
	MoPa	-	-	-	-	-	-	-	-	-	-
	MoPa + S	21.88	34.81	43.25	56.56	76.00	21.12	33.06	40.88	53.75	69.50
	MonSTeR	22.93	36.81	44.37	56.18	74.68	19.75	32.25	40.62	52.18	68.31

Table 7. Retrieval Results on TRUMANS+ [6]: t2m and m2t

		t2m					*m2t*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	TMR	7.19	12.44	15.50	20.69	31.94	6.44	11.38	15.38	21.75	33.62
	TMR + S	1.50	3.06	4.81	6.94	11.94	8.50	10.06	11.31	12.69	17.56
	MoPa	1.50	2.75	3.62	5.19	8.06	1.81	4.56	6.12	8.44	11.94
	MoPa + S	1.06	1.88	2.88	4.25	7.56	1.19	2.38	3.19	4.62	7.88
	MonSTeR	1.50	3.31	4.69	7.00	11.88	2.44	3.75	4.81	7.25	11.25
Small Batches	TMR	57.31	78.12	85.06	92.93	97.56	57.87	77.43	85.93	91.87	97.56
	TMR + S	28.06	42.37	51.31	64.31	80.87	30.93	43.31	53.18	65.68	82.50
	MoPa	23.25	36.56	45.38	58.25	79.31	29.56	45.75	55.31	67.50	84.06
	MoPa + S	19.38	30.38	39.94	51.38	69.94	19.94	31.38	38.94	50.81	70.69
	MonSTeR	26.00	39.31	47.50	59.87	74.75	26.93	39.56	49.06	59.68	75.43

Table 8. Retrieval Results on TRUMANS+ [6]: m2s and s2m

		m2s					*s2m*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	TMR	-	-	-	-	-	-	-	-	-	-
	TMR + S	2.12	3.50	4.69	7.56	11.56	0.12	0.44	0.75	1.25	3.00
	MoPa	-	-	-	-	-	-	-	-	-	-
	MoPa + S	6.00	7.69	10.31	12.69	17.25	0.50	0.94	1.44	2.25	5.19
	MonSTeR	4.31	5.38	6.75	8.56	14.38	0.50	0.94	1.62	2.38	4.25
Small Batches	TMR	-	-	-	-	-	-	-	-	-	-
	TMR + S	9.12	15.62	21.25	30.50	49.87	8.43	13.87	19.56	28.93	45.43
	MoPa	-	-	-	-	-	-	-	-	-	-
	MoPa + S	19.94	31.12	40.25	53.88	72.94	16.44	27.12	35.94	48.00	69.56
	MonSTeR	14.56	23.93	30.50	41.31	59.12	12.43	20.50	27.12	36.68	55.31

Table 9. Retrieval Results on TRUMANS+ [6]: s2t and t2s

		s2t					*t2s*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	TMR	-	-	-	-	-	-	-	-	-	-
	TMR + S	0.31	0.88	1.44	2.25	3.38	3.62	4.75	5.94	7.81	11.56
	MoPa	-	-	-	-	-	-	-	-	-	-
	MoPa + S	0.12	0.44	0.75	1.38	2.50	3.88	5.69	7.06	10.00	14.88
	MonSTeR	0.81	1.50	2.19	3.50	6.00	5.50	7.12	8.94	12.31	17.88
Small Batches	TMR	-	-	-	-	-	-	-	-	-	-
	TMR + S	8.87	15.37	21.31	30.62	50.68	11.68	19.25	26.06	36.75	56.56
	MoPa	-	-	-	-	-	-	-	-	-	-
	MoPa + S	9.88	18.00	24.69	36.56	56.50	13.06	21.00	26.69	38.12	59.06
	MonSTeR	15.75	25.06	31.37	42.18	61.06	16.37	26.06	33.81	46.12	65.87

Table 10. Retrieval Results on HUMANISE+ [11]: st2m and m2st

		st2m					*m2st*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	TMR	0.83	1.59	2.17	2.68	6.37	-	-	-	-	-
	TMR + S	1.27	2.04	3.06	4.78	9.37	0.83	1.78	2.61	4.08	7.20
	MoPa	1.21	2.74	4.02	5.48	8.99	-	-	-	-	-
	MoPa + S	0.64	1.15	1.53	2.55	4.65	0.57	1.40	1.78	2.80	5.67
	MonSTeR	5.23	8.54	11.66	17.02	27.09	4.72	8.41	11.41	16.57	24.60
Small Batches	TMR	21.62	38.5	50.62	73.12	96.12	-	-	-	-	-
	TMR + S	27.06	42.75	55.68	70.25	89.68	23.68	40.43	53.18	69.06	91.12
	MoPa	26.59	44.83	58.48	78.38	97.32	-	-	-	-	-
	MoPa + S	21.88	35.33	45.85	60.84	80.23	16.52	28.12	37.18	50.19	71.62
	MonSTeR	50.93	71.68	82.31	91.81	99.00	51.06	71.31	81.37	92.18	99.12

Table 11. Retrieval Results on HUMANISE+ [11]: ms2t and t2sm

		ms2t					*t2sm*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	TMR	0.45	1.15	1.66	2.61	4.40	-	-	-	-	-
	TMR + S	1.27	2.61	4.21	7.52	13.45	1.21	2.42	3.63	6.12	10.58
	MoPa	1.15	1.72	2.49	4.14	8.16	-	-	-	-	-
	MoPa + S	0.57	0.83	1.15	1.91	3.63	0.38	0.89	1.59	2.36	4.46
	MonSTeR	2.55	5.10	7.01	10.33	17.34	3.70	6.05	8.22	12.11	21.86
Small Batches	TMR	21.5	37.43	52.12	72.68	96.56	-	-	-	-	-
	TMR + S	39.62	60.12	72.68	85.56	95.87	33.75	54.68	67.93	83.12	95.75
	MoPa	28.51	46.56	58.99	77.10	95.79	-	-	-	-	-
	MoPa + S	16.71	27.61	36.10	49.36	72.51	16.20	27.04	35.33	46.88	70.73
	MonSTeR	44.43	66.87	80.18	91.50	98.50	47.43	69.00	81.31	90.68	98.68

Table 12. Retrieval Results on HUMANISE+ [11]: tm2s and s2mt

		tm2s					*s2mt*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	TMR	-	-	-	-	-	-	-	-	-	-
	TMR + S	0.06	0.38	0.57	1.34	3.06	0.38	0.89	1.59	2.42	4.59
	MoPa	-	-	-	-	-	-	-	-	-	-
	MoPa + S	0.83	1.53	2.29	3.76	7.97	0.70	1.27	2.17	4.02	7.14
	MonSTeR	0.83	1.85	3.06	4.91	9.82	0.96	2.23	3.44	5.67	9.94
Small Batches	TMR	-	-	-	-	-	-	-	-	-	-
	TMR + S	13.25	22.06	29.37	42.06	62.56	17.00	25.75	34.31	44.62	64.62
	MoPa	-	-	-	-	-	-	-	-	-	-
	MoPa + S	21.56	35.78	46.49	59.76	78.25	15.37	27.42	36.22	51.91	72.70
	MonSTeR	30.75	46.81	56.12	69.87	84.18	30.06	44.75	56.06	69.37	85.06

Table 13. Retrieval Results on HUMANISE+ [11]: t2m and m2t

		t2m					*m2t*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	TMR	0.83	1.59	2.17	2.68	6.37	0.45	1.15	1.66	2.61	4.40
	TMR + S	0.76	1.59	2.80	3.82	8.03	0.57	1.72	2.87	4.02	6.95
	MoPa	1.21	2.74	4.02	5.48	8.99	1.15	1.72	2.49	4.14	8.16
	MoPa + S	0.32	0.57	0.89	1.34	2.61	0.19	0.45	0.45	0.89	1.53
	MonSTeR	0.96	1.66	2.55	4.40	8.54	0.89	1.47	2.36	3.70	7.14
Small Batches	TMR	21.62	38.5	50.62	73.12	96.12	21.5	37.43	52.12	72.68	96.56
	TMR + S	27.62	42.50	55.81	75.12	96.06	25.31	41.68	54.12	73.50	96.25
	MoPa	26.59	44.83	58.48	78.38	97.32	28.51	46.56	58.99	77.10	95.79
	MoPa + S	9.06	16.58	23.79	35.20	57.72	8.74	16.26	22.58	33.93	53.51
	MonSTeR	29.06	48.00	61.31	80.68	96.93	26.75	45.50	59.81	79.87	97.50

Table 14. Retrieval Results on HUMANISE+ [11]: m2s and s2m

		m2s					*s2m*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	TMR	-	-	-	-	-	-	-	-	-	-
	TMR + S	0.13	0.70	0.83	1.27	2.17	0.38	0.57	1.02	1.66	3.44
	MoPa	-	-	-	-	-	-	-	-	-	-
	MoPa + S	0.38	1.27	1.59	2.74	5.04	0.51	1.02	1.59	2.29	3.89
	MonSTeR	0.32	0.70	0.96	1.59	2.23	0.57	0.83	1.02	1.15	2.04
Small Batches	TMR	-	-	-	-	-	-	-	-	-	-
	TMR + S	8.87	14.81	21.50	31.37	52.12	11.68	19.25	24.68	32.62	50.68
	MoPa	-	-	-	-	-	-	-	-	-	-
	MoPa + S	14.35	24.36	32.78	44.83	66.96	14.60	24.30	32.08	45.15	67.86
	MonSTeR	12.06	23.68	30.87	43.37	62.37	13.68	22.50	30.18	42.18	62.00

Table 15. Retrieval Results on HUMANISE+ [11]: s2t and t2s

		s2t					*t2s*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	TMR	-	-	-	-	-	-	-	-	-	-
	TMR + S	0.38	0.70	0.96	1.40	3.44	0.38	0.51	0.76	1.59	2.93
	MoPa	-	-	-	-	-	-	-	-	-	-
	MoPa + S	0.19	0.32	0.57	1.34	3.51	0.32	0.51	0.57	1.34	2.61
	MonSTeR	0.70	1.27	1.72	3.25	5.61	0.25	0.64	0.76	2.04	4.72
Small Batches	TMR	-	-	-	-	-	-	-	-	-	-
	TMR + S	14.50	24.56	34.50	49.43	69.25	13.87	25.81	33.75	47.81	70.31
	MoPa	-	-	-	-	-	-	-	-	-	-
	MoPa + S	13.07	24.17	32.33	46.94	70.73	12.56	22.32	30.48	43.24	67.22
	MonSTeR	22.25	35.62	47.37	62.12	81.43	21.81	36.12	47.50	63.37	82.12

Table 16. Retrieval Results on TRUMANS+ [6]: st2m and m2st

		st2m					*m2st*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	MonSTeR w/o cross-modal	1.875	3.4375	4.75	7.0625	12.3125	2.1875	3.9375	4.8125	7.75	12.625
	MonSTeR w trimodal	1.8125	3.3125	4.6875	7.25	11.0625	1.6875	2.8125	4.0	5.875	10.1875
	MonSTeR w/o single	4.25	8.5	12.0	16.3125	23.6875	5.4375	8.875	11.375	15.1875	24.1875
	MonSTeR	3.8125	6.75	9.5625	13.0625	19.5	3.9375	6.8125	8.75	12.5	19.4375
Small Batches	MonSTeR w/o cross-modal	26.62	41.06	49.75	60.5	76.87	26.56	40.37	48.87	62.56	77.25
	MonSTeR w trimodal	23.62	36.31	45.25	56.93	74.62	24.81	37.43	46.43	59.43	75.5
	MonSTeR w/o single	42.87	59.18	67.68	76.43	88.62	43.5	59.93	69.37	78.93	90.43
	MonSTeR	36.25	50.93	59.81	71.31	85.68	36.75	52.06	60.75	71.87	85.56

Table 17. Retrieval Results on TRUMANS+ [6]: ms2t and t2sm

		ms2t					*t2sm*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	MonSTeR w/o cross-modal	2.31	4.37	5.81	7.87	12.5	1.94	3.75	5.38	7.81	12.62
	MonSTeR w trimodal	2.06	3.87	5.43	7.25	12.56	2.06	4.06	5.69	7.06	12.81
	MonSTeR w/o single	2.62	4.06	6.00	8.50	13.87	1.88	3.69	5.12	7.44	13.25
	MonSTeR	3.31	5.31	7.37	10.75	16.18	3.25	5.56	7.75	10.75	15.25
Small Batches	MonSTeR w/o cross-modal	25.06	38.37	48.43	61.25	77.5	26.37	40.56	49.00	60.00	77.81
	MonSTeR w trimodal	26.25	40.37	49.37	61.12	76.75	27.12	38.93	47.75	58.81	75.43
	MonSTeR w/o single	31.43	45.56	55.37	67.75	81.81	31.62	46.68	54.62	66.81	82.56
	MonSTeR	32.75	47.31	57.31	69.06	84.56	31.56	47.87	57.68	68.81	83.37

Table 18. Retrieval Results on TRUMANS+ [6]: tm2s and s2mt

		tm2s					*s2mt*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	MonSTeR w/o cross-modal	6.31	8.94	10.69	13.94	20.19	0.88	1.69	2.44	4.00	6.50
	MonSTeR w trimodal	5.38	7.25	9.31	12.00	17.50	0.88	1.81	2.62	3.69	6.25
	MonSTeR w/o single	6.31	8.69	11.31	15.31	23.31	1.38	2.75	3.94	6.00	9.75
	MonSTeR	7.44	9.69	11.44	14.44	20.56	1.19	2.31	3.31	4.75	8.12
Small Batches	MonSTeR w/o cross-modal	17.62	27.56	35.43	45.81	64.18	15.12	24.37	30.62	41.56	58.25
	MonSTeR w trimodal	18.56	31.06	38.68	50.50	68.18	16.43	26.12	32.93	43.93	63.00
	MonSTeR w/o single	27.75	40.87	49.00	61.50	78.81	23.43	37.87	46.37	58.31	75.93
	MonSTeR	22.93	36.81	44.37	56.18	74.68	19.75	32.25	40.62	52.18	68.31

Table 19. Retrieval Results on TRUMANS+ [6]: t2m and m2t

		t2m					*m2t*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	MonSTeR w/o cross-modal	2.56	4.12	5.75	8.50	13.81	3.06	4.62	5.88	7.62	13.00
	MonSTeR w trimodal	2.38	3.94	5.75	7.62	11.88	2.75	4.31	5.12	7.50	11.81
	MonSTeR w/o single	0.00	0.06	0.12	0.25	0.50	0.00	0.00	0.00	0.00	0.31
	MonSTeR	1.50	3.31	4.69	7.00	11.88	2.44	3.75	4.81	7.25	11.25
Small Batches	MonSTeR w/o cross-modal	28.18	43.68	52.25	63.06	76.62	28.93	42.93	52.12	63.18	77.37
	MonSTeR w trimodal	26.43	38.25	46.68	58.93	74.18	25.56	38.50	47.50	60.06	74.56
	MonSTeR w/o single	3.00	5.68	9.50	15.31	29.00	2.81	5.56	8.18	15.18	30.43
	MonSTeR	26.00	39.31	47.50	59.87	74.75	26.93	39.56	49.06	59.68	75.43

Table 20. Retrieval Results on TRUMANS+ [6]: m2s and s2m

		m2s					*s2m*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	MonSTeR w/o cross-modal	4.19	5.56	7.50	10.44	17.56	0.62	1.00	1.62	2.31	4.31
	MonSTeR w trimodal	4.19	5.62	7.00	10.00	15.00	0.50	1.19	1.75	2.50	4.25
	MonSTeR w/o single	1.94	2.62	3.50	4.38	5.94	0.19	0.31	0.38	0.56	1.00
	MonSTeR	4.31	5.38	6.75	8.56	14.38	0.50	0.94	1.62	2.38	4.25
Small Batches	MonSTeR w/o cross-modal	13.37	22.68	30.18	41.18	60.37	12.06	20.31	26.00	36.93	54.25
	MonSTeR w trimodal	14.37	24.81	31.75	42.12	63.12	12.87	21.06	27.62	38.00	57.12
	MonSTeR w/o single	5.43	10.12	15.81	26.18	46.75	4.43	8.68	12.62	20.68	37.81
	MonSTeR	14.56	23.93	30.50	41.31	59.12	12.43	20.50	27.12	36.68	55.31

Table 21. Retrieval Results on TRUMANS+ [6]: s2t and t2s

		s2t					*t2s*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	MonSTeR w/o cross-modal	0.44	1.38	1.69	2.62	4.19	4.25	5.62	7.00	9.62	14.81
	MonSTeR w trimodal	0.88	1.38	1.88	3.50	5.25	4.00	5.38	6.94	9.88	14.75
	MonSTeR w/o single	0.00	0.00	0.00	0.12	0.25	1.44	1.81	2.00	3.19	5.12
	MonSTeR	0.81	1.50	2.19	3.50	6.00	5.50	7.12	8.94	12.31	17.88
Small Batches	MonSTeR w/o cross-modal	12.43	21.06	27.87	38.62	55.43	14.56	23.12	30.68	41.18	59.12
	MonSTeR w trimodal	14.18	23.37	29.87	41.12	59.06	16.06	25.93	33.81	46.06	63.56
	MonSTeR w/o single	2.56	5.12	8.31	14.31	30.62	3.56	8.00	11.87	20.06	36.00
	MonSTeR	15.75	25.06	31.37	42.18	61.06	16.37	26.06	33.81	46.12	65.87

Table 22. Retrieval Results on HUMANISE+ [11]: st2m and m2st

		st2m					*m2st*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	MonSTeR w/o cross-modal	1.78	3.00	4.27	6.25	10.71	1.08	1.91	2.87	4.27	8.73
	MonSTeR w trimodal	1.78	3.37	4.97	7.77	12.81	1.72	3.44	4.65	7.45	12.74
	MonSTeR w/o single	3.57	7.20	10.71	15.42	22.63	4.40	8.22	10.90	16.06	25.05
	MonSTeR	5.23	8.54	11.66	17.02	27.09	4.72	8.41	11.41	16.57	24.60
Small Batches	MonSTeR w/o cross-modal	33.18	51.75	64.31	81.50	96.62	29.00	49.25	62.56	79.31	96.18
	MonSTeR w trimodal	33.18	49.87	62.31	75.93	91.12	33.31	49.75	61.43	75.68	92.12
	MonSTeR w/o single	48.56	67.06	77.25	87.43	96.25	49.56	68.68	78.25	87.87	96.68
	MonSTeR	50.93	71.68	82.31	91.81	99.00	51.06	71.31	81.37	92.18	99.12

Table 23. Retrieval Results on HUMANISE+ [11]: ms2t and t2sm

		ms2t					*t2sm*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	MonSTeR w/o cross-modal	3.19	4.72	6.63	10.58	19.44	2.29	4.46	5.93	9.56	17.53
	MonSTeR w trimodal	2.74	4.46	6.05	9.56	14.97	0.63	1.52	2.80	3.82	6.81
	MonSTeR w/o single	2.04	4.59	6.88	10.01	18.23	2.68	5.04	7.07	10.71	19.31
	MonSTeR	2.55	5.10	7.01	10.33	17.34	3.70	6.05	8.22	12.11	21.86
Small Batches	MonSTeR w/o cross-modal	48.81	69.87	81.18	91.87	98.81	46.50	66.75	79.75	92.06	98.75
	MonSTeR w trimodal	39.12	58.75	71.06	82.87	93.31	21.62	35.50	45.75	59.62	78.50
	MonSTeR w/o single	46.81	67.68	79.81	90.31	97.50	45.37	67.50	78.43	89.31	96.87
	MonSTeR	44.43	66.87	80.18	91.50	98.50	47.43	69.00	81.31	90.68	98.68

Table 24. Retrieval Results on HUMANISE+ [11]: tm2s and s2mt

		tm2s					*s2mt*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	MonSTeR w/o cross-modal	0.25	0.96	1.59	3.38	6.25	0.70	1.34	2.04	3.06	5.80
	MonSTeR w trimodal	1.01	1.52	2.54	3.69	6.30	1.59	2.23	3.12	4.97	8.85
	MonSTeR w/o single	1.02	2.10	3.12	5.67	9.75	0.96	1.91	3.00	5.42	10.26
	MonSTeR	0.83	1.85	3.06	4.91	9.82	0.96	2.23	3.44	5.67	9.94
Small Batches	MonSTeR w/o cross-modal	22.25	36.25	46.56	61.50	82.87	21.43	37.06	47.50	61.06	79.50
	MonSTeR w trimodal	22.18	36.00	46.75	59.68	77.06	28.81	47.12	60.68	76.31	94.18
	MonSTeR w/o single	26.31	42.56	52.87	65.37	81.68	27.87	41.56	51.43	65.43	83.00
	MonSTeR	30.75	46.81	56.12	69.87	84.18	30.06	44.75	56.06	69.37	85.06

Table 25. Retrieval Results on HUMANISE+ [11]: t2m and m2t

		t2m					*m2t*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	MonSTeR w/o cross-modal	0.89	2.17	3.51	5.16	10.01	0.70	1.21	2.04	3.82	8.29
	MonSTeR w trimodal	1.59	2.23	3.12	4.97	8.85	1.40	2.23	3.37	5.73	9.11
	MonSTeR w/o single	0.00	0.13	0.13	0.19	0.64	0.00	0.06	0.25	0.51	0.64
	MonSTeR	0.96	1.66	2.55	4.40	8.54	0.89	1.47	2.36	3.70	7.14
Small Batches	MonSTeR w/o cross-modal	30.12	47.31	62.81	81.50	98.18	28.75	47.25	60.06	80.18	97.56
	MonSTeR w trimodal	28.81	47.12	60.68	76.31	94.18	27.31	44.93	57.62	75.56	94.56
	MonSTeR w/o single	7.06	11.87	16.62	25.68	43.50	4.62	7.75	11.31	19.87	42.18
	MonSTeR	29.06	48.00	61.31	80.68	96.93	26.75	45.50	59.81	79.87	97.50

Table 26. Retrieval Results on HUMANISE+ [11]: m2s and s2m

		m2s					*s2m*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	MonSTeR w/o cross-modal	0.32	0.45	0.57	0.96	2.23	0.25	0.45	0.70	1.27	2.36
	MonSTeR w trimodal	0.57	0.89	1.65	2.48	4.39	0.38	0.82	1.14	1.84	4.07
	MonSTeR w/o single	0.00	0.13	0.19	0.32	0.64	0.00	0.00	0.00	0.13	0.32
	MonSTeR	0.32	0.70	0.96	1.59	2.23	0.57	0.83	1.02	1.15	2.04
Small Batches	MonSTeR w/o cross-modal	12.25	20.93	27.87	40.31	61.25	12.87	21.12	27.37	39.81	60.00
	MonSTeR w trimodal	15.87	24.31	32.31	44.68	65.68	15.68	25.56	33.50	45.18	65.81
	MonSTeR w/o single	4.56	7.18	9.62	15.56	28.12	5.25	8.06	10.81	16.75	31.43
	MonSTeR	12.06	23.68	30.87	43.37	62.37	13.68	22.50	30.18	42.18	62.00

Table 27. Retrieval Results on HUMANISE+ [11]: s2t and t2s

		s2t					*t2s*				
Protocol	**Method**	rank1	rank2	rank3	rank5	rank10	rank1	rank2	rank3	rank5	rank10
All	MonSTeR w/o cross-modal	0.89	1.21	1.85	3.38	5.86	0.57	1.34	1.78	2.74	6.18
	MonSTeR w trimodal	0.44	0.95	1.27	2.67	5.16	0.31	0.76	1.21	1.78	4.07
	MonSTeR w/o single	0.00	0.13	0.19	0.32	0.57	0.13	0.25	0.25	0.70	1.34
	MonSTeR	0.70	1.27	1.72	3.25	5.61	0.25	0.64	0.76	2.04	4.72
Small Batches	MonSTeR w/o cross-modal	21.25	33.93	46.18	61.37	82.00	21.43	35.43	46.31	62.56	82.56
	MonSTeR w trimodal	19.50	33.56	43.93	56.50	75.50	18.75	31.00	41.87	56.56	75.31
	MonSTeR w/o single	5.18	8.25	10.43	16.12	34.12	6.81	10.37	13.00	19.43	35.50
	MonSTeR	22.25	35.62	47.37	62.12	81.43	21.81	36.12	47.50	63.37	82.12

References

[1] Léore Bensabath, Mathis Petrovich, and Gül Varol. Tmr++: A cross-dataset study for text-based 3d human motion retrieval. In *CVPRW*, 2024. 7

[2] 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. 1

[3] 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. 7

[4] 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, 7

[5] 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. 1

[6] 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. 1, 4, 10, 11, 14, 15

[7] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10985–10995, 2021. 7

[8] Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *ICCV*, 2023. 4, 7

[9] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 7

[10] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 1, 4

[11] 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, 12, 13, 16, 17

[12] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao 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. 4, 5, 6

[13] Qing Yu, Mikihiro Tanaka, and Kent Fujiwara. Exploring vision transformers for 3d human motion-language models with motion patches. In *CVPR*, 2024. 4

[14] Zhu Ziyu, Ma Xiaojian, Chen Yixin, Deng Zhidong, Huang Siyuan, and Li Qing. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *ICCV*, 2023. 7