



SAPIENZA
UNIVERSITÀ DI ROMA

Modeling Virtual Humans and Scenes

Department of Computer, Control and Management Engineering
National PhD Program in Artificial Intelligence (XXXVIII cycle)

Massimiliano Pappa

ID number 1661249

Advisor

Prof. Fabio Galasso

Co-Advisor

Prof. Indro Spinelli

Academic Year 2025/2026

Abstract

Human body features like size, shape, and pose are key elements in Human-Centric Computer Vision. They determine where people are, what they do, and how much space they occupy in the environment. This thesis investigates three aspects of computer vision, all connected by human presence: crowd analysis, motion generation, and human interaction with scenes. Perceiving and counting people in a crowd is useful for public safety and incident prevention. Motion generation helps creators improve the quality and smoothness of their content with minimal effort. Human-Scene Interaction enables motions to realistically interact with environments, and the evaluation of such types of solutions is critical. Current methods in these domains often fail to meet key user requirements.

Practical applications demand an understanding of the physical space people occupy, a requirement that mere counting fails to provide. To address this, we first introduce the new task of Crowd Volume Estimation: predicting the total human body volume present in a scene from a single RGB image. We release ANTHROPOS-V, a photorealistic benchmark with per-person and per-part supervision derived from anatomically plausible meshes and based upon anthropometric priors. Training on per-part volume density maps spreads supervision beyond heads to torsos and limbs, making estimates robust to occlusion and scale. Our proposed model, STEERER-V, outperforms counting and human mesh recovery baselines on the Crowd Volume Estimation task, and can easily transfer its capabilities to real images, advancing crowd analysis from simple counts to a more practical, volume-based analysis for safety, planning, and comfort.

A central challenge in motion generation is ensuring that outputs are plausible and coherent, not just diverse. To tackle the problem of implausible and uncontrolled generation, we propose MoDiPO, a method to align text-to-motion diffusion models. MoDiPO adapts Direct Preference Optimization to the motion domain and replaces costly human labels with AI feedback. For each prompt, an AI ranker constructs preference sets over candidate motions; We then align the generator toward winner motions, repulsing loser motions, while preserving diversity. With a new dataset made of motion-preference pairs, named Pick-a-Move, MoDiPO improves FID, human preference, and prompt faithfulness without any mode collapse, turning diffusion-based text-to-motion models into preference-aligned motion generators.

True understanding requires contextual coherence across motion, intention, and the environment. To achieve this, we propose MonSTeR, a tri-modal retrieval model that embeds motion, scene, and text in a unified latent space. This space is trained via coupled unimodal and cross-modal encoders, supporting flexible, all-direction retrieval. This versatile representation enables downstream tasks like zero-shot in-scene object placement and enhanced scene-aware motion captioning. MonSTeR serves as a powerful evaluator for Human-Scene Interaction models. We validate this capability by showing that its scores correctly penalize physically implausible interactions such as those derived from path rotations and scene collisions.

Together, these contributions move from population-level capacity (volume) to individual-level plausibility (aligned motion) to contextual coherence (motion-scene-text alignment).

Contents

List of Figures	v
List of Tables	ix
1 Introduction	1
2 ANTHROPOS-V	5
2.1 Overview	5
2.2 Introduction	5
2.3 Related Works	7
2.3.1 Single subject VE in controlled environments	7
2.3.2 Crowd Counting	8
2.3.3 Human Mesh Recovery for Few Individuals	8
2.4 Measuring Crowds Volumes	9
2.4.1 Problem Formalization	9
2.4.2 Proposed Metrics	9
2.5 Crowd Volume Estimation	10
2.5.1 From HMR to CVE	10
2.5.2 From Crowd Counting to CVE	10
2.5.3 Per-part Volume Density Maps	11
2.6 Experiments	12
2.6.1 Experimental Results	12
2.6.2 Qualitative Evaluation	14
2.6.3 From ANTHROPOS-V to real images	15
2.7 The ANTHROPOS-V dataset	17
2.7.1 Dataset Generation	17
2.7.2 Alignment to real-world body-types	18
2.7.3 SMPL Fitting	19
2.7.4 Dataset Statistics	20
2.8 Limitations and Future Works	20
2.9 Conclusion	21
3 MoDiPO	23
3.1 Overview	23
3.2 Introduction	23
3.2.1 Aligning Models with Human Feedback	25
3.2.2 Aligning Models with Synthetic Feedback	26

3.2.3	Text-to-Motion	26
3.3	Background	27
3.3.1	Variational Auto-Encoder for Human Representation	27
3.3.2	DDPMs	27
3.3.3	Latent DDPMs	28
3.3.4	Preference Alignment Overview	28
3.4	Methodology	29
3.4.1	Motion Preferences Dataset	29
3.4.2	Motion Alignment Pipeline	30
3.5	Experimental Evaluation	31
3.5.1	Evaluation Metrics	31
3.5.2	Quantitative Evaluation	33
3.5.3	Qualitative Evaluation	36
3.5.4	User Study	36
3.6	Ablation Studies	37
3.6.1	Pair Selection	38
3.6.2	Ranking Method	38
3.6.3	Alignment with generations from other models	38
3.6.4	Online Alignment	39
3.7	Conclusions	39
4	MonSTeR: a Unified Model for Motion, Scene, Text Retrieval	41
4.1	Overview	41
4.2	Introduction	41
4.3	Related Works	43
4.3.1	Text-Motion Retrieval	43
4.3.2	Beyond Two-Modal Alignment	44
4.3.3	Modality Aggregation	44
4.4	MonSTeR	44
4.4.1	Data Representation	44
4.4.2	Designing MonSTeR’s latent space	45
4.4.3	Training MonSTeR	46
4.5	Experiments	47
4.5.1	Datasets	48
4.5.2	Retrieval Tasks	49
4.5.3	Ablation Studies on the Model Design	53
4.5.4	MonSTeR for Evaluation	54
4.5.5	Downstream Tasks	56
4.6	Limitations	58
4.7	Conclusion	58
5	Conclusions	59
	Publications	61
1	International Conferences	61
2	Under Submission	61

List of Figures

2.1	ANTHROPOS-V is the first dataset for the novel task of Crowd Volume Estimation. It features human crowds engaged in activities within real-world environments (<i>left</i>). Each individual in the dataset is labeled with ground-truth SMPL shape parameters (<i>center</i>) along with their volume labels. We employ novel Per-Part Volume Density Maps as a superior supervision signal for training models to address this task (<i>right</i>).	6
2.2	STEERER-V's PP-MAE/MAE computed on ANTHROPOS-V test set.	9
2.3	Bayesian+ and MAN modified architectures. The Transformer Encoder blocks with <i>Learnable Region Attention</i> are used only by MAN. The green layers are kept as they are in the original architectures. Orange layers are part of the additional branch tailored on the CVE task. The estimated density map (D^{est}) is reduced with summation to a single value and counting losses are calculated on it. D^{est} is also used as input for our additional volume-related branch to regress the volume occupied in a frame. Finally, we calculate the L1 loss between regressed and ground truth volumes.	11
2.4	Zero-shot results of STEERER-V on CrowdHuman  images when predicting the volume of adults. On each predicted volume map we superimpose the difference between the average real-world per-person volume and the predicted per-person one.	16
2.5	Zero-shot results of STEERER-V on CrowdHuman  images when predicting the volume of builds that have not been seen at train time, e.g., kids. On each predicted volume map we superimpose the difference between the average real-world per-person volume and the predicted per-person one.	17
2.6	Examples from ANTHROPOS-V, showcasing several lighting and weather conditions, camera angles, and a variety of physiques. The crops in the zoomed boxes depict persons with differences in statures and body shapes.	18

2.7	Statistical analysis of the distributions of mass, height, and <i>Body Mass Index</i> (BMI) of the individuals in ANTHROPOS-V. Solid curves depict the empirical distributions, while dashed curves refer to the theoretically expected ones [2]. (2.7a) Distribution of the body features of the characters' meshes in GTA-V without any manipulation. (2.7b) Distribution of the body features of the characters' meshes in GTA-V after applying some geometrical transformations. (2.7c) Distribution of the body features of the resulting fitted SMPL meshes.	19
2.8	A SMPL mesh from ANTHROPOS-V. Different body parts are highlighted in different colors. Orange dots represent the keypoints associated with each body part.	20
3.1	MoDiPO (in yellow) leverages the DPO alignment to generate enhanced human motions.	24
3.2	MoDiPO Schematics. Starting with an input prompt, we generate K motions with a frozen reference model. These motions are then ranked according to their relevance with the input description. This motions-text-ranking tuple constitutes a sample in our "Pick-a-Move" (PaM) preferential dataset. At train time, we divide motions into two sets of best and worst generations, according to these rankings. By sampling from these sets, we form a winner-loser pair that is used to refine the target model using DPO.	32
3.3	Qualitative Examples	35
3.4	User Study Interface	37
4.1	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.	42
4.2	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})	45
4.3	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 $C_{i,j}$	47
4.4	Comparison between HUMANISE captions and HUMANISE+ recapitoning.	49
4.5	Qualitative examples for $st2m$ (4.5a, 4.5b, 4.5c) and $ms2t$ (4.5d, 4.5e, 4.5f). First , second , and third retrieved samples are shown. In each pictorial, GT=color (<i>top-left corner</i>) indicates the correct corresponding motion (<i>top row</i>) and text (<i>bottom row</i>)	51

4.6	Qualitative examples for $t2m$. First , second , and third retrieved samples are shown.	52
4.7	Qualitative examples for $mt2s$. First , second , and third retrieved samples are shown.	53
4.8	MonSTeR's FID $\downarrow$ (a) and Recall@1 $\uparrow$ (b) trends when motions are increasingly rotated, from 0 to π radians.	55
4.9	In-Scene Object Placement scores visualized	56
4.10	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).	57

List of Tables

2.1	Results on ANTHROPOS-V, reported in dm^3 . Inference time (ms) is measured on an NVIDIA A100. Methods are divided into HD+HMR, Crowd Counting, and our proposed approach. $C_{B+}(I)$ refers to the headcount given by [3], while $V_{\mathcal{D}}$ is the average human volume. Grayed-out lines use oracular information and should not be directly compared with the other results.	13
2.2	Results on ANTHROPOS-V's S1, S2 and whole test set (FT). All the results are reported in dm^3	14
2.3	Evaluation on 3DPW and CrowdHumans (CH). Reported metrics are MAE/PP-MAE (3DPW) and the difference between the average real-world per-person volume and the predicted per-person one (CH). B-CLIFF stands for BEDLAM-CLIFF.	15
2.4	Visual results of baseline models and STEERER-V on ANTHROPOS-V, along with the Ground Truth Per-Part-Volume Density Maps (GT Per Part). STEERER's concentrates the volume on the heads, whereas STEERER-V distributes it across the entire body.	15
3.1	Results for HumanML3D. Each group of results corresponds to the base models, aligned models, and current SOTA backed by a RAG framework. <i>Italic</i> is for best results overall, while bold represents best results among base models and MoDiP0-aligned ones.	33
3.2	Results for KIT-ML. Each group of results corresponds to the base models, aligned models, and current SOTA backed by a RAG framework. <i>Italic</i> is for best results overall, while bold represents best results among base models and MoDiP0-aligned ones.	34
3.3	Performance of Winner (W) and Loser (L) among the 8 generations per-prompt of MLD and MLD aligned with MoDiP0 (Ours) for HUMANML3D.	34
3.4	Performance of Winner (W) and Loser (L) among the 8 generations per-prompt of MLD and MLD aligned with MoDiP0 (Ours) for KIT-ML.	34
3.5	Ablation study with MLD model on HumanML3D dataset. Stoch stands for stochastic selection of pairs, default is edge selection. GT_p is the probability of ground truth being used as winner. Online for online generation and ranking with the reference model. PaM+ is PaM with generations of both MLD and MDM used for the winner-loser pair selection.	37

4.1	Tasks and protocols' results on HUMANISE+ [4] 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.	48
4.2	Tasks and protocols' results on TRUMANS+ [5] 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.	50
4.3	Ablation Studies for MonSTeR performed on HUMANISE+ [4]. We report the average mRecall computed across ranks.	50
4.4	Ablation Studies for MonSTeR performed on TRUMANS+ [5]. We report the average mRecall computed across ranks.	51
4.5	Generation Results on HUMANISE+ [4].	55
4.6	Captioning performances on HUMANISE+ [4].	58
4.7	Captioning performances on TRUMANS+ [5].	58

Chapter 1

Introduction

Accurately perceiving and modeling human beings is a foundational challenge in artificial intelligence, essential for systems that interact with and operate in human-centric environments. The body's physical characteristics, like size, shape, pose, and motion, are the fundamental primitives through which we interpret human attributes and behaviors.

In the field of Human-Centric Computer Vision, these features are paramount, as they determine where people are, what they do, and how much space they occupy within an environment. Unlike rigid objects, the human form is highly articulated, deformable, and presents immense variability in appearance, making its analysis a persistent and challenging frontier. A deep understanding of these elements is therefore critical for both perception tasks that can range from ensuring public safety to analyzing nuanced social behaviors, as well as generation tasks, like creating realistic and interactive digital agents for virtual worlds. This thesis explores three distinct yet interconnected domains within this field: the analysis of human crowds, the synthesis of human motion, and the contextual interplay between humans and their environments.

The real-world importance of these domains demands reliable solutions. The ability to perceive and analyze crowds is crucial not only for ensuring public safety in high-density environments but also for optimizing architectural design, managing pedestrian flow in public transport hubs, and ensuring structural integrity under heavy pedestrian loads.

The generation of realistic human motion is a vital tool for animators, game developers, and creators, enabling the production of high-quality digital content while reducing manual effort. Beyond entertainment, it is essential for creating digital twins, performing ergonomic simulations, and synthesizing training data for other AI models.

Finally, modeling the intricate relationship between a person's actions, their intentions, and their physical surroundings are a foundational requirement for developing the next generation of intelligent systems, from assistive robotics and augmented reality to advanced autonomous driving systems that must predict pedestrian behavior.

However, while significant progress has been made in these domains, current solutions often fail to meet key user requirements for practical, real-world applications.

This thesis argues that there are three critical gaps between the state-of-the-art and what is truly needed. First, in crowd analysis, a simple headcount is a shallow metric that ignores critical factors like weight and occupancy, which are vital for preventing structural overloading and ensuring comfort. A true understanding thus requires quantifying the physical space people occupy. Second, in motion generation, the inherent stochasticity of powerful models like diffusion models often leads to physically implausible or undesirable motions. This unrestricted diversity must be aligned with user preferences for plausibility. Third, in human-scene interaction, a unified method for evaluating the realism of a generated motion within its environment is completely absent. Existing metrics are too weak, failing to capture the contextual coherence between motion, intention, and the scene.

This thesis addresses these three weaknesses directly, advancing the field of human-centric computer vision along each of these axes. Our first contribution moves beyond traditional crowd analysis by introducing the novel task of Crowd Volume Estimation, which uses volume as a robust proxy for weight and occupancy. Since no datasets existed for this task, we developed ANTHROPOS-V, the first large-scale, photorealistic synthetic dataset with per-person volume annotations. To obtain this dataset, we manipulated the 3D meshes of GTA-V characters to align their anthropometrics with real-world human height and weight distributions, significantly reducing the sim-to-real gap. We then propose STEERER-V, a model that outperforms both counting and Human Mesh Recovery baselines by using a novel, volume-specific supervision signal, providing a more meaningful and capacity-based understanding of crowds.

Second, we tackle the problem of plausibility and control in generative motion models. We introduce MoDiPO, a framework that aligns text-to-motion diffusion models with user preferences using Direct Preference Optimization. To circumvent the costly process of gathering human preferences, we pioneer the use of AI feedback from retrieval models to automatically create a motion-preference dataset, which we name Pick-a-Move. MoDiPO leverages this data with a dynamic sampling strategy that prevents overfitting and mode collapse. Our experiments demonstrate that this alignment process yields significantly more realistic motions, securing a threefold increase in preference from human evaluators compared to the original model’s outputs.

Finally, to address the need for a unified evaluation metric in human-scene interaction, we present MonSTeR, the first tri-modal retrieval model designed to assess the coherence between motion, scene, and text. Using coupled unimodal and cross-modal encoders to capture higher-order relationships, MonSTeR embeds all three modalities into a unified latent space. While this space supports flexible retrieval, its most critical application is serving as a holistic, scene-aware evaluator. We validate this by showing that its scores strongly correlate with human judgments of plausibility, correctly penalizing contextually absurd interactions like interpenetrations, thus providing the first reliable, global metric for contextual coherence in this domain.

The remainder of this thesis is structured to detail these contributions. Chapter 2 introduces the task of Crowd Volume Estimation and presents our ANTHROPOS-V dataset beside STEERER-V that demonstrates its superior performance over traditional counting methods. Chapter 3 addresses the challenge of plausibility

in motion generation, detailing our approach for text-to-motion models alignment. Chapter 4 presents MonSTeR, a model for tri-modal retrieval and validates its effectiveness as a tool for evaluating contextual coherence. Finally, Chapter 5 will conclude with a summary of the contributions.

Chapter 2

ANTHROPOS-V: benchmarking the novel task of Crowd Volume Estimation

2.1 Overview

We introduce the novel task of **Crowd Volume Estimation (CVE)**, defined as the process of estimating the collective body volume of crowds using only RGB images. Besides event management and public safety, CVE can be instrumental in approximating body weight, unlocking weight-sensitive applications such as infrastructure stress assessment, and assuring even weight balance. We propose the first benchmark for CVE, comprising *ANTHROPOS-V*, a synthetic photorealistic video dataset featuring crowds in diverse urban environments. Its annotations include each person’s volume, SMPL shape parameters, and keypoints. Also, we explore metrics pertinent to CVE, define baseline models adapted from Human Mesh Recovery and Crowd Counting domains, and propose a CVE-specific methodology that surpasses baselines. Although synthetic, the weights and heights of individuals are aligned with the real-world population distribution across genders, and they transfer to the downstream task of CVE from real images.

2.2 Introduction

Dealing with large gatherings in public spaces presents significant challenges in crowd management: *overcrowding* can jeopardize the safety, health, and comfort of individuals, while the assembly of crowds on structures not designed for high capacity poses risks of structural damage or collapse due to *overloading* [6].

Currently, the monitoring of crowds’ risks based on head count [7, 8] tends to disregard potentially critical factors such as weight, occupancy, heat dissipation, and oxygen consumption, which are strongly correlated with individuals’ body build [6, 9]. Additionally, these factors exhibit significant variability based on age, gender, ethnicity, and health conditions [10, 11, 12].

A precise estimate of the *crowd’s total volume* offers a more reliable method for detecting space underuse or overcrowding by leveraging a priori knowledge of the



Figure 2.1. ANTHROPOS-V is the first dataset for the novel task of Crowd Volume Estimation. It features human crowds engaged in activities within real-world environments (*left*). Each individual in the dataset is labeled with ground-truth SMPL shape parameters (*center*) along with their volume labels. We employ novel Per-Part Volume Density Maps as a superior supervision signal for training models to address this task (*right*).

available in-place volume. Additionally, this approach can significantly mitigate the risks associated with overloading, as volume serves as a robust proxy for estimating weight [13, 14].

Motivated by these insights, this paper introduces the novel *Crowd Volume Estimation* (CVE) task, which aims to estimate the collective volume occupied by groups of individuals directly from single RGB images.

Recently, volume estimation (VE) has garnered significant research attention. However, previous works [15, 16, 17, 14, 18] have focused primarily on estimating the volume of single individuals in controlled environments, often relying on expensive annotations. These works rely on requirements and assumptions that render them impractical for estimating the volume of large crowds. In addition, there are no datasets tailored for CVE, as existing datasets [19, 20] featuring scenes with multiple people only encompass a limited number of individuals (≤ 15) or lack necessary volume annotations [21, 22]. We thus propose the first CVE benchmark, including baselines, a novel dataset, and metrics.

As for the baselines, we outline two research directions for crafting models for CVE: (1) grounding on methods from the Crowd Counting [23, 24, 25, 3] domain, or (2) repurposing the pipeline of Human Mesh Recovery (HMR) [26, 27, 20] combined with a human detector. In the case of (1), we extend the density-map strategy to a more specific form of supervision, the *Volume Density Maps*, tailored to regress the human volumes in an image. For these models, we further define a more fine-grained form of supervision which we dub *Per-Part Volume Density Maps* (rightmost picture in Fig. 2.1). This supervision allows the model to learn how to regress the volume of different human subparts (e.g., arms, chest, legs), resulting in a more accurate estimate of each individual’s volume. In the case of (2), we show that HMR models with a human detection preprocessing can serve as volume estimators without any modification, as the volume estimate comes as a by-product after postprocessing the predicted meshes (middle picture in Fig. 2.1).

Moreover, to enable the training of CVE-specific models, we introduce “ANTHROPometrics POse Shape and Volume estimation dataset” (ANTHROPOS-V), a synthetic, large-scale, and video-realistic dataset representing large crowds in urban scenarios and reporting for the first time the annotated volume of each individual appearing in the scenes. ANTHROPOS-V is generated using the videogame engine

of Grand Theft Auto V (GTA-V), which includes a large variety of realistic urban environments and a broad diversity of characters’ ethnicities.

It is worth noting that acquiring real-world data with annotated anthropometric features for crowds requires gathering sensitive information from thousands of individuals in a wide range of environments, hence posing severe issues of feasibility, bias, and personal details disclosure. Thus, we collect a synthetic dataset to train models on CVE, evaluating their learned knowledge of both synthetic and real-world data.

Aiming at reducing the domain gap from synthetic to real images, we enhance the game’s appearance and perform an in-depth analysis of the GTA-V default characters. Our findings reveal that the default characters exhibit a restricted and repetitive assortment of human anthropometrics, such as body sizes and heights. Therefore, we directly manipulate characters’ 3D meshes to align them with the real-world human size distributions [28]. As a result, we improve both the realism of the original GTA-V scenes and their annotations, featuring virtual characters whose height and weight distributions closely follow the authentic human variations [2].

In summary, our contributions are four-fold:

- we propose the novel task of *Crowd Volume Estimation* (CVE) to regress the volume of large groups of people from RGB images;
- we release the first CVE benchmark, including metrics and baselines;
- we introduce ANTHROPOS-V, a dataset explicitly devised for CVE but also encompassing annotations for other human-centric tasks, with careful attention to mirror real-world anthropometric and gender statistics;
- we experiment with a novel volume-specific form of supervision, namely Per-Part Volume Density Maps, and use it to train our proposed model, STEERER-V, achieving superior results.

2.3 Related Works

In this section, we review the existing literature that relates to the proposed CVE task. We discuss studies on single-subject Volume Estimation (VE) (Sec. 2.3.1), Crowd-Counting (Sec. 2.3.2) and Human Mesh Recovery (Sec. 2.3.3).

2.3.1 Single subject VE in controlled environments

Previous literature explored volume estimation, targeting single-bodies [15, 16, 17, 14, 18] or objects [29, 30, 31, 32, 33, 34, 35] for applications in healthcare and nutrition. In particular, [17] rely on 3D scans, while [15, 31, 32, 14] exploit depth-maps or point clouds data. To deal with scale ambiguity, [30, 36, 33, 34, 35] make use of reference objects, while [29, 18] employ multiple images of the same object in different views. While all the mentioned works tightly depend on controlled environments, scans, or multiple inputs, and apply to a single subject at a time, we aim at estimating the total volume of human crowds in the wild. Notably, [16] proposed a large-scale video dataset displaying individual textured SMPL meshes [37], paired with body-part

volume ground truths. However, scenes are designed by superimposing a single mesh onto 2D bedroom images, lacking realism and scale consistency between humans and the background. On the contrary, we propose a dataset of realistic scenes featuring large crowds.

2.3.2 Crowd Counting

Crowd Counting aims to estimate the number of people in images or videos. Typically, datasets in this domain showcase large crowds from bird-eye views [38, 39, 40, 41, 42, 43]. While seminal works [44, 45, 46] cast this problem as a regression task, recent literature [23, 47, 25] address crowd counting as a localization task. These methods regress the 2D positions of the heads in the images and estimate the total number by summing the retrieved outcomes after filtering the more uncertain predictions. Density-Map-based methods [48, 49, 50, 51, 52, 53, 54, 55, 24, 56] express the ground truth density map y for an image x as a single-channel image of the same size, where each pixel is assigned 1 if it contains the center of a person’s head, 0 otherwise; y is subsequently smoothed with a Gaussian filter. The Gaussian filtering operation is common in Counting tasks [55, 57, 58, 59, 60], as it allows to treat the GT density map as a continuous function [39], which, in turn, allows the end-to-end training of the network. Bayesian-based approaches differ from conventional density-based methods as they estimate density maps without supervision upon ground truth density maps. For instance, [25, 3] employ a bayesian-loss to construct a density contribution probability model starting from bare annotations. We evaluate strong Density-Map [24], Localization [23], and Bayesian models [3, 25] for the CVE task. We introduce baselines capitalizing on an adaptation of the Density Maps approach for CVE, namely Volume Density Maps: each pixel containing a person’s head is assigned with the whole person’s body volume instead of 1.

2.3.3 Human Mesh Recovery for Few Individuals

Human Mesh Recovery (HMR) regresses the human 3D shape and pose from single RGB images. CLIFF [27] pairs cropped image features with their bounding box information, enabling the accurate prediction of global rotations. In BEDLAM-CLIFF [20], the authors train CLIFF on their dataset and improve its performance. ReFit [26] exploits a recurrent updater that iteratively adjusts a parametric human model to align with image features. The recent TokenHMR [61] uses a tokenized representation of the human pose and reformulates the problem as a token prediction. DPMesh [62] leverages a diffusion model to meliorate robustness to occlusions. Crowd3DNet [63] focuses on mesh reconstruction of people within crowds in wide-field images, though tightly assuming the existence of a common plane where all the actors lie; such assumption does not hold for the complex scenes of ANTHROPOS-V. Similarly, [64] exploits pseudo-GT to model the relations and interactions of the individuals and improve pose and localization estimation; however, this work does not focus on human shape, as no shape-related metric is employed. Contrarily, CVE requires precise volume/shape GT for correct computation (Sec. 2.4.2). Thus, we repurpose [27, 20, 26] for CVE, pairing them with a human detector.

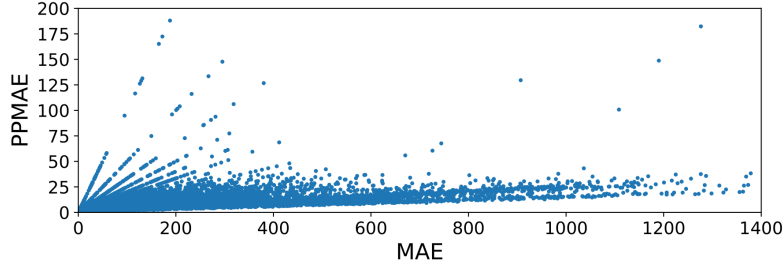


Figure 2.2. STEERER-V’s PP-MAE/MAE computed on ANTHROPOS-V test set.

2.4 Measuring Crowds Volumes

In this section, we formalize CVE (cf. Sec. 2.4.1) and define the metrics for the novel CVE benchmark (cf. Sec. 2.4.2).

2.4.1 Problem Formalization

We define **Crowd Volume Estimation** as the task of estimating the undergarment total body volume occupied by human bodies in a given scene. While CVE can be applied to videos (that we make available in ANTHROPOS-V), we define the CVE task to be benchmarked per frame.

Let I be an image and V_{tot} the label of the actual total volume of human bodies represented in I . We define the objective of CVE as $\min_{\theta} ||V_{tot} - M_{\theta}(I)||$, where M_{θ} is a crowd volume estimation function parameterized by θ . This definition is intentionally general and designed to be independent of any specific methodology. Indeed, this formulation enables its application as an objective for both Crowd Counting and HMR models, as well as for our proposed method.

2.4.2 Proposed Metrics

Leveraging earlier research in Crowd Counting [39, 40, 43, 24], we propose to measure the volume estimation error with the following suite of metrics:

Mean Absolute Error (MAE), as a standard measure to assess the quality of the estimations. Given a set of images $\{I_k\}$, we indicate with $\{V_k\}$ the total volume associated to each image and with $\{\hat{V}_k\}$ the estimated one. Hence:

$$\text{MAE} = \frac{1}{K} \sum_{k=1}^K |V_k - \hat{V}_k| \quad (2.1)$$

This measure estimates the accuracy of the predictions on the whole test set as K represents the total count of the images belonging to the test set.

Per-Person Mean Absolute Error (PP-MAE), to measure the average error for each individual. Its formulation can be expressed as:

$$\text{PP-MAE} = \frac{1}{K} \sum_{k=1}^K \frac{|V_k - \hat{V}_k|}{n_k} \quad (2.2)$$

where n_k is the number of persons in the k -th image. PP-MAE highlights the estimator’s mean error per person, enabling evaluation across scenes and datasets with different numbers of individuals. This is advantageous in CVE, as crowd sizes vary widely among frames. It is worth noting that PP-MAE is related to the Normalized Absolute Error metric used in Crowd Counting [24], which normalizes the mean counting error by the total number of persons.

Although PP-MAE and MAE appear to be closely related, Fig. 2.2 confirms the absence of a direct correlation between the two metrics. The lines composed of aligned points in Fig. 2.2 correspond to frames with a fixed human count (as their slope MAE/PP-MAE equals the number of individuals). Thus, comparing MAE vs. PP-MAE facilitates identifying the frames whose estimation error marginally depends on the number of individuals and more likely stems from other latent variables (e.g. camera position, weather and lighting condition).

2.5 Crowd Volume Estimation

In this section, we outline how we repurpose Human Mesh Recovery (HMR) (Sec. 2.5.1) and adapt Crowd Counting (Sec. 2.5.2) baselines for our task. In Sec. 2.5.3, we describe the intuitions and methodologies of our proposed approach.

2.5.1 From HMR to CVE

We adapt CLIFF [27], BEDLAM-CLIFF [20], and ReFit [26] to the task of CVE according to the following pipeline: (1) identifying human occurrences using a human detector (HD) model, (2) determining the mesh for each individual in the scene, and (3) calculating the volume of each mesh. Finally, the total crowd volume is obtained by aggregating the individual volumes. We dub these baselines as HD+HMR. A shortcoming of this approach is that these methods rely upon an upstream human detector, which can fail when multiple human instances populate an image, as in the case of crowds. To marginalize this issue, we also consider an oracular baseline that replaces the predicted bounding-box locations of humans in the scene with the ground truth ones.

2.5.2 From Crowd Counting to CVE

To assess whether CVE can be naively solved without the adoption of any specific strategy, we set a baseline whose volume estimation stems from $C_{B+}(I) \times \bar{V}_{\mathcal{D}}$, where C_{B+} is a Crowd Counting model, I is the input image, and $\bar{V}_{\mathcal{D}}$ is the average per-person volume in the dataset $\mathcal{D}$. As a statistical reference, we further experiment with an oracular version of this baseline that replaces $C_{B+}(I)$ with the ground-truth count of image I , namely Oracular $C(I) \times \bar{V}_{\mathcal{D}}$.

Additionally, we adapt relevant *Localization*, *Bayesian*, and *Density Map* approaches from Crowd Counting (cf. Sec. 2.3.2).

As for the *Localization* approach, we select P2P-Net [23]. For CVE purposes, we adjust its architecture to predict an array of $(2+1)$ scalars (x, y, v) per detected head, where the additional coordinate v represents the volume of the target person. To encourage accurate v estimates rather than solely emphasizing (x, y) localization,

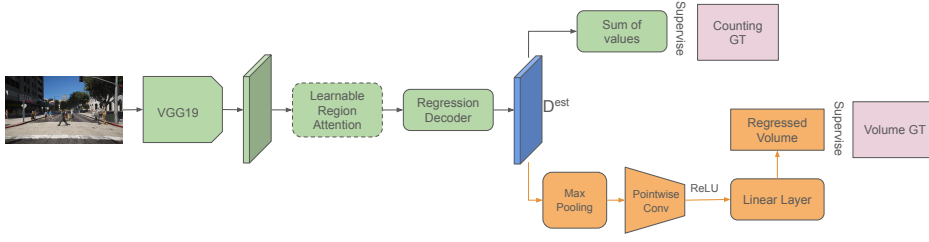


Figure 2.3. Bayesian+ and MAN modified architectures. The Transformer Encoder blocks with *Learnable Region Attention* are used only by MAN. The green layers are kept as they are in the original architectures. Orange layers are part of the additional branch tailored on the CVE task. The estimated density map (D^{est}) is reduced with summation to a single value and counting losses are calculated on it. D^{est} is also used as input for our additional volume-related branch to regress the volume occupied in a frame. Finally, we calculate the L1 loss between regressed and ground truth volumes.

we introduce an additional supervision term: an $\mathcal{L}_1$ loss between predicted and ground-truth volumes, weighted by $\lambda = 10^{-4}$.

As *Bayesian* approaches we consider Bayesian+ [3] and MAN [25]. We adapt them for CVE by appending an additional branch that takes the estimated density map as input and regresses the total volume in the input frame. These two models share the same architecture, with the only exception being MAN employing the Transformer Encoder with the Learnable Region Attention block. The original architecture is shown by the green blocks in Fig. 2.3, while the orange blocks illustrate the novel branch. Since these models are trained on 512×512 image crops, Max Pooling is employed for computing volume on larger-sized images, while Pointwise Convolution compresses tensors to a single dimension. Furthermore, we alter the pre-processing pipeline to compute both counting and volume-related ground truths on which both models are supervised, i.e., the total number of persons in the frame and the total volume occupied by them, respectively.

For the *Density Map* method, we adopt the recent STEERER [24]. Our adaptation preserves the original network architecture while modifying the model’s supervision technique: instead of using conventional counting density maps that label a pixel representing a person’s head with the value 1, we use *Volume Density Maps*, where we annotate the pixel with the person’s total volume. This Volume Density Map is then smoothed using a Gaussian filter.

2.5.3 Per-part Volume Density Maps

In our proposed approach, we leverage ANTHROPOS-V per-part volume annotations, discussed in Sec. 2.7.3. Driven by the insight that volume is distributed throughout the human body, we enhance the proposed Volume Density Maps approach to incorporate this concept. Specifically, since ANTHROPOS-V provides fine-grained annotations of body parts volumes of each character, we introduce *Per-Part Volume Density Maps*, where specific keypoints of each person are assigned a portion of the total body volume. For instance, each of the five torso keypoints will be attributed with $\frac{1}{5}$ of the torso-only volume. After smoothing this local map, the volume is distributed over the interested body parts (see the second column of Table 2.4 for

visualization). We train a STEERER-like model from scratch with these annotations and dub this model as STEERER-V.

2.6 Experiments

In this section, we evaluate all methods’ performance on ANTHROPOS-V quantitatively and qualitatively (Secs. 2.6.1 and 2.6.2). Sec. 2.6.3 provides the results of our best model on real-world datasets.

2.6.1 Experimental Results

Table 2.1 reports results on the test set of ANTHROPOS-V for the CVE task. All the baselines are trained on ANTHROPOS-V. The HD+HMR baselines’ human detectors are YOLOv7 [65] instances fine-tuned on our dataset.

HD-HMR methods do not perform well. The upper part of Table 2.1 shows that HD+HMR methods report suboptimal performance in CVE tasks. These approaches are limited by the heterogeneous scales of individuals within crowd scenes and the limitations of the HD, whose accuracy is significantly marred by severe occlusions and challenging environmental conditions. Note that HD not only fails to generate a bounding box for some individuals, but it can also propose multiple bounding boxes for the same person, resulting in redundant volume estimations for the same individual. Replacing the HD with an oracle that provides GT bounding boxes yields a marked reduction in volume estimation error, still reporting a rather large PPMAE with respect to the Crowd Counting adapted baselines. This is due to the elevated number of occlusions in ANTHROPOS-V, which hampers the exact body shape reconstructions. Our proposed STEERER-V demonstrates superior performance, surpassing all the oracle-enhanced HMR approaches.

Density maps help in CVE. STEERER demonstrates superior performance among the methodologies adapted from Crowd Counting and trained on Volume Density Maps (second block in Table 2.1). This result suggests that Bayesian [3, 25] and localization [23] techniques exhibit suboptimal efficacy in CVE compared to architectures purely based on density prediction. Indeed, STEERER-V, a model built up from STEERER that benefits from the proposed Per-Part Volume Density Maps during training (cf. Sec. 2.5.3), reports the best performance. This superiority is attributed to its capacity for fine-grained predictions, enabling effective management of significant occlusions inherent in crowded scenarios. Quantitatively, STEERER-V reports a minimal average error of 6.73 dm^3 per individual, representing a 53.36% improvement over the most effective Crowd Counting adapted model, STEERER.

Crowd Counting is not enough for CVE. Additionally, Table 2.1 presents the results of the $C_{B+}(I) \times \bar{V}_D$ approach (Sec. 2.5.2). This method underperforms when compared to STEERER-V and its oracular counterpart. This is due to the compounded error arising from substituting individual body volume estimations with the average volume, $\bar{V}_D$, as well as the inherent detection inaccuracies of the counting model, C_{B+} . This comparison underscores the significant performance degradation that would result *in practice* from naively applying a Crowd Counting strategy to CVE, highlighting the necessity for a specialized approach designed specifically for VE.

Table 2.1. Results on ANTHROPOS-V, reported in dm^3 . Inference time (ms) is measured on an NVIDIA A100. Methods are divided into HD+HMR, Crowd Counting, and our proposed approach. $C_{B+}(I)$ refers to the headcount given by [3], while $\bar{V}_{\mathcal{D}}$ is the average human volume. Grayed-out lines use oracular information and should not be directly compared with the other results.

	Model	Oracle	MAE	PPMAE	Inf. time
HMR	CLIFF [27]		673.7	21.41	145.7
	BEDLAM-CLIFF [20]		656.4	21.17	137.9
	ReFit [26]		595.2	18.79	170.0
	CLIFF [27]	✓	370.2	12.89	56.68
	BEDLAM-CLIFF [20]	✓	364.7	12.15	49.73
	ReFit [26]	✓	346.8	11.31	108.1
Counting	Bayesian+ [3]		578.09	17.31	37.50
	P2P [23]		590.91	17.07	61.16
	MAN [25]		557.90	17.03	81.84
	STEERER [24]		506.94	14.43	105.1
	$C_{B+}(I) \times \bar{V}_{\mathcal{D}}$		507.97	14.39	37.50
	Oracular $C(I) \times \bar{V}_{\mathcal{D}}$	✓	191.50	5.32	-
	STEERER-V [24]		205.59	6.73	105.1

When considering Oracular $C(I) \times \bar{V}_{\mathcal{D}}$, which mimics a perfect human detector, something unattainable in practical applications, the error is reduced. This comparison emphasizes the magnitude of the error introduced by the imperfect detection carried out by $C_{B+}(I) \times \bar{V}_{\mathcal{D}}$.

It is worth noting that although STEERER-V does not leverage any privileged information and consequently exhibits imperfect detection, it is comparable with Oracular $C(I) \times \bar{V}_{\mathcal{D}}$. This indicates that STEERER-V compensates for its detection inaccuracies achieving a robust per-person volume estimation, making it the best candidate for practical CVE application. This quality primarily originates from STEERER-V’s training strategy, which integrates body part detection, mitigating false negatives caused by occlusions, with expert knowledge of the volume contribution of each body segment.

Decoupling Crowd Counting from Volume Estimation. The CVE error metrics presented in this paper (MAE/PPMAE) compare per-frame ground truth volumes with model predictions. However, this error stems from two main sources: missed detections and incorrect volume estimations of individuals. To improve CVE models, both of these factors must be addressed. To assess the contribution of each error source to the overall error, we conduct an additional experiment.

We aim to design an evaluation method applicable to all the models proposed in this paper. HD+HMR models are straightforward to adapt by removing the HD component and using ground truth bounding boxes (bbox). However, density-based models require additional steps. For these models, we predict densities for the entire image, then crop the corresponding ground truth bboxes for the individuals involved, isolating their respective volumes. For all models, we consider only non-overlapping bboxes to prevent volume duplication and report the resulting PPMAE.

To avoid penalizing models for detection errors, we exclude any predicted volumes

Table 2.2. Results on ANTHROPOS-V’s S1, S2 and whole test set (FT). All the results are reported in dm^3 .

Model	PPMAE (S1)	PPMAE (S2)	PPMAE (FT)
ReFit [26]	17.94	18.25	18.79
STEERER [24]	12.46	13.68	14.43
STEERER-V	6.67	3.39	6.73

under 10 dm^3 from being counted as positives. Additionally, to facilitate easier detection, we removed scenes from the ANTHROPOS-V test set that contain significant occlusions or challenging lighting conditions, creating a refined test set called S1. Furthermore, we constructed an additional test set, S2, consisting exclusively of bird’s-eye view scenes, which minimize occlusion and further simplify detection.

In Table 2.2, we present the results of this experiment with simplified detection. Notably, ReFit and STEERER show error levels comparable to those on the full test set (FT), highlighting that the primary source of their error lies in evaluating individuals’ volume. In contrast, STEERER-V’s performance improves by 50% on the easier detection set (S2), suggesting that its error is equally divided between detection and volume estimation. Moreover, as shown in Table 1 of the main paper, when comparing both $C(I)_{B+} \times \bar{V}_D$ and ReFit with their oracular counterparts, a similar ratio emerges, further confirming that volume estimation error accounts for half of the total error.

2.6.2 Qualitative Evaluation

Table 2.4 presents the qualitative results of our proposed method, STEERER-V, alongside the Per-Part Volume Density Map, which serves as its training supervision. Furthermore, we provide qualitative results of STEERER and BEDLAM-CLIFF. STEERER-V stands out as the top performer because of its robustness to occlusion and its capacity to generalize. The first row demonstrates that both STEERER and STEERER-V perform well when heads are visible and occlusions are minimal. However, STEERER tends to hallucinate volume along the branches of trees, probably because the model learned that such an object may hide human heads. Contrarily, STEERER-V, designed to distribute volume across the entire body, does not suffer from this side effect, as it does not detect bodies in such scenarios. As occlusions intensify, particularly with multiple people overlapping at a distance (second and third rows), the performance gap between STEERER and STEERER-V becomes more pronounced, with STEERER-V being notably better. In the case of the dark image in the fourth row, STEERER fails to recognize the volume of the person in the foreground because their head merges with the background, while STEERER-V focuses on visible body parts, such as arms or legs, thus reducing the error.

Table 2.3. Evaluation on 3DPW and CrowdHumans (CH). Reported metrics are MAE/PPMAE (3DPW) and the difference between the average real-world per-person volume and the predicted per-person one (CH). B-CLIFF stands for BEDLAM-CLIFF.

	$C_{B+}(I) \times \bar{V}_D$	Refit	B-CLIFF	CLIFF	STEERER	STEERER-V
3DPW	71.3/43	125/75	64.0/40	69.8/43	102/89	40.4/25
CH	-	-30.0	-10.1	+1.00	-7.30	-3.40

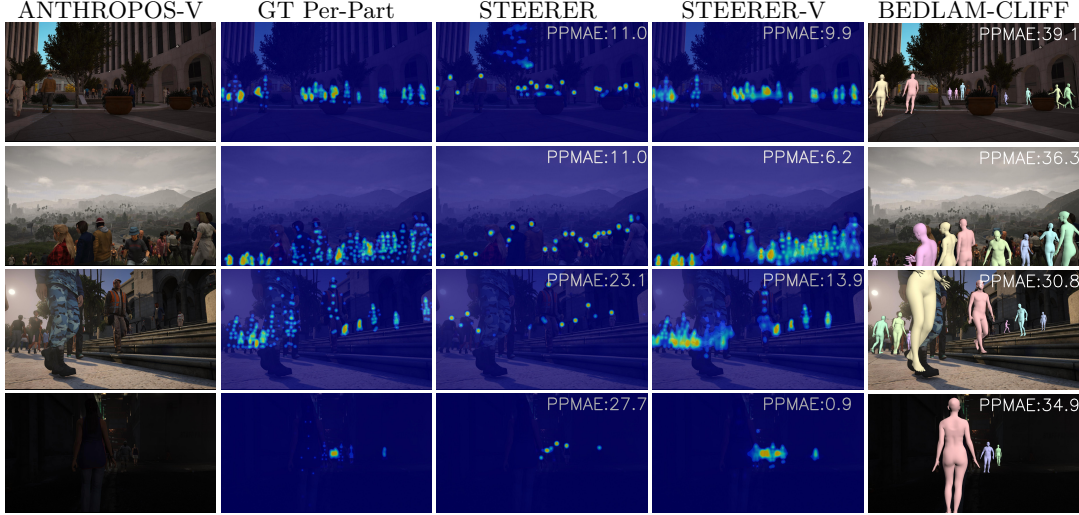


Table 2.4. Visual results of baseline models and STEERER-V on ANTHROPOS-V, along with the Ground Truth Per-Part-Volume Density Maps (GT Per Part). STEERER’s concentrates the volume on the heads, whereas STEERER-V distributes it across the entire body.

2.6.3 From ANTHROPOS-V to real images

We assess the transferability of models trained on ANTHROPOS-V to real imagery for CVE. Given the absence of suitable real-world datasets with volume annotations for crowds, we employ a bifurcated evaluation approach: we use crowd-centric real-world datasets, such as CrowdHumans [1], which lack volume annotations, and mesh-based real-world datasets, such as 3DPW [66], which allow ground-truth volume computation but do not feature crowds.

For 3DPW [66], we compare each model’s predictions against ground-truth mesh volumes. However, several 3DPW images include unannotated persons, such as cameramen or unscripted passers-by. Since no ground-truth is available for these individuals, we manually excluded these images from our test set, reducing the original test set to 6989 images. Results indicate that STEERER-V trained on ANTHROPOS-V outperforms all baseline models (Table 2.3), with MAE and PPMAE registering at 40.40 and 25.28, respectively.

Additionally, we evaluated STEERER-V trained on datasets from [20] and [19] on 3DPW. In this scenario, STEERER-V continues to showcase superior results, with its counterparts presenting increased MAE and PPMAE to (59.72, 37.43) and (44.47, 29.15), respectively.



Figure 2.4. Zero-shot results of STEERER-V on CrowdHuman [1] images when predicting the volume of adults. On each predicted volume map we superimpose the difference between the average real-world per-person volume and the predicted per-person one.

For CrowdHumans [1], we address the lack of volume labels by imputing the average real-world volume [2] for each individual in the images. It is worth noting that such an approximation strongly assumes that people are all similar in size to the average adult and that genders are equally represented in crowds. We compare these estimates with each model’s predictions (Table 2.3). This experiment assesses the alignment of each model’s predictions with expected crowd volumes, with STEERER-V and CLIFF being the most aligned. STEERER-V underestimates the expected volume by 3.40 dm^3 per person, while CLIFF overestimates it by 1.00 dm^3 . As Fig. 2.4 shows, STEERER-V reasonably estimates the volume of adults in indoor environments, large crowds, and with severe occlusions, diverging from the mean to take into account diverse builds, e.g., the men in the foreground of the first row, and uneven balance of genders, e.g., the crowds displayed in the second and third row. Our model fairly performs even with crowds of kids, totally absent from the GTA-V game. The image in the second row of Fig. 2.5 stresses how the model assigns greater volume to the adults in the right foreground than to the surrounding children.



Figure 2.5. Zero-shot results of STEERER-V on CrowdHuman [1] images when predicting the volume of builds that have not been seen at train time, e.g., kids. On each predicted volume map we superimpose the difference between the average real-world per-person volume and the predicted per-person one.

2.7 The ANTHROPOS-V dataset

Here we describe the generation of the proposed ANTHROPOS-V (Sec. 2.7.1). We detail how we align in-game meshes to the real-world statistics (Sec. 2.7.2) and how we obtain SMPL meshes (Sec. 2.7.3). We also comment on ANTHROPOS-V statistics and annotations (Sec. 2.7.4).

2.7.1 Dataset Generation

We construct ANTHROPOS-V exploiting the tools introduced in [21, 22], which, leveraging the game engine from Grand Theft Auto V (GTA-V), allow us to create densely crowded scenes within photorealistic environments. GTA-V provides several 3D urban settings, with different weather and lighting conditions during day and night, and a broad array of characters with diverse appearances, as depicted in Fig. 2.6. In addition, differently from previous GTA-based datasets [21, 22], to achieve a higher degree of photorealism, we use a professionally designed mod [67] that enhances the game graphics and improves the behavior and interaction among characters. Moreover, it offers additional atmospherical conditions and improves the physics in the scenes.

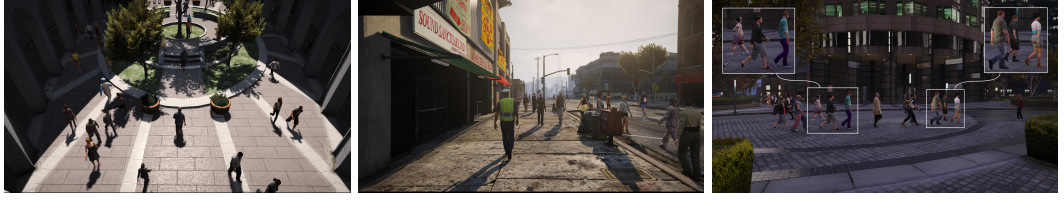


Figure 2.6. Examples from ANTHROPOS-V, showcasing several lighting and weather conditions, camera angles, and a variety of physiques. The crops in the zoomed boxes depict persons with differences in statures and body shapes.

2.7.2 Alignment to real-world body-types

The original GTA-V meshes exhibit a narrow range of variations in anthropometric features, with utterly repetitive heights and volumes and a noticeable imbalance in gender representation. To address this limitation, we carefully revise the in-game meshes and code and generate a distribution of individuals that closely mirrors the real-world one [28] concerning height, volume, and gender.

To achieve this purpose, first, we conduct an in-depth statistical analysis of the distribution of the characters’ anthropometrics in GTA-V. We consider the mass¹, height, and *Body Mass Index* (BMI) of the male and female characters in the game (from now on referred to as “in-game meshes”). We estimate the mass by multiplying the body volume by the average body density (1000 kg/m³ as in [13, 14]).

As theoretically proven by [2], such body features can be represented with random variables M, H, B that follow a log-normal distribution $\Lambda(\mu, \sigma^2)$:

$$M \sim \Lambda(\mu_M, \sigma_M^2), \quad H \sim \Lambda(\mu_H, \sigma_H^2), \quad B = \frac{M}{H^2} \quad (2.3)$$

Fig. 2.7 shows the empirical distributions (solid lines) as opposed to the expected distributions (dashed lines).

The body features of the original in-game meshes do not adhere to the theoretically expected ones, especially for the height that varies in a narrow range around the mean, as evident in the middle plot in Fig. 2.7a. To mitigate such mismatch and increase the variance, we scale the in-game meshes along the three axes with scaling factors α, β, γ that we independently sample from truncated normal distributions; we carefully choose the hyperparameters for this step to avoid unfeasible and unnatural bodies and to end up with meshes that appear realistic. The anthropometrics of the resulting meshes follow a distribution that improves the approximation (Fig. 2.7b). Quantitatively, the Kullback-Leibler divergence between the empirical and the expected distributions, averaged across genders, decreases by 27.9%, 63.3%, and 19.8% for mass, height, and BMI, respectively. The SMPL fitting process (Sec. 2.7.3) disregards the clothing, thereby producing meshes that more closely match the real-world distributions of height [28] and BMI [68] (Fig. 2.7c). As a final remark, the BMI of the SMPL meshes in ANTHROPOS-V ranges in [10, 50] kg/m², representing also underweight and obese individuals.

¹medical literature refers to the body mass as “weight”, which in physics refers to another quantity; we stick with the physics definition.

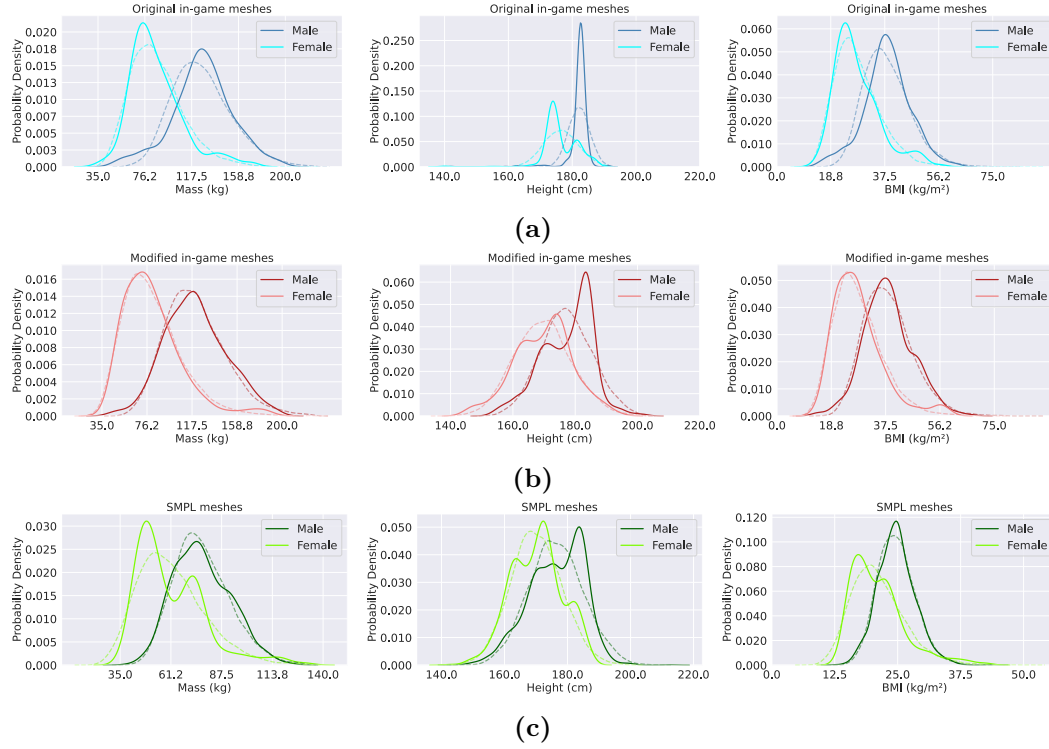


Figure 2.7. Statistical analysis of the distributions of mass, height, and *Body Mass Index* (BMI) of the individuals in ANTHROPOS-V. Solid curves depict the empirical distributions, while dashed curves refer to the theoretically expected ones [2]. (2.7a) Distribution of the body features of the characters’ meshes in GTA-V without any manipulation. (2.7b) Distribution of the body features of the characters’ meshes in GTA-V after applying some geometrical transformations. (2.7c) Distribution of the body features of the resulting fitted SMPL meshes.

2.7.3 SMPL Fitting

To label each character with accurate ground truth volume, we employ a technique akin to the one described in [19]. The fitting procedure ensures that the SMPL mesh tightly conforms to the character mesh’s uncovered body parts while allowing a looser fit on clothed parts. To perform this fitting process, we need both a 3D pose prior and knowledge of which vertices in the GTA-V mesh represent skin or clothing. Thus, we initiate the process by generating 10 visual renders of the GTA-V 3D characters. This is achieved by moving the camera around the textured 3D mesh of the characters. Then, the pose estimation process exploits [69] to predict the character’s 2D pose in each rendered image, pose that is then lifted into a 3D space to give a complete spatial representation of the character’s posture. Next, the process of dividing a character’s mesh into skin and clothes vertices leverages [70] to segment each of the 10 renderings. The resulting segmentation is reprojected onto the mesh to label each vertex. The SMPL fitting involves aligning a standard human body model to the character’s 3D pose and shape. This step requires meticulous adjustments to ensure that the SMPL mesh accurately follows the contours and posture of the character. Indeed, following [19], we employ two different loss functions to constrain

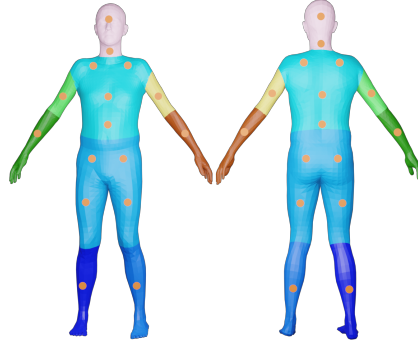


Figure 2.8. A SMPL mesh from ANTHROPOS-V. Different body parts are highlighted in different colors. Orange dots represent the keypoints associated with each body part.

the SMPL within the GTA-V mesh. The first loss is applied to the retrieved skin vertices, where we impose a severe fitting. The second applies to the clothes vertices, where we aim to have a looser fit so that these vertices would not penetrate the original mesh while remaining sufficiently close to it. We report that our SMPL meshes have an average per skin vertex error of 7.32 mm and a penetration error of 10 mm for clothed vertices, where a looser fit is desired. This measure indicates how much these vertices extend beyond the GTA-V mesh. Finally, we use the obtained meshes to compute ground-truth volume labels for each character using Blender [\[71\]](#) Python API, which calculates the volume within a mesh. Notably, besides offering labels for the total body volume, ANTHROPOS-V includes annotations for the volume of individual body parts obtained by slicing the SMPL meshes. We divide the estimated meshes into nine sections: head, torso, thighs, left and right arms, forearms, and calves, as shown in Fig. [2.8](#). We then calculate the volume of each of these parts separately.

2.7.4 Dataset Statistics

ANTHROPOS-V features 768 FHD videos with annotated volumes, SMPL shape parameters, keypoints, and camera parameters and position. Videos are recorded at 30 fps and display crowds moving in diverse urban scenarios. ANTHROPOS-V features 701 distinct characters, each with a variable number of outfits, resulting in over 3k unique appearances, interacting in 384 diverse scenarios with different camera angles and weather conditions. To propose a fair split, we divide characters into three disjoint sets of 495, 64, and 142 that we distribute in different train, validation, and test videos, respectively. Within crowded scenes, characters engage with each other and with the environment, undertaking interrelated actions. For instance, they avoid collisions and form queues to navigate stairs or enter confined areas.

2.8 Limitations and Future Works

As the first endeavor to establish a benchmark for Crowd Volume Estimation (CVE), our work lays the groundwork for this emerging field. However, we acknowledge

some aspects of our work that present opportunities for future refinement.

We introduced ANTHROPOS-V aiming to bridge the gap between synthetic and real-world data. While testing the transferability of the learned knowledge on real images without fine-grained and precise volume annotations may suffice to make an initial point on the validity of the dataset, future work should embark on acquiring detailed volume estimates of real images. Moreover, it may pursue even larger crowds, increasingly complex and diverse interactions, and estimates of objects other than people (e.g. backpacks, bags, etc.).

The current output of our model provides a single per-frame number representing the total crowd volume. While suitable for many applications, this approach encourages exploration into more granular spatial analyses that could further benefit fields such as civil engineering, where detailed volume distribution information might be valuable.

Future research should also investigate the integration of monocular depth estimation. These methods could serve as a complementary modality to refine volume estimation, reducing issues such as scale ambiguity and occlusion induced errors, potentially yielding further performance improvements.

Finally, we acknowledge that the ethical implications of CVE from images present complex challenges. Primary among these is the privacy issue in public spaces, which intersects with concerns about data security and the potential for misuse, as the underlying data could be adapted for unintended surveillance purposes.

Furthermore, bias in volume estimates due to potential underrepresentation in training data could lead to discriminatory applications. As CVE technology evolves, these ethical considerations underscore the critical need for robust guidelines and transparent deployment protocols to ensure that the benefits of CVE can be realized while safeguarding individual rights.

2.9 Conclusion

In this study, we have established the first benchmark for Crowd Volume Estimation. We introduced relevant metrics and developed a dataset specifically designed for this task, focusing on human crowds in real-world-like environments. Additionally, we evaluated baseline and oracular models adapted from Crowd Counting and Human Mesh Recovery domains. Furthermore, we proposed a novel supervision approach called Per-Part Volume Density Maps, which we utilized to train STEERER-V, achieving superior results. Given the challenges in gathering real-world datasets for CVE, we anticipate that introducing this new task and benchmark will ignite interest in the research community and inspire future endeavors in the field.

Chapter 3

MoDiPO: text-to-motion alignment via AI-feedback-driven Direct Preference Optimization

3.1 Overview

Diffusion Models have revolutionized the field of human motion generation by offering exceptional generation quality and fine-grained controllability through natural language conditioning. Their inherent stochasticity, that is the ability to generate various outputs from a single input, is key to their success. However, this diversity should not be unrestricted, as it may lead to unlikely generations. Instead, it should be confined within the boundaries of text-aligned and realistic generations. To address this issue, we propose **MoDiPO** (Motion Diffusion DPO), a novel methodology that leverages Direct Preference Optimization (DPO) to align text-to-motion models. We streamline the laborious and expensive process of gathering human preferences needed in DPO by leveraging AI feedback instead. This enables us to experiment with novel DPO strategies, using both online and offline generated motion-preference pairs. To foster future research we contribute with a motion-preference dataset which we dub Pick-a-Move. We demonstrate, both qualitatively and quantitatively, that our proposed method yields significantly more realistic motions. In particular, **MoDiPO** not only achieves a substantial improvement in the Fréchet Inception Distance (FID) but also secures a threefold increase in preference from human evaluators compared to the original models' outputs.

3.2 Introduction

Synthesizing realistic human behavior is a fundamental challenge in computer animation, holding significant relevance in gaming, film, and virtual reality contexts. Recent advancements have shifted towards text-driven generation methods, aiming to bypass time-consuming traditional modeling processes requiring specialized

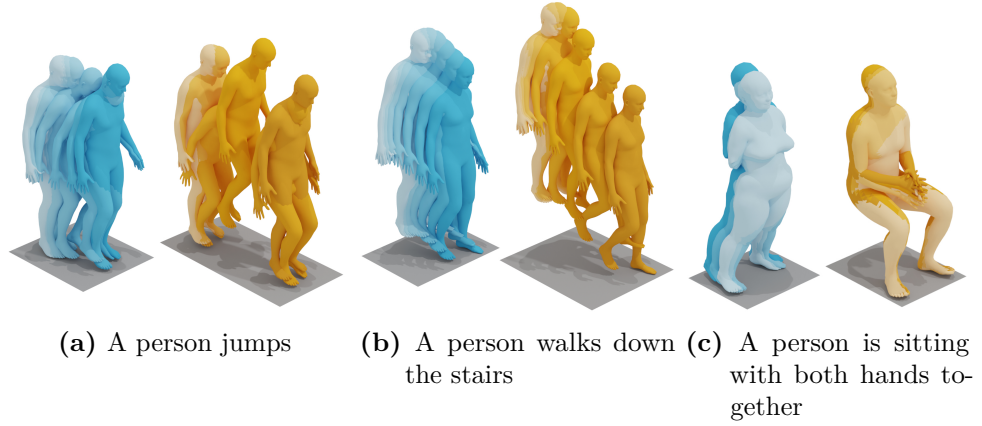


Figure 3.1. MoDiPO (in yellow) leverages the DPO alignment to generate enhanced human motions.

equipment and artists’ domain knowledge. Instead, leveraging natural language descriptions provides sufficient semantic details to describe generic human motion. One key advantage of text-driven generative models is their probabilistic nature, allowing for multiple diverse generations from the same prompt, each characterized by unique nuances. Yet, this also introduces the risk of producing motions that are either unrealistic or too different from the intended ones. To our knowledge, no research has addressed this aspect so far.

Recently, [72, 73] introduced alignment methodologies allowing users to instill desired behaviors in pre-trained generative models by leveraging curated sets of human preferences. This approach has succeeded in various tasks, including question answering [74], dialogue support, and image generation [75]. Despite substantial efforts in developing text-driven motion synthesis models, no existing research has attempted to align these models with human preferences, mostly due to the cost and time required for collecting human-annotated data. In contrast, leveraging synthetic or AI-generated feedback overcomes these limitations and enables scaling up preference datasets.

To exploit the creativity of the generative pipeline, we would like to bias our model toward high-quality motions, avoiding altogether regions of the learned latent manifold that lead to unfeasible motions. We base our work on the adaptation of Direct Preference Optimization [76] for DDPMs [77] to directly optimize the models on the preferential data. For this purpose, we introduce “Pick-a-Move”, a dataset designed to mimic human preferences for motion. To achieve this, we first generate motions using state-of-the-art text-to-motion models, then pair these outputs with preference scores obtained through text-to-motion retrieval models [78, 79]. These retrieval models leverage a shared latent space where paired text and motion are placed closer together, allowing them to express preferences for specific text-motion combinations. We use Pick-a-Move to align text-driven diffusion-based generative models for motion synthesis [80, 81]. Specifically, MLD [81] is selected as it strikes a compromise between efficiency and realism, the latter witnessed by its generations’ FID. Given that many motions can equally represent the input text, training models on fixed motion-preference pairs may limit their diversity. Pick-a-Move provides

more flexibility in training by offering up to N motion options for each descriptive text. These motions are ranked from most to least preferred, which allows training on pairs of text and motion where the selection of *winners* and *losers* can be adjusted during the training process.

Additionally, we exploit the availability of ground truth motions by selecting them as winners with a low probability: this approach helps prevent overfitting to the training set and reduces the artifacts or text-misalignment issues that can arise when training exclusively on generated sequences. As a result, this supervised fine-tuning element within our pipeline further enhances the quality of the generated outputs (Fig. 3.1).

Pioneering the combination of text-driven motion generation and alignment, we perform extensive evaluations on two datasets with two different models. We achieve improvements in FID of up to 39% while retaining consistent RPrecision and Multi-Modality performances. To determine whether these quantitative improvements translate into visibly better motion quality, we conducted a user study. Our results show that motions generated by MoDiPO are preferred three times more often than those from the not-aligned models.

To summarize, our contributions are the following.

- We propose MoDiPO, a framework for aligning text-to-motion generative models.
- We introduce Pick-a-Move, a motion-preference dataset annotated leveraging synthetic feedback, and conducted a user study to assess alignment between AI-driven and human preferences.
- We use a dynamic winner-loser sampling strategy that draws pairs from different candidates and ground truth motions. This training signal prevents mode collapse and reduces overfitting to the alignment dataset.
- We experiment with both offline and online alignment strategies, that is, leveraging pre-collected motion sequences as well as ones generated on the fly.

Related Works In this section we discuss the main approaches for alignment, distinguishing between methods driven by human input (Sec. 3.2.1) and those relying on synthetic data (Sec. 3.2.2). We then continue with an overview of Text-to-Motion models (Sec. 3.2.3).

3.2.1 Aligning Models with Human Feedback

Early applications of model alignment employed a complex reinforcement learning pipeline, requiring preferential data built through human feedback (RLHF) [72]. This technique has been applied to text generation tasks such as summarization [82]. Advancements in the reinforcement learning pipeline by InstructGPT [83], such as the use of Proximal Policy Optimization [84], have enabled the finetuning of models for tasks requiring evidence-supported question answering [74] and dialogue support [85].

Several solutions have been proposed to simplify this complex pipeline: Implicit Language Q-Learning (ILQL) [86] utilizes Q-learning to learn value functions that guide language model generation to maximize user-specified utility functions. Similarly, HIVE [87] leverages offline reinforcement learning for instruction editing alignment. Direct Preference Optimization (DPO) [76] offers a reinforcement learning-free approach by directly optimizing the model using a supervised clas-

sification objective on the preferential data, eliminating the need for an explicit reward model. This approach has inspired several RL-free methods for text-to-image alignment [88] including Diffusion-DPO [77], which specializes DPO’s objective for DDPMs [89], a popular choice for text-driven human motion generative models. While we draw inspiration from these approaches, our method differs by specifically tailoring the alignment process to motion data while also using AI Feedback, addressing challenges unique to this domain, and achieving improved motion quality.

3.2.2 Aligning Models with Synthetic Feedback

Recent works have focused on the possibility of replacing human annotators with Synthetic Feedback (also known as AI Feedback) to avoid the high costs and effort required to collect high-quality datasets. One approach involves self-supervision (self-improvement), as demonstrated in [73]. In this method, human “supervision” only defines a list of prompts, that the generative model should respect. The model then self-evaluates the generated sentences by computing a reward based on these human-specified guidelines. This synthetically labeled dataset is then used to train a preference model, which a policy aims to maximize using an RL algorithm. The approach in [90] utilizes synthetic ranking by employing different models with varying levels of complexity to generate sentences based on the same input prompt, where the outputs from models with higher parameter counts, which tend to generate better content, are assigned higher rewards. Similar techniques have been applied to text-to-image generation, as seen in [91]. Additionally, [92] demonstrated through extensive experiments that RLAIFF (Reinforcement Learning from AI Feedback) can be comparable to, or even outperform, RLHF. For these reasons, we utilize synthetic feedback in our pipeline by leveraging a retrieval model that scores generated motions based on their alignment with an input text.

3.2.3 Text-to-Motion

Early approaches like TEMOS [93] directly used VAEs to generate motion based on textual conditioning. Currently, two leading generation strategies dominate: Autoregressive methods [94], such as MotionGPT [95] and T2M-GPT [96], and Denoising Diffusion Probabilistic Models (DDPMs) [89]. The latter category includes models like MDM [80] and MLD [81]. MDM applies DDPM directly on raw motion data to capture the motion-text relationship for a specific number of frames. MLD, instead, improves performances by utilizing a two-step approach: First, they learn a compact motion latent space using VAEs; then, they apply a diffusion process on the learned latent space leveraging Latent DDPMs [97]. Furthermore, Retrieval-Augmented Generation [98, 99] demonstrates performance improvements when conditioning motion generation on a sample retrieved from the entire training set. However, this approach requires storing and searching on a large database, the quality of which significantly impacts the generation. Conversely, aligning a model does not introduce additional computational complexity.

3.3 Background

Given our aim to enhance motion diffusion models, in Sec. 3.3.1 we provide an overview of VAE latent representation, which is a mandatory requirement for latent diffusion models. In Sec. 3.3.2 and Sec. 3.3.3 we briefly describe the DDPMs and latent-DDPMs frameworks. In Sec. 3.3.4 we present an overview on preference alignment.

3.3.1 Variational Auto-Encoder for Human Representation

VAE consists of an encoder-decoder generative architecture trained to minimize the reconstruction error. It is typically used to reduce the dimensionality of the pose vector V , projecting it into a manifold of feasible poses. In this framework, the encoder network, parameterized by ϕ , generates lower-dimensional embeddings $\mathbf{z} \in \mathbb{R}^{n \times D}$ as outputs based on input poses $\mathbf{p} = \{\mathbf{p}_k\}_{k=1}^F$. Following VAE literature [100], $q_\phi(\mathbf{z}|\mathbf{p})$ approximates the true posterior distribution of the latent space with a multivariate Gaussian with a diagonal covariance structure:

$$q_\phi(\mathbf{z}|\mathbf{p}) = \mathcal{N}(\mathbf{z}|\mu_\phi(\mathbf{p}), \sigma_\phi^2(\mathbf{p})\mathbf{I}), \quad (3.1)$$

where $\mu_\phi(\cdot)$ is the mean, $\sigma_\phi(\cdot)$ the standard deviation given as output by the encoder. We sample from the approximated posterior $\mathbf{z}_i \sim q_\phi(\mathbf{z}|\mathbf{p}_i)$ using:

$$\mathbf{z}_i = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i^2 \odot \boldsymbol{\rho}, \quad (3.2)$$

where $\mathbf{z}_i$ is a one-dimensional vector of size D , and $\boldsymbol{\rho}$ is sampled from a standard multivariate Gaussian distribution. The transformer decoder [81, 93] parametrized by θ maps the sampled values back to the body pose representation $p_\theta(\mathbf{p}|\mathbf{z})$, which may be further processed into body meshes with the differentiable SMPL model for visualization purposes. The network parameters are obtained by optimizing the Evidence Lower Bound (ELBO) objective, as described in [100].

3.3.2 DDPMs

Denoising Diffusion Probabilistic Models (DDPMs) [89] work in two stages. First, through a **forward process**, they gradually add noise to the input data according to a variance schedule $\beta_t \in (0, 1)$ for $t \in \{1, \dots, T\}$, transforming any data distribution $q(x_0)$ into a simple prior (e.g., Gaussian). This forward process can be represented as

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t\mathbb{I}) \quad (3.3)$$

Then, a separate **reverse process** also dubbed as *denoising* process, removes this noise, effectively cleaning the data, and is formulated as

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \beta_t\mathbb{I}) \quad (3.4)$$

with $\mu_\theta(x_t, t)$ being the posterior mean predicted by the trained neural network, which behaves as a denoiser.

The loss term to optimize is thus

$$Loss := \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \mathbf{I}), t} [\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2]. \quad (3.5)$$

3.3.3 Latent DDPMs

In Latent DDPMs [97, 81] the diffusion process takes place on a condensed, low-dimensional latent representation of motion, denoted as z . This representation is derived by training a VAE to encode and reconstruct input motions. Contrary to the simple DDPMs that work on raw features, for latent DDPMs the approximated posterior $g(\mathbf{z}_t|\mathbf{z}_{t-1})$ gradually converts latent representations $\mathbf{z}_0 = \mathbf{z}$ into random noise $\mathbf{z}_T$ in T timesteps. Then, the reverse process gradually refines the noised vector to a suitable latent representation $\mathbf{z}_0$.

Following [81, 97, 89, 101], we use the notation $\{\mathbf{z}_t\}_{t=0}^T$ to denote the sequence of noised latent vectors, with $\mathbf{z}_{t-1} = \epsilon_\psi(\mathbf{z}_t, t)$ representing the denoising operation at time step t . Here, ϵ_ψ refers to a denoising autoencoder trained to predict the denoised version of its input. The forward and reverse processes are identical to those described in equations 3.3 and 3.4, but they are directly applied to latent vectors.

At inference, $\epsilon_\psi(\mathbf{z}_t, t)$ predicts $\hat{\mathbf{z}}_0$ through a series of T iterative denoising steps and the frozen decoder translates $\hat{\mathbf{z}}_0$ into poses in a single forward step.

3.3.4 Preference Alignment Overview

Preference-based alignment has recently emerged in natural language processing to steer Large Language Models' outputs toward human-preferred responses. One of the first techniques proposed to achieve this is Reinforcement Learning from Human Feedback (RLHF), where a reward model r is trained to approximate the Bradley-Terry model, which defines the human preference distribution as:

$$p(x_0^w \succ x_0^l | c) = \frac{\exp(r(c, x_0^w))}{\exp(r(c, x_0^w)) + \exp(r(c, x_0^l))} \quad (3.6)$$

Given a preference dataset, a reward model $r_\phi(c, x_0)$ can be trained via maximum likelihood. Interpreting it as a binary classification problem leads to the negative log-likelihood loss:

$$L_R(\phi) = -\mathbb{E}_{c, x_0^w, x_0^l} [\log \sigma(r_\phi(c, x_0^w) - r_\phi(c, x_0^l))] \quad (3.7)$$

The RLHF framework improves a conditional distribution $p_\theta(x_0|c)$ via a policy optimizer aiming to maximize rewards. To ensure training stability, retain the pretrained model's knowledge, and avoid mode collapse effects, a KL-Divergence regularizer is introduced, controlled by β , with respect to a reference distribution:

$$\max_{p_\theta} \mathbb{E}_{c \sim D_c, x_0 \sim p_\theta(x_0|c)} [r(c, x_0)] - \beta \mathbb{D}_{\text{KL}}[p_\theta(x_0|c) || p_{\text{ref}}(x_0|c)] \quad (3.8)$$

Direct Preference Optimization (DPO) requires neither an explicit reward model nor a dedicated policy optimizer. Indeed, the DPO objective becomes a simple weighted maximum-likelihood loss. The optimal solution p^* from [76] takes the form of:

$$p_\theta^*(x_0|c) = \frac{1}{Z(c)} p_{\text{ref}}(x_0|c) \exp\left(\frac{1}{\beta} r(c, x_0)\right) \\ \text{where } Z(c) = \sum_{x_0} p_{\text{ref}}(x_0|c) \exp\left(\frac{1}{\beta} r(c, x_0)\right) \quad (3.9)$$

Rewriting it in terms of reward function this translates to:

$$r(c, x_0) = \beta \log \frac{p_\theta^*(x_0|c)}{p_{ref}(x_0, |c)} + \beta \log Z(c) \quad (3.10)$$

Given the Bradley-Terry formulation in Eq. 3.7, the DPO objective can then be written as:

$$L_{DPO} = -\mathbb{E}_{(c, x_0^w, x_0^l) \sim \mathcal{D}} \log \sigma \left[\left(\beta \log \frac{p_\theta(x_0^w|c)}{p_{ref}(x_0^w|c)} - \beta \log \frac{p_\theta(x_0^l|c)}{p_{ref}(x_0^l|c)} \right) \right] \quad (3.11)$$

This formulation highlights how DPO streamlines preference alignment by relying on a simple likelihood-based objective without needing a separate reward model and RL optimizers. In DiffusionDPO [77], DPO is adapted to diffusion models. The central difficulty is that diffusion models create samples via a lengthy chain of denoising steps, so $p_\theta(x_0|c)$ cannot be computed directly without evaluating the likelihood across every possible intermediate trajectory for both the preferred x^w and the unpreferred x^l samples. For this reason, a reward computation over denoising paths is introduced:

$$r(c, x_0) = \mathbb{E}_{p_\theta(x_{1:T}|x_0, c)} [R(c, x_{0:T})] \quad (3.12)$$

with $R(c, x_{0:T})$ being the reward assigned to the entire denoising chain $x_{0:T}$. By using Eq. 3.7, DiffusionDPO’s objective can then be written as follows

$$L_{DDPO}(\theta) = -\mathbb{E}_{(c, x_0^w, x_0^l) \sim \mathcal{D}} \log \sigma \left(\beta \cdot \mathbb{E}_{\substack{x_{1:T}^w \sim p_\theta(x_{1:T}^w|x_0^w, c) \\ x_{1:T}^l \sim p_\theta(x_{1:T}^l|x_0^l, c)}} \left[\log \frac{p_\theta(x_{0:T}^w|c)}{p_{ref}(x_{0:T}^w|c)} - \log \frac{p_\theta(x_{0:T}^l|c)}{p_{ref}(x_{0:T}^l|c)} \right] \right) \quad (3.13)$$

Intuitively, this loss increases the likelihood for the preferred denoising paths $x_{0:T}^w$ while decreasing it for non-preferred ones $x_{0:T}^l$.

3.4 Methodology

In this section we describe our approach to DPO-based text-to-motion alignment: in Sec. 3.4.1 we introduce a synthetic dataset comprising text-conditioned motion generations, where all motions are paired with AI preferences, dataset that is then used by the alignment framework described in Sec. 3.4.2. Notice that, in our work, we follow [78] and represent a motion sequence as $\mathbf{p} = \{\mathbf{p}_k\}_{k=1}^F \in \mathbb{R}^{F \times V}$ with F being the number of frames and V the pose vector’s features.

3.4.1 Motion Preferences Dataset

Unlike text and images, motion data often suffers from scarcity. At the same time, a dataset with motion preferences does not exist. To address this lack, we propose a synthetic dataset named Pick-a-Move, PaM for short. The dataset is partitioned into two parts, one sourced from HumanML3D and the other from KIT-ML. Each

subset employs its original text prompts and native motion-data format to steer motion generation. For every motion in HumanML3D and KIT-ML, we sample one associated caption c , generate $K = 8$ motion sequences with both MLD and MDM conditioned on c , and link the resulting sequences with the unsampled captions, since they describe the same underlying motion. As a result, the HumanML3D portion of the PaM dataset contains 365'672 motion sequences, each of which is associated with an average of 2.99 descriptions. In contrast, the KIT-ML subset comprises 74'192 motion sequences, each paired to an average of 2.12 motion descriptions.

Next, given the set of $N = 2K$ motion sequences $m = \{m_k\}_{k=1}^N$, we compute scores using either TMR [79] or the model proposed in [78], which we dubbed as “Guo”. Using either of these models, we encode the motion description and each of the k generated motions into embeddings e_t and e_m^k ; then we define the set of preference scores as $s = \{s_k\}_{k=1}^N$ where $s_k = \frac{e_t \cdot e_m^k}{\|e_t\| \|e_m^k\|}$, corresponding to the cosine similarity between e_t and e_m^k . We use s to order m into $m^* = \{m_k^*\}_{k=1}^N$, representing a list where motions are presented from the most coherent with c , namely *winner*, to the least, *loser*. By employing this synthetic method, we can automate the process of ranking motion sequences, eliminating the need for human judgments. To validate the reliability of our AI-based preferences, we conducted a user study on the HumanML3D subset of PaM: we randomly sampled 100 pairs of motions generated from the same prompts. We asked 10 evaluators with Computer Science backgrounds to choose their preferred motion in each pair. To quantify the agreement between AI-based and Human preferences, we collected TMR and Guo models’ preferences and calculated Cohen’s Kappa, using majority voting for human choices. The resulting Cohen’s Kappa score, 0.65 for TMR and 0.61 for Guo, demonstrates a substantial agreement between the Human and AI-driven preferences.

Thanks to this dataset, we can then take a pair (x_0^w, x_0^l) representing a winner and a loser, such that by using their associated scores they comply with $(x_0^w \succ x_0^l | c)$ where c is the conditional signal corresponding to the motion description. Thus this process identifies which motion within a pair better reflects the textual description. With this dataset, it is possible to achieve:

- **Increased Pairwise Comparisons:** By generating N motions, we achieve up to $\binom{N}{2}$ potential winner-loser pairs per prompt. A high number of pairs is crucial for alignment methods.
- **Preserved Multi-Modality:** The high number of pairs associated with each prompt helps maintain Multi-Modality. This prevents the alignment process from homogenizing the generated motions, reducing issues such as mode collapse.

3.4.2 Motion Alignment Pipeline

We apply L_{DDPO} to both DDPM and latent DDPM for motion synthesis using PaM as preference dataset. We rely on two identical diffusion models, a reference model p_{ref} and a trainable target model p_θ , with the only distinction being that the reference’s parameters remain fixed throughout the optimization. For latent diffusion architectures such as MLD, we limit updates to the denoiser’s weights that

operate in latent space, thus preventing unwanted behavior shifts and avoiding the cost of tuning the entire network. In contrast, for MDM, where denoising occurs directly on raw motion sequences, we find it most effective to fine-tune only the network’s final layers.

Intuitively, during training, the model generates examples, latent embeddings in the case of latent-diffusion models or raw motions for non-latent architectures and evaluates them against preferred and non-preferred motions drawn from PaM. The denoiser is thus optimized to produce motions that more closely match the “winning” sequences identified by the ranking model while steering away from the “losing” ones. This process ensures that generated outputs align with the AI-derived preferences and faithfully match the characteristics specified by the textual prompts.

Building on DiffusionDPO [77], we replace its static pairing strategy with a dynamic winner-loser selection over N candidate motions. At each training step, we sample N motions from PaM, each generated starting from the same prompt, and apply a pair selection criteria (see 3.6.1) to choose two motions as winner and loser. Moreover, with probability GT_p , we inject a ground-truth motion (from the HumanML3D or KIT-ML training sets) as the winner:

$$x_0^w \sim \begin{cases} \mathcal{D}_{GT}, & \text{with probability } GT_p, \\ \mathcal{D}_{PaM}, & \text{with probability } 1 - GT_p. \end{cases}$$

This results in a more stable training and yields measurable improvements in FID and overall motion quality, as shown in Table 3.5. Beyond the usage of PaM, we also explore an online DPO variant: for each prompt, we dynamically generate K sequences, use the aforementioned sampling criteria to determine a winner-loser pair, and then update the target model p_θ . This streaming alignment method leverages generated sequences and continuous synthetic feedback, enabling us to scale the alignment pipeline without relying on the existence of an offline preference dataset. Fig. 3.2 provides a schematic overview of our complete motion alignment pipeline.

3.5 Experimental Evaluation

This section discusses the used evaluation metrics in Sec. 3.5.1 and provides a quantitative (Sec. 3.5.2) and qualitative (Sec. 3.5.3) evaluation of our framework. In addition, in Sec. 3.5.4 we report the results of a user study to evaluate the quality of the generated motions. We evaluate on the HumanML3D and KIT-ML test sets, enabling direct comparisons between the baseline models, MLD and MDM, and their MoDiPO-aligned versions.

3.5.1 Evaluation Metrics

We evaluate the generated motion sequences employing all metrics adopted in text-to-motion literature.

The RPrecision evaluates how effectively a model generates motion sequences that are relevant to a given textual description. It is computed by combining the generated motion with 31 unrelated motions and evaluating the top-3 accuracy in matching the text to the corresponding motion.

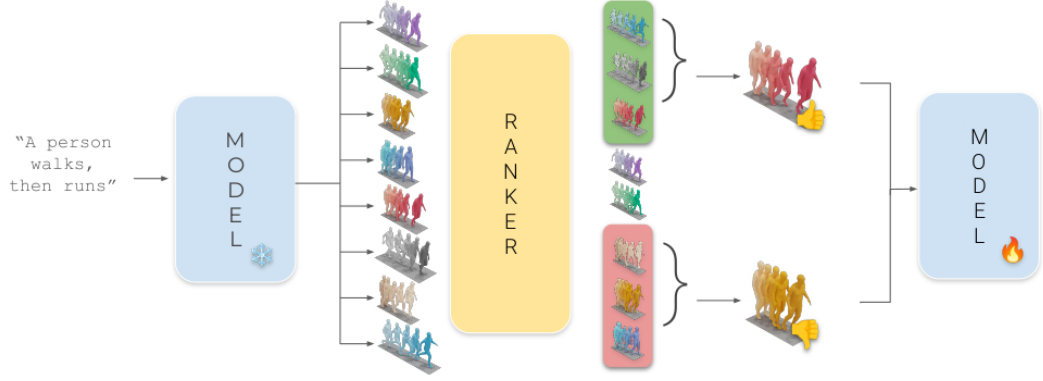


Figure 3.2. MoDiPO Schematics. Starting with an input prompt, we generate K motions with a frozen reference model. These motions are then ranked according to their relevance with the input description. This motions-text-ranking tuple constitutes a sample in our “Pick-a-Move” (PaM) preferential dataset. At train time, we divide motions into two sets of best and worst generations, according to these rankings. By sampling from these sets, we form a winner-loser pair that is used to refine the target model using DPO.

In contrast, Multimodal Distance (MM-Dist) quantifies the degree of alignment between the generated motions and their associated textual descriptions.

The Fréchet Inception Distance (FID) [102] assesses the similarity between the generated motions and ground truth distributions. FID is a well-established metric [103, 101], especially in image synthesis, which is more consistent with the human judgment of image quality compared to the Inception Score (IS) [104] and is used as a main performance indicator by capturing both diversity and fidelity to the ground truth dataset.

Diversity evaluates the variety of motions generated for a specific text input. To compute it, the data is randomly split into two equal-sized subsets, denoted as $\{x_1, \dots, x_{X_d}\}$ and $\{x'_1, \dots, x'_{X_d}\}$, each containing motion feature vectors, where X_d represents the number of motions in each subset. Diversity is then defined as:

$$\text{Diversity} = \frac{1}{X_d} \sum_{i=1}^{X_d} \|x_i - x'_i\| \quad (3.14)$$

Multi-Modality quantifies the variation of generated motions across the entire dataset for different text descriptions. This is calculated by first selecting a set of T_d text descriptions from the dataset. For each description, two subsets of generated motions, each with X_d elements, are randomly sampled. These subsets are represented as $x_{t,1}, \dots, x_{t,X_d}$ and $x'_{t,1}, \dots, x'_{t,X_d}$.

Multi-Modality is then defined as:

$$\text{MM} = \frac{1}{T_d \cdot X_d} \sum_{t=1}^{T_d} \sum_{i=1}^{X_d} \|x_{t,i} - x'_{t,i}\| \quad (3.15)$$

3.5.2 Quantitative Evaluation

We aligned MLD and MDM using PaM, with results for HumanML3D and KIT-ML presented in Table 3.1 and Table 3.2, respectively. To align MLD with MoDiPO for the KIT-ML dataset, a re-trained MLD was used because of the unavailability of the original model. For HumanML3D, using both TMR and Guo as rankers, we achieve improvements in FID and Multi-Modality for MLD, while for MDM we improve FID whereas other metrics remain consistent, a trend that also holds for KIT-ML. We consistently achieve improvements in FID while preserving Multi-Modality and RPrecision performance. This is particularly significant as alignment processes can sometimes lead to mode collapse, a problem that, as our results demonstrate, does not occur.

Table 3.1. Results for HumanML3D. Each group of results corresponds to the base models, aligned models, and current SOTA backed by a RAG framework. *Italic* is for best results overall, while **bold** represents best results among base models and MoDiPO-aligned ones.

Model	Top3 $\uparrow$	MMDist $\downarrow$	Diversity $\rightarrow$	FID $\downarrow$	MM $\uparrow$
Real	.797 $\pm$.002	2.974 $\pm$.008	9.503 $\pm$.065	.002 $\pm$.0	-
MotionGPT [95]	.778 $\pm$.002	3.096 $\pm$.008	9.528 $\pm$.071	.232 $\pm$.008	2.008 $\pm$.084
T2M-GPT [96]	.775 $\pm$.002	3.118 $\pm$.011	9.761 $\pm$.081	.116 $\pm$.004	1.856 $\pm$.011
MotionDiffuse [105]	.782 $\pm$.001	3.113 $\pm$.001	9.410 $\pm$.049	.630 $\pm$.001	1.553 $\pm$.042
MLD [81]	.755 $\pm$.003	3.292 $\pm$.01	9.793 $\pm$.072	.459 $\pm$.011	2.647 $\pm$.11
Ours MLD-TMR	.758$\pm$.002	3.267$\pm$.01	9.747 $\pm$.077	.303 $\pm$.007	2.663 $\pm$.111
Ours MLD-GUO	.753 $\pm$.003	3.294 $\pm$.010	9.702$\pm$.075	.281$\pm$.007	2.736$\pm$.111
MDM [80]	.703 $\pm$.005	3.658 $\pm$.025	9.546 $\pm$.066	.501 $\pm$.037	2.874$\pm$.070
Ours MDM-TMR	.706$\pm$.004	3.634$\pm$.026	9.531 $\pm$.073	.451$\pm$.031	2.822 $\pm$.069
Ours MDM-GUO	.704 $\pm$.005	3.641 $\pm$.025	9.495$\pm$.071	.486 $\pm$.031	2.832 $\pm$.066
ReMoDiffuse [99]	.795 $\pm$.004	2.974 $\pm$.016	9.018 $\pm$.075	.103 $\pm$.004	1.795 $\pm$.043

To demonstrate that models aligned with MoDiPO are more consistent, even when generating multiple motions for the same prompt, we synthesized eight different motions for each prompt in the HumanML3D and KIT-ML test sets and ranked them using the Guo ranker, with the results shown in Table 3.3 and Table 3.4 respectively. Here we computed all evaluation metrics specifically for the worst and best motion sequences from each set of eight generations. MLD exhibits a significant difference in FID scores between the worst and best motions, while MoDiPO-aligned MLD significantly reduces the difference between the FID scores for loser and winner motions. In particular, we observe a gap reduction of 15.05% (from 0.711 to 0.604) for HumanML3D and of 96.51% (from 0.172 to 0.006) for KIT-ML. It is worth noting that the FID score for losers decreases from 0.997 to 0.789 (an improvement of 20.86%) for HumanML3D and from 1.037 to 0.769 (an improvement of 25.84%) for KIT-ML. Besides the reduced gap between loser and winner, the lower overall FID for the aligned model suggests that MoDiPO is a practical solution for improving the consistency of motion quality in motion diffusion models.

Table 3.2. Results for KIT-ML. Each group of results corresponds to the base models, aligned models, and current SOTA backed by a RAG framework. *Italic* is for best results overall, while **bold** represents best results among base models and MoDiPO-aligned ones.

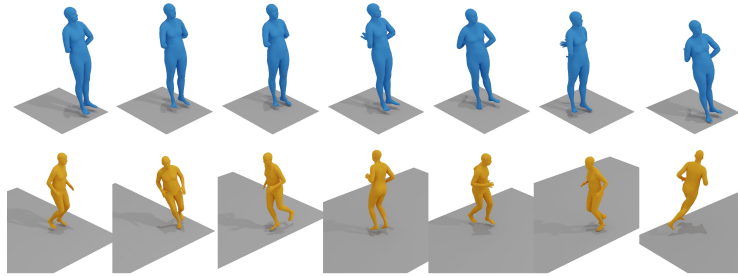
Model	Top3 $\uparrow$	MMDist $\downarrow$	Diversity $\rightarrow$	FID $\downarrow$	MM $\uparrow$
Real	.779 $\pm$.006	2.788 $\pm$.012	11.08 $\pm$.097	.031 $\pm$.004	-
MotionGPT [95]	.680 $\pm$.005	3.527 $\pm$.021	10.35 $\pm$.084	.510 $\pm$.016	2.328 $\pm$.117
T2M-GPT [96]	.745 $\pm$.006	3.007 $\pm$.023	10.92 $\pm$.108	.514 $\pm$.029	1.570 $\pm$.039
MotionDiffuse [105]	.739 $\pm$.004	2.958 $\pm$.005	11.10 $\pm$.143	1.954 $\pm$.062	0.730 $\pm$.013
MLD [81]	.718 $\pm$.005	3.156$\pm$.023	10.921 $\pm$.108	.733 $\pm$.032	1.595 $\pm$.087
Ours MLD-GUO	.720$\pm$.006	3.166 $\pm$.03	11.022$\pm$.074	.623$\pm$.037	1.660$\pm$.068
MDM [80]	.726$\pm$.005	3.098 $\pm$.023	10.690$\pm$.088	.496 $\pm$.028	1.829$\pm$.037
Ours MDM-GUO	.725 $\pm$.005	3.097$\pm$.023	10.681 $\pm$.086	.450$\pm$.026	1.821 $\pm$.038
ReMoDiffuse [99]	.765 $\pm$.055	2.814 $\pm$.012	10.80 $\pm$.105	.155 $\pm$.006	1.239 $\pm$.028

Table 3.3. Performance of Winner (W) and Loser (L) among the 8 generations per-prompt of MLD and MLD aligned with MoDiPO (Ours) for HUMANML3D.

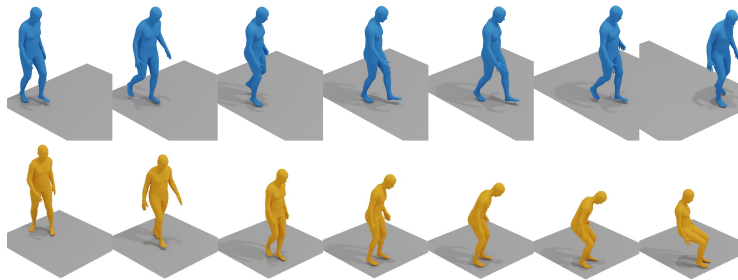
Model	Top3 $\uparrow$	MMDist $\downarrow$	Diversity $\rightarrow$	FID $\downarrow$
Real	.797 $\pm$.002	2.974 $\pm$.008	9.503 $\pm$.065	.002 $\pm$.0
MLD	.755 $\pm$.003	3.292 $\pm$.010	9.793 $\pm$.072	.459 $\pm$.011
MLD L	.511 $\pm$.003	4.822 $\pm$.014	9.249 $\pm$.079	.997 $\pm$.026
MLD W	.914 $\pm$.001	2.341 $\pm$.004	9.942 $\pm$.075	.286 $\pm$.007
Ours	.753 $\pm$.003	3.294 $\pm$.010	9.702 $\pm$.075	.281 $\pm$.007
Ours L	.487 $\pm$.002	4.945 $\pm$.014	9.056 $\pm$.069	.789 $\pm$.021
Ours W	.916 $\pm$.001	2.303 $\pm$.007	9.873 $\pm$.079	.185 $\pm$.005

Table 3.4. Performance of Winner (W) and Loser (L) among the 8 generations per-prompt of MLD and MLD aligned with MoDiPO (Ours) for KIT-ML.

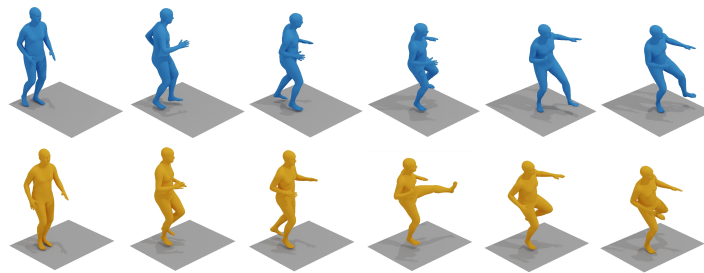
Model	Top3 $\uparrow$	MMDist $\downarrow$	Diversity $\rightarrow$	FID $\downarrow$
Real	.779 $\pm$.006	2.788 $\pm$.012	11.080 $\pm$.097	.031 $\pm$.004
MLD	.718 $\pm$.005	3.156 $\pm$.023	10.921 $\pm$.108	.733 $\pm$.032
MLD L	.619 $\pm$.007	4.001 $\pm$.034	10.829 $\pm$.100	1.037 $\pm$.053
MLD W	.799 $\pm$.006	2.672 $\pm$.024	11.302 $\pm$.115	.865 $\pm$.039
Ours	.720 $\pm$.006	3.166 $\pm$.030	11.022 $\pm$.074	.623 $\pm$.037
Ours L	.615 $\pm$.004	4.048 $\pm$.026	10.805 $\pm$.106	.769 $\pm$.047
Ours W	.802 $\pm$.005	2.646 $\pm$.022	11.297 $\pm$.114	.763 $\pm$.034



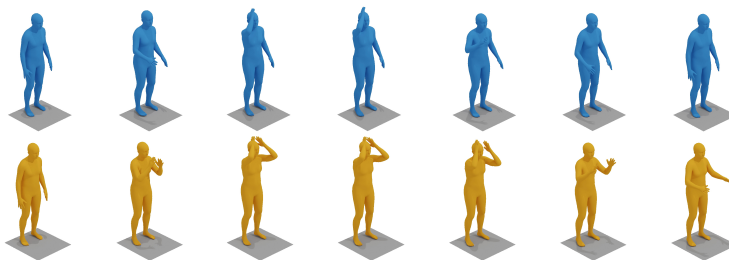
(a) A person is jogging around



(b) A person walks forwards, sits



(c) A man steps forward vigorously and then kicks something low with his left leg



(d) A man touches his head with both hands before lowering them

Figure 3.3. Qualitative Examples

3.5.3 Qualitative Evaluation

To perform qualitative evaluation, we generate motions using our best-performing model on HumanML3D. Fig. 3.3 compares MLD (in blue) with the MoDiPO-aligned MLD (in yellow). In Fig. 3.3a, the motion generated by MLD is visibly unnatural, with the character’s arms crossed behind its back. In contrast, the motion produced by MoDiPO is significantly more realistic and avoids this issue. Similarly, in Fig. 3.3b, the aligned model better captures the transition from standing to sitting. Moreover, the motion is more coherent with the prompt, and the same holds for Fig. 3.3c and Fig. 3.3d. In particular, Fig. 3.3c illustrates how MoDiPO places greater emphasis on the execution of the kick. Additionally, Fig. 3.3d demonstrates the model’s enhanced ability to capture subtle prompt details. For instance, when prompted with “the man moves both of his hands” MoDiPO-aligned MLD generates a motion in which both hands indeed move, unlike the MLD model, which only moves one hand. These examples demonstrate the effectiveness of MoDiPO in enhancing the quality and realism of generated motions.

3.5.4 User Study

To evaluate the quality of the motions generated by MLD aligned with MoDiPO, compared with those produced by MLD, we conducted a User Study.

The annotation interface, depicted in Fig. 3.4, includes:

- A detailed description of the task, including instructions for the evaluators and judging criteria for choosing the “best” GIF;
- A textual prompt describing a motion;
- A pair of GIFs depicting the corresponding motion as generated by MLD and MoDiPO-aligned MLD, in a random order;
- Four buttons for selecting the user’s preferred motion:
 - “*First*” or “*Second*” to choose the motion that best suits the given prompt, preferring the motion with greater smoothness and plausibility or smaller presence of foot sliding effect if the two motions equally represent the prompt;
 - “*Both*” if both of the motions match the prompt and the user cannot see differences in other aspects;
 - “*None*” if none of the motions match the prompt and the user cannot see differences in other aspects.

We collect as feedback 751 votes from 17 evaluators over a sample of 85 pairs of motions. We select people with expertise in Computer Vision and Deep Learning, best suited to a full understanding of their task. The results show that motions from MoDiPO-aligned MLD are selected as the best one 53.6% of the time, MLD’s motions are preferred 15.5% of the time, while 30.9% of the time the users cannot choose between the two motions. These results show that the alignment with MoDiPO improves the quality of the generated motion, including its realism, smoothness,

Table 3.5. Ablation study with MLD model on HumanML3D dataset. Stoch stands for stochastic selection of pairs, default is edge selection. GT_p is the probability of ground truth being used as winner. Online for online generation and ranking with the reference model. PaM+ is PaM with generations of both MLD and MDM used for the winner-loser pair selection.

Scorer	Stoch.	GT_p	Online	PaM+	Top3 $\uparrow$	MMDist $\downarrow$	Diversity $\rightarrow$	FID $\downarrow$	MM $\uparrow$
without alignment					.755 $\pm$.003	3.292 $\pm$.010	9.793 $\pm$.072	.459 $\pm$.011	2.647 $\pm$.110
Guo	$\checkmark$				.753 $\pm$.003	3.294 $\pm$.010	9.702 $\pm$.075	.281 $\pm$.007	2.736 $\pm$.111
TMR	$\checkmark$				.758 $\pm$.002	3.267 $\pm$.010	9.747 $\pm$.077	.303 $\pm$.007	2.663 $\pm$.111
Guo					.752 $\pm$.003	3.300 $\pm$.010	9.714 $\pm$.077	.336 $\pm$.009	2.736 $\pm$.110
TMR					.759 $\pm$.002	3.262 $\pm$.010	9.714 $\pm$.079	.325 $\pm$.007	2.630 $\pm$.107
Guo	$\checkmark$	.25			.745 $\pm$.002	3.324 $\pm$.010	9.674 $\pm$.077	.261 $\pm$.007	2.834 $\pm$.119
Guo	$\checkmark$	.5			.745 $\pm$.003	3.327 $\pm$.010	9.614 $\pm$.076	.273 $\pm$.007	2.826 $\pm$.117
Guo	$\checkmark$		$\checkmark$		.677 $\pm$.003	3.701 $\pm$.013	9.241 $\pm$.079	.276 $\pm$.008	3.477 $\pm$.144
Guo	$\checkmark$	.25	$\checkmark$		.714 $\pm$.003	3.496 $\pm$.010	9.489 $\pm$.074	.293 $\pm$.009	3.179 $\pm$.128
Guo	$\checkmark$			$\checkmark$	.761 $\pm$.003	3.252 $\pm$.010	9.851 $\pm$.071	.335 $\pm$.008	2.610 $\pm$.110
TMR	$\checkmark$			$\checkmark$	.759 $\pm$.002	3.266 $\pm$.010	9.798 $\pm$.073	.345 $\pm$.007	2.592 $\pm$.104

and adherence to the given textual prompt. Furthermore, the human evaluation supports the evidence given by the FID reported in Sec. 3.5.2, which is better for MoDiPO-aligned MLD than for MLD.

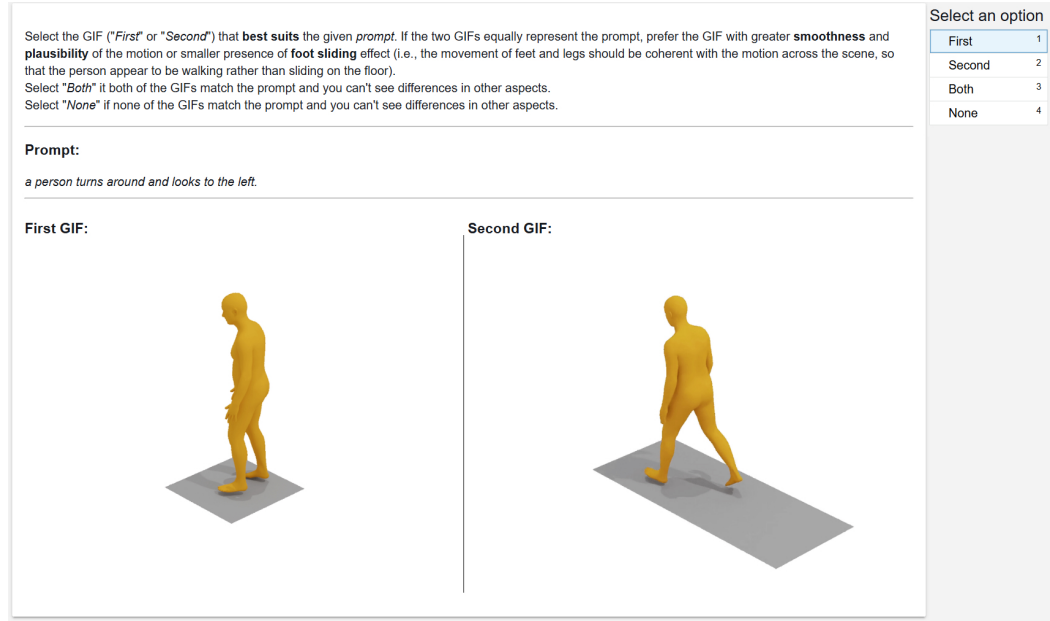


Figure 3.4. User Study Interface

3.6 Ablation Studies

We conducted a series of experiments to analyze the impact of specific choices within the proposed method. In Sec. 3.6.1, we investigated the effects of using stochastic or edge sampling of winner-loser pairs for the ranking process. Sec. 3.6.2 covers the evaluation of performances for different ranking methods used within

the framework. In Sec. 3.6.3, we describe the advantages and disadvantages of using motions generated by models other than the model we aim to align. Finally, Sec. 3.6.4 illustrates an online alignment experiment.

3.6.1 Pair Selection

Leveraging the multiple motions generated for each prompt in PaM, we develop two strategies, “edge” and “stochastic” selection, for determining winners and losers. In the edge selection strategy, we take the highest-ranked sequence as the “winner” and the lowest-ranked sequence as the “loser” from the ordered list of generated motions. In the stochastic selection, instead, we divided the list into two parts with potential winners in the first half, and potential losers in the second. According to our experiments, reported in Table 3.5, this second strategy performs better if compared to the edge selection. Subjected to different generations for the same textual prompt, models aligned with MoDiPO retain their RPrecision and Multi-Modality performances while also reducing the FID. We further experiment by using the ground truth motions as winners with a probability of 0.25 or 0.5, to enhance the generation of realistic motions. However, as depicted in Table 3.5, even if this setting helps decrease FID and increase Multi-Modality, it deteriorates RPrecision performances.

3.6.2 Ranking Method

As described in Sec. 3.4.1, we propose to synthetically rank generated motions starting from two models: TMR and Guo. While TMR has been proposed to tackle retrieval tasks, Guo was originally introduced for evaluation purposes. For both the used datasets, KIT-ML and HumanML3D, we annotated each set composed of text motions with Guo. Instead, TMR was used only for HumanML3D annotation, since it uses a data representation for KIT-ML which is incompatible with the one used by MLD and MDM. This prevents the use of TMR for KIT-ML synthetic ranking. Results of experiments done with both the used rankers are available in Table 3.5.

3.6.3 Alignment with generations from other models

In this section, we investigate whether increasing the number of winner-loser pairs helps maintain both Multi-Modality and RPrecision performance. To this end, we enhance the diversity of the training sequences by incorporating motions generated by MDM during the tuning of MLD, namely PaM+, evaluating whether this broader motion set yields additional benefits. However, as shown in Table 3.5, this approach did not yield superior results. The performance of the target model decreases if compared to using only its own generated motions for tuning. This is because by employing models with generations that have different distributions, we encourage our model to generate motions that fall between the two distributions, which results in decreased performance.

3.6.4 Online Alignment

We explored an alternative approach using online alignment with MLD. This approach is feasible due to MLD’s latent-space operation, which significantly reduces the time required to generate output motions if compared to MDM.

In this scenario, the frozen reference model generates motions which are then ranked and from which we select winner-loser pairs using the stochastic selection. This entire process occurs online, offering potential benefits in terms of efficiency and adaptability.

However, initial results show a trade-off between different metrics. While the target model’s FID decreases, suggesting improved quality in distribution similarity, the RPrecision Top3 metric also decreases. One possible explanation for these results is that, in the online setting, each motion is encountered only once during the entire alignment process. On the contrary, processing the same motion multiple times, as occurs when using PaM as the alignment dataset, is necessary to achieve better performances across all metrics.

3.7 Conclusions

The findings of this study demonstrate that preference-based alignment can substantially enhance the quality of text-to-motion generation. By applying Direct Preference Optimization with AI feedback, MoDiPO achieves improvements in motion plausibility while preserving semantic accuracy and generative diversity. The reductions in FID reflect not only statistical gains but also perceptual improvements, with outputs that are smoother and more human-like. Perhaps more importantly, the narrowing gap between best and worst-case generations suggests that aligned models consistently avoid implausible outputs, reducing the need for users to filter through multiple samples (Sec. 3.5.2). This consistency is particularly significant for downstream applications such as animation, gaming, and virtual reality, where unrealistic movements can break immersion and undermine usability.

These contributions position MoDiPO within a broader effort to align generative models with human intent. In text-to-motion research, earlier approaches leveraged stochastic diversity as a strength but were prone to producing suboptimal outputs [93, 80, 81]. Retrieval-based methods offered improvements by grounding generations in real motion samples [99], yet required large databases and introduced additional computational cost [79]. MoDiPO provides an alternative: it internalizes alignment directly into the generative process, avoiding reliance on external retrieval at inference time while still ensuring plausible, text-faithful results. From a methodological perspective, this work extends DPO, originally explored in natural language and image generation [76, 77], to the motion domain. It also joins a growing body of research showing that synthetic, AI-generated preferences can effectively substitute for expensive human annotations [90, 91, 88]. Our results confirm that rankings derived from motion-text retrieval models not only correlate with human judgments but also transfer effectively into alignment, reinforcing synthetic feedback as a practical and scalable resource.

Beyond demonstrating feasibility, this work opens several directions for future exploration. Improving the quality of ranking mechanisms remains a central challenge.

Current retrieval-based rankers [79, 78] focus primarily on semantic correspondence, but future models could incorporate dimensions of physical realism, style consistency, or biomechanical plausibility. Such richer evaluators could provide more precise feedback signals, pushing generative models toward higher fidelity across multiple criteria. Expanding the framework beyond binary winner–loser comparisons toward graded or multi-aspect evaluations would also allow finer control over alignment.

In summary, MoDiPO demonstrates that AI-feedback-driven alignment is an effective and scalable strategy to improve the quality, realism, and reliability of generative models. By constraining stochastic diversity within AI-preferred boundaries, it ensures both variety and plausibility, laying a foundation for future systems that are both creative and dependable across a range of generative applications.

Chapter 4

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

4.1 Overview

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-TEExt 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.

4.2 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 4.1). This highlights the necessity for strong **coherence** and plausibility among intentions (text), movements (motion), and environments (scene) to create a scenario where

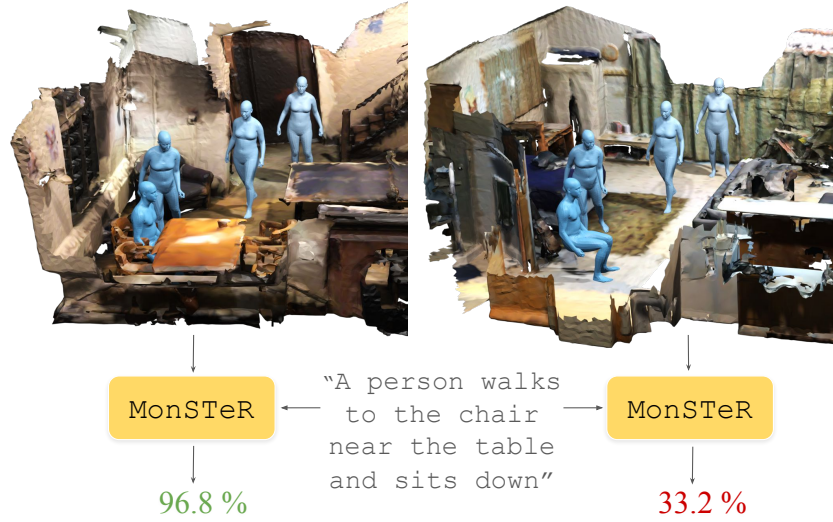


Figure 4.1. MonSTeR can estimate coherence among text, motion, and scene by embedding them into a unified latent space. In the left image, all three modalities are coherent. However, in the right image, this coherence decreases, as there is no chair available in the scene.

human-scene interaction can naturally take place. Also, multiple captions could potentially describe the motion in Fig. 4.1 in isolation, but given the scene context, it is evident that the provided one is the most accurate. This confirms that there is 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 [78, 79, 106, 107] 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, 108, 109]. 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 evaluations 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-TEText 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 [110], 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.

4.3 Related Works

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

4.3.1 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 [78, 111, 79, 112, 106, 107]. In particular, [111] initiated this field by aligning texts, motions, and image renderings of the motions. Additionally, [79, 112] proposed an effective and lightweight model to address this task and introduced several evaluation protocols for benchmarking. Existing approaches [79] represented motions through velocities and rotations in root space [78] or tackled this by encoding motions as image patches [106] 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.

4.3.2 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, [113, 114, 115, 116, 117] align all modalities to one modality (or a subset) chosen as a reference. Alternatively, [118, 119, 120] align each modality with all the others, creating intricate interactions among modalities. Particularly relevant to our task, [120] 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.

4.3.3 Modality Aggregation

Early approaches used separate encoders for each modality, combining them via latent averaging [113, 119], or adopted conditional generation—e.g., generating scenes from motion [121, 122, 123] or motion from scenes [124, 125]. 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 [126, 127], 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, [128] aligns global (multi-modal) and per-modality encoders, but lacks the flexibility to encode partial modality combinations into a unified representation. Finally, [129] 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.

4.4 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.

4.4.1 Data Representation

Following [79], we use DistilBERT [130] 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 [131, 132], 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.

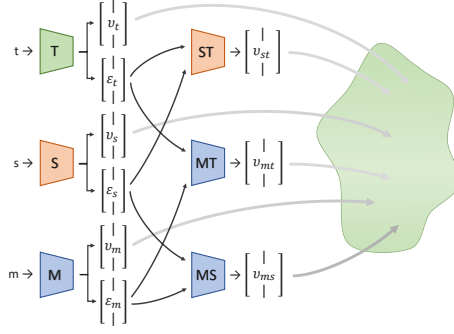


Figure 4.2. MonSTeR’s Architecture Overview. Each input modality t, s, m is processed by its single-modality encoder. From the first output tokens of T, M, S we sample vectors v_t, v_m , and v_s . The remaining tokens of each encoder’s output, namely $\varepsilon_t, \varepsilon_m, \varepsilon_s$, are pairwise concatenated and passed through cross-modal encoders to generate joint latent vectors (v_{st}, v_{mt}, v_{ms}) .

4.4.2 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 ambiguity¹. Unlike pairwise alignments, this triplet captures higher-order interactions between the modalities, providing a richer and more coherent representation of the underlying event [79].

Higher-Order Relations in Topological Spaces. Topology provides a robust framework to analyze and extract insights from data [133, 110]. 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 [134] 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. 4.5.3.

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. 4.2 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

¹Some ambiguity still exists, as underlying intentions and personal preferences influence human movement

Algorithm 1 Encoders' Input–Output Processing**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 $\mathbf{v}_{ij} \in \mathbb{R}^{256}$

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

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 $\epsilon_t, \epsilon_m, \epsilon_s$, are pairwise concatenated and processed by cross-modal encoders (e.g. ϵ_s and ϵ_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} . Algorithm 1 illustrate the complete encoding pipeline used in MonSTeR².

4.4.3 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 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. 4.3 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 [135], which we aggregate in the final loss:

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

²For 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.

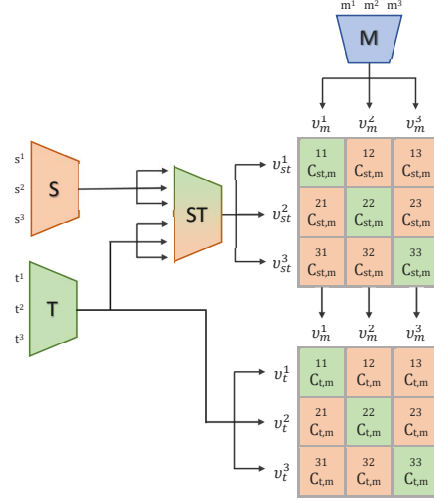


Figure 4.3. Composition of $C_{st,m}$ and $C_{t,m}$ similarity matrices. For $C_{st,m}$, motion latents v_m^n are compared with the cross-modal scene-text latents v_{st}^n , while for $C_{t,m}$, they are compared only with text latents v_t^n . Green cells are the locations of similarity scores between positive samples. Our optimization objective promotes assigning higher scores to these cells in all matrices $C_{i,j}$.

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 4.2, and then calculate its cosine similarity with the outputs of the M encoder.

Implementation Details. In MonSTeR, T and M encoders are based on [136], while we use [137] as S , augmented with Spatial Attention layers. For the cross-modal encoders, we start from the unified encoder from [131] and modify it to produce a single joint modalities’ representation via cross-attention between modalities and projection layers. Our model incorporates a motion decoder [136] 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 [79].

4.5 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.

Table 4.1. Tasks and protocols’ results on HUMANISE+ [4] 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 [79]	2.73	—	2.05	—	—	—	2.73	2.05	—	—	—	—	—
	MoPa [106]	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	<u>1.41</u>	1.02	<u>1.24</u>	<u>1.38</u>	<u>2.72</u>
	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	<u>3.62</u>	3.11	1.12	<u>1.16</u>	1.68	2.51	4.80
Small Batches	TMR [79]	56.0	—	56.01	—	—	—	56.0	56.01	—	—	—	—	—
	MoPa [106]	<u>61.12</u>	—	61.39	—	—	—	<u>61.12</u>	<u>61.39</u>	—	—	—	—	—
	TMR + S	57.08	<u>55.49</u>	70.77	67.05	33.86	37.26	59.42	58.17	27.78	25.73	38.31	38.45	<u>47.44</u>
	MoPa + S	48.83	40.73	40.46	39.23	<u>48.37</u>	<u>40.73</u>	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	<u>34.11</u>	<u>34.47</u>	50.18	49.76	60.00

4.5.1 Datasets

To benchmark MonSTeR, we use two recent datasets, [4, 5]. 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 [4] 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 LLAMA3-8B-it [138] to recaption the original texts to include references to nearby objects and express spatial context around the motion path. 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. This approach substantially enhances the grounding of text descriptions within the scene as shown in Fig. 4.4. We refer to this dataset as HUMANISE+.

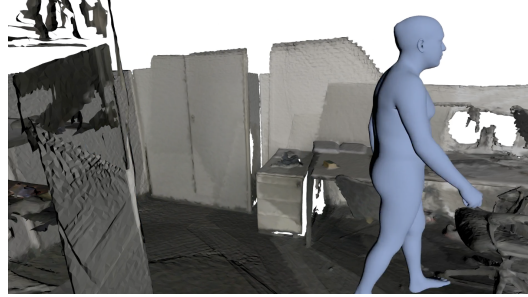
TRUMANS+. TRUMANS [5] comprises 15 hours of motion-captured data representing human interacting with 100 varied indoor scenes and 20 distinct object types. While TRUMANS texts’ describe the motion accurately, their connection to the scene is not strong, primarily focusing on the object of interaction. 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. A recaptioning process identical to that of HUMANISE+ was not feasible for TRUMANS due to lack of detailed scene segmentation labels. For this reason, we recaption using only the objects mentioned in the caption and explicitly present in the mesh file, while keeping the rest of the pipeline the same, to reinforce alignment between text, motion, and scene content. We refer to this dataset as TRUMANS+.



(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 4.4. Comparison between HUMANISE captions and HUMANISE+ recaptioning.

4.5.2 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

Table 4.2. Tasks and protocols’ results on TRUMANS+ [5] 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 [79]	17.55	—	17.71	—	—	—	17.55	17.71	—	—	—	—	—
	MoPa [106]	4.22	—	6.58	—	—	—	4.22	6.58	—	—	—	—	—
	TMR + S	4.04	4.44	<u>9.01</u>	<u>4.67</u>	9.14	1.61	5.65	<u>12.03</u>	1.11	5.89	6.74	<u>1.65</u>	<u>5.49</u>
	MoPa + S	4.83	<u>5.20</u>	4.11	4.39	13.20	<u>2.14</u>	3.52	3.85	2.06	10.79	<u>8.30</u>	1.04	5.28
	MonSTeR	<u>10.54</u>	10.29	8.59	8.51	<u>12.71</u>	3.94	<u>5.67</u>	5.90	<u>1.94</u>	<u>7.88</u>	10.35	2.80	7.42
Small Batches	TMR [79]	82.2	—	82.13	—	—	—	82.2	82.13	—	—	—	—	—
	MoPa [106]	48.55	—	56.44	—	—	—	48.55	<u>56.44</u>	—	—	—	—	—
	TMR + S	48.46	<u>47.63</u>	54.05	51.85	30.38	26.7	<u>53.38</u>	55.12	23.24	25.27	30.06	25.37	39.29
	MoPa + S	42.40	44.84	52.76	<u>52.99</u>	<u>46.50</u>	43.66	42.20	42.35	39.41	43.62	<u>31.59</u>	<u>29.12</u>	<u>42.60</u>
	MonSTeR	<u>60.8</u>	61.4	<u>58.2</u>	57.86	46.99	<u>42.62</u>	49.49	50.13	<u>30.41</u>	<u>33.88</u>	37.65	35.08	47.05

Table 4.3. Ablation Studies for MonSTeR performed on HUMANISE+ [4]. 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	13.91	13.14	<u>8.46</u>	10.39	<u>4.09</u>	4.45	3.62	3.11	<u>1.12</u>	<u>1.16</u>	1.68	<u>2.51</u>	5.63
	- w/o cross-modal	5.20	3.77	8.91	7.95	2.49	2.59	4.35	<u>3.21</u>	1.01	0.91	2.52	2.64	3.79
	- w/o single	<u>11.91</u>	<u>12.93</u>	8.35	<u>8.96</u>	4.33	<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>	4.37	2.00	1.66	<u>2.10</u>	1.63	4.14
Small Batches	MonSTeR	79.15	79.01	76.30	77.42	57.55	57.06	<u>63.20</u>	<u>61.89</u>	<u>34.11</u>	<u>34.47</u>	50.18	49.76	60.00
	- w/o cross-modal	65.47	63.26	78.11	<u>76.76</u>	49.89	49.31	63.98	62.76	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	36.57	37.15	45.80	44.70	53.70

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 [129], which represents the average Recall across ranks {1, 2, 3, 5, 10}.

Protocols. The evaluation adheres to the retrieval protocols outlined in [79, 78]. 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.

Retrieval Results. In Table 4.1 and Table 4.2, 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 [79] and MoPa [106]. 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 [113].

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.

Table 4.4. Ablation Studies for MonSTeR performed on TRUMANS+ [5]. 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	10.54	10.29	8.59	8.51	12.71	3.94	5.67	5.9	1.94	7.88	10.35	2.83	7.42
	- w/o cross-modal	5.89	6.26	6.58	6.3	12.01	3.1	6.95	6.84	<u>1.98</u>	9.05	<u>8.26</u>	2.06	<u>6.27</u>
	- w/o single	12.95	13.01	7.01	6.28	12.99	4.76	0.19	0.06	0.49	3.67	2.71	0.07	5.34
	- w tri-modal	5.62	4.91	6.24	<u>6.34</u>	10.29	3.05	<u>6.31</u>	<u>6.3</u>	2.04	<u>8.36</u>	8.19	<u>2.58</u>	6.02
Small Batches	MonSTeR	<u>60.8</u>	<u>61.4</u>	58.2	57.86	<u>46.99</u>	<u>42.62</u>	<u>49.49</u>	<u>50.13</u>	<u>30.41</u>	<u>33.88</u>	37.65	35.08	47.04
	- w/o cross-modal	50.96	51.12	50.12	50.75	38.12	33.98	52.76	52.91	29.91	33.56	33.73	31.08	42.41
	- w/o single	66.96	68.43	<u>56.38</u>	<u>56.46</u>	51.59	48.38	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	31.33	35.23	<u>37.08</u>	<u>33.52</u>	<u>46.72</u>

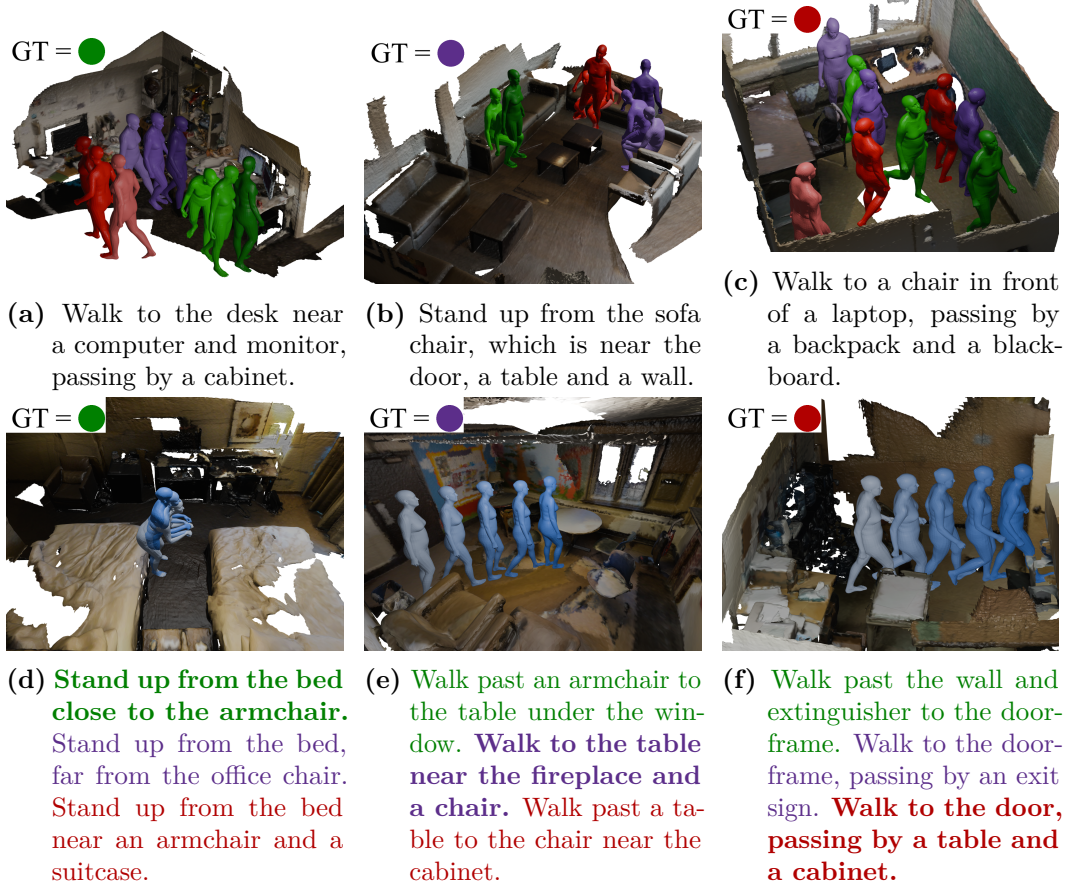


Figure 4.5. Qualitative examples for *st2m* (4.5a, 4.5b, 4.5c) and *ms2t* (4.5d, 4.5e, 4.5f). **First**, **second**, and **third** retrieved samples are shown. In each pictorial, GT=color (top-left corner) indicates the correct corresponding motion (top row) and text (bottom row)

TMR capitalizes on this with Guo features [78], which ignore spatial ties between scene and motion and thus make tasks other than *t2m*/*m2t* impossible. 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 average, MonSTeR



(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.



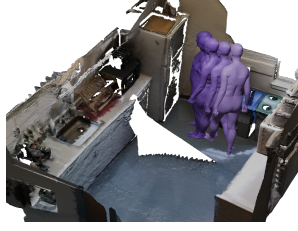
(c) A person is walking to the toilet, passing by a sink.

Figure 4.6. Qualitative examples for $t2m$. **First**, **second**, and **third** retrieved samples are shown.

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. 4.5 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.

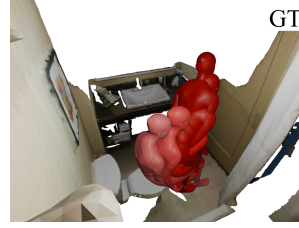
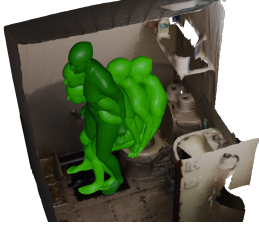
Fig. 4.6 presents qualitative results of MonSTeR’s rankings for the $t2m$ task and Fig. 4.7 for the $mt2s$ task. For each example in the first task, we show the first three motions retrieved from the given text prompt. Note that here they are shown together with the respective scene, even if it is not 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.



(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 4.7. Qualitative examples for *mt2s*. **First**, **second**, and **third** retrieved samples are shown.

4.5.3 Ablation Studies on the Model Design

In Table 4.3 and Table 4.4 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. 4.4.3, 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 compute the cross-modal representations [113]. 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 4.3 and Table 4.4). 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 [129]. 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.

4.5.4 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 [81, 78, 139, 80]. 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. 4.2). We extend the Recall [78], 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 [4, 5], starting at an angle of 0 radians and increasing incrementally up to π , 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. 4.8 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 [5] 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. 4.8 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

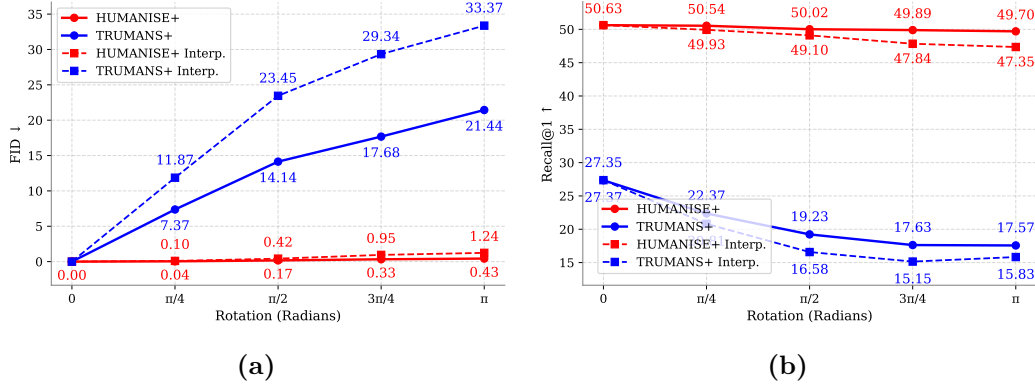


Figure 4.8. MonSTeR’s FID $\downarrow$ (a) and Recall@1 $\uparrow$ (b) trends when motions are increasingly rotated, from 0 to π radians.

Table 4.5. Generation Results on HUMANISE+ [4].

Model	goal distance $\downarrow$	contact $\uparrow$	non-collision $\uparrow$	Recall@1 $\uparrow$	Recall@2 $\uparrow$	Recall@3 $\uparrow$	FID $\downarrow$
MaYS [108]	0.168	0.94	0.996	0.43 ± 0.002	0.62 ± 0.003	0.74 ± 0.002	2.157
cVAE [4]	0.596	0.855	0.997	0.33 ± 0.004	0.49 ± 0.003	0.59 ± 0.002	32.857

enable motion rotations that would lead to interpenetration. In Fig. 4.8, 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.

Human Scene Interaction Evaluation We evaluate human scene interaction methods using Recall@1,2,3 and FID calculated using MonSTeR, alongside commonly used metrics for this task [4, 108], including goal distance, contact, and non-collision. The evaluation is conducted on two state-of-the-art models: a two-stage diffusion model (dubbed as MaYS) [108], and the conditional variational auto-encoder (cVAE) from [4].

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 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 4.5 summarizes the results. Both MaYS and cVAE demonstrate strong performance in maintaining appropriate motion-scene contact and avoiding interpenetration with 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. MaYS exhibits superior performance in goal distance, reflecting its ability to generate

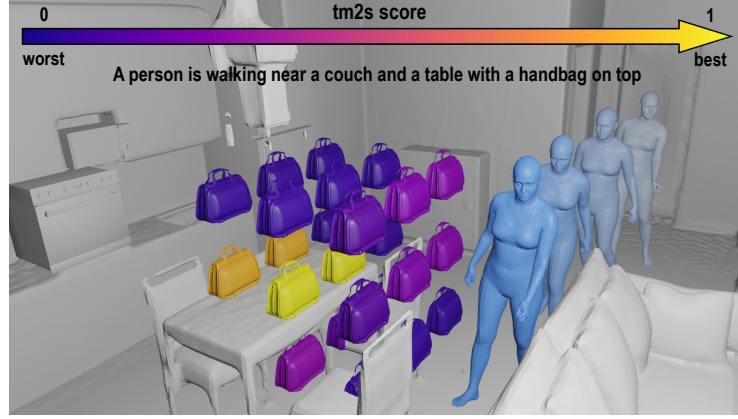


Figure 4.9. In-Scene Object Placement scores visualized

motion trajectories that more closely align with the target destination.

These trends are further corroborated by the FID scores in Table 4.5, 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. As seen in the path plausibility evaluation, MonSTeR-based FID is sensitive to implausible or incorrect paths followed by motion trajectories, an aspect not captured by goal distance, contact, or non-collision metrics.

Alignment to Human Preferences. We perform a user study to test the alignment of MonSTeR’s ranking with human preferences. We generate motions using [4] and [108]. 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.

4.5.5 Downstream Tasks

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. 4.9. To this aim, we extrapolate objects of interest from the scene, and set up a $5 \times 5 \times 5$ grid around their ground-truth positions. Fig. 4.10 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;

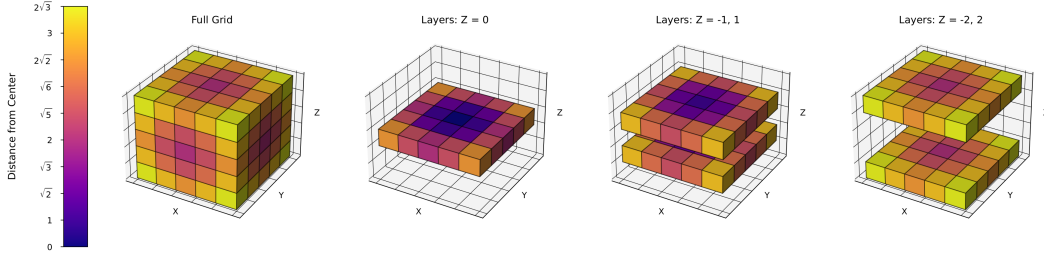


Figure 4.10. Visualization of the $5 \times 5 \times 5$ grid used for the in-scene object placement evaluation. The full grid (leftmost) shows all cells. Subsequent panels depict sliced layers along the z -axis: $z = 0$ (second from left), $z = (-1, 1)$ (third), and $z = (-2, 2)$ (rightmost). Distances from the center represented with colors, with values going from 0 (darker colors) to $2\sqrt{3}$ (brighter colors).

- 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.

Using the defined grid, we shift the object into each of the 125 cells, generating an equal number of scene variants. 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 $((\text{text}, \text{motion}), \text{scene}_i)$, with $i \in [1, 125]$. We locate the object of interest by selecting the highest score provided by our model and compute the $L2$ distance to its real position.

In this scenario, being L the translation step, the *average error* is given by

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

Using a translation step of 25 cm the error ranges in $[0 \text{ cm}, 86.6 \text{ cm}]$, with the maximum corresponding to two steps taken along a diagonal path in the grid. While the average error corresponds to 58.98 cm, our zero-shot prediction scores only 18 cm. MonSTeR can accurately localize objects in a zero-shot setting, leveraging spatial knowledge grounded in the text-motion latent representations.

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 [140] which we then train to caption motions. We then compare our results with MotionGPT [95], retrained from scratch on both HUMANISE+ and TRUMANS+, adopting the evaluation protocol from [95]. As shown by Table 4.6 and Table 4.7, 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_L and Bert_{F1} 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.

Table 4.6. Captioning performances on HUMANISE+ [4].

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

Table 4.7. Captioning performances on TRUMANS+ [5].

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

4.6 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.

4.7 Conclusion

We introduced a novel retrieval model that unifies text, motion, and scene within a shared latent space. Inspired by topological deep learning, MonSTeR is designed to capture higher-order interactions across all three modalities. We validated it by performing retrieval tasks on individual and paired modalities, 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 analyzed the use of MonSTeR as a tool for grounding Human-Scene Interaction models within the scene context and evaluated its alignment with human preferences. We further demonstrated the versatility of our model on in-Scene Object Placement and Motion Captioning, showing its ability to generalize beyond traditional retrieval scenarios.

Chapter 5

Conclusions

This thesis explored three critical aspects of Human-Centric Computer Vision: the analysis of human crowds, the synthesis of human motion, and the contextual interplay between humans, their intentions, and their environments. We were motivated by a foundational gap between the capabilities of current AI models and the practical requirements of real-world applications. True human-centric computer vision must move beyond shallow metrics and instead quantify the physical space people occupy, ensure the plausibility of generated actions, and evaluate the coherence of human-scene interactions.

First, to address the flaws of traditional crowd analysis, Chapter 2 moved beyond simple headcounts to instead quantify human presence. We introduced the novel task of Crowd Volume Estimation (CVE) as a more meaningful task for public safety, comfort, and structural load assessment. To enable this, we developed ANTHROPOS-V, the first large-scale photorealistic dataset for this task, with careful alignment of its synthetic anthropometrics to real-world human data. We then proposed STEERER-V, a novel model whose volume-specific supervision signal demonstrated superior performance over both counting and Human Mesh Recovery baselines, successfully providing a more meaningful, capacity-based understanding of crowds.

Second, to tackle the challenge of implausible and uncontrolled motion generation, Chapter 3 focuses on an alignment method that produces more plausible and text-coherent motions. We introduced MoDiPO, a framework that successfully leverages Direct Preference Optimization to align text-to-motion diffusion models. By employing AI feedback to create the Pick-a-Move dataset, we avoided the expensive and slow process of human annotation. Our experiments confirmed that this alignment process yields significantly more realistic and desirable motions, as evidenced by a threefold increase in human evaluators' preferences, proving that the creative diversity of generative models can be successfully harnessed for plausibility.

Third, to provide a necessary metric for human-scene interaction, Chapter 4 introduced a method for retrieving motion-scene-text tuples. We presented MonSTeR, the first tri-modal model to embed motion, scene, and text into a unified latent space. While this space proved effective for flexible retrieval, its most crucial application is as a scene-aware evaluator. We validated its effectiveness by showing its scores strongly correlate with human judgments of plausibility and correctly penalize

interpenetrations and absurd interactions. At the same time, MonSTeR can also be used for downstream tasks such as motion captioning and in-scene object placement. This work provides the first reliable, global metric for assessing the complex coherence between a person’s intention, their actions, and their physical surroundings.

These three contributions represent a significant step toward more physically-grounded, preference-aligned, and context-aware solutions. Future work can build upon these foundations. For CVE, richer datasets incorporating diverse weather and lighting conditions could improve real-world robustness. For motion alignment, future work could move beyond binary preferences to incorporate richer, multi-aspect feedback, leading to generative models with even higher fidelity. For HSI, the unified latent space of MonSTeR could be leveraged as a generative prior to guide the creation of scene-aware motions directly, in a Retrieval Augmented Generation fashion.

In summary, this thesis delivers foundational tools such as models, datasets, objectives, and evaluators for building AI systems that can reason about humans in space: from their physical volume to the plausibility of their motion and their coherence within a scene.

Publications

1 International Conferences

1. *ANTHROPOS-V: benchmarking the novel task of Crowd Volume Estimation*
Luca Collorone¹, Stefano D'Arrigo¹, **Massimiliano Pappa**¹, Guido MD'Amely Di Melendugno, Giovanni Ficarra, Fabio Galasso; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2025, pp. 5284-5294.
2. *MonSTeR: a Unified Model for Motion, Scene, Text Retrieval*
Luca Collorone¹, Matteo Gioia¹, **Massimiliano Pappa**, Paolo Leoni, Giovanni Ficarra, Or Litany, Indro Spinelli, Fabio Galasso; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2025, pp. 10940-10949.

2 Under Submission

1. *MoDiPO: text-to-motion alignment via AI-feedback-driven Direct Preference Optimization*
Massimiliano Pappa¹, Luca Collorone¹, Giovanni Ficarra, Indro Spinelli, Fabio Galasso.

¹Equal Contribution.

Bibliography

- [1] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, “Crowdhuman: A benchmark for detecting human in a crowd,” *arXiv preprint arXiv:1805.00123*, 2018.
- [2] M. P. Silverman, “Exact statistical distribution of the body mass index (bmi): Analysis and experimental confirmation,” *Open Journal of Statistics*, vol. 12, no. 3, 2022.
- [3] Z. Ma, X. Wei, X. Hong, and Y. Gong, “Bayesian loss for crowd count estimation with point supervision,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6142–6151.
- [4] Z. Wang, Y. Chen, T. Liu, Y. Zhu, W. Liang, and S. Huang, “Humanise: Language-conditioned human motion generation in 3d scenes,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 14 959–14 971, 2022.
- [5] N. Jiang, Z. Zhang, H. Li, X. Ma, Z. Wang, Y. Chen, T. Liu, Y. Zhu, and S. Huang, “Scaling up dynamic human-scene interaction modeling,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1737–1747.
- [6] P. A. Thompson, “Developing new techniques for modelling crowd movement,” *KB thesis scanning project 2015*, 1994.
- [7] D. Sudharson, J. Srinithi, S. Akshara, K. Abhirami, P. Sriharshitha, and K. Priyanka, “Proactive headcount and suspicious activity detection using yolov8,” *Procedia Computer Science*, vol. 230, pp. 61–69, 2023.
- [8] M. Fiandero, T. T. Nguyen, H. Wong, and E. B. Hsu, “Modernized crowd counting strategies for mass gatherings—a review,” *Journal of Acute Medicine*, vol. 13, no. 1, p. 4, 2023.
- [9] C. Kappagoda, R. Linden, and J. Newell, “A comparison of the oxygen consumption/body weight relationship obtained during submaximal exercise on a bicycle ergometer and on a treadmill,” *Quarterly Journal of Experimental Physiology and Cognate Medical Sciences: Translation and Integration*, vol. 64, no. 3, pp. 205–215, 1979.
- [10] K. R. Jenkins, N. H. Fultz, S. J. Fonda, and L. A. Wray, “Patterns of body weight in middle-aged and older americans, by gender and race, 1993–2000,” *Sozial-und Präventivmedizin/Social and Preventive Medicine*, vol. 48, pp. 257–268, 2003.
- [11] D. Gallagher, M. Visser, D. Sepúlveda, R. N. Pierson, T. Harris, and S. B. Heymsfield, “How Useful Is Body Mass Index for Comparison of Body Fatness across Age, Sex, and Ethnic Groups?” *American Journal of Epidemiology*, vol. 143, no. 3, pp. 228–239, 02 1996. [Online]. Available: <https://doi.org/10.1093/oxfordjournals.aje.a008733>
- [12] P. Deurenberg, M. Yap, and W. A. Van Staveren, “Body mass index and percent body fat: a meta analysis among different ethnic groups,” *International journal of obesity*, vol. 22, no. 12, pp. 1164–1171, 1998.

- [13] J. V. Durnin and J. Womersley, "Body fat assessed from total body density and its estimation from skinfold thickness: measurements on 481 men and women aged from 16 to 72 years," *British journal of nutrition*, vol. 32, no. 1, pp. 77–97, 1974.
- [14] C. Pfitzner, S. May, C. Merkl, L. Breuer, M. Köhrmann, J. Braun, F. Dirauf, and A. Nüchter, "Libra3d: Body weight estimation for emergency patients in clinical environments with a 3d structured light sensor," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 2888–2893.
- [15] P. Hu, X. Dai, R. Zhao, H. Wang, Y. Ma, and A. Munteanu, "Point2partvolume: Human body volume estimation from a single depth image," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–12, 2023.
- [16] F. Leinen, V. Cozzolino, and T. Schön, "Volnet: estimating human body part volumes from a single rgb image," *arXiv preprint arXiv:2107.02259*, 2021.
- [17] X. Liu, J. Niu, L. Ran, and T. Liu, "Estimation of human body volume (bv) from anthropometric measurements based on three-dimensional (3d) scan technique," *Aesthetic Plastic Surgery*, vol. 41, pp. 971–978, 2017.
- [18] K. Pirker, M. Rüther, H. Bischof, F. Skrabal, and G. Pichler, "Human body volume estimation in a clinical environment," *AAPR/OAGM: challenges in the biosciences: image analysis and pattern recognition aspects*, Stainz Austria, 2009.
- [19] P. Patel, C.-H. P. Huang, J. Tesch, D. T. Hoffmann, S. Tripathi, and M. J. Black, "Agora: Avatars in geography optimized for regression analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 468–13 478.
- [20] M. J. Black, P. Patel, J. Tesch, and J. Yang, "Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8726–8737.
- [21] M. Fabbri, F. Lanzi, S. Calderara, A. Palazzi, R. Vezzani, and R. Cucchiara, "Learning to detect and track visible and occluded body joints in a virtual world," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 430–446.
- [22] M. Fabbri, G. Brasó, G. Maugeri, O. Cetintas, R. Gasparini, A. Ošep, S. Calderara, L. Leal-Taixé, and R. Cucchiara, "Motsynth: How can synthetic data help pedestrian detection and tracking?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 849–10 859.
- [23] Q. Song, C. Wang, Z. Jiang, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Wu, "Rethinking counting and localization in crowds: A purely point-based framework," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3365–3374.
- [24] T. Han, L. Bai, L. Liu, and W. Ouyang, "Steerer: Resolving scale variations for counting and localization via selective inheritance learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 848–21 859.
- [25] H. Lin, Z. Ma, R. Ji, Y. Wang, and X. Hong, "Boosting crowd counting via multifaceted attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 628–19 637.
- [26] Y. Wang and K. Daniilidis, "Refit: Recurrent fitting network for 3d human recovery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14 644–14 654.

- [27] Z. Li, J. Liu, Z. Zhang, S. Xu, and Y. Yan, "Cliff: Carrying location information in full frames into human pose and shape estimation," in *European Conference on Computer Vision*. Springer, 2022, pp. 590–606.
- [28] M. Roser, C. Appel, and H. Ritchie, "Human height," *Our World in Data*, 2021, <https://ourworldindata.org/human-height>.
- [29] N. Bándi, R.-B. Tunyogi, Z. Szabó, E. Farkas, and C. Sulyok, "Image-based volume estimation using stereo vision," in *2020 IEEE 18th International Symposium on Intelligent Systems and Informatics (SISY)*, 2020, pp. 000 055–000 060.
- [30] J. Dehais, M. Anthimopoulos, S. Shevchik, and S. Mougiakakou, "Two-view 3d reconstruction for food volume estimation," *IEEE Transactions on Multimedia*, vol. 19, no. 5, pp. 1090–1099, 2017.
- [31] F. P.-W. Lo, Y. Sun, J. Qiu, and B. Lo, "Food volume estimation based on deep learning view synthesis from a single depth map," *Nutrients*, vol. 10, no. 12, p. 2005, 2018.
- [32] F. P.-W. Lo, Y. Sun, J. Qiu, and B. P. Lo, "Point2volume: A vision-based dietary assessment approach using view synthesis," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 577–586, 2019.
- [33] M. Puri, Z. Zhu, Q. Yu, A. Divakaran, and H. Sawhney, "Recognition and volume estimation of food intake using a mobile device," in *2009 Workshop on Applications of Computer Vision (WACV)*, 2009, pp. 1–8.
- [34] C. Xu, Y. He, N. Khanna, C. J. Boushey, and E. J. Delp, "Model-based food volume estimation using 3d pose," in *2013 IEEE International Conference on Image Processing*. IEEE, 2013, pp. 2534–2538.
- [35] Z. Yang, H. Yu, S. Cao, Q. Xu, D. Yuan, H. Zhang, W. Jia, Z.-H. Mao, and M. Sun, "Human-mimetic estimation of food volume from a single-view rgb image using an ai system," *Electronics*, vol. 10, no. 13, p. 1556, 2021.
- [36] H. Hassannejad, G. Matrella, P. Ciampolini, I. D. Munari, M. Mordonini, and S. Cagnoni, "A new approach to image-based estimation of food volume," *Algorithms*, vol. 10, no. 2, p. 66, 2017.
- [37] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, oct 2015. [Online]. Available: <https://doi.org/10.1145/2816795.2818013>
- [38] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2547–2554.
- [39] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 589–597.
- [40] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah, "Composition loss for counting, density map estimation and localization in dense crowds," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [41] V. A. Sindagi, R. Yasarla, and V. M. Patel, "Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1221–1231.

- [42] —, “Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2594–2609, 2020.
- [43] Q. Wang, J. Gao, W. Lin, and X. Li, “Nwpu-crowd: A large-scale benchmark for crowd counting and localization,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 6, pp. 2141–2149, 2020.
- [44] S.-Y. Cho, T. W. Chow, and C.-T. Leung, “A neural-based crowd estimation by hybrid global learning algorithm,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 29, no. 4, pp. 535–541, 1999.
- [45] A. N. Marana, S. Velastin, L. Costa, and R. Lotufo, “Estimation of crowd density using image processing,” *IET Conference Proceedings*, 1997.
- [46] D. Kong, D. Gray, and H. Tao, “A viewpoint invariant approach for crowd counting,” in *18th International Conference on Pattern Recognition (ICPR’06)*, vol. 3. IEEE, 2006, pp. 1187–1190.
- [47] D. Liang, W. Xu, and X. Bai, “An end-to-end transformer model for crowd localization,” in *European Conference on Computer Vision*. Springer, 2022, pp. 38–54.
- [48] S. Zhang, G. Wu, J. P. Costeira, and J. M. Moura, “Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3667–3676.
- [49] Y. Li, X. Zhang, and D. Chen, “Csnet: Dilated convolutional neural networks for understanding the highly congested scenes,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1091–1100.
- [50] Q. Wang, J. Gao, W. Lin, and Y. Yuan, “Learning from synthetic data for crowd counting in the wild,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [51] B. Wang, H. Liu, D. Samaras, and M. H. Nguyen, “Distribution matching for crowd counting,” *Advances in neural information processing systems*, vol. 33, pp. 1595–1607, 2020.
- [52] Z. Ma, X. Wei, X. Hong, H. Lin, Y. Qiu, and Y. Gong, “Learning to count via unbalanced optimal transport,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 2319–2327.
- [53] J. Wan, Z. Liu, and A. B. Chan, “A generalized loss function for crowd counting and localization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1974–1983.
- [54] J. Cheng, H. Xiong, Z. Cao, and H. Lu, “Decoupled two-stage crowd counting and beyond,” *IEEE Transactions on Image Processing*, vol. 30, pp. 2862–2875, 2021.
- [55] C. Liu, Y. Zhong, A. Zisserman, and W. Xie, “Countr: Transformer-based generalised visual counting,” *arXiv preprint arXiv:2208.13721*, 2022.
- [56] Y. Ranasinghe, N. G. Nair, W. G. C. Bandara, and V. M. Patel, “Crowddiff: Multi-hypothesis crowd density estimation using diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 809–12 819.
- [57] V. Lempitsky and A. Zisserman, “Learning to count objects in images,” *Advances in neural information processing systems*, vol. 23, 2010.

- [58] V. Ranjan, U. Sharma, T. Nguyen, and M. Hoai, “Learning To Count Everything,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3394–3403. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Ranjan_Learning_To_Count_Everything_CVPR_2021_paper.html
- [59] Z. You, K. Yang, W. Luo, X. Lu, L. Cui, and X. Le. Few-shot Object Counting with Similarity-Aware Feature Enhancement. [Online]. Available: <http://arxiv.org/abs/2201.08959>
- [60] P. Thanasutives, K.-i. Fukui, M. Numao, and B. Kijirikul, “Encoder-Decoder Based Convolutional Neural Networks with Multi-Scale-Aware Modules for Crowd Counting,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2020, pp. 2382–2389. [Online]. Available: <http://arxiv.org/abs/2003.05586>
- [61] S. K. Dwivedi, Y. Sun, P. Patel, Y. Feng, and M. J. Black, “Tokenhmr: Advancing human mesh recovery with a tokenized pose representation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1323–1333.
- [62] Y. Zhu, A. Li, Y. Tang, W. Zhao, J. Zhou, and J. Lu, “Dpmesh: Exploiting diffusion prior for occluded human mesh recovery,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1101–1110.
- [63] H. Wen, J. Huang, H. Cui, H. Lin, Y.-K. Lai, L. Fang, and K. Li, “Crowd3d: Towards hundreds of people reconstruction from a single image,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8937–8946.
- [64] B. Huang, J. Ju, Z. Li, and Y. Wang, “Reconstructing groups of people with hypergraph relational reasoning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14 873–14 883.
- [65] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 7464–7475.
- [66] T. Von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, “Recovering accurate 3d human pose in the wild using imus and a moving camera,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 601–617.
- [67] G. Redux, “Gta - redux,” gta5redux.com, 2023, accessed on 17 October 2023.
- [68] W. H. Organization *et al.*, “Mean bmi (kg/m²)(crude estimate).[internet]. who. 2017,” 2017.
- [69] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, “Openpose: Realtime multi-person 2d pose estimation using part affinity fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [70] K. Gong, Y. Gao, X. Liang, X. Shen, M. Wang, and L. Lin, “Graphonomy: Universal human parsing via graph transfer learning,” in *CVPR*, 2019.
- [71] B. O. Community, *Blender - a 3D modelling and rendering package*, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. [Online]. Available: <http://www.blender.org>
- [72] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, “Deep reinforcement learning from human preferences,” *Advances in neural information processing systems*, vol. 30, 2017.

- [73] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon *et al.*, “Constitutional ai: Harmlessness from ai feedback,” *arXiv preprint arXiv:2212.08073*, 2022.
- [74] J. Menick, M. Trebacz, V. Mikulik, J. Aslanides, F. Song, M. Chadwick, M. Glaese, S. Young, L. Campbell-Gillingham, G. Irving *et al.*, “Teaching language models to support answers with verified quotes,” *arXiv preprint arXiv:2203.11147*, 2022.
- [75] X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li, “Human preference score: Better aligning text-to-image models with human preference,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2096–2105.
- [76] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, “Direct preference optimization: Your language model is secretly a reward model,” in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [77] B. Wallace, M. Dang, R. Rafailov, L. Zhou, A. Lou, S. Purushwalkam, S. Ermon, C. Xiong, S. Joty, and N. Naik, “Diffusion model alignment using direct preference optimization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8228–8238.
- [78] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, “Generating diverse and natural 3d human motions from text,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5152–5161.
- [79] M. Petrovich, M. J. Black, and G. Varol, “Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9488–9497.
- [80] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano, “Human motion diffusion model,” in *The Eleventh International Conference on Learning Representations*, 2023.
- [81] X. Chen, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, and G. Yu, “Executing your commands via motion diffusion in latent space,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18000–18010.
- [82] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano, “Learning to summarize with human feedback,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 3008–3021, 2020.
- [83] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, “Training language models to follow instructions with human feedback,” *Advances in neural information processing systems*, 2022.
- [84] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.
- [85] A. Glaese, N. McAleese, M. Trebacz, J. Aslanides, V. Firoiu, T. Ewalds, M. Rauh, L. Weidinger, M. Chadwick, P. Thacker *et al.*, “Improving alignment of dialogue agents via targeted human judgements,” *arXiv preprint arXiv:2209.14375*, 2022.
- [86] C. V. Snell, I. Kostrikov, Y. Su, S. Yang, and S. Levine, “Offline RL for natural language generation with implicit language q learning,” in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=aBH_DydEvoH
- [87] S. Zhang, X. Yang, Y. Feng, C. Qin, C.-C. Chen, N. Yu, Z. Chen, H. Wang, S. Savarese, S. Ermon *et al.*, “Hive: Harnessing human feedback for instructional visual editing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9026–9036.

- [88] K. Lee, H. Liu, M. Ryu, O. Watkins, Y. Du, C. Boutilier, P. Abbeel, M. Ghavamzadeh, and S. S. Gu, “Aligning text-to-image models using human feedback,” *arXiv preprint arXiv:2302.12192*, 2023.
- [89] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [90] S. Kim, S. Bae, J. Shin, S. Kang, D. Kwak, K. Yoo, and M. Seo, “Aligning large language models through synthetic feedback,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 13 677–13 700.
- [91] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine, “Training diffusion models with reinforcement learning,” in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: <https://openreview.net/forum?id=YCWjhGrJFD>
- [92] H. Lee, S. Phatale, H. Mansoor, K. Lu, T. Mesnard, C. Bishop, V. Carbune, and A. Rastogi, “Rlaif: Scaling reinforcement learning from human feedback with ai feedback,” *arXiv preprint arXiv:2309.00267*, 2023.
- [93] M. Petrovich, M. J. Black, and G. Varol, “TEMOS: Generating diverse human motions from textual descriptions,” in *European Conference on Computer Vision (ECCV)*, 2022.
- [94] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, “Improving language understanding by generative pre-training,” 2018.
- [95] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen, “Motiongpt: Human motion as a foreign language,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [96] J. Zhang, Y. Zhang, X. Cun, S. Huang, Y. Zhang, H. Zhao, H. Lu, and X. Shen, “T2m-gpt: Generating human motion from textual descriptions with discrete representations,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [97] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.
- [98] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” *Advances in Neural Information Processing Systems*, 2020.
- [99] M. Zhang, X. Guo, L. Pan, Z. Cai, F. Hong, H. Li, L. Yang, and Z. Liu, “Remodifuse: Retrieval-augmented motion diffusion model,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 364–373.
- [100] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [101] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.
- [102] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” *Advances in neural information processing systems*, vol. 30, 2017.

- [103] A. Brock, J. Donahue, and K. Simonyan, “Large scale GAN training for high fidelity natural image synthesis,” in *International Conference on Learning Representations*, 2019. [Online]. Available: <https://openreview.net/forum?id=B1xsqj09Fm>
- [104] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training gans,” *Advances in neural information processing systems*, vol. 29, 2016.
- [105] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu, “Motiondiffuse: Text-driven human motion generation with diffusion model,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 6, pp. 4115–4128, 2024.
- [106] Q. Yu, M. Tanaka, and K. Fujiwara, “Exploring vision transformers for 3d human motion-language models with motion patches,” 2024.
- [107] K. Fujiwara, M. Tanaka, and Q. Yu, “Chronologically accurate retrieval for temporal grounding of motion-language models,” in *European Conference on Computer Vision*. Springer, 2024, pp. 323–339.
- [108] Z. Wang, Y. Chen, B. Jia, P. Li, J. Zhang, J. Zhang, T. Liu, Y. Zhu, W. Liang, and S. 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*, 2024, pp. 433–444.
- [109] Z. Cen, H. Pi, S. Peng, Z. Shen, M. Yang, S. Zhu, H. Bao, and X. Zhou, “Generating human motion in 3d scenes from text descriptions,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1855–1866.
- [110] T. Papamarkou, T. Birdal, M. M. Bronstein, G. E. Carlsson, J. Curry, Y. Gao, M. Hajij, R. Kwitt, P. Lio, P. Di Lorenzo, V. Maroulas, N. Miolane, F. Nasrin, K. Natesan Ramamurthy, B. Rieck, S. Scardapane, M. T. Schaub, P. Veličković, B. Wang, Y. Wang, G. Wei, and G. Zamzmi, “Position: Topological deep learning is the new frontier for relational learning,” in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Eds., vol. 235. PMLR, 21–27 Jul 2024, pp. 39 529–39 555.
- [111] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or, “Motionclip: Exposing human motion generation to clip space.” Springer, 2022, pp. 358–374.
- [112] L. Bensabath, M. Petrovich, and G. Varol, “Tmr++: A cross-dataset study for text-based 3d human motion retrieval,” 2024.
- [113] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, “Imagebind: One embedding space to bind them all,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 180–15 190.
- [114] L. Xue, M. Gao, C. Xing, R. Martín-Martín, J. Wu, C. Xiong, R. Xu, J. C. Niebles, and S. 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*, 2023, pp. 1179–1189.
- [115] L. Xue, N. Yu, S. Zhang, A. Panagopoulou, J. Li, R. Martín-Martín, J. Wu, C. Xiong, R. Xu, J. C. 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*, 2024, pp. 27 091–27 101.
- [116] W. Lei, Y. Ge, K. Yi, J. Zhang, D. Gao, D. Sun, Y. Ge, Y. Shan, and M. Z. Shou, “Vit-lens: Towards omni-modal representations,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 647–26 657.

- [117] R. Liu, J. Zhang, K. Peng, J. Zheng, K. Cao, Y. Chen, K. Yang, and R. 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*, 2023, pp. 1857–1867.
- [118] S. Mai, Y. Zeng, S. Zheng, and H. Hu, “Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis,” *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2276–2289, 2022.
- [119] Y. Ruan, H.-H. Lee, Y. Zhang, K. Zhang, and A. 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)*, January 2024, pp. 5815–5825.
- [120] K. Yin, S. Zou, Y. Ge, and Z. Tian, “Tri-modal motion retrieval by learning a joint embedding space,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1596–1605.
- [121] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and J. Sivic, “People watching: Human actions as a cue for single view geometry,” in *Computer Vision – ECCV 2012*, ser. Lecture Notes in Computer Science, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., vol. 7576. Springer, Berlin, Heidelberg, 2012, pp. 732–745. [Online]. Available: https://doi.org/10.1007/978-3-642-33715-4_53
- [122] T. Brooks and A. A. Efros, “Hallucinating pose-compatible scenes,” 2022. [Online]. Available: <https://arxiv.org/abs/2112.06909>
- [123] Y. Nie, A. Dai, X. Han, and M. Nießner, “Pose2room: Understanding 3d scenes from human activities,” 2022. [Online]. Available: <https://arxiv.org/abs/2112.03030>
- [124] S. Kulal, T. Brooks, A. Aiken, J. Wu, J. Yang, J. Lu, A. Efros, and K. Kumar Singh, “Putting people in their place: Affordance-aware human insertion into scenes,” in *Computer Vision – CVPR 2023*, 06 2023, pp. 17 089–17 099.
- [125] J. Lee and H. Joo, “Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3d environments,” 2023. [Online]. Available: <https://arxiv.org/abs/2301.02667>
- [126] Y. Gong, A. Rouditchenko, A. H. Liu, D. Harwath, L. Karlinsky, H. Kuehne, and J. Glass, “Contrastive audio-visual masked autoencoder,” *arXiv preprint arXiv:2210.07839*, 2022.
- [127] J. Chen, D. Barath, I. Armeni, M. Pollefeys, and H. Blum, “Where am i? scene retrieval with language,” *European Conference on Computer Vision 2024*, 2024.
- [128] P. Poklukar, M. Vasco, H. Yin, F. S. Melo, A. Paiva, and D. Kragic, “Geometric multimodal contrastive representation learning,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 17 782–17 800.
- [129] G. Delmas, P. Weinzaepfel, F. Moreno-Noguer, and G. Rogez, “Poseembroider: Towards a 3d, visual, semantic-aware human pose representation,” *European Conference on Computer Vision 2024*, 2024.
- [130] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter,” 2020. [Online]. Available: <https://arxiv.org/abs/1910.01108>
- [131] Z. Ziyu, M. Xiaojian, C. Yixin, D. Zhidong, H. Siyuan, and L. Qing, “3d-vista: Pre-trained transformer for 3d vision and text alignment,” in *ICCV*, 2023.

- [132] B. Jia, Y. Chen, H. Yu, Y. Wang, X. Niu, T. Liu, Q. Li, and S. Huang, “Sceneverse: Scaling 3d vision-language learning for grounded scene understanding,” in *European Conference on Computer Vision*. Springer, 2025, pp. 289–310.
- [133] C. Bodnar, F. Frasca, Y. Wang, N. Otter, G. F. Montufar, P. Lió, and M. Bronstein, “Weisfeiler and lehman go topological: Message passing simplicial networks,” in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 1026–1037.
- [134] M. Papillon, S. Sanborn, M. Hajij, and N. Miolane, “Architectures of topological deep learning: A survey of message-passing topological neural networks,” 2024. [Online]. Available: <https://arxiv.org/abs/2304.10031>
- [135] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [136] M. Petrovich, M. J. Black, and G. Varol, “Action-conditioned 3d human motion synthesis with transformer vae,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021, pp. 10 965–10 975.
- [137] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” *Advances in neural information processing systems*, vol. 30, 2017.
- [138] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024.
- [139] C. Guo, Y. Mu, M. G. Javed, S. Wang, and L. Cheng, “Momask: Generative masked modeling of 3d human motions,” 2023. [Online]. Available: <https://arxiv.org/abs/2312.00063>
- [140] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” 2019.