HuTuMotion: Human-Tuned Navigation of Latent Motion Diffusion Models with Minimal Feedback

Gaoge Han[1], Shaoli Huang[2*], Mingming Gong[3,4], Jinglei Tang [1*]
[1]College of Information Engineering, Northwest A&F University
[2]Tencent AI Lab
[3]School of Mathematics and Statistics, The University of Melbourne
[4]Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
hangaoge@nwafu.edu.cn, sshaol.huang@gmail.com, mingming.gong@unimelb.edu.au, tangjinglei@nwsuaf.edu.cn

Abstract

We introduce HuTuMotion, an innovative approach for generating natural human motions that navigates latent motion diffusion models by leveraging few-shot human feedback. Unlike existing approaches that sample latent variables from a standard normal prior distribution, our method adapts the prior distribution to better suit the characteristics of the data, as indicated by human feedback, thus enhancing the quality of motion generation. Furthermore, our findings reveal that utilizing few-shot feedback can yield performance levels on par with those attained through extensive human feedback. This discovery emphasizes the potential and efficiency of incorporating few-shot human-guided optimization within latent diffusion models for personalized and style-aware human motion generation applications. The experimental results show the significantly superior performance of our method over existing state-of-the-art approaches.

Introduction

Human motion generation, a rapidly growing area of research (Li et al. 2022; Raab et al. 2022; Guo et al. 2020; Ling et al. 2020; Petrovich, Black, and Varol 2021; Guo et al. 2022a,c,b; Zhang et al. 2022a; Tevet et al. 2022) in computer vision and artificial intelligence, has gained significant attention due to its wide-ranging applications in animation, gaming, and robotics. Recent techniques typically encompass the process of sampling latent variables z from a standard normal prior distribution $p(z)$, followed by generating data x using the generative probability distribution $p(x|z)$. These approaches, often based on deep generative models like Variational Autoencoders (VAEs) (Chen et al. 2023) or Generative Adversarial Networks (GANs) (Lee et al. 2019), have made significant strides in motion synthesis. However, they often fall short of capturing the real data characteristics and generating human motions that accurately reflect the input semantics. For instance, consider generating a motion from a textual input like "an old person walking at an average pace forward." Existing methods (Chen et al. 2023; Tevet et al. 2022; Guo et al. 2022a) may generate a generic "walking at an average pace forward" motion but fail to capture the specific nuance of "old". This limitation has hindered the advancement of more sophisticated applications, such as personalized and editable motion generation, which call for a deeper grasp of the underlying data distribution and effective incorporation of human feedback.

In light of these challenges, we propose HuTuMotion, a novel approach that seeks to improve the quality of motion generation by leveraging latent diffusion models and incorporating few-shot human feedback. The central idea of HuTuMotion is to adjust the prior distribution $p(z)$ based on human feedback rather than existing approaches of drawing from a standard normal prior distribution. This adjustment allows the latent space to capture the characteristics of the data better, thereby improving the quality and realism of the generated motions. Our method does not solely rely on an arbitrary sampling of the latent space. Instead, we carefully optimize the selection of regions in the latent space that yield more realistic and semantically meaningful human motions. To achieve this, we first strategically identify representative and diverse motion descriptions. We then utilize a unique feedback mechanism that incorporates a few-shot learning approach. In this paradigm, minimal yet effective human feedback guides the optimization process, refining the link between the descriptions and their corresponding latent distributions. Furthermore, to ensure semantic alignment between the input text and output motion, we employ a text similarity measure. During testing, this measure assesses the similarity between the input text and the representative motion descriptions. The most similar representative prior distribution is then used to sample a latent, which ultimately generates the corresponding human motion.

In addition to enhancing the general text-driven motion generation, HuTuMotion also introduces a new capability to support personalized and style-aware motion generation. This functionality allows the users to provide their specific motion style preferences, which are then incorporated into the motion generation process. Through this mechanism, our method can generate unique, individualized motions that better reflect the users' intentions and preferences, thereby opening up new possibilities for applications in areas such as interactive gaming and personalized animation. In our quantitative experiments on both the HumanML3D and KIT datasets, HuTuMotion significantly outperforms exist-

*Corresponding author: Shaoli Huang, Jinglei Tang.

ing state-of-the-art methods. Additionally, through qualitative experiments, we observe that our method generates more natural and semantically correct motions. Importantly, when comparing the effectiveness of few-shot and extensive human feedback within our method, our results show that using few-shot human feedback achieves comparable performance to extensive feedback. This underscores the efficiency and potential of our few-shot human-guided optimization approach in the field of human motion generation applications.

The key contributions of our work are:

- We introduce HuTuMotion, a novel approach to improve the quality of motion generation using latent diffusion models and few-shot human feedback. To the best of our knowledge, this work is the first attempt to leverage few-shot human feedback to enhance motion generation quality.
- HuTuMotion uniquely adjusts the prior distribution based on human feedback, optimizing the selection of regions in the latent space to yield more realistic and semantically meaningful human motions.
- We propose a unique feedback mechanism that incorporates a few-shot learning approach and a text similarity measure to refine the link between motion descriptions and their corresponding latent distributions, enhancing semantic alignment between the text and motion.
- HuTuMotion also supports personalized and style-aware motion generation, enabling users to provide specific motion style preferences that are incorporated into the motion generation process.

Related Work

Human Motion Diffusion Model. The impressive performance of diffusion models on text-to-image tasks (Ho and Salimans 2021; Rombach et al. 2022a,b) has recently inspired the creation of diffusion-based human motion models (Tevet et al. 2022; Zhang et al. 2022a; Chen et al. 2023; Zhang et al. 2023a) that are trained on the motion capture datasets using the human motion estimation methods (Kanazawa et al. 2018; Yu et al. 2023; Cheng et al. 2023). MotionDiffuse (Zhang et al. 2022a) pioneered this field as the first text-based motion diffusion model with fine-grained instructions on body parts. MDM (Tevet et al. 2022) and MLD (Chen et al. 2023) followed suit, with the former proposing a motion diffusion model on raw motion data to understand the relationship between motion and input conditions, and the latter performing the motion diffusion process in the latent space to significantly reduce computational overhead during training and inference stages. Unlike these models, our approach guides initial latent generation on a latent-based diffusion model, achieving state-of-the-art performance in text-to-motion tasks with minimal cost.

Reinforcement Learning with Human Feedback. Reinforcement Learning with Human Feedback (RLHF) is an expanding field demonstrating substantial potential in aligning human references and model performance across language tasks (Stiennon et al. 2020; Ouyang et al. 2022; OpenAI 2022). It typically involves using human-gathered ranking data to train a reward model, which is then used to fine-tune a Supervised Fine-Tuning (SFT) model via policy gradients. Recent applications of RLHF in the text-to-image field (Zhang et al. 2023b; Lee et al. 2023; Tang, Rybin, and Chang 2023; Xu et al. 2023; Wu et al. 2023) have shown promising text-image alignment performance. However, unlike previous works, our method uses a novel online human feedback approach, avoiding the need for large-scale human-ranked data to optimize the quality of generated human motion.

Few-Shot Learning and Generation. Most few-shot learning methods fall into three categories: meta-learning (Finn, Abbeel, and Levine 2017; Oreshkin, López, and Lacoste 2018; Vinyals et al. 2016), transfer-learning (Yang, Wang, and Zhu 2022; Zhang et al. 2022b; Hu et al. 2022), and feature augmentations (Lazarou, Stathaki, and Avrithis 2022; Chen et al. 2018; Ye et al. 2020). These methods use textual descriptions of novel classes to generate and align images, promoting the effective use of synthetic images in training few-shot learners. Recently, few-shot generation has been employed in the text-to-image task using diffusion models (Gal et al. 2023; Ruiz et al. 2023; Samuel et al. 2023). Specifically, (Gal et al. 2023; Ruiz et al. 2023) learn to map a set of images to a corresponding "word" in the low-dimensional embedding space using a pre-trained model. (Samuel et al. 2023) addresses long-tail learning in the presence of highly unbalanced training data by selecting optimal generation seeds from the noise space.

In our approach, we generate natural human motions by incorporating few-shot feedback and text similarity, requiring only a few-shot human feedback for searching the prior distribution of latent, thus achieving a network-free method during the inference period.

Method

As depicted in Figure 1, our method, HuTuMotion, is characterized by two principal steps: 1) Representative Distribution Optimization and 2) Semantic Alignment and Motion Generation. We begin with a concise introduction to the Latent Diffusion Model, followed by an in-depth explanation of these pivotal steps.

Overview of Latent Diffusion Models

Latent diffusion models (LDMs) has achieved great success in text-to-motion task, such as Stable Diffusion (Rombach et al. 2022b). Different from the diffusion model, LDMs perform diffusion process in latent space using an extra denoising U-Net. Low-dimensional space is better suited for likelihood-based generative models, as it allows them to concentrate on the crucial semantic representation of the data and trains in a lower-dimensional space that is computationally more efficient. MLD (Chen et al. 2023) is the first latent-based motion diffusion model for text-to-motion synthesis. MLD design a transformer-based conditional denoiser ϵ_θ. Its conditional objective can be expressed as:

$$\mathcal{L}_{\mathbf{MLD}} := \mathbb{E}_{\epsilon,t,c}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, t, \tau_\theta(c)\right)\right\|_2^2\right], \qquad (1)$$

where $\epsilon \sim \mathcal{N}(0, I)$, z_t is the latent in time step t and $\tau_\theta(c)$ denotes the CLIP (Radford et al. 2021) text encoder

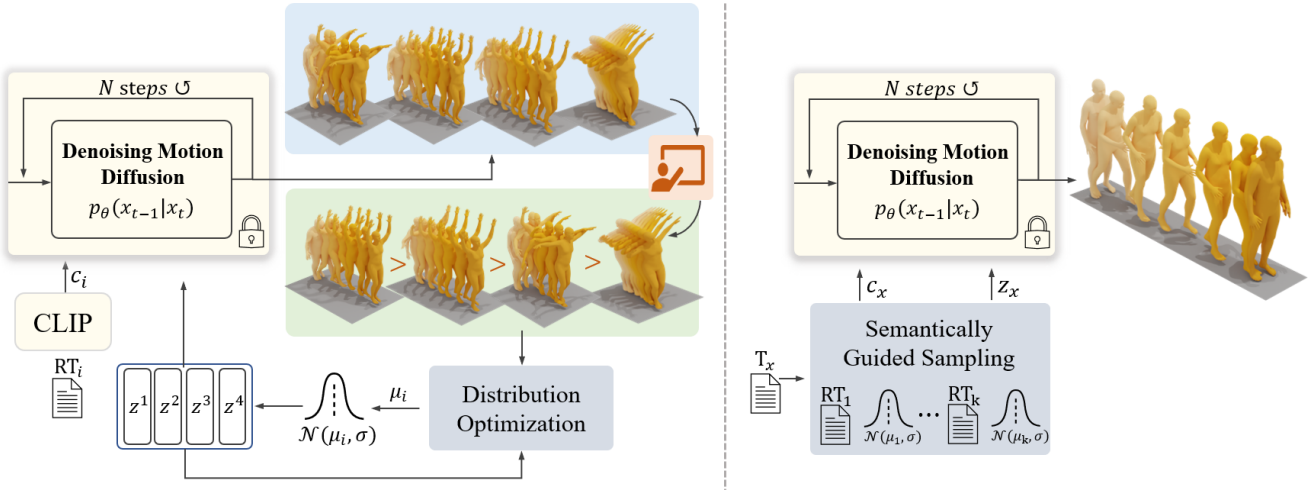

Figure 1: An overview of our framework. For simplicity, we omit the motion decoder. σ denotes the standard deviation. It consists of two stages. In representative distribution optimization, we obtain the optimized latent distribution corresponding to representative texts. In a semantically guided generation, we select one latent distribution by computing text similarity.

with embedded conditions c. MLD uses the DDIM (Song, Meng, and Ermon 2020) as the sampler. It is worth noting that the denoising diffusion process becomes deterministic and solely on the latent embedding for ODE-based diffusion samplers such as DDIM (Song, Meng, and Ermon 2020), DPM-solver (Lu et al. 2022a) and DPM-solver++ (Lu et al. 2022b). We optimize the latent distribution based on MLD.

Representative Distribution Optimization

Our methodology starts by identifying a set of representative and diverse motion descriptions. We then incorporate few-shot human feedback to adjust the prior distribution through a two-stage optimization process.

Representative Motion Descriptions. Representative motion descriptions can be defined as descriptors that capture the wide range and diversity of human motions. A straightforward approach for selecting these descriptions involves employing a K-means clustering algorithm on the training dataset. In this process, the number of clusters denoted as K, is predetermined based on the diversity present in the data. Each cluster's centroid is then considered a representative motion description. However, our experimental observations indicate that a representative motion description does not necessarily need to come from a specific training dataset. Asking a large language model (ChatGPT (OpenAI 2022)) to generate representative texts can also yield comparable results. The detailed process can be found in the supplementary material [‡].

Latent Optimization via Human Feedback. The central innovation of HuTuMotion lies in the optimization of the latent variables from the prior distribution. Unlike conventional methods that draw samples from a standard normal prior distribution, we adjust the latent input from the prior distribution based on few-shot human feedback in the form of rankings.

Our goal is to ascertain an optimal value of z that yields the minimum score for the function $f(z, c, t)$, wherein c represents the text embedding and t signifies the diffusion steps. Given that c and t remain constant throughout the optimization process, we will omit them in the ensuing discussions. Here, the score of $f(z)$ should reflect the quality of the generated motion from the input z according to human judgment. The lower the score, the better the generated motion. The specific form of $f(z)$ is not predefined and is implicitly determined by human feedback. To formalize this, we can express it as an optimization problem: $\min_{z \in \mathbb{R}^d} f(z)$.

Given that $f(z)$ functions as a black box, we do not have direct access to its internal workings. Instead, we can interface with it through a ranking oracle. This oracle provides insight into the function by sorting the scores of generated outcomes, which can be thought of as motions in a particular context.

Due to the constraints imposed by the ranking oracle, the optimization problem at hand can be recast as an (m, k)-ranking oracle optimization problem. This type of problem involves determining an optimal ranking of a set of items constrained by the limited information provided by the oracle. Therefore, the challenge lies not only in optimizing $f(z)$ but also in intelligently querying the oracle to explore the search space efficiently.

To solve this problem, we adopt the zeroth-order optimization algorithm (Tang, Rybin, and Chang 2023) that obtains the descent direction using a rank-based random estimator. The estimator converts the (m, k)-ranking oracles' input and output into a directed acyclic graph (DAG), $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N} = \{1, ..., m\}$ and $\mathcal{E} =$

[‡]The supplementary material is included in the arxiv version https://arxiv.org/abs/2312.12227

$(i,j) \mid f(z_i) < f(z_j)$. With access to an (m,k)-ranking oracle $O(m,k)$ and a starting point z, we query $O(m,k)$ using $x_i = x + \mu\xi_i$, where $\xi_i \sim \mathcal{N}(0,I)$ for $i = 1, ..., m$. The rank-based gradient estimator, constructed using the ranking information from $O(m,k)$, is:

$$\tilde{g}(z) = \frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} \frac{\xi_j - \xi_i}{\mu}. \quad (2)$$

Upon the completion of the initial optimization process, we continue to refine the optimal latent variable z. The comprehensive steps for this procedure are outlined in Algorithm 1.

Optimal Representative Prior Distribution. We have noted a significant correlation in our observations: texts that are closely aligned in the embedding space also demonstrate similar proximity in their optimal latents. Detailed results supporting this observation can be found in the accompanying supplementary material [‡]. Based on this, we propose a method to construct an optimal prior distribution for texts that are near a given representative text (RT). We use the optimal latent of a representative text as the mean and specify a relatively small standard deviation to construct a Gaussian distribution. This Gaussian distribution then serves as the optimal prior distribution for texts that are close to the representative text in the embedding space.

Semantic Alignment and Motion Generation

By employing representative distribution optimization coupled with human ranking information, we acquire pair sets consisting of representative texts and their corresponding latent distributions. Given an input text, we initially calculate its cosine similarity with the representative texts. Subsequently, we select the latent distribution that corresponds to the representative text closest in similarity. This approach serves as an efficient strategy to enhance overall motion quality, as it does not necessitate significant resources to fine-tune the model.

Semantically Guided Sampling. Contrary to the MLD approach which samples latent from $\mathcal{N}(0,I)$, we instead sample latent from $\mathcal{N}(z_m^{**}, \sigma)$, where z_m^{**} is the optimal latent determined for the representative text RT_m. When given an input text, we compute the index m by measuring the cosine similarity between c_x and $\{c_1, ..., c_k\}$ corresponding to the input text T_x and the set of representative texts $\{RT_1, ..., RT_k\}$, respectively. In this context, c_x and $\{c_1, ..., c_k\}$ denote the text embeddings obtained from the CLIP text encoder. The maximum index m is then computed using the following equation:

$$m = \arg\max_i \frac{c_x \cdot c_i}{||c_x|| \cdot ||c_i||}. \quad \text{for } i \in 1, ..., k \quad (3)$$

This allows us to select the representative text that aligns most closely with the given input text, ensuring a more effective and relevant sampling of the latent variable.

Inference (Motion Generation). By sampling from the optimal distribution $\mathcal{N}(z_m^{**}, \sigma^2)$, we obtain z_x. Subsequently, we input z_x and c_x into the DDIM sampler (Song, Meng, and Ermon 2020) to facilitate the denoising motion diffusion process. It's important to note that the standard deviation σ acts as our hyper-parameter. We further examine and discuss its impact in the Ablation Studies section.

Algorithm 1: Distribution optimization for representative texts

Require: Objective function f (Evaluated by human), number of queries m, stepsize η, smoothing parameter μ_1, μ_2, μ_3, shrinking rate $\gamma \in (0,1)$.

Initialize the reference point z^* with all-zero vectors.
Initialize the gradient memory $\bar{g}$ with all-zero vectors.
Set $\tau = 0$.
Choose one sample from the representative texts provided by ChatGPT.
Sample m i.i.d. starting point input $\mathcal{X}_1 = \{\xi_1, \cdots, \xi_m\}$ from $\mathcal{N}(0, \mu_1 I)$.
while not select the best motion by human **do**
  Query $O_f^{(m,k)}$ with input $\mathcal{X}_1$ for $2 \le k \le m$. Denote $\mathbb{I}_1$ as the output.
  Set z^* to be the weighted $\mathcal{X}_1$ using the ranking information $\mathbb{I}_1$.
  Compute the gradient $\tilde{g}$ using the ranking information $\mathbb{I}_1$ according to the equation 2.
  $\bar{g} = (\tau\bar{g} + \tilde{g})/(\tau + 1)$
  $\tau = \tau + 1$
  Sample m i.i.d. direction $\{\psi_1, \cdots, \psi_m\}$ from $\mathcal{N}(0, \mu_2 I)$.
  $\mathcal{X}_1 = \{z^* - \eta\bar{g} + \psi_1, z^* - \eta\gamma\bar{g} + \psi_2, ..., z^* - \eta\gamma^{m-1}\bar{g} + \psi_m\}$
end while
Set z^{**} to be the best point in $\mathcal{X}_1$ with minimal objective value using the ranking information $\mathbb{I}_1$.
while not exit by human **do**
  Sample m i.i.d. direction $\{\psi_1, \cdots, \psi_m\}$ from $\mathcal{N}(0, \mu_3 I)$.
  Query $O_f^{(m,1)}$ with input $\mathcal{X}_2 = \{z^{**} + \psi_1, z^{**} + \psi_2, ..., z^{**} + \psi_m\}$. Denote $\mathbb{I}_2$ as the output.
  Set z^{**} to be the best point in $\mathcal{X}_2$ with minimal objective value using the ranking information $\mathbb{I}_2$.
end while

Expanding the Scope: Personalized and Style-Aware Generation

Our method is primarily designed to enhance the generation quality of general human motions. However, its versatility allows for straightforward extensions to accommodate new tasks. These include but are not limited to, personalized and style-aware motion generation, further demonstrating the adaptability and potential of our approach.

Personalized Motion Generation. Personalized Motion Generation is a task in human motion generation that focuses on generalizing motions that align with user preferences. In this context, suppose we have an output set $\{(RT_1, z_1^{**}), ..., (RT_k, z_k^{**})\}$ derived from our method and a user-provided text set $T_U^1, ..., T_U^l$ that reflects a specific preference for a desired motion style. To implement Person-

Methods	R Precision (top3) ↑	FID↓	MM Dist↓	Diversity↑	MModality
Real	$0.797\pm.002$	$0.002\pm.000$	$2.974\pm.008$	$9.503\pm.065$	-
Seq2Seq (Plappert, Mandery, and Asfour 2018)	$0.396\pm.002$	$11.75\pm.035$	$5.529\pm.007$	$6.223\pm.061$	-
LJ2P (Ahuja and Morency 2019)	$0.486\pm.002$	$11.02\pm.046$	$5.296\pm.008$	$7.676\pm.058$	-
T2G(Bhattacharya et al. 2021)	$0.345\pm.002$	$7.664\pm.030$	$6.030\pm.008$	$6.409\pm.071$	-
Hier (Ghosh et al. 2021)	$0.552\pm.004$	$6.532\pm.024$	$5.012\pm.018$	$8.332\pm.042$	-
TEMOS (Petrovich, Black, and Varol 2022)	$0.722\pm.002$	$3.734\pm.028$	$3.703\pm.008$	$8.973\pm.071$	$0.368\pm.018$
T2M (Guo et al. 2022a)	$0.740\pm.003$	$1.067\pm.002$	$3.340\pm.008$	$9.188\pm.002$	$2.090\pm.083$
MDM (Tevet et al. 2022)	$0.611\pm.007$	$0.544\pm.044$	$5.566\pm.027$	$9.559\pm.086$	$2.799\pm.072$
MLD (Chen et al. 2023)	$0.772\pm.002$	$0.473\pm.013$	$3.196\pm.010$	$9.724\pm.082$	$2.413\pm.079$
Ours*	$\underline{0.782}\pm.002$	$\mathbf{0.224}\pm.006$	$\mathbf{3.058}\pm.009$	$\underline{9.745}\pm.073$	$0.966\pm.046$
Ours	$\mathbf{0.785}\pm.002$	$\underline{0.295}\pm.006$	$\underline{3.093}\pm.007$	$\mathbf{9.828}\pm.091$	$1.019\pm.054$

Table 1: Comparison of text-to-motion synthesis on HumanML3D (Guo et al. 2022b) dataset. $*$ means using the texts of cluster's centroid of K-means. These metrics are evaluated by the motion encoder from (Guo et al. 2022a). For each metric, we repeat the evaluation 20 times and report the average with a 95% confidence interval. We employ real motion as a reference and sort all approaches by descending FIDs. Bold and underline indicate the best and the second best result. The complete R Precision metric is in the supplementary material‡.

Methods	R Precision (top3) ↑	FID↓	MM Dist↓	Diversity↑	MModality
Real	$0.779\pm.006$	$0.031\pm.004$	$2.788\pm.012$	$11.08\pm.097$	-
Seq2Seq(Plappert, Mandery, and Asfour 2018)	$0.241\pm.006$	$24.86\pm.348$	$7.960\pm.031$	$6.744\pm.106$	-
T2G(Bhattacharya et al. 2021)	$0.338\pm.005$	$12.12\pm.183$	$6.964\pm.029$	$9.334\pm.079$	-
LJ2P (Ahuja and Morency 2019)	$0.483\pm.005$	$6.545\pm.072$	$5.147\pm.030$	$9.073\pm.100$	-
Hier (Ghosh et al. 2021)	$0.531\pm.007$	$5.203\pm.107$	$4.986\pm.027$	$9.563\pm.072$	$2.090\pm.083$
TEMOS (Petrovich, Black, and Varol 2022)	$0.687\pm.005$	$3.717\pm.051$	$3.417\pm.019$	$10.84\pm.100$	$0.532\pm.034$
T2M (Guo et al. 2022a)	$0.693\pm.007$	$2.770\pm.109$	$3.401\pm.008$	$10.91\pm.119$	$1.482\pm.065$
MDM (Tevet et al. 2022)	$0.396\pm.004$	$0.497\pm.021$	$9.191\pm.022$	$10.85\pm.109$	$1.907\pm.214$
MLD (Chen et al. 2023)	$0.734\pm.007$	$0.404\pm.027$	$3.204\pm.027$	$10.80\pm.117$	$2.192\pm.071$
Ours*	$\underline{0.766}\pm.005$	$\mathbf{0.201}\pm.064$	$\mathbf{3.082}\pm.025$	$\underline{10.88}\pm.086$	$0.901\pm.035$
Ours	$\mathbf{0.768}\pm.006$	$\underline{0.224}\pm.045$	$\underline{3.098}\pm.025$	$\mathbf{10.96}\pm.090$	$0.914\pm.039$

Table 2: Comparison of text-to-motion synthesis on KIT (Plappert, Mandery, and Asfour 2016) dataset. $*$ means using the texts of cluster's centroid of K-means. Reported metrics are the same as Table 1. Bold and underline indicate the best and the second best result. The complete R Precision metric is in the supplementary material‡.

alized Motion Generation, we begin by identifying the closest RT_i for each text T_U^j and use z_i^{**} as the starting point in Algorithm 1 to conduct feedback optimization. The final optimized result z_i^{**U} will then replace the original z_i^{**}. By processing all the user sets in this manner, we generate the final output set $\{(RT_1, z_1^{**U}), ..., (RT_k, z_k^{**U})\}$ which can be used to generate motions that align with the user's preferences.

Style-Aware Motion Generation. Style-aware Motion Generation pertains to the task of producing motions that embody a specified style, given input text containing style descriptions. Different from personalization, the stylized prompt presents a long-tail distribution in the HumanML3D and KIT datasets. We observe that using the stylized prompt as input, the MLD model fails to generate motion consistent with the stylistic semantics, as shown in Fig. 3. This differs from Personalized Motion Generation in that it need to feedback the text with style Words. Initially, it identifies diverse texts corresponding to the same style descriptor ST_i. Following this, it carries out optimization from scratch to secure the optimal latent. In this context, we can amass a variety of styles and establish a comprehensive set $\{(ST_1, z_1^{**}), ..., (ST_M, z_M^{**})\}$ by employing the previously mentioned optimization process. This approach allows us to generate a broad spectrum of stylized motions.

Experiments

In this section, we provide extensive experimental results. Firstly, we introduce the datasets, implementation details, and evaluation metrics. Secondly, we show the qualitative and quantitative results compared with the state-of-the-art approaches. Finally, we perform ablation studies. More qualitative results are provided in the supplementary material ‡.

Datasets and Evaluation Metrics

We experiment with two text-to-motion synthesis datasets: HumanML3D (Guo et al. 2022b) and KIT (Plappert, Mandery, and Asfour 2016).

KIT comprises 3,911 motion sequences and 6,278 text annotations, with each motion linked to one to four descriptions. The dataset downsampled to 12.5 FPS, is partitioned into 80% training, 5% validation, and 15% test sets.

HumanML3D is the largest 3D human motion-language dataset, containing 14,616 human motions and 44,970 text descriptions. Each motion pairs with at least three descriptions. Motions, re-scaled to 20 FPS and spanning between 2 and 10 seconds, are divided similarly to KIT.

We evaluate text-to-motion models based on metrics as described in (Guo et al. 2022b): **R Precision**: This measures the accuracy of the top-1, top-2, and top-3 ranked Euclidean distances between one motion sequence and 32 text descriptions. **Frechet Inception Distance (FID):** This assesses the feature distribution distance between generated and real motions using a feature extractor. **Multimodal Distance (MM-Dist):** This metric calculates the average Euclidean distance between the generated motion feature and each text feature. **Diversity:** We measure the diversity within a motion set by calculating the average Euclidean distance between features of randomly selected motion pairs. **Multimodality (MModality):** This evaluates the diversity of motion generated for one text description by averaging the Euclidean distances between features of generated motion pairs.

Implementation Details

We implement our method using the state-of-the-art latent human diffusion model, MLD (Chen et al. 2023). Our representative distribution optimization and semantically guided generation are conducted on a single NVIDIA GeForce RTX 2080 Ti GPU, with text embedding and latent dimensions set to 768 and 256, respectively. We set σ to 0.2 for latent sampling and use DDIM (Song, Meng, and Ermon 2020) as the denoising motion diffusion sampler. All other settings are consistent with MLD (Chen et al. 2023). We obtain five representative textual descriptions using ChatGPT (OpenAI 2022) (refer to supplementary material ‡) and optimize the representative distribution with human feedback according to Algorithm 1. The experimental details considering Algorithm 1 are provided in the supplementary material ‡.

Comparison with State-of-the-art Methods

In this section, we present the quantitative and qualitative results compare to existing state-of-the-art methods (Plappert, Mandery, and Asfour 2018; Ahuja and Morency 2019; Bhattacharya et al. 2021; Ghosh et al. 2021; Petrovich, Black, and Varol 2022; Guo et al. 2022a; Tevet et al. 2022; Chen et al. 2023) on the test set of HumanML3D (Guo et al. 2022b) and KIT (Plappert, Mandery, and Asfour 2016). Our method is implemented based on MLD (Chen et al. 2023).

Quantitative Results Comparison. The comparison results presented in Table 2 and Table 1 on the HumanML3D (Guo et al. 2022b) and KIT (Plappert, Mandery, and Asfour 2016) test sets illustrate the superior performance of our approach. It significantly outperforms other state-of-the-art methods, achieving the best scores in R Precision, FID, MM Dist, and Diversity metrics. This performance consistency across both datasets underlines the robustness of our proposed method. Although we didn't observe a consistent improvement in MModality, a higher MModality doesn't necessarily denote superior algorithm performance, as it can lead to semantically incorrect motions. Also, our MModality metric is significantly influenced by the hyperparameter σ, which we explore further in Ablation Studies section.

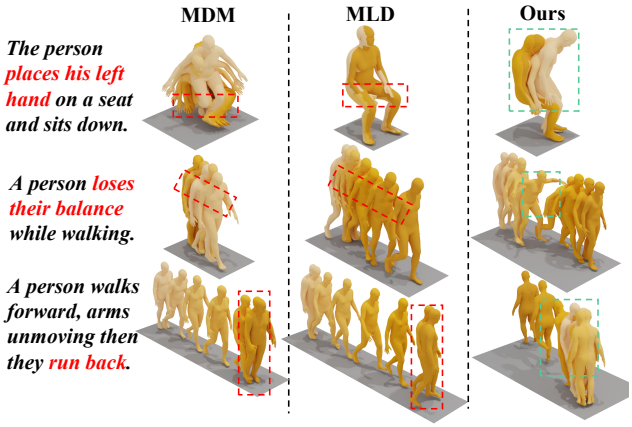

Figure 2: Qualitative results on HumanML3D (Guo et al. 2022b) dataset. The darker colors indicate the later frame in time.

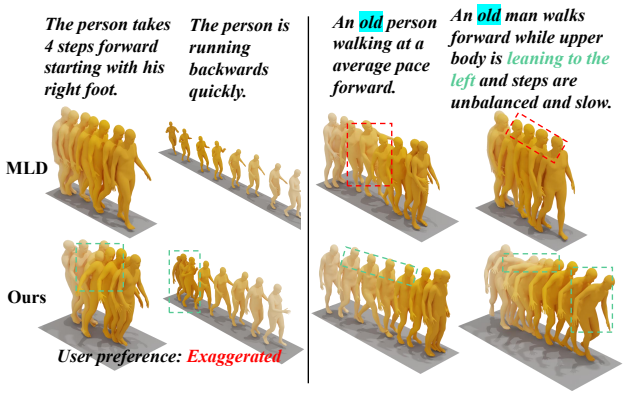

Figure 3: Personalized (left) and style-aware (right) motion generation. The darker colors indicate the later frame in time.

Qualitative Results Comparison. Figure 2 displays qualitative results on the HumanML3D dataset (Guo et al. 2022b). The examples clearly demonstrate that our method generates more natural and semantically accurate motions compared to the lower-quality motions produced by MLD and MDM. For instance, given the text "a person loses their balance while walking." MLD and MDM fail to capture the "losing balance" action. Similarly, with the input, "the person places his left hand on a seat and sits down." MLD generates a static motion while MDM exhibits distorted movements. And for the phrase "a person walks forward, arms unmoving, then they run back." MLD and MDM cannot produce the "run back" motion sequence.

Methods	R Precision ↑			FID↓	MM Dist↓	Diversity↑	MModality
	Top 1	Top 2	Top 3				
Real	0.511±.003	0.703±.003	0.797±.002	0.002±.000	2.974±.008	9.503±.065	-
Ours ($\sigma = 0.1$)	<u>0.496</u>±.002	0.687±.002	<u>0.784</u>±.001	**0.258**±.005	**3.091**±.006	9.780±.089	0.637±.038
Ours ($\sigma = 0.2$)	**0.497**±.002	**0.689**±.002	**0.785**±.002	<u>0.295</u>±.006	<u>3.093</u>±.007	9.828±.091	1.019±.054
Ours ($\sigma = 0.3$)	0.494±.001	<u>0.688</u>±.001	**0.785**±.002	0.356±.007	3.103±.006	<u>9.867</u>±.092	1.276±.062
Ours ($\sigma = 0.4$)	0.491±.002	0.685±.001	0.781±.002	0.437±.006	3.130±.007	**9.882**±.087	1.480±.070
Ours ($\sigma = 0.5$)	0.481±.002	0.674±.001	0.772±.001	0.554±.010	3.190±.009	9.822±.085	1.680±.076

Table 3: Effect of σ on HumanML3D (Guo et al. 2022b) dataset. Reported metrics are the same as Table 1. Bold and underline indicate the best and the second best result.

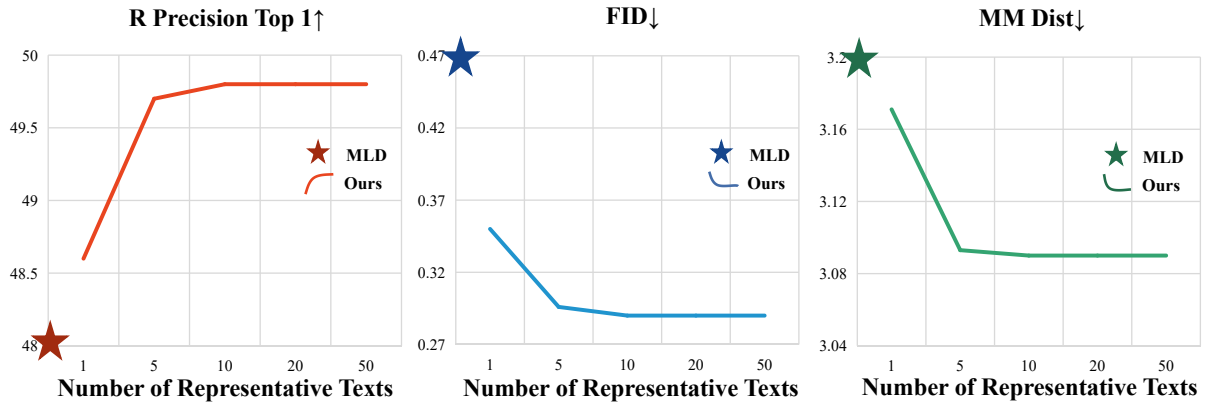

Figure 4: Effect of varying the number of representative texts on R Precision Top 1, FID and MM Dist.

Personalized and Style-aware Generation. Figure 3 showcases the personalized and style-aware motion generation capabilities of our method. In personalization, given the same text input, MLD produces standard motions, whereas our method can generate exaggerated actions in line with human preferences. Regarding text-guided stylization, MLD fails to generate style-aware motion. For instance, given the input "an old man walks forward while the upper body is leaning to the left and steps are unbalanced and slow", MLD fails to generate a sequence exhibiting the desired "leaning to the left" and "old" characteristics. In contrast, our method can generate semantically consistent motions while also reflecting an elderly style.

Ablation Studies

In this section, we first examine the influence of the number of human feedback samples. Following that, we investigate the optimal hyper-parameters for σ.

Effect of The Number of Representative Texts. Fig. 4 examines the impact of varying the number of representative texts on semantically guided generation using the HumanML3D dataset. The metrics of R Precision Top 1, FID, and MM Dist are reported for a range of 1 to 50 representative texts. We found that increasing the text count does not guarantee performance enhancement. To balance performance and distribution optimization, we opted for 5 representative texts as our default implementation setting.

Effect of The Standard Deviation σ. Table 3 reveals the effects of varying σ between 0.1 and 0.5. Our findings showed optimal R Precision Top 1 at $\sigma = 0.2$, best FID at $\sigma = 0.1$, and peak diversity at $\sigma = 0.4$. While increasing σ did enhance MModality, it also resulted in some semantically incorrect motions at higher values. Balancing all factors, we opted for $\sigma = 0.2$ in our experimental setup.

Conclusion and Limitation

In conclusion, our research advances human motion generation with HuTuMotion, an innovative method that utilizes latent diffusion models and few-shot human feedback. Unlike traditional methods, HuTuMotion uniquely adjusts the prior distribution based on human feedback, enhancing data realism. We have also established a feedback mechanism that ensures semantic alignment between motion descriptions and corresponding latent distributions. Importantly, HuTuMotion accommodates personalized and style-aware motion generation, broadening its potential applications. Quantitative and qualitative experiments affirm HuTuMotion's superiority over existing methods. Due to the higher dimensionality of an explicit diffusion model (e.g., MDM) is hard to optimize, the proposed method is limited to latent motion diffusion (i.e. MLD). In additiion, when processing long prompts with numerous action descriptions, MLD tends to miss some actions. Our method to improve this issue is relatively limited.

Acknowledgments

This work was supported by Key Research and Development projects in Shaanxi Province (No. 2023-YBNY-121), Xi'an Science and Technology Plan Project (No. 22NYYF013), Xianyang Key Project of Research and Development Plan (No. L2022ZDYFSF050), and the National Key Research and Development Program of China (No. 2021YFD1600704).

References

Ahuja, C.; and Morency, L.-P. 2019. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, 719–728. IEEE.

Bhattacharya, U.; Rewkowski, N.; Banerjee, A.; Guhan, P.; Bera, A.; and Manocha, D. 2021. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, 1–10. IEEE.

Chen, X.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; and Yu, G. 2023. Executing Your Commands via Motion Diffusion in Latent Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18000–18010.

Chen, Z.; Fu, Y.; Zhang, Y.; Jiang, Y.-G.; Xue, X.; and Sigal, L. 2018. Multi-Level Semantic Feature Augmentation for One-Shot Learning. *IEEE Transactions on Image Processing*, 28: 4594–4605.

Cheng, Y.; Huang, S.; Ning, J.; and Shan, Y. 2023. BoPR: Body-aware Part Regressor for Human Shape and Pose Estimation. *arXiv preprint arXiv:2303.11675*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *ICML*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023. An image is worth one word: Personalizing text-to-image generation using textual inversion. *ICLR*.

Ghosh, A.; Cheema, N.; Oguz, C.; Theobalt, C.; and Slusallek, P. 2021. Synthesis of compositional animations from textual descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1396–1406.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022a. Generating Diverse and Natural 3D Human Motions From Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5152–5161.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022b. Generating Diverse and Natural 3D Human Motions From Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5152–5161.

Guo, C.; Zuo, X.; Wang, S.; and Cheng, L. 2022c. TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts. In *ECCV*.

Guo, C.; Zuo, X.; Wang, S.; Zou, S.; Sun, Q.; Deng, A.; Gong, M.; and Cheng, L. 2020. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2021–2029.

Ho, J.; and Salimans, T. 2021. Classifier-free diffusion guidance. *NeurIPS workshop on Deep Generative Models and Downstream Applications*.

Hu, S. X.; Li, D.; Stühmer, J.; Kim, M.; and Hospedales, T. M. 2022. Pushing the Limits of Simple Pipelines for Few-Shot Learning: External Data and Fine-Tuning Make a Difference. In *CVPR*.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7122–7131.

Lazarou, M.; Stathaki, T.; and Avrithis, Y. 2022. Tensor feature hallucination for few-shot learning. In *WACV*.

Lee, H.-Y.; Yang, X.; Liu, M.-Y.; Wang, T.-C.; Lu, Y.-D.; Yang, M.-H.; and Kautz, J. 2019. Dancing to music. *Advances in neural information processing systems*, 32.

Lee, K.; Liu, H.; Ryu, M.; Watkins, O.; Du, Y.; Boutilier, C.; Abbeel, P.; Ghavamzadeh, M.; and Gu, S. S. 2023. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*.

Li, B.; Zhao, Y.; Zhelun, S.; and Sheng, L. 2022. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1272–1279.

Ling, H. Y.; Zinno, F.; Cheng, G.; and van de Panne, M. 2020. Character Controllers Using Motion VAEs. *ACM Trans. Graph.*, 39(4).

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022a. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35: 5775–5787.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022b. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*.

OpenAI. 2022. ChatGPT,https://openai.com/ blog/chatgpt/.

Oreshkin, B. N.; López, P. R.; and Lacoste, A. 2018. TADAM: Task dependent adaptive metric for improved few-shot learning. In *NeurIPS*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Petrovich, M.; Black, M. J.; and Varol, G. 2021. Action-Conditioned 3D Human Motion Synthesis with Transformer VAE. In *International Conference on Computer Vision (ICCV)*.

Petrovich, M.; Black, M. J.; and Varol, G. 2022. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*.

Plappert, M.; Mandery, C.; and Asfour, T. 2016. The KIT Motion-Language Dataset. *Big Data*, 4(4): 236–252.

Plappert, M.; Mandery, C.; and Asfour, T. 2018. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems*, 109: 13–26.

Raab, S.; Leibovitch, I.; Li, P.; Aberman, K.; Sorkine-Hornung, O.; and Cohen-Or, D. 2022. MoDi: Unconditional Motion Synthesis from Diverse Data. *arXiv preprint arXiv:2206.08010*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022a. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022b. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *CVPR*.

Samuel, D.; Ben-Ari, R.; Raviv, S.; Darshan, N.; and Chechik, G. 2023. It is all about where you start: Text-to-image generation with seed selection. *arXiv preprint arXiv:2304.14530*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021.

Tang, Z.; Rybin, D.; and Chang, T.-H. 2023. Zeroth-Order Optimization Meets Human Feedback: Provable Learning via Ranking Oracles. *arXiv preprint arXiv:2303.03751*.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Bermano, A. H.; and Cohen-Or, D. 2022. Human Motion Diffusion Model. *arXiv preprint arXiv:2209.14916*.

Vinyals, O.; Blundell, C.; Lillicrap, T.; kavukcuoglu, k.; and Wierstra, D. 2016. Matching Networks for One Shot Learning. In *NeurIPS*.

Wu, X.; Sun, K.; Zhu, F.; Zhao, R.; and Li, H. 2023. Better Aligning Text-to-Image Models with Human Preference. *arXiv preprint arXiv:2303.14420*.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. *arXiv preprint arXiv:2304.05977*.

Yang, Z.; Wang, J.; and Zhu, Y. 2022. Few-Shot Classification with Contrastive Learning. In *ECCV*.

Ye, H.-J.; Hu, H.; Zhan, D.-C.; and Sha, F. 2020. Few-Shot Learning via Embedding Adaptation With Set-to-Set Functions. In *CVPR*.

Yu, Z.; Huang, S.; Fang, C.; Breckon, T. P.; and Wang, J. 2023. ACR: Attention Collaboration-based Regressor for Arbitrary Two-Hand Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12955–12964.

Zhang, J.; Huang, S.; Tu, Z.; Chen, X.; Zhan, X.; Yu, G.; and Shan, Y. 2023a. TapMo: Shape-aware Motion Generation of Skeleton-free Characters. *arXiv preprint arXiv:2310.12678*.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2022a. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. *arXiv preprint arXiv:2208.15001*.

Zhang, R.; Zhang, W.; Fang, R.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2022b. Tip-Adapter: Training-free Adaption of CLIP for Few-shot Classification. In *ECCV*.

Zhang, S.; Yang, X.; Feng, Y.; Qin, C.; Chen, C.-C.; Yu, N.; Chen, Z.; Wang, H.; Savarese, S.; Ermon, S.; et al. 2023b. HIVE: Harnessing Human Feedback for Instructional Visual Editing. *arXiv preprint arXiv:2303.09618*.